Data Mining Case Studies

Proceedings of the Third International Workshop on Data Mining Case Studies

held at the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining in Paris France

Edited by

Peter van der Putten, Chordiant Software
Gabor Melli, Simon Fraser University
Brendan Kitts, Microsoft

ISBN 978-1-60558-674-8

The Association for Computing Machinery, Inc.
2 Penn Plaza, Suite 701
New York, NY 10121-0701

Notice to Past Authors of ACM-Published Articles

Additional copies may be ordered prepaid from:

ACM Order Department
PO Box 11405
Church Street Station
New York, NY 10286-1405

Phone 1-800-342-6626 (USA and Canada)
+1-212-626-0500
(all other countries)
Fax: +1-212-944-1318
Email: acmhelp@acm.org

Contents

Organizers

Chairs

Peter van der Putten, Chordiant Software
Gabor Melli, PredictionWorks
Brendan Kitts, Microsoft

Program Committee

Gregory Piatetsky-Shapiro, PhD., KDNuggets
Karl Rexer, PhD., Rexer Analytics
Gang Wu, PhD. Microsoft
Dean Abbott, PhD., Abbott Analytics
Venky Ganti, PhD., Microsoft
Jing Ying Zhang, PhD. Microsoft
Arber Xu, PhD., Summba
David Duling, PhD., SAS Corporation
Kunal Punera, PhD., Yahoo! Inc.
Moninder Singh, PhD., IBM Corporation
Felix Chen, Predictionworks,
Richard Bolton, PhD., KnowledgeBase Marketing
Robert Grossman, PhD., University of Illinois at Chicago
Ed Freeman, DS-IQ
Bamshad Mobasher, PhD., DePaul University
Antonia de Medinaceli, Elder Research

Sponsors

The Association for Computing Machinery (ACM)
Chordiant
Microsoft
PredictionWorks

Participants

A. van der Zanden	*Amsterdam Police Force*
Abdellatif Bey-Temsamani	*Flanders Mechatronics*
Agusmian P. Ompusunggu	*Flanders Mechatronics*
Aishwarya Meenakshi Sundar	*U. Minnesota*
Albert Roux	*Microsoft*
Andy Motten	*Flanders Mechatronics*
Balaji Krhnapuram	*Siemens*
Brendan Kitts	*Microsoft*
C. Hue	*GFI Informatique*
Carlos Ordonez	*U. Houston*
Eric Bjorklund	*Minnesota Dept. of Revenue*
Glenn Fung	*Siemens*
Greg Tschida	*Minnesota Dept. of Revenue*
H. T. Roos	*Amsterdam Police Force*
Jaideep Srivastava	*U. Minnesota*
Jaideep Srivastava	*U. Of Minnesota*
Javier Garc´ıa-Garc´ıa	*UNAM*
Jinbo Bi	*Siemens*
Jing Ying Zhang	*Microsoft*
Kamran Kanany	*Microsoft*
Kuo-Wei Hsu	*U. Of Minnesota*
Marc Engels	*Flanders Mechatronics*
Matthew Rice	*Microsoft*
Michael J. Rote	*Teradata*
Murat Dundar	*Siemens*
Nishith Pathak	*U. Of Minnesota*
Pavel Brusilovsky	*BI Solutions*
R. Bharat Rao	*Siemens*
R. van der Veer	*Sentient*
Roger Longbotham	*Microsoft*
Romer Rosales	*Siemens*
Ron Kohavi	*Microsoft*
Ron Mills	*Microsoft*
Sriram Krishnan	*Siemens*
Steve Vandenplas	*Flanders Mechatronics*
Thiery Vallaud	*Socio Logiciels*
Thomas Crook	*Microsoft*
V. Lemaire	*Orange Labs*
Vikas Raykar	*Siemens*

The Data Mining Case Studies Workshop

From its inception the field of Data Mining has been guided by the need to solve practical problems. Yet few articles describing working, end-to-end, real-world case studies exist in our literature. Success stories can capture the imagination and inspire researchers to do great things. The benefits of good case studies include:

1. Education: Success stories help to build understanding.
2. Inspiration: Success stories inspire future data mining research.
3. Public Relations: Applications that are socially beneficial, and even those that are just interesting, help to raise awareness of the positive role that data mining can play in science and society.
4. Problem Solving: Success stories demonstrate how whole problems can be solved. Often 90% of the effort is spent solving non-prediction algorithm related problems.
5. Connections to Other Scientific Fields: Completed data mining systems often exploit methods and principles from a wide range of scientific areas. Fostering connections to these fields will benefit data mining academically, and will assist practitioners to learn how to harness these fields to develop successful applications.

The *Data Mining Case Studies Workshop* was established in 2005 to showcase the very best in data mining case studies. We also established that *Data Mining Practice Prize* to attract the best submissions, and to provide an incentive for commercial companies to come into the spotlight.

This first workshop was followed up by a *SIGKDD Explorations Special Issue on Real-world Applications of Data Mining* edited by Osmar Zaiane in 2006 and Data Mining Case Studies 2007 at KDD2009.

It is our pleasure to continue the work of highlighting significant industrial deployments with the *Third Data Mining Case Studies Workshop* in 2009.

Like its predecessors, Data Mining Case Studies 2009 has highlighted data mining implementations that have been responsible for a significant and measurable improvement in business operations, or an equally important scientific discovery, or some other benefit to humanity. Data Mining Case Studies papers were allowed greater latitude in (a) range of topics - authors may touch upon areas such as optimization, operations research, inventory control, and so on, (b) page length - longer submissions are allowed, (c) scope - more complete context, problem and solution descriptions will be encouraged, (d) prior publication - if the paper was published in part elsewhere, it may still be considered if the new article is substantially more detailed, (e) novelty - often successful data mining practitioners utilize well established techniques to achieve successful implementations and allowance for this will be given.

The Data Mining Practice Prize

Introduction

The Data Mining Practice Prize is awarded to work that has had a significant and quantitative impact in the application in which it was applied, or has significantly benefited humanity. All papers submitted to Data Mining Case Studies will be eligible for the Data Mining Practice Prize, with the exception of members of the Prize Committee. Eligible authors consent to allowing the Practice Prize Committee to contact third parties and their deployment client in order to independently validate their claims.

Award

Winners and runners up receive an impressive array of honors including

a. Plaque awarded at the KDD conference General Session on August 12th 2007.
b. Prize money comprising $500 for first place, $300 for second place, $200 for third place.
c. Winner announcements to be published in the journal SIGKDD Explorations
d. Awards Dinner with organizers and prize winners.

We wish to thank ACM for making our competition and workshop possible.

The Association for Computing Machinery

The Association for Computing Machinery (ACM) is an international scientific and educational organization dedicated to advancing the arts, sciences, and applications of information technology. With a world-wide membership ACM is a leading resource for computing professionals and students working in the various fields of Information Technology, and for interpreting the impact of information technology on society.

ACM is the world's oldest and largest educational and scientific computing society. Since 1947 ACM has provided a vital forum for the exchange of information, ideas, and discoveries. Today, ACM serves a membership of computing professionals and students in more than 100 countries in all areas of industry, academia, and government. ACM's 34 Special Interest Groups (SIGs) address the varied needs of today’s IT and computing professionals, including computer graphics, human interfaces, artificial intelligence, data mining, mobile communications, computer education, software engineering, and programming language. Each SIG is organized around specific activities that best serve its practitioner and research-based constituencies. Many SIGs sponsor leading conferences and workshops, produce newsletters and publications, and support email forums for information exchange. ACM can be found on the web at http://www.acm.org

Online Experimentation at Microsoft

Ron Kohavi
ronnyk@microsoft.com

Thomas Crook
tcrook@microsoft.com

Roger Longbotham
rogerlon@microsoft.com

Microsoft, Experimentation Platform, One Microsoft Way, Redmond, WA 98052

ABSTRACT

Knowledge Discovery and Data Mining techniques are now commonly used to find novel, potentially useful, patterns in data. Most KDD applications involve post-hoc analysis of data and are therefore mostly limited to the identification of correlations. Recent seminal work on Quasi-Experimental Designs (Jensen, et al., 2008) attempts to identify causal relationships. Controlled experiments are a standard technique used in multiple fields. Through randomization and proper design, experiments allow establishing causality scientifically, which is why they are the gold standard in drug tests. In software development, multiple techniques are used to define product requirements; controlled experiments provide a way to assess the impact of new features on customer behavior. The Data Mining Case Studies workshop calls for describing completed implementations related to data mining. Over the last three years, we built an experimentation platform system (ExP) at Microsoft, capable of running and analyzing controlled experiments on web sites and services. The goal was to accelerate innovation through trustworthy experimentation and to enable a more scientific approach to planning and prioritization of features and designs (Foley, 2008). Along the way, we ran many experiments on over a dozen Microsoft properties and had to tackle both technical and cultural challenges. We previously surveyed the literature on controlled experiments and shared technical challenges (Kohavi, et al., 2009). This paper focuses on problems not commonly addressed in technical papers: cultural challenges, lessons, and the ROI of running controlled experiments.

1. INTRODUCTION

We're here to put a dent in the universe.
Otherwise why else even be here?
-- Steve Jobs

On Oct 28, 2005, Ray Ozzie, Microsoft's Chief Technical Officer at the time, wrote *The Internet Services Disruption* memo(Ray Ozzie, 2005). The memo emphasized three key tenets that were driving a fundamental shift in the landscape: (i) The power of the advertising-supported economic model; (ii) the effectiveness of a new delivery and adoption model (discover, learn, try, buy, recommend); and (iii) the demand for compelling, integrated user experiences that "just work." Ray wrote that the "web is fundamentally a self-service environment, and it is critical to design websites and product 'landing pages' with sophisticated closed-loop measurement and feedback systems… This ensures that the most effective website designs will be selected…" Several months after the memo, the first author of this paper, Ron Kohavi, proposed building an Experimentation Platform at Microsoft. The platform would enable product teams to run controlled experiments.

The Workshop on Data Mining Case Studies calls for papers that "describe a completed implementation" that are "guided by the need to solve practical problems." In this paper, we intentionally avoid covering the technical aspects of controlled experiments, as these were covered elsewhere (Kohavi, et al., 2009)—rather the paper focuses on the cultural challenges, lessons, and the ROI of running controlled experiments through real examples. Over the last three years, we built an experimentation platform system (ExP) at Microsoft, capable of running and analyzing controlled experiments on web sites and services. Experiments ran on 18 Microsoft properties, including MSN home pages in several countries (e.g., www.msn.com, uk.msn.com), MSN Money, MSN Real Estate, www.microsoft.com, , support.microsoft.com, Office Online, several marketing sites, and Windows Genuine Advantage.

The "story" we tell should **inspire** others to use controlled experiments, whether by implementing their own system or using a 3rd party system. The humbling statistics we share about the percentage of ideas that pass all the internal evaluations, get implemented, and fail to improve the metrics they were designed to improve are **humbling**. The cultural challenges we faced in deploying a methodology that is foreign to many classical Microsoft teams should help others foster similar cultural changes in their organizations. We share stories on **education** campaigns we ran and how we raised awareness in a large company with close to 100,000 employees. We ran numerous controlled experiments on a wide variety of sites and analyzed the data using statistical and machine learning techniques. Real-world examples of experiments open people's eyes as to the potential and the return-on-investment. In this paper we share several interesting examples that show the power of controlled experiments to improve sites, establish best practices, and resolve debates with

DMCS'09, June 28, Paris, France.
Copyright 2009 ACM 978-1-60558-674-8…$5.00.

data rather than deferring to the HIghest-Paid-Person's Opinion (HiPPO) or to the loudest voice.

Our mission at the Experimentation Platform team is to accelerate software innovation through trustworthy experimentation. We have made a small dent in Microsoft's universe and would like to share the learnings so you can do the same in yours.

In Section 2, we briefly review the concept of controlled experiments. In Section 3, we describe the progress of experimentation at Microsoft over the last three years. In Section 4, we look at successful applications of experiments that help motivate the rest of the paper. In Section 5, we share some humbling statistics about the success and failure of ideas. In Section 6, we review the Application Implementation Continuum and discuss the sweet-spot for experimentation. Section 7 reviews the cultural challenges we faced and how we dealt with them. We conclude with a summary. Lessons and challenges are shared throughout the paper.

2. Controlled Experiments

> *It's hard to argue that Tiger Woods is pretty darn good at what he does. But even he is not perfect. Imagine if he were allowed to hit four balls each time and then choose the shot that worked the best. Scary good.*
> -- *Michael Egan, Sr. Director, Content Solutions, Yahoo*

In the simplest controlled experiment, often referred to as an A/B test, users are randomly exposed to one of two variants: Control (A), or Treatment (B) as shown in Figure 1 (Kohavi, et al., 2009; Box, et al., 2005; Holland, et al., 2005; Eisenberg, et al., 2008). The key here is "random." Users cannot be distributed "any old which way" (Weiss, 1997); no factor can influence the decision.

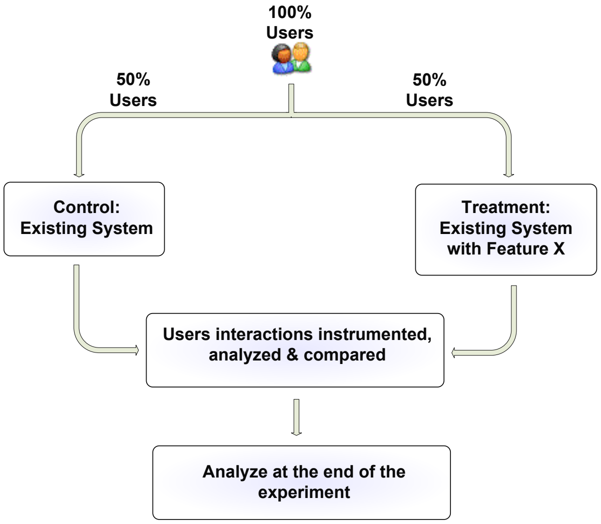

100% Users
50% Users
50% Users
Control: Existing System
Treatment: Existing System with Feature X
Users interactions instrumented, analyzed & compared
Analyze at the end of the experiment

Figure 1: High-level flow for an A/B test

Based on observations collected, an Overall Evaluation Criterion (OEC) is derived for each variant (Roy, 2001). The OEC is sometimes referred to as a Key Performance Indicator (KPI) or a metric. In statistics this is often called the Response or Dependent Variable.

If the experiment was designed and executed properly, the only thing consistently different between the two variants is the change between the Control and Treatment, so any statistically significant differences in the OEC are the result of the specific change, establishing causality (Weiss, 1997 p. 215).

Common extensions to the simple A/B tests include multiple variants along a single axis (e.g., A/B/C/D) and multivariable tests where the users are exposed to changes along several axes, such as font color, font size, and choice of font.

For the purpose of this paper, the statistical aspects of controlled experiments, such as design of experiments, statistical tests, and implementation details are not important. We refer the reader to the paper *Controlled experiments on the web: survey and practical guide* (Kohavi, et al., 2009) for more details.

3. Experimentation at Microsoft

> *The most important and visible outcropping of the action bias in the excellent companies is their willingness to try things out, to experiment.*
> *There is absolutely no magic in the experiment…*
> *But our experience has been that most big institutions have forgotten how to test and learn. They seem to prefer analysis and debate to trying something out, and they are paralyzed by fear of failure, however small.*
> *-- Tom Peters and Robert Waterman,*
> *In Search of Excellence*

In 2005, when Ron Kohavi joined Microsoft, there was little use of controlled experiments at Microsoft outside Search and the MSN US home page. Only a few experiments ran as one-off "split tests" in Office Online and on microsoft.com. The internet Search organization had basic infrastructure called "parallel flights" to expose users to different variants. There was appreciation for the idea of exposing users to different variant, and running content experiments was even patented (Cohen, et al., 2000). However, most people did not test results for statistical significance. There was little understanding of the statistics required to assess whether differences could be due to chance. We heard that there is no need to do statistical tests because "even election surveys are done with a few thousand people" and Microsoft's online samples were in the millions. Others claimed that there was no need to use sample statistics because all the traffic was included, and hence the entire population was being tested.

In March 2006, the Experimentation Platform team (ExP) was formed as a small incubation project. By end of summer we were seven people: three developers, two program managers, a tester, and a general manager. The team's mission was dual-pronged:

1. Build a platform that is easy to integrate
2. Change the culture towards more data-driven decisions

In the first year, a proof-of-concept was done by running two simple experiments. In the second year, we focused on advocacy and education. More integrations started, yet it was a "chasm" year and only eight experiments ultimately ran successfully. In the third year, adoption of ExP, the Experimentation Platform, grew significantly, with a new experiment starting about once a week. The search organization has evolved their parallel flight infrastructure to use statistical techniques and is executing a large number of experiments independent of the Experimentation Platform, but using the same statistical evaluations. Over 15 web properties at Microsoft ran at least one experiment with ExP, and several more properties are adopting the platform.

Testimonials from ExP adopters show that groups are seeing the value. The purpose of sharing the following testimonials isn't self-promotion, but rather to share actual responses showing that cultural changes are happening and ExP partners are finding it highly beneficial to run controlled experiments. Getting to this point required a lot of work and many lessons that we will share in the following sections. Below are some testimonials.

- I'm thankful everyday for the work we've done together. The results of the experiment were in some respect counter intuitive. They completely changed our feature prioritization. It dispelled long held assumptions about <area>. Very, very useful.
- The Experimentation Platform is essential for the future success of all Microsoft online properties… Using ExP has been a tremendous boon for <team name>, and we've only just begun to scratch the surface of what that team has to offer.
- For too long in <team name>, we have been implementing changes on <online site> based on opinion, gut feeling or perceived belief. It was clear that this was no way to run a successful business…Now we can release modifications to the page based purely on statistical data
- We are partnering with the ExP...and are planning to make their system a core element of our mission

The next section reviews several successful applications of controlled experiments.

4. Applications of Controlled Experiments at Microsoft

One of the best ways to convince others to adopt an idea is to show examples that provided value to others, and carry over to their domain. In the early days, publicly available examples were hard to find. In this section we share recent Microsoft examples.

4.1 Which Widget?

The MSN Real Estate site (http://realestate.msn.com) wanted to test different designs for their "Find a home" widget. Visitors to this widget were sent to Microsoft partner sites from which MSN Real estate earns a referral fee. Six different designs, including the incumbent, were tested.

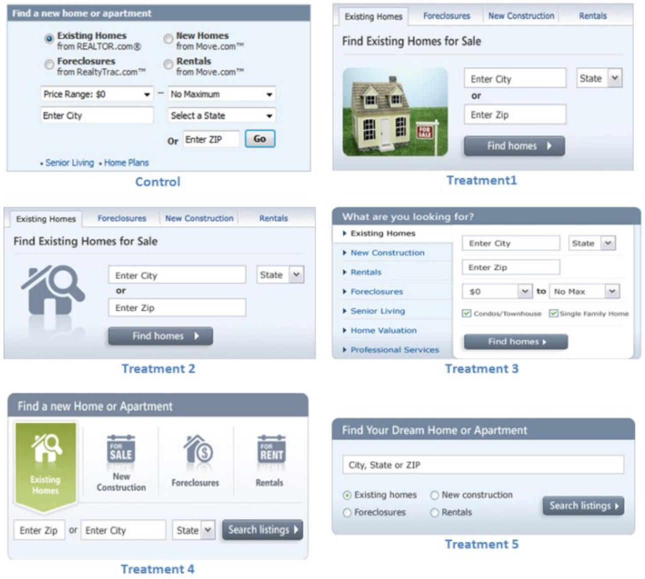


Figure 2 Widgets tested for MSN Real Estate

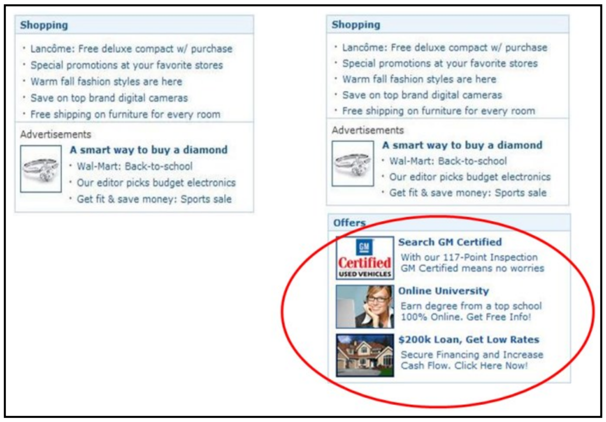
A "contest" was run by Zaaz, the company that built the creative designs, prior to running an experiment with each person guessing which variant will win. Only three out of 21 people guessed the winner, and the three were from the ExP team (prior experience in experiments seems to help). All three said, among other things, that they picked Treatment 5 because it was simpler. One person said it looked like a search experience.

The winner, Treatment 5, increased revenues from referrals by almost 10% (due to increased clickthrough). The Return-On-Investment (ROI) was phenomenal.

4.2 MSN Home Page Ads

A critical question that many site owners face is how many ads to place. In the short-term, increasing the real-estate given to ads can increase revenue, but what will it do to the user experience, especially if these are non-targeted ads? The tradeoff between increased revenue and the degradation of the end-user experience is a tough one to assess, and that's exactly the question that the MSN home page team at Microsoft faced.

The MSN home page is built out of modules. The Shopping module is shown on the right side of the page above the fold. The proposal was to add three offers right below it, as shown in Figure 3, which meant that these offers would show up below the fold for most users. The Display Ads marketing team estimated they could generate tens of thousands of dollars per day from these additional offers.



Shopping

- Lancôme: Free deluxe compact w/ purchase
- Special promotions at your favorite stores
- Warm fall fashion styles are here
- Save on top brand digital cameras
- Free shipping on furniture for every room

Advertisements

A smart way to buy a diamond

- Wal-Mart: Back-to-school
- Our editor picks budget electronics
- Get fit & save money: Sports sale

Shopping

- Lancôme: Free deluxe compact w/ purchase
- Special promotions at your favorite stores
- Warm fall fashion styles are here
- Save on top brand digital cameras
- Free shipping on furniture for every room

Advertisements

A smart way to buy a diamond

- Wal-Mart: Back-to-school
- Our editor picks budget electronics
- Get fit & save money: Sports sale

Offers

Search GM Certified
With our 117-Point Inspection
GM Certified means no worries

Online University
Earn degree from a top school
100% Online. Get Free Info!

$200k Loan, Get Low Rates
Secure Financing and Increase
Cash Flow. Click Here Now!

Figure 3: MSN Home Page Proposal.
Left: Control, Right: proposed Treatment

The interesting challenge here is how to compare the ad revenue with the "user experience." We refer to this problem as the OEC, or the Overall Evaluation Criterion. In this case, we decided to see if page views and clicks decreased, and assign a monetary value to each. (No statistically significant change was seen in visit frequency for this experiment.) Page views of the MSN home page have an assigned value based on ads; clicks to destinations from the MSN home page were estimated in two ways:

1. Monetary value that the destination property assigned to a click from the MSN home page. These destination properties are other sites in the MSN network. Such a click generates a visit to an MSN property (e.g., MSN Autos or MSN Money), which results in multiple page views.
2. The cost paid to search engines for a click that brings a user to an MSN property but not via the MSN home page (Search Engine Marketing). If the home page is driving less traffic to the properties, what is the cost of regenerating the "lost" traffic?

As expected, the number from #2 (SEM) was higher, as additional value beyond direct monetization is assigned to a click that may represent a new user, but the numbers were close enough to get agreement on the monetization value to use.

A controlled experiment was run on 5% of the MSN US home page users for 12 days. Clickthrough rate decreased by 0.35% (relative change), and the result was statistically significant. Page views per user-day decreased 0.35%, again a result that was highly statistically significant.

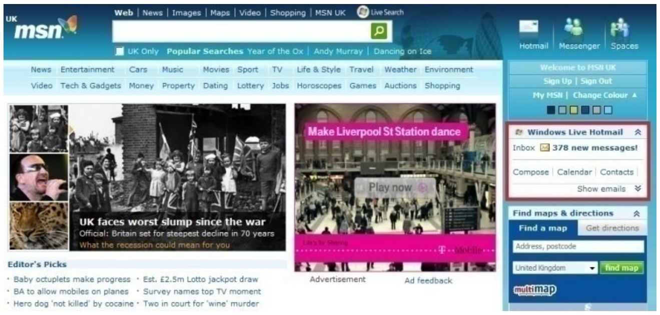
Translating the lost clicks to their monetary value, it was higher than the expected ad revenue. The estimated loss, had this feature been deployed, was millions of dollars per year.

4.3 Open in Place or in a Tab?

When a visitor comes to the MSN home page and they are recognized as having a Hotmail account, a small Hotmail convenience module is displayed. Prior to the experiment, if they clicked on any link in the module, Hotmail would open in the same tab/window as the MSN home page, replacing it. The MSN team wanted to test if having Hotmail open in a new tab/window would increase visitor engagement on the MSN because visitors will reengage with the MSN home page if it was still present when they finished reading e-mail.

The experiment included one million visitors who visited the MSN UK home page, shown in Figure 4 and clicked on the Hotmail module over a 16 day period. For those visitors the number of clicks per user on the MSN homepage increased 8.9% and the percentage of visitors who clicked on the homepage after opening Hotmail increased 6.6%. This change resulted in significant increase in user engagement and was implemented in the UK and in the US shortly after the experiment was completed.

One European site manager wrote: "This report came along at a really good time and was VERY useful. I argued this point to my team and they all turned me down. Funny, now they have all changed their minds."



UK
msn
Web | News | Images | Maps | Video | Shopping | MSN UK
Live Search
UK Only
Popular Searches Year of the Ox | Andy Murray | Dancing on Ice
Hotmail
Messenger
Spaces
News Entertainment Cars Music Movies Sport TV Life & Style Travel Weather Environment
Video Tech & Gadgets Money Property Dating Lottery Jobs Horoscopes Games Auctions Shopping
Welcome to MSN UK
Sign Up | Sign Out
My MSN | Change Colour
Make Liverpool St Station dance
Play now
T Mobile
Advertisement
Ad feedback
Windows Live Hotmail
Inbox
Compose | Calendar | Contacts
Show emails
Find maps & directions
Find a map
Get directions
Address, postcode
United Kingdom
find map
multimap

UK faces worst slump since the war
Official: Britain set for steepest decline in 70 years
What the recession could mean for you

Editor's Picks

- Baby octuplets make progress
- BA to allow mobiles on planes
- Hero dog 'not killed' by cocaine
- Est. £2.5m Lotto jackpot draw
- Survey names top TV moment
- Two in court for 'wine' murder

378 new messages!

Figure 4 Hotmail Module highlighted in red box

4.4 Personalize Support?

The support site for Microsoft (http://support.microsoft.com) has a section near the top of the page that has answers to the most common issues. The support team wanted to test whether making those answers more specific to the user would be beneficial. In the Control variant, users saw the top issues across all segments. In the Treatment, users saw answers specific to their particular browser and operating system. The OEC was the click-through rate (CTR) on the links to the section being tested. The CTR for the treatment was over 50% higher the Control, proving the value of simple personalization.

This experiment ran as a proof of concept with manually generated issue lists. The support team now plans to add this functionality to the core system.

4.5 Pre-Roll or Post-Roll Ads?

Most of us have an aversion to ads, especially if they require us to take action to remove them or if they cause us to wait for our content to load. We ran a test with MSN Entertainment and Video Services (http://video.msn.com) where the Control had an ad that ran prior to the first video and the Treatment post-rolled the ad, after the content. The primary business question the site owners had was "Would the loyalty of users increase enough in the

Treatment to make up for the loss of revenue from not showing the ad up front?" We used the first two weeks to identify a cohort of users that was then tracked over the next six weeks. The OEC was the return rate of users during this six week period. We found that the return rate increased just over 2% in the Treatment, not enough to make up for the loss of ad impressions, which dropped more than 50%.

5. Most Ideas Fail to Show Value

> *The fascinating thing about intuition is that a fair percentage of the time it's fabulously, gloriously, achingly wrong*
>
> -- *John Quarto-vonTivadar, FutureNow*

It is humbling to see how bad experts are at estimating the value of features (us included). Every feature built by a software team is built because *someone* believes it will have value, yet many of the benefits fail to materialize. Avinash Kaushik, author of *Web Analytics: An Hour a Day*, wrote in his Experimentation and Testing primer (Kaushik, 2006) that "80% of the time you/we are wrong about what a customer wants." In Do It Wrong Quickly (Moran, 2007 p. 240), the author writes that Netflix considers 90% of what they try to be wrong. Regis Hadiaris from Quicken Loans wrote that "in the five years I've been running tests, I'm only about as correct in guessing the results as a major league baseball player is in hitting the ball. That's right - I've been doing this for 5 years, and I can only "guess" the outcome of a test about 33% of the time!" (Moran, 2008).

We in the software business are not unique. QualPro, a consulting company specializing in offline multi-variable controlled experiments, tested 150,000 business improvement ideas over 22 years and reported that 75 percent of important business decisions and business improvement ideas either have no impact on performance or actually hurt performance (Holland, et al., 2005). In the 1950s, medical researchers started to run controlled experiments: "a randomized controlled trial called for physicians to acknowledge how little they really knew, not only about the treatment but about disease" (Marks, 2000 p. 156). In *Bad Medicine: Doctors Doing Harm Since Hippocrates,* David Wootton wrote that "For 2,400 years patients have believed that doctors were doing them good; for 2,300 years they were wrong." (Wooton, 2007). Doctors did bloodletting for hundreds of years, thinking it had a positive effect, not realizing that the calming effect was a side effect that was unrelated to the disease itself. When President George Washington was sick, doctors extracted about 35%-50% of his blood over a short period, which inevitably led to preterminal anemia, hypovolemia, and hypotension. The fact that he stopped struggling and appeared physically calm shortly before his death was probably due to profound hypotension and shock (Kohavi, 2008). In an old classic, Scientific Advertising (Hopkins, 1923 p. 23), the author writes that "[In selling goods by mail] false theories melt away like snowflakes in the sun... One quickly loses his conceit by learning how often his judgment errs--often nine times in ten."

When we first shared some of the above statistics at Microsoft, many people dismissed them. Now that we have run many experiments, we can report that Microsoft is no different. Evaluating well-designed and executed experiments that were designed to improve a key metric, **only about one-third were successful at improving the key metric!**

There are several important lessons here

1. Avoid the temptation to try and build optimal features through extensive planning without early testing of ideas. As Steve Kurg write: "The key is to start testing early (it's really never too early) and test often, at each phase of Web development" (Krug, 2005).
2. Experiment often. Because under objective measures most ideas fail to improve the key metrics they were designed to improve, it is important to increase the rate of experimentation and lower the cost to run experiments. Mike Moran phrased this lesson as follows: "You have to kiss a lot of frogs to find one prince. So how can you find your prince faster? By finding more frogs and kissing them faster and faster" (Moran, 2007).
3. A failure of an experiment is not a mistake: learn from it. Badly executed experiments are mistakes (Thomke, 2003), but knowing that an idea fails provides value can save a lot of time. It is well known that finding an error in requirements is 10 to 100 times cheaper than changing features in a finished product (McConnell, 2004). Use experimentation with software prototypes to verify requirements in the least costly phase of the software development lifecycle. Think of how much effort can be saved by building an inexpensive prototype and discovering that you do not want to build the production feature at all! Such insights are surprisingly common in organizations that experiment. The ability to fail fast and try multiple ideas is the main benefit of a customer-driven organization that experiments frequently. We suggest that development teams launch prototype features regularly, and extend them, making them more robust, and then fully deploy them only if they prove themselves useful. This is a challenging proposition for organizations whose development culture has been to "do it right the first time".
4. Try radical ideas and controversial ideas. In (Kohavi, et al., 2009), we described the development of Behavior-Based Search at Amazon, a highly controversial idea. Early

experiments by an intern showed the surprisingly strong value of the feature, which ultimately helped improve Amazon's revenue by 3%, translating into hundreds of millions of dollars in incremental sales. Greg Linden at Amazon created a prototype to show personalized recommendations based on items in the shopping cart (Linden, 2006). Linden notes that "a marketing senior vice-president was dead set against it," claiming it will distract people from checking out. Greg was "forbidden to work on this any further." Nonetheless, Greg ran a controlled experiment and the rest is history: the feature was highly beneficial. Multiple sites have copied cart recommendations. Sir Ken Robinson made this point eloquently when he said: "If you're not prepared to be wrong, you will not come up with anything original" (Robinson, 2006).

6. The Sweet-Spot for Experiments

> *When optimizing for conversion, we often find clients trying to improve engine torque while ignoring a flat tire*
> *-- Bryan Eisenber and John Quarto-vonTivadar in Always Be Testing* (2008)

We now describe software development environments that present the sweet spot for controlled experiments along a continuum of product types: the Application Implementation Continuum, shown in Figure 5.

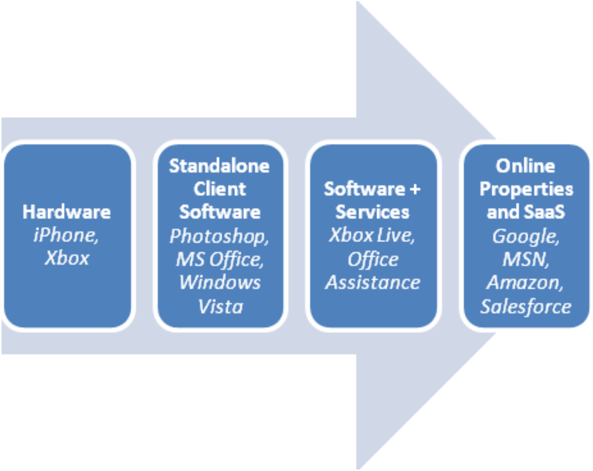

Hardware
iPhone, Xbox
Standalone Client Software
Photoshop, MS Office, Windows Vista
Software + Services
Xbox Live, Office Assistance
Online Properties and SaaS
Google, MSN, Amazon, Salesforce

Figure 5: The Application Implementation Continuum ranges from hardware devices, which are hard to change and experiment on, to online properties and SaaS, which are easy to change and experiment on.

Products on the right side of the continuum are more amenable to experimentation and agile development methodologies than those on the left. Microsoft, which had historically developed complex software products delivered on physical media, evolved development methodologies appropriate for the left side of the continuum. Younger companies, such as Amazon and Google, developed methodologies appropriate to the right side of the continuum. Below we review the ways to identify customer preferences and the ingredients necessary for running effective experiments.

6.1 Identifying Customer Preferences

Forty to sixty percent of software defects are due to errors in requirements (Wiegers, 2003). Historically, this motivated developers to increase the amount and depth of customer research carried out before design and coding commenced in earnest. Unfortunately, there are natural limits to what can be learned from this type of research:

- We cannot completely know customers' needs and usage environments before a feature is deployed. Some information about customer requirements is "sticky." i.e., it can only be discovered at the time and place where the customer uses the product (von Hippel, 1994). Beta testing can identify missing and erroneous requirements at a later stage, but such late discoveries are costly, and often arrive too late for developers to make the necessary changes.
- Customer research findings are not necessarily predictive of actual customer behavior. In the real world, customers must make tradeoffs that are not adequately captured in focus group or laboratory settings. Furthermore, what customers say in a focus group setting or a survey may not truly indicate what they prefer. A well-known example of this phenomenon occurred when Philips Electronics ran a focus group to gain insights into teenagers' preferences for boom box features. The focus group attendees expressed a strong preference for yellow boom boxes during the focus group, characterizing black boom boxes as "conservative." Yet when the attendees exited the room and were given the chance to take home a boom box as a reward for their participation, most chose black (Cross, 2005).
- Traditional customer research techniques are expensive. Teams must design usability studies and surveys, recruit participants, conduct focus groups, run user experience sessions, conduct in-depth interviews, and finally, compile,

analyze and report on the results. Most studies are done with a dozen or so participants, leading to anecdotal evidence and little statistical significance.

Thomke wrote that organizations will recognize maximal benefits from experimentation when it is used in conjunction with an "innovation system" (Thomke, 2003). Agile software development is such an innovation system.

In contrast to development on the left side of the Application Implementation Continuum, development on the right side can leverage controlled experiments and fast iterations to converge on designs that please customer relatively quickly and inexpensively. Agile development teams can deploy prototypes to web sites and services as experiments, enabling the organization to learn from the customer behavior.

Ninety percent of the time to code features is typically spent in handling the edge cases of small populations. In online environments, these can be excluded from the experiments. For example, when implementing JavaScript, the browser support matrix is enormous and compatibility testing is very time consuming. For implementing a prototype, it may be enough to support a few most common browser versions. If 80% of the users do not behave as desired, it's time to go back to the drawing board. Conversely, if there is a significant boost in metrics of interest, above what was expected, the feature should be prioritized higher and possibly expanded. The development team can iterate quickly to converge on an optimized customer experience if they only have to develop the software for a few browsers. Then, once a good solution is found, they can spend the time to roll out the new experience to the long tail.

6.2 Necessary Ingredients

Controlled experiments are not applicable everywhere. In order to use agile development with controlled experiments, several ingredients have to exist.

1. A clear objective that can be practically evaluated. Controlled experiments require an Overall Evaluation Criterion, or OEC. Organizations that have not agreed what to optimize should first get agreement on that (which is sometimes hard), but as Lewis Carroll said, "If you don't know where you are going, any road will take you there." It is important to note that for many sites the OEC must represent a long-term customer value, not a short-term gain. For example, time on site and frequency of visit are better criteria than ads clicked, which is a short term metric that will lead to short-term gains and long-term doom as the site plasters itself with ads.
2. Easy to collect data about the user behavior. With client software, user behavior is hard to track and usually requires consent. As more of the user experience moves online, it becomes easier to track user behavior since server-side logging and client-side JavaScript logging are commonly used in industry and accepted as reasonable practices.
3. Easy to change and experiment with real users. As you move along the Application Implementation Continuum, experimentation becomes easier. At the left, we have hardware devices, which are hard to change and therefore make it harder to experiment with real users outside of focus groups and prototypes (although see several great examples in *Experimentation Matters* (Thomke, 2003)). At the other extreme are online properties, such as MSN, Amazon.com, Google and eBay, and Software as Services (Saas) implementations such as Salesforce.com, which are very easy to change. Iterative experiments with real users are easy to carry out at this end of the continuum. In between the extremes, we have standalone client software and Software+Services. In the former, experiments have to be planned and the opportunity to experiment may be limited to beta cycles. For the latter, experimentation capability has to be properly baked into the client so that server-side changes can be made to impact the client. For example, assistance/help for Microsoft Office products sends user queries to a service and receives articles to display. Software for hardware devices with connectivity can also use experimentation
4. Sufficient users exist. Very small sites or products with no customer base cannot use experimentation, but such sites typically have a key idea to implement and they need quick feedback after going live. Because new sites are aiming for big improvements, the number of users needed to detect the desired effects can be relatively small (e.g., thousands of users). Large sites, which are typically better optimized, benefit even from small improvements and therefore need many

customers to increase their experiment's sensitivity level.

Most non-trivial online properties meet, or could meet, the necessary ingredients for running an agile development process based on controlled experiments. Many implementations of software+services could also meet the requirements relatively easily. For example, we are currently working with the Microsoft Zune team to use experiments to find the best music recommendation algorithms.

7. Cultural Challenges

> *There were three ways to get fired at Harrah's:*
> *steal, harass women, or institute a program or policy*
> *without first running an experiment*
>
> *-- Gary Loveman, quoted in*
> *Hard Facts* (Pfeffer, et al., 2006 p. 15)

Microsoft clearly knows how to build and ship classical "shrink-wrapped" or "client" software. There have been over 120 million Office licenses sold since the launch of Office 2007 to July 2008 (Elop, 2008). Office releases are well planned and executed over three to four years. But in the evolving world of the web and services, there is a different way of "shipping" software. Mark Lucovsky described it well (Lucovsky, 2005):

> *When an Amazon engineer fixes a minor defect, makes something faster or better, makes an API more functional and complete, how do they "ship" that software to me? What is the lag time between the engineer completing the work, and the software reaching its intended customers? A good friend of mine investigated a performance problem one morning, he saw an obvious defect and fixed it. His code was trivial, it was tested during the day, and rolled out that evening. By the next morning millions of users had benefited from his work*

Websites and services can iterate faster because shipping is much easier. In addition, getting implicit feedback from users through online controlled experiments is something that could not be done easily with shrink-wrapped products, but can easily be done in online settings. It is the combination of the two that can make a big difference in the development culture. Instead of doing careful planning and execution, one can try many things and evaluate their value with real customers in near-real-time.

Linsky and Heifetz in *Leadership on the Line* (Linsky, et al., 2002) describe *Adaptive Challenges* as those that are not amenable to standard operating procedures and where the technical know-how and procedures are not sufficient to address the challenge. We faced several non-technical challenges that are mostly cultural. It is said that the only population that likes change consists of wet babies. We share the things we did that were useful to nudge the culture toward an experimentation culture.

7.1 Education and Awareness

People have different notions of what "experiment" means, and the word "controlled" in front just doesn't help to ground it. In 2005, no Microsoft groups that we are aware of ran proper controlled experiments with statistical tests. In the few groups that ran "flights," as they were called, traffic was split into two or more variants, observations were collected and aggregated, but no tests were done for statistical significance, nor were any power calculations done to determine how large a sample was needed and how long experiments should run. This led to overfitting the noise in some cases.

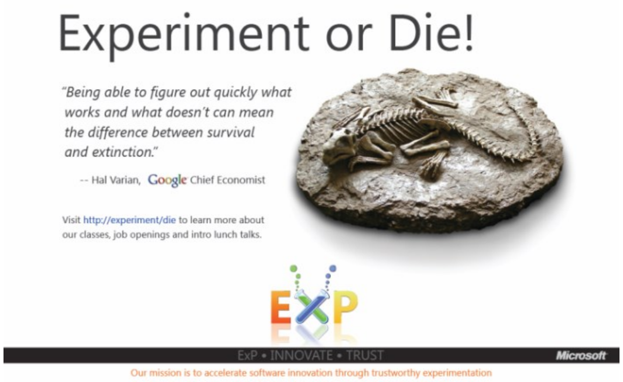
One of our first challenges was education: getting people to realize that what they have been doing was insufficient. Upton Sinclair wrote that "It is difficult to get a man to understand something when his salary depends upon his not understanding it." People have found it hard to accept that many of their analyses, based on raw counts but no statistics, have been very "noisy," to put it mildly.

We started teaching a monthly one-day class on statistics and design of experiments. Initially, we couldn't fill the class (of about 20), but after a few rounds interest grew. To date more than 500 people at Microsoft have attended our class, which now commonly has a waiting list.

The challenge is ongoing, of course; we still find people who test ideas by comparing counts from analytical reporting tools without controlling for many factors and without running statistical tests.

We wrote the KDD paper *Practical Guide to Controlled Experiments on the Web* (Kohavi, et al., 2007) in our first year to help give the team credibility as "experts" in the field. The paper is now part of the class reading for several classes at Stanford University (CS147, CS376), USCD (CSE 291), and at the University of Washington (CSEP 510). It is getting referenced by dozens of articles and some recent book, such as King (2008).

We put posters across the Microsoft campus with examples of A/B tests or with quotations. One of the more controversial and successful ones was "Experiment or Die!," shown in Figure 6, with a fossil and a quotation from Hal Varian at Google.



Experiment or Die!

"Being able to figure out quickly what works and what doesn't can mean the difference between survival and extinction."

-- Hal Varian, Google Chief Economist

Visit http://experiment/die to learn more about our classes, job openings and intro lunch talks.

EXP
ExP • INNOVATE • TRUST
Microsoft
Our mission is to accelerate software innovation through trustworthy experimentation

Figure 6: Example of a Poster: Experiment or Die!

We ate our own dog food and A/B tested our posters by creating two designs for each promotion. Each design was tagged with a unique URL offering more information about our platform, services and training classes. We compared page views for each URL to determine the effectiveness of each design.

One of our most successful awa campaigns featured a HiPPO stress imprinted with our URL. HiP stands for the Highest Paid Person' Opinion (Kohavi, et al., 2007). We gave away thousands of HiPPOs at the annual Microsoft employee company meetings, in our training classes, introductory talks, and through a HiPPO FAQ web site.[1] The campaign went viral, spawning word of m awareness and even a small fan club in Microsoft India.

We created an internal Microsoft e-mail distribution list for those interested in experimentation. There are now over 700 people on the alias.

In late 2008, enough experiments started to execute across groups that we decided to share interesting results and best practices. An internal Microsoft e-mail distribution list was created for sharing results, similar to the experiments we shared earlier in this paper.

7.2 Perceived Loss of Power

Linsky and Heifetz wrote that "People do not resist change, per se. People resist loss" (Linsky, et al., 2002). Some people certainly viewed experimentation as a risk to their power and/or prestige. Some believed it threatened their job as decision makers. After all, program managers at Microsoft select the next set of features to develop. Proposing several alternatives and admitting you don't know which is best is hard. Likewise, editors and designers get paid to create a great design. In some cases an objective evaluation of ideas may fail and hurt their image and professional standing.

It is easier to declare success when the feature launches and not *if* it is liked by customers. We have heard statements such as "we know what to do. It's in our DNA," and "why don't we just do the right thing?"

This was, and still is, a significant challenge, despite the humbling statistics about the poor success rate of ideas when evaluated objectively (see Section 5).

What we found was that a great way to convince people that we are not good at predicting the outcomes of experiment is to challenge them. We created a survey with eight A/B tests, and offered a nice polo shirt for anyone who could correctly guess 6 out of 8 (the options were: A is statistically significantly better, B is statistically significantly better, or there's no statistically significant difference between them). With over 200 responses, we didn't have to hand out a single shirt! 6 out of 200 had 5 answers correct; the average was 2.3 correct answers. Humbling! At the 2008 CIKM conference (Pasca, et al., 2008), Kohavi gave an invited talk on controlled experiments and challenged the audience to predict the outcome of three actual A/B tests that ran. Out of about 150 people in the audience who stood up to the challenge, only 1 correctly guessed the outcome of two challenge questions. Note that with three options to each question, this is much worse than random (150/9 = 16 people).

7.3 Reward Systems

Lee et al. (2004) write about the mixed effects of inconsistency on experimentation in organizations. They note that management can support experimentation and highlight it as a value (normative influence), but inconsistent reward systems that punish failure lead to aversion, especially in organizations that are under constant evaluation for perfect execution.

At Microsoft, as in many other large companies, employees are evaluated based on yearly goals and commitments. Conventional wisdom is that the best goals and commitments need to be SMART: specific, measurable, attainable, realistic and timely[2]. Most goals in software development organizations at Microsoft are around "shipping" products, not about their impact on customers or key metrics. In most projects, the classical triangular tradeoff exits between features, time, and quality. Some teams, such as Microsoft Office, traditionally focused on time and quality and cut features; others focused on features and quality and delayed the release schedule. Either way, features are commonly defined by their perceived value and are prioritized by program managers. Controlled experiments and the humbling results we shared bring to question whether a-priori prioritization is as good as most people believe it is. One possible change to goal definitions is to avoid tying them to products and features, but rather tie them to key metrics, and empower the development organizations to regularly test their ideas using controlled experiments. The feature development pace will undoubtedly slow down, but more corrections will be made on the way, ultimately leading to a better customer experience in shorter time.

It is hard for us to judge whether we are making any change in people's goals; cultural changes take time and it is unlikely that we have made a dent in many people's yearly performance goals. This is an ongoing challenge worth highlighting.

7.4 Incorrect Reasons Not to Experiment

Controlled experiments are a tool that has its limitations, which we discussed in *Controlled experiments on the web: survey and practical guide* (Kohavi, et al., 2009). A recent article by Davenport (2009) points out that controlled experiments are best suited for strategy execution, not strategy formulation; they are not suited for assessing a major change in business models, a large merger or acquisition (e.g., you can't run a randomized experiments on whether Microsoft should acquire Yahoo!). We agree, of course. However, we have also heard many incorrect reasons why not to experiment and would like to address them.

1. Claim: Experimentation leads to incremental innovations.
   While it is true that one can limit experiments to trivial UI changes like choosing colors, there is no reason experiments can't be used for radical changes and non-UI changes. Amazon makes heavy use of experimentation

[1] See http://exp-platform.com/whatsahippo.aspx

[2] Guy Kawasaki in Reality Check (2008 p. 94) suggests that goals be "rathole resistant," to avoid short-term target that dead-ends. We agree, and we have emphasized the importance of setting OECs for long-term customer lifetime-value.

and its page design has evolved significantly—its first home page did not even have a search box. Multiple industry-leading innovations came from experimenting with prototypes that showed significant value and were reprioritized quickly once their value was apparent. Two such examples were described in Section 5 (item 4).
One lesson that we learned is that many of our initial examples were indeed highlighting a big difference achieved through a small UI change, something that may have solidified the thinking that experimentation is best used for small incremental changes. Now we emphasize more sophisticated examples, such as whether to show ads (Section 4.5) and backend changes (Section 4.4).

2. Claim: Team X is optimizing for something that is not measurable. Here we need to differentiate between not measurable and non-economical to measure. We believe that the former is a bad way to run a business. If you can't articulate what you're optimizing for, how can the organization determine if you are doing a good job? If you are not improving a measurable metric, perhaps the other direction is also true: no measurable change will be observable without you in the organization!
The other interpretation is more reasonable: it may be non-economical to measure the change. While this is valid at times, we would like to point to Amazon as an example of a company that did decide to measure something hard: the value of TV ads. After a 15-month-long test of TV advertising in two markets, it determined that TV ads were not a good investment and stopped them (Bezos, 2005). Is your organization avoiding experiments whose answer they would rather not know?

3. Claim: It's expensive to run experiments. Holland (2005) wrote that based on 150,000 business improvement ideas over 22 years, "there is no correlation between what people in the organization think will work and what actually does work… The lack of correlation between what people think will work and what does work has nothing to do with the level of the people in the organization who make these judgments. The experts are no better than the front-line workers or senior executives in determining which ideas will improve results." While we think Holland's sample is biased because his consulting company, QualPro, is brought in to help evaluate more controversial ideas, we do believe that people and organizations are overly confident of their ideas, and the poor success rate described in Section 5 strongly supports that. While it is expensive to experiment, it is more expensive to continue developing and supporting features that are not improving the metrics they were supposed to improve, or hurting them, and at Microsoft, these two cases account for 66% of experiments.

   The flip side is to reduce costs and develop infrastructure to lower the cost of experimentation, and that's why we embarked on building the Experimentation Platform.

8. SUMMARY

It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment[s], it's wrong.
-- Richard Feynman

Experimentation lies at the heart of every company's ability to innovate (Thomke, 2001; Thomke, 2003). Running physical experiments is relatively expensive, so companies have had to be parsimonious with the number of experiments. The electric light bulb required more than 1,000 complex experiments. In modern times, with the magic of software, experimentation is much cheaper, and the ability to test innovative ideas unprecedented.

Changing the culture at a large company like Microsoft, with over 95,000 employees, is not easy. As more software is written in the form of services and web sites, the value of running controlled experiments and getting direct feedback in near-real-time rises. In the last three years, experimentation at Microsoft grew significantly in usage, but we are only at the early stages. We presented successful applications of experimentation, the many challenges we faced and how we dealt with them, and many lessons. The humbling results we shared in Section 5 bring to question whether a-priori prioritization is as good as most people believe it is. We hope this will help readers initiate similar changes in their respective organizations so that data-driven decision making will be the norm, especially in software development for online web sites and services.

ACKNOWLEDGMENTS

We would like to thank members of the Experimentation Platform team at Microsoft, especially Randy Henne and Lisa Ambler. Special thanks to David Treadwell and Ray Ozzie; without their support the experimentation platform would not have existed.

REFERENCES

Bezos, Jeff. 2005. The Zen of Jeff Bezos. [ed.] Chris Anderson. *Wired Magazine.* January 2005, 13.01. http://www.wired.com/wired/archive/13.01/bezos.html.

Box, George E.P., Hunter, J Stuart and Hunter, William G. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery.* 2nd. s.l. : John Wiley & Sons, Inc, 2005. 0471718130.

Cohen, Jules S, Kromann, Paul K and Reeve, Thomas S. 2000. *Systems and methods for conducting internet content usage experiments. Patent 7,343,390* December 20, 2000. http://www.google.com/patents?vid=USPAT7343390.

Cross, Robert G. and Dixit, Ashutosh. 2005. Customer-centric pricing: The surprising secret for profitability. *Business Horizons.* 2005, Vol. 48, p. 488.

Davenport, Thomas H. 2009. How to Design Smart Business Experiments. *Harvard Business Review.* 2009, February.

Eisenberg, Bryan and Quarto-vonTivadar, John. 2008. *Always Be Testing: The Complete Guide to Google Website Optimizer.* s.l. : Sybex , 2008. 978-0470290637 .

Elop, Stephen. 2008. Financial Analyst Meeting 2008. *Microsoft Investor Relations.* [Online] July 24, 2008. http://www.microsoft.com/msft/speech/FY08/ElopFAM2008.mspx.

Foley, Mary-Jo. 2008. Microsoft looks to make product planning more science than art. *ZDNet: All About Microsoft.* [Online] April 16, 2008. http://blogs.zdnet.com/microsoft/?p=1342.

Holland, Charles W and Cochran, David. 2005. *Breakthrough Business Results With MVT: A Fast, Cost-Free, "Secret Weapon" for Boosting Sales, Cutting Expenses, and Improving Any Business Process* . s.l. : Wiley, 2005. 978-0471697718 .

Hopkins, Claude. 1923. *Scientific Advertising.* New York City : Crown Publishers Inc., 1923.

Jensen, David D, et al. 2008. Automatic Identification of Quasi-Experimental Designs for Discovering Causal Knowledge. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2008, pp. 372-380.

Kaushik, Avinash. 2006. Experimentation and Testing: A Primer. *Occam's Razor.* [Online] May 22, 2006. http://www.kaushik.net/avinash/2006/05/experimentation-and-testing-a-primer.html.

Kawasaki, Guy. 2008. *Reality Check: The Irreverent Guide to Outsmarting, Outmanaging, and Outmarketing Your Competition.* s.l. : Portfolio Hardcover , 2008. 978-1591842231 .

King, Andrew. 2008. *Website Optimization: Speed, Search Engine & Conversion Rate Secrets* . s.l. : O'Reilly Media, Inc, 2008. 978-0596515089 .

Kohavi, Ron. 2008. Bloodletting: Why Controlled Experiments are Important . [Online] May 19, 2008. http://exp-platform.com/bloodletting.aspx.

Kohavi, Ron, et al. 2009. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery.* February 2009, Vol. 18, 1, pp. 140-181. http://exp-platform.com/hippo_long.aspx.

Kohavi, Ron, Henne, Randal M and Sommerfield, Dan. 2007. Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO. [ed.] Rich Caruana, and Xindong Wu Pavel Berkhin. August 2007, pp. 959-967. http://exp-platform.com/hippo.aspx.

Krug, Steve. 2005. *Don't Make Me Think: A Common Sense Approach to Web Usability.* 2nd. s.l. : New Riders Press, 2005. 978-0321344755 .

Lee, Fiona, et al. 2004. The Mixed Effects of Inconsistency on Experimentation in Organizations. *Organization Science.* 2004, Vol. 15, 3, pp. 310-326.

Linden, Greg. 2006. Early Amazon: Shopping cart recommendations. *Geeking with Greg.* [Online] April 25, 2006. http://glinden.blogspot.com/2006/04/early-amazon-shopping-cart.html.

Linsky, Martin and Heifetz, Ronald. 2002. *Leadership on the Line: Staying Alive Through the Dangers of Leading.* s.l. : Harvard Business School Press, 2002. 978-1578514373 .

Lucovsky, Mark. 2005. Shipping Software . *Markl's Thoughts* . [Online] February 12, 2005. http://mark-lucovsky.blogspot.com/2005/02/shipping-software.html.

Marks, Harry M. 2000. *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900-1990.* s.l. : Cambridge University Press, 2000. 978-0521785617 .

McConnell, Steven C. 2004. *Code Complete.* 2nd Edition. s.l. : Microsoft Press, 2004.

Moran, Mike. 2007. *Do It Wrong Quickly: How the Web Changes the Old Marketing Rules* . s.l. : IBM Press, 2007. 0132255960.

—. 2008. Multivariate Testing in Action. *Biznology Blog by Mike Moran.* [Online] December 23, 2008. http://www.mikemoran.com/biznology/archives/2008/12/multivariate_testing_in_action.html.

Pasca, Marius and Shanahan, James G. 2008. Industry Event. *ACM 17th Conference on Information and Knowledge Management* . [Online] Oct 29, 2008. http://cikm2008.org/industry_event.php#Kohavi.

Pfeffer, Jeffrey and Sutton, Robert I. 2006. *Hard Facts, Dangerous Half-Truths, and Total Nonsense: Profiting from Evidence-Based Management.* s.l. : Harvard Business School Press, 2006. 978-1591398622 .

Ray Ozzie. 2005. Ozzie memo: 'Internet services disruption'. [Online] 10 28, 2005. http://news.zdnet.com/2100-3513_22-145534.html.

Robinson, Ken. 2006. Do schools kill creativity? *TED: Ideas Worth Spreading.* Feb 2006. http://www.ted.com/index.php/talks/ken_robinson_says_schools_kill_creativity.html.

Roy, Ranjit K. 2001. *Design of Experiments using the Taguchi Approach : 16 Steps to Product and Process Improvement.* s.l. : John Wiley & Sons, Inc, 2001. 0-471-36101-1.

Thomke, Stefan. 2001. Enlightened Experimentation: The New Imperative for Innovation. Feb 2001.

Thomke, Stefan H. 2003. Experimentation Matters: Unlocking the Potential of New Technologies for Innovation. 2003.

von Hippel, Eric. 1994. "Sticky information" and the locus of problem solving: Implications for innovation. *Management Science.* 1994, Vol. 40, 4, pp. 429-439.

Weiss, Carol H. 1997. *Evaluation: Methods for Studying Programs and Policies.* 2nd. s.l. : Prentice Hall, 1997. 0-13-309725-0.

Wiegers, Karl E. 2003. *Software Requirements.* 2nd Edition. s.l. : Microsoft Press, 2003.

Wooton, David. 2007. *Bad Medicine: Doctors Doing Harm Since Hippocrates* . s.l. : Oxford University Press, 2007. 978-0199212798 .

Data Mining Based Tax Audit Selection: A Case Study from Minnesota Department of Revenue

Kuo-Wei Hsu Nishith Pathak Jaideep Srivastava
Department of Computer Science and Engineering,
University of Minnesota
Minneapolis, MN USA
{kuowei, npathak , srivastava}@cs.umn.edu

Greg Tschida
Department of Revenue,
State of Minnesota
St. Paul, MN USA
greg.tschida@state.mn.us

Eric Bjorklund
Computer Sciences Corporation
Falls Church, VA USA
ebjorklu@csc.com

ABSTRACT

In 2001 the Internal Revenue Service (IRS) estimated the tax gap, i.e. the gap between revenue owed and revenue collected, to be approximately $345 billion, of which they were able to recover only $55 billion. It is critical for the government to reduce the tax gap and an important process for doing so is audit selection. In this paper, we present a case study where data mining based methods are used to improve the audit selection procedure at the Minnesota Department of Revenue. We describe the current tax audit selection process, discuss the data from various sources as well as the issues regarding feature selection, and explain the data mining techniques employed. On evaluation data, data mining methods showed an increase of 63.1% in efficiency. We also present results from actual field experiments (i.e. results of field audits performed by auditors at the Minnesota Department of Revenue) validating the effectiveness of data mining for audit selection. The impact of this study will be a refinement of the current audit selection and tax collection procedures.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data mining.*

J.1 [**Computer Applications**]: Administrative Data Processing – *Financial, Government.*

General Terms

Management.

Keywords

Tax audit selection, Data mining, MultiBoosting, Naïve Bayes

9. 1. INTRODUCTION

The Internal Revenue Service (IRS) employs the concept of "tax gap" to estimate the amount of non-compliance with tax laws. The tax gap measures the difference between the amount of tax that taxpayers should pay and the amount that taxpayers actually pay on time. For Tax Year 2001, the estimated tax gap was about $345 billion and only a small portion of it was eventually collected. The IRS recovered about $55 billion, reducing the tax gap for 2001 to $290 billion. This means that, ultimately, the IRS recovered only 15.9% of the total tax gap. Former IRS Commissioner Mark W. Everson has stated that "the magnitude of this tax gap highlights the critical role of enforcement in keeping our system of tax administration healthy".

Since, tax is the primary source of revenue for the government, it is extremely important for the government to reduce the tax gap. The first step is to understand the sources of this tax gap. These include non-filing of tax returns, underreporting of tax and underpayment of tax. Underreporting of tax is the single largest factor, and the amount underreported is much larger than the sum of the other two. Underreporting of taxes presents a challenge to the government and discovering these cases requires considerable effort and work from multiple departments. The data mining community has made contributions towards solving this problem. Related work include using artificial neural network (ANN) to determine if an audit case requires further audit [17], using classification techniques to assist in strategies for audit planning [2, 3], and using machine learning and statistical methods to identify high-income individuals taking advantage of abusive tax shelters [5, 6]. Most research models underreporting of tax as fraudulent behavior, and discovering unreported taxes is realized as a fraud detection problem. In this paper, however, we focus on a different problem, viz. audit selection. Especially, we focus on a specific form of tax called Use tax. However, our data mining based approach also helps audit selection for Sales tax because audits for Use and Sales taxes are usually conducted together. While this is also an important part of tax administration, so far little research has addressed it.

The audit selection problem is that of effectively selecting audit cases from a pool of candidates such that the selected cases will result in significant revenue gains. Audit selection is the very first step in any tax audit project. Improving the efficiency of audit

[1] http://www.irs.gov/newsroom/article/0,,id=158619,00.html

selection is a key strategic priority to drive government revenue growth. The more potentially profitable cases are selected for audits, the less unsuccessful audits could be expected, the more cost savings will be achieved, and more revenue will be brought into the government. Audit selection is important for all audit processes and requires intensive efforts as well as knowledge from domain experts.

It is impractical to audit all taxpayers with the limited time and resources. Moreover, there is always a cost associated with an audit and the generated revenue might not cover it. Due to these factors, in any audit project analysts evaluate certain audit cases and determine taxpayers who are at potential risk for underreporting or underpaying taxes. The final results (i.e. the revenue generated by audit cases) are dependent on the quality of the initial pool of selected taxpayers or audit cases. Although there is a systematic approach, it serves more as a guideline and audit cases are generally evaluated by analysts based on their experience. Nevertheless, there is room for improvement and data mining has significant potential to improve the audit selection process.

The current process for audit selection is human-intensive and depends heavily on the experience of tax analysts. It proceeds as follows - first, rules derived from tax research are used to filter out several thousand candidate cases from a database (or data warehouse). This list of candidate cases is then refined and several hundred of them are selected for field audits. In the refinement phase analysts evaluate candidate cases partially based on some pre-specified set of rules, but mostly based on their experience and expertise. If audit cases are well selected in the two stages, there is a better chance of generating cost savings and additional revenue. If an audit case turns out to be unsuccessful (i.e., to generate revenue less than the efforts associated with the audit process), the cost is not only a waste of time but also the loss of an opportunity to work on a successful case. The goal of this project is to utilize data mining to improve audit selection, particularly the second stage. As for the first stage, where an initial pool of candidates is generated, the audit selection criteria and process are confidential.

Figure 1 illustrates the problem: The y-axis is the ratio of the average benefit (or revenue) obtained from an audit case to the average cost of an audit; the x-axis is the number of audits that are ranked according to some criteria, such as the business size.

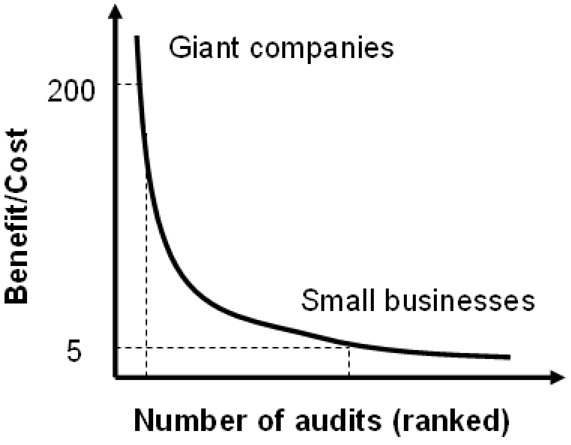


Figure 1. Audit selection problem.

Large businesses are usually selected because there are a few of them and they have higher B/C (Benefit/Cost) values, and therefore, auditing them generates more revenue. In contrast, small businesses are selected more or less at random. Improvement can be obtained by 1) getting higher B/C values (i.e. creating a lift) in the tail area, or 2) extending the curve with the same B/C values (i.e. making an extension) in the tail area.

In this project, we employ data mining to analyze a collection of candidate cases filtered from the database, and aim to identify a smaller set of more profitable candidate cases. The identified candidates (i.e. those classified as “Good” by a data mining process) would be good cases for field audit. We pose this a classification problem, and use supervised learning methods along with historical data for training. Focusing on a particular tax type and a specific group of taxpayers, we describe our approach and report validation results. We analyze tax data and audit results collected during a certain period of time. The tax audit data reflects taxpayer behavior and poses certain challenges for data mining. This data contains latent subgroups and suffers from the imperfect nature of real-world data. Models trained on this data are evaluated using data from real audits conducted during a subsequent period of time. It must be noted that our approach has been fully validated by the Department of Revenue, based on actual field audits performed by auditors. The proposed approach has been used for more than one tax audit. Thus, this presents a unique and valuable case study for mining tax data.

The rest of this paper is organized as follows - Section 2 briefly introduces domain knowledge and background information. Sections 3 and 4 describe our approach and results for data mining on tax data, respectively. Section 5 presents validation results from field audits and section 6 concludes the paper with a discussion on the impact of this study.

10. 2. BACKGROUND

The Department of Revenue (DOR) is responsible for executing and enforcing the tax laws defined by the legislative process. Enforcement of the tax laws is one key piece of this process. Carrying out audits to identify taxpayers that are furthest from tax compliance is a central component of this process. The DOR has a limited number of resources to allocate to this process, so there is always interest in finding better ways to identify taxpayers furthest from compliance and therefore allocate resources more efficiently. There are significant opportunities within the DOR to improve the efficiency of compliance activities and subsequently increase revenue. Data mining can identify these. In looking at the existing compliance efforts, there are essentially two avenues for improving efficiency and increasing revenue: cost savings and revenue collection.

Cost savings could be understood through the following example. In a real audit project, the final results indicated that approximately 27% of the audits generated less than $1,000 of assessment while an assessment greater than $1,000 was defined as the cutoff for a successful audit. The cost of that work was nearly a Full-time equivalent (FTE) for half a year doing work that did not generate revenue for the DOR. However, as for the project, a 73% success ratio was one of the more successful audit projects. This leads to the conclusion, that if successful in reducing costs for this project, data mining has an even greater potential to reduce costs for audit projects that currently generate a higher number of unsuccessful audits. Moreover, revenue collection is the other essential way to improve efficiency and increase revenue. For every given audit executed, there is potentially a better audit candidate that is not being audited that could provide a higher return for the Department than the audit this is currently being performed. Data mining is the solution that best enables increased efficiency and revenue collection and is applicable to most or all of compliance activities. Nevertheless, data mining is not the only solution to improve audit selection at the DOR.

In this section, the audit process in each division at the DOR is described. Given that each tax type has unique characteristics and varying degrees of reliance on data sources other than its own tax return, the audit process varies across divisions at the DOR. It was observed that all divisions are motivated to increase efficiency throughout the audit process. Details relating to the current state and needs of each division at the DOR with respect to audit are outlined below.

10.1 2.1 Individual Income Tax

Part of the Individual Income Tax (IIT) audit workload is reduced by filtering out non-compliance that can be identified by looking at the returns alone or in comparison with available federal return data as the return is being processed. This filtering is done using audit selectors. Once the return has posted and all the federal return data has been received, automated audits are generated by comparing the state and federal return data. This process generates tax orders with little user intervention. The IIT audit teams are responsible for generating audit leads from the remaining returns. The IIT audit group does have a limited number of pre-defined queries that they run each year to identify non-compliant taxpayers, but these queries require several manual merges of data. In addition to the pre-defined queries, the supervisors and lead workers are responsible for using their experience and data resulting from ad-hoc queries to identify audit leads. These queries may also require several manual merges depending on the data requested.

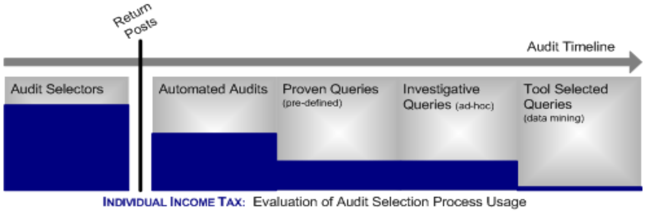


Figure 2. Individual Income Tax (IIT).

10.2 2.2 Sales and Use Tax

Nearly all Sales and Use Tax returns are processed electronically which allows them to make use of audit selectors as the return is being processed. Tax experts currently use a few dozen selectors and do not believe there is much room to increase this number due to the limited number of fields on the return itself. Automated audits (i.e., comparing the Sales Tax return to another electronic return) are not possible at this time due to the inaccessibility to the necessary comparison return data. Once the returns are in the system, tax experts request worksheets listing subsets of taxpayer financial information. They then sort this list as necessary and add other sources as necessary to look up business taxpayers attempting to find good leads. Audit selection process uses both pre-defined queries and ad-hoc queries to an extent. However, most audit selection work is delegated to each region to extract independently using the data available in spreadsheets and the related system.

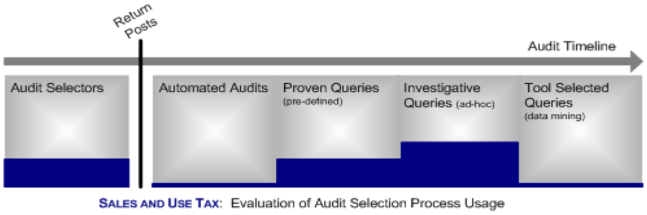


Figure 3. Sales and Use Tax.

10.3 2.3 Corporate Franchise Tax

Processing the Corporate Tax Returns and getting the paper filed returns into electronic entry format is a difficult task. The returns come in many different formats and it is a challenge for data entry personnel to fit the data into the standard tax return data structure. The electronic data is therefore somewhat suspect at this time. As a result, audit selection process for Corporate Tax returns is partially paper-based. Automated audits are not being used today, but if investigated further there may be some potential for using them. The audit group makes use of some proven queries and tracks the success of these queries.

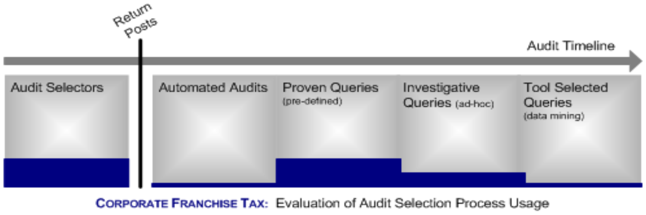


Figure 4. Corporate Franchise Tax.

10.4 2.4 Partnership, Estate, Fiduciary, S-Corporation Tax

Partnership, Estate, Fiduciary, S-Corporation Tax (PEFS) filings are all entered into the related system. Some audit selectors called audit flags are implemented. Proven queries and investigative queries play a part in audit selection. Federal, Individual, and other data are merged with PEFS data to identify audit candidates. Most of the investigative project work has come about when new resources were added to the audit group; otherwise the majority of audit work revolves around proven audit projects. Additional staff and additional data sources could lead to more investigative audits and ultimately better audit selection quality.

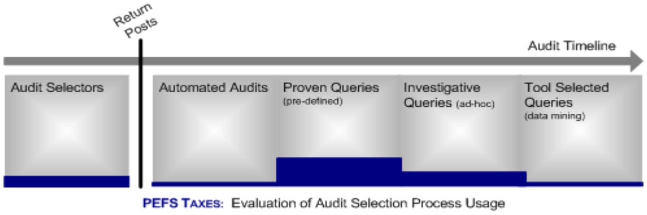


Figure 5. Partnership, Estate, Fiduciary, S-Corporation Tax (PEFS).

10.5 2.5 Other Taxes

Audit selection process for the Withholding Tax Division makes use of the data available to them. They do not have automatic audit selectors in the withholding return processing system.

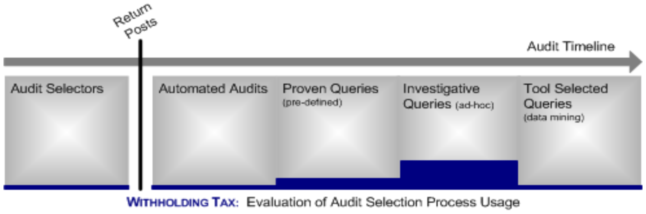


Figure 6. Withholding Tax.

For the majority of the tax types under the jurisdiction of the Special Taxes Division, there is not a significant need to improve the efficiency of audit. Each tax type has a relatively small and static taxpayer population. Most of the external data is received on paper, not via electronic submission.

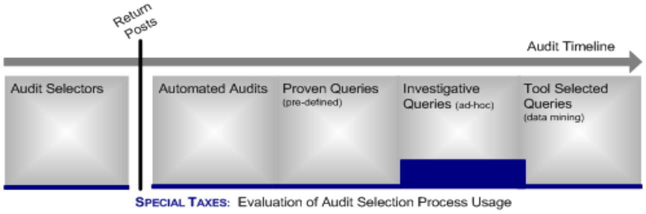


Figure 7. Special Taxes.

The Petroleum Tax Division has an audit group that primarily uses the information available on the petroleum tax return to guide audit selection. They too have a relatively small number of filers and feel that their audit process is well handled today and not in need of additional efficiency.

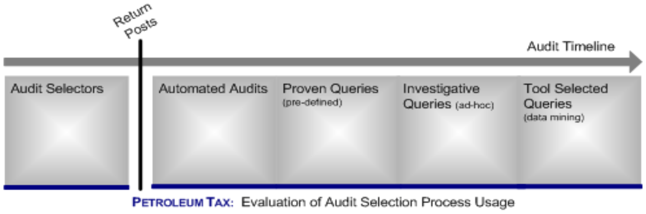


Figure 8. Petroleum Tax.

The Property Tax Division does not process returns in the same way that the other tax divisions do. Property Tax administers the Property Tax system; however, the counties are responsible for the processing of returns.

10.6 2.6 Summary

This paper reports the results of pilot studies for tax audit selection that were conducted by the Minnesota Department of Revenue (DOR). There was a pilot study for Individual Income tax, while the DOR also planned to have a pilot study for Corporate tax but did not have enough (electronic) data. Since Sales and Use taxes are generally available in electronic format, the data is mostly machine-readable and we could concentrate our attention on data analysis. Currently, the DOR is implementing an integrated tax system.

11. 3. APPROACH

11.1 3.1 Business Understanding

In the final pool of field audits, about 50% of the audits are fixed irrespective of their outcome. These include the largest companies in every state zone (where state zones are divisions of the state, created and used internally by the DOR for efficient work-load balancing) that are audited regularly. Also included are a group of audits dedicated to research conducted internally by analysts and tax specialists. The other 50% are handpicked by experts using the audit selection process. These hand-picked audits fall under a general category called APGEN, and typically consist of cases involving small to medium scale sized businesses.

Figure 9 presents the process for audit selection for the APGEN category.

Step 1: Auditors pose a database query (and potentially refine the query) to select several thousand records.

Step 2: Auditors look at data and select several hundred cases to audit; this step could take a few days up to a week.

Step 3: Auditors go out in the field and audit selected cases.

Step 4: Audit accuracy and ROI are calculated based on the results.

Figure 9. The current process for audit selection.

Initially, analysts pose a database query in order to select several thousand candidate cases out of all the businesses in the state. The query depends on tax type, the related systems and other information, as described earlier. Additionally, analysts can potentially refine the query for candidates satisfying certain other criteria (most likely criteria depending on the current state of the economy, industries and other situational factors). Next the analysts examine some of the candidate cases in more detail and select about 10% of them. They may also select certain cases based upon suspicions, recommendations from other experts, tip-offs etc. The final selection is based on case by case subjective evaluations by the experts and takes about a few days up to a week, depending on the number of experts. Finally, field audits are conducted on those selected cases and success is measured using audit accuracy along with return on investment (ROI) (which represents efficiency).

The more “potentially good” audit cases that are selected in the first two stages, the higher the revenue from the audits. If an audit case turns out to be unsuccessful, then - 1) Time and effort put in the audit is wasted, 2) The resources could have been directed at a successful case, thus, potential revenue from a successful audit is lost.

In this project, we apply data mining to audit selection process and use it to select field audits from the pool of several thousand candidate cases resulting from the database query. As mentioned earlier, the selection criteria and process to generate the initial pool is confidential so that our goal is to analyze the generated pool rather than to generate such a pool of candidates. We focus on Use and Sales taxes and our goal is to rank candidate cases based upon chances of success of a sales and use tax audit.

“Individuals owe the use tax on goods and services purchased outside their state of residence, by mail order, or over the Internet,” as described in the article [11]. Sales tax is for the government of the state where the good or services are sold, while Use tax is for the government of the state where the goods or services are used. However, the law regarding Use tax is quite complex and its collection is challenging.

Use tax is an important source of revenue and it comes from small to large scale state business. Use tax can be described as follows: If a company bought items from outside the state or any other entity that charges lower Sales tax than the region where the goods are consumed, the company then owes the difference in Sales tax as Use tax to the region where the goods are consumed.

11.2 3.2 Data Understanding

This project focused only on audits in the APGEN category since audits in all other categories are pre-determined. A binary definition of goodness of an audit was used and defined as the following: In this project, if use tax assessed is greater than $500 per year during the audit period then the audit is “Good”; if use tax assessed is less than $500 per year during the audit period then the audit is “Bad”. Audit period is the number of years in the past (starting from the current fiscal year) for which tax compliance is checked. In almost all cases the audit period is three years. In this project, a use tax assessment of $1,500 is considered a successful audit. This criterion was determined by the domain experts at the DOR.

Taxpayer behavior varies across many diverse factors and as a result, even though the focus is on use tax, data sources from other tax returns were considered as features. Figure 10 illustrates these data sources which include business registration, income, sales and use tax return data. Use tax field audits and their results conducted over the last three years were used to construct the training and test data. Multiple data sources are used for audit selection. This is because business tax data are complicated and related, with certain data sources having potential information regarding use tax compliance.

The training data set consisted of APGEN use tax audits and their results for the years 2004, 2005 and 2006. The test data consisted of APGEN use tax audits conducted in 2007 and was used to evaluate or test models built upon the training data set, while validation was done by actually conducting field audits on predictions made by the model on 2008 use tax return data.

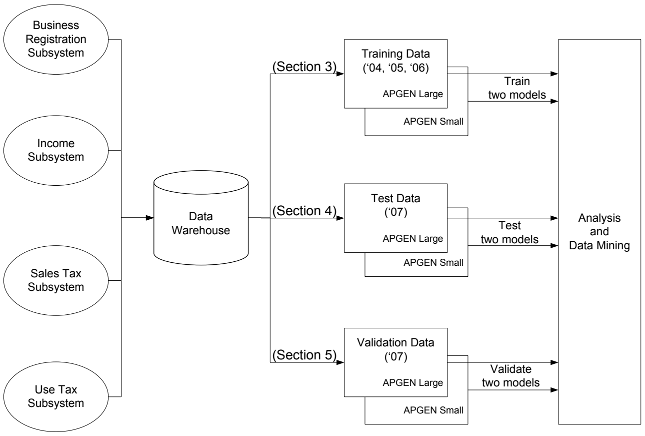


Figure 10. Data sources for data mining.

11.3 3.3 Data Preparation

The quality of the information embedded in data sources is associated with the quality of the knowledge discovered by data mining and it plays a critical role in the success of a data mining based process. It is increasingly recognized that the realization of the true potential of data mining requires that (training) data sets are properly pre-processed and are tied in with sufficiently good quality.

We started with cleaning the training data set by removing inadequate cases i.e. cases with none or at most one year of tax return data. These were generally new businesses or businesses that did not file tax returns; these cases had null values for most, if not all, of the features and were removed from the training data. The data set originally consisting of 11,083 cases was cut down to 10,943 cases after this step. The domain experts and data mining practitioners selected an initial list of about 220 features from the various data sources. After iterative cycles of feature selection and expert consultations a handful of features were selected. These features consisted of two categories: (i) Features related to business characteristics obtained from business registration data, geographic location and type of business. (ii) Features correlated with the size of the business, for the three years of the audit period. Nevertheless, details of features that were actually used in this study cannot be reported here due to the confidential nature of the tax audit process. Doing so can increase the potential for re-engineering of the audit process, which is clearly undesirable

Initial models built using the refined feature set were leaning heavily towards rule-sets that predicted successful audits for larger businesses. This was not surprising given that large businesses have complex tax returns and are more likely to make mistakes in ensuring tax compliance. Moreover for sufficiently large businesses, even trivial mistakes in tax compliance can result in significant revenues for the DOR. During the evaluation stage (using n-fold cross-validation on training data as well as results on test data) it was observed that the initial models did well for roughly half the population which consisted of relatively larger businesses and badly on the other half which consisted of smaller businesses. The pattern correlating business size with audit success was so dominant that almost all other patterns did not have sufficient relative support to be detected. Therefore, it was decided to divide the original modeling task into two parts ñ (i) build one DM model for audit prediction on larger businesses (for which the initial models seemed to be doing well) and (ii) build a second DM model for the smaller businesses. We chose to label these two categories as APGEN large and APGEN small respectively.

Businesses in the training set were ranked from largest to smallest, with average annual withholding amount being used as an indicator for business size. The annual withholding amount is directly related to the number of employees and is a very strong indicator of business size. Statistical t-tests [7, 13] were used to determine a withholding amount threshold such that for all businesses below it, there was no significant difference between annual withholding amounts of good and bad audit cases. For cases larger than the threshold value the business size played an important role in picking good audits (these were mainly the larger businesses in the dataset). The actual withholding amount determined for such a division was determined and it served as a (constant) threshold in this project. Thus the threshold was used to divide the dataset into the APGEN large and APGEN small categories.

We next performed feature selection for APGEN Large and Small individually and obtained different feature sets for each. The following figure presents the feature selection process (note: a similar feature selection process was used to build the initial models as well). Figure 11 presents the feature selection process.

Starting from the original feature set, we construct a working feature set and utilize it as training data to build some classification models. If the results are sufficiently good, we continue evaluating the model with test data. Here, good results mean those achieving reasonably high precision and recall with improved estimated ROI. In addition, the feature sets corresponding to good results are expected to be consistent with the knowledge and experience of domain experts. However, at any point where results were not adequately good, the models and their results, along with help from domain experts were examined to identify and remove inadequate features and/or derive new ones. This process was repeated iteratively in order to best utilize information embedded in the data. Deriving new features was suitable for this purpose and provided a chance for the classification algorithm to analyze the data from different perspectives.

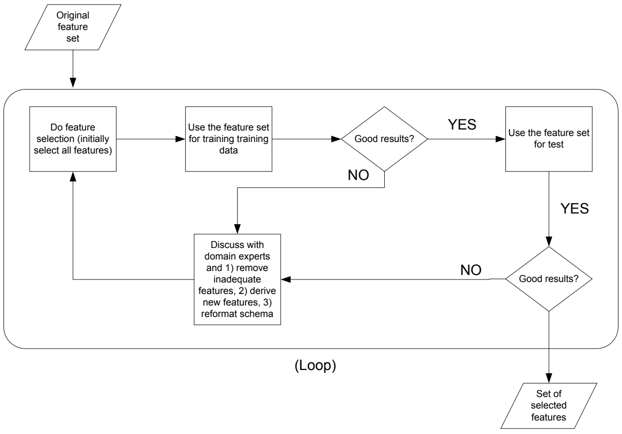

Original feature set
Do feature selection (initially select all features)
Use the feature set for training training data
Good results?
YES
Use the feature set for test
NO
YES
Discuss with domain experts and 1) remove inadequate features, 2) derive new features, 3) reformat schema
NO
Good results?
(Loop)
Set of selected features

Figure 11. The feature selection process.

Reformatting data schema was quite beneficial in this project. Before doing so, we could obtain good results in training data but not in test data, thus, indicating a potential over-fitting problem. It was realized that this problem was due to feature construction working with different tax filing cycles for various businesses. The problem existed from using actual timestamps and the data schema had to be reformatted to accommodate relative ones.

With the help of domain experts, we also recognized two categories of features. One category is composed of features from pre-audit information, which is compiled before auditors perform field audits; the other category consists of features from post-audit information, which is collected in or after field audits. However, post-audit information could not be used since it is not available until after auditors perform the field audits. If erroneously employed, it could misguide the classification algorithm and create models that succeed in training and/or test (evaluation) but fail in validation. In fact, any model created using post-audit information is not a predictive model but a descriptive one. While such models were found to be useful to help auditors better understand cases that had already been audited, they were unsuitable for our final objective and therefore, not explored further.

11.4 3.4 Handling Missing Values and Noise

The data cleaning process early on filtered out cases that had very little or no data associated with them, however, the training data-set still consisted of quite a few missing values. Generally speaking, a NULL-value can mean 1) not existing, 2) not applicable, or 3) existing and applicable but unknown. There are techniques designed to handle missing values in data mining. For example, we could replace “NULL” representing unknown value (i.e., the third case) with the mean or median of values in the corresponding column. As for the first two cases (i.e., not existing and non-applicable), we could employ classification algorithms that are not sensitive to such missing values or discard the corresponding column and/or instances. Features derived from various tax returns data fell in the first two cases. For example, in case of annual withholding amount, a NULL value could mean not existing (the business did not file a return that year or did not report a particular value) or not applicable (the business did not exist in that year or is not required to report the withholding amount). Features derived from business registration data (such as location, type of industry etc.) fell into the third case, however, for such features the domain experts felt that the median and mean values are a poor substitute for missing values. The dataset also consisted of some noise primarily arising due to errors in reporting from businesses filing their returns or errors due to incorrect data recording from the DOR. However, this was not a significant issue as the expected fraction of noisy records was very low. Hence, it was decided to ignore missing values and focus on classifiers that are robust towards handling them as well as noise.

11.5 3.5 Modeling

Concretizing the criterion of using business size as an important indicator of a good use tax audit, business size was related to some combination of withholding amounts and number of employees. Since the analysis discovered two subsets in the APGEN data set, it was accordingly divided into two parts. These parts were separately examined as they represented different groups of taxpayer behaviors. Consequently, models for each of them were built and these models primarily used various rules for business size along with industrial categorization and location of the businesses.

For modeling we used WEKA [8, 16], an open source data mining package. Several algorithms have been experimented with and the two with the best performance are reported here. For the APGEN large data set, MultiBoosting [15] using Naïve Bayes [10, 12] as the base algorithm was used, and for the APGEN small data set a Naïve Bayes model was trained. Naïve Bayes models have the following characteristics [14]: 1) They are robust to noise, 2) they handle missing values naturally, and 3) they are robust to irrelevant features. However, their performance can suffer if they are working with correlated features.

The Naïve Bayes algorithm assumes independence among the features. This assumption is unrealistic in many situations; however, in spite of it, Naïve Bayes has been successfully employed in many real-world applications. The success of Naïve Bayes algorithm might be related to the following explanation. First, the algorithm outputs the predicted class label rather than the actual probability for the class, which reduces the effect of the imprecisely estimated class probabilities. Second, the algorithm can still be useful if the effect due to dependent features is canceled [18].

MultiBoosting is an ensemble technique that forms a committee to use group wisdom to make a decision (i.e., classification or prediction). It is different from other ensemble techniques in the sense that it forms a committee of sub-committees. Each sub-committee is formed using the AdaBoost [9] algorithm and wagging [1] is used to further combine all these sub-committees into a single committee. It manipulates the given data set to generate different training sets and construct diverse member models (i.e., classifiers). The most important characteristic of MultiBoosting is that it exploits the bias reduction capability from AdaBoost as well as the variance reduction capability from wagging. Bias is the average difference between a created model and the (theoretical) underlying model, while variance represents the average difference among created models. Here the difference arises from using different training sets and it can be measured using error rates. AdaBoost has been shown to have effective bias as well as variance reduction, while it is primarily used for bias reduction. On the other hand, wagging (short for “weight aggregation”) is a variant of bagging [4] (short for “bootstrap aggregation”) and is used to reduce variance. Thus, MultiBoosting leverages both methods and forms a committee that is close to the (theoretical) underlying model and is stable i.e. less variance/more consistent results on unseen data.

12. 4. EVALUATION AND RESULTS

Models were built on tax audit data collected in 2004, 2005 and 2006, while they were tested (evaluated) by predicting goodness of audits for 2007 APGEN audit cases. We use some instead of all such cases because of certain restrictions (mainly those imposed by tax laws). The same withholding amount threshold was used to split the 2007 APGEN data set into large and small businesses and the corresponding models were used. The predictions made by the models on both large and small businesses were compared to the actual audit results.

12.1 4.1 APGEN large

To begin with, let us focus on large APGEN businesses or APGEN large. Figure 12 illustrates the evaluation procedure for data mining based audit selection on APGEN large.

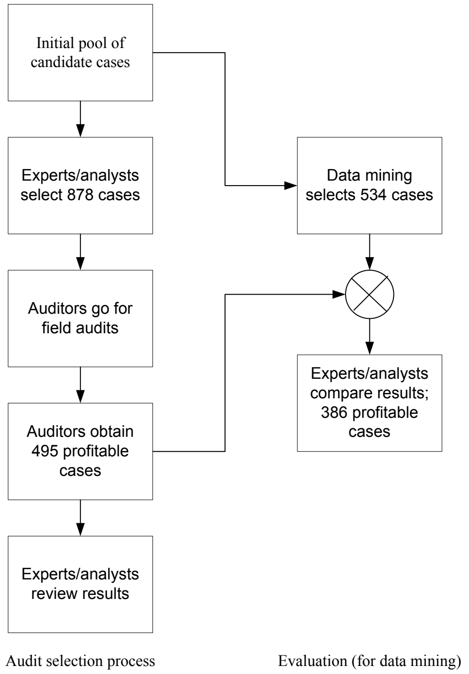

Initial pool of candidate cases
Experts/analysts select 878 cases
Data mining selects 534 cases
Auditors go for field audits
Experts/analysts compare results; 386 profitable cases
Auditors obtain 495 profitable cases
Experts/analysts review results
Audit selection process
Evaluation (for data mining)

Figure 12. Evaluation for APGEN large.

APGEN large for 2007 consists of 878 cases of field audits. Analysts predicted these 878 cases to be good audits and after the actual audits 495 of them actually turned out to be good audits. On applying data mining models to the same data set and comparing the results to those from field audits, it was observed that DM models predicted 534 out of the 878 to be good audits, out of which 386 cases (or 72.3%) were actually good audits. Along with traditional precision-recall measures, results were also evaluated on the ROI metric.

Table 1 summarizes results from the current audit selection process adopted by the DOR, while Table 2 presents results using the DM based model. In both tables, the first row (disregarding the header) indicates the number of audits (and corresponding dollar amounts) that were selected by the process. For evaluation purposes the following were used (these were estimates suggested by domain experts and not the actual numbers): The average number of hours spent conducting a use tax audit is 23 (hrs), while the average pay of tax specialist is $20 per hour. Thus, the collection cost of k audits was $460k. In Tables 1 and 2, the second and the third rows report revenue generated and collection cost, respectively.

Table 1. Business analysis for the current audit selection process for APGEN large.

	Good Audits	Bad Audits	Total
Number of	495	383	878

audits	(56.4%)	(43.6%)	(100%)
Revenue generated	$6,502,724 (97.4%)	$170,849 (2.6%)	$6,673,573 (100%)
Collection cost	$227,700 (56.4%)	$176,180 (43.6%)	$403,880 (100%)

Table 2. Business analysis for the data mining model created for APGEN large.

	Good Audits	Bad Audits	Total
Number of audits	386 (72.3%)	148 (27.7%)	534 (100%)
Revenue generated	$5,577,431 (98.7%)	$72,744 (1.3%)	$5,650,175 (100%)
Collection cost	$177,560 (72.3%)	$68,080 (27.7%)	$245,640 (100%)

The ROI value for the current audit selection process for APGEN large is 1,652% and the same for DM based models is 2,300%. This represents a 39.2% increase in efficiency. Here, we use increase in ROI as a measure of efficiency, as shown in Eq. 1 below.

$$\text{Efficiency} = \text{ROI} = \frac{\text{Total revenue generated}}{\text{Total collection cost}} \qquad \text{Eq. 1}$$

Figure 13 illustrates audit resource deployment efficiency (and so does Figure 15 in the next subsection). In Figure 13, the x-axis and y-axis respectively represent the number of audits performed (i.e. audit effort) and the number of audits that are successful and generate revenue. Theoretically best situation is that all audits performed turn out to be profitable. This is captured by the left-most (solid) line. Additionally, the current audit selection process is represented by the right-most (solid) line. Based on these two lines, we divide the space into three regions, as shown in both figures: The left-most region, A, represents the impossible situation of having more successful audits than the actual number of audits conducted. The right-most region, C, represents the situation where, compared to the current process, fewer successful audits are obtained. Any model in this region is more inefficient than the current process. The data mining based process is located in the region B. Any solution in region B represents a method better than the current one and is closer to the theoretically best process.

From Figure 13, the theoretically best process will find 495 successful audits when 495 audits are performed, while the current process will need 878 audits in order to obtain the same number of successful audits. If we project the (solid) line presenting the data mining based process, we observe that in order to obtain 495 successful audits, the number of audits performed will be lower than 878 (which is better than the current process). It can be estimated as 534*495/386, which is approximately 685 audits. Alternatively, if the current (manual) process selected only 534 cases, the number of successful audits would be lower than the number 386, found by the data mining based process. It can be estimated as 495*534/878, which is approximately 301. The former number shows that with data mining less effort is required for the same degree of tax compliance, while the latter number shows that higher tax compliance is achievable for the same effort.

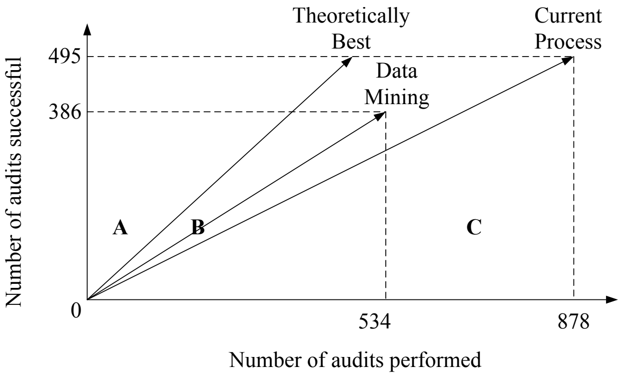


Figure 13. Audit resource deployment efficiency for APGEN large (for Tables 1 and 2).

Furthermore, Table 3 presents the confusion matrix for the DM model on APGEN large data set. Columns and rows are for predictions and actual results, respectively. We also report revenue and collection cost associated with each element. The top-left element is use tax assessment collected, the top-right element is use tax assessment lost i.e. cases predicted as bad turning out to be actually good, the bottom-left element is collection costs wasted due to audits incorrectly predicted as good and the bottom-right element is collection costs saved as predicted bad audits are not assessed. Notice that the model eliminated cases which consumed 26.7% of collection resources but generated only 1.4% of revenue, thus, significantly improving efficiency.

Table 3. The confusion matrix for APGEN large.

	Predicted as Good	Predicted as Bad

Actually Good	386 (use tax collected)
	R = $5,577,431 (83.6%)
	C = $177,560 (44%)
		109 (use tax lost)
		R = $925,293 (13.9%)
		C = $50,140 (12.4%)

Actually Bad	148 (costs wasted)
	R = $72,744 (1.1%)
	C = $68,080 (16.9%)
		235 (costs saved)
		R = $98,105 (1.4%)
		C = $108,100 (26.7%)

Note: R stands for "Revenue; C stands for "Collection cost".		

12.2 4.2 APGEN small

Figure 13 illustrates the process used to evaluate the data mining based process for APGEN small. Here, analysts predict 473 cases as good audits and after conducting field audits only 99 of them are actually good. For businesses in this subgroup, only one-fifth of cases selected by the process currently adopted by the DOR generate revenue greater than the pre-defined threshold value. In contrast, 47 out of 140 cases (33.6%) selected by the classification model are truly good audits.

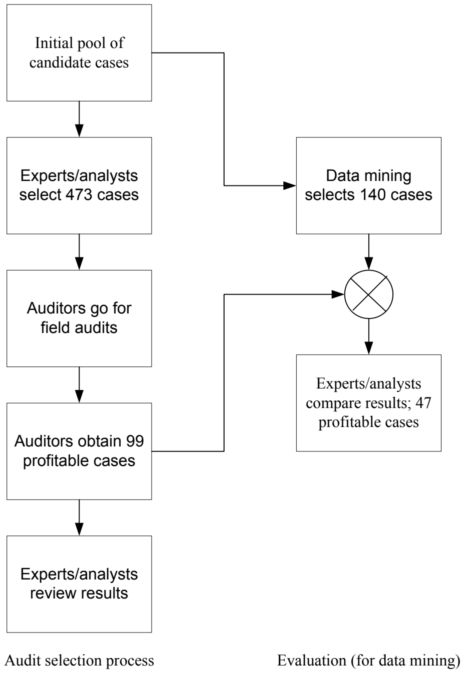

Initial pool of candidate cases
Experts/analysts select 473 cases
Data mining selects 140 cases
Auditors go for field audits
Experts/analysts compare results; 47 profitable cases
Auditors obtain 99 profitable cases
Experts/analysts review results
Audit selection process
Evaluation (for data mining)

Figure 14. Evaluation for APGEN small.

Table 4 presents results for the current audit selection process and Table 5 summarizes results for the DM model. Apart from the increase in precision, the ROI value for the current process for APGEN small is 285% and the same for the DM model is 447%, indicating a 57% increase in efficiency.

Table 4. Business analysis for the current audit selection process for APGEN small.

	Good Audits	Bad Audits	Total
Number of audits	99 (20.9%)	374 (79.1%)	473 (100%)
Revenue generated	$527,807 (85%)	$93,259 (15%)	$621,066 (100%)
Collection cost	$45,540 (20.9%)	$172,040 (79.1%)	$217,580 (100%)

Table 5. Business analysis for the data mining model created for APGEN small.

	Good Audits	Bad Audits	Total
Number of audits	47 (33.6%)	93 (66.4%)	140 (100%)
Revenue generated	$263,706 (91.5%)	$24,441 (8.5%)	$288,147 (100%)
Collection cost	$21,620 (33.6%)	$42,780 (66.4%)	$64,400 (100%)

Similar to Figure 13, Figure 15 illustrates audit resource deployment efficiency. From Figure 15, when 99 audits are performed the theoretically best process will find 99 profitable audits. However, the current process will need 473 audits in order to obtain the same number of successful audits. For using our data mining based approach to obtain 99 successful audits, the number of audits performed will be lower than 473. It can be estimated as 140*99/47, which is approximately 295 audits. This number shows that with data mining less effort is required to obtain the same degree of tax compliance. Moreover, 47 out of 140 cases selected by the data mining based process turned out to be successful audits. If the current process is employed to select 140 audits, the number of successful audits would be lower than 47. It can be estimated as 99*140/473, which is approximately 29. This number shows that with data mining higher tax compliance is achievable for the same effort.

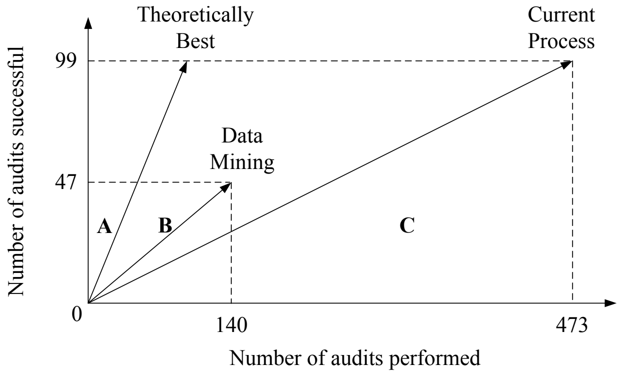


Figure 15. Audit resource deployment efficiency for APGEN small (for Tables 4 and 5).

The confusion matrix for the classification model for APGEN small is presented in Table 6. The 47 good audits correctly identified correspond to cases that consume 9.9% of collection costs but generate 42.5% of revenue. Note that the 281 bad audits correctly predicted by the DM model represent significant collection cost savings. These are associated with 59.4% of collection costs generating only 11.1% of the revenue.

Table 6. The confusion matrix for APGEN small.

	Predicted as Good	Predicted as Bad

Actually Good	47 (use tax collected)
	R = $263,706 (42.5%)
	C = $21,620 (9.9%)
		52 (use tax lost)
		R = $264,101 (42.5%)
		C = $23,920 (11%)

Actually Bad	93 (costs wasted)
	R = $24,441 (3.9%)
	C = $42,780 (19.7%)
		281 (costs saved)
		R = $68,818 (11.1%)
		C = $129,260 (59.4%)

Note: R stands for “Revenue; C stands for “Collection cost”.		

12.3 4.3 Summary

The following two tables summarize the evaluation results. Table 7 and Table 8 present results from data mining and business analysis perspectives, respectively. It is clear from Table 7 that the overall accuracy of the data mining based process is significantly higher than that of the current process. The data mining process achieves 64.2% in accuracy while the current process just shows 49.7%. Furthermore, based on the evaluation discussed above and Table 8, for the APGEN category, the net use tax ROI for the data mining based process is 1,915% while that for the current audit selection process is 1,174%. Consequently, the net increase in efficiency is 63.1%. The results demonstrate that data mining based methods are more efficient than the current manual process.

Table 7. Summary of evaluation results from data mining perspective.

APGEN		Large	Small	Combined
Current process	Number of selected cases	878	473	1351
	Good audits	495	99	594
	Accuracy	56.4%	20.9%	49.7%
Data mining	Number of selected cases	534	140	674
	Good audits	386	47	433
	Accuracy	**72.3%**	**35.6%**	**64.2%**

Table 8. Summary of evaluation results from business analysis perspective.

APGEN		Large	Small	Combined
Current process	Revenue generated	$6,673,573	$621,066	$7,294,639
	Collection cost	$403,880	$217,580	$621,460
	Efficiency (ROI)	1,652%	285%	1,174%
Data mining	Revenue generated	$5,650,175	$288,147	$5,938,322
	Collection cost	$245,640	$64,400	$310,040
	Efficiency (ROI)	**2,300%**	**447%**	**1,915%**

From the perspective of data mining practitioners, the following lessons have been learned:

(1) Data preparation and formatting is very important, e.g. for DM methods it is useful for data schemas to accommodate relative time-stamps rather than absolute ones. This allows one to generalize the developed models and apply them to data that will be collected in the future.

(2) Analysis of businesses in the APGEN category revealed two latent subgroups. Furthermore, different subgroups represent different behaviors and thus require different models. A single model for the entire group of taxpayers is too generic and performs poorly.

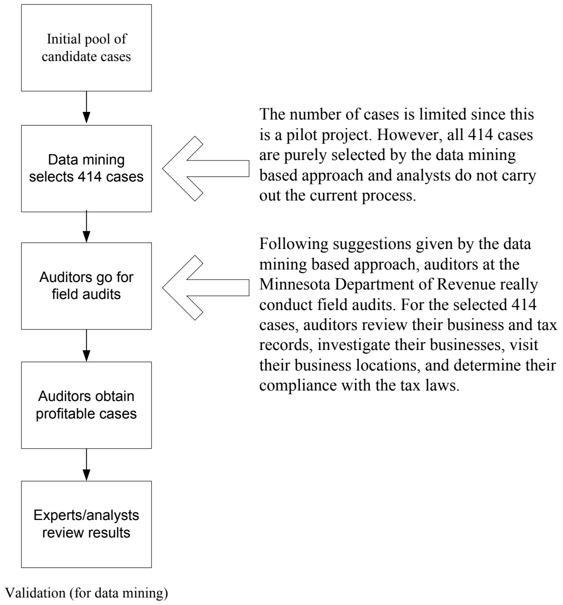
(3) Data preprocessing efforts must be devoted towards identifying latent sub-groups as well as their behaviors in order to better segment the data set for modeling (see previous point).

(4) Constant iterative refinement of models and feature sets, along with regular consultation with domain experts, is very important for achieving desired levels of performance.

13. 5. VALIDATION

In this section, we report results from field audits conducted by auditors at the DOR. In order to evaluate the pilot data mining project, the DOR validates models by using them to select cases for field audits. The audit selection criteria and process to generate the initial pool of candidate cases are confidential due to DOR regulations. Thus, the data mining based approach is to select cases from the pool for field audits. Figure 16 illustrates the process used to validate the data mining based process.

The DOR used the presented data mining models to select 414 tax cases for which auditors conducted actual field audits. For this project, the DOR defines a productive audit as an audit resulting in an assessment of at least $500 per year, or $1,500 per case.


Initial pool of candidate cases
Data mining selects 414 cases
Auditors go for field audits
Auditors obtain profitable cases
Experts/analysts review results
Validation (for data mining)

The number of cases is limited since this is a pilot project. However, all 414 cases are purely selected by the data mining based approach and analysts do not carry out the current process.

Following suggestions given by the data mining based approach, auditors at the Minnesota Department of Revenue really conduct field audits. For the selected 414 cases, auditors review their business and tax records, investigate their businesses, visit their business locations, and determine their compliance with the tax laws.

Figure 16. Validation for the data mining based process.

The models are used to analyze the collected tax data and the top 414 most likely predicted good audits are selected. For these selected cases, auditors review their business and tax records, visit their business locations, and determine their compliance with the tax laws. Our data mining approach has been accepted in the DOR and therefore the DOR did not carry out the current (manual) audit selection process in parallel. Consequently, there is no direct comparison between our data mining approach and the current process. However, as shown in the last step of Figure 16, audit results from the data mining based approach were reviewed by experts and analysts.

The DOR reviewed the results of the actual audits and compared the actual results to the predicted ones. Table 9 reports results in success rate, i.e. accuracy, while Table 10 reports results in dollars. Both tables present results for Use tax as well as Sales tax even though we focus on Use tax in our earlier discussions. On the one hand, auditors would simultaneously do Use tax and Sales tax when they decide to audit a taxpayer, no matter if the decision is based on their analysis for Sales tax or Use tax for the taxpayer. On the other hand, auditors would probably concentrate on Sales tax even though the initial decision was from the analysis of Use tax. This depends on their experiences, the possible collection cost, and the potential profits of these selected audits.

As we can see from Table 9, only 29% of audits for Sales tax are thought to be profitable by the current process (named “pre data mining” process) while 38% of audits are predicted as profitable by the data mining based process. After tax experts performed actual field audits, 37% of audits turn out to be successful and generate revenue for the DOR. Similarly, 56% of audits for use tax are predicted as profitable by the data mining based process, while only 39% of audits are thought to be profitable by the current process. For Use tax, the actual success rate is 51%, which is closer to the rate predicted by the data mining based process. These numbers validate that the data mining based process has better precision and consequently better efficiency.

Table 9. Validation results in success rate.

	Pre Data Mining Average Success Rate	Data Mining Predicted Success Rate	Actual Success Rate
Sales	29%	38%	37%
Use	39%	56%	51%

Table 10. Validation results in dollars.

	Pre Data Mining Average Success Rate	Data Mining Predicted Success Rate	Actual Success Rate
Sales	$6,497	$11,976	$8,186
Use	$5,019	$8,623	$10,829

Table 10 reports revenue in dollars predicted by the current process as well as the data mining based process and also the revenue actually collected by tax experts after they performed these field audits. The “pre data mining average success rate” is from historical data but not from conducting the current audit selection process in parallel.

Experts and analysts provided qualitative assessment of the cases that were selected, as mentioned earlier. The assessment results are not reported here since they are protected. Nevertheless, details are presented below. Results in dollars for different categories are reported in Table 11. As we described earlier, audits would concentrate their attention on Use tax, Sales tax, or both once they decided to conduct field audits. Therefore, there are different categories shown in Table 9. These results clearly show that the data mining based process actually generated more revenue for the DOR. For example, in Table 11, the DOR assessed Sales tax of $23,776 in average for relatively large businesses. Recall that the threshold for being a profitable audit is set to $1,500 per case (for a 3-year audit period). The result achieved by the data mining based process clearly proves that it is able to not only save costs and efforts but also generate more revenue. What is more, the current process is struggling with relatively small businesses and usually most cases would generate less than $1,500. Nevertheless, the average amount of assessed Sales and Use taxes achieved by the data mining based process is $2,504. Furthermore, if auditors decided to concentrate on Sales tax, the average assessed amount for relatively large businesses is $23,776 while that for relatively small businesses is $1,998.

Considering the threshold is set to $1,500 per case, and the average assessed amount of dollars is $18,848, we know that the revenue generated by the data mining based approach is over 12 times of the threshold that is associated with the average collection cost, These results demonstrate that data mining has the potential to efficiently and effectively perform more sophisticated tax audit selection.

Table 11. Validation results (of 414 filed audits) in dollars for different categories.

	Overall Total Assessed	Overall Average Assessed
Large Use and Sales	$1,399,436	$19,437
Small Use and Sales	$72,605	$2,504
Large Sales	$6,229,248	$23,776
Small Sales	$101,895	$1,998
Combined Totals	$7,803,184	$18,848

From the perspective of the DOR, the following list summarizes lessons that have been learned in this project -

(1) Mining may first uncover what is already known.

(2) The DOR may need multiple models to best capture taxpayer behavior. The DOR has to create separate medium and small business models for both sales and use tax.

(3) Do not count anything out. Analysts at the DOR used as much data as they could from other business returns.

(4) Do not be discouraged at preliminary results; iterative refinement will eventually get you where you want to be.

(5) Data mining identified a large number of small businesses that analysts at the DOR would otherwise have chosen only by chance.

(6) Predicting the actual dollar amount of likely payment is possible but is a much more difficult problem.

(7) The DOR is still testing the models we developed to see how they perform compared to the old methods of audit selection. The DOR still needs to complete a number of audits based on the data mining findings. Some models turn out to be better at telling the DOR who not to audit rather than point out good audit candidates (thus saving significant collection costs).

(8) Overall, there appears to be promise in the use of data mining for tax compliance.

From the perspective of the DOR, this was a pilot study and the first attempt to apply data mining to audit selection for Use and Sales taxes. While the data mining based process is not become routine, it has been used for more than one tax audit. In the future, the DOR intends to explore more applications and possibilities for data mining. The DOR is currently implementing an integrated tax system which will have much more comprehensive data collection. Given the success of this pilot study, the DOR will fully expects to use data mining much more comprehensively in the next couple of years.

14. 6. CONCLUSIONS AND THE IMPACT OF THE ANALYSIS

In this paper, we have shared our experiences and the lessons we learned over the past two years of utilizing data mining to improve tax audit selection. Additionally, we have described some practical challenges when applying data mining to audit selection problems. Improving the efficiency of audit selection (and further the productivity of the tax collection process) is an essential component of driving revenue growth for the DOR as well as the government.

The current audit selection process is an intensive, time consuming manual effort required of knowledgeable analysts. However, apart from being cumbersome, the current process is also inefficient. Bad audits not only waste auditors' time and resources but also erode revenue. Our proposed approach is to use data mining models for the purpose of improving audit selection. Since data plays a vital role in any data mining technique, we paid much attention to data preprocessing, cleaning, and reformatting. We tested our models using an independent test set. The time period for this test data was different from that of the training data. Our results show that DM models achieve a significant increase in efficiency (63.1%). The most important part of this study is the validation from actual field audits which demonstrate the usefulness of data mining for improving audit selection in terms of precision as well as revenue generated.

The presented study provides sufficient evidence that data mining based methods are an effective solution for audit selection. Improving government efficiency is important for effective governance, while improving tax collection efficiency is essential for economic activities. This is especially critical in a tough economy. Such analysis provides a further impact of increased interest among the government for effective applications of data mining. The direct impact of this study is a reexamination and refinement of other tax collection processes that are currently in use but may be inefficient.

15. ACKNOWLEDGMENTS

The research was supported in part by a grant from the Minnesota Department of Revenue. This collaboration would not have been possible without the sponsorship and support of a number of organizations and individuals. Specifically, the authors would like to thank the Minnesota Department of Revenue and the University of Minnesota for providing the institutional support for this work. Finally, we would like to acknowledge the support of the Office of Enterprise Technology, State of Minnesota, and the vision and leadership of Commissioner Gopal Khanna for creating a novel relationship which allows University faculty and students to interact with government departments to mutually collaborate on using advanced technologies to address the State of Minnesota's information technology needs. The opinions expressed in this paper, however, are solely those of the authors, and do not represent, directly or by implication, the policies of their respective organizations.

16. REFERENCES

[1] Bauer, E. and Kohavi, R. 1999. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. Machine Learning, 36, 105-139.

[2] Bonchi, F., Giannotti, F., Mainetto, G., and Pedreschi, D. 1999. A classification-based methodology for planning audit strategies in fraud detection. In Proceedings of the Fifth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (San Diego, California, United States, August 15 - 18, 1999). KDD '99. ACM, New York, NY, 175-184. DOI= http://doi.acm.org/10.1145/312129.312224

[3] Bonchi, F., Giannotti, F., Mainetto, G., and Pedreschi, D. 1999. Using Data Mining Techniques in Fiscal Fraud Detection. In Proceedings of the first international Conference on Data Warehousing and Knowledge Discovery (Florence, Italy, August 30 - September 1, 1999). DaWaK '99. LNCS 1676, 369-376.

[4] Breiman, Leo. 1996. Bagging predictors. Machine Learning, 24(2), 123-140.

[5] DeBarr, D. and Eyler-Walker, Z. 2005. Closing the Gap: Automated Screening of Tax Returns to Identify Egregious Tax Shelters. In Proceedings of the first workshop on Data Mining Case Studies in conjunction with IEEE 2005 international Conference on Data Mining (Houston, Texas, USA, November 27 - 30, 2005). ICDM '05. 34-40.

[6] DeBarr, D. and Eyler-Walker, Z. 2006. Closing the gap: automated screening of tax returns to identify egregious tax shelters. SIGKDD Explor. Newsl. 8, 1 (Jun. 2006), 11-16. DOI= http://doi.acm.org/10.1145/1147234.1147237

[7] DeGroot, M. H. and Schervish, M. J. 2001. Probability and Statistics, 3rd Edition, October, 2001. Addison Wesley. Chapter 10.

[8] Frank, E., Hall, M. A., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., and Trigg, L. 2005. Weka - a machine learning workbench for data mining. The Data Mining and Knowledge Discovery Handbook, 1305-1314, Springer, 2005.

[9] Freund, Y. and Schapire, R. E. 1996. Experiments with a new boosting algorithm. In Proceedings of the international Conference on Machine Learning (San Francisco, California, USA, 1996). 148-156.

[10] John, G. H. and Langley, P. 1995. Estimating Continuous Distributions in Bayesian Classifiers. The eleventh Conference on Uncertainty in Artificial Intelligence. 338-345.

[11] Manzi, Nina. 2007. Use Tax Collection On Income Tax Returns In Other States, St. Paul, Minnesota: Policy Brief, Research Department, Minnesota House of Representatives.

[12] Mitchell, T. 1997. Machine Learning, 1st edition, October, 1997. McGraw Hill Higher Education. Chapter 6.

[13] Moore, D. S. and McCabe, G. P. 2005. Introduction to the Practice of Statistics, 5th edition, February, 2005. W. H. Freeman. Chapter 12.

[14] Tan, P.-N., Steinbach, M., and Kumar, V. 2005. Introduction to Data Mining, US ed edition, May, 2005. Addison Wesley. Chapter 5 (Section 5.3).

[15] Webb, G. I. 2000. MultiBoosting: A Technique for Combining Boosting and Wagging. Machine Learning. 40(2).

[16] Witten, I. H. and Frank, E. 2005. Data Mining: Practical machine learning tools and techniques, 2nd Edition, 2005. Morgan Kaufmann, San Francisco.

[17] Wu, R. C. 1994. Integrating Neurocomputing and Auditing Expertise. Managerial Auditing Journal, 9(3), 20-26.

[18] Zhang, H. 2004. The optimality of Naïve Bayes. In Proceedings of the 17th international FLAIRS conference (FLAIRS '04), AAAI Press.

A practical approach to combine data mining and prognostics for improved predictive maintenance

Abdellatif Bey-Temsamani, Marc Engels, Andy Motten, Steve Vandenplas, Agusmian P. Ompusunggu
Flanders' MECHATRONICS Technology Centre.
Celestijnenlaan 300D, B-3001, Leuven, Belgium
abdellatif.bey-temsamani@fmtc.be, marc.engels@ fmtc.be, andy.motten@ fmtc.be, steve.vandenplas@ fmtc.be, agusmian.partogi@ fmtc.be

ABSTRACT

Original Equipment Manufacturer companies (OEMs) are facing more and more the challenge to increase the efficiency and reduce the cost for the service of their equipment over their lifecycle. A predictive maintenance strategy, where the optimal time to schedule a service visit is forecasted based on the condition of the equipment, is often proposed as an answer to this challenge. However, predictive maintenance approaches are frequently hampered. First, by the lack of knowledge of the features those give a good indication of the condition of the equipment. Second, by the processing power needed for prediction algorithms to forecast the future evolution of the selected features, especially, when large measurements are collected. In most cases, this processing power is not available from the machine's processor.

To overcome these problems, we propose in this paper to combine two approaches that are currently used separately: data mining and prognostics.

The proposed method consists of two steps. First, data mining and reliability estimation techniques are applied to historical data from the field in order to optimally identify the relevant features for the condition of the equipment and the associated thresholds. Secondly, a prediction model is fitted to the live data of the equipment, collected from customer's premises, for predicting the future evolution of these features and forecasting the time interval to the next maintenance action. To overcome the limited processing power of the machine's processor, this prediction part is computed on a local server which is remotely connected to the machine.

The proposed method proved always to retrieve, from the datasets, the relevant feature to be forecasted. Validation has been done for two different industrial cases.

A first prototype of the predictive module is implemented in some copiers and is running in live conditions, since November 2008, in order to check the forecast robustness. First results showed that this module offers a very good indication on when part replacement would be required, with some level of uncertainty which decreases over time.

Calculated business cases showed that this module will be highly beneficial for the company. Savings of approximately €4,8 million/year worldwide are estimated. This estimate was mainly calculated by reducing labour in reactive service visits and stock costs.

DMCS'09, June 28, Paris, France.
Copyright 2009 ACM 978-1-60558-674-8…$5.00.

Categories and Subject Descriptors

D.2.7 [**Distribution, Maintenance and Enhancement**]

D.4.8 [**Performance**]

C.4 [**Performance of systems**]

General Terms

Performance, Reliability, Algorithms, Management

1. INTRODUCTION

Condition based maintenance (CBM), also called Predictive maintenance (PdM), has evident benefits to OEMs, including reducing maintenance cost, increasing machine reliability and operation safety, and improving time and resources management [1]. A side to side comparison between PdM and traditional corrective or preventive maintenance programs is made in [2]. From this survey, it was concluded that major improvements can be achieved in maintenance cost, unscheduled machine failures, repair downtime, spare parts inventory, and both direct and indirect overtime premiums, by using PdM.

Although these benefits are well illustrated, two major problems hamper the implementation of predictive maintenance in industrial applications. First, the lack of knowledge about the right features to be monitored and second, the required processing power for predicting the future evolution of features, which is often not available on the machine's processor.

For the last problem, the present work fits in an architecture where the machines at the customer side are connected to a central office. The live measurement data are processed by a server at this central office. This allows using standard computers with off the shelf software tool, with virtually unlimited processing power and storage capacity. Next to condition monitoring of the machine and predictive maintenance scheduling, the central server can also be used for providing other services to the customer.

For the former problem, data mining techniques proved to be useful for relevant features extraction. It has been proved in [3][4] that application of data mining techniques to data, such as acoustic emission for the monitoring of corrosion processes, is very useful for extracting relevant features which can be used as parameters for machine diagnosis and/or prognostics. However, in many other industrial applications, no clear physical understanding of the process is available and therefore to retrieve these relevant features, a clear methodology is required.

The main contribution of this paper is such a practical methodology which combines data mining and prediction techniques used consecutively in order to perform an accurate predictive maintenance scheduling. This methodology is called the IRIS-PdM approach (see Section 6).

Unlike standards such as Cross Industry Standard Process for Data Mining (CRISP-DM) [5], which indicated, in a high level, the different steps to be followed in order to apply data mining to industrial data, the IRIS-PdM approach described in this paper proposes specific algorithms which can be applied directly in different industrial applications. Furthermore, up-till-now prognostic has been tackled as an independent step from data mining, assuming the prediction of a completely known parameter [6]. On the contrary, in the IRIS-PdM approach, prognostics is an integral part of the flowchart and makes use of the relevant features extracted in data mining step. In this way, the IRIS-PdM approach enables the possibility to compare different features evolution and combine them to improve the accuracy of remaining lifetime forecast.

The IRIS-PdM approach consists mainly of: (i) A data mining step on historical data where data preparation, data reduction and relevant features extraction are performed. (ii) A prognostics step, where optimal thresholds are retrieved using reliability estimation methods and prediction algorithms are applied on live data to estimate the remaining time for the relevant features to reach the thresholds. This remaining life time can be used to establish an optimal predictive maintenance.

This paper is organized as follows. In Section 2, general descriptions of the remote connectivity platform and the IRIS-PdM approach are given.

In Section 3, the application of the IRIS-PdM approach to an industrial dataset is illustrated, with a review of data preparation steps, data reduction techniques and the most important data mining modeling techniques used to optimally identify relevant features.

In Section 4, a description of the prognostics step is given, with a review of reliability estimation techniques to determine optimal thresholds and prediction algorithms to forecast the optimal time to schedule a maintenance action.

Finally, conclusions are given in Section 5.

2. GENERAL DESCRIPTION

2.1 Remote Connectivity Platform

Remote connectivity to customer premises offers the OEMs a continuous monitoring of their products, which results in a long service life of the products. This is thanks to the recommendations that OEMs can give to their customers for an optimal usage of the machines as well as the optimal maintenance scheduling before the machine or a sub-part of the machine approaches the end of life. Furthermore, improvement of maintenance logistics planning may be achieved.

The work presented in this paper fits within the framework of such a platform, schematically shown in figure 1. A human expert in a remote assistance center, at the machine builder side connects remotely to the customer premises, through a secure internet connection, and collects the data continuously or periodically from the machine.

Using different local software tools, including data mining software, different intelligent services can be provided such as Predictive Maintenance (PdM).

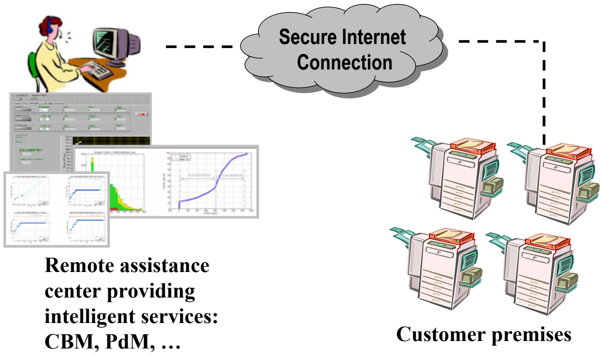

Secure Internet Connection
Remote assistance center providing intelligent services: CBM, PdM, …
Customer premises

Figure 1. Schematic of the remote connectivity platform.

2.2 IRIS-PdM Approach

In order to optimally forecast the predictive maintenance actions, a practical approach has been developed where different steps are followed starting from the available historical database of machines running in the field to the PdM scheduling. The different steps of the IRIS-PdM approach are summarized in figure 2.

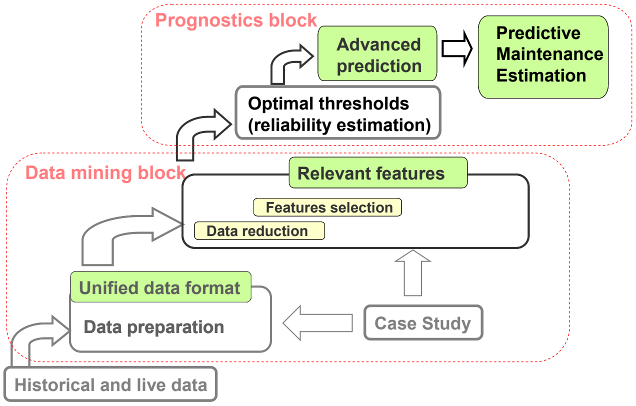


Figure 2. IRIS-PdM approach steps.

The IRIS-PdM approach proved to be easy to transfer from one industrial application to another. It was mainly tested for predictive maintenance of copy machines, but it also proved to be feasible for predictive maintenance of high-end microscopes.

The starting point is the historical database of the machines. In general, such a database exists for almost all machine manufacturers, but only a limited amount of information is currently used.

The next step consists of data preparation, including data transformation from the original to unified data formats and data cleaning such as removing outliers or calculating missing values.

Note that the data preparation step may be time consuming since the unified data format should be compatible with the data mining software tool and still retain a physical interpretation of the data.

The data modeling step consists of two sub-steps. Firstly, the data reduction sub-step where the Independence Significance Feature method is used to reduce significantly the number of attributes. This method is described in Section 3.2.1. Secondly, the features selection sub-step where the relevant features are extracted out of the selected attributes using a decision tree (DT) method. This sub-step is described in Section 3.2.2.

The choice of these methods was based on requirements of processing speed, accuracy, and interpretability of the results.

The final step consists of prognostics, itself divided into two sub-steps. First, reliability estimation methods are applied to historical database in order to identify the optimal thresholds of the selected features. These techniques are described in Section 4.1. Secondly, a prediction algorithm is applied to live data in order to forecast the time to schedule the optimal PdM. The prediction model used in this paper is based on slope calculation and called Weighted Mean Slope (WMS) model. This model is described in Section 4.2.

In next sections the different steps of IRIS-PdM approach are described in details and illustrated for an industrial dataset.

3. APPLICATION OF IRIS-PdM APPROACH TO A DATASET EXAMPLE

In this section, we illustrate the different steps of the IRIS-PdM approach, described in the previous section, on an industrial maintenance database. The dataset used in this section, is a large historical maintenance database of more than 2000 copiers.

Every copier has one or more data files corresponding to the history of service visits. One data file contains measurements of the sensors/counters that the copiers are equipped with and information added by the service technician, that illustrates the maintenance action type performed during his visits. This maintenance type can be corrective, in case a replacement of a broken part is done, or preventive, in case the part is changed before it is actually broken. As mentioned previously, the preventive maintenance is decided by the service technician and not always based on a physical understanding of the degradation process.

3.1 Data Preparation Step

The data preparation consists of transforming the information contained in the original data files of different machines to a unified format. The original data is structured in different files of different formats. The unified format is a matrix containing columns corresponding to the different attributes which represent sensors/counters and lines corresponding to observations at a service visit. Two extra columns with the component replacement information and maintenance action type are added. The different data files are concatenated into one single file. Since enough data was available and missing data can be considered as missing completely at random in the studied application, the missing values were simply discarded from the dataset. A schematic of such a data set format is given in table 1.

Based on the dataset, a matrix of more than 1000 attributes (features) and 1 million objects was generated.

3.2 Data Modeling Step

Data mining algorithms are generally evaluated according to three different criteria [7]

1. Interpretability: how well the model helps to understand the data
2. Predictive accuracy: how well the model can predict unseen situations
3. Computational efficiency: how fast the algorithm is and how well it scales to very large databases

The importance of these criteria differs from one application to another. For our use, the interpretability and scaling to large databases were essential. A summary of the techniques implemented in popular data mining software, is given in table 2.

Table 1. Schematic of a unified data format

Attribute_1	Attribute_2	…	…	Attribute_N	Output_1	Output_2
Object_{F11}	Object_{F21}			Object_{FN1}	Part not replaced	No maintenance
Object_{F12}	Object_{F22}			Object_{FN2}	Part replaced	Corrective maintenance
Object_{F13}	Object_{F23}			Object_{FN3}	Part not replaced	No maintenance
...	...			...		
…	…			…		
…	…			…		
Object_{F1M}	Object_{F2M}			Object_{FNM}	Part replaced	Preventive maintenance

Table 2. Important data mining methods and associated algorithms

Technique	Description	Method(s)
Clustering	Unsupervised machine learning to group objects in different classes	K-Means, AutoClass
Classification	Assign unknown objects to well established classes	Neural Networks, K-NN, SVM
Conceptual clustering	Qualitative language to describe the knowledge used for clustering	Decision Tree
Dependency modeling	Describes the dependency between variables	PCA, Dendrogram, ISF
Summarization	Provides a compact description of a subset of data	Statistical reporting Visualization
Regression	Determines functions that links a given continuous variable to other	ANN, Regression Tree, ML Regression
Rules based modeling	Generate rules that describe tendency of data	Association rules

The data modeling step is divided into two sub-steps (i) data reduction, and (ii) features selection.

3.2.1 Data reduction

The method chosen for data reduction is called Independence Significance Feature (ISF) technique. This method, initially described in [8], is meant to quickly and inexpensively discard features which seem obviously useless for the division of the data in multiple classes.

Unlike the Principal Components Analysis (PCA) method, the ISF method does not generate new features and therefore retaining the physical meaning of the original features.

The ISF method proved also to be much quicker than other methods like correlation or entropy reduction. For example, the processing time made by ISF method to calculate significance of the top 100 attributes out of the database described in Section 3.1 is approximately 2.5s while Spearman correlation method made 20s and Entropy reduction method made around 130s.

The ISF method consists on measuring the mean of a feature for all classes without worrying about the relationship to other features. The larger the difference between means, the better separation between classes (better significance).

In our case, this dramatically reduces the number of candidate predictors to be considered in the final selection process.

The ISF method reduces the data from more than 1000 features to ~100 features, using as an output of the two classes corresponding to the replacement information (Output_1 in table 1). The mathematical formula to calculate the significance, for every attribute, is given as:

$$Sig = \frac{\overline{X_1} - \overline{X_2}}{S_{\overline{X_1}-\overline{X_2}}} \qquad (1)$$

With

$$S_{\overline{X_1}-\overline{X_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (2)$$

Where

Sig : Significan ce

$\overline{X_i}$: mean value of the input objects in class i

s_i^2 : the variance of the input objects in class i

n_i : number of samples in class i

Only features with high significance measure are retained for further processing.

When the attributes are ordered according to decreasing significance, this results in the graph of figure 3.

As can be seen on this graph, only the first 100 attributes have higher significance than 0.2. This threshold was chosen, since the most related attributes to the classification output have significance higher than this value. These attributes are selected to be further processed in the next step.

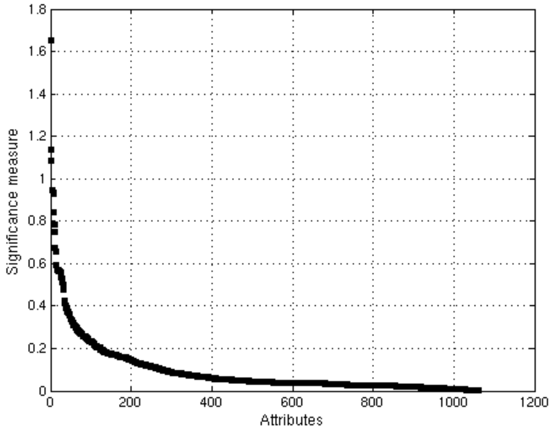


Figure 3. Significance measure versus attributes

3.2.2 Features selection

Once the data reduction is done as explained in the previous section, the relevant features extraction method is applied.

In this paper the Decision Trees (DT) method has been chosen. The main advantage of this method is the ease to interpret the results and transform them to if-then rules, which can be easily understood by an industrial user. Furthermore, decision trees algorithm implementations are quite robust and can deal with multivariate and multi-classes analyses.

In [9] a description of a decision tree method is given. It is a top-down tree structure consisting of internal nodes, leaf nodes, and branches. Each internal node represents a decision on a data attribute, and each outgoing branch corresponds to a possible outcome. Each leaf node represents a class.

Inducing a decision tree is done by looking to the optimal attribute to be used at every node of the tree. In order to do that, the following steps are followed [10]:

(1) a quantity known as Entropy (H) is calculated for every attribute

$$H(s_1, s_{2,} ..., s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{3}$$

with s_i the number of samples belonging to the class C_i, (m possible classes) for the calculated attribute. p_i is the ratio of number of samples in each class divided by the total number of samples (can be understood as a probability).

(2) the expected information $E(A)$ is computed for each attribute

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + ... + s_{mj}}{s} H(s_{1j}, s_{2j}, ..., s_{mj}) \tag{4}$$

where v is the number of distinct values in the attribute A and s_{ij} is the number of samples of class C_i in a subset S_j obtained by partitioning on the attribute A.

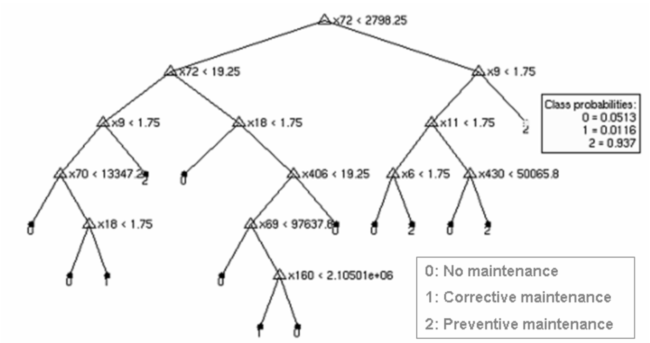


Figure 4. Decision tree model for relevant feature extraction

(3) Compute the information gain $G(A)$ using

$$G(A) = H(s_1, s_2, ..., s_m) - E(A) \tag{5}$$

(4) select attribute having highest information gain to be test attribute

(5) Split the node and repeat steps (1) to (4) till all values belong to the same class C_i or all attributes were used.

In this work, the CART decision tree implementation in Matlab statistics toolbox was used for a relevant feature selection.

Figure 4 shows a decision tree used to retrieve the relevant features from the list of the selected features obtained in Section 3.2.1. The purpose of this classification problem, is to identify from the list of features, which features are the most relevant to predict a maintenance action. The classification output used in this step is the maintenance type information (Output$_2$ in table 1).

Note that in order to reduce the size of the three a pruning method is used. This pruning method calculates the statistical significance of a further split of a data and stops the process when this significance becomes too low. By keeping the size of the three under control, it remains feasible to physically interpret the classification.

In the resulting decision tree for our dataset, more than 95% of preventive maintenances were performed when the feature with number x72 is higher than the value ~2800 (right hand branch).

The pattern proposed by decision tree can be visualized also by the scatter plot in figure 5. This figure shows a clear separation between values of feature x72 for the preventive maintenance class and the rest of the classes, at the threshold of ~2800. The preventive maintenance observations around zero value of y-axis are performed by service technicians for testing purposes.

By looking back to the physical meaning of attribute x72, a meaningful relationship to preventive maintenance can be carried out.

In order to check the accuracy of decision trees (DT) method towards other well-known data mining methods, such as k-NN (Nearest Neighbor) [11], a 5 fold cross validation check is carried out with both methods and visualized using confusion matrices. The corresponding results are shown in table 3 and table 4, respectively for the k-NN and DT method. An overall accuracy of 88% ± 0.5% was achieved for decision trees algorithm versus only 77% ± 0.9% for k-NN algorithm.

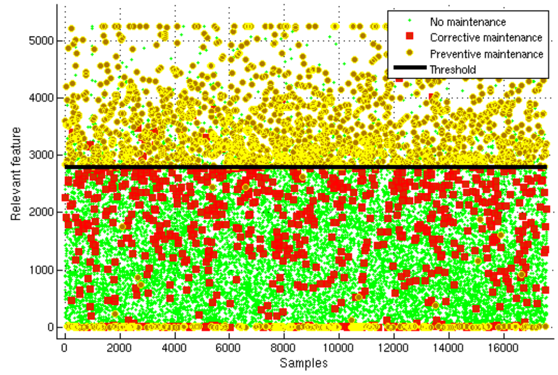


Figure 5. Classification using the relevant feature extracted by DT

Table 3. Confusion matrix for k-NN method

		Actual		
		NM	CM	PM
Prediction	NM	67.6%	57.3%	49%
	CM	9.9%	15.9%	11.6%
	PM	21.7%	26.7%	39.3%
Total		100%	100%	100%

Table 4. Confusion matrix for DT method

		Actual		
		NM	CM	PM
Prediction	NM	92.9%	73.3%	20.1%
	CM	4.6%	21.3%	1.9%
	PM	2.4%	5.3%	77.8%
Total		100%	100%	100%

Where NM, CM and PM stand, respectively, for no maintenance, corrective and preventive maintenances.

It is clearly shown from the tables that misclassification using the DT method is much lower than for the k-NN method. As an example, 77.8% of preventive maintenances are correctly classified using the DT method, versus only 39.3% using the k-NN method.

Note that the low percentage of the correct classification using both methods for corrective maintenance is mainly due to the corresponding low number of samples in the complete data set.

The value 2800 shown in figure 5 can be used as a threshold for the monitoring of the selected feature. Note that, in order to fine tune this threshold in an optimal way, statistical methods, such as reliability estimation could be used.

This latter together with prediction algorithms are discussed in the next Section.

4. PROGNOSTICS

Prognostics in the IRIS-PdM approach consists on two steps. (i) the reliability estimation step where optimal thresholds are calculated form the historical database for the selected features and (ii) the prediction step where a prediction algorithm is applied to live data to forecast the time for scheduling a maintenance action.

The former step is described in Section 4.1, while the latter step is described in Section 4.2.

4.1 Reliability Estimation

In the IRIS-PdM approach, reliability estimation is applied to the historical database in order to estimate the optimal thresholds for the selected features. These thresholds are going to be used for forecasting the remaining time until the next maintenance action. This estimation consists of fitting a life time distribution model on the data and identifying the failure rate at a given feature's value [12].

For our dataset, a Weibull distribution model fits quite well the data. Figure 6 shows the Weibull distribution model compared to the life measurements of the machines.

This model can be used to determine the optimal threshold by looking to an acceptable failure rate of the studied components. In our case we retain the threshold of ~2800 for the feature x72, which corresponds to ~20% failure rate (right side graph).

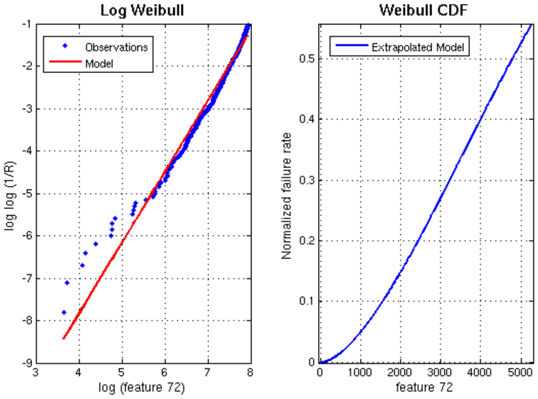


Figure 6. Weibull distribution model fitted to the data

4.2 Prediction Algorithm

In this step, time series of live data are analyzed. The main goal is to obtain a model of the evolution of the selected feature based on past observations in order to predict the future evolution. Combining this prediction with the optimal threshold, as defined in the previous section, allows estimating the remaining time before replacement of the part.

The accuracy of this remaining time depends strongly on the choice of the model. Two classes of models can be identified in literature: the models which allow a non-linear trend prediction and the ones which allow a linear trend prediction of time series.

For the former case, neural network is a good example, which is extensively used in stock market prediction [13]. It uses at least two parameters: the number of hidden nodes and the weight decay. These two settings are dynamically changing to adapt to the observations.

For linear trend prediction methods, exponential smoothing techniques have been used since long time [14][15]. In these techniques some model parameters need to be optimized in order to minimize the error between model and data. This can be problematic in the presence of steep changes in the data evolution versus time.

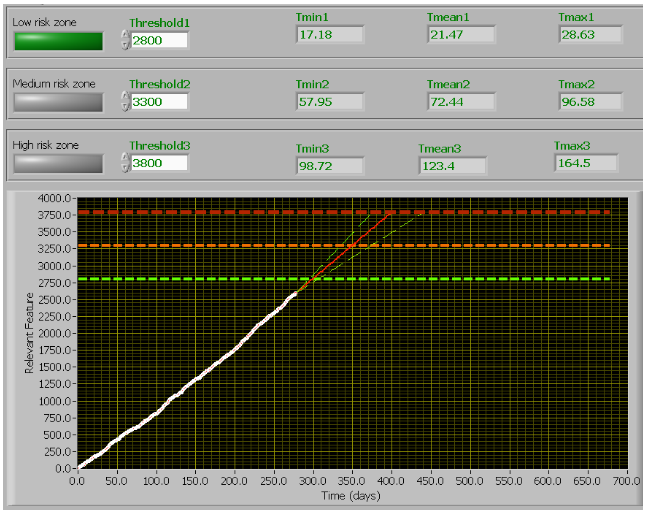
Therefore, a model based on weighted mean slopes (WMS) estimation has been developed. This model works robustly, even when data changes abruptly. In this model, the prediction is performed by looking to the recent observations and fit an optimal model to them.

Suppose the primary data $Y = \{y_1, y_2, \ldots, y_n\}$ is the live time sequence of the selected feature at time $T = \{t_1, t_2, \ldots, t_n\}$. The weighted mean slope (WMS) is calculated as:

$$\begin{cases} S_k = \dfrac{y_k - y_{k-1}}{t_k - t_{k-1}}; \qquad \{k = 2, \ldots, n\} \\ WMS = \dfrac{\sum_k k.S_k}{\sum_k k} \end{cases} \tag{6}$$

The prediction value at time $t + m$, is given by

$$y_{t+m} = y_t + WMS.m \tag{7}$$


Low risk zone
Threshold1 2800
Tmin1 17.18
Tmean1 21.47
Tmax1 28.63

Medium risk zone
Threshold2 3300
Tmin2 57.95
Tmean2 72.44
Tmax2 96.58

High risk zone
Threshold3 3800
Tmin3 98.72
Tmean3 123.4
Tmax3 164.5

Figure 7. Forecast of remaining time to maintenance action

An example of WMS prediction model applied to live data of one copier is shown in figure 7. The uncertainty bounds are based on variances of different machines usages, which have been calculated off-line. The WMS model has the ability to follow flat zones, where machines are not operational, which is not possible to achieve with a simple regression or exponential smoothing models.

In this figure three zones corresponding to low, medium and high risk zones are used. For each zone a different threshold can be set depending criticality of an unscheduled repair for the customer. The remaining days to perform a maintenance action are displayed with a confidence interval for each zone.

5. CONCLUSIONS

In this paper, a practical methodology called the IRIS-PdM approach has been presented and discussed. This approach consists of different steps making use of data mining, reliability estimation techniques and prediction algorithms in order to extract the relevant feature and use it in prognostics to predict the remaining time until a preventive maintenance action is required.

Independent Significance Feature (ISF) was successfully applied on an industrial data set for data reduction. Next, the most relevant features were extracted by means of the Decision Tree classification method. Comparison between the k-NN and the DT methods proves that DT is an accurate classification method.

A Weighted Mean Slopes (WMS) model was applied for prediction of the remaining time to schedule a maintenance action. This model works robustly, even for abruptly changing data.

The methods described in this approach can be broadly and robustly applied to different industrial data sets and maintenance databases. The results can also be easily interpreted and transformed to if-then rules, allowing insight and easy interaction with the results.

6. ACKNOWLEDGMENT

This research fits in the framework of IRIS (Intelligent Remote Industrial Services) project which is financially supported by the Dutch governmental agencies: Ministry of Economy Affairs, Province of Noord-Brabant, Province Limbrug and 'Samenwerkingsverband Regio Eindhoven, in the framework of Pieken in de Delta (PID) program (Project N^o. EZ 1/5262). IRIS is a cooperation of the project partners Sioux Remote Solutions, Assembléon, Océ Technologies, FEI Company, FMTC (Belgium) and the Technical University of Eindhoven. The IRIS consortium was created in cooperation with the BOM (Brabantse Ontwikkelings Maatschappij) and BO4.

The authors wish to thank all the IRIS project partners for their valuable advices and review of this research.

7. REFERENCES

[1]. J. Blair & A. Shirkhodaie, 'Diagnosis and prognosis of bearings using data mining and numerical visualization techniques', p.395-399, Proceedings of the 33rd Southeastern Symposium on System theory, 2001.

[2]. R. K. Mobley, 'An introduction to predictive maintenance', Van Nostrand Reinhold, 1990.

[3]. G. Van Dijck, 'Information theoretic approach to feature selection and redundancy assessment', PhD thesis, Katholieke Universiteit Leuven, April 2008.

[4]. R. Isermann, ‘Fault-diagnosis systems’, Springer 2006.

[5]. C. Shearer, ‘The CRISP-DM model: the new blue print for data mining’, Journal of data warehousing, Vol. 5, Nr. 4, p. 13-22, 2000.

[6]. K. M. Goh, T. Tjahjono & S. Subramaniam, ‘A review of research in manufacturing prognostics’, p.417-422, IEEE International Conference on Industrial Informatics, 2006.

[7]. R. Duda, P. Hart & D. Stork, ‘Pattern classification’, 2nd edition, 2001 by John Wiley & Sons, Inc.

[8]. W. Sholom, I. Nitin, Predictive data mining: a practical guide, Morgan Kaufmann, 1998.

[9]. P. Geurts, ‘Contributions to decision tree induction’, PhD thesis, University of Liège, Belgium, 2002.

[10]. J. You, L. & S. Olafsson, ‘Multi-attribute decision trees and decision rules’, Chap. 10, Springer, Heidelberg, Germany, pp. 327-358, 2006

[11]. Y. Zhan, H. Chen, G. Zhang, ‘An optimization algorithm of K-NN classification’, Proceeding of the fifth international conference on machine learning and cybernetics, Dalian, 13-16 August 2006. p. 2246-2251

[12]. P. Yadav, N. Choudhary, C. Bilen, ‘Complex system reliability estimation methodology in the absence of failure data’, Quality and reliability engineering international, 2008; 24: 745-764

[13]. D. Komo, C. J. Cheng & H. Ko, ‘Neural network technology for stock market index prediction’, p. 534-546, International symposium on speech, image processing and neural networks, 13-16 April 1994, Hong Kong.

[14]. R. G. Brown, F. R. Meyer, ‘The fundamental theorem of exponential smoothing’, A. D. Little, Cambridge, 673-685, 1961

[15]. E. S. Gardner, ‘Exponential smoothing: the state of the arts – Part II’, International journal of forecasting 22 (2006) 637-666.

Mining Medical Images

Glenn Fung, Balaji Krshnapuram, Jinbo Bi, Murat Dundar, Vikas Raykar, Romer Rosales, Sriram Krishnan, R. Bharat Rao,
Siemens Medical Solutions USA, Inc., 51 Valley Stream Pkwy, Malvern, PA-19355. Ph: 610-448-4819.
Glenn.Fung@Siemens.com , Bharat.Rao@siemens.com

ABSTRACT

Advances in medical imaging technology have resulted in a tremendous increase in information density for a given study. This may result from increased spatial resolution, facilitating greater anatomical detail, or increased contrast resolution allowing evaluation of more subtle structures than previously possible. An increased temporal image acquisition rate also increases the study information content. Finally, new technologies enable visualization or quantification of additional tissue properties or contrast mechanisms.

However, such technological advances, while potentially improving the diagnostic benefits of a study typically result in “data overload” overwhelming the ability of radiologists to process this information. This often manifests as increased total study time, defined as the combination of acquisition, processing and interpretation times; even more critically, the vast increase in data does not always translate to improved diagnosis/treatment selection. This paper describes a related series of clinically motivated data mining algorithms & products that extract the key, actionable information from the vast amount of imaging data in order to ensure an improvement in patient care (via more accurate/early diagnosis) and a simultaneous reduction in total study time.

In addition, these applications yield decreased inter-user variability and more accurate quantitative image-based measurements. While each application targets a specific clinical task, they share the common methodology of transforming raw imaging data, through knowledge-based data mining algorithms, into clinically relevant information. This enables users to spend less time interacting with an image volume to extract the clinical information it contains, while supporting improved diagnostic accuracy by reducing the risk of accidental oversight of critical information.

General Terms

Algorithms, Measurement, Performance, Experimentation.

DMCS’09, June 28, Paris, France.
Copyright 2009 ACM 978-1-60558-674-8…$5.00.

Keywords

Computer Aided Diagnosis, CAD, machine learning, data mining, image processing.

17. INTRODUCTION

The invention of the X-ray by William Röntgen in 1895 [50] revolutionized medicine. Thanks to the science of *in-vivo* imaging, doctors were able to look inside a patient’s body without resorting to dangerous procedures – the term “exploratory surgery” has all but vanished from our lexicon today.

The fundamental value of the X-ray remains the same today, as it was over 100 years ago – different structures (bone, cartilage, tissue, tumor, metal, etc.) can be identified based on their ability to block the X-ray/Röntgen beam. The initial uses of *in-vivo* imaging were to diagnose broken bones and locate foreign objects, such as, bullets, inside a patient’s body. As imaging techniques and resolutions improved, physicians began to use these methods to locate medical abnormalities (e.g., cancer), both for planning surgery and for diagnosing the disease. The state-of-the-art of medical imaging improved to the point that it soon required its own specialty, namely, radiologists, who were skilled at interpreting these increasing complex images.

The introduction of computers and the subsequent invention of computed tomography [51] in the 1970s created another paradigm – that of 3-dimensional imaging. X-ray beams were used to compute a 3-d image of the inside of the body from several 2-d X-ray images taken around a single axis of rotation. Radiologists were now not only able to detect subtle variations of structures in the body, they were now able to locate them within a fixed frame of reference. Early CT’s generated images or slices orthogonal to the long axis of the body, modern scanners allow this volume of data to be reformatted in various planes or even visualized as volumetric (3D) representations of structures.

A question that is often asked, is why medical imaging is moving towards increased finer resolution, and whether this has clinical value. This is illustrated in example below, for using CTs to diagnose lung cancer.

Lung cancer is the most commonly diagnosed cancer worldwide, accounting for 1.2 million new cases annually. Lung cancer is an exceptionally deadly disease: 6 out of 10 people will die within one year of being diagnosed. The expected 5-year survival rate for all patients with a diagnosis of lung cancer is merely 15%, compared to 65% for colon, 89% for breast and 99.9% for prostate cancer. In the United States, lung cancer is the leading cause of cancer death for both men and women, causes more deaths than the next three most common cancers combined, and costs almost $10 billion to treat annually.

However, lung cancer prognosis varies greatly depending on how early the disease is diagnosed; as with all cancers, *early detection* provides the best prognosis. At one extreme are the patients diagnosed with distant tumors (that have spread far from the lung, Stage IV patients), for whom the 5-year survival rate is just 2%. The prognosis of early stage lung cancer patients (Stage I) is more optimistic with a mean 5 year survival rate of about 49%. This follows logically from the fact that early detection implies the cancer is found when it is still relatively small in size (thus, fewer cancer cells in the body) and localized (before it has spread): therefore, many treatment options (surgery, radiotherapy, chemotherapy) are viable.

In order to identify and characterize increasingly minute lung lesions the resolution of the image must be improved. The recent development of multi-detector computed tomography (MDCT) scanners has made it feasible to detect lung cancer at very early stages, and the number of lung nodules routinely identified in clinical practice is steadily growing. The key factor in CT is the slice thickness, the distance between two axial cross-sectional X-rays, and decreases slice thickness means increased resolution. Today's MDCT's are capable of locating lung nodules that are 2mm-8mm in size, and cancers found at this early stage have excellent prognosis. Yet, despite these technologies, only 24% of lung cancer cases are diagnosed at an early stage [9,10], and many potentially clinically significant lesions still remain undetected [11].

One contributing factor could be the explosion of MDCT imaging data: just 8 years ago, the 2-slice CT could acquire 41 axial images of the thorax in a 30-second scan (single breath hold); the state-of-the-art 64-slice dual-source CT acquires up to 3,687 axial images in 30 seconds for each patient. Figure 1 illustrate 2 such images for a single patient, and each image must then be carefully examined by a radiologist to identify which of the marks on the image correspond to normal structures (air passage), benign tumors, lung diseases other than cancer, and early-stage lung cancer. It should be noted that despite the exponential increase in data in a few years, radiologists have roughly the same case load (or in some cases greater) than was the case 20 years ago when they examined a handful of images per patient.

There is a growing consensus among clinical experts that the use of computer-aided detection/diagnosis (CAD) software can improve the performance of the radiologist. The proposed workflow is to use CAD as a second reader (i.e., in conjunction with the radiologist) – the radiologist first performs an interpretation of the image as usual, and then runs the CAD algorithm (typically a set of image processing algorithms followed by a classifier), and displays "CAD marks" – structures identified by the CAD algorithm as being of interest to the radiologist. The radiologist examines these marks and concludes the interpretation. Figure 1 shows super-imposed CAD marks on the images. Clinical studies have shown that the use of CAD software not only offers the potential to improve the detection and recognition performance of a radiologist, but also to reduce mistakes related to misinterpretation [12,13].

The principal value of CAD is determined *not* by its stand-alone performance, but rather by carefully measuring the *incremental value* of Computer-Aided Diagnosis in normal clinical practice, such as the number of additional lesions detected using CAD. Secondly, CAD systems must not have a negative impact on patient management (for instance, false positives which cause the radiologist to recommend unnecessary and potentially dangerous follow-ups).

This explosion in radiologist data is not confined to CT alone. The invention of the CT was rapidly followed by the development of 3-d magnetic resonance imaging (MRI). MRI uses a powerful magnetic field to align the water molecules in the body, and thus provides much greater contrast between the different soft tissues of the body than does CT. Positron emission tomography (PET) and Single photon emission computed tomography (SPECT) use radioactive isotopes to provide functional imaging. Recently, medicine has been moving towards fusion of these different imaging modalities to combine functional and structural properties in a single image. As should be obvious, the ability to identify and characterize increasingly minute structures and subtle variations in bodily function in 3-d images, has resulted in an enormous explosion in the amount of data that must be processed by a radiologist. It is estimated that in a few years, medical images will constitute fully 25% of all the data stored electronically.

This vast data store also provides additional opportunities for data mining. CAD algorithms have the ability to automatically extract features, quantify various lesions and bodily structures, and create features than can be subsequently mined to discover new knowledge. This new knowledge can be further fed back into medicine as CAD progresses from detecting abnormal structures, to characterizing structures (identifies structures of interest, and also indicating whether they are malignant or not). This discussion is beyond the scope of this paper, which focuses on practical application that are deployed today. Another area of interest, is the use of CAD for change detection – for instance, to automatically measure tumors from images taken at different point in times and determine if the tumor size has changes. Such methods can be used both for diagnosis (malignant tumors grow quickly) and for therapy monitoring (is the tumor shrinking with the treatment).

One final area we discuss in this paper is the automatic quantification of ultrasound images. So far, all the modalities we have discussed take a snapshot of the body (MR compares various images to determine the change). Cardiac ultrasound captures the very fast motion of the heart; we describe mining software that tracks the motion of the heart and automatically measures key clinical variables (ejection fractions) that characterize the function of the heart.

The rest of the paper is organized as follows: Section 2 provides the clinical motivation for the image-mining systems listed in this paper. While traditional approaches were tried initially, we quickly realized the need for first principles research in order to achieve clinically acceptable levels of accuracy. Section 3 describes some of the original research in data-mining and machine learning that was necessary to develop systems with a clinically acceptable level of accuracy. Section 4 summarizes some of examples of the results obtained from clinical validation studies. Section 5 concludes the paper by summarizing the key lessons learnt while developing such real-world data mining applications in order to impact millions of patients.

18. CLINICAL MOTIVATION

18.1 Lung

Lung cancer is the most commonly diagnosed cancer worldwide, accounting for 1.2 million new cases annually. Lung cancer is an exceptionally deadly disease: 6 out of 10 people will die within one year of being diagnosed. The expected 5-year survival rate for all patients with a diagnosis of lung cancer is merely 15%, compared to 65% for colon, 89% for breast and 99.9% for prostate cancer. One of the promising but controversial ideas is to screen at risk patients (eg smokers) with CT scans.

One important factor when designing CAD systems for mining lung images is the relative difficulty in obtaining ground truth for lung cancer. Whereas, for example, in breast cancer virtually all suspicious lesions are routinely biopsied (providing definitive histological ground truth), a lung biopsy is a dangerous procedure, with a 2% risk of serious complications (including death). It makes obtaining definitive lung cancer ground truth infeasible, particularly for patients being evaluated for early signs of lung cancer.

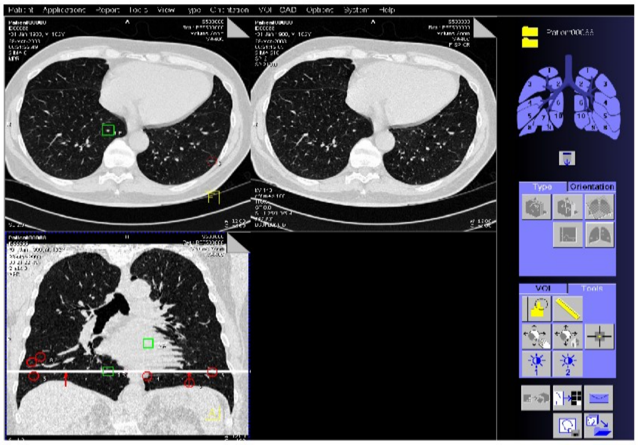


Figure 1: suspicious regions highlighted in lung (CT)

18.2 Colon

Colorectal cancer (CRC) is the third most common cancer in both men and women. It is estimated that in 2004, nearly 147 000 cases of colon and rectal cancer will be diagnosed in the USA, and more than 56 730 people would die from colon cancer, accounting for approximately 11% of all cancer deaths. Early detection of colon cancer is key to reducing the 5-year survival rate. In particular, since it is known that in over 90% of cases the progression stage for colon cancer is from local (polyp adenomas) to advanced stages (colorectal cancer), it is critical that major efforts be devoted to screening of colon cancer and removal of lesions (polyps) when still in a early stage of the disease.

Colorectal polyps are small colonic findings that may develop into cancer at a later stage. Screening of patients and early detection of polyps via Optical Colonoscopy (OC) has proved to be efficient as the mortality rate from colon cancer is currently decreasing despite population aging. CT Colonoscopy (CTC), also known as Virtual Colonoscopy (VC) is an increasingly popular alternative to standard OC. In virtual colonoscopy, a volumetric CT scan of the distended colon is acquired, and is reviewed by the physician by looking at 2D slices and/or using a virtual fly-through in the computer-rendered colon, searching for polyps. The interest in VC is bound to increase due to its many advantages over OC (better patient acceptance, lower morbidity, possibility of extra-colonic findings, etc.), with only a small penalty on sensitivity if the reader is a trained radiologist.

Computer-Aided Detection (CAD) systems are getting a lot of attention lately from the VC community. These various systems take advantage of having the full 3-D volume of the colon and use specific algorithms to detect polyps and presenting them to the physician, boosting their sensitivity [15, 16].

Generally, CAD systems for detecting colonic polyps are usually composed of three major successive parts. In the first part, the original image is segmented in different parts corresponding to anatomical organs. In the second part, a fast and efficient colorectal polyp detection algorithm finds as many polyps as possible — along with some other irrelevant points on the colon wall. This is the candidate generation step. In the third part, the candidates are filtered by complex descriptors that rule out most non-polyp structures.

Since the candidate generation runs on the entire colon wall, it is usually much simpler than the filters downstream. It should be as sensitive as possible while generating a reasonable amount of candidates (typically orders of magnitude less than the original number of voxels on the colon wall). In the next stage, filters use much more elaborate, specific polyp descriptors. They are also much slower, but they run only on the points selected by the candidate generator, not on the entire colon surface.

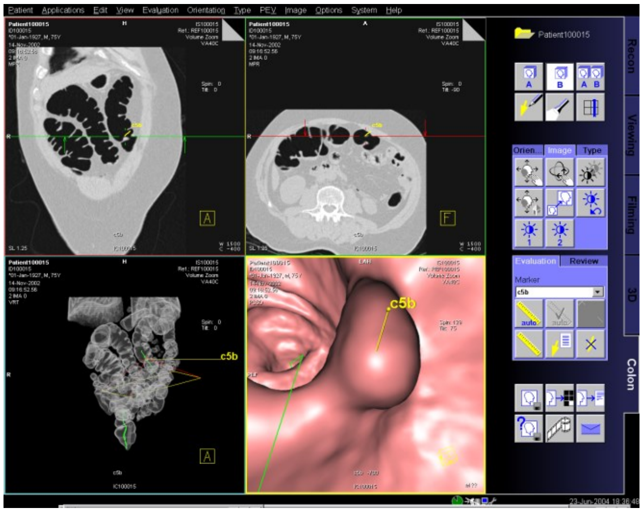

Patient Applications Edit View Evaluation Orientation Type PEV Image Options System Help
Patient100015
Recon
Viewing
Filming
3D
Colon
Orien. Image Type
Evaluation Review
Marker
c5b
c5b

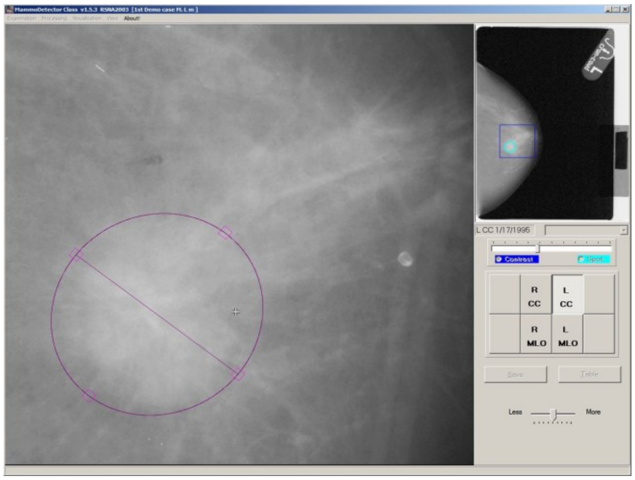
Figure 2: CT Scan identifying a polyp in the Colon.

18.3 Breast

Breast cancer is the second most common form of cancer in women, after non-melanoma skin cancer [17]. Breast cancer is the number one cause of cancer death in Hispanic women. It is the second most common cause of cancer death in white, black, Asian/Pacific Islander, and American Indian/Alaska Native women.

In 2005 alone 186,467 women and 1,764 men were diagnosed with breast cancer; 41,116 women and 375 men died from the disease.

X-ray Mammography is now accepted as a valid method for breast cancer screening, after many years in which its effectiveness was questioned [38], but spite FDA approval and despite securing reimbursement from most insurance payers, the role of Computer Aided Diagnosis (CAD) in screening mammography still remains controversial [39]. It has been clearly demonstrated that the use of CAD increased the sensitivity of detection [40], and could advance the time of detection in previously overlooked cancers [41]. However, it is unclear whether earlier detection can significantly impact morbidity and mortality of breast cancer [42]. Other studies suggested that a 2 Mammography CAD algorithm with varying filtering levels better measure of the effectiveness of mammography would be the identification of small invasive cancers [43], since it would impact mortality from breast cancer. However, small cancers are more difficult to identify and characterize because of their non-specific appearance in conventional mammography, and a CAD algorithm should, in particular, assist the radiologist in diagnosing these cancers. Since CAD devices also generate false marks, the radiologist frequently dismisses CAD prompts for these small cancers. The chances that the radiologist would accept CAD prompts for small cancers should be increased by reducing the number of false marks generated by the CAD algorithm, as the radiologist's attention is drawn to the true malignant findings. Furthermore, the addition of classification capabilities could potentially improve the efficacy of such systems by calculating the level of suspicion of any finding either detected by the first tier of the system, or considered suspicious by the radiologist.



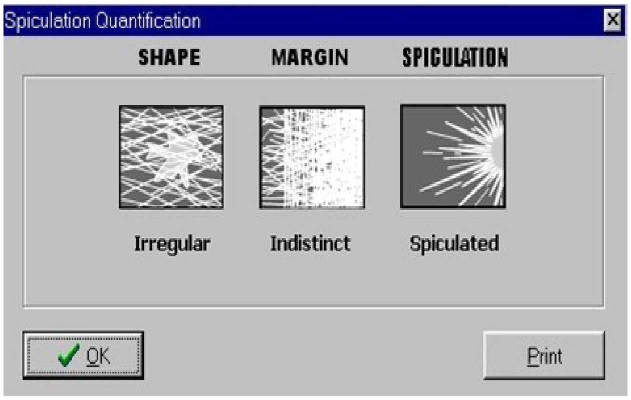


Figure 3: The system-generated BI-RAD descriptors displayed to the user for a malignant mass

18.4 PE

In recent years, there has been active research and development in computer aided diagnosis (CAD) in medical imaging across academia and industry, leading to the FDA approval for clinical use of CAD systems. We developed a fast yet effective approach for computer aided detection of pulmonary embolism (PE) in CT pulmonary angiography (CTPA). Our research has been motivated by the lethal, emergent nature of PE and the limited accuracy and efficiency of manual interpretation of CTPA studies.

PE is a sudden blockage in a pulmonary artery caused by an embolus that is formed in one part of the body and travels to the lungs in the bloodstream through the heart. PE is the third most common cause of death in the US with at least 600,000 cases occurring annually. It causes death in about one-third of the cases, that is, approximately 200,000 death annually. Most of the patients who die do so within 30 to 60 minutes after symptoms start; Many cases are seen in the emergency department.

Treatment with anti-clotting medications is highly effective, but sometimes can lead to subsequent hemorrhage and bleeding, therefore, the anti-clotting medications should be only given to those who really need. This demands a very high specificity in PE diagnosis. Unfortunately, PE is among the most difficult conditions to diagnose because its primary symptoms are vague, non-specific, and may have a variety of other causes, making it hard to separate out the critically ill patients who suffer from PE. PE cases are missed in diagnosis more than 400,000 times in the US each year. If pulmonary embolism can be diagnosed and appropriate therapy started, the mortality can be reduced from approximately 30 percent to less than ten percent; roughly 100,000 patients die who would have survived with the proper and prompt diagnosis and treatment. A major clinical challenge, particularly in an ER (Emergency Room) scenario, is to quickly and correctly diagnose patients with PE and then send them on to treatment. A prompt and accurate diagnosis of PE is the key to survival.

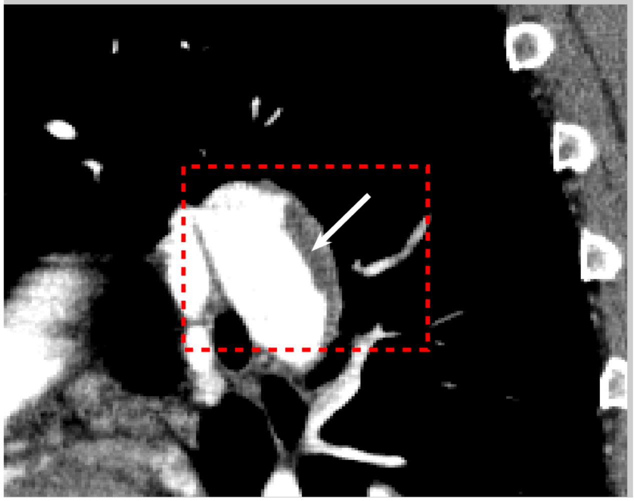


Figure 4: Highlighted Pulmonary embolism in the Lung

18.5 Cardiac

Cardiovascular Disease (CVD) is a global epidemic that is the leading cause of death worldwide (17mil. deaths per year) [8]. It is the single largest contributor to “Disability Adjusted Life Years” - 10% of DALY’s in low and middle income nations, and 18% of DALY’s in high-income nations. Hence the WHO and CDC agree that “CVD is no longer an epidemic. It is a pandemic”. In the United States, CVD accounted for 38% of all deaths in 2002 [7] and was the primary or contributing cause in 60% of all deaths. *Coronary Heart Disease* (CHD) accounts for more than half the CVD deaths (roughly 7.2 mil. deaths worldwide every year, and 1 of every 5 deaths in the US), and is the *single* largest killer in the world.

The reliable delineation of the left ventricle (LV) for robust quantification requires years of clinical experience and expertise by echocardiographers and sonographers. Even with acceptable image quality, issues such as trabeculations of the myocardium, fast-moving valves, chordi and papillary muscles, all contribute to the challenges associated with delineation of the LV. Technical issues, such as the fact that a 2D plane is acquired on a twisting 3D object, make this problem even more difficult. Limited success has been achieved in automated quantification based on LV delineation with methods that simply look for a border between black and white structures in an image.

One important application of automatic LV delineation is the automatic assessment of the left ventricular (LV) ejection fraction (EF). EF is a relevant criterion for pharmacologic, defibrillator, and resynchronization therapy, therefore, being able to automatically calculate a robust EF measure is of interest to improve clinical workflow. Currently, the method widely used in clinical practice consists on a subjective visual estimation of EF, even though it is prone to significant variability.

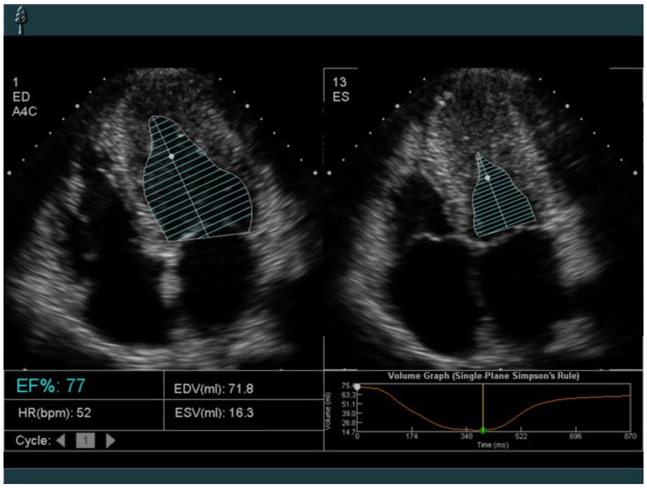


Figure 5: Automated Measurement of Ejection Fraction

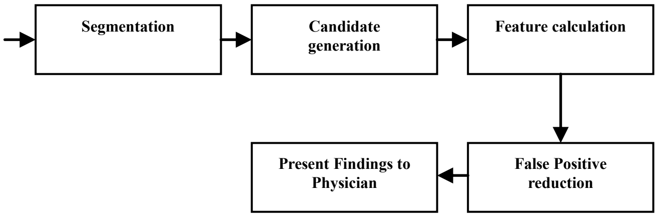


Figure 6: Common data-flow architecture of software for mining clinical-image data

19. OVERVIEW OF APPROACH

19.1 Common CAD software paradigm

In an almost universal paradigm, this problem is addressed by a 4 stage system (see fig 6):

1. Candidate generation stage which identifies suspicious unhealthy candidate regions of interest (ROI) from a medical image;

2. Image processing & Feature extraction that computes descriptive features for each candidate so that each candidate is represented by a vector of numerical values or attributes.

3. A classification stage that differentiates candidates that are true nodules from the rest of the candidates based on candidate feature vectors;

4. Visual presentation of CAD findings to the radiologist.

To keep the discussion focused on data mining, in this article we will focus on the technical challenges confronted in learning the classifier in Step 3.

Automated learning algorithms comprise an important part of the modern methodology for efficiently designing computer aided diagnostic products. In addition to improving the diagnostic accuracy, these technologies greatly reduce the time required to develop deployable systems that act as ``second readers". In the context of CAD, many standard algorithms, such as support vector machines (SVM), back-propagation neural nets, kernel Fisher discriminant, have been used to learn classifiers for detecting malignant structures [12 ,13, 18, 19]. However, these general-purpose learning methods either make implicit assumptions that are commonly violated in CAD applications, or cannot effectively address the difficulties involved in learning a CAD system, which often results in sub-optimal performance of the classifiers that they construct.

For example, traditional learning methods almost universally assume that the training samples are independently drawn from an identical — albeit unobservable — underlying distribution (the IID assumption), which is almost never the case in CAD systems. Moreover, it is common that only an extremely small portion of the candidates identified in the candidate generation stage are actually associated with true malignant lesions, leading to an unbalanced sample distribution over the normal and abnormal classes. When searching for descriptive features, researchers often deploy a large amount of experimental image features to describe the identified candidates, which consequently introduces irrelevant or redundant features. Sparse model estimation is often desired and beneficial. Medical domain knowledge which may not be learnable from limited sample data often sheds a light on the essential learning process.

Efficiently incorporating related medical knowledge into the automatic learning algorithms yields better CAD systems. Extensive research is hence required to address all these challenges in order to produce clinically acceptable CAD products.

In the rest of this section, we present our solutions to these challenges, including a general sparsity treatment for feature selection in section 3.1.6 and cascaded classification schemes in section 3.1.4 as a solution to balancing skewed data distribution as well as to the speed requirement of the CAD system. A few strategies for dealing with non-IID data will be discussed in section 3.1.5. Section 3.1.5. will discuss the problem of noisy ground truth. Scalable algorithms for optimizing large scale problems on massive candidate sets have been explored and presented in section 3.1.2. An automatic gating network is used and discussed in section 3.1.2.

19.1.1 MIL

One of the challenges in LungCAD training is to obtain reliable ground truth due to difficult lung biopsy procedure. The current process to recognize the true nodule candidates often leads to noisy labels. Candidates are labeled positive if they are within some pre-determined distance from a radiologist mark, some of the positively labeled candidates may actually refer to healthy structures that just happen to be near a mark, thereby introducing labeling errors in the training data. These labeling errors can potentially sabotage the learning process by ``confusing" a classifier that is being trained with faulty labels, resulting in classifiers with poor performance.

As shown in [20] multiple-instance-learning is one of the effective ways to deal with this problem. Using the fact that candidate generators tend to generate multiple candidates associated with the same nodule. The set comprised of the feature vectors representing candidates associated with the same nodule can be seen as a bag of multiple instances of the specific nodule, therefore the problem can be modeled as a multiple instance problem. A novel and effective general approach is thus developed in [28] and more recently [20] by forming convex hulls of the instances in each individual bag.

Let X_i represent the feature matrix associated with a specific structure in the image, such as the region corresponding to the *i*-th nodule in our ground truth database. Note that X_i consists of n_i feature vectors as columns, each representing a candidate within a critical distance to a radiologist mark. Then (X_i λ) represents a point in the convex hull formed by all column vectors in X_i given $\lambda_j \geq 0$ and $\sum_{j=1}^{n_i} \lambda_j = 1$. The main idea in [20] is to search a good representative point in the convex hull and correctly classify this point in contrast to the conventional method where all the points in X_i need to be correctly classified. This allows the classifiers to have certain degree of tolerance to noisy labels and makes use of the practical observation that not all candidates close to a nodule mark need to be identified. Mathematically, the convex-hull representation idea can be combined to many existing classification formulations depending on choices of loss functions and regularization operators. For example, the 1-norm SVM (see [21]) can be revised to incorporate the convex hull representation as follows:

$$\begin{array}{ll} \min & \mu||w||_1 + \sum_{i=1}^{K} \xi_i + \sum_{i \in C^-} \xi_i \\ \text{s.t.} & w^T X_i \lambda_i + b \geq 1 - \xi_i, \ \xi_i \geq 0, \\ & \lambda_i \geq 0, \ e^T \lambda_i = 1, \ i = 1, \cdots, K, \\ & -(w^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0, \ i \in C^-, \end{array}$$

(1)

where it has been assumed there are K nodules, and for the *i*-th nodule, multiple instances included in X_i exist.

A similar principle can be applied to other learning formulations such as linear discriminant analysis, for which a fast, globally convergent algorithm has been developed, showing competitive performance with other state-of-the-art MIL methods.

19.1.2 Gating engineering for incorporating domain knowledge

Incorporating medical knowledge and prior observations can be critical to improve the performance of the CAD system. In LungCAD, nodules have various characteristics in their shapes, sizes and appearances. The most simple example is that nodules can be very big or small. Many of the image features are calculated by averaging over the voxels within a nodule segmentation. Features calculated on a large nodule will hence more accurate than those evaluated on a small one. Consequently, it may be more insightful to construct classifiers with separate decision boundaries respectively for large nodule candidates and small candidates. ``Gating" is a technique used to automatically learn meaningful clusters among candidates and construct classifiers, one for each cluster, to classify true nodule candidates from false detections. This can obviously be extended to incorporate different kinds of knowledge, for instance, to exploit differences between the properties of central versus peripheral nodules.

A novel Bayesian hierarchical mixture of experts (HME) has been developed [22] and tested in our lungCAD study. The basic idea behind the HME is to decompose a complicated task into multiple simple and tractable subtasks. The HME model consists of several domain experts and a stochastic gating network that decides which experts are most trustworthy on any input pattern. In other words, by recursively partitioning the feature space into sub-regions, the gating network probabilistically decides which patterns fall in the domain of expertise of each expert. Each expert node is represented by a probabilistic model $p(y|\boldsymbol{x},\boldsymbol{w})$, the conditional distribution of the output class label y, given an input pattern $\boldsymbol{x}$ and the expert parameters $\boldsymbol{w}$.

Each gating node can be interpreted by a binary indicator $z \in \{0,1\}$. When $z = 1$ an input pattern x goes down the left branchof the node; otherwise, it goes down the right branch. The random variable $\boldsymbol{z}|\boldsymbol{x}$ follows a binomial distribution,

$$p(z|\mathbf{x},\mathbf{v}) = \sigma(\mathbf{v}^T\mathbf{x})^z(1-\sigma(\mathbf{v}^T\mathbf{x}))^{1-z}.$$

Where $\mathbf{v}$ are parameters of the gating node and $\sigma(\cdot)$ is the sigmoid function $\sigma(x) = 1/(1+\exp(-x))$.

Given the experts W and the 0/1 state of every gating node $z = (z_1,..,z_K)^T$, the conditional distribution of y is

$$p(y|\mathbf{x},\mathbf{z},W) = \sum_{j=1}^{M} \eta_j p(y|\mathbf{x},\mathbf{w}_j),$$

where $\boldsymbol{\eta}$=[η_1,..., η_M]T is an all-zero vector except that the *j*-th entry is equal to one, if $\boldsymbol{x}$ is assigned to expert E_j. The vector $\boldsymbol{\eta}$ encodes the gating variable $\boldsymbol{z}$. For example, the vector $\boldsymbol{\eta}=[0,1,0]^T$ if $z_1=0$, $z_2=1$, and $z_3=0$ indicating that the input pattern $\boldsymbol{x}$ goes to expert E_2.

19.1.3 Scalability for massive data

Often a great amount of candidates are commonly produced in the candidate generation stage to uncover any suspicious regions, which results in large massive training data. This impose a requirement for the scalability of the learning algorithms.

Boosting algorithms are efficient to scale up with large data. A boosting approach is proposed to incrementally solve our classification formulation based on column generation techniques [23]. The column generation approach has been widely used for solving large-scale linear programs or difficult integer programs since the 1950s [24]. The column generation boosting has been explored in [25, 26] for linear programs and later extended to solve quadratic programs with piece-wise quadratic loss functions or the l_2-norm regularization [23].

The optimization problems formed by the 1-norm nonlinear ("kernelized") formulation can be written as:

$$\begin{array}{ll} \min_{\mathbf{w},b,} & \mu||\mathbf{w}||_1 + \sum\limits_{i=1}^{\ell} \nu_i \xi_i \\ \text{s.t.} & y_i(\sum\limits_{j=1}^{d} K_{ij} w_j + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \; i = 1, \cdots, \ell, \end{array} \qquad \textbf{(2)}$$

In the context of boosting, the hypothesis $K_{\bullet j}\, w_j$ which is based on a single column of the matrix K (or the j-th feature) is regarded as a weak model or base classifier. The formulation presented above can be seen as a model that constructs the optimal classifier based on the linear combination of the weak models.

The dual of the optimization problem (2) can be derived and written as the following optimization problem in terms of the Lagrange multipliers (the dual variables) $\boldsymbol{\beta}$ [23].

$$\begin{array}{ll} \max & \sum\limits_{i=1}^{\ell} \beta_i \\ \text{s.t.} & -\mu \leq \sum\limits_{i=1}^{\ell} \beta_i y_i K_{ij} \leq \mu, \quad j = 1, \cdots, d, \\ & \sum\limits_{i=1}^{\ell} \beta_i y_i = 0, \\ & 0 \leq \beta_i \leq \nu_i, \; \forall i \end{array} \qquad \textbf{(3)}$$

Let the variables $\boldsymbol{w} = [w_1, w_2, ..., w_d]^T$ be partitioned into two sets, the working set $\boldsymbol{w}^V$ used to build the model and the remaining set denoted as $\boldsymbol{w}^N$ that is eliminated from the model as the corresponding features are not generated. In the primal space, the column generation method solves the linear program (2) restricted on a subset of $\boldsymbol{w}$ variables, i.e. $\boldsymbol{w}^V$, which means not all features (columns of the feature matrix) are evaluated at once and used to construct the model. Columns are generated and added sequentially to the problem to achieve optimality. In the dual space, the columns in the primal problem correspond to the constraints in the dual problem. When a column is not included in the primal, the corresponding constraint does not appear in the dual. If a constraint is absent from the dual problem, it is violated by the solution to the restricted problem, then, this constraint (a column in the primal) needs to be included in the restricted problem in order to obtain optimality.

19.1.4 Cascade

Typical CAD training data sets are large and extremely unbalanced between positive and negative classes. In the candidate identification stage, high sensitivity (ideally close to 100 %) is essential, because any cancers missed at this stage can never be found by the CAD system, this high sensitivity at the candidate generation stage is achieved at the cost of a high false positives (less than 1% of the candidates are true lesions), making the subsequent classification problem highly unbalanced. Moreover, a CAD system has to satisfy stringent real-time performance requirements in order for physicians to use it during their diagnostic analysis.

These issues can be addressed by employing a cascade framework in the classification approach as discussed in [27]. The method in [27] investigates a cascaded classification approach that solves a sequence of linear programs, each constructing a hyperplane (linear) classifier. The linear programs are derived through piece-wise linear cost functions together with the l_1-norm regularization condition. The resulting linear program works in the same principle as for the 1-norm SVM. The main idea is to incorporate the computational complexities of individual features into the feature selection process.

A weighted l_1 norm $\|w\|_1 = \sum\limits_j \gamma_j |w_j|$ is employed where each weight γ_j is determined by the computational cost of the j-th feature. Each linear program employs an asymmetric error measure that penalizes false negatives and false positives with different costs. An extreme case is that the penalty for a false negative is infinity, which is used in the early stage of the cascade design to alleviate the skewed class distribution and preserve high detection rates.

This approach has been compared with the well-known cascade AdaBoost, and is superior with advantages of multiple folds:

1. Easy classification: the detection problem becomes much more balanced at later stages, facilitating advanced classification algorithms to be applied and perform well at

these stages when prediction accuracy becomes more demanding at later stages.

2. High computational efficiency: early stages weed out many non-target candidates, so most stages are not evaluated for a typical negative candidate. Computationally expensive features are only calculated for a small portion of the candidates at later stages.

3. Robust system: the linear program with a l_1-norm regularization at each stage is a robust system. Although no theoretical justification is derived, a cascade of very few stages is unlikely to harm the robustness of linear classifiers, as opposed to a cascade of over 20 stages as often obtained via cascade AdaBoost.

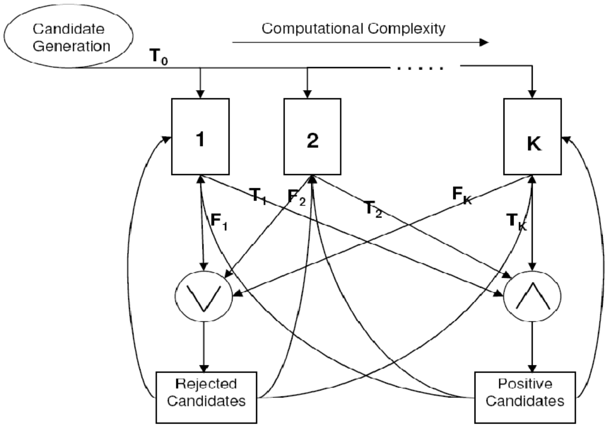


Figure 7: **The proposed cascade framework for off line training of classifiers**

in [27] follows standard cascade procedure to train classifiers sequentially for each different stage, which amounts to a greedy scheme, meaning that the individual classifier is optimized only toward the specific stage given the candidates survived from the stages prior to it. The classifiers are not necessarily optimal to the overall structure where all stages are taken into account. A novel AND-OR cascade training strategy as illustrated in Figure 7 is proposed to optimize all the classifiers in the cascade in parallel by minimizing the regularized risk of the entire system. By optimizing classifiers together, it implicitly provides mutual feedback to different classifiers to adjust parameter design. This strategy takes into account the fact that in a classifier cascade, a candidate is classified as positive only if all classifiers say it is positive, which amounts to an AND relation among classifiers. Nevertheless, a candidate is labeled as negative as long as one of the classifiers views it as negative, an OR relation of classifiers.

Let the cascade consist of *M* stages, and correspondingly *M* linear classifiers need to be constructed. To this end, we optimize the following cost function to find the optimal [$\boldsymbol{w}_1, \boldsymbol{w}_{2,\ \ldots}\boldsymbol{w}_M$]:

$$\begin{aligned} \mathcal{J}(\mathbf{w}_1,\ldots,\mathbf{w}_M) &= \mu\sum_{k=1}^{M}||\mathbf{w}_k||_1 \\ &+ \nu_1\sum_{i\in C^-}\prod_{k=1}^{M}(e_{ik})_+ \\ &+ \nu_2\sum_{i\in C^+}\max(0,e_{i1},\ldots,e_{iK}) \end{aligned} \qquad \textbf{(4)}$$

where μ is the regularization parameter, $(e_{ik})_+ = \max(0,1-y_i w_k^T x_i^k)$defines the hinge loss of the *i*-th training example *($\boldsymbol{x}_i$,y_i)* induced by classifier *k*. The index sets C^+ and C^- contain, respectively, the indices of the positive and negative examples. Notice that classifier *k* may use a specific subset of the features in $\boldsymbol{x}_i$ so we denote the feature vector used by classifier *k* as $\boldsymbol{x}_i^k$. The first term in (4) is a summation of the regularizers for each of classifiers in the cascade and the second and third terms accounts for the losses induced by the negative and positive samples respectively. Unlike traditional 1-norm SVM, the loss here is different for positive and negative samples. The loss induced by a positive sample is zero only if $\forall k : 1 - y_i w_k^T x_i^k \leq 0$ which corresponds to the ``AND" operation in Figure 2, and the loss induced by a negative sample is zero as long as $\exists k : 1 - y_i w_k^T x_i^k \leq 0$ which corresponds to an ``OR" operation.

The objective function in (4) is non convex and nonlinear, which by itself is computationally expensive to solve. An efficient alternating optimization algorithm is hence developed to solve this problem. At an iteration, we fix all the classifiers in the cascade but the classifier *k*. After dropping the fixed terms in (4) we obtain,

$$\mathcal{J}(\mathbf{w}_k) = \mu||\mathbf{w}_k||_1 + \textstyle\sum_{i\in C^-}\bar{\nu}_1\ (e_{ik})_+ \cdots \\ + \textstyle\sum_{i\in C^+}\nu_2\max(0,e_{i1},\ldots,e_{ik},\ldots,e_{iM})$$

Where $\bar{\nu}_1 = \nu_1\prod_{j=1,j\neq k}^{M}\max(0,e_{ij})$. This can be formulated as a mathematical programming problem as follows:

$$\begin{aligned} \min_{(\mathbf{w}_k,\xi_k)} \quad & \mu||\mathbf{w}_k||_1 + \bar{\nu}_1\textstyle\sum_{i\in C^-}\xi_i + \nu_2\textstyle\sum_{i\in C^+}\xi_i \\ \text{s.t.} \quad & y_i\mathbf{w}_k^T\mathbf{x}_i^k \geq 1-\xi_i,\ \forall i \\ & \xi_i \geq 0,\ \forall i. \end{aligned} \qquad \textbf{(5)}$$

The subproblem (5) is convex and differs from Problem (4) by fixing $w_j, j \neq k$. The weight assigned to the loss induced by the negative samples is now adjusted by the term $\prod_{j=1, j\neq k}^{M} \max(0, e_{ij})$. This imposes a zero loss on the candidate $\boldsymbol{x}_i^k$ in the optimization problem as long as one of the other classifiers classifies $\boldsymbol{x}_i^k$ correctly. The problem (5) can be easily optimized by any standard linear program solver.

19.1.5 Internal correlations/Non-IID

Most classification systems assume that the data used to train and test the classifier are independently drawn from an identical underlying distribution. For example, samples are classified one at a time in a support vector machine (SVM), thus the classification of a particular test sample does not depend on the features from any other test samples. Nevertheless, this assumption is commonly violated in many real-life problems where sub-groups of samples have a high degree of correlation amongst both their features and their labels. Due to spatial adjacency of the regions identified by a candidate generator, both the features and the class labels of several adjacent candidates can be highly correlated. This is true in both the training and test sets. The newly proposed batch-wise classification algorithms [29] account for these correlations explicitly.

In this setting, correlations exist among both the features and the labels of candidates belonging to the same (batch) image both in the training data-set and in the unseen testing data. Furthermore, the level of correlation can be captured as a function of the pairwise-distance between candidates: the disease status (class-label) of a candidate is highly correlated with the status of other spatially proximate candidates, but the correlations decrease as the distance is increased. Most conventional CAD algorithms classify one candidate at a time, ignoring the correlations amongst the candidates in an image. By explicitly accounting for the correlation structure between the labels of the test samples, the algorithms proposed in [29] connect between each other the class assignments of spatially nearby candidates to improve the classification accuracy significantly.

The approach in [29] derives a probabilistic batch classification model by specifying *a-priori* guess on the candidate labels with a covariance matrix **Σ** that encodes the spatial-proximity-based correlations within an image. The spatial proximity matrix $\mathbf{\Sigma}_i$ for image *i* is defined as $\Sigma_i = \exp(-\alpha D_i)$ where α is a scaling parameter and $\mathbf{D}_i$ is the matrix of Euclidean distances between candidates detected from the *i*-th CT image.

Many existing classification methods can be revised to incorporate correlations into the formulations by exploring the probabilistic batch model. For example, an SVM-based batch learning formulation solves the following optimization problem:

$$\begin{array}{ll} \min_{\mathbf{w},\xi} & \mu||\mathbf{w}||_1 + \sum_{i=1}^{P} \mathbf{e}^T \xi_i \\ \text{s.t.} & Y_i(\theta\Sigma_i + I)(X_i\mathbf{w} + b) \geq \mathbf{e} - \xi_i, \\ & \quad \xi_i \geq 0,\ i = 1, \cdots, P, \end{array}$$

where *P* is the number of training images, X_i consists of the feature vectors (as rows) of all candidates identified from the *i*-th image, and Y_i is a diagonal matrix with diagonal element equal to the labels of corresponding candidates. The matrix *I* is an identity matrix of proper dimension and $\boldsymbol{e}$ is the vector of ones. The parameter θ is tuned to adjust the influence of the proximity matrix $\mathbf{\Sigma}_i$.

The labels of the candidates from a test image *i* can be obtained by applying the model $z_i = (\theta\Sigma_i + I)\ (X_i\ \boldsymbol{w} + b)$ and taking the sign $sgn(z_i)$. For instance, the candidate generator produces 3 candidates from an image volume, $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$. Let us predict the label of the first candidate. The label depends on the sign of $z_1 = s_{11}\ y_1 + s_{12}\ y_2 + s_{13}\ y_3$ where *s* are the elements in $(\theta\Sigma + I)$ measuring the similarities between the first candidate and other candidates, and each predicted value $y_i = \boldsymbol{x}_i^T \boldsymbol{w} + b$.

19.1.6 Feature selection (building sparse models)

Feature selection has long become an important problem in statistics [30 ,31] and machine learning [32, 33], and is highly desired in CAD applications. It is a well-known that a reduction of classifier feature dependencies improves the classifier's generalization capability. However, the problem of selecting an "optimal" subset of features from a large pool (in the orders of up to hundreds) of potential image features is known to be NP-hard. An early LungCAD system [18] utilizes a greedy forward selection approach. Given a subset of features S, the greedy approach consists of finding a new single feature x_j from the feature pool that improves classification performance when considering the expanded subset of features S.

This procedure begins with an empty set of features and stops when classification performance does not improve significantly when any remaining feature is added to S. At each step, classification performance is measured based on Leave-One-Patient-Out (LOPO) cross-validation procedure [34].

Recent research has focused more on general sparsity treatments to construct sparse estimates of classifier parameters, such as in LASSO [30], the 1-norm SVM [21, 35], and sparse Fisher's discriminant analysis [36]. Assume that a linear function $f(x) = \boldsymbol{x}^T\boldsymbol{w} + b$ is used, where $\boldsymbol{w}$ and *b* are parameters to be determined, to classify candidates according to its sign *sgn(f(**x**))*. The fundamental idea for achieving sparsity is to replace the classic l_2-norm regularization term $||\boldsymbol{w}||^2$ (e.g. classic SVMs, ridge regression) with a sparse-favoring regularization term, such as the l_1-norm regularization $||\boldsymbol{w}||_1$ or the so-called 0-norm regularization condition [37]. In our formulations devised in later sections, we explore the l_1 norm regularization since the 0-norm introduces complicated combinatorial optimization problems whereas the l_1 norm often leads to scalable linear programs. The l_1 norm regularization inherently performs feature selection since penalizing on the 1-norm regularization of $\boldsymbol{w}$ drives the resulting optimal $\hat{w}$ to be sparse, meaning only a few features receive a non-zero weight. One of the well-known sparse classification approaches is the so-called 1-norm SVM [21,25] which is written as follows:

$$\begin{array}{ll} \min & \mu||\mathbf{w}||_1 + \sum_{i \in C^+ \bigcup C^-} \xi_i \\ \text{s.t.} & \mathbf{w}^T \mathbf{x}_i + b \geq 1 - \xi_i,\ \xi_i \geq 0,\ i \in C^+, \\ & -(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i \in C^-, \end{array}$$

where μ is a tuning parameter and index sets C^+ and C^- contain, respectively, the indices of the positive and negative examples.

19.2 Alternative approach for detection of shapes

Accurate analysis of the myocardial wall motion of the left ventricle is crucial for the evaluation of the heart function. This task is difficult due to the fast motion of the heart muscle and respiratory interferences. It is even worse when ultrasound image sequences are used since ultrasound is the noisiest among common medical image modalities such as MRI or CT. Figure 8a illustrates the difficulties of the tracking task due to signal dropout, poor signal to noise ratio or significant appearance changes.

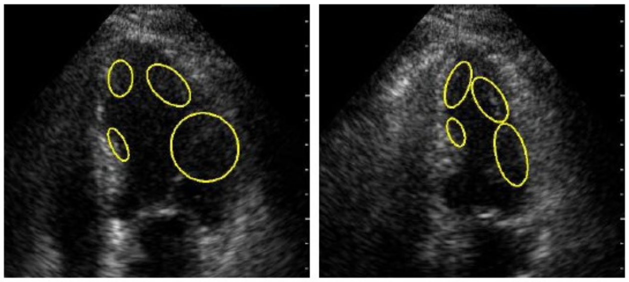


Figure 8a. Echocardiography images with area of acoustic drop-out, low signal to noise ratio and significant appearance changes. Local wall motion estimation has covariances (depicted by the solid ellipses) that reflect noise.

Several methods have been proposed for myocardial wall tracking. Model-based deformable templates [46], Markov random fields [47], optical flow methods [48], or combinations of above, have been applied for tracking left ventricle (LV) from 2-D image sequences. Jacob et al. provided a brief recent review in [49].

In [52] a unified framework was introduced for fusing motion estimates from multiple appearance models and fusing a subspace shape model with the system dynamics and measurements with heteroscedastic noise. The appearance variability is modeled by maintaining several models over time. This amounts for a nonparametric representation of the probability density function that characterizes the object appearance. Tracking is performed by obtaining independently from each model a motion estimate and its uncertainty through optical flow.

The diagram of the proposed robust tracking proposed in [52] is illustrated in Figure 8b. the approach is robust in two aspects: in the measurement process, Variable-Bandwidth Density-based funsion (VBDF) is used for combining matching results from multiple appearance models and in the filtering process, fusion is performed in the shape space to combine information from measurement, prior knowledge and models while taking advantage of the heteroscedastic nature of the noise.

To model the changes during tracking we propose to maintain several exemplars of the object appearance over time which is equivalent to a nonparametric representation of the appearance distribution.

A component-based approach is more robust that a global representation, being less sensitive to structural changes thus being able to deal with nonrigid shape deformations.

Each component is processed independently; its location and covariance matrix is estimated in the current image with respect to all of the model templates. For example, one of the components is illustrated by the rectangle in Figure 2 and its location and uncertainty with respect to each model is shown in the motion estimation stage.

The VBDF robust fusion procedure is applied to determine the most dominant motion (mode) with the associated uncertainty. The location of the components in the current frame is further adapted by imposing subspace shape constraints using pre-trained shape models. Robust shape tracking is achieved by optimally resolving uncertainties from the system dynamics, heteroscedastic measurements noise and subspace shape model. By using the estimated confidence in each component location reliable components contribute more to the global shape motion estimation.

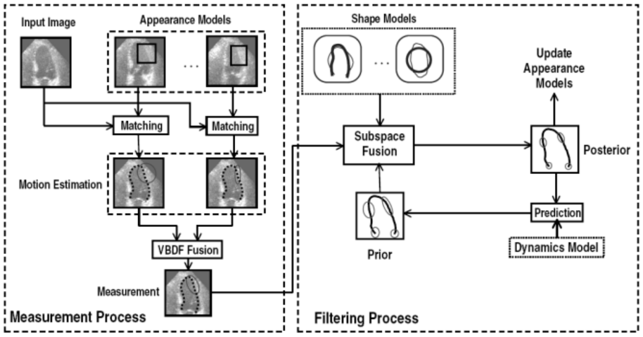


Fig 8b: **The block diagram of the robust tracker with the measurement and filtering processes**

20. CLINICAL SYSTEMS

20.1 Lung

Clinical studies have done on the LungCAD system which automatically detects lung nodules by effectively combining techniques from image analysis and machine learning. At Radiology Society of North America, several independent evaluations have been reported. One of them [1] used 102 thin-section (0.67-1mm) MDCT scans: 101 Philips (Brilliance 64, and Mx8000 IDT 16 with the D and L reconstructions), and 1 SIEMENS (Sensation 16 with the B50f convolution kernel). Each image volume was reviewed and marked by an experienced thoracic radiologist with size and certainty (or confidence). The radiologist also defined the types of each lesion as either GGNs or PSNs. In this process, 168 nodules were found.

There were 112 nodules (67 GGNs and 45 PSNs) out of 168 based on the criteria of confidence (>=50%) and size (6-30mm for GGNs and 3-30mm for PSNs) for the assessment of the lung nodule detection algorithm. 61 lesions (22 GGNs and 39 PSNs) out of 112 were proved as premalignant or malignant lesions including atypical adenomatous hyperplasia (AAH, n=6), bronchioloalveolar cell carcinoma (BAC, n=12), and adenocarcinoma (ADC, n=43). The other 51 lesions were not confirmed pathologically.

The sensitivity was estimated for GGNs and PSNs. The overall sensitivity for all ground truth is 81.25% (91/112 nodules) at 5 FPs per case. The sensitivities for each category of GGNs and PSNs are 77.61% (52/67 nodules) and 86.67% (39/45 nodules) at 5FPs, respectively. The pathology proven nodules showed 83.61% (51/61 nodules) sensitivity at 5 FPs. The pathology proven missed lesions consist of 1 AAH, 2 BACs, and 7 ADCs.

Another independent study reported at RSNA in [2] used 54 chest CT scans with known ground-glass containing nodules which were randomly selected from a pool of cases with lung nodules from two institutions (SNUBH and NYU). Two chest radiologists reviewed the cases, marking all lung nodules and characterizing them by maximum diameter, type (pure ground-glass, part-solid, and pure solid nodules), location (contact with pleural surface, vessels, or isolated), level of confidence, and level of nodule conspicuity. The radiologists afterwards reviewed CAD results, classifying them as true or false positives. Ground truth for CAD performance was provided by a third experienced chest radiologist.

A total of 447 nodules were detected by the radiologists and/or CAD. 395 nodules were considered true nodules, ranging in size between 0.8 to 41.7 mm (mean 5.5 mm) and including 234 solid, 29 part-solid and 132 ground-glass nodules. The sensitivity of CAD for nodules ≥3 mm was 50.9 % (overall), 67.5 % (solid), 60.7 % (part-solid) and 29.7 % (ground-glass). CAD resulted in an increase in reader sensitivity from 56.2 to 66 % (overall), 55.6 to 73.8% (solid) and 50.5 to 53.2% (ground-glass) for reader 1; and from 79.2 to 89.8 %, 72.2 to 91.3 % and 84.7 to 88.3 % for reader 2, respectively. False positive rates per scan were 0.89, 0.15 and 0.05 for CAD, reader 1 and reader 2, respectively. The use of CAD did not increase the number of false positives for any of the readers.

Based on these and many other clinical studies, we see technological advances have improved CAD detection of non-solid and part-solid nodules without significant increase in false positive rate. The use of CAD as a second reader improves radiologist's sensitivity for detection of pulmonary nodules, including ground-glass nodules.

20.2 Colon

In [44], the performance of the Polyp enhanced viewer system (PEV) prototype was developed as part of a study involving data sets obtained from two sites, New York University Medical Center (NYU) and the Cleveland Clinic Foundation (CCF). Both sites granted Independent Review Board approvals and all cases were de-identified (all patient identification information was removed) prior to their transfer to Siemens.

The database consisted of 150 data sets, 292 volumes from high resolution CT scanners. These included patients with polyps (positive cases; *n*=64) and patients without polyps (negative cases; *n*=86). The positive cases include a total of 92 unique polyps greater that 3.0 mm. These cases were partitioned into a working set (training set) and an unseen set (test set). The sensitivity and specificity (number of false positives) of PEV as a tool to aid in polyp detection was established with respect to radiologists' findings on CTC confirmed by concurrent fiber optic colonoscopy. The locations and dimensions of the lesions were then used in subsequent stages to automatically compute sensitivity and specificity with respect to polyp size.

There were 88 cases with 171 volumes in the training set (some cases did not have both prone and supine studies). A total of 53 unique polyps were identified in this set: 19

small (less than 6 mm), 25 mid-size (6–9 mm) and 9 large (10–20 mm).

The candidate generation stage generated an average of 48.2 candidates per volume while missing 3 small sized polyps.

In this training set, after adequate validation, we obtained a median false positive (fp) rate of 3.5 per volume. The sensitivities obtained for different ranges of polyps are as follows: small=63.1%, mid-size = 92.0%, large = 88.9%, overall = 81.1% and overall ⩾ge;5 mm = 91.2%.

For the testing set, there were 62 cases with 121 volumes. A total of 39 unique polyps were identified in this set: 18 small, 11 mid-size and 10 large.

The candidate generation stage generated an average of 51.4 candidates per volume while missing 5 small polyps and 1 medium polyp.

for this testing set, we obtained a median false positive rate of 3 per volume. The sensitivities obtained for different ranges of polyps are as follows: small = 66.7%, mid-size = 81.8%, large = 100%, overall = 82.1% and overall ⩾ ge;5 mm = 90.5%.

The PEV results of 90% sensitivity for detection of medium–large colon polyps compares favorably both with detection rates in published CTC studies and with published results from other systems.

20.3 Breast

In [45] the goal was to assess the performance of a new generation mammography algorithm designed to detect clusters, deemed actionable by expert radiologists.

In this study, 53 cases with clusters were culled from 212 biopsy proven cases collected consecutively from 3 digital mammography screening facilities. 45 of the 53 cases (25 malignant) were considered actionable by at least 2 of 3 independent expert radiologists who interpreted the cases retrospectively. These 45 cases together with 208 normal cases collected consecutively from the above facilities were run on 2 versions of a prototype detection algorithm (Siemens), designed to detect all clusters deemed actionable by expert radiologists. In the advanced algorithm, multi-step filtration is replaced by global classification of candidate micro-calcifications. Moreover, the advanced algorithm considers interdependence between various stages of the parametric clusterization process and implements automatic performance optimization. The performance of the algorithms was compared.

The advanced algorithm improved the sensitivity from 80% to 98%. For malignant lesions, the sensitivity improved from 92% to 100% while for benign lesions deemed actionable by experts it increased from 67% to 94%. The first algorithm yielded a sensitivity of 80% in dense breasts and 81% in non-dense breasts while the sensitivity of the advanced algorithm was 100% and 96%, respectively. The calcification false mark rate per view was reduced from 0.27 to 0.14.

In conclusion, it was shown that the new generation algorithm achieved the goal of reproducing the performance of expert radiologists with 98% sensitivity and very few false marks. The algorithm performed equally well in dense and non-dense breasts.

20.4 PE

Several independent evaluations of our medical images miner system for PE CAD have been performed in real clinical settings. Dr. Das and his team conducted a clinical study as reported in [3], whose objectives were to assess the sensitivity of our PE CAD system for the detection of pulmonary embolism in CTPA examinations with regard to vessel level and to assess the influence on radiologists' detection performance. Forty-three patients with suspected PE were included in this retrospective study. MDCT chest examinations with a standard PE protocol were acquired at a 16-slice MDCT. All patient data were read by three radiologists (R1, R2, R3), and thrombi were marked and stored in a database. Our MIMS-PE was applied to all patient cases, and each finding of the software was analyzed with regard to vessel level. The standard of reference was assessed in a consensus read. Sensitivity for the radiologists as well as CAD software was assessed. Thirty-three patients were positive for PE, with a total of 215 thrombi. The mean overall sensitivity for the CAD software alone was 83% (at specificity, of 80%). Radiologist sensitivity was 87% (R1), 82% (R2), and 77% (R3). With the aid of the CAD software as a second reader, sensitivities of radiologists increased to 98% (R1),

93% (R2), and 92% (R3) (p$<$0.0001). CAD performance at the lobar level was 87%, at the segmental 90% and at the subsegmental 77%. The study concluded that the detection performance of radiologists can be improved with the use of CAD for PE.

Dr. Buhmann and his team [4] evaluated our system with 40 clinical cases. The 40 cases were read by six general radiologists to mark any PE, while they were simultaneously, automatically processed by our PE CAD system in the background. An expert panel consisting of two chest radiologists analyzed all PE marks from the six readers and our CAD system, also searching for additional finding primarily missed by both, forming the ground truth. The ground truth consisted of 212 emboli. Of these, 65 (31%) were centrally and 147 (69%) were peripherally located. The readers detected 157/212 emboli (74%) leading to a sensitivity of 97% (63/65) for central and 70% (103/147) for peripheral emboli with 9 false-positive findings. CAD detected 168/212 emboli (79%), reaching a sensitivity of 74% for central (48/65) and 82%(120/147) for peripheral emboli. A total of 154 CAD candidates were considered as false positives, yielding an average of 3.85 false positives per case. Our system was designed for detecting peripheral emboli, because the central emboli can be easily detected by radiologists. The evaluation concluded that CAD detection of findings incremental to the radiologists suggests benefits when used as a second reader, especially for peripheral emboli.

Dr. Lake and his colleagues [5] examined our PE CAD system performance and investigated its influence on interpretation by resident readers. 25 patients with suspected pulmonary embolus were included in this study. 4 radiology residents first independently reviewed all CT scans on a dedicated workstation and recorded sites of suspected PE on a segmental and subsegmental arterial level, then reanalyzed all studies for a second time with the aid of the PE CAD prototype. The residents' performance for diagnosis of PE with and without PE CAD were compared to the expert read of a board-certified thoracic radiologist. At the finding level on a vessel-by-vessel basis, the performance changes are: Reader 1 from 46.7% to 52.3%; Reader 2 from 57.9% to 59.8%, Reader 3: from 100% to 100%, and Reader 4 from 91.7% to 100%, and the case level on a per-patient basis, the performance improvements are Reader 1 from 70.8% to 83.3%, Reader 2 from 79.2% to 87.5%, Reader 3 from 100% to 100% and Reader 4 from 91.7% to 100%. Overall, mean detection of PE on a vessel-by-vessel basis by resident readers was increased from 53.5 vs 58.9 (p<0.028). On a per-patient basis, the PE CAD system enabled correct PE diagnosis for an additional 4 patients overall. The mean false positive rate of the PE CAD was 2.4 per case with respect to the expert read.

We have also received feedbacks from other clinical sites with our PE CAD installations for evaluations. The consensus is that our PE CAD system is of special value in the emergency room, as it boots the physicians' confidence in negative reports and reduces missing diagnosis -- a critical issue in current PE patient managements, as diagnosis has been missed in about 70% of the cases.

20.5 Cardiac

To address these problems, Auto EF, an automatic system for automatic EF measurement from 2D images was created using state-of-the-art artificial intelligence techniques. In [8] a clinical study is presented were they test three hypothesis regarding auto EF:

1) EF produces similar results to manually traced biplane Simpson's rule;

2) EF performs with less variability than visual EF calculated by expert and novice readers;

3) EF correlates favorably with EF calculated by using magnetic resonance imaging (MRI).

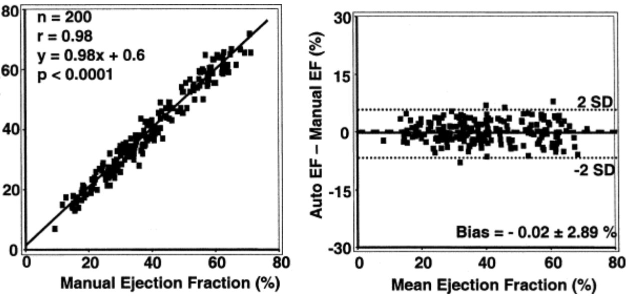
The study consisted in a group of 218 patients referred for routine transthoracic echocardiography and consisted of 165 consecutive patients with LV dysfunction (aged 65 ± 14 years, 50 women) and 53 consecutive patients with preserved LV function (aged 51 ± 20 years, 22 women). The available Data was by visual EF, by manual EF, and by Auto EF by readers that did not know the results from the other approaches.

EF was calculated by visual assessment by 1 of 6 expert readers using all available views, including a subset with contrast injections. The EFs were reported in a range of 5% EF units, from 5–10% to 65–70%, which is the clinical routine for their laboratory.

When comparing auto EF to manual biplane Simpsons' rule the two methods were closely related: r = 0.98; p < 0.01. The relationship was also favorable for single-plane EF in apical 4- and 2-chamber views (r = 0.95 and r = 0.92, respectively; p < 0.01). Absolute volumes by Auto EF also correlated with those by manual tracings (biplane LV end-diastolic volume: r = 0.94; p < 0.01; biplane LV end-systolic volume: r = 0.96; p < 0.01)

Auto EF correlated well with visual EF by expert readers (r = 0.96; p < 0.001), with a bias of 2%. The novice readers achieved similar results to that of the experts when using Auto EF (r = 0.96; p < 0.001), even though they operated Auto EF for their first time. There was significantly lower interobserver and intraobserver variability using Auto EF.

A favorable correlation was observed between Auto EF and MRI EF: r = 0.95; 95% confidence interval 0.88 to 0.98, bias −0.3%, and limits of agreement 12%.



Auto EF Versus Manual Biplane Simpson's Rule

Ejection fraction (EF) scatter plots with linear regression **(left)** and Bland-Altman **(right)** analyses of assessment of biplane left ventricular EF by Auto EF versus manual tracings using Simpson's rule, demonstrating a close correlation and narrow limits of agreement.

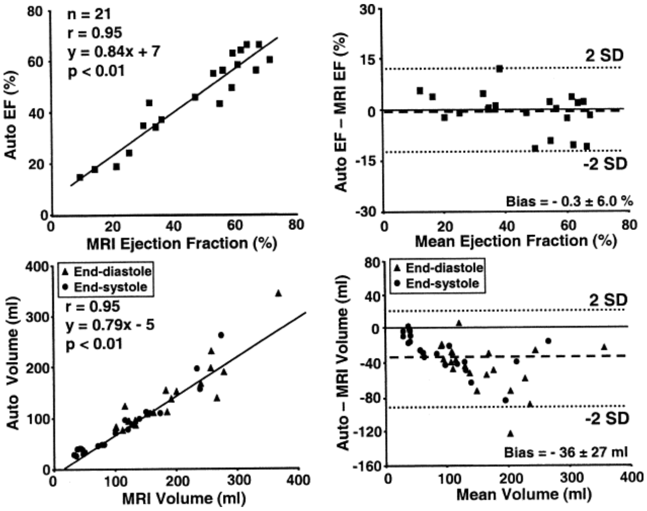


Figure 9. **Auto EF Versus Magnetic Resonance Imaging.** Scatter plots with linear regression **(left)** and Bland-Altman **(right)** analyses of assessment of left ventricular ejection fraction (EF) by Auto EF versus magnetic resonance imaging (MRI)**(top)** and combined end-systolic and end-diastolic volumes from the same patients **(bottom)**, demonstrating significant correlations but underestimation of absolute volumes.

21. CONCLUSIONS & LESSONS LEARNT

In an era of dramatic medical advances, radiologists now have access to orders of magnitude more data for diagnosing patients. Paradoxically, the deluge of data makes it more difficult & time consuming to identify key clinical findings for improving patient diagnosis & for therapy selection. This paper describes a suite of related software for mining medical images to identify & evaluate suspicious structures such as nodules, possible polyps, possibly early stage breast cancers (masses, clusters of micro-calcifications etc), pulmonary emboli, etc. This paper also described software for change quantification, and the quantification of key clinical information hidden in raw image data.

These systems are now commercially available worldwide from Siemens Medical Solutions USA, Inc and they have been deployed in several (tens of) thousands of hospitals. They have analyzed data from several hundred thousand patients: for example every woman above the age of 40 undergoes X-ray screening for breast cancer in many developed countries, and in the US more than half of all the X-ray screening mammography cases in the US are currently read by physicians assisted by a second-reader CAD software. With radiology data (images) expected to reach 25% of all data in the hospital within the next few years, it is critical to have key enablers like knowledge driven, role-based context sensitive data mining software in this domain. The field is undergoing explosive growth, and there is a key opportunity for data mining technologies to impact patient care worldwide.

Along the way, while developing these systems we learnt several key points that are absolutely critical for large scale adoption of data mining systems in an area where there is initially a lot of skepticism about the abilities of computerized systems. One of the key lessons was that the systems are not successful just by being more accurate. Their true measure of impact is in terms of how much they improve the radiologists in *their diagnoses of patients*, assisted by software. This raises the need for extensive validation of how the radiologists accuracy changes while using the system.

One system described in this paper (finding nodules in the lung CT) is one of only a handful of FDA Pre-Market Approved (PMA) systems. Others were deemed to have a lower risk and did not have the regulatory cost & burden of a PMA. Nevertheless every system underwent extensive clinical validation (including FDA approval) showing clinical benefit to radiologists in terms of improving their accuracy while reading medical image data.

Another key lesson learnt was the need for first principles research innovation specific to the data domain. While we initially tried off-the-shelf methods like SVMs, we quickly learnt the need to focus on the specific data domain and the key data characteristics & requirements therein. We learnt that changing from an SVM to a boosting algorithm or a neural network really was not what improved system performance in a significant manner, it was absolutely essential to carefully analyze data, visualize and re-think the fundamental assumptions, evaluate which assumptions are appropriate for the problem, and study how we can change them while still retaining mathematical tractability. For example, we realized that the data is never independent and identically distributed (i.i.d.), a key assumption that is almost universal in most of the traditional classifier design technologies such as SVMs, neural networks etc.

Driven by the needs of our data and our problem, we re-evaluated the assumptions and re-thought systems from first principles. This resulted in huge domain-specific improvements in system accuracy measures that are relevant for our products (as opposed to accuracy measures used in the data mining community based off the 0-1 loss for example). In all honesty, the initial approach of throwing a bunch of data mining algorithms at a problem and seeing what stuck simply led to initial disasters until we were humble enough to work on the problem we had rather than the method we had. This was a second key lesson for us (most of the authors were practitioners who often came fresh from grad school trained in the data mining).

A final lesson learnt from our work in this area was the need for securing buy in (and leadership) from key clinical subject matter experts in order to have them drive the product features and capabilities. Many of the key product definition ideas were a result of collaborating with radiologists who identified the key capabilities in system that should be developed – our best guesses as data mining researchers were based on what we found technologically challenging or exciting, but often a feature which was much less time consuming and "cool" added much more value to our customers. The lesson was that while there is a huge value addition that data mining can bring to the table, it should be defined in collaboration with the end-user in order to fully exploit it.

22. ACKNOWLEDGMENTS

Any commercial product relies on the dedication of a large team. We are particularly to our Siemens colleagues: Alok Gupta, Ph.D., Ingo Schmuecking, MD, Arun Krishnan, Ph.D., Marcos Salganicoff, Ph.D., Jonathan Stoeckel, Ph.D., Harald Steck, Ph.D., Maleeha Qazi, Shipeng Yu, Ph.D., Stefan Niculescu, Ph.D., and the entire SISL team.

23. REFERENCES

[1] M C Godoy, T Kim, J P Ko, C Florin; A K Jerebko, D P Naidich, et al. 2008. Computer-aided Detection of Pulmonary Nodules on CT: Evaluation of a New Prototype for Detection of Ground-glass and Part-Solid Nodules, Abstract at Radiology Society of North America**.**

[2] S Park, PhD, Malvern, PA; T Kim, MD, PhD; V C Raykar, PhD; V Anand; A K Jerebko, PhD; M Dewan, PhD; et al, 2008, Assessment of Computer-aided Nodule Detection (CAD) Algorithm on Pathology Proved CT Data Sets. Abstract at Radiology Society of North America.

[3] M. Das, et. al. Computer-aided detection of pulmonary embolism: Influence on radiologists' detection performance with respect to vessel segments. *European Radiology*, Volume 18, Issue 7, July 2008, 1350–1355.

[4] S. Buhmann, P. Herzog, J. Liang, M. Wolf, M. Salganicoff, C. Kirchhoff, M. Reiser, C. Becker. Clinical evaluation of a computeraided diagnosis (CAD) prototype for the detection of pulmonary embolism. *Academic Radiology*, Volume 14, Issue 6, June 2007,651-658.

[5] Lake, et al. Computer-aided detection of peripheral pulmonary embolus on multi-detector row CT: Initial experience and impact on resident diagnosis. *The 106th annual meeting of the American Roentgen Ray Society*, April 30 - May 5, 2006. Vancouver, BC, Canada.

[6] American Heart Association, *Heart Disease and Stroke Statistics 2005 Update*, http://www.americanheart.org/downloadable/heart, 2005.

[7] World Health Organization, *The Atlas of Global Heart Disease and Stroke*, http://www.who.int/ cardiovascular diseases/ resources/atlas/ , 2004

[8] M. Cannesson, M. Tanabe, M. Suffoletto, D. McNamara, S. Madan, J. Lacomis, J. Gorcsan. A Novel Two-Dimensional Echocardiographic Image Analysis System Using Artificial Intelligence-Learned Pattern Recognition for Rapid Automated Ejection Fraction. Journal of the American College of Cardiology 16 January 2007 (volume 49 issue 2 Pages 217-226 DOI: 10.1016/j.jacc.2006.08.045)

[9] American Lung Association, Trends in lung cancer morbidity and mortality report. 2006

[10] A. Jemal and R. Siegel and E. Ward and T. Murray and J. Xu and M. J Thun. Cancer Statistics. CA Cancer J. Clin. Volume 57, pages 43-66,2007.

[11] Swensen, Stephen J. and Jett, James R. and Hartman, Thomas E. and Midthun, David E. and Mandrekar, Sumithra J. and Hillman, Shauna L. and Sykes, Anne-Marie and Aughenbaugh, Gregory L. and Bungum, Aaron O. and Allen, Katie L., CT screening for lung cancer: five-year prospective experience. Radiology, volume 235, 1, pages 259-265, 2005.

[12] S. G. Armato-III and M. L. Giger and H. Mac Mahon, Automated detection of lung nodules in {CT} scans: preliminary results, Medical Physics, volume 28,8 , pages 1552-1561, 2001.

[13] D. P. Naidich and J. P. Ko and J. Stoechek, Computer aided diagnosis: Impact on nodule detection amongst community level radiologist. A multi-reader study. Proceedings of {CARS 2004} Computer Assisted Radiology and Surgery, pages 902 -907, 2004

[14] B. Levin, Colorectal cancer screening: from fecal DNA to virtual colonoscopy, in Proc. of the 95th Annual Meeting of the AACR, 2004.

[15] P. Cathier, S. Periaswamy, A. Jerebko, M. Dundar, J. Liang, G. Fung, J. Stoeckel, T. Venkata, R. Amara, A. Krishnan, B. Rao, A. Gupta, E. Vega, S. Laks, A. Megibow, M. Macari, and L. Bogoni, CAD for polyp detection: an invaluable tool to meet the increasing need for colon cancer screening," in Proc. of CARS'04, pp. 978–982, 2004.

[16] . L. Bogoni, P. Cathier, A. Jerebko, S. Lakare, M. Dundar, J. Liang, S. Periaswamy, M. Baker, and M. Macari, Computer-aided detection (cad) for ct colonography: a tool to address a growing need," BJR 78, pp. 52–62, 2005.

[17] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2005 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2009. Available at*: http://www.cdc.gov/uscs.*

[18] M. Wolf, A. Krishnan, M. Salganicoff, J. Bi, M. Dundar, G. Fung, J. Stoeckel, S. Periaswamy, H. Shen, P. Herzog, and D. P. Baidich. CAD performance analysis for pulmonary nodule detection on thin-slice MDCT scans. In H.U. Lemke, K. Inamura, K. Doi, M.W. Vannier, and A.G. Farman, editors, *Proceedings of CARS 2005 Computer Assisted Radiology and Surgery*, pages 1104–1108, 2005.

[19] J. Fu, S. Lee, S. Wong, J. Yeh, A. Wang, and H. Wu. Image segmentation, feature selection and pattern classification for mammographic microcalcifications. *Comput. Med. Imaging Graph*, 9: 419–429, 2005.

[20] G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao. Multiple instance algorithms for computer aided diagnosis. In *Advances in Neural Information Processing Systems*, 2006.

[21] P. S. Bradley , O. L. Mangasarian and W. N. Street. Feature Selection via Mathematical Programming,, INFORMS Journal on Computing}, volume 10, 2, pages 209-217, 1998.

[22] Y. Xue, X. Liao, B. Krishnapuram, and L. Carin, Bayesian Hierarchical Mixture of Experts for Pattern Classification, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), October 2005.

[23] J. Bi and T. Zhang and K. Bennett, Column-generation boosting methods for mixture of kernels, Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 521-526. 2004.

[24] S. G. Nash and A. Sofer, Linear and Nonlinear Programming, McGraw-Hill publisher, New York, NY, 1996.

[25] Kristin P. Bennett, Ayhan Demiriz and John Shawe-Taylor. A Column Generation Algorithm for Boosting, Proceedings of the 17th International Conference on Machine Learning , pages 65--72, 2000.

[26] Ayhan Demiriz, Kristin P. Bennett and John Shawe-Taylor, Linear Programming Boosting via Column Generation, Machine Learning , 46, 1--3, pages 2 25—254, 2002.

[27] J. Bi, S. Periaswamy and K. Okada, T. Kubota, G. Fung, M. Salganicoff and R. B. Rao", Computer aided detection via asymmetric cascade of sparse hyperplane classifiers, Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.

[28] O. L. Mangasarian and E. W. Wild, Multiple Instance Classification via Successive Linear Programming, Journal of Optimization Theory and Applications 137(1), 2008, 555-568.

[29] V. Vural, G. Fung , B. Krishnapuram, J. Dy and R. B. Rao, Batch-wise classification with applications to computer aided diagnosis, Proceedings of European Conference on Machine Learning, 2006.

[30] R. Tibshirani, Regression selection and shrinkage via the lasso, Journal of the Royal Statistical Society Series B, 58, 1, pages 267--288, 1996.

[31] T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction, publisher New York: Springer, 2001.

[32] I. Guyon and A. Elisseeff. An introduction to variable and feature selection, J. Mach. Learn. Res., 3, 2003

[33] M. Dash 1, H. Liu2, Feature Selection for Classification, Intelligent Data Analysis, 131–156. 1997.

[34] M. Dundar, G. Fung, L. Bogoni, M. Macari, A. Megibow and R. B. Rao, A methodology for training and validating a CAD system and potential pitfalls, Proceedings of CARS 2004 Computer Assisted Radiology and Surgery, 2004.

[35] Ji Zhu, Saharon Rosset, Trevor Hastie and Rob Tibshirani. 1-norm Support Vector Machines, Advances in Neural Information Processing Systems 16, 2004.

[36] M. Dundar, G. Fung , J. Bi, S. Sandilya and R. B. Rao, Sparse Fisher discriminant analysis for computer aided detection, Proceedings of {SIAM} International Conference on Data Mining, 2005.

[37] J. Weston, A. Elisseeff , B. Scholkopf and M. Tipping, Use of the Zero-Norm with Linear Models and Kernel Methods, Journal of Machine Learning Research, pages 1439-1461, 2003.

[38] Gotzsche PC, Olsen O. Is screening for breast cancer with Mammography justifiable?. Lancet, 355 ,129-134. 2000.

[39] Fenton JJ, Taplin SH, Carney PA, et al, Influence of computer aided detection on performance of screening mammography, N Engl J Med, 356, 1399-1409, 2007.

[40] Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center, Radiology, 220, 781-786, 2001.

[41] Warren Burhenne LJ, Wood SA, D'Orsi CJ et al. Potential Contribution of Computer-aided Detection to

the Sensitivity of Screening Mammography, Radiology, 215,554-562, 2000.

[42] Feig SA, Auditing and benchmarks in screening and diagnostic mammography, Radiol Clin North Am, 45, 791-800, 2007.

[43] Tabar L, Tony Chen HH, Amy Yen MF et al. mammographic tumor features can predict long-term outcomes reliably in women with 1-14 mm invasive breast cancers, Cancer, 101, 1745-1759, 2004.

[44] L Bogoni, P Cathier, M Dundar, A Jerebko, S Lakare, J Liang, S Periaswamy, M E Baker and M Macari, Computer-aided detection (CAD) for CT colonography: a tool to address a growing need, British Journal of Radiology 78, S57-S62, 2005.

[45] Bamberger P., Leichter I., Merlet N., Fung, G. and Lederman R, A New Generation Algorithm for Digital Mammography Designed to Reproduce the Performance of Expert Radiologists in Detecting Actionable Clusters. European Radiology Supplements, 18, 1, 2008.

[46] Cootes, T., Taylor, C.: Statistical models of appearance for medical image analysis and computer vision. In: Proc. SPIE Medical Imaging. 236–248, 2001.

[47] Mignotte, M., Meunier, J., Tardif, J.C.: Endocardial boundary estimation and tracking in echocardiographic images using deformable templates and markov random fields. Pattern Analysis and Applications,4, 256–271, 2001.

[48] Mailloux, G.E., Langlois, F., Simard, P.Y., Bertrand, M.: Restoration of the velocity field of the heart from two-dimensional echocardiograms. IEEE Trans. Medical Imaging, 8, 143–153, 1989.

[49] Jacob, G., Noble, J., Behrenbruch, C., Kelion, A., Banning, A.: A shape-space-based approach to tracking myocardial borders and quantifying regional left-ventricular function applied in echocardiography. IEEE Trans. Medical Imaging, 21, 226–238, 2002.

[50] Stanton, Arthur "Wilhelm Conrad Röntgen On a New Kind of Rays: translation of a paper read before the Würzburg Physical and Medical Society, 1895" Nature 53 (1369): 274–6

[51] Ambrose J, Hounsfield G. Computerized transverse axial tomography. Br J Radiol. 1973 Feb;46(542):148-9

[52] Bogdan Georgescu, Xiang Sean Zhou, Dorin Comaniciu, Bharat Rao, Real-Time Multi-Model Tracking of Myocardium in Echocardiography using Robust Information Fusion. MICCAI, 2004.

Data mining for intelligence led policing

ir. RCP van der Veer
Sentient

H.T. Roos MSc
Amsterdam Police force

A. van der Zanden MSc
Amsterdam Police force

Singel 160, 1015 AH, Amsterdam, The Netherlands
+31 (0)20 5 300 325
rvdveer@sentient.nl

ABSTRACT

The benefit of data mining for police seems tremendous, yet only a few limited applications are documented. This paper starts with describing the implementation problems of police data mining and introduces a new approach that tries to overcome these problems in the form of a data mining system with associative memory as the main technique. This technique makes the system easier to use, allows uncomplicated data handling and supports many different data types. Consequently, data preparation becomes easier and results contain more information. A number of Dutch police forces have already been using this system for several years with over 30 users. Since the analytical process within the police is very knowledge-intensive, a high level of domain expertise is essential, which makes it harder to find a police data miner with sufficient domain knowledge plus technical skills in the area of databases, statistics and data mining. The police domain also has data quality issues and a very diverse information need. This is why the system design tries to reduce the need for technical skills as much as possible by working with one standard datawarehouse, techniques that can be configured automatically and active user guidance. The ease of use is also ensured by integrating many tools and techniques from statistics, business intelligence and data mining into one interactive environment that does not require the analytical process to be designed beforehand. Instead, the analysis is performed through step by step interaction. This paper discusses the benefit of police data mining, the design of the system, a number of practical applications, best practices and success stories. Experiments have shown a factor 20 efficiency gain, a factor 2 prediction accuracy increase, a 15% drop in crime rate, and 50% more suspect recognition.

General Terms

Algorithms, Management, Measurement, Design, Experimentation, Security, Human Factors.

Keywords

crime, police, data mining, prediction, GIS, hot spots, spatial, public safety, analysis

24. INTRODUCTION

In this digital era, police forces have access to a rapidly growing amount of data. Combined with the dynamic nature and complexity of criminal behavior, this sets the stage for successful data mining applications. Still, examples of consistently used police data mining implementations are scarce. In this paper we discuss the practical application of a standard police data mining system as used by a growing number of Dutch police forces. The system has been developed in a group effort of data mining software company Sentient and the Dutch police forces *Amsterdam-Amstelland, Midden en West Brabant* and *Brabant-Noord.* It consists of an integrated data mining tool called *DataDetective*, which is being developed and applied by Sentient since 1992, and an extensive datawarehouse containing data from various police systems and external sources, such as weather data, geographical data and socio-demographics. During the eight years of practical application, the data mining system has continuously been evaluated, improved and extended. The examples given in this paper illustrate how data mining now plays an important role at operational, tactical and strategic levels of decision making.

The main key to success has been a strong focus on simplicity. After a short training, selected police officers can quickly discover patterns and trends, make forecasts, find relationships and possible explanations, map criminal networks and identify possible suspects. Expert knowledge on statistics or data mining is not required. Additionally, a much larger audience for data mining results is reached through weekly reports containing statistics, prediction maps, crime clusters, trends and lists of suspects. These reports are automatically produced by the data mining system.

In Section 2 we start off by discussing why it is important for police to apply data mining, followed by a description of the shortcomings of traditional systems in section 3. Section 4 then describes the system we have built including how its use is organized. In section 5 we discuss various applications of the system, followed by a number of success stories in section 6. We finish with discussions in section 7 about the added value of data mining, the challenges for implementation and future work.

25. POLICE NEED FOR DATA MINING

The increasing adoption of the *Intelligence-Led Policing* model [1] puts analysis at the heart of operational, tactical and strategic decision-making. In this model, intelligence serves as a guide to operations, rather than the reverse. Therefore, it is now more important than ever to find out how data mining can help create better understanding and predictions.

Traditionally, police systems focus on small parts of the available data (e.g. year, month, type of crime) for a specific purpose (e.g. monitoring crime rates for strategy). Without data mining, the amount of data used in analysis is limited by the time that analysts have to go through it, step by step. It is simply unfeasible to analyze all the potentially useful data by hand. However, for many policing problems it is important to use as much data as possible, to be able to explain, understand, link and predict. The explanation of a phenomenon (e.g. the sudden increase of pickpocket activity) typically lies in small details, for example in the fact that in the recent period street festivals took place, with many potential pickpocket victims on the streets. This shows that by using more data, patterns offer more contextual information and help analysts reach the right conclusions. So, to understand crime, data is needed that goes beyond simple aspects of an incident or person, e.g. the type of neighborhood, the Modus Operandi (MO or manner of working), witness descriptions, stolen goods, vehicles involved and the background of the people involved (history, crime profile, socio-demographic profile). Linking crimes by similarity also benefits from rich data for the same reasons. Finding links can help to detect crime series, connect cases and solve them. This is where data mining comes in. Automated pattern recognition is necessary to turn the data overload into a manageable flow of information that matters.

Apart from handling the volume of data, data mining techniques also help to deal with the dynamic nature and complexity of criminal behavior. This can be seen apart from the data volume subject discussed above, simply because complex patterns can be present in just a few data elements. For example, the following question is very hard to answer using conventional techniques: where and when do crimes take place during the week looking at X- Y co-ordinates of the incidents, the time of day and the day of the week. A proven method to answer this question is to use clever clustering techniques (see 5.1).

It is a common misconception that data mining requires large data volumes in order to add value. On the contrary, it is our experience that when implementing data mining it is best to start with one, maybe two data sources to get acquainted with the possibilities and to manage expectations. Many organizations are surprised by how much information they can gain from data mining on just a small part of their data. Nevertheless, more data is always better for providing more depth and more context.

To conclude, in theory data mining allows the police to better understand and predict crime because many data sources can be analyzed and complex patterns can be found. In 2004, an extensive study by the program bureau of the Dutch Police (ABRIO) concluded that 'data mining enables more effective and goal driven decision making on strategic, tactical and operational levels' (internal report).

26. SHORTCOMINGS OF CRIME ANALYSIS SYSTEMS

Traditional crime analysis tool sets suffer from the following issues:

1. Based on selecting variables

Traditional analytical tools require the analyst to look at variables one by one. This way of working is not viable for rich data sets containing many variables.

2. Static results

Existing systems usually generate static reports that do not allow interaction. They cannot be used to find the explanations behind the numbers they present.

3. Based on simple patterns

When an analyst focuses on one or two variables, the traditional tools allow only the analysis of those individual variables. In rare cases, the interaction between two chosen variables is analyzed, but all other combinations are not used which is why useful interactions between two or more variables can be overlooked.

4. Difficult extraction

It is typically hard to extract data from police source systems because of old and diverse database systems with data models based on transactions instead of analysis. There are several standard analytical applications working on small extractions, but when analysis needs to go a step further, the analyst faces a challenge in getting the right data from the systems. The fact that there are many different systems in the organization adds to the challenge and so does poor data quality. Typically, extraction, linking, correction and preparation need to be carried out for each analysis. To address this, a small number of Dutch police forces have implemented datawarehouses.

5. Diverse tool sets

Analysts have access to a range of tools, each with their own focus, look and feel and different method of reading data and creating results. There are tools for geographic visualization, for statistical analysis, for creating charts, for defining queries, for making reports, for analyzing criminal networks, for monitoring crime rates, etc. This requires analysts to learn all these tools and slows down the process because of the effort needed to transfer data between tools. There is no tight coupling that allows analysts to jump from technique to technique.

6. Tool complexity

The available tools for statistical analysis require special training and sometimes even an education in mathematics or statistics. Often, analysts do not have this background.

Do data mining tools solve these issues? The previous section argues that data mining offers tremendous added value for police. Indeed, data mining tools solve some of the problems mentioned above, by not requiring the selection of variables and the ability to find complex patterns, but they also make other problems worse by increasing tool diversity (yet another tool),

tool complexity (datamining expertise required) and difficulties with extraction (techniques put new demands on data). The result is that there are many strong criteria for police data mining users to be effective: they need to be well-trained in IT, data mining, statistics, and have domain knowledge. In other words; they have to know police databases, how to extract databases, how to prepare data, how to use different analytical tools, to design a mining process, to select variables, to correct missing values, to pick the right techniques, to set the right parameters, and to know about psychology, criminals, society and police work. In addition, police data typically challenges the user because of quality issues.

We believe that the high demands for users of standard data mining tools are the main reasons why there seem to be just a few successful police mining applications ([2][3][4][5][6]) and most of these applications are either academic endeavors or small applied projects - not continuous activities. Furthermore, when these applications do appear to be ongoing activities they seem to be limited to a single police force.

27. SYSTEM OVERVIEW

The previous section argues that standard data mining tools are difficult to make continuously useful for police experts. With this in mind, our design philosophy for the data mining system has been to just require users to know their domain and to have analytical skills. No more knowledge and skills are required. The developed system brings various techniques from business intelligence, statistics, machine learning and GIS together in a comprehensive data mining infrastructure. This infrastructure includes a datawarehouse, a reporting module and a desktop tool to provide easy query definition, matching, data visualization, basic statistics, clustering, modeling for predicting, modeling for explaining, link analysis, geographic profiling and geographic visualization.

The following subsections discuss the key system elements, the contents of the database and the way in which the data mining system is used.

27.1 Key system elements

The shortcomings listed in the previous section were used to define requirements during the design of the data mining system. The resulting system has the following characteristics:

- **Ready database**

The requirement for expert IT and database knowledge is reduced by providing a single all-embracing database in which all the data has been extracted, linked, cleaned and augmented. The aim is that the database covers 99% of the information need. The remaining 1% needs dedicated data extraction and preparation. The extensive database works like a *single point of truth*: all analysts use the same standard data and definitions.

- **Automated data mining**

The requirement for expert statistical and data mining knowledge is reduced by automated selection and configuration of data mining techniques. Based on the task and the data, the tool chooses the right technique and optimizes parameters based on the data. Furthermore, the tool assists the user by watching for typical pitfalls, such as unreliable patterns and too many missing values. In some cases, a data mining expert would perhaps surpass the automated selection and configuration. This is the compromise that must be made to let data mining novices mine data. Still, when data mining experts use the tool, they save time and their quality of work is more consistent.

- **User friendly interface**

An intuitive and graphical user interface is provided, including a task-based setup, instead of a technique-based design.

- **Interactive analysis**

The system works like an interactive analytical instrument in which every part of the results is clickable to 'zoom in'. In this way, the user can simply embark on an analytical journey without the need to first design the process, as is required in typical workflow-based data mining tools. To support this intuitive and ad hoc process, visualization is an important aspect of the user interface. These interactive possibilities support the discovery process and enhance the creativity and instinct of the analyst. In addition, they allow for interactive sessions with people requiring the information (e.g. someone in charge of an investigation). By working together at critical moments in the analytical process, the real question can be refined, new questions can be answered immediately and patterns can be selected based on their relevance.

- **Traceability**

Although the user is working interactively, it is important to keep track of the steps that were taken to reach a result, especially because such documentation can be required in court. Therefore, the system keeps track of the history of each result.

- **Data flexibility**

Associative memory [14][7] is used as the main technique for prediction, clustering and matching in the data mining system. This means that input data is matched to a representation of training data using a similarity principle, as in *Self Organizing Maps* [17] but with a much wider acceptance of different data types. Because of this, the mining process becomes easier and data usage richer:

1. Hardly any data preparation is required because the associative memory can handle a wide range of data types, such as symbolic data, cyclic ordinals (e.g. day of the week), lists, texts and categories with many values. As long as similarity between values can be calculated, a data type can be used. Furthermore, missing values do not need to be removed or guessed since similarity metrics are able to leave out the missing values in the calculation.
2. Because the technology can handle a wide range of data types, much more information is included in the mining process than would be feasible with other techniques because of their more strict data requirements. Working with non-standard data types when using other techniques is either impossible, introduces performance or training problems, or requires much work for data preparation.
3. The associative memory is able to explain how it reached a result by displaying relevant cases or persons from the memory - which is an intuitive way of explaining a decision to any user.

4. Building associative memory models converges well and fast, compared to for example back propagation neural networks [14].
5. Associative memories are robust against suboptimal parameters (*graceful degradation*) and therefore suitable to use in a situation where parameters are set automatically and the user is not an expert on optimizing the technique [14]

In other words: the user does not need to worry about fine tuning parameters, retraining, solving missing values, variable selection, and decoding variables into a usable form.

There is no free lunch, so there is a price to pay here. First: execution time of associative prediction models is not as fast as other techniques. In police practice, this does not pose a problem because the longer waiting times typically occur at the end of an analytical process, e.g. when there is time to wait for results from a batch prediction job. Furthermore, associative clustering is very fast. Second, other techniques produce more accurate models in some cases and in some cases the associative memory outperforms the rest. We found that the occasional differences in model quality do not outweigh the advantages of saving analyst time, the ability to use rich data and that non-experts are enabled to mine data.

- **Integration**

By integrating most of the tools and techniques into one tool, there is more consistency in usability, the user is no longer required to install and learn several tools and no longer needs to exchange results between tools by exporting and importing. The system features more than just techniques from data mining, ranging from simple data browsing to advanced OLAP analysis. For some types of results, popular standard tools have been linked in such a way that results can be exchanged automatically. Examples of these linked applications are Excel^{TM}, MapInfo^{TM}, Microsoft Word^{TM}, Cognos Reportnet^{TM}, Analyst's Notebook^{TM}, Weka and Google Maps$^{TM.}$. Many analysts are already familiar with these tools.

- **Geo-spatial analysis**

The spatial aspect of crime is obviously important and therefore the data mining system is able to visualize results on maps and to use spatial aspects (e.g. co-ordinates, ground use, and census data) in models.

- **Automated routine work**

The data mining system features a reporting module that creates a report for every district and for every priority crime type. These reports contain the following elements:

1. Hot spot maps of the recent period.
2. Temporal hot spot maps to show what changed.
3. Prediction maps of the upcoming period.
4. A where/when analysis with description of the clusters found.
5. A predicted week-distribution of crime over the upcoming period: on what days and times is the highest crime rate to be expected?
6. Crime rate graphs with basic statistics, trends and key performance indicators.
7. Hot shot lists of the most frequent offenders with their social network status and photographs.
8. Maps showing residence and activity areas of offenders.

- **Best practice sharing**

The system allows users to store their best practices in the form of recipes: descriptions of problems with the steps taken to solve them. These best practices can be looked up and reused by all users.

27.2 Datawarehouse

The data mining system provides insight into thousands of variables from various police systems, census data, spatial information, weather, and lifestyle data. All this data is extracted, linked, cleaned, augmented and made available in an open datawarehouse by an automated module. This means that data does not have to be collected and cleaned before every analysis.

The datawarehouse is based on the following data sources:

1. BPS/BVH/GIDS: the main transaction systems containing incidents, goods, persons, vehicles etc.
2. HKS: a system containing more details about offenders and a longer history of crimes
3. SHERPA: extensive geographic material, used to provide more information about the scene of the crime (type of area, surrounding infrastructure etc.) and the home addresses of persons
4. CBS: Dutch socio-demographic information on neighborhood level
5. Experian: extensive socio-demographic information on zip-code level
6. Sun/Moon: information about light conditions at a given date and time
7. Events: events taking place in certain areas (e.g. football matches, festivities, holidays)
8. KNMI: Dutch weather information at a given date and province

From these sources, information is gathered at the level of incidents and persons (victims, offenders, witnesses, others involved). Because the system is able to handle many columns and works with many different data types, relational data is also gathered, e.g. a list of crime types in an offender history.

27.3 Types of use

Three types of use of the data mining system can be distinguished: personal desktop analysis, group sessions and reports. The group sessions are special cases of the personal desktop analysis in which the analyst operates the system in the company of domain experts and stakeholders, to interactively look at a problem. In this way, new questions and theories can instantly be addressed and validated.

By distributing results of automated data mining reports, a large audience can benefit from the discovery of complex patterns without having to operate the data mining system themselves. If questions rise from the reports, readers can request an interactive data mining session with a trained user.

27.4 User organization

The Amsterdam police force started using the first version of the data mining system in 2001 and now has around 30 authorized and trained users. In order to match the available techniques to the responsibilities of the users, each user has access to a specific selection of functions in the tool.

In Amsterdam, data mining users are organized in teams of two for each district. These small teams are supported by a central team of analysts and one person responsible for managing the functionality of the system and communication with the user community. This user community consists of domain experts who are either police analysts, or researchers or detectives.

The amount of training required for these users varies from one to three days. New users are assessed to determine if they meet the requirements necessary to become a successful data mining user. This assessment is more focused on intellectual abilities plus analytical and communicative skills, rather than on technical knowledge and education. Simply put, the profile of the data mining user is: a clever person with analytical insight, good with numbers, experience with computer, good verbal and reporting skills, and knowledge of the police domain. Users are not required to have a background in statistics, data analysis or databases. It is a plus though, when they know the data model.

Importance of domain experts using data mining

Extensive police domain knowledge is essential for finding useful patterns and interpreting them. An example: a data mining expert from outside the police force was asked to analyze the activity pattern of drug-related crimes in and around a shopping centre. The expert found regularity but thought it wasn't interesting because there didn't seem to be a cause and effect. However, an experienced police officer noticed that this particular behavior corresponded with the time table of the so-called *methadone bus* that provides prescribed drug replacements to addicts. Furthermore, because the *methadone bus* is a helpful program, the solution should not be to stop the bus from visiting the area, but to find out the problem elements in the drug-related crimes at those times and locations. It turned out that a few addicts on the methadone program consistently misbehaved in the vicinity of the bus. The plan was made to confront these individuals, letting them know their prescriptions were on the line.

28. SYSTEM APPLICATIONS

The following subsections describe the main methods by which the data mining system is applied in practice.

28.1 Spatio-temporal clusters: where and when

Knowing when and where the risk of crime is highest allows proactive and effective deployment of resources. The traditional method for spatial risk analysis is to either list the neighborhoods with the highest recent activity or create a map that essentially does the same. This method ignores the fact that the crime rate in an area depends very much on the time of day and the day of the week. For example: some areas are crowded around rush hour, some areas are crowded when pubs close and some areas suffer from burglaries on Saturday evenings because many inhabitants are not at home. To find these patterns, one could create a giant cross table to combine neighborhood, time of day and day of the week, which would be hard to interpret. Also, more location detail is required for good tactical planning. Police presence in one street may not prevent problems a street further. Therefore it is important to use exact co-ordinates, since offenders do not take the boundaries of areas into account. When co-ordinates are used instead of neighborhood, traditional methods and visual inspection fall short.

One of the most popular applications of the data mining system is the so-called 'where-and-when-analysis' that clusters recent incidents and incidents from similar past seasons on co-ordinates, time of day and day of the week. Every cluster is reported with a typical profile of the incidents including MO and descriptions of the offenders. Together with the location and time, police employees have all the information to know where and when to go and what to look for.

The where-and-when-analysis employs an associative cluster technique that we based on the principle of Metric Multi Dimensional Scaling (MMDS) [8], projecting the high dimensional space onto two dimensions in a non-linear way. This process tries to preserve the distances between incidents as determined by the associative memory technique. Contrary to factor analysis, all variance is represented in the two-dimensional result space. This leads to some distortion, but that is not a problem since the purpose of the technique is to visualize and identify clusters. The advantage of the MMDS approach is that the results are easy to interpret for non-data mining experts, as it appears as a regular scatter plot, visualizing the clusters and the relations between them.

Our cluster algorithm first employs the associative memory to calculate the distances between incidents. These distances are used to define gravitational forces between these incidents: the more incidents are alike, the stronger they want to move to another in a two-dimensional plane. Furthermore, a rotational force is introduced, as well as a force that drives away all incidents from the weighted center of the plane. These forces are then used to iterate through an optimization process in which the incidents that are alike move closer and create clouds in the plane (see figure 4). When movement drops below a threshold, the iteration stops and the resulting distances in two dimensions should resemble the multi-dimensional distances.

The results from where-and-when-analysis are of continuous value in the prevention and repression of priority crime types, which is why cluster overviews are generated automatically and included in the weekly standard reports. The cluster overview shows a hot spot map of the district with ellipses drawn where the clusters are, followed by a list of clusters. Each cluster is presented by a hot spot map, a description of time and day of the week plus a crime profile of what is typical for the cluster (e.g. type of offender, type of stolen property).

In Amsterdam, this analysis is also used to determine when and where to plan public search actions aimed at finding weapons on people in the streets. This is a co-operative activity by the city council and the Amsterdam police. Since the start of these data mining guided searches, weapon possession has dropped by 27%.

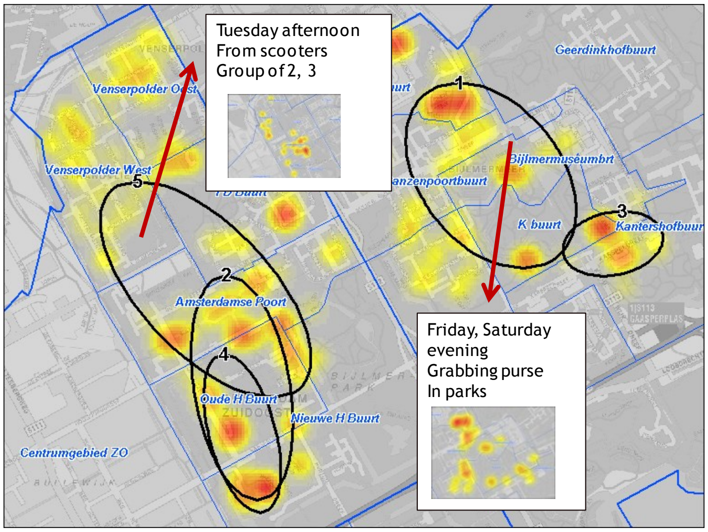


Figure1: Spatio-temporal clusters with automated profiles

28.2 Associative spatial prediction

Spatio-temporal clusters (see 5.1) find regularities in space and time that can be used as a general tactical plan for a specific period, e.g. a different suggested patrol route for every weekday during the month of November 2009. Associative spatial prediction is another crime prediction method, aimed at providing an optimal crime risk map for a specific short period; say November 1st 2009 during the 16:00 to 20:00 shift. Such a moment in time provides more context because there is a weather forecast for that day, in November the sun sets early and it is Halloween. Associative spatial prediction applies an associative memory to search for relevant situations in the past, after which these situations are superimposed to create a detailed predictive hot spot map, showing the spatial risk distribution of crime for the given future situation.

This approach is based on the theories of *repeat victimization* [9][10], *routine activity* [11][12] and *prospective hotspotting* [13]. Associative spatial prediction takes this work a step further by taking more into account than just location: recency, trends, seasonal influence, weather, time of day, day of the week, and events or holidays. Experiments have shown that for burglaries the hottest spots in the predicted maps contain 50% of all future incidents, whereas traditional methods (a hot spot map of the recent period) contain only 25% in the hottest spots.

Evaluation

Associative spatial prediction has been evaluated by specifying several random combinations of time periods for training and testing. For each combination, the prediction model was trained on one time period and tested on the other. The performance of the model was measured by calculating the error between the predicted number of crimes and the actual (future) number of crimes for each cell in a 30x30 grid covering the total area to be predicted. The total model error is the average error over all cells, averaged over the various test sets. The table below shows the model results compared to what is considered the standard method for geographic anticipation; creating a hotspot map of the crimes in the recent period (*Kernel density* – see 5.4):

Table 1: performance results of associative spatial prediction

Region	Technique	Total model mean square error
City centre of Tilburg	Kernel density	0.11
City centre of Tilburg	Associative	0.052
North Tilburg	Kernel density	0.086
North Tilburg	Associative	0.044

The table shows that the associative model outperforms the standard for these situations with about half the error. Further work is underway in looking at more situations and applying a performance measure that takes into account how the technique increases effectiveness of police work.

28.3 Analyze trends or behavior

This subsection discusses several ways to describe and explain trends or behavior using the data mining system.

28.3.1 Explain trends

Police forces respond to changes in crime rates. Often it is necessary to understand the reason behind a trend to know whether the trend can be explained by a known phenomenon or whether it needs special attention. Also, finding a probable cause allows the cause itself to be addressed.

The data mining system provides trend explanations by comparing the recent period with the period before, performing chi square and Student T-tests for all the available data columns and then sorting on the index. The result is a *profile analysis*; an enumeration of the most significant differences between the periods, to offer clues for an explanation. The system takes this a step further by providing the option to build a decision tree using the same tests in order to find *combinations* of factors as trend explanations. For example: in the end of last year, more burglaries took place in the early evening in the down town area and on Saturday mornings in the western suburbs.

The same method can be used to determine the success of counter measures. For example: the effect of installing surveillance cameras in a public area was analyzed by comparing the period after the installation with the period before. The results showed that the destruction of public property was reduced, but reports of street robberies in parking lots (out of the cameras' sight) increased.

The example below shows how it can be explained why last December had more violent crimes than November. The tree starts out by looking at all violent crimes in November and December, where December is 60% of the total. The tree algorithm finds that street robberies show the most difference: 80% of all violent crime in December was a street robbery, which was 60% in November. Next, it turns out that robberies in Public transport have relatively increased, especially between 17:00 and 18:00. The latter pattern covers 40% of 75% of 80% is 24% of all violent crimes in December.

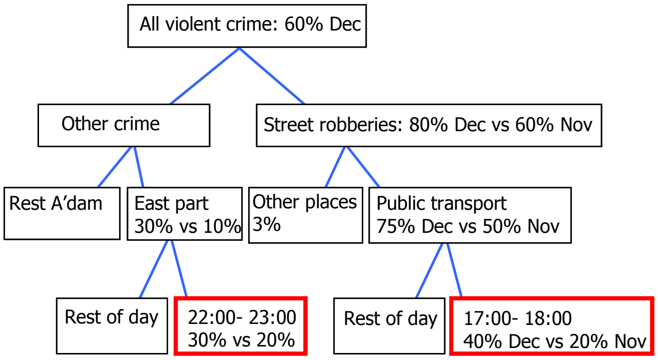

All violent crime: 60% Dec
Other crime
Street robberies: 80% Dec vs 60% Nov
Rest A'dam
East part 30% vs 10%
Other places 3%
Public transport 75% Dec vs 50% Nov
Rest of day
22:00- 23:00 30% vs 20%
Rest of day
17:00- 18:00 40% Dec vs 20% Nov

Figure 2: Decision tree explaining a trend

28.3.2 Find contextual trends

An alternative use of comparing periods is to find trends in context; not the trends in crime rates going up or down, but trends in the aspects of crime. For example: there is no strong increase of total number of burglaries but there is a sudden increase of flat screen theft from apartments, or a trend where a specific type of tool is used. These trends can be used to point out activity by a specific offender or offender group, or a more general phenomenon, such as an increase in motor cycle theft because spring has just begun.

28.3.3 Explain or describe behavior

The same approach can be used to explain behavior, by comparing the activities and/or the features of a person or group against a reference group. For example: compare violent young criminals in a specific neighborhood against all young criminals in that neighborhood. What makes these violent persons different? Are they from specific parts of town, are there patterns in their careers? What is their typical social background? Such patterns can be found with a decision tree analysis. By gaining insight in related aspects of such behavior, preventative measures can be designed. Another application is creating a description of the *signature* (typical way of working) of a person or group.

28.3.4 Explain spatial relations

Detecting relations between spatial aspects (e.g. type of neighborhood) and behavior (e.g. crime rates) is traditionally done by looking at areas: the crime rate for each area is used as a variable that needs to be explained by the aspects of the area (average income etc.). A new method of doing this was created, by using local crime density as the variable to be explained. In this way, the level of detail can be much higher, say on address level, without the requirement of having numerous incidents on that level. First, a density analysis is done (see paragraph 5.4), calculating the density of the crime for each address. Then, the techniques for explanation are applied. This allows for better and detailed explanations since crime rates and spatial aspects do not need to be averaged over large areas.

28.3.5 The use of text

Using textual information can be very helpful for gaining insight. For example: the system was asked to analyze the Ramadan period (Islamic month of fasting) and it found that this period typically has a strong increase of destruction of property. Before we had the chance to think about possible relations between fasting and this pattern, the system also showed that the words *firecracker* and *fireworks* occurred more often in incidents during Ramadan. Now, Ramadan takes place during the ninth month of the Islamic calendar, causing it to sometimes take place in December. This explains the pattern of fireworks and thus the destruction of property resulting from incidents during the days before New Year's Eve, which was confirmed by the resulting decision tree.

28.4 Kernel density estimation - hot spot maps

Knowing where crimes take place is crucial information for the police and is best shown on a map for optimal interpretation. The basic way of visualization is to plot single incidents as dots. If dots are too close to each other, they can be combined into larger dots. This method seems fairly obvious. However, such dot maps can be hard to interpret, especially where the concentration of dots is high. Hot spot maps provide a solution by interpolating incidents for each cell of a detailed grid on the map, resulting in a color-coded *hot spot* map. The data mining system uses kernel density estimation [11] for the interpolation on three detail levels, each for a different zoom range on the map. Thus, the hot spot map shows more detail when the user zooms in. An extension of this technique was implemented by allowing the map to incorporate parts of streets in addition to exact addresses, since a large percentage of incidents have been registered with just a street name.

28.5 Temporal hot spots

The data mining system allows temporal hot spots to be created for visualizing spatial trends in time. This is done by creating a kernel density grid for the recent period and one for the period before, which are then subtracted, resulting in a density map with positive (red) areas with increased crime rates and negative (blue) areas with decreased crime rates.

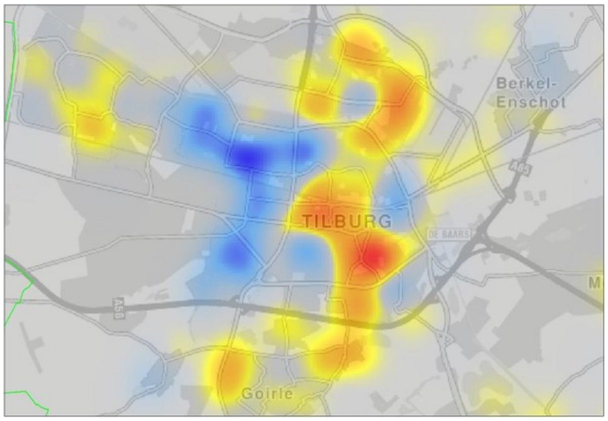


Figure 3: Temporal hot spot analysis of burglaries in Tilburg

28.6 Cluster series of crimes

A typical police analyst's task is to link crimes in order to either solve a case or to find interesting series that point to the activity of a single offender or a group. This linking process is supported by the data mining system by allowing crimes to be clustered on location, MO, time, day of the week, weather, and suspect descriptions.

Associative clustering (see paragraph 5.1) is used to create the result, showing clouds of incidents on a two-dimensional chart. The closer incidents are on the chart, the more similar they are.

This chart can be used to detect series and to match a specific unsolved case to similar cases simply by looking it up on the chart. Especially the *solved* similar cases are interesting. This associative approach is similar to the application of Self Organizing Maps (or Kohonen maps [14]) for clustering crimes as discussed in [2]. The associative approach however is easier to use, has less data requirements and faster execution times.

28.7 Cluster sub problems

Spatio-temporal clustering, as discussed in paragraph 5.1, detects clusters in space and time, typically for one specific type of crime. By adding contextual information, the data mining system is able to detect clusters of similar incidents that together define a *sub problem*. In this way it is possible to cluster a wider range of incidents (e.g. all incidents) to find out what structural problems exist in order to address them individually through prevention and/or repression. Finding the right preventative measures is supported by a deeper analysis of the cluster to find its probable causes. Repression is done by using the tactical information just as with the spatio-temporal clusters: location, time, day, and clues to look for.

An example of a situation where this approach was successful is the analysis of sub problems on Queen's day. Queen's day in the Netherlands is a national street festival with its own typical safety challenges. In order to prepare for this day, the Amsterdam police ran a sub problem clustering on Queen's day data from recent years. The results showed the typical Amsterdam everyday problems plus typical Queen's day problems, such as pickpocket crimes near the Rembrandt Square in the early evening, inside bars and restaurants, with mostly tourist victims. This pattern, along with many other sub problems, allowed the police to define very specific instructions for officers to address these problems on the upcoming Queen's day.

Another example is the application of sub problem clustering for a specific shopping area in the South East of Amsterdam. It took place in an interactive group session where experts and stakeholders were present: neighborhood police officer, neighborhood watch, shop owners and drug expert. The data mining system was operated by an analyst visualizing the problems, so questions could be addressed immediately.

Experiments with extensive socio-demographic data in problem clusters have demonstrated that causes of problems can be explained better by combining tactical data with household data. For example: clusters were found containing theft of flat screen televisions from low-income households, who have a tendency of showing off material possessions. This resulted in the theory that this group makes itself vulnerable by positioning the televisions so they can clearly be seen from the outside, making them more attractive to burglars. Such a theory also helps to determine other geographic areas that suffer from the same risks.

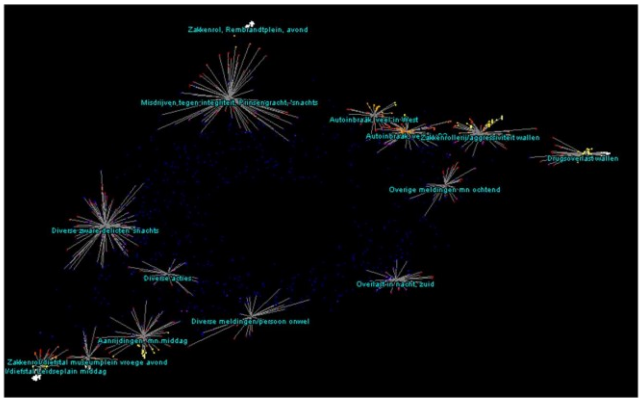

Zakkenrol, Rembrandtplein, avond
Misdrijven tegen integriteit, Prinsengracht, 's nachts
Autoinbraak veel in West
Autoinbraak
Zakkenrollerij/aggressiviteit wallen
Drugsoverlast wallen
Overige meldingen mn ochtend
Diverse zware delicten 's nachts
Diverse acties
Overlast in nacht, zuid
Diverse meldingen/persoon onwel
Aanrijdingen mn.middag
Zakkenrol/diefstal museumplein vroege avond
l/diefstal Leidseplain middag

Figure 4: Clustering of sub problems on Queen's day

28.8 Geographic profiling

Geographic profiling is a method that uses locations of crimes to find the most likely areas of the offender's home address [15]. This is done by applying theories about the activity radius of offenders combined with theories about offenders usually not operating in the vicinity of their own home address. By combining this activity theory for every crime location, the data mining system creates a probability map that can be used to narrow down a selection process while solving a case. The parameters are calculated automatically based on the data.

28.9 Link analysis

Connecting crimes and criminals is important when solving cases and when looking into the networks of criminals to see what groups there are, who knows who, and what their roles seem to be. The data mining system provides a link analysis tool to visualize the structure and content of a collection of interconnected entities and to calculate measurements based on theories from *Social Network Analysis*, tailored to the criminal domain [16]. These measurements help in finding roles (e.g. who maintains the most connections). The data mining system is able to generate a network and to visualize it using the same dimensional scaling algorithm as in clustering (See 5.1). This way, the suspects and known offenders in the network are organized so that their closeness on the screen resembles closeness in links. This saves analysts hours of combining data. Networks can be exported to Analysts's Notebook.

28.10 Find possibly connected suspects or incidents

One of the standard methods for solving a case is the search for similar cases in the past to see if they contain useful information. Especially when those similar cases have a known offender, it can be interesting to consider that offender as a possible suspect of the new case. Performing such a search is a difficult task because there is hardly ever an exact match. For example: suppose the case to be solved is a violent street robbery. The same offender may have been a suspect for a very similar incident a few weeks ago, but the similarity may be complex: the past crime has been entered as a pickpocket incident and not as a street robbery; the pickpocket location was near the robbery location, but not in the same street; witness descriptions state blond hair for the pickpocket, but dark blond for the robber. There may be many similarities making the pickpocket case interesting, but it is typically not found using conventional query methods.

The police data mining system features associative techniques that find similar cases based on a search case. This is used as part of the algorithms for predictive modeling and clustering, and it can be used as a technique for flexible search as well. Such a search takes a case as input and creates an output of similar cases, ordered by similarity. These similar cases can then be used to find the related suspects. If one of those suspects has more than one similar case, that suspect becomes more interesting. This entire process has been automated in the system, also allowing multiple cases (for example a series of crimes) to be used as the input for a search. For example: a series of 10 crimes is used in an associative search that finds 33 similar cases in the past of which 4 have one and the same suspect and these 4 cases are similar to 8 out of the 10 crimes in the search series.

There are many applications for associative searching, e.g. looking for suspect photographs based on a witness description, or looking for incidents that are most similar to the entire criminal history of an offender in order to find other crimes that may be committed by this offender.

29. DATA MINING SUCCESS STORIES

The following subsections present success stories of applying the data mining system in practice.

29.1 Robbery

In the beginning of 2009, the district of Tilburg in the Netherlands suffered from a serious wave of robberies at gas stations, restaurants etc. Analysts used the data mining system to visualize the locations of these robberies and applied associative spatial prediction to determine when and where police actions (e.g. roadblocks) would be optimal. They applied link analysis on past robberies to determine which suspects were important to keep an eye out for. In addition, the system produced top X lists of offenders, ordered by their past robbery activity. These selected people were brought in if they had any outstanding fines. Police officers visited the rest and their photographs were used in briefings. Within two weeks, the wave of robberies was stopped.

29.2 Car burglaries

In 2006, a trend analysis in Amsterdam indicated an increase of motor vehicle theft in District 2 for the month of May. A profile analysis of this trend showed that the increase could be explained by thefts from private garages. More officers than normal were deployed and their actions were supported by spatio-temporal cluster analysis plus the top 10 list of car burglars. The result was that a repeat offender was caught in the act within one hour after the data mining information was provided and crime went down 90% in the first week.

29.3 Burglary

The following scenario illustrates how data mining techniques are combined in solving a problem. The analysis actually was performed by police to better understand the underlying problems.

1. The standard report indicates a sudden increase of burglaries in October, which is atypical for the time of year.
2. A time series chart is created to visualize the trend and validate the indication.
3. By clicking on October in the chart, an explanation of the trend is requested in the form of a decision tree to show what combinations of factors have changed (see 5.3). The tree shows that there are more break-ins at the back of houses than before, especially in the early evening. It also shows that the burglaries occurred more often after sunset. This leads to the theory that it is getting dark earlier and therefore there is more opportunity in the early evening especially in back yards where there is hardly any street lighting. This happens especially in a specific neighborhood. The pattern can be used purely to explain the trend, but it can also be used to start a program in that neighborhood to encourage placing motion-triggered lights in back yards.
4. The decision tree also shows a strong geographic difference. This leads to a temporal hot spot analysis (see 5.5) that shows a number of hot spots where crime increased.
5. The largest hot spot is chosen by zooming in on the map and asking the system to create an explanation of this trend. This shows a very specific time of day, specific streets and some witness descriptions; useful information to support police officers on patrol in that area.
6. To determine who the police officers should look for, the hot spot selection is used to perform *associative searching* (see 5.10). This matches the incidents in the hot spot with the entire known crime history. The best matches and most recent incidents are selected to see who the connected suspects were, in order to provide their photographs to the police officers.
7. Next, link analysis is used to find the people that are directly or indirectly connected to the selected probable suspects, to also add their photographs.

30. DISCUSSION

In this section we will discuss added value for police work, the challenges involved and future work.

30.1 Added value

Section two discusses the various reasons why data mining is important for police analysis. The organizational impact of the discussed data mining system is that a large number of users are now able to gain more insight and predict criminal behavior. These users are responsible for providing information to the organization, so in other words, data mining has made the organization more intelligent. The growing number of users of the data mining system is a clear sign of the acceptance of data mining as an important tool for the police. But how can the added value be measured?

Five methods to assess the added value of data mining can be distinguished: model accuracy, experiments in practice, mimic practice, analyst efficiency and qualitative comparison.

30.1.1 Measure model accuracy

Measuring model accuracy can be done by basing the model on one part of the data and testing it on the rest. The resulting accuracy can then be translated to the predicted increase in police effectiveness using a simple model of police practice. For example: we calculated that the top 5% hotspot areas from associative spatial prediction eventually contain 50% of all crimes in the predicted period, whereas the top 5% areas from traditional tactical information contain 25%. We assume the police only have the capacity to be present in those 5% areas and that police presence reduces the risk of an incident taking place by 50%. This way we can calculate the projected reduction of crime. For the spatial prediction, 50% of all crime is reduced by 50%, so total crime is reduced by 25%, were the traditional approach reduces 50% of 25%, which is a reduction of the total crime rate by 12.5%. This method of projecting effectiveness typically suffers from many assumptions but it can be useful to illustrate the impact of these predictive models.

30.1.2 Run experiments in police practice

The proof of the pudding is in the eating and therefore the best way to measure effectiveness is to experiment in practice. However, because public safety is generally considered not something to experiment with, pure tests are very difficult to arrange. Furthermore, crime and environment are so dynamic that an exclusive explanation of success cannot be based on one experiment alone. Crime rates may change during an experiment because of many reasons other than data mining being used or not. Therefore, multiple experiments are required for a good measurement. Another challenge in measuring added value is how to measure safety, security and fear. Crime rates are useful but cover just a part of the whole picture.

Recently, the police force of Midden en West Brabant addressed a region-wide wave of crime. One district used data mining for providing information on these crimes, the others did not. After two months, the data mining district reduced crime by more than 15% whereas the other districts showed constant crime levels. There are more encouraging results like this, but it is no solid proof. Because of the difficulties with measuring this type of added value, we also use the other methods in this section to convince police organizations.

30.1.3 Mimic police practice in laboratory settings

An alternative to experimenting in practice is to mimic practice. We performed a field test with the Rotterdam and Haarlem Police forces in which volunteers were asked to sit in a waiting room, where a laptop was 'stolen' by an actor. The volunteers witnessed this act and were asked to provide descriptions that were then used to search for photos to show for identification. This was done using two systems: the standard police search system and an associative search tool, similar to the one used in the current data mining system. The result was that 50% more witnesses recognized photos of criminals selected by the associative system than from the standard police search system. The disadvantage of such tests is that they are expensive.

30.1.4 Measure analysts' efficiency

An alternative to measuring increase of effectiveness is to measure the gain in *efficiency* as a result of data mining. Measuring analysts' efficiency is more simple than measuring police effectiveness and the work situation is easier to control than criminals and their environment. Research in the region of Midden en West Brabant showed that analysts generally reduced their analysis time by a factor 20 when using the data mining system. The added value of that comes down to cost reduction or increased effectiveness because analysts can get more work done. In all police organizations we have seen, analysts always have much more to do than they have time for, which means that when the data mining system is used, the organization is better informed.

30.1.5 Qualitative comparison of results

In our experience, qualitative assessment of data mining advantages is often convincing to decision makers. Once analysts and their chiefs experience what data mining can offer them, they typically do not need quantitative proof that it will make the organization more effective. An analogy is the implementation of geographic information systems in police forces. Once officers start working with maps, the added value seems clear and there is no need for conducting experiments in which this new situation is compared with situations without maps. Research by Abrio [internal report] and Midden en West Brabant has shown that data mining analysis is more productive and the results better satisfy the information needs because of their level of detail and explanation.

30.1.6 Cost and return on investment

The costs for data mining can be divided in implementation and running costs. Implemenation requires the availability of a datawarehouse in which data sources are extracted, combined, cleaned and extended by calculating variables (e.g. domestic violence yes/no). Such a datawarehouse provides a *single point of truth* for the organization, and its merits go far beyond data mining. Therefore, police forces in the Netherlands see the availability of a datawarehouse as a general *must* and there are many datawarehouse initiatives. We have developed a datawarehouse based on the dominant standard police systems in the Netherlands, thereby reducing datawarehouse implementation costs. Remaining costs are for database server hardware and software, installation and configuration.

Other data mining implementation costs are: application server hardware, software license and training. Running costs are software license, retraining, functional management and operating cost of batch modules that update the datawarehouse and create standard reports.

We have reduced these costs greatly by making the data mining system easy to learn and easy to use. Training is done in two to three days and no expert users need to be hired.

It is possible to argue a positive return on investment, based on only the benefit of analyst efficiency (7.1.4). The efficiency gain is a factor 20, so theoretically the capacity in analyst personnel can be reduced strongly while maintaining the same level of information delivery. This would create a cost reduction that would more than make up for the costs of implementing and using data mining. However, police organizations do not intend to reduce analyst capacity because the information need strongly exceeds the information production. In other words: in a simplified model, using data mining can be compared to hiring more analysts for a fraction of the costs. This makes the business case for data mining, based on analyst efficiency alone, leaving all other benefits out of the equation.

30.2 Challenges

The implementation and use of the data mining system pose a number of challenges.

30.2.1 User skill management

Even though much attention has been given to making the system user friendly, it is important to maintain the user's level of skill. Applying data mining is an inherently complex task, even if all choices regarding parameters and techniques are done automatically. Maintaining the skill level is managed by proactive user assistance, sharing best practices, stimulation of user communities and regular training sessions.

30.2.2 Data quality

The higher the data quality, the better the results from data mining are. Managing the quality of police data is a notoriously difficult problem because of the diversity of people who enter the data, the varying circumstances in which this is done and because of difficulties in applying definitions when registering facts. This has not stopped data mining applications to be successful but it is obvious that results will be better as data quality increases.

It is interesting to note that data quality can be improved by applying data mining because of two effects: 1) all data is used (demonstrating to the organization how important it is to register all data correctly), and 2) some patterns found by data mining illustrate data problems. For example: an increase in specific MO may not be caused by an increase of that type of crime but rather by new instructions to the people that entered the data.

Two principles guide the handling of data quality problems by the data mining system:

1. Deal with potential data problems as early as possible while creating the datawarehouse by combining data sources and by using software to correct data errors, for example by simply removing a piece of information when rules detect it is erroneous.
2. Document data problems and make this documentation easily accessible from within the data mining system.

30.2.3 Managing expectations

It is our experience that in some cases the expectations of the impact of data mining are too high. This requires clear communication of what can be expected. Data mining will not solve all data problems. Data mining will not make analysts dispensable. Data mining will not be able to interpret textual information perfectly. Data mining will not easily succeed in bringing together all the available data because of privacy laws and the amount of effort it takes to create reusable extractions.

30.3 Future work

Although the data mining system has been operational for years, new ideas still come up to improve the system and to apply it in new ways.

30.3.1 More prediction models

Currently, prediction models are used to predict where crimes are likely to take place and to explain trends or behavior. Many other applications of predictive models are being studied:

1. Criminal career: who have the highest risk of becoming a repeat offender?
2. Criminal profiling: predict probable motive, offender age, gender, etc., based on a crime or series of crimes.
3. Weapon possession: what is the risk of a person carrying a weapon, based on personal profile and history?
4. Domestic violence: what is the chance of a reported domestic violence situation getting out of hand, to support the decision to pay extra attention to the case?
5. Predict crime rates based on infrastructure, buildings and socio-demographics in a new neighborhood, allowing what-if experiments.
6. What are the risks for a specific type of crime for an area based on the properties of that area and the situation (time, day, weather)? This alternative approach to associative spatial prediction (see 5.2) is useful when there are too few examples to make the spatial prediction work.

30.3.2 Dashboard

The standard reports generated by the data mining system are currently distributed in document form, created by transforming XML into HTML. There are plans to offer this information in the form of a clickable dashboard that starts with an overview of crime rates and allows clicking on rates to zoom in on areas and types of crime, showing more details such as maps and trends. In this way, the wealth of available report information could be navigated more effectively.

30.3.3 Text

Text is an important source of information within police forces because, for example, statements have more detail than the information recorded in the structured data. The implemented data mining system allows textual information to be included and used for selection and for pattern analysis. However, a system to regularly collect textual information from the source systems still needs to be built. A different way of working with text is to extract entities (license plates, phone numbers, addresses and names) to use them as a method for linking people and cases. When such entity extraction systems are implemented within police forces, we will add the entity information to the data mining system.

30.4 To conclude

In this paper we showed that data mining is important for police and that it has been made a continuous activity, performed by police employees without extensive expertise in databases or data mining. This has been realized by developing a data mining system in co-operation with police, bringing together interaction, visualization, tool integration, automated algorithms, a large and diverse datawarehouse and ease of use. Especially the latter is a key success element because a combination of factors make data mining for police relatively difficult: analytical diversity, strong need for domain expertise, data quality problems and difficult data extraction. The system has proven to provide more depth in analysis and in some cases an efficiency gain of a factor 20. Furthermore, crime rates dropped in field tests. After eight years of use, the system is still being extended constantly based on new ideas, which shows great promise for the future of crime fighting.

31. REFERENCES

[53] Ratcliffe, J.H. 2003. Intelligence-led Policing, Australian Institute of Criminology, Canberra, Australia, 248.

[54] Adderley, R. & P.B. Musgrove, P.B. 2001. Data Mining Case Study: Modeling the Behavior of Offenders Who Commit Serious Sexual Assaults, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.

[55] McCue, C. 2007. Data mining and predictive analysis: Intelligence gathering and crime analysis, Butterworth-Heinemann.

[56] Mena, J. 2003. Investigative Data Mining for Security and Criminal Detection, Elsevier Science (USA).

[57] Brown D.E. 1998. The Regional Crime Analysis Program (RECAP): A framework for mining data to catch criminals, University of Virginia.

[58] Bruin, J.S. de, Cocx, T.K.., Kosters, W.A., Laros, J.F.J., Kok, J.N. 2006. Data Mining Approaches to Criminal Career Analysis, Proceedings of ICDM '06.

[59] Uyl, M.J. den. 1986. Representing Magnitude by Memory Resonance, Proceedings of the 6th Annual Conference of the Cognitive Science Society, pp. 63-71.

[60] Borg, I. & Groenen, P. 2005. Modern Multidimensional Scaling: theory and applications (2nd ed.), SpringerlVerlag New York.

[61] Farrell, G. 2005. Progress and prospects in the prevention of repeat victimization.

[62] Eck, J.E., Chainey, S., Cameron, J.G., Letiner, M. & Wilson, R.E. 2005. Mapping crime: Understanding hot spots, Technical report 209393, National Institute of Justice.

[63] Chainey, S. & Ratcliffe, J. 2005. GIS and Crime Mapping, Mastering GIS: Technology, applications and management, West Sussex, England: Wiley.

[64] Clarke, R. V. & M. Felson, Eds. 2004. Routine Activity and Rational Choice (Advances in Criminological Theory), Transaction Publishers, New Brunswick (U.S.A).

[65] Bowers, K.J., Johnson, S.D. & Pease, K. 2004. Prospective hot-spotting: The future of crime mapping?, Br. J. Criminol 44(5), pp. 641-658.

[66] Kohonen, T. 1984. Self-organisation and associative memory, Springer series in information sciences, Vol8. Springer Verlag, New York, USA.

[67] Rossmo,D.K. 2000. Geographic profiling, Boca Raton, Fla, CRC Press.

[68] Freeman, L. 2004. The development of social network analysis, Vancouver, CA: Empirical Press.

[69] Kohonen, T. 2001. Self-Organizing maps, Springer series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York

Identification of Independent Individualistic Predictors of Mail Order Pharmacy Prescription Utilization of Healthcare Claims

Aishwarya Meenakshi Sundar
3035 Eagandale Place, Apt #119,
Eagan, Minnesota - 55121
612-308-9608, 001
sunda031@gmail.com

Professor. Jaideep Srivastava
Department of Computer Science and Engineering
University of Minnesota, Twin Cities
(612) 625-4012, 001
srivasta@cs.umn.edu

ABSTRACT

Health care research has lead to the development and availability of safe and effective drugs to treat debilitating chronic diseases. The impact of this advancement has been the ever increasing costs to obtain these medications. Prime Therapeutics, a pharmacy benefit manager (PBM), strives to ameliorate some of the financial burden through the delivery of medications via mail order pharmacy. Mail order pharmacy provides cost savings to both the insurer and their members; in addition to improving overall customer satisfaction through a convenient home delivery system. The objective of this project was to apply various Data Mining techniques, like Classification, Clustering and Association analysis, to member profile and health care claims data, to identify the links between member characteristics and their mail order behavior. Identifying the individual characteristics influencing mail order acceptance behavior will help Prime Therapeutics in creating target marketing programs to improve mail order utilization.

General Terms

Health Care Related Terms [8][9]

Pharmacy/Prescription claims (rxclms)

A prescription filled by the member and paid through the PBM's claim processing system. This gives information about the prescription purchase, including the dosage and strength of the drug and other purchase related details. Unique characteristics of the pharmacy claim include identification of the generic or brand status of the medication, formulary or non-formulary status, maintenance drug status, specialty drug status and lastly whether the prescription was filled as 90 day supply through the retail pharmacy also called extended day supply.

Generic Drug

A generic drug is one, which is produced and distributed without patent protection [10]. Thus any company can manufacture it. The generic manufacturer must submit for FDA testing and approval prior to selling a generic drug.

Formulary Drug

A drug list, or formulary, is a list of drugs that the insurer would like doctors to use when writing prescriptions for plan members. The drugs on the formulary are considered the best choices based upon safety, effectiveness and reasonable cost. [10]

Non-Formulary Drug

Doctors may prescribe a prescription drug that is not on your health plans formulary; in that case the member may have to pay the full price for the medication when they pick it up at the pharmacy.

Maintenance Drug

A drug prescribed for chronic conditions like diabetes, arthritis, high blood pressure, or heart conditions.

Specialty Drug

Specialty drugs are used to treat serious or chronic medical conditions such as multiple sclerosis, hemophilia, hepatitis and rheumatoid arthritis. They are typically injectables and can be self-administered by a patient or family member.

Extended Supply at Retail

The prescription was filled at a retail network pharmacy and it was a 90-day supply of drug.

DMCS'09, June 28, Paris, France.
Copyright 2009 ACM 978-1-60558-674-8…$5.00.

Medical claims (mdclms)

A medical claim is for a service provided by a health care professional, usually in a medical clinic or hospital. Medical claims provide information about the member's medical diagnoses and health care utilization.

Pharmacy Benefit Characteristics

Deductible Flag

Individual's insurance has a deductible requirement or not. Deductible is defined as an amount of money an individual has to pay themselves before their health insurance benefits become active. For example, an individual may have a health insurance benefit of a fixed $20 prescription copayment for any prescription but this only becomes active after they have spent $100 on prescription drugs.

Benefit Max Flag

Individual's insurance has a benefit maximum or not. Benefit Maximum is defined as the maximum amount the health insurer will pay for medical costs after which the individual is liable for the entire payment of all medical costs above the maximum. For example an individual may have a prescription benefit maximum of $100,000 and once the prescription costs exceed $100,000 then the individual must pay the entire amount for prescriptions.

Out of Pocket Amount Flag

Individual's insurance has an out of pocket amount or not. This is defined as the amount an individual pays out of pocket after which the health insurer generally pays 100% of the costs up to a benefit maximum (if a benefit maximum exists). For example, an individual may have a $1000 out of pocket amount and once their out of pocket prescription co-payments accumulate to $1000 then they will no longer have to pay anything out of pocket for their prescription medications.

Tools Employed

Weka provides a suite of sampling and feature selection functions, used to derive representative samples and test the suitability of various data mining algorithms on these samples, which were widely employed in the preprocessing steps of this project.

Microsoft SQL Server 2005: Analysis and Data Mining Services [8]

All of the algorithms employed and the predictive models built in this project are using the Microsoft SQL Server 2005 Analysis Services (SSAS). The Analysis Services tools provide the capability to design, create, and manage data mining models and to provide client access to data mining data. SSAS provides storage capabilities to support very large datasets, a suite of data transformation functionalities, discretization and binarization techniques and a collection of popular and powerful data mining algorithms like decision trees, naïve bayes, neural networks, clustering techniques and association analysis.

1. INTRODUCTION

In the past decades the domain of health care has seen great amount of research and advancement, resulting in the introduction of new drugs to treat debilitating chronic diseases. With the advent of all these new medications, it has become increasingly expensive to obtain these drugs, especially for long-term usage. This escalation in cost affects both the consumers and the health care insurers. When consumers, with chronic health disorders, purchase long term medication or specialty drugs from retail outlets they potentially have higher out of pocket expenses than purchasing through mail order pharmacy. Managed care organizations (MCOs) like health plans, health maintenance organizations (HMOs), pharmacy benefit managers (PBMs) etc, take many initiatives to lower drug expenditures. One of the ways to tackle this cost hike is for the MCOs to sign mail order contracts directly with the drug manufacturers, which results in much lower drug purchasing costs. Consumers who choose to fill their prescriptions via mail can save money and receive their medications at their doorstep. Mail order utilization is a powerful means for capturing cost savings for both the insurer and their members, while potentially improving the members' overall satisfaction with their pharmacy benefit.

Prime Therapeutics is a thought leader in pharmacy benefit management (PBM) strategies, which provides health care management related products and services to clients, and their employers, enrolled with Blue Cross Blue Shield (BCBS) Plans. The members of Prime Therapeutics include BCBS of Florida, Illinois, Kansas, Minnesota, Nebraska, New Mexico, North Dakota, Oklahoma, Texas and Wyoming. Their clients include the people who have enrolled in health care policies with either of the above ten providers. At Prime Therapeutics, they provide collaborative and workable strategies that enable their members and clients to effectively and efficiently manage pharmacy benefits. Their services cover up to 14.6 million lives; 36 million weighted lives across 16 different Blue Cross and Blue Shield Plans. [10]

PrimeMail is a Mail order pharmacy utilization program by Prime Therapeutics, which provides safe, convenient, cost-effective prescription delivery services. They offer a fully integrated mail service solution backed by an expert pharmacy staff and account support. Completely understanding the escalating costs of medication and the cost incurred by its members in providing health care benefits to its clients, Prime Therapeutics encourages its clients to utilize mail order pharmacy service. [10]

The current mail order utilization rate of Prime Therapeutics is approximately 5%. Comparable pharmacy benefits managers (Caremark, Express Scripts, Medco) have mail order utilization rates greater than 10%. Having provided pharmacy benefit management services for over ten years, Prime has a repository of member demographic and health profile date, their enrollment policy details and pharmacy claims (both via retail and mail order). These comprehensive datasets could provide invaluable insights into the characteristics of individuals who have utilized mail order pharmacy in the past. Identifying individual characteristic links could help Prime Therapeutics predict mail order behavior of new members and channel their marketing funds to improve their mail order utilization rate.

Situations that involve sorting through large amounts of data and extracting useful information have always been a challenge when using traditional data analysis tools and techniques. The domain of Data Mining is known to address these issues by applying computer-based methodologies, including new techniques for knowledge discovery from data.

The capabilities of Data Mining and Prime Therapeutics' requirements provided a new opportunity to map data mining techniques on to the domain of health care, to identify strong predictors of mail order pharmacy prescription utilization.

Objective of this study: Perform data mining on member profiles and their pharmacy claims, to identify individual characteristics associated with mail order pharmacy prescription utilization and persistency of mail order utilization.

Prime Therapeutics' Goal: To take individual characteristic links identified through data mining and create target marketing programs and messages to improve the mail order utilization rate.

2. DATASETS

The data available consists of

1. Pharmacy prescription claims for approximately

(i) 9 million commercially insured individuals

(ii) 850,000 individuals enrolled in the federal government Medicare Part D program.

2. Medical claims available for only Blue Cross Blue Shield clients (1.8 million individuals)

3. Integrated medical and pharmacy prescription claims for 1.8 million individuals.

The above data was compiled into three different data sets for analysis. All these three datasets are stripped of any individual identifiers. The members are referred using a dummy member identifier.

2.1 Enrollment Information

This data set provides information about the member eligibility for health care benefits on every 15th and 30th day of each month, ranging from January 2006 to May 2007, thus setting the time frame of the analysis as eighteen months. This information helps us identify the consistent members i.e., who are insured for all eighteen months, and concentrate our efforts on these members while building our predictive models.

2.2 Member Profile Information

This data set provides information about member demographics and health conditions for each of the members listed in the enrollment data set above.

1. Dummy Memid – Unique de-identified member identification that associates each member with the enrollment information in the previous dataset. Thus this dataset presents an eighteen month enrollment summary for about 2.1 million

2. Age

3. Year of Birth

4. Sex

5. Zipcode of the member location

The various health indicators include,

(i) Nine health conditions indicated by prescription claims (rxclms) for about 1.4 million users. All these features are binary attributes i.e., a 0 / 1 indicating the absence / presence of the various health conditions.

Table 1: Health Indicators from Prescription Claims (rxclms)

Afibrx	Atrial fibrillation prescription
Dmrx	Diabetes prescription
Asthrx	Asthma prescription
Lipidrx	Hyperlipidemia prescription
Chfrx	Congestive heart failure prescription
Osteorx	Osteoporosis prescription
Htnrx	Hypertension prescription
Deprx	Depression prescription
Cadrx	Coronary artery disease prescription

(ii) Nine health conditions indicated by medical claims (mdclms) for about 1.3 million users. These are also binary attributes and correspond to the conditions in the prescription claims eg: Afibdx, Dmdx, Asthdx etc.

(iii) Twelve severe diseases and their corresponding Charlson weights. Charlson weighing score of a disease indicates its severity and the score range varies from disease to disease. Certain disorders like Depression have just a 0/1 Charlson score, while for some diseases like AIDS the score could range between 0 and 6. The higher the score, the greater is severity of the illness. It is a weighted index that takes into account the number and the seriousness of comorbid disease, which helps in predicting the risk of death from comorbid disease. [1]

(iv) Overall Charlson Score is a number between 0 and 22, indicating the overall health condition of a person. It is a cumulative indicator of the severity of a person's health disorders. Closer the score to 22, greater is the probability for the person to die.

Table 2: Severe illness with Charlson weights

Pvdch	Peripheral vascular disease Charlson weighting score = 1
Cbdch	Cerebrovascular disease Charlson weighting score = 1
Dementiach	Dementia Charlson weighting score = 1
Rheumch	Rheumatologic Charlson weighting score = 1
Pudch	Peptic ulcer disease Charlson weighting score = 1
Livermildch	Mild liver disease Charlson weighting score = 1
Hemiparach	Hemiparalysis Charlson weighting score = 2
Renaldxch	Renal disease Charlson weighting score = 2
Cancerch	Cancer Charlson weighting score = 2
Livermodch	Moderate liver disease Charlson weighting score = 3
Cancermetch	Cancer with metastases Charlson weighting score = 6

Thus the entire member profile data available can be summarized as follows

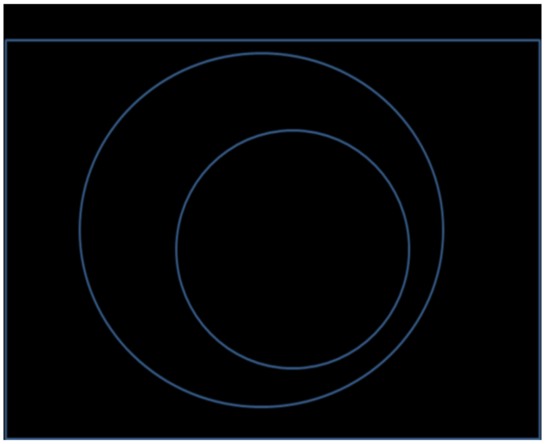


Figure 1: Summary of Data Available

2.3 Prescription Claim Information

This dataset provides detailed information about the prescription filled by each member in the two datasets. It consists of the following details for approximately 21 million prescriptions filled.

1. Policy Features (Refer Page 1)

- Deductible Flag
- Benefit Max Flag
- Out of Pocket Amount Flag

2. Purchase Details

Table 3: Features indicating Purchase Details

dw_unique _key	Uniquely identifies each prescription.
Affliation Id	Identifies whether the prescription was filled at a chain store (like Walmart / Walgreens) or at a independent pharmacy
Pharmacy Name	Name of pharmacy where prescription was filled
Pharmacy Zipcode	5 digit zipcode of the pharmacy
Date of Service	

3. Drug Details (Refer Page 9)

Table 4: Features indicating Drug Details

GPINumber	Unique fourteen digit drug identification number
NameNameExt	Actual drug name
DrugGeneric Name	Drug name inclusive of its strength
Quantity dispensed	
DaysSupply	
Generic	0 /1 indicating whether the drug purchased is a generic
Formulary / Non Formulary	F/ NF indicating whether the drug is formulary or not
Maintenance	0 /1 indicating whether the drug is a maintenance drug
Speciality	0 /1 indicating whether the drug is a specialty drug

4. Financial Features indicating the cost of a particular purchase

Table 5: Financial Features

Totpaid	Total cost of prescription
Membpaid	Amount paid by member
Planpaid	Amount paid by plan

5. Order details which encompasses the following

Table 6: Order Features

ExtendedSupplyatRetail	1/0 whether the prescription is a long term purchase (90 day supply) made at a retail store
Mailorder	1/0 indicating whether the claim was filled through a mail order pharmacy or not

2.4 Acxiom Data

Acxiom is a customer and data information management company which, among its other services, collects and provides detailed information about people such as their ethnicity, socio economic status, and education. For our analysis, we obtained six such Acxiom features, which we believed would provide more information about the customer's buying behavior. These features were integrated into the models to enhance the predictive accuracy. All these features are stripped off any identifiers and are matched with the previous three data sets using the dummy memid.

The following Acxiom features were available for 730,205 members, among with 28,224 of these members have utilized mail order pharmacy in the past.

Table 7: Acxiom Features

Education	Individuals education status 1 = Completed High School 2 = Completed College 3 = Completed Graduate School 4=Attending Vocational / Technical Default is blank
PC Internet	Indicates whether the member has a PC and Internet 1 = True

Marital Status	M = Married, S = Single, A = Inferred Married,
	B = Inferred Single. Default is blank

No of Children	0 to 7. 8 = 8 or greater
Length at residence	Number of years spent at the current residence 0 to 14 years 15 = 14 years or greater
Mailorder Buyer	Indicated whether the individual had made any purchase through mail (not just medication) B = Mailorder Buyer Default is blank

3. DATA PREPROCESSING

3.1 Data Filtering

All the datasets provided were very large and required some amount of filtering to

(A) Narrow down the dataset to obtain a subset of records which are more important for the analysis

(B)Eliminate records that might have a confounding effect on the target of the analysis

(C)Eliminate records that have missing or inconsistent values i.e., data cleaning

Some of the filtering employed in this project is:

1. Eliminate members who are not enrolled for the entire 18 months of our analysis i.e., retain only consistently enrolled members. The rationale behind this was to study the characteristics of continuously enrolled members first and build models that are representative of the persistent characteristics. These members are the most important subjects of our analysis. (This is in accordance to point (A) above)

2. In the third data set, retain purchase information of only Maintenance drugs i.e., long term usage drugs. This is, again, in accordance to point (A), because members under maintenance medications have greater need for cost savings and hence greater probability of mail order utilization.

3. Eliminate members who work for companies that force their employees to utilize mail order pharmacy. Such member characteristics are not indicative of voluntary mail order behavior. Such records if not purged will give rise to a 'false positive' conclusion that the dependent variable is in a causal relationship with the features of these members. Thus, confounding is a major threat to the validity of inferences made. (Point (B))

4. Very few members had a negative age of -1, indicating that the information is missing. Such records were purged from the analysis. (Point (C))

3.2 Working Dataset

Thus the final data sets available for analysis are as follows:

Member profile information	: 757,219 members

Members utilizing mail order in past : 28,685 members

Prescription Claim information : 6,220,276 records

With 28,685 (3.8%) of 757,219 members utilizing mail order pharmacy the data shows that a minority problem exists. Therefore, class balancing and re-sampling techniques were employed in the data preprocessing stage prior to actual analysis.

3.3 Sampling [3][5]

Sampling is a commonly used approach for selecting a highly representative subset of the objects to be analyzed. The various motivations for sampling include:

1. Reduction of data load on the algorithm
2. Faster and better performance
3. Choice of more complex and powerful algorithms that require lesser load eg: SVM and Neural Networks

Candidate samples are chosen using the SQL Server Integration Services' Percentage Sampling. The Percentage Sampling transformation creates a sample data set by selecting a specified percentage of the input rows. The sample data set is a random selection of rows from the transformation input, to make the resultant sample representative of the input. For our study, multiple candidate sets were chosen by repeatedly selecting twenty 2% and 5% samples from the dataset without replacement.

3.4 Re-sampling to balance minority class

One of the key requirements of sampling is to choose a representative sample. A sample is said to be representative if it has approximately the same property of interest as the original data set. In our scenario the aspect of interest is the characteristics of members who have utilized mail order in the past. As we had noted that it is a minority class problem, in order to preserve the characteristics of the minority positive examples we re-sampled the candidate sets chosen above to balance the minority samples. This eliminates the inherent skewness in the data and enables us to obtain better results. The re-sampling technique employed here, to achieve a uniform class distribution, was the 'Supervised Re-sampling' in Weka.

3.5 Feature Creation [5]

It is possible to create, from original attributes, a new set of attributes that captures the important information in a data set much more effectively. Some of the techniques employed here to construct new features were:

3.5.1 Aggregation

This is a technique used to reduce the granularity of data and analyze it at a higher level. Sometimes such analysis can act as a change of scope or scale that will tend to produce much more meaningful results than highly granular data. Some of the data aggregations performed were:

(i) Translating the member zipcode to indicate whether the member belongs to an Urban, Town or Rural locality. The rationale behind this translation was that the purchase behaviors of people vary broadly with the size of their locality and the presence of amenities in their area e.g.: Chain stores, good pharmacies etc.The mapping from zipcode values to urban, town and rural was made using the US Census Board Population Threshold values

Population >= 10000 => Urban
1000 <= Population < 10000 => Town
Population < 1000 => Rural

For each zipcode, the Census 2000 and 2005 population data of the zipcode was used to classify the area. E.g. 2005 census population of 55414 was 24,453 people. Since its population > 10,000, it is classified as an Urban area. Thus the feature constructed is,

Member location : {Urban, Town, Rural }

(ii) Translating the 14 digit GPI number to a 2 digit GPI drug family. This would help us identify the drug families having a positive affinity towards mail order pharmacy. There were totally 39 different drug families that were identified. eg: Anti-Asthmatic and Bronchodilators Agents, Antidiabetics, Cardiovascular Agents etc

(iii) Compute Annualized cost (Total cost, plan paid cost and member paid costs averaged over 18 months) paid by each member for medications over the duration of eighteen months. These new features enable us to identify various cost thresholds above which members tend to take up to mail order pharmacy. Hence a new table Financial_Detail was created to capture the annualized medical expense of each member. The attributes of the table are:

Table 8: Annualized Cost Features

Dummy Memid	Uniquely indentifies each member
Cummulative Totpaid	Annual cost of medication of the member
Cummulative Planpaid	Annual cost paid by the member's policy
Cummulative Membpaid	Annual cost paid by the member

3.5.2 Discretization

The 'Charlson score' attribute that indicates the health condition of a member has a large range of values between 0 and 22. Moreover, records with higher values of Charlson score i.e., above five are very sparse and infrequent. Hence it was necessary to discretize the score into fewer categories. Due to the skewed nature of the attribute, we could not apply Equal Width or Equal Frequency techniques to discretize. The discretization technique employed was the Clustering discretization method within Weka that splits the data into different clusters until the entropy is minimized. This technique provided better accuracy than other techniques since it takes the nature of the dataset into account while clustering the records.

Thus the Charlson score is converted into three binary features as follows

Table 9: Charlson Score Binary Features

Sc02	0/1 indicating if the Charlson score of the member is between 0 to 2
Sc37	0/1 indicating if the Charlson score of the member is between 3 to 7
Sc8	0/1 indicating if the Charlson score of the members 8 or greater

3.5.3 Normalization

Generally attributes values are normalized to make the entire set of values have a particular property. In our pharmacy claim dataset, the cost of each purchase i.e., totpaid, planpaid and membpaid, had to be divided by the days of supply to avoid having a long term purchase (which will cost more) dominate the results. These normalized costs give a rough estimate of the amount spent on medication per day. Hence for each purchase we computed the normalized costs.

Table 10: Normalized costs

Normalized_Totpaid	Totpaid / days of supply. Total medication cost per day
Normalized_Planpaid	Planpaid / days of supply. Amount paid by plan per day
Normalized_Membpaid	Membpaid / days of supply. Amount paid by member per day

3.6 Dimensionality Reduction / Feature Selection

The given data sets have a large number of features. Dimensionality reduction techniques serve to identify and eliminate redundant, irrelevant and highly correlated features that might affect the model building [5]. The next step was to apply dimensionality reduction to select features for analysis.

Principle Component Analysis is a linear algebra technique for continuous attributes that finds principle components (attributes) that are linear combination of original attributes and capture the maximum amount of variation in the data [5][11].

Information Gain Attribute Evaluation – This technique evaluates the worth of an attribute by measuring the information gain with respect to the class label. It ranks the attributes according to their ability to discriminate between the different class labels i.e., in our case between mail order =1 and mail order = 0.

X2 test [7] - Evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. In statistics, the Chi-squared test is applied to test the independence of two events. In feature selection, the two events are occurrence of a feature and occurrence of the class. We then rank terms with respect to the following quantity:

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

N – Observed frequency of the various combinations of occurrence of the feature and the class

E – Expected frequency of the various combinations of occurrence of the feature and the class

A high value of Chi-Square indicates the greater dependence between the feature and the class. The occurrence of the feature makes the occurrence of the class more likely (or less likely), so it should be helpful as a feature. This is the rationale for Chi squared test. By applying all of the above techniques and by ranking the attributes with respect to the class label of mail order = 1, the features selected for analyses were the following:

Member Profile Features:

- Demographic features {Memid, Age, Sex, Member location, Sc02, Sc37, Sc8, Mailorder}
- Health features from prescription claims {Afibrx, Dmrx, Asthrx, Htnrx, Deprx, Lipidrx, Chfrx, Osteorx, Cadrx}
- Health Features from medical claims {Dmdx, Htndx, Depdx, Osteodx, Lipiddx}

Prescription Claim Features:

- Policy Features {Deductible Flag, Benefit Max Flag, Out of Pocket Amount Flag}

- Drug Features
  {GPI2, Generic, Formulary, Specialty, Maintenance drug}

- Financial Features
  {Normalized_Totpaid, Normalized_Planpaid, Normalized_Membpaid}

- Order Details
  {dw_unique _key , Extended Supply at Retail, Mailorder}

3.7 Algorithm Selection

Once we had the candidate samples and the feature set selected the next step was to run various classification algorithms to choose the most suitable techniques to operate on the datasets available. The suite of algorithms offered by Weka – Decision Trees, Bagging, Boosting, Naives Bayes and Neural Network – was applied on to the candidate samples selected. Both cross validation and percentage split methods were applied to train and test the algorithms on the various samples. The precision and the recall of each technique on each of the samples were compared to decide on both the representative samples i.e., those with good performance, and the algorithms suitable for these samples.

3.8 Representative Samples and Algorithms

From the analysis above, representative samples were formed by aggregating group of candidate samples with high accuracy and recall.

3.8.1 Member Profile Dataset

From the various member profile samples, two representative samples were formed:

- Sample1 : Aggregation of highly representative 2% samples (139,416 records)
- Sample2 : Aggregation of highly representative 5% samples (350,068 records)

3.8.2 Prescription Claims Dataset

Again for this dataset, two representative samples were formed by aggregating the 2% samples and 5% samples.

- rxclms_2per : Aggregation of highly representative 2% samples (604,677 records)
- rxclms_5per : Aggregation of highly representative 5% samples (3,082,793 records)

The algorithms selected to operate on these representative samples were Decision Trees, Naives Bayes and Neural Networks (Average accuracy of 88% and recall of 87% in Weka). Now that we have the completely preprocessed samples, we chose SQL Server 2005 Analysis Services to build the predictive models.

4. ANALYSIS OF MEMBER PROFILE INFORMATION

As we have seen earlier this dataset provides demographic and health related information of each member. The analyses were performed on the two representative samples selected and the 22 features chosen. The analysis is broken into three main categories as follows:

(1) Building predictive models using Classification techniques
(2) Clustering analysis of the prescription claim health conditions
(3) Association analysis

4.1 Classification Analysis

In our scenario the target of our classification analysis is to decide whether a member, given his/her characteristics, will utilize mail order pharmacy or not. Hence the class label is Mailorder = 1 or Mailorder = 0. The predictive models are built by running the training samples through the various classification algorithms chosen. In the future given a new object, the model will classify it into either a positive or a negative object.

The predictive models were built at 3 different levels of analysis

4.1.1 Predictive Model using the entire dataset

This higher-level analysis helps us to identify the dependence of mail order on each attribute and also the interdependence of demographics and the health condition features. The analyses were run on both the 2% and 5% samples to corroborate the consistency of the results obtained.

a. Dependency Network

This indicates the effect each attribute has on the class label, 'Mailorder = 1'. This could be either a positive or a negative effect i.e., presence of a feature value could boost mail order utilization rate or vice versa. The dependency network does not tell us anything about the nature of influence. It just indicates how strong of an effect (if at all any) an attribute has on the class label. The numbers on the links rank the influence, 1 being the strongest and 8 being the weakest.

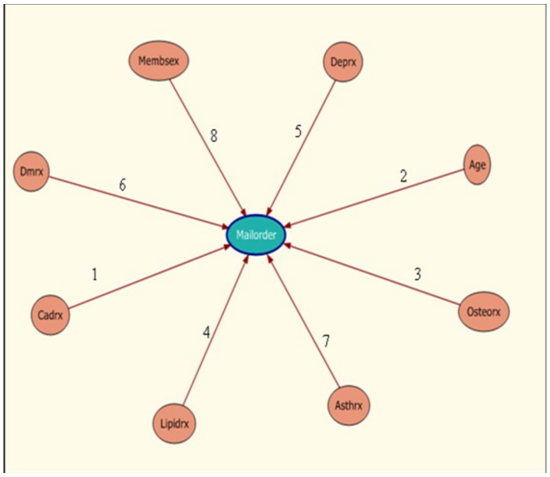


Figure 2: Dependency Network of Member Profiles on Mail order

From the figure we see that the pharmacy health condition 'Cadrx' has the strongest effect on Mailorder =1, followed by the member's Age. We also see a host of other health conditions having a stronger effect on the class label rather than the demographic features. Finally we observe that the gender of the member also seems to have an effect on mail order behavior. To analyze these effects in detail i.e., the nature of the influence, the attribute values that favor mail order etc, we proceed by running the various classification algorithms on the samples and studying the models built.

b. Decision Trees

Decision tree builds a model that gives a bigger picture of the strong links or paths that exist within the dataset. This helps us identify the combination of features that have an effect on mail order. The algorithm however does not show the effect of all of the attributes on the class label. It only comes up with the top attributes or nodes that provide high information gain. The algorithm prunes the tree to show only the strong predictors i.e., leaves having at least a minimum support.

Each node in the decision tree below has a horizontal line, an indicator of the percentage of records in that node that has a positive class label. The red portion indicates 'Mailorder' = 1 while blue indicates 'Mailorder = 0'. From the decision tree we can see that, 'Cadrx' – Coronary artery disease, which was the strongest influencer of mail order in the dependency network, comes up as the root node. We see that presence of Cadrx i.e., Cadrx = 1, has a very positive influence on Mailorder =1. 20% of the members having Cadrx utilize mail order pharmacy. In other words we can say that, if a member has Cadrx, then there is a 20% probability that the member will utilize mail order pharmacy. Such a high percentage of information gain makes Cadrx the strongest predictor of mailorder characteristic and hence the root node.

If we take the upper branch i.e., where Cadrx = 1, we see that next strongest predictor is the health condition Lipidrx – hyperlipidemia. We see that the presence of Lipidrx =1 positively affects mail order since the probability of 'Mailorder=1' increases to 23% when both Cadrx =1 and Lipidrx =1.

On other hand if Lipidrx = 0 and the member has diabetes (Dmrx = 1), then the probability of 'Mailorder=1' is 19%. We see the probability decrease from 20% (when Cadrx =1) to 19% (when Cadrx =1 and Dmrx =1) because the number of cases with Cadrx =1 alone is much higher than the number of records having a combination of Cadrx =1 Lipidrx =0 and Dmrx = 1.

Similarly when members have Cadrx and not Lipidrx or Dmrx (Cadrx =1, Lipidrx =0 and Dmrx =0), then the probability decreases to 11%.

Now if we follow the lower branch where Cadrx =0, we see that Age is the next strong predictor used to classify the objects. The algorithm splits age into 4 groups {Age<10, 10 <= Age < 30, 30<= Age < 50, Age >= 50}.

The top rules identified by the algorithm are as follows:

- If members are of Age < 10 and have asthma i.e., Asthrx =1, then there is a probability of 4% that such members will utilize mail order pharmacy. Though the rule has a low probability, it has a good information gain for decision tree to list it.

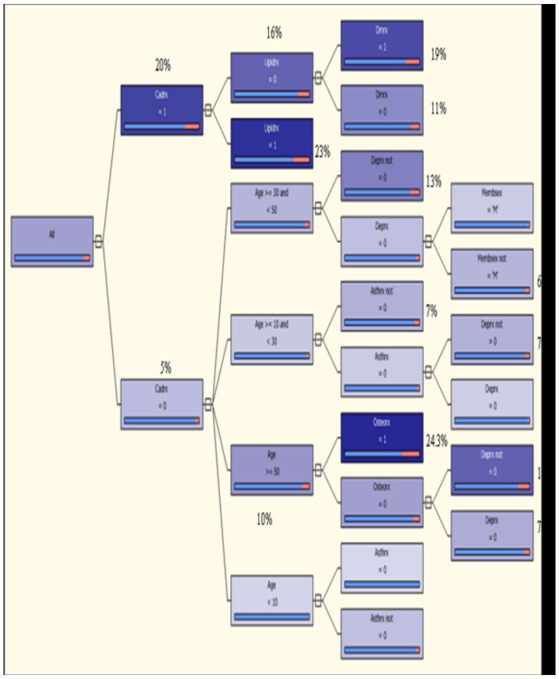


Figure 3: Decision Tree of Entire Member Profile Dataset

- Members with 10 <= Age < 30 and Asthrx =1 or Deprx =1 tend to utilize mail order pharmacy 7% of the times.
- Among members having 30 <= Age < 50 and depression Deprx =1, 13% of them utilize mail order pharmacy. In the same age group, Female members

seem to utilize mail order 6.4% of the time more than male. This explains the effect of Gender on Mailorder depicted in the dependency network.

- Finally older members with Age >= 50 having osteoporosis Osteorx =1 utilize mail order pharmacy 23.4% of the time. This is the strongest link identified by the algorithm. Thus we could guess with almost one-fourth probability that a member with the above condition is a good candidate for mail order. In the same age group members with depression Deprx =1 also seem to utilize mail order 16% of the time.

Thus decision tree helps us to identify potential combination of features, which are not so obvious from the raw dataset, that influence mail order behavior positively. Once these characteristics are identified, the marketing team can choose to act on certain combinations to improve mail order utilization rates.

c. Neural Networks

Next representative samples with all the features were run through neural network for various reasons. First and foremost reason being to corroborate the results seen in decision trees. Secondly decision trees do not capture the effect of each attribute on the class label. It only lists the top attributes and their combined effects. NN, on the other hand indicates how each feature relates to 'Mailorder=1'. It also indicates attribute values that have a negative effect on mail order pharmacy, unlike Decision trees that indicate only positive factors. NN is an adaptive system that changes its structure based on external or internal information that flows through the network.

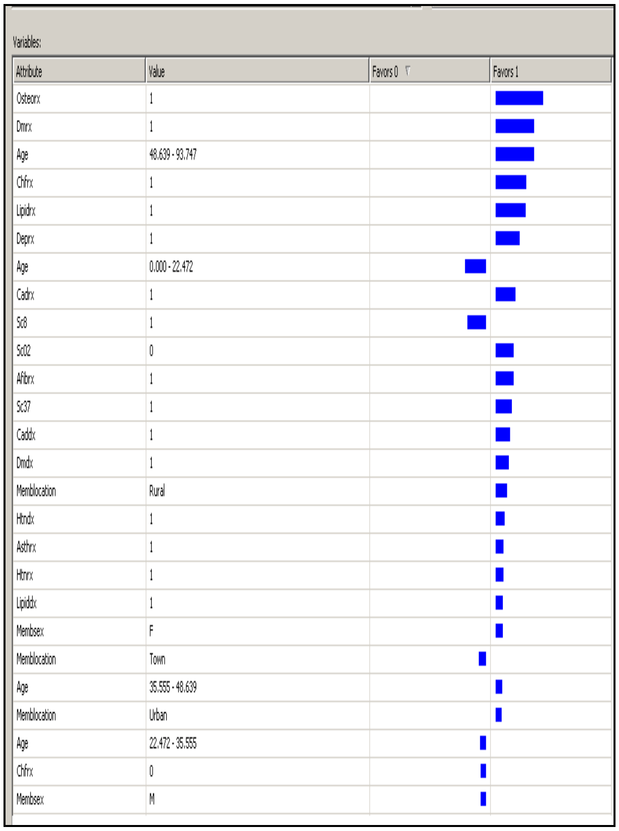
In the figure below, the two rightmost columns, Favors 0 and Favors 1, indicate which of the class label value each of the attribute values favors. Eg: Presence of Osteorx favors the class 'Mailorder =1'. This means that whenever a member has osteoporosis i.e., Osteorx =1, they tend to utilize mailorder pharmacy. The blue indicator gives value of lift that Osteorx =1 has on Mailorder =1. In simple words, it gives the probability that of all members having Osteorx =1 how many utilize mail order pharmacy.

Thus from the figure we see that all the health conditions that came up in the decision tree, Osteorx, Dmrx, Lipidrx, Cadrx Deprx, also come up as strong predictors in neural network. Here Osteorx seems to have a greater effect than Cadrx because NN indicates the lift of each attribute rather than the information gain. We saw in decision tree that cases with Osteorx =1 had a greater probability, of 23%, of mail order utilization than those with Cadrx =1, even though Cadrx is a much stronger predictor.

We also see that 'Age' between 48 to 94 years seem to have higher affinity towards mail order pharmacy than 'Age' between 22 and 48. Whereas 'Age < 22 years' has a negative effect on mail order. This is in agreement with the results of decision tree where higher affinity towards mail order is observed among members with Age > 10. NN also corroborates that mail order pharmacy is popular among female members much more than that of male

We find that a Charlson score of 0,1,2 or >=8 have a negative influence on 'Mailorder=1', while members with Charlson score between 3 to 7 i.e., Sc37 =1, tend to favor mail order pharmacy much more. We also find that members living in Rural and Urban areas support mail order much more than members living in towns. This could be due to the absence of chain stores or any good retail pharmacy in a rural area and due to the busy life styles of urban members. Thus NN gives us some unseen and highly insightful results about the influence of member location and health score on mail order utilization rate.

d. Naive Bayes Classifier This classifier is used to build a predictive model assuming independence among attributes. It captures the effect of each attribute on the class labels independent of the influence of other attributes. Like the other two techniques, Naïve Bayesian also identifies health conditions to have the strongest influence on mail order.



Variables:

Attribute	Value	Favors 0	Favors 1
Osteorx	1		■
Dmrx	1		■
Age	48.639 - 93.747		■
Chfrx	1		■
Lipidrx	1		■
Deprx	1		■
Age	0.000 - 22.472	■	
Cadrx	1		■
Sc8	1	■	
Sc02	0		■
Afibrx	1		■
Sc37	1		■
Caddx	1		■
Dmdx	1		■
Memblocation	Rural		■
Htndx	1		■
Asthrx	1		■
Htnrx	1		■
Lipiddx	1		■
Membsex	F		■
Memblocation	Town	■	
Age	35.555 - 48.639		■
Memblocation	Urban		■
Age	22.472 - 35.555	■	
Chfrx	0	■	
Membsex	M	■	

Figure 4: Results of Neural Network technique on member profile data

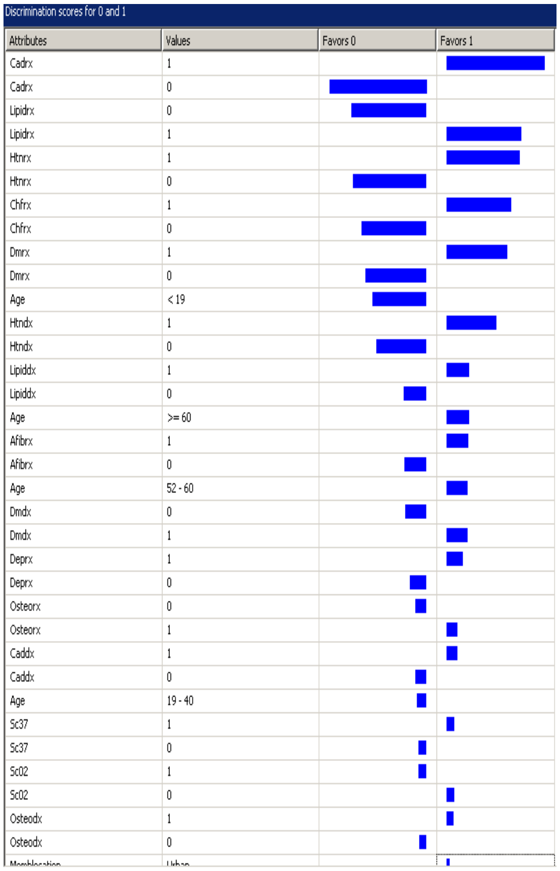
Member age of 50 and higher is associated with a positive effect on mail order utilization compared to age < 19. It also finds a Charlson score between 3 to 7 and Urban locations favorable to the class Mailorder =1. (Figure 5)

e. Predictive Performance of the Model

The above models built by Decision tree, NN and Naïve Bayes algorithms were tested on a completely new sample taken from the original dataset. The model built should not only perform well on the data set used to train it, but also show acceptable accuracy in predicting the class label of unseen records i.e., test data. A test sample of 35,000 records were used to test the predictive accuracy of the models

Table 11: Predictive Accuracy of models built on member profile data

🟩	Decision Tree	70%
🟪	Neural Network	71%
🟨	Naïve Bayes	70%
🟥	Ideal Model	100%
🟦	Random	50%

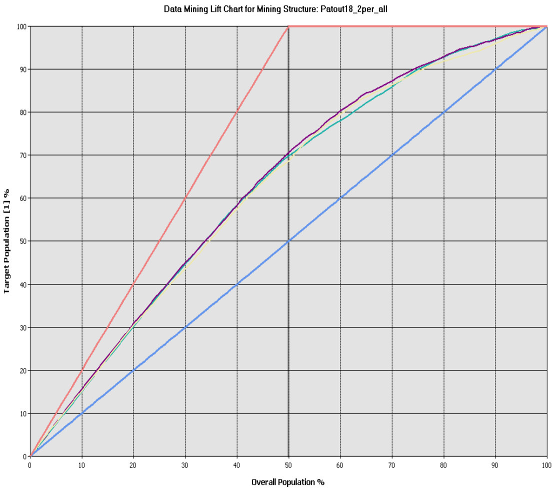


Discrimination scores for 0 and 1

Attributes	Values	Favors 0	Favors 1
Cadrx	1		■
Cadrx	0	■	
Lipidrx	0	■	
Lipidrx	1		■
Htnrx	1		■
Htnrx	0	■	
Chfrx	1		■
Chfrx	0	■	
Dmrx	1		■
Dmrx	0	■	
Age	< 19	■	
Htndx	1		■
Htndx	0	■	
Lipiddx	1		■
Lipiddx	0	■	
Age	>= 60		■
Afibrx	1		■
Afibrx	0	■	
Age	52 - 60		■
Dmdx	0	■	
Dmdx	1		■
Deprx	1		■
Deprx	0	■	
Osteorx	0	■	
Osteorx	1		■
Caddx	1		■
Caddx	0	■	
Age	19 - 40	■	
Sc37	1		■
Sc37	0	■	
Sc02	1	■	
Sc02	0		■
Osteodx	1		■
Osteodx	0	■	
Memblocation	Urban		■

Figure 5: Result of Naïve Bayes technique on member profile data

Figure 6: Lift Chart giving Predictive Accuracy of models built on member profile data



4.1.2 Predictive Model of the Demographic Features

The previous analysis combined the demographic and health features to identify the strong predictors and also the interrelation between the demographic and health features. As we saw in the previous analysis, the health condition features seem to dominate the demographic features. In the decision trees, we were able to find combined factors mostly consisting of health condition features and 'age'. Hence the aim of this analysis was to isolate the demographic information from the health condition features and try to build models that help identify rules involving just the demographic information that were obscured in the previous analysis. However one has to bear in mind that except for Age, the other member demographic features are only secondary predictors when compared to the pharmacy claim health conditions. But the combined factors mined in this analysis could add on to our knowledge base and enrich our rule set.

a. Decision Trees:

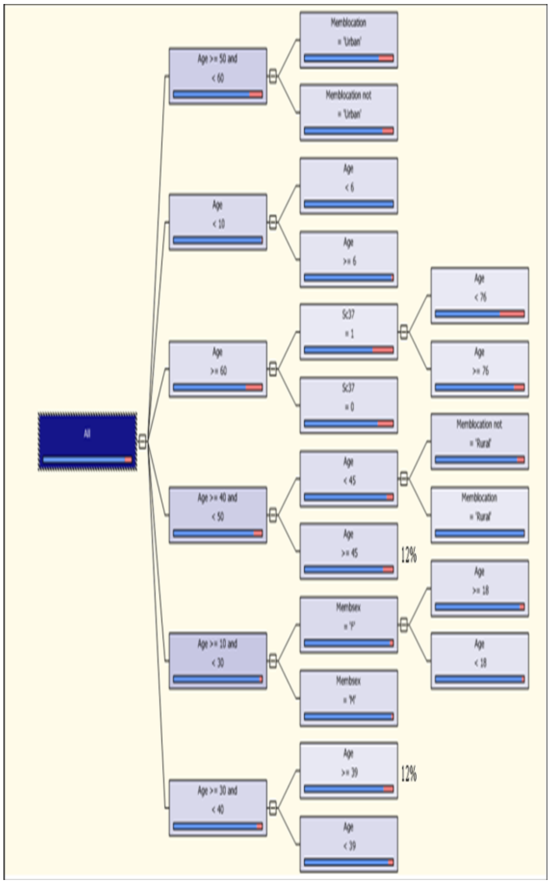


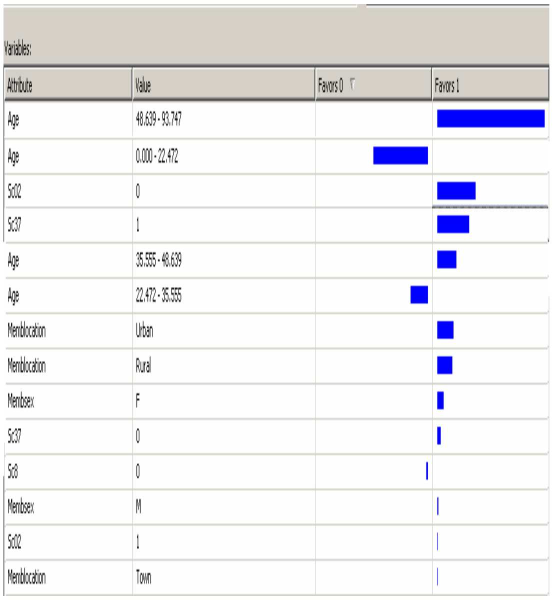
Figure 7: Decision Tree of Member Demographics

Here again we see that 'Age' seems to be the strongest predictor of all the features, appearing at the root level. However we see that Score, Member Location and Member sex appear in the decision tree, which was not the case in the previous analysis. When the health features are removed, the effects of the other demographic features come to the fore. We see that Member location of Urban and Rural have a good percentage of positive samples. We find that Sc37 positively affects mail order. The other node which says 'Sc37 = 0' consists mostly of samples where Sc8 =1. Thus members with Charlson score > =8 also utilize mail order, but in smaller percentages when compared members with Charlson score between 3 and 7.

With this decision tree we can find new combined factors such as:

- Members with Age >=60 and Sc37 =1 or Sc8 =1 are potential targets for mail order
- Members with 50<= Age <60 and Location = Urban / Rural are also good candidates
- Female members with Age > 18 are more likely to respond to mail order pharmacy.

b. Neural Network

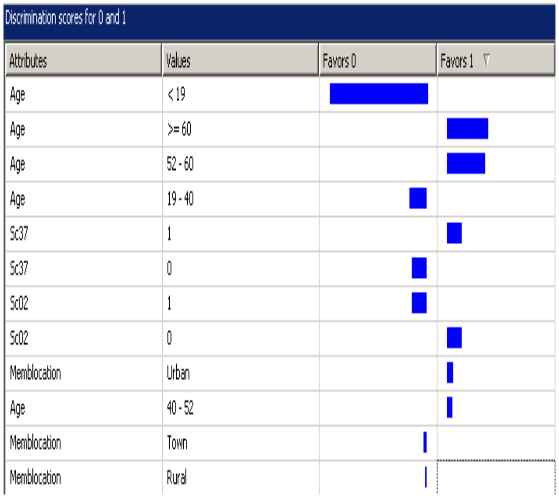


Variables:

Attribute	Value	Favors 0	Favors 1
Age	48.639 - 93.747		■
Age	0.000 - 22.472	■	
Sc02	0		■
Sc37	1		■
Age	35.555 - 48.639		■
Age	22.472 - 35.555	■	
Memblocation	Urban		■
Memblocation	Rural		■
Membsex	F		■
Sc37	0		■
Sc8	0	■	
Membsex	M		■
Sc02	1		■
Memblocation	Town		■

Figure 8: Neural Network model of Member Demographics

c. Naïve Bayes:



Discrimination scores for 0 and 1

Attributes	Values	Favors 0	Favors 1
Age	< 19	■	
Age	>= 60		■
Age	52 - 60		■
Age	19 - 40	■	
Sc37	1		■
Sc37	0	■	
Sc02	1	■	
Sc02	0		■
Memblocation	Urban		■
Age	40 - 52		■
Memblocation	Town	■	
Memblocation	Rural	■	

Figure 9: Naïve Bayes model of Member Demographics

d. Predictive Performance

The models built are tested with a test sample of 35,000 records. The predictive accuracy of the models is as follows:

Table 12: Predictive Accuracy of models built on member demographics

🟩	Decision Tree	66%
🟪	Neural Network	67%
🟨	Naïve Bayes	66%
🟥	Ideal Model	100%
🟦	Random	50%

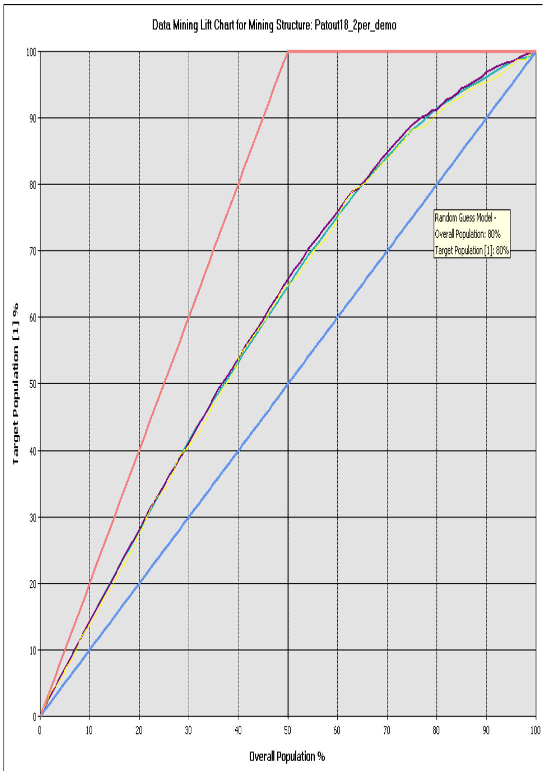


Figure10: Lift Chart giving Predictive Accuracy of models built on member demographics

We see that the accuracy of the models have gone down when compared to the first analysis. This is only to be expected considering that we have omitted the features related to health condition, which are the strongest predictors. Though this model has decreased in performance a little, it helped us identify new combination of demographic factors that did not come up in the first analysis.

4.2 Clustering Analysis of Member Health Features

The health conditions of the members indicated in the member profile dataset can be used to group members according to the similarity of their health conditions.

Motivation: Clustering members on their health conditions and analyzing their group characteristics can help us identify the combination of health features that have a strong link to mail order pharmacy. In the future when we see a new member with certain set of health conditions we can immediately identify the member's group and hence ascertain the probability of the member utilizing mail order pharmacy.

Before proceeding with the clustering analysis, we first formed a vector of the nine prescription claim health conditions for each member along with their past mail order behavior. For each member we computed a vector of the form: {Memid, Cadrx, Lipidrx, Deprx, Dmrx, Htnrx, Chfrx, Osteorx, Afibrx, Asthrx, Mailorder}. Each of these is a binary feature, with a '0' indicating absence of the health condition / past mail order behavior, while a '1' indicating the presence.

The clustering techniques employed are those provided by SQL Server 2005

- K – means clustering
- EM Clustering

a. Cluster Diagram

This diagram graphically displays the clusters that the algorithm discovered in the data set. The layout in the diagram represents the relationships of the clusters, where similar clusters are grouped close together [3]. The shade of the node color represents the number of positive samples (Mailorder =1) in the cluster—the darker the node, the more positive samples it contains.

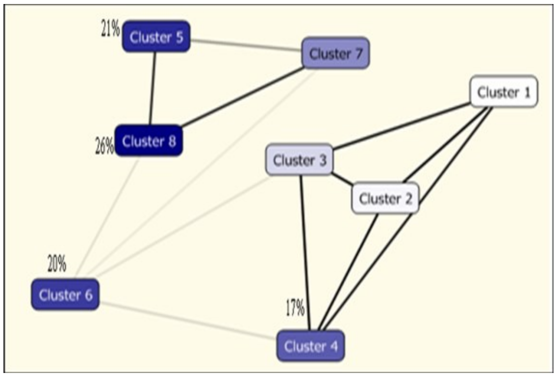


Figure 11: Cluster Diagram of the Prescription Health Conditions

The algorithm identifies eight natural clusters i.e., eight different combinations of health conditions that exist in the data set. Of these, clusters 8, 5, 6, 4 and 7 are made up of a good percentage of positive samples.

b. Cluster Profiles

Figure 12 displays the combinations of various health conditions that make up each cluster. The clusters make up individual columns and each row represents the various health features and the last row is the class label. The aqua color in each of the cell represents the negative samples i.e., absence of the characteristic, while the red color indicates the presence. From the last row in the cluster profile diagram, which gives the percentage of positive mail order samples in each cluster, we find that clusters 8, 5, 6, 4 and 7 have greater affinity towards mail order pharmacy and hence are of greater interest.

Cluster 8: The members belonging to this cluster have the following health conditions {Cadrx, Lipidrx, Afibrx, Chfrx, Dmrx, Htnrx} and they utilize mail order 26% of the time. Based on this information, in the future if we find a member with a combination of any or all of these conditions, we can predict with a 26% confidence that he/she will utilize mail order pharmacy.

Cluster 7: This is a subset of cluster 8 with very small percentages of Lipidrx and Dmrx. Therefore, the information given by this cluster is subsumed by Cluster 8 and can be omitted.

Cluster 5: The members belonging to this cluster have the following health complaints {Cadrx, Lipidrx, Chfrx, Dmrx, Htnrx} and they utilize mail order 21% of the time.

Cluster 6: This cluster consists of members having {Cadrx, Lipidrx} complaints alone and they utilize mail order 20% of the time. This combination of Cadrx and Lipidrx seems to be pretty dominant and hence appeared as the top predictors in the decision tree built as well.

Cluster 4: This cluster consists of {Deprx, Asthrx} in smaller percentages and has a mail order rate of 17%.

We also see that the health conditions making up these clusters are those that were identified as strong predictors in the classification analysis.

c. Performance of Clustering Analysis

To test the accuracy of the clusters identified and the performance of the model on unseen records, a set of 35,000 completely different members' profiles was provided as the test set. We see a **predictive accuracy of 70%** on the test data.

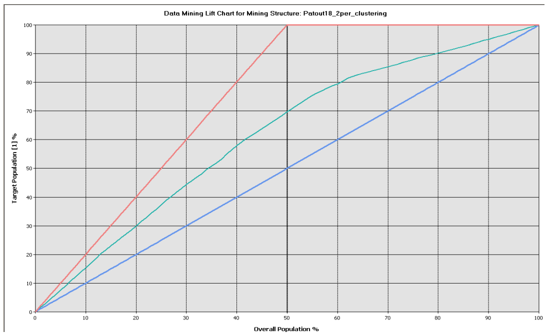


Figure13: Lift Chart of the Clustering model

4.3 Association Analysis of Member Profile Information

This analysis was performed as a follow up on the previous two analyses. Association analysis generates hundreds of rules from the dataset and gives the co-occurrence probabilities of class labels and the various features. However what is of greater interest is the effect of each attribute value on the class label Mailorder =1.

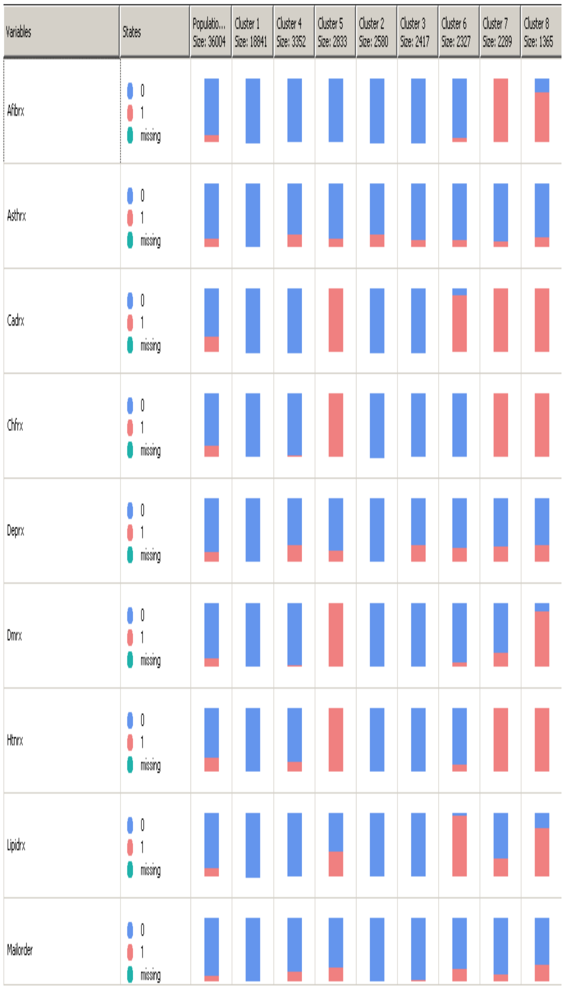


Figure 12: Cluster Profiles of the Prescription Health Conditions

From the dependency network (Figure 14) below we see that association analysis also identifies the health conditions and 'age' (greater than 40) as the strong predictors of mail order characteristics. Below is also a snapshot of some of the co-occurrence rules (Figure 15) generated by the algorithm. The predictive accuracy of this analysis on the test samples is also 70%.

5. ANALYSIS OF PHARMACY CLAIMS

Once we identified the member level predictors of mail order pharmacy, the next set of analyses involved mining the pharmacy claims data set to identify the claim level features that promote mail order acceptance rate. The analyses were performed on the two representative samples selected from the claims dataset during the preprocessing step. (Refer page 7 for the data set features).

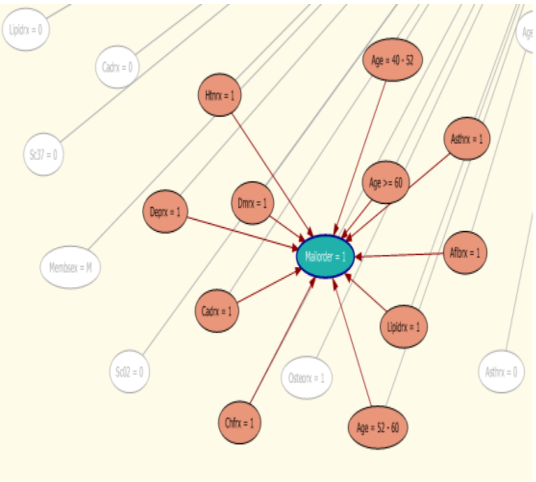
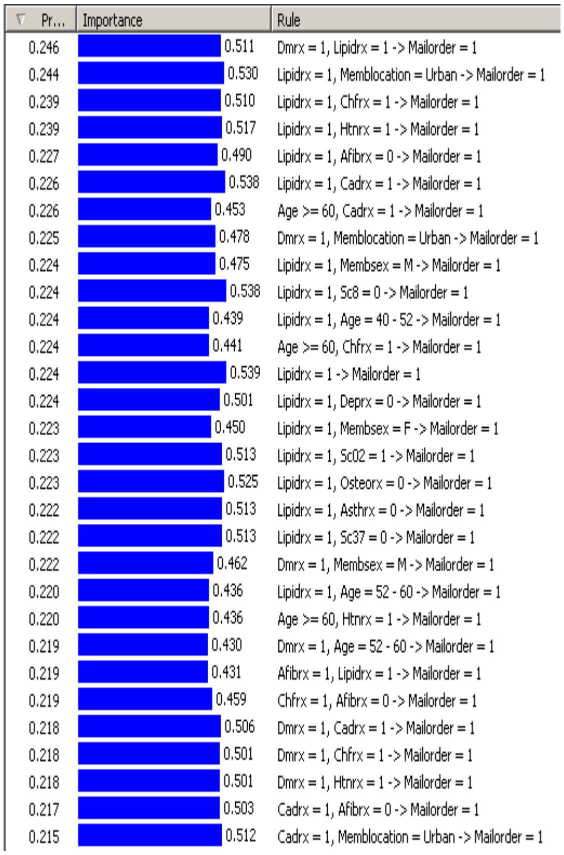


Figure 14: Dependency Network of Association Analysis of Member Profiles



Pr...	Importance	Rule
0.246	0.511	Dmrx = 1, Lipidrx = 1 -> Mailorder = 1
0.244	0.530	Lipidrx = 1, Memblocation = Urban -> Mailorder = 1
0.239	0.510	Lipidrx = 1, Chfrx = 1 -> Mailorder = 1
0.239	0.517	Lipidrx = 1, Htnrx = 1 -> Mailorder = 1
0.227	0.490	Lipidrx = 1, Afibrx = 0 -> Mailorder = 1
0.226	0.538	Lipidrx = 1, Cadrx = 1 -> Mailorder = 1
0.226	0.453	Age >= 60, Cadrx = 1 -> Mailorder = 1
0.225	0.478	Dmrx = 1, Memblocation = Urban -> Mailorder = 1
0.224	0.475	Lipidrx = 1, Membsex = M -> Mailorder = 1
0.224	0.538	Lipidrx = 1, Sc8 = 0 -> Mailorder = 1
0.224	0.439	Lipidrx = 1, Age = 40 - 52 -> Mailorder = 1
0.224	0.441	Age >= 60, Chfrx = 1 -> Mailorder = 1
0.224	0.539	Lipidrx = 1 -> Mailorder = 1
0.224	0.501	Lipidrx = 1, Deprx = 0 -> Mailorder = 1
0.223	0.450	Lipidrx = 1, Membsex = F -> Mailorder = 1
0.223	0.513	Lipidrx = 1, Sc02 = 1 -> Mailorder = 1
0.223	0.525	Lipidrx = 1, Osteorx = 0 -> Mailorder = 1
0.222	0.513	Lipidrx = 1, Asthrx = 0 -> Mailorder = 1
0.222	0.513	Lipidrx = 1, Sc37 = 0 -> Mailorder = 1
0.222	0.462	Dmrx = 1, Membsex = M -> Mailorder = 1
0.220	0.436	Lipidrx = 1, Age = 52 - 60 -> Mailorder = 1
0.220	0.436	Age >= 60, Htnrx = 1 -> Mailorder = 1
0.219	0.430	Dmrx = 1, Age = 52 - 60 -> Mailorder = 1
0.219	0.431	Afibrx = 1, Lipidrx = 1 -> Mailorder = 1
0.219	0.459	Chfrx = 1, Afibrx = 0 -> Mailorder = 1
0.218	0.506	Dmrx = 1, Cadrx = 1 -> Mailorder = 1
0.218	0.501	Dmrx = 1, Chfrx = 1 -> Mailorder = 1
0.218	0.501	Dmrx = 1, Htnrx = 1 -> Mailorder = 1
0.217	0.503	Cadrx = 1, Afibrx = 0 -> Mailorder = 1
0.215	0.512	Cadrx = 1, Memblocation = Urban -> Mailorder = 1

Figure 15: Snapshot of Association Rules Generated

5.1 Classification Analysis

The aim of this analysis was to build predictive models using claim level features to identify the independent effect of each attribute, such as the policy details, cost of the purchase, the details of the drugs purchased etc, on mail order pharmacy. The unit of analysis here are the individual prescriptions filled by the customers. Once a predictive model is built using the training samples, then whenever a new prescription is filled in the future, the various attributes of the purchase could be run through the model and analyzed to decide whether the member could be motivated to utilize mail order for that purchase in the future or not.

5.1.1 Analysis of the Policy Features and Normalized Financial Features

The nature of the health care policy owned by each individual and the amount they spend on each of their medications could greatly influence the mail order behavior of the individual.

Motivation: To focus more closely on the policy and the financial features alone to identify the attribute values and rules having a positive influence on mail order.

The attributes of interest for this analysis are: {Deductible Flag, Out of Pocket Amount Flag, Benefit Max Flag, Normalized_totpaid, Normalized_Membpaid, Normalized_planpaid Mailorder}

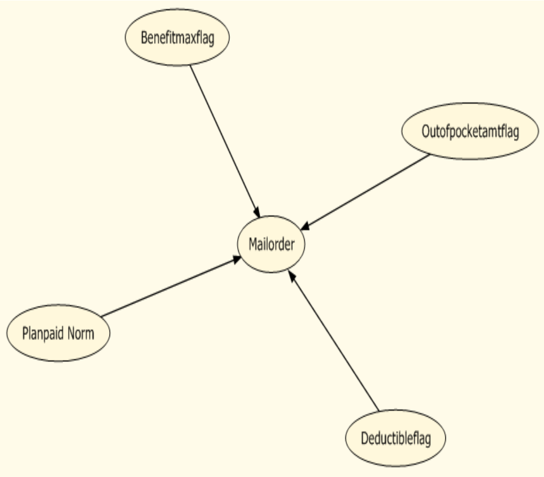


Figure 16: Dependency Network of Policy and Financial Features

a. Decision trees

We found that 'Out of Pocket Amount' flag is strongest predictor of mail order behavior. The absence of 'Out of Pocket Amount' in the pharmacy benefit design was associated with higher the mail order utilization. Among members whose policy does not have an 'Out of Pocket Amount', 12% of members utilize mail order pharmacy, while the probability is only 4% among members who have an 'Out of Pocket Amount'. Thus, presence of 'Out of Pocket Amount Maximum' has a negative effect on mail order pharmacy utilization.

Similarly the presence of a deductible in the pharmacy benefit design was also associated with a decrease in the probability of members utilizing mail order pharmacy. In cases where both 'Out of Pocket Amount' flag and 'Deductible' flag are absent, the Mailorder =1 probability goes up to 14%.

A pharmacy benefit design with 'Benefit Max' positively influences mail order utilization rate. Members with policies having the 'Benefit Max' clause invariably utilize mail order much more than members without 'Benefit Max' clause. Thus, members who have policies containing the following combination of features were the most likely to utilize mail order.

{OutofPocketAmount = absent, Deductible = absent, BenefitMax = X} (20%)

b. Neural Network

The model built by Neural Network also emphasizes the same facts about the influence of the various policy features on mail order rate. However they also evaluate the threshold values of the various normalized costs that benefit mail order utilization. Thus we can, with considerable confidence, believe that members whose daily medication costs exceed the evaluated thresholds are more likely to utilize mail order pharmacy.

Table 13: Summary of Results for Policy and Financial Features Analysis

Technique	Strong Predictors	Moderate Predictors

Policy Features	Out of Pocket Amount Flag = absent
	DeductibleFlag = absent
	BenefitMaxFlag = present	

Financial Features	\$3.325 <= Plan paid cost / day <= 9.535 \$4 <= Total cost / day <= 11 Member paid cost / day <= 0.5	\$1.52 <= Plan paid cost / day <= \$3.325 Total cost / day < 2 \$1.524 <= Member paid cost / day <= \$3.373

c. Predictive Accuracy of the Model

The effectiveness of the model built depends on its ability to predict the class label of new prescriptions based on the policy and cost features. Hence the model was tested on a completely unseen test sample with 134,876 prescription claims. (Refer Figure 17 and Table 14)

5.1.2 Analysis Involving the Drug Details

Once we have analyzed the influence of policy and financial features on mail order, the next step was to include the details of the drugs purchased to identify the combined effect they have on the mail order purchase rate. The feature set included:

{Out of pocket amount flag, Deductible flag, Benefit max flag, GPI2, Normalized_Totpaid, Normalized_Planpaid, Normalized_Membpaid, Generic, Formulary, Specialty, Extended_supply_retail, Mailorder}

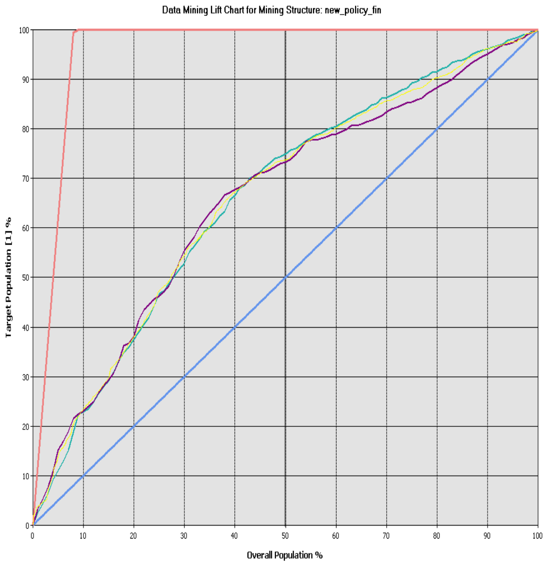


Figure 17: Predictive Performance of the Classification model built using Policy and Financial Features

Table 14: Predictive Accuracy of model built using Policy and Financial Features

🟩	Decision Tree	75%
🟪	Neural Network	73%
🟨	Naïve Bayes	74%
🟥	Ideal Model	100%
🟦	Random	50%

a. Dependency Network

Once again we find that the policy features seem to have the most dominant effect on mail order (Figure 18). However with the inclusion of the drug details, the drug family and the nature of the drug (generic, formulary etc) seem to have a greater influence on the mail order behavior than the financial features.

b. Decision Tree

The whole decision tree built for this analysis has 14 levels, indicating various combinations of policy and drug features that positively influence mail order. Below is a summary of the various rules generated by the tree (Table 15), along with the support and confidence of each rule indicating its usefulness.

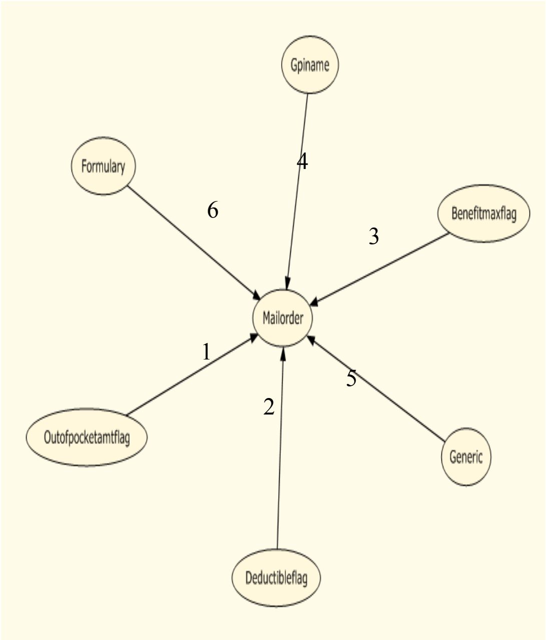


Figure 18: Dependency Network involving the Drug Details

Table 15: Rules generated by Decision Tree using Policy and Drug Details

Out of Pocket Amt Flag	Deductible Flag	Benefit Max Flag	GPI2	Formulary	Support (cases)	Confidence (%)
X			Analgesics Anti-Inflammatory		5500	3
X			Beta-Blockers		10745	3
X			Anticonvulsants		5450	3
X	X		Antineoplastics		10	50
	X		Antihyperlipidemics			6
		X	Anticonvulsants		1290	8
		X	ADHD/Antinarcolepsy/Anti…		812	6
		X	Analgesics Anti-Inflammatory		1516	11
		X	Thyroid	NF	560	18
		X	Thyroid Agents	F	1404	7
		X	Endocrine Metabolic		546	36
		X	Antihyperlipidemics		3950	26
		X	Antineoplastics		104	48

Each row in the table identifies a particular combination of prescription claim features that promote mail order. For example consider the following cases:

- The row with Thyroid drugs indicates that 18% of all Non Formulary Thyroid Drugs purchases by members whose health care policies have Benefit Max flag, go through mail order pharmacy.

- Whereas if the drug purchased is a Formulary Thyroid Agent then, there is a 7% probability that such a purchase will be through mail order.

Table 16: Rules generated by Decision Tree using Drug Details

Generic	GPI2	Formulary	Support (cases)	Confidence (%)
0	ADHD/Anti-narcolepsy/Anti-Obesity…			5
	Anti-Asthmatic Bronchodilators		4815	10
	Thyroids Agents		1685	10
	Anti- Diabetics		4471	12
	Anticonvulsants			
	Anticonvulsants	NF	354	21
		F	1581	7
	Antiparkinson			
		F	373	16
1	Anticonvulsants		2164	7
	Analgesic Anti-Inflammatory	NF	274	13
	Analgesic Anti-Inflammatory	F	3383	4

- If the drug purchased is a Non – Generic (=0), Non Formulary Anticonvulsant, then there we can guess with a 21% confidence that the member making the purchase will utilize mail order pharmacy.
- Similarly if the purchase is a Generic (=1) and Non Formulary version of Analgesic Anti-Inflammatory drug, then there is a 13% probability that the member will make the purchase through mail order.

c. Neural Network and Naïve Bayes

The Neural network and Naïve Bayes model also depict the same relationship between the various drug characteristics and mail order utilization behavior.

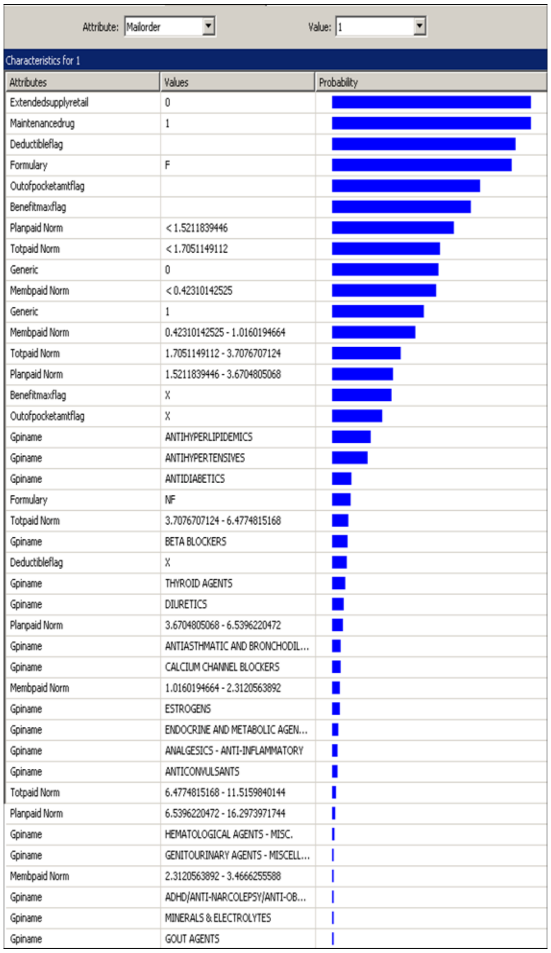
The drugs families that came up in the decision tree are seen to favor the class Mailorder =1, while the others favor class Mailorder = 0. As seen in earlier analysis, the effect of the various policy features on mail order and the cost thresholds identified remained the same in this analysis as well.

Additionally we found that if the drug purchased is a Specialty i.e., Specialty =1, then it favors mail order pharmacy greatly. There is a very high probability that people buying specialty drugs utilize mail order often.

For formulary and non formulary drugs, we found that NF drugs favor mail order more compared to Formulary drugs. We see that Formulary drugs favor class Mailorder =0, but only to a small extent. E.g.: If there are 100 cases of formulary drugs purchased, may be 60 purchases are made at retail store while 40 are through mail. Thus we can see that formulary drug purchases also go through mail order, but the number of retail purchases are slightly higher. But in case of NF drugs the majority of their purchases are made through mail, but some also go through retail stores. This is why we find positive rules for both NF and F generated by the decision tree. Thus NF/F is not an independent predictor. Their ability to influence mail order rate depends on other features like the drug purchased, the policy owned by the member etc.

This argument holds also for the Generics. If the drug purchased is a Non-Generic (Generic = 0), then it favors mail order most of the time, whereas majority of Generic drugs are bought at retail pharmacy. However there is some percentage of cases that go the opposite way for both Generic and Non Generic drugs. Again these are also not independent predictors.

Naïve Bayes model also indicates the same effect for Formulary and Generics. We see that F, NF, Generics and Non-Generics promote mail order, but the extent to which they favor mail order depends on other attributes. Naïve Bayes also indicates that the attribute 'Extended supply at retail' is an independent predictor of mail order pharmacy. If the purchase is a long-term medication bought at a retail store i.e., Extended supply at retail = 1, then such purchases never occur through mail. Hence it is a strong predictor of anti-mail order behavior.



Attribute: Mailorder Value: 1

Characteristics for 1

Attributes	Values	Probability
Extendedsupplyretail	0	
Maintenancedrug	1	
Deductibleflag		
Formulary	F	
Outofpocketamtflag		
Benefitmaxflag		
Planpaid Norm	< 1.5211839446	
Totpaid Norm	< 1.7051149112	
Generic	0	
Membpaid Norm	< 0.42310142525	
Generic	1	
Membpaid Norm	0.42310142525 - 1.0160194664	
Totpaid Norm	1.7051149112 - 3.7076707124	
Planpaid Norm	1.5211839446 - 3.6704805068	
Benefitmaxflag	X	
Outofpocketamtflag	X	
Gpiname	ANTIHYPERLIPIDEMICS	
Gpiname	ANTIHYPERTENSIVES	
Gpiname	ANTIDIABETICS	
Formulary	NF	
Totpaid Norm	3.7076707124 - 6.4774815168	
Gpiname	BETA BLOCKERS	
Deductibleflag	X	
Gpiname	THYROID AGENTS	
Gpiname	DIURETICS	
Planpaid Norm	3.6704805068 - 6.5396220472	
Gpiname	ANTIASTHMATIC AND BRONCHODIL...	
Gpiname	CALCIUM CHANNEL BLOCKERS	
Membpaid Norm	1.0160194664 - 2.3120563892	
Gpiname	ESTROGENS	
Gpiname	ENDOCRINE AND METABOLIC AGEN...	
Gpiname	ANALGESICS - ANTI-INFLAMMATORY	
Gpiname	ANTICONVULSANTS	
Totpaid Norm	6.4774815168 - 11.5159840144	
Planpaid Norm	6.5396220472 - 16.2973971744	
Gpiname	HEMATOLOGICAL AGENTS - MISC.	
Gpiname	GENITOURINARY AGENTS - MISCELL...	
Membpaid Norm	2.3120563892 - 3.4666255588	
Gpiname	ADHD/ANTI-NARCOLEPSY/ANTI-OB...	
Gpiname	MINERALS & ELECTROLYTES	
Gpiname	GOUT AGENTS	

Figure 19: Naïve Bayes model using Drug Details

d. Performance of the model

The predictive accuracy of the models were tested using a test set of 134,876 prescription claims different from those used to build the model. The lift chart is given below:

Table 17: Predictive Performance of the model using Drug Details

🟩	Decision Tree	75%
🟪	Neural Network	75%
🟨	Naïve Bayes	75.2%
🟥	Ideal Model	100%
🟦	Random	50%

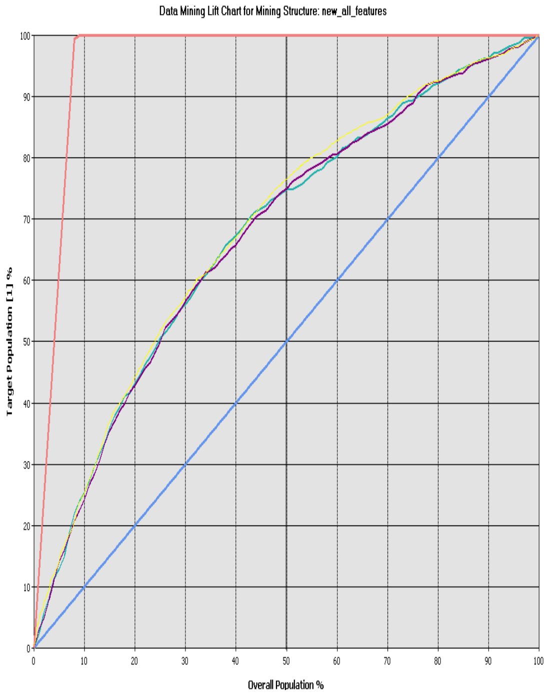


Figure 20: Predictive Performance of the Classification model using Drug Details

Thus we find that the details about the drugs purchased helped improve the accuracy of the predictive models by a percent or two.

5.2 Clustering Analysis of the GPI2 drug family

The drugs purchased by a member, indicated by the two-digit GPI drug family, can be used to group members according to their health conditions and the type of medications purchased.

Motivation: Clustering members in this manner and analyzing their group characteristics can help us identify the combination of drugs families that have strong link to mail order pharmacy. In the future when we see a member purchase a particular drug we can immediately ascertain how likely is it for the member to respond to a mail order campaign. Before proceeding with the clustering analysis, we first formed a vector of the 39 different drug families (refer Page 6) for each purchase. Thus for each record in the representative sample, we computed a vector of the form:

{Dw_Unique_key, Antidiabetics, Thyroid, Antihypertensives, Antihyperlipidemics, …, }

Each of these is a binary feature, with a '1' indicating that a drug belonging to that particular drug family was included in the prescription filled. The clustering techniques employed are those inbuilt within SQL Server 2005

- K – means clustering
- EM Clustering

a. Cluster Profiles

The algorithms identified ten natural clusters i.e., ten different combinations of drug purchased, which positively affect mail order acceptance rate. Some clusters are made up of only one drug family, while some are made up of a group of drug families indicating co-occurence relationship between the various drug families. The drug families frequently purchased through mail order are:

- Antihyperlipidemics
- Antihypertensives
- Antidiabetics
- Thyroid
- Beta-Blockers
- Calcium Channel Blockers
- Diuretics
- Estrogens, Analgesics Anti-Inflammatory
- Anti-Asthmatic Brocholidators, Anticonvulsants
- Endrocrine Metabolics Agents, Analgesics Anti-Inflammatory, Adhd

Correlation between the GPI2 drug families and the Rxclms

We find that most of the drugs identified through the clustering analysis are those that address the various health conditions, identified to have a strong influence on mail order behavior in earlier clustering analysis of member health profile details.

Table 18: Correlation between Health Claims and Drugs Purchased

Health Conditions (Rxclms)	Drug Family
Cadrx, Chfrx, Afibrx	Beta-Blockers, Calcium Channel Blockers
Lipidrx	Antihyperlipidemics
Dmrx	Antidiabetics
Htnrx	Antihypertensives
Asthrx	Anti-Asthmatic Brocholidators, Anticonvulsants

b. Performance of Clustering Analysis

To test the accuracy of the clusters identified and the performance of the model on unseen records, a set of 134,876 completely different prescription claims data, consisting of both positive and negative samples, was provided as the test set. We see a predictive accuracy of 72% on the test data.

6. ANALYSIS AND RESULTS OF FINANCIAL FEATURES

This analysis of the financial features aimed at identifying the thresholds of the annualized costs spent by each member and their health care provider on their medication. The feature set consists of total amount, member paid amount and plan paid amount aggregated over 18 months for each member. The prediction model was built using Decision Trees and Neural Network algorithms, as they are the most potent technique for working with continuous attributes. The NN model is depicted below.

The feature set is as follows:

Financial_det{Memid,Cumm_Totpaid, Cumm_Membpaid, Cumm_PlanPaid, Mailorder}

6.1 Neural Network Model of Annualized Medication Costs

- **Annual Member Paid Amount Threshold** From the model, we find that when the annual medication cost spent by a member exceeds $690, then there is a very high probability, almost 1, that the member will adopt mail order pharmacy. If the annual member expense is between $335 and $690, then there is reasonable probability that the member will accept mail order purchase.

- **Annual Plan Paid Amount Threshold** If the member's health care provider ends up spending more than $3468, then such members will utilize mail order more than half of the time. Hence the threshold for Cumm_Planpaid is $3468 and greater.

- **Annual Total Medication Cost Threshold** If the total annual medication cost of a member exceeds $3992, then there is a high probability (63%) these members will utilize mail order pharmacy. If total annual costs are less than $3392 and greater than $1500, these members utilize mail order, but with lesser probability (27%).

Table 19: Results of Financial Analysis

Feature	Strong Predictors	Moderate Predictors
Annualized Total Cost	>= $4000	$1500 to $4000
Annualized Member Costs	>= $690	$335 to $690
Annualized Plan Costs	>= $3470	

6.2 Performance of the Model

The predictive accuracy of the model was evaluated by running a test sample with 35,000 records consisting of member annualized cost information.

Table 19: Predictive Performance of Annualized Cost Model

🟩	**Decision Tree**	**71%**
🟪	**Neural Network**	**72%**
🟥	**Ideal Model**	**100%**

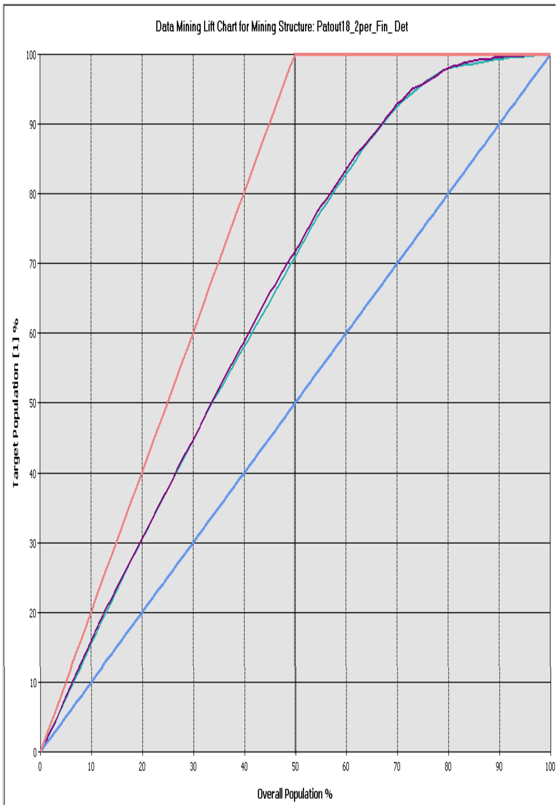


Figure 21: Predictive Performance of Annualized Cost Model

7. ANALYSIS USING THE ACXIOM FEATURES

Six different Acxiom member related elements, listed in pages 4 and 5, were integrated into the previous models with the aim of mining more member profile information that could promote mail order pharmacy utilization.

The rationale behind incorporating Acxiom data in the models is that if at all there is some relationship (positive or negative) between a member's education, marital status, mail order buying behavior, length at current residence, number of members in household, and PC Internet access to his/her mail order utilization rate, then it would come to fore.

The two representative samples derived earlier from the member profile dataset were used as training data sets for this analysis. The Acxiom data, corresponding to the members in these representative samples, were appended to the member profile information. The details of the policy held by a member i.e., Out of pocket amount flag, deductible flag, benefit max flag, were appended to these samples from the prescription claims data set. The various policy flags were set for a member if at least one of their prescription claims had these flags set.

Thus the final feature set used for these analyses is as follows:

- **Member demographics and policy features:** {Memid, Age, Sex, Memblocation, Sc02,Sc37,Sc8, Out of pocket amount flag, Deductible flag, Benefit Max flag, Mailorder}
- **Acxiom Data:** {Education, PCInternet, Marital Status, No of Children, Length at Residence, Mailorder buyer}

a. Dependency Network

From the dependency network we see that once again **Age,** followed by the **policy features** have the strongest influence on mail order. Then comes the **Educational status** of the member followed by their demographics like gender, health scores (Sc37) and their location. Then among the Acxiom features, the 'marital status' and 'mail order buyer' seem to exercise an influence on mail order behavior.

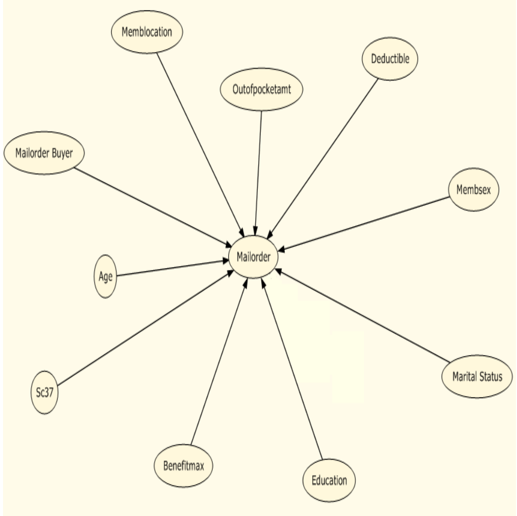


Figure 22: Dependency Network of Member Profile with Acxiom data

b. Consolidated Results

Building predictive models for the training data using Decision Trees, Neural Network and Naïve Bayes classifiers yield the following consolidated results.

Table 20: Result of Analysis using Acxiom Data

Strong Predictors	Moderate Predictors
Mailorder_buyer = B	
PCInternet = 1	
Marital_status = Married	
No of Children = 0	No of Children = 4 or 7
Length of residence >= 14 years	Length of residence >= 7yrs

Some of rules generated that promote mail order utilization are:

- 11 <= Age <= 44, Out of Pocket Amount flag = 0, Deductible flag =0, Benefit Max flag, Mail order Buyer = B
- 33 <= Age <= 44, Deductible flag =0, Benefit Max flag = 0, Charlson score > = 8
- 44 <= Age <= 55, Out of pocket amount flag = 0, Deductible flag = 1, Benefit Max flag = 0, Member Location = Urban

- Out of pocket amount flag = 0, Deductible flag = 1, Benefit Max flag = 1, Charlson score between 3 to 7

- Out of pocket amount =0, Benefit Max =0, Educational status =1 i.e., member has completed high school

- Out of pocket amount =0, Benefit Max =0, Member Location = Urban

Although the Acxiom attributes influence mail order behavior, the extent to which they have an impact on mail order utilization is weak, unlike the member demographics. They are not alone sufficient to predict mail order behavior. However models built integrating the Acxiom features with the member profile information improves the predictive performance by 8% to 9%. The predictive accuracy of this model on a test data with 35,000 records is as follows:

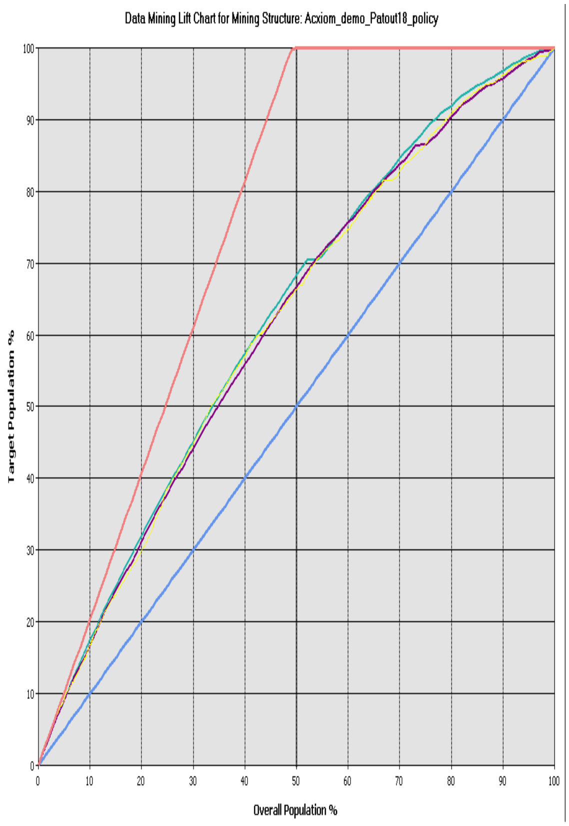


Figure 23: Performance of the Integrated Model - Member Profile with Acxiom data

Table 21: Performance of Integrated Model - Member Profile with Acxiom data

🟩	Decision Tree	69%
🟪	Neural Network	67%
🟨	Naïve Bayes	67%
🟥	Ideal Model	100%

8. CONCLUSION AND FUTURE WORK

Through all the data mining techniques employed and models built, we found that member level details and their prescription claims provide great insights into the mail order behavior of the people. From all the analyses performed so far, we particularly identified the following features as strong and good predictors of mail order utilization

1. Member demographics – Age, Location, Sex, Charlson score of 3 to 7
2. Health Indicators – Pharmacy claim health conditions
3. Policy details and Annual expense on medication
4. Drug details

Some of the future directions which we see worth pursuing are:

a. Modeling complex and better performing algorithms for SQL Server:

The algorithms employed for the above analyses are the ones existing in SQL Server i.e., Decision Trees, Naïve Bayes and Neural Networks. Since this was our first attempt at mapping Data Mining and KDD to health care claims, we concentrated on experimenting various approaches that could maximum knowledge, rather on employing different algorithms. Much stronger predictive models could be built by using algorithms like SVM, Boosting and Bagging etc on this data. In fact building combination of classifiers like Decision trees and boosting, could lead to a better predictive performance. Of course all these involve great labor and time in terms of coding up these classifiers in SQL Server or even building plug-ins for SQL Server.

b. Designing Target Marketing Programs

From the different potential member profiles identified by the analyses, the marketing team at Prime could design different marketing programs to target various member profiles. Eg: Sending out mails to attract potential members with Age < 30 or sending snail mails for people in rural areas etc.

c. Clustering members to identify the programs that suit them

Once these various marketing programs are in place, we could use Clustering techniques to cluster the potential members according to the programs that would suit them the most. This helps us identify the appropriate marketing program for each member we would like to contact.

d. Building Scoring models

Once we know the potential target members and the program corresponding to them, we could build some scoring models or models with good visualization abilities, to rank the members. This would help us come up with an order in which to contact the members. The members with high scores are more likely to respond positively to mail order campaigns, than those with low scores. Hence depending upon the funds available we can evaluate how many of the top rank members to contact to achieve our goal mail order utilization rate.

e. Studying Inconsistent Members

So far in all of the above analyses, we only concentrated on members enrolled for the entire 18 month duration. We could also study, in much more detail, the behavior of inconsistent customers. This could give us some information such as

(i) Which quarter of the year the members remain enrolled in the policy,

(ii) Does their enrollment pattern suggest something about their total medical expense and hence the mail order behavior

(iii) Is there any relation between mail order behavior and yearly quarters i.e., do people tend to take to mail order for purchases made at the end of the year or beginning of the year etc.

Analyzing such trends can help us know when to contact a particular member with the target marketing programs.

A Complete Prediction Framework for Mail-order Utilization:

This would be something really interesting and very useful to implement. We could build an entire system, which when given a set of new members,

1. Applies the chosen classification algorithms (Step a) to identify the potential members who can be targeted. i.e., predicts mail order behavior for each member

2. Of the positive members filtered from the above step, it goes on to identify the most suitable marketing program for each potential member (Step c)

3. Then orders the potential members according to their ranks given by the scoring model.

4. Feedback loop: The system could also be designed to take feedback about the actual success of the program on each member and use the information to learn more about the member and improve the model dynamically.

9. REFERENCES

[1] Charlson, M.E., Pompei, P., Ales, K.L., & Mackenzie, C.R. (1987). A new method of classifying prognostic co morbidity in longitudinal studies: Development and validation. Journal of Chronic Disease, 40(5), 373-383.

[2] Factors Associated With Choice of Pharmacy

Setting Among DoD Health Care Beneficiaries Aged 65 Years or Older: Andrea Linton, MS; Mathew Garber, PhD; Nancy K. Fagan, DVM, PhD; and Michael Peterson, DVM, DrPH

[3] Google

[4] Impact of Alternative Interventions on Changes in

Generic Dispensing Rates: A. James O'Malley, Richard G. Frank, Atheer Kaddis, Barbara M. Rothenberg, and Barbara J. McNeil

[5] Introduction to Data Mining: Pang-Ning Tan;

Michael Steinbach and Vipin Kumar

[6] Introduction to the Practice of Statistics:

Davis.S.Moore; George.P.McCabe

[7] Maintaining the Affordability of the Prescription Drug Benefit: How Managed Care Organizations Secure Price Concessions from Pharmaceutical Manufacturers

[8] Microsoft SQL Server 2005 Tutorial

[9] Prime Mail Brochure, Article, Sell Sheet,

Newsletter

[10] PrimeTherapeutics Website

[11] Wikipedia – For definitions and concepts

ACKNOWLEDGMENTS

I would like to extend my thanks to Professor Jaideep Srivasta, for giving me the opportunity to work on such an interesting project and for all his valuable insights, guidance and unrelenting encouragement throughout the course of this project. I would like to sincerely acknowledge and thank Dr.Patrick Gleason, Prime Therapeutics LLC, for his support and readiness to help me at each and every stage during this work. I would also like to thank the Prime Therapeutics team for all the time they have devoted to me and for sharing the enthusiasm to apply new techniques to the healthcare domain.

Last, but not the least, a big thanks to the fellow students of the Data Mining Research group, for their help during the initial stages of this project.

Data Mining Approach to Credit Risk Evaluation of Online Personal Loan Applicants

Pavel Brusilovsky
Business Intelligence Solutions
http://www.bisolutions.us
150 Borrows Street,
Thornhill, ON L4J 2W8
Canada
T: 647.588.7777
Pavel@BIsolutions.us

ABSTRACT

We describe the problem of credit risk evaluation of online personal loan applicants. Credit risk scoring is implemented within the data mining universe, using the stochastic gradient boosting algorithm. Discussion is concentrated around the specificity of the data and problem, the selection of an appropriate modeling method, determining drivers of the probability of being a good customer, and estimation of the impact of different predictors on this probability. The synergy of data mining and spatial techniques is useful for this type of problems.

General Terms

Algorithms, Performance, Design, Verification.

Keywords

Credit scoring, credit risk, probability of being a good customer, stochastic gradient boosting, predictor importance, GIS

32. 1. INTRODUCTION

As a rule, a lender must decide whether to grant credit to a new applicant. The methodology and techniques that provide the answer to this question is called credit scoring. This paper is dedicated to the development of credit scoring models for online personal loans.

Taking into account the non-linearity of the relationship between overall customer risk and predictors, the primary objective is to develop a non-parametric credit scoring model within data mining paradigm that will predict overall customer risk with maximum possible accuracy. This objective implies several goals:

1. Create a regression type credit scoring model that predicts overall customer risk on a 100 point scale, using the binary assessment of customer risk (*good* customer/*bad* customer).
2. Identify the importance of the predictors, and the drivers of being a good customer in order to separate good behavior from bad.
3. Develop the basis for a customer segmentation model that uses overall customer risk assessment to predict high (*H*), medium (*M*) and low (*L*) risk customers.
4. Show the fruitfulness of the synergy of credit scoring modeling and Geographical Information Systems (GIS).

The outcome of the regression scoring model can be treated as the probability of being a *good* customer. The segmentation rule depends on two positive thresholds $h1$ and $h2$, $h2<h1<1$. If for a given customer the probability of being a good customer is greater than $h1$, where $h1$ is a large enough threshold (e.g., 0.75), then the customer belongs to the low risk segment. If, however, the probability of being a good customer is less than $h1$ but greater than $h2$ (e.g., $h2$=0.5), then the customer belongs to the medium risk segment. Finally, if the probability that the customer is a good customer is less than $h2$, he belongs to the high risk segment. The thresholds $h1$ and $h2$ should be provided by an owner of a database, or their optimal values can be determined by BIS as a result of minimization of the corresponding cost matrix.

DMCS'09, June 28, Paris, France.
Copyright 2009 ACM 978-1-60558-674-8…$5.00.

Risk scoring is a tool that is widely used to evaluate the level of credit risk associated with a customer. While it does not identify "good" (no negative behavior) or "bad" (negative behavior expected) applicants on an individual basis, it provides the statistical odds, or probability, that an applicant with any given score will be "good" or "bad" [6, p.5].

Scorecards are viewed as a tool for better decision making. There are two major types of scorecards: traditional and non-traditional. The first one, in its simplest form, consists of a group of “attributes” that are statistically significant in the separating good and bad customers. Each attribute is associated with some score, and the total score for applicant is the sum of the scores for each attribute present in the scorecard for that applicant.

Traditional scorecards have several advantages [6, p.26-27]:

- Easy to interpret (there is no requirement for Risk Managers to know in depth statistics or data mining);
- Easy to explain to a customer why an application was rejected;
- Scorecard development process is transparent (not a black box) and is widely understood;
- Scorecard performance is easy to evaluate and monitor.

The disadvantage of traditional scorecards is their accuracy. As a rule, non-traditional scorecards (that can be represented as a data mining non-parametric logistic regression) outperform traditional scorecards. Since each percent gained in credit assessment accuracy can lead to a huge savings, this disadvantage is crucial for credit scoring applications. Modern technology allows us to easily employ a very complex data mining scoring model to new applicants, and to dramatically reduce the misclassification rate for Good – Bad customers.

This paper is dedicated to non-traditional scorecard development within a data mining paradigm.

2. METHODS

2.1 Data Description

This study is based on the Strategic Link Consulting (SLC) sample of 5,000 customers, including 2,500 Good customers and 2,500 Bad customers, one record per customer [1]. According to the rule of thumb [6, p. 28], one should have at least 2,000 bad and 2,000 good accounts within a defined time frame in order to get a chance to develop a good scorecard. Therefore, in principle, the given sample of accounts is suitable for scorecard development.

Each customer is characterized by 50 attributes (variables) that are differently scaled. The following variable types are present in the data:

- numeric (interval scaled) variables such as *age*, *average salary*, credit *score* (industry specific credit bureau), etc;
- categorical (nominal), with a small number of categories such as *periodicity* (reflects payroll frequency) with just 4 categories;
- categorical, with a large number of categories (e.g., *employer name*, *customer’s bank routing number*, *e-mail domain*, etc)
- date variables (*application date*, *employment date*, *due date*, etc)

The data also include a geographic variable (*customer ZIP*), and several customer identification variables such as *customer ID*, *user ID*, *application number*, etc. All ID variables were contaminated by random noise by the data owner. Unfortunately, the data does not include psychographic profiling variables.

There are several specific variables that we would like to mention:

- *BV Completed* is a variable that answers whether the customer had a bank verification completed by the loan processor. A value of 1 means the bank verification was completed. A missing value or 0 means it was not. Bank verification involves a 3 way call with the customer and their bank to confirm deposits, account status, etc.
- *Score* is an industry specific credit bureau score.
- *Email Domain* is a variable that reflects an ending part of the email address after the @ symbol.
- The variable *Monthly* means monthly income in dollars.
- *Required Loan Amount* is the principal amount of a loan at the time of origination.
- *Credit Model* is a predictor that can take the following values:
    - New customer scorecards – there are three credit bureau scorecards that exists, each with more stringent approval criteria. The baseline scorecard has only identity verification and an OFAC check while the tightest scorecard has a variety of criteria including inquiry restrictions, information about prior loan payment history, and fraud prevention rules. They are limited to standard loan amounts with standard fees, subject to meeting income requirements.
    - Returning customers have minimal underwriting and are eligible for progressively larger loan amounts with a fee below the standard fee for new customers.
- *Isoriginated* is either 1 for originated loans or 0 for unoriginated loans. Withdrawn applications and denied applications will have values of 0. Loans that were

funded and had a payment attempt will have a value of 1.

- *Loan Status* is the status of the loan. Loan statuses are grouped as follows:
  - D designates the class of Good Customers (a loan is paid off successfully with no return).
  - P, R, B, and C designate the class of Bad Customers.

Other variable names are self-explanatory.

It makes sense to consider a two-segment analysis of risk. In two-segment analysis, the target is a binary variable Risk (*Good, Bad*), or Risk Indicator (1, 0), where 1 corresponds to a Good customer (Risk = Good) and 0 corresponds to a Bad customer (Risk = Bad).

As we mentioned before, each target is associated with a unique optimal regression type model. The outcome of each model can be treated as the corresponding probability of target = 1 which, in turn, can be interpreted as a credit score on a 100 point scale. In other words, the model under consideration serves to estimate probability/credit score of being a Good customer.

2.2 Exploratory Data Analysis and Data Preprocessing

Exploratory Data Analysis (EDA) and data preprocessing are time consuming but necessary steps of any data analysis and modeling project [9]. The steps can significantly improve the quality of the model.

The objectives of EDA include understanding the data better, evaluating the feasibility and accuracy of overall customer risk assessment, estimating the predictability of Good/Bad customers, and identifying the best modeling methodology of credit scoring modeling.

Typical data preprocessing might include reduction of the number of categories, creation of new variables, treatment of missing values, etc. For example, in the categorical variable *Application Source*, the first four characters indicate the market source. It turned out that this variable has 45 distinct values, but only 18 categories are large. The rest of the categories were grouped into a new category, OTHER. it turns out that the constructed variable *Market Source Grouped* is selected as an important predictor, whereas the original *Market Source* variable is not.

Another problem with the data is the misspelling and/or double name of one and the same category for some categorical variables. In particular, the variable *Email Domain* has a lot of errors in the correct spelling of a domain. For instance, there are 5 different spelling versions of yahoo.com:

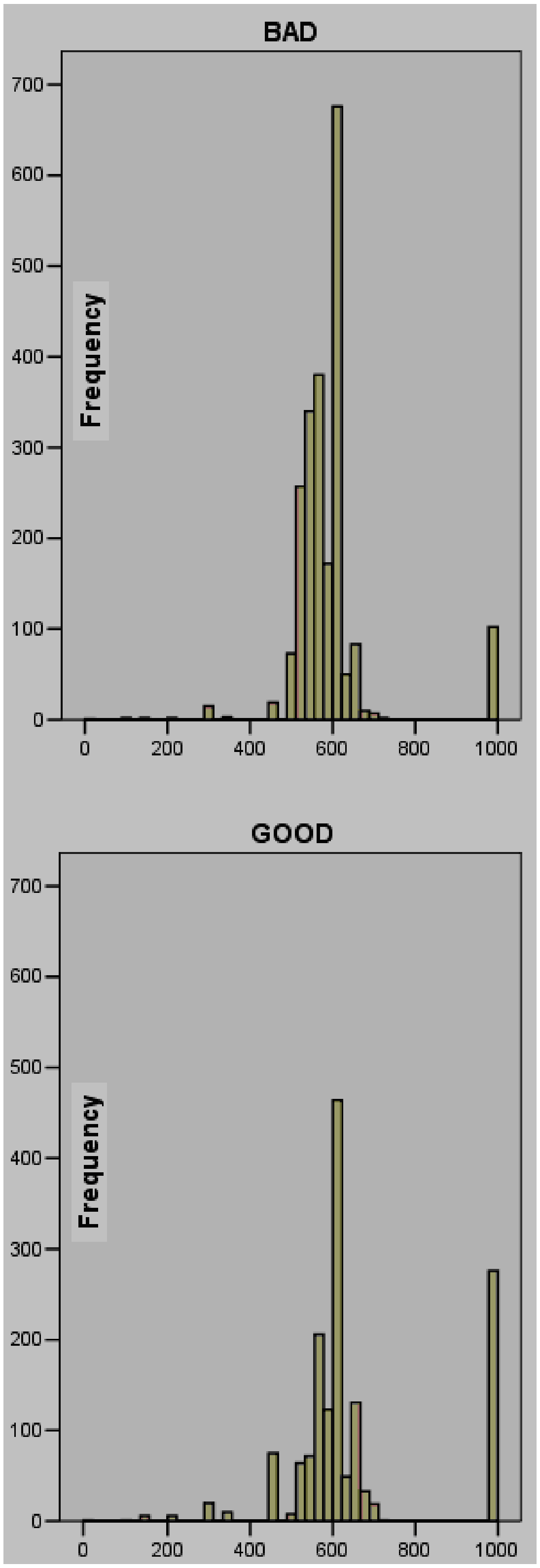
Email Domain	Number of Customers
yaho.com	2
yahoo.com	2023
yhaoo.com	3
Yahoo.com	6
YAHOO.COM	402
YAOO.COM	1

and 7 different versions of the domain sbcglobal.net:

Email Domain	Number of Customers
sbcglobal.ne	1
sbcglobal.net	194
sbcgloblal.net	1
sbcgolbal.net	1
sbclobal.net	1
SBCGLOBA.NET	1
SBCGLOBAL.NET	33

In order to produce meaningful results, all misspellings should be corrected.

According to our intuition, the variable *Score* is the most important to correctly predict the probability of being a good customer. The first thing that can be done is discriminating between customers, using just the *Score* predictor. *Graph 1* shows that it is not easy to do manually.



Graph 1. Distribution of the Variable Score for Bad and Good Customers

Construction of additional variables can dramatically improve the accuracy of risk prediction. New time duration variables

orig_duration = Origination Date – Application Date
emp_duration = Origination Date – Employment Date
due_duration = Loan Due Date – Origination Date

serve as examples of new variable creation. For the sake of illustrating the importance of data preprocessing, we can mention here that the latter two of these three variables were important predictors selected by the Stochastic Gradient Boosting algorithm.

The complexity the data under consideration can be characterized by:

- High dimensionality (about 50 predictors)
- Uncharacterizable non-linearities
- Presence of differently scaled predictors (numeric and categorical)
- Missing values for some predictors
- Large percentage of categorical predictors with extremely large numbers of categories and extremely non-uniform frequency distributions
- Non-normality of numeric predictors.

Therefore, complex sophisticated methods should be employed to separate good and bad accounts in the SLC data.

2.3 Analysis

Data and problem specificity limit the number of algorithms that can be used for SCL data analysis. Any traditional parametric regression modeling approach (such as statistical logistic regression) and any traditional nonparametric regression (such as Lowess, Generalized Additive Models, etc.) are inadequate for such problems. The main reason for this is the presence of a large number of categorical variables with huge numbers of categories. The inclusion of such categorical information in a multidimensional dataset imposes a serious challenge to the way researchers analyze data [10].

Any approach based on linear, integer or non-linear programming (see, for example, [7], Chapter 5), is also not the best approach for the same reasons.

Within the data mining universe, only some algorithms can be applicable to SLC data. For example, data mining cluster analysis algorithms available in data mining software such as SAS Enterprise Miner and SPSS Clementine are based on Euclidean distance and cannot be used for the same reasons as above.

On the other hand, preliminary analyses and modeling that we had conducted have shown that the accuracy of models generated by

Stochastic Gradient Boosting and Random Forest algorithms are acceptable.
We should note that the use of each of the applicable methods implies that the original data are randomly separated into two parts: the first is for *training* (model development) and the second is for *validation* of the model. Validation is the process of testing the developed model on unseen data.

Since a small improvement in model accuracy can lead to huge increase in ROI, the tradeoff between prediction accuracy and model representation simplicity we resolve in favor of the accuracy.

2.3.1 Stochastic Gradient Boosting Overview

Stochastic gradient boosting was invented in 1999 by Stanford University Professor Jerome Friedman [2, 3]. The first commercial tool – TreeNet - was released in 2002 by Salford Systems [4]. The intensive research has shown that stochastic gradient boosting models are among the most accurate of any known modeling techniques.

We will use both terms TreeNet and Stochastic Gradient Boosting interchangeably. Corresponding models are usually complex, consisting of hundreds (or even thousands) of trees, and require special efforts to understand and interpret the results. The algorithm generates a number of special reports with visualization to extract the meaning of the model, such as a ranking of predictors according to their importance on a 100 point scale, and graphs of the relationship between inputs and target.

2.3.2 TreeNet Risk Assessment models

For the purpose of our analysis we randomly selected the data sample into two subsamples. 60% of the sample builds the first subsample, the LEARN data. It will only be used for model estimation. The second subsample, the TEST data, will be used to estimate model quality.

The TreeNet algorithm has about 20 different options that can be controlled by a researcher. As a rule, usage of default options does not produce the best model. Determination of the best options/optimal model is time consuming and requires experience and expertise.

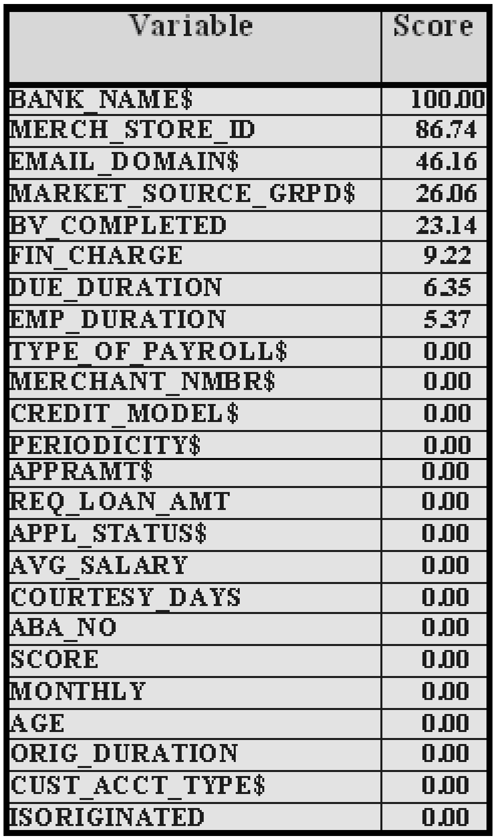
Models that are quite different (see First and Second models below) can have similar accuracy, and the interpretability criterion should be used to select the best model.

3. RESULTS

3.1 First Risk Assessment model

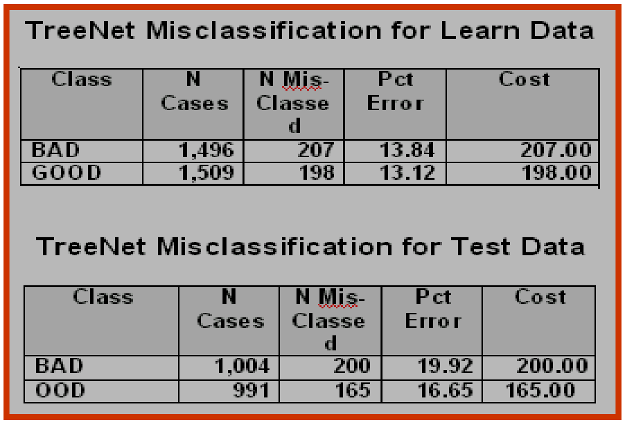
The target is a binary variable Risk with two possible values: *Good* and *Bad*. The *Good* value of the target was selected as a focus event. All predictors are listed in Table 1. The second column reflects an importance score on a 100 point scale with the highest score of 100 corresponding to the most important predictor. If the score equals 0, the predictor is unimportant at all for the target.

This particular model is based on just 8 predictors, but has a risk prediction error of about 14% on learning data, and a risk prediction error of about 19% on validation data. If the TreeNet algorithm did not select the *Score* variable, it means that within this model the variable Score is not important. It does not mean that the credit score is superfluous or irrelevant in customer credit risk assessment. It just means that the useful information provided by the variable *Score* is covered by 8 important predictors, selected by TreeNet (see Table 1). The misclassification rate is presented in Table 2.



Variable	Score
BANK_NAME$	100.00
MERCH_STORE_ID	86.74
EMAIL_DOMAIN$	46.16
MARKET_SOURCE_GRPD$	26.06
BV_COMPLETED	23.14
FIN_CHARGE	9.22
DUE_DURATION	6.35
EMP_DURATION	5.37
TYPE_OF_PAYROLL$	0.00
MERCHANT_NMBR$	0.00
CREDIT_MODEL$	0.00
PERIODICITY$	0.00
APPRAMT$	0.00
REQ_LOAN_AMT	0.00
APPL_STATUS$	0.00
AVG_SALARY	0.00
COURTESY_DAYS	0.00
ABA_NO	0.00
SCORE	0.00
MONTHLY	0.00
AGE	0.00
ORIG_DURATION	0.00
CUST_ACCT_TYPE$	0.00
ISORIGINATED	0.00

Table 1. Variable importance



TreeNet Misclassification for Learn Data

Class	N Cases	N Mis-Classed	Pct Error	Cost
BAD	1,496	207	13.84	207.00
GOOD	1,509	198	13.12	198.00

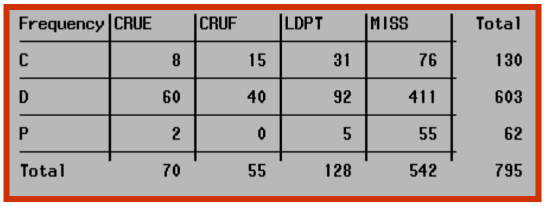
TreeNet Misclassification for Test Data

Class	N Cases	N Mis-Classed	Pct Error	Cost
BAD	1,004	200	19.92	200.00
OOD	991	165	16.65	165.00

Table 2. Misclassification Rate

Graph 2 presents the impact of the variable *MarketSourceGrouped* on the probability of being of a good customer. The Y-axis is a log odds of the event Risk =Good. Therefore, 0 corresponds to the situation when odds are 1-to-1, or the probability of an event equals the probability of a non-event. In other words, the X-axis corresponds to the base line that reflects an equal chance to be a good or bad customer.

The impact of *Market Source Grouped* is significant and varies across different values. All values of the *Market Source Grouped* variable with bars above the X-axis increase the probability of being a good customer, and all bars that are below the X-axis decrease the probability of being a good customer. We can say that the value of LDPT has the highest positive impact on the probability of being a good customer, and the value of CRSC has the highest negative impact on the probability of being a good customer.



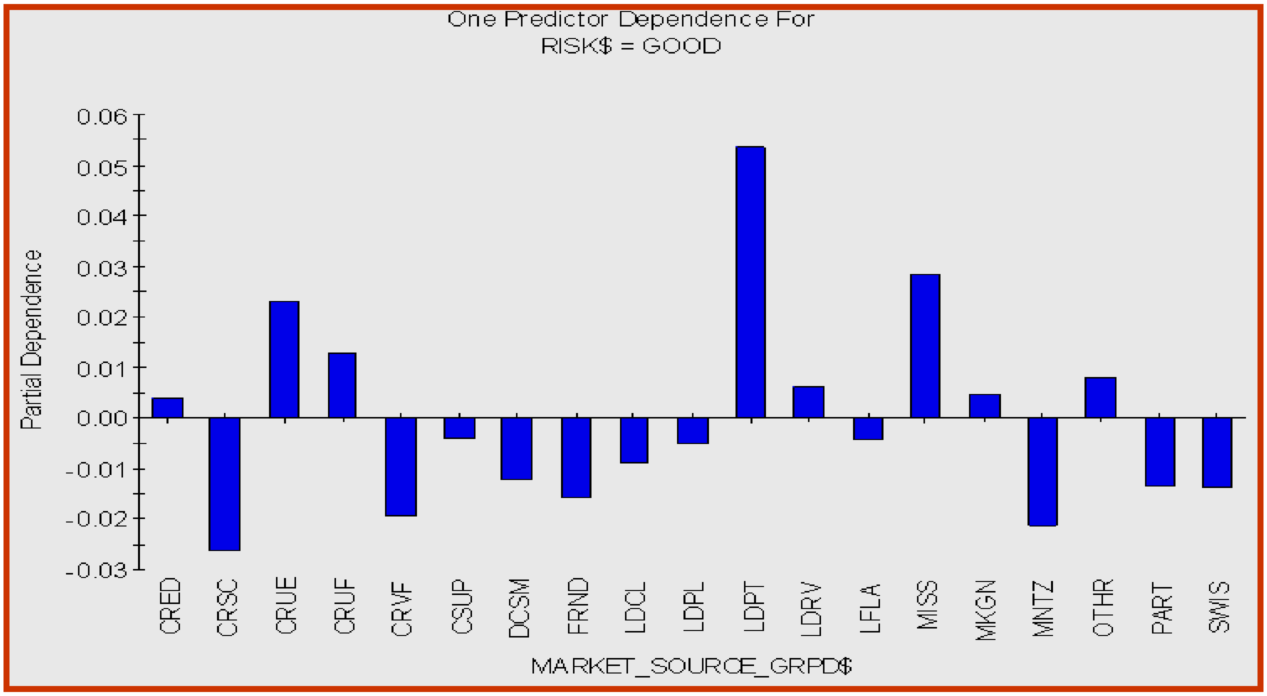
Frequency	CRUE	CRUF	LDPT	MISS	Total
C	8	15	31	76	130
D	60	40	92	411	603
P	2	0	5	55	62
Total	70	55	128	542	795

Table 3. Frequency of *Loan Status* by *Market Source Grouped*

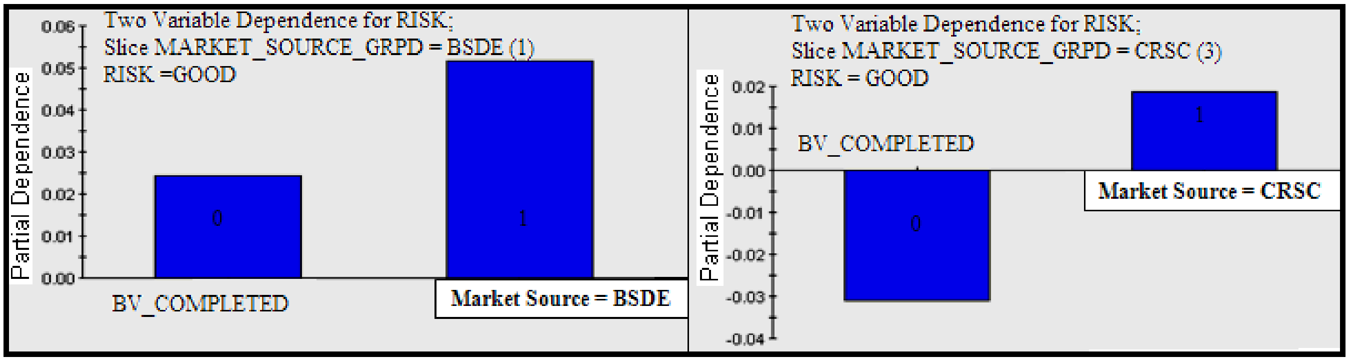
The left column of Table 3 depicts the values of the *Loan Status* predictor, and the upper row of the table depicts the values of the *Market Source Grouped* predictor. Table 3 pictured the frequency of customers that have the following values of the *Market Source Grouped* variable: CRUE, CRUF, LDPT, and MISS. These values are matched to the tallest positive bars on Graph 2 (the values with the highest positive impact on the probability of being a good customer). Since the value of D of the *Loan Status* predictor designates a Good customer, and the values C and P correspond to a Bad customer (see Data Structure section), we can infer that there is a good agreement between the model (Graph 3) and the data (Table 3).

There is a significant difference between the information presented in Table 3 and in Graph 2. If we forget for a minute about the existence of all other predictors, and consider just two of them (*Loan Status* and *Market Source Groupe*) using available data, then we can arrive at the conclusion that the majority of customers with *Market Source Grouped* values of CRUE, CRUF, LDPT, and MISS are Good customers. Again, we considered the join frequency distribution of only these two predictors, and disregarded the impact of all other predictors. In other words, there is no control for other predictors at all, and it is data induced information.

The information, represented by Graph 2, on the contrary, was produced by the developed TreeNet model, and it is model induced information. The relationship between target (log odds of being a good customer) and *Market Source Grouped* was mapped, controlling for all other predictors.



Graph 2. Impact of *Market Source Grouped* predictor on the probability of being a good customer: Risk = Good, controlling for all other predictors.



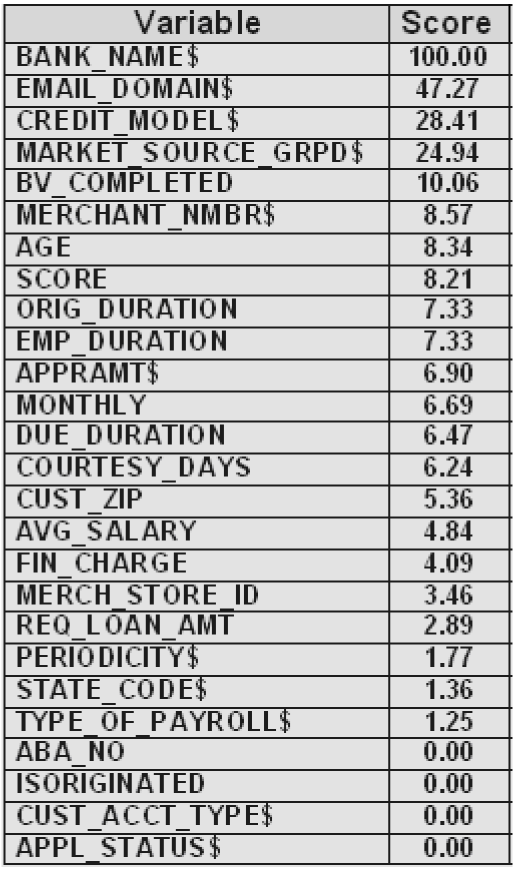
Graph 3. Impact of *BV Completed* and *Market Source Grouped* predictors on Probability of Being a Good Customer (controlling for all other predictors).

Graph 3 represents an example of the non-linear interaction between *BV Completed* and *Market Source* predictors: for different values of one predictor, the impact on the probability of being a good customer has different directions. Actually, for the value of *Market Source Grouped = BSDE* both values of *BV Completed* predictor have a positive impact on the probability of being a good customer. On the other hand, for the value of *Market Source Grouped* =CRSC*,* the value 0 of the *BV Completed* predictor accords with negative impact, but the value 1 accords with a positive impact on the probability of being a good customer.

3.2 Second Risk Assessment Model

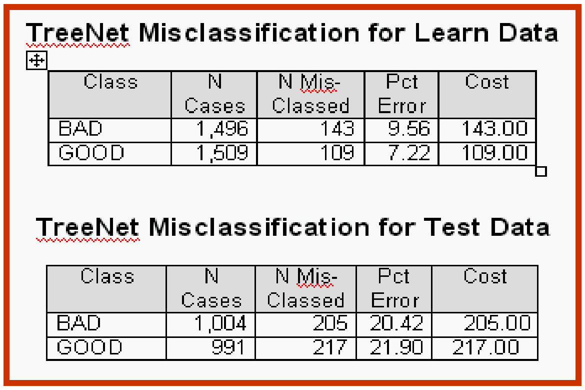
As in the First TreeNet model construction, 60% of the data are randomly selected to be used for model development (learning), and the remaining 40% of data used for model *validation* (holdout observations, or test data).

This particular model is based on 17 predictors, and has a risk prediction error of about 9% on learning data, and a risk prediction error of about 20% on validation data.



Variable	Score
BANK_NAME$	100.00
EMAIL_DOMAIN$	47.27
CREDIT_MODEL$	28.41
MARKET_SOURCE_GRPD$	24.94
BV_COMPLETED	10.06
MERCHANT_NMBR$	8.57
AGE	8.34
SCORE	8.21
ORIG_DURATION	7.33
EMP_DURATION	7.33
APPRAMT$	6.90
MONTHLY	6.69
DUE_DURATION	6.47
COURTESY_DAYS	6.24
CUST_ZIP	5.36
AVG_SALARY	4.84
FIN_CHARGE	4.09
MERCH_STORE_ID	3.46
REQ_LOAN_AMT	2.89
PERIODICITY$	1.77
STATE_CODE$	1.36
TYPE_OF_PAYROLL$	1.25
ABA_NO	0.00
ISORIGINATED	0.00
CUST_ACCT_TYPE$	0.00
APPL_STATUS$	0.00

Table 4. Variable importance



TreeNet Misclassification for Learn Data

Class	N Cases	N Mis-Classed	Pct Error	Cost
BAD	1,496	143	9.56	143.00
GOOD	1,509	109	7.22	109.00

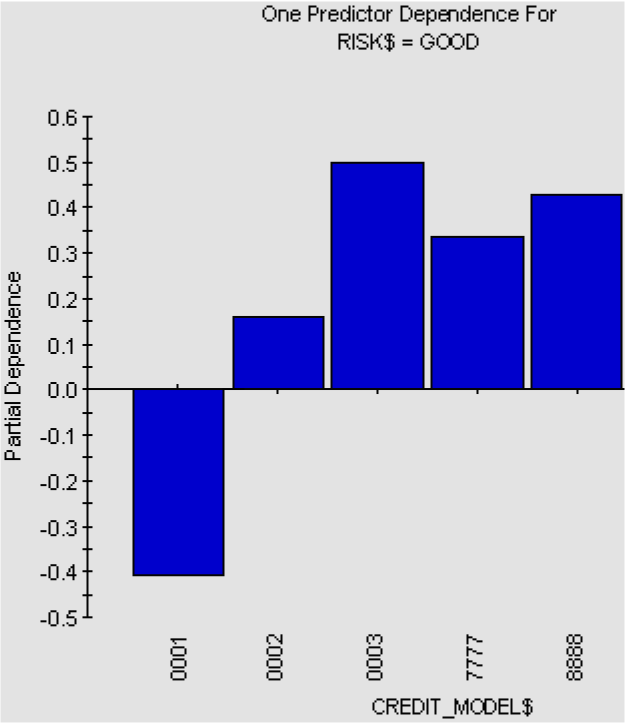
TreeNet Misclassification for Test Data

Class	N Cases	N Mis-Classed	Pct Error	Cost
BAD	1,004	205	20.42	205.00
GOOD	991	217	21.90	217.00

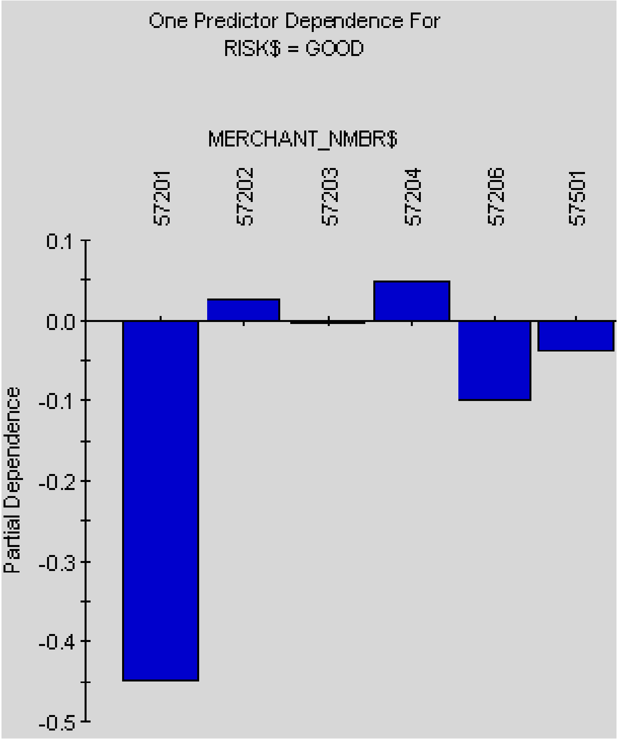
Table 5. Misclassification rate

The impact of *Email Domain* is extremely significant, but has different directions for different values. We can mention several distinctive segments of *Email Domain* values with different impacts on the probability of being a good customer:

1. Extremely positive impact
2. Modest positive impact
3. Practically no impact
4. Modest negative impact
5. Extremely negative impact

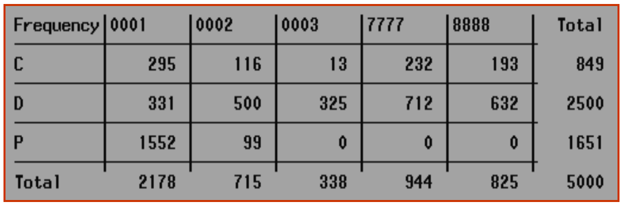


Graph 4. TreeNet Modeling: Impact of *Credit Model* predictor on log odds of Being a Good Customer, controlling for all other predictors (Second Model).



Graph 5. Impact of *Merchant Number* predictor on log odds of Being a Good Customer, controlling for all other predictors (Second Model)

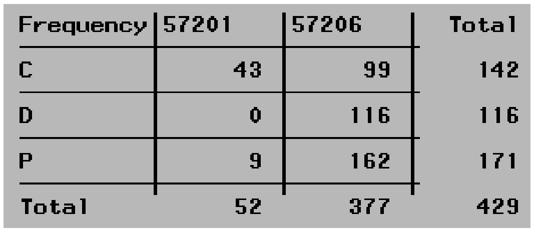
The only value 0001 of *Credit Model* is associated with a strong negative impact on the Probability of being a Good customer. The value of 0003 is associated with the strongest positive impact on Probability of being a Good customer.



Frequency	0001	0002	0003	7777	8888	Total
C	295	116	13	232	193	849
D	331	500	325	712	632	2500
P	1552	99	0	0	0	1651
Total	2178	715	338	944	825	5000

Table 6. Frequency of Loan *Status* by *Credit Model*

The left column of Table 6 depicts values of the *Loan Status* predictor (D designates a class of Good customers, and C and P designate a class of Bad customers), and the upper row of the table depicts the values of the *Credit Model* predictor (see section Data Structure for meaning of *Credit Model* values). The data supports the directions and strength (size) of impact on the probability of being a Good customer, induced by the model (Graph 5).



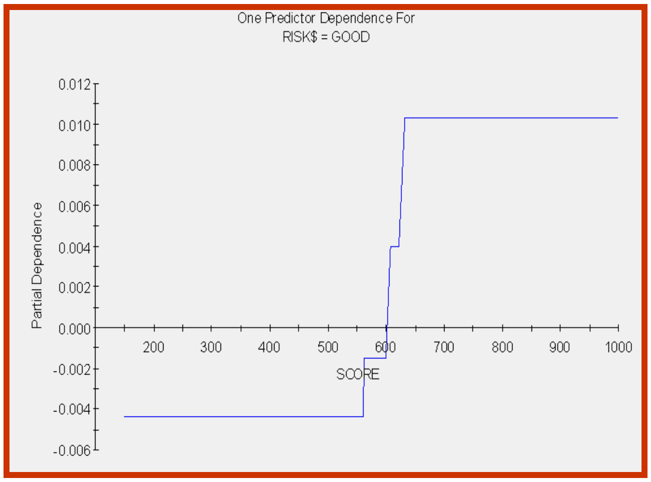
Frequency	57201	57206	Total
C	43	99	142
D	0	116	116
P	9	162	171
Total	52	377	429

Table 7. Frequency of Loan *Status* by *Merchant Number*

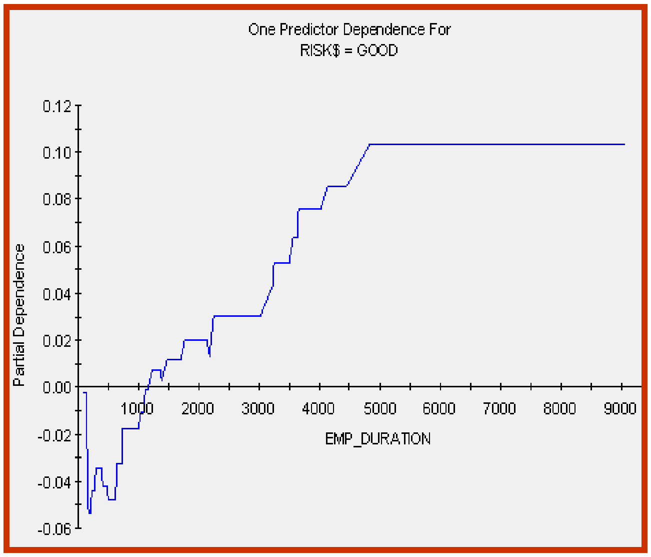
The left column of Table 7 depicts values of the *Loan Status* predictor (symbol D corresponds to a Good customer, and symbols C and P correspond to a Bad customer), and the upper row of the table depicts the values of the *Merchant Number* predictor. Table 7 pictured the frequency of customers that have the following values of *Merchant Number* variable: 57201 and 57206. We can observe that the majority of customers with these values of *Merchant Number* belong to the segment of Bad customers. Again, the model induced knowledge (Graph 5) and the data are in a good agreement.

Credit bureau score (*Score* predictor) has a binary impact on the probability of being a good customer: if the score is 600 and higher, then the probability jumps up, and if an applicant score is less than 600, then the probability jumps down, but this negative jump is not very large. In other words, for scores of less than 600 the impact on the probability of being a good customer is very modest.

If *Employment Duration* is less than 1,500 days, we can say that there is no strong impact on the probability of being a Good customer (we can treat the part of Graph 7 curve for *Employment Duration* is less than 1,500 days as a noise). The real impact of *Employment Duration* starts from 1,500 days, and is linear up to 5,000 days. Then the probability of being a Good customer has a diminishing returns effect when *Employment Duration* becomes greater than 5000 days.



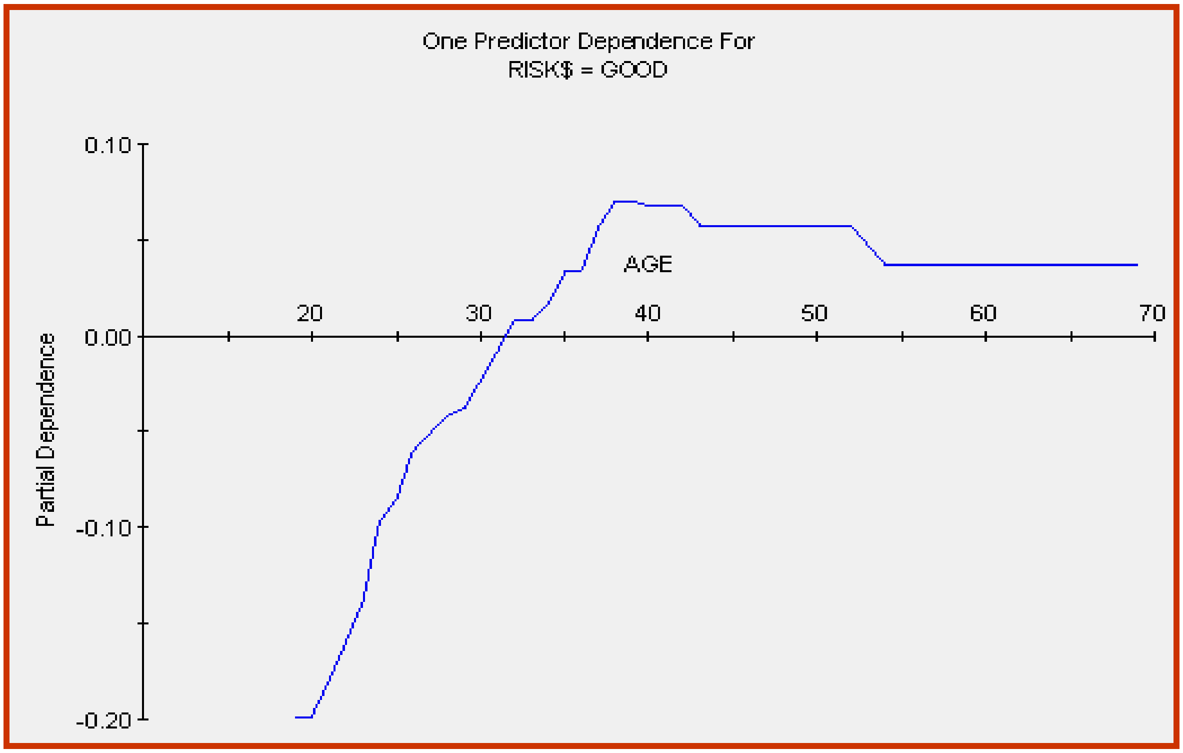
Graph 6. TreeNet Modeling: Impact of *Score* predictor on Probability of Being a Good Customer, controlling for all other predictors (Second Model)



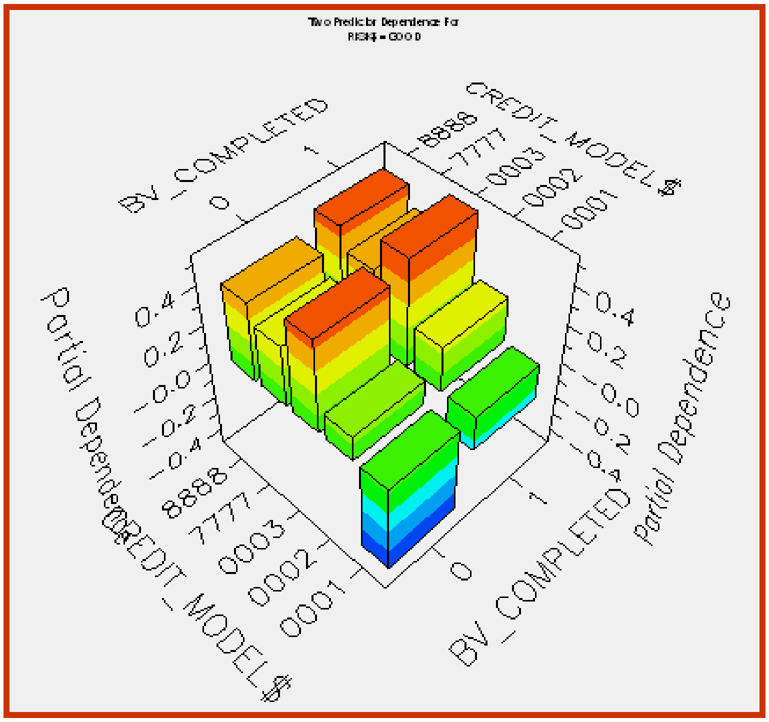
Graph 7. TreeNet Modeling: Impact of *Employment Duration* and *Score* on Probability of Being a Good Customer, controlling for all other predictors (Second Model)

It turned out that the best probability of being a Good customer occurs with applicants between ages 38 and 42 years. If the age is less than 32, then the impact is negative, and the younger an applicant the lower the probability of being a good customer. On the other hand, the strength of positive impact on the probability of being a Good customer goes down when age is increasing.

The vertical axis of Graph 9 maps log odds of the event Risk = Good. Two other axes are matched by values of the *BV Completed* and *Credit Model* predictors. The combination of *BV Completed* =1 and *Credit Model* = 7777 has the strongest positive impact on the probability of being a Good customer. On the other hand, the combination of *BV Completed* = 0 and *Credit Model* = 0001 has the strongest negative impact on the probability at hand.



Graph 8. TreeNet Modeling: Impact of *Age* on Probability of Being a Good Customer, controlling for all other predictors (SecondModel).



Graph 9. TreeNet Modeling: Impact of the interaction of *BV Completed* and *Credit Model* predictors on Probability of Being a Good Customer, controlling for all other predictors (Second Model)

4. GIS APPLICATION

In order to improve the quality of the data mining predictive models, it is useful to enrich SLC's data with additional region-level demographic, socioeconomic and housing variables that can be obtained from the US Bureau of the Census. These variables include

- median household income
- education
- median gross rent
- median house value, etc.

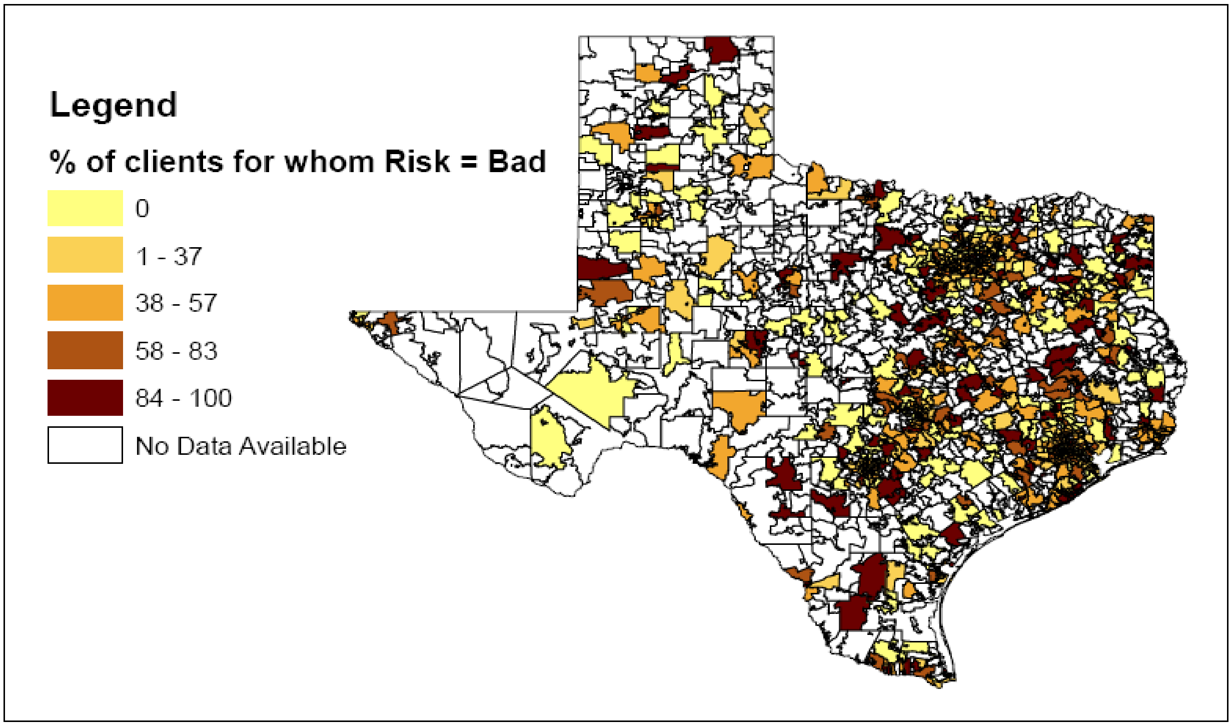
The variables can be obtained at different geographic levels, namely at the ZIP Code and the Census block levels.

Because the SLC data include ZIP code as one of the variables, it will be possible to merge the Zip level Census data to the SLC data directly.

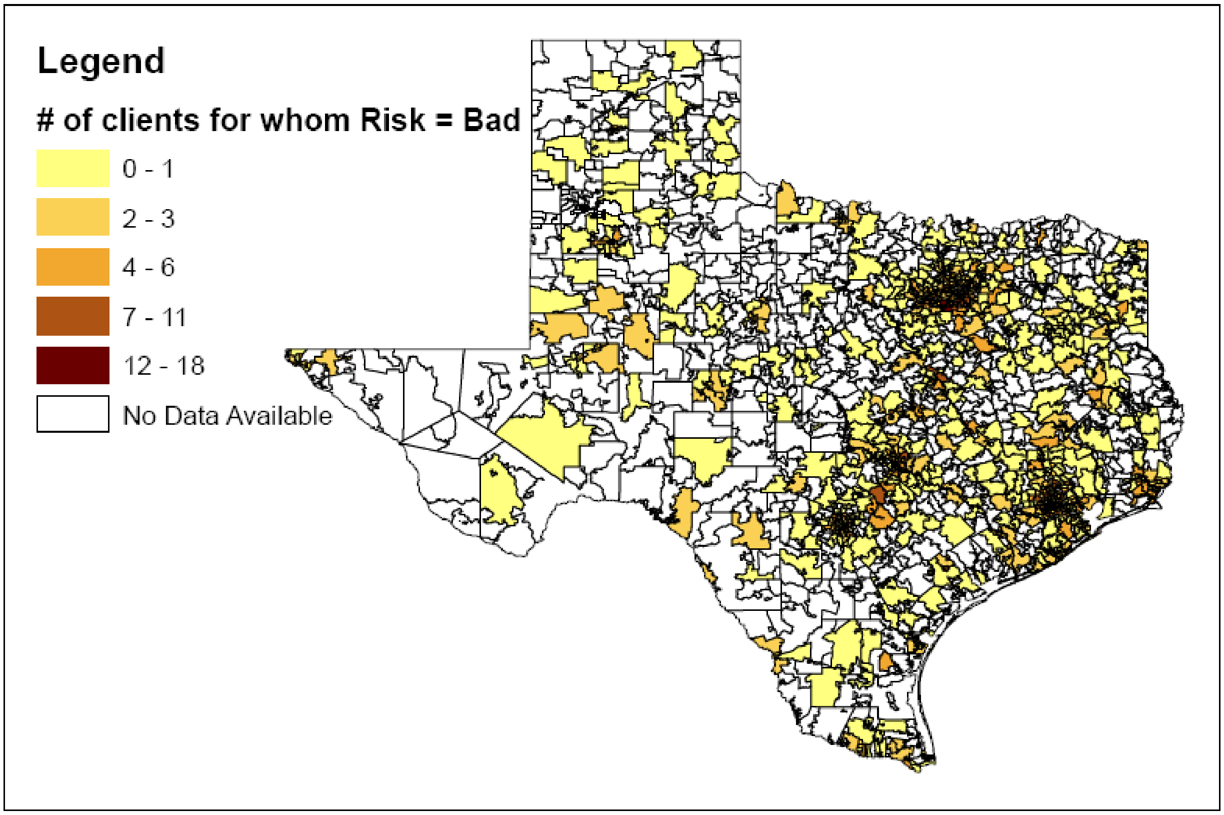
However, if customer address data are available, it will be advisable to obtain the Census data for smaller geographic regions (namely, Census blocks). Because Census blocks are in general much smaller than Zip codes, the Census estimates for these areas will be much more precise and much more applicable than for their ZIP code counterparts.

Using the Geographic Information Systems software, customer addresses can be geocoded (i.e. the latitude and longitude of the addresses can be determined, and the addresses can be mapped). Then, it will be possible to spatially match the addresses to their respective Census blocks (and Census block data). Demographic, socioeconomic, and housing data can then be obtained at the Census Block level. Although geocoding is a time intensive procedure, enriching the SLC data with the Census block level data will make the accuracy of the credit score even higher.

Employing dissimilar data mining tools, it is easy to determine which Census variables are crucial for customer risk assessment. When corresponding data become available, maps produced by the GIS will enable us to visually identify zip codes with many bad (high) risk customers and zip codes with many good (low) risk customers (Graphs 10 and 11).



Graph 10. Percent of Bad Customers in Each ZIP Code



Graph 11. Number of Bad Customers in Each ZIP Code

5. CONCLUSION

Knowledge generated by the models considered is in good compliance with the data and readily interpretable. The accuracy of stochastic gradient boosting models is superior. Stochastic gradient boosting is an appropriate tool for non-traditional scorecard development. Knowledge obtained from a stochastic gradient boosting model is a great asset for traditional scorecard development that can help to reduce both types of error and significantly improve ROI.

33. 6. ACKNOWLEDGMENTS

Our thanks to David Johnson (SLC) and Eugene Brusilovskiy (BISolutions) for the help and useful comments.

7. REFERENCES

[1] P. Brusilovskiy and D. Johnson 2008. Credit Risk Evaluation of Online Personal Loan Applicants: A Data Mining Approach, http://bisolutions.us/web/graphic/Credit-Risk-Evaluation-of-Online-Personal-Loan.pdf

[2] J. Friedman 1999. Greedy Function Approximation: A Gradient Boosting Machine

http://www.salford-systems.com/doc/GreedyFuncApproxSS.pdf

[3] J. Friedman 1999. Stochastic Gradient Boosting

http://www.salford-systems.com/doc/StochasticBoostingSS.pdf

[4] Dan Steinberg 2006. Overview of TreeNet Technology. Stochastic Gradient Boosting

http://perseo.dcaa.unam.mx/sistemas/doctos/TN_overview.pdf

[5] Boosting Trees for Regression and Classification, StatSoft Electronic Text Book

http://www.statsoft.com/textbook/stbootres.html

[6] Naeem Siddiqi 2005. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring (Wiley and SAS Business Series), Wiley, 208 p.

[7] Thomas, L.C., Edelman, D.B., Crook, J.N, 2002. Credit Scoring and its Applications, SIAM, 250 p.

[8] Matignon, R. 2007. Data Mining Using SAS Enterprise Miner. Wiley Publishing.

[9] Myatt, G. J. 2006. Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining. Wiley Publishing.

[10] Seo, J. and Gordish-Dressman, H. 2007. *Exploratory Data Analysis With Categorical Variables: An Improved Rank-by-Feature Framework and a Case Study*. International Journal of Human-Computer Interaction. Available online at http://www.informaworld.com/smpp/title~content=t775653655

Migration of Data Mining Preprocessing into the DBMS

Carlos Ordonez
University of Houston
Dept. of Computer Science
Houston, TX, USA

Javier García-García
UNAM[*]
Facultad de Ciencias
Mexico City, Mexico

Michael J. Rote
Teradata
Data Mining Solutions
San Diego, CA, USA

ABSTRACT

Nowadays there is a significant amount of data mining work performed outside the DBMS. This article discusses recommendations to push data mining analysis into the DBMS paying attention to data preprocessing (i.e. data cleaning, summarization and transformation), which tends to be the most time-consuming task in data mining projects. We present a discussion of practical issues and common solutions when transforming and preparing data sets with the SQL language for data mining purposes, based on experience from real-life projects. We then discuss general guidelines to create variables (features) for analysis. We introduce a simple prototype tool that translates statistical language programs into SQL, focusing on data manipulation statements. Based on experience from successful projects, we present actual time performance comparisons running SQL code inside the DBMS and outside running programs on a statistical package, obtained from data mining projects in a store, a bank and a phone company. We highlight which steps in data mining projects are much faster in the DBMS, compared to external servers or workstations. We discuss advantages, disadvantages and concerns from a practical standpoint based on users feedback. This article should be useful for data mining practitioners.

Categories and Subject Descriptors

H.2.8 [Database Management]: Systems—Relational Databases; H.2.8 [Database Management]: Database Applications—Data Mining

General Terms

Languages, Performance

Keywords

Data preprocessing, denormalization, SQL, translation

[*]Universidad Nacional Autonoma de México

DMCS'09, June 28, Paris, France.
Copyright 2009 ACM 978-1-60558-674-8 ...$5.00.

1. INTRODUCTION

In a modern IT environment transaction processing [7] and data warehousing [10] are managed by database systems. Analytics on the database are a different story. Despite the data mining [8, 10] functionality offered by the DBMS there exist many statistical tasks that are performed outside [11, 10]. This is due to the existence of sophisticated statistical tools and libraries [11], the lack of expertise of statistical analysts to write correct and efficient SQL queries, the limited set of statistical methods available offered by the DBMS and legacy code. In such environments users rely on the database to provide data, and then they write SQL queries to extract data joining several tables. Once data sets are exported they are summarized and transformed depending on the task at hand, but outside the DBMS. Finally, when the data set has the desired variables (features), statistical models are computed, interpreted and tuned. In general, models include both descriptive and predictive techniques, coming from machine learning [14] and statistics [11]. From all the tasks listed above preparing the data set for analysis is the most time consuming task [8] because it requires significant effort to transform typical normalized tables in the database into tabular data sets appropriate for analysis. Unfortunately, in such environments manipulating data sets outside the DBMS creates many data management issues: data sets must be recreated and re-exported every time there is a change, models need to be deployed inside the DBMS, different users may have inconsistent version of the same data set in their own computers, security is compromised. Therefore, we defend the idea of transforming data sets and computing models inside the DBMS, exploiting its extensive functionality. Our motivation to migrate statistical analysis into the DBMS, yields the following benefits. The DBMS server is generally a fast computer and thus results may be obtained sooner. The DBMS provides powerful querying capabilities through SQL and 3rd party tools. The DBMS has extensive data management functionality (maintaining referential integrity, transaction processing, fault-tolerance, security).

We assume the DBMS provides basic statistical and data mining functionality. That is, the DBMS is capable of performing common data mining tasks like regression [11], clustering [11], PCA [14] or association rules, among others. The commercial data mining tool used in our real-life projects is called Teradata Warehouse Miner (TWM) [25]. On the other hand, we assume the DBMS supports the SQL language, which is used to store and retrieve information from the DBMS [7, 10]. Based on experience from data mining projects we have become aware that statistical analysts have a hard time writing correct and efficient SQL code to extract data from the DBMS or to transform their data sets for the data mining task at hand. To overcome such issues, statistical analysts resort to statistical packages to manipulate and transform their data sets. Since most statistical packages provide a programming language, users end up creating long scripts mixing data transformation and modeling tasks together. In such scripts most of the programming effort is spent on transforming the data set: this is the main aspect studied in this article. We discuss practical issues and common solutions for data transformation tasks. Finally, we present evidence transforming data sets and building models on them is more efficient to do inside the DBMS than outside with an external statistical program.

This article presents a prototype tool capable of translating common statistical language (e.g. SAS [6]) programs into SQL [7]. The tool automates the task of translating statistical language scripts into equivalent SQL scripts, producing the same results. There are certain similarities and differences between both languages that make the problem interesting and challenging. The statistical language is an imperative language that manipulates data sets as tables, but not strictly as relational tables. The language includes syntactic constructs to specify arithmetic expressions, flow control and procedural calls. On the other hand, SQL is a set oriented language that also manipulates data sets as tables, but which allows specifying relationships among tables with primary/foreign keys. A set of tables is manipulated with join operations among their keys. Both languages identify data set columns with names and not with subscripts and both automatically scan all rows in a data set without the need of a loop construct. Query optimization is a fundamental issue in SQL.

The article is organized as follows. Section 2 discusses important practical issues when preparing and transforming data sets for statistical analysis as well as common solutions. Section 3 presents a translator of data transformation statements into SQL. Section 4 presents performance comparisons and users feedback from projects where statistical programs were successfully migrated. Related work is discussed in Section 5. Finally, Section 6 concludes the article.

2. PRACTICAL ISSUES PREPARING DATA SETS IN SQL

We present important practical issues to write and optimize SQL code for data transformation tasks. These issues have been collected from several projects helping users migrate statistical analysis performed in external statistical tools into the DBMS.

In general, the main objection from users to do so is to translate existing code. Commonly such code has existed for a long time (legacy programs), it is extensive (there exist many programs) and it has been debugged and tuned. Therefore, users are reluctant to rewrite it in a different language, given associated risks. A second complaint is that, in general, the DBMS provides elementary statistical functionality, compared to sophisticated statistical packages. Nevertheless, such gap has been shrinking over the years. As explained before, in a data mining environment most user time is spent on preparing data sets for analytic purposes. We discuss some of the most important database issues when tables are manipulated and transformed to prepare a data set for data mining or statistical analysis. We pay particular attention to query optimizations aspects [7]. These issues represent experience that has been used to improve and optimize SQL code generated by a data mining tool tightly integrated with the DBMS [25].

Throughout this section we present example of typical SQL queries. Some examples refer to retail (store sales) databases, whereas some others refer to a bank.

2.1 Main Issues

Summarization

Unfortunately, most data mining tasks require dimensions (variables) that are not readily available from the database. Such dimensions typically require computing aggregations at several granularity levels. This is because most columns required by statistical or machine learning techniques require measures (or metrics), which translate as sums or counts computed with SQL. Unfortunately, granularity levels are not hierarchical (like cubes or OLAP [7]) making the use of separate summary tables necessary (e.g. summarization by product or by customer, in a retail database). A straightforward optimization is to compute as many dimensions in the same statement exploiting the same group-by clause, when possible. In general, for a statistical analyst it is best to create as many variables (dimensions) as possible in order to isolate those that can help build a more accurate model. Then summarization tends to create tables with hundreds of columns, which make query processing slower. However, most state-of-the-art statistical and machine learning techniques are designed to perform variable (feature) selection [11, 14] and many of those columns end up being discarded.

A typical query to derive dimensions from a transaction table is as follows:

```
SELECT
  customer_id
 ,count(*) AS cntItems
 ,sum(salesAmt) AS totalSales
 ,sum(case when salesAmt<0 then 1 end)
  AS cntReturns
FROM sales
GROUP BY customer_id;
```

Denormalization

It is required to gather data from many tables and store data elements in one place. It is well-known that on-line transaction processing (OLTP) DBMSs update normalized tables. Normalization makes transaction processing faster and ACID [7] semantics are easier to ensure. Queries that retrieve a few records from normalized tables are relatively fast. On the other hand, analysis on the database requires precisely the opposite: a large set of records is required and such records gather information from many tables. Such processing typically involves complex queries involving joins and aggregations. Therefore, normalization works against efficiently building data sets for analytic purposes. One solution is to keep a few key denormalized tables from which specialized tables can be built. In general, such tables cannot be dynamically maintained because they involve join computation with large tables. Therefore, they are periodically recreated as a batch process.

```
SELECT
  customer_id
 ,customer_name
 ,product.product_id
 ,product_name
 ,department.department_id
 ,department_name
FROM sales
    JOIN product
      ON sales.product_id=product.product_id
    JOIN department
      ON product.department_id
         =department.department_id;
```

Time window

In general, for a data mining task it is necessary to select a set of records from one of the largest tables in the database based on a date range. In general, this selection requires a scan on the entire table, which is slow. When there is a secondary index based on date it is more efficient to select rows, but it is not always available. The basic issue is that such transaction table is much larger compared to other tables in the database. For instance, this time window defines a set of active customers or bank accounts that have recent activity. Common solutions to this problem include creating materialized views (avoiding join recomputation on large tables) and lean tables with primary keys of the object being analyzed (record, product, etc) to act as filter in further data preparation tasks.

```
SELECT
  customer_id
 ,product_id
 ,sales_amt
FROM sales
WHERE cast(salesDate AS char(10))>= '2009-01-01'
      and cast(salesDate AS char(10))< '2009-03-01';
```

Dependent SQL statements

A data transformation script is a long sequence of SELECT statements. Their dependencies are complex, although there exists a partial order defined by the order in which temporary tables and data sets for analysis are created. To debug SQL code it is a bad idea to create a single query with multiple query blocks. In other words, such SQL statements are not amenable to the query optimizer because they are separate, unless it can keep track of historic usage patterns of queries. A common solution is to create intermediate tables that can be shared by several statements. Those intermediate tables commonly have columns that can later be aggregated at the appropriate granularity levels.

```
CREATE TABLE X AS (
 SELECT
   A,B,C
 FROM
) WITH DATA;

SELECT *
FROM X JOIN T1 ...

SELECT *
FROM X JOIN T2 ...

..

SELECT *
FROM X JOIN TN ...
```

Computer resource usage

This aspect includes both disk and CPU usage, with the second one being a more valuable resource. This problem gets compounded by the fact that most data mining tasks work on the entire data set or large subsets from it. In an active database environment running data preparation tasks during peak usage hours can degrade performance since, generally speaking, large tables are read and large tables are created. Therefore, it is necessary to use workload management tools to optimize queries from several users together. In general, the solution is to give data preparation tasks a lower priority than the priority for queries from interactive users. On a longer term strategy, it is best to organize data mining projects around common data sets, but such goal is difficult to reach given the mathematical nature of analysis and the ever changing nature of variables (dimensions) in the data sets.

Views vs temporary tables

Views provide limited control on storage and indexing. It may be better to create temporary tables, especially when there are many primary keys used in summarization. Nevertheless, disk space usage grows fast and such tables/views need to be refreshed when new records are inserted or new variables (dimensions) are created.

Level of detail

Transaction tables generally have two or even more levels of detail, sharing some columns in their primary key. The typical example is store transaction table with individual items and the total count of items and total amount paid. This means that many times it is not possible to perform a statistical analysis only from one table. There may be unique pieces of information at each level. Therefore, such large tables need to be joined with each other and then aggregated at the appropriate granularity level, depending on the data mining task at hand. In general, such queries are optimized by indexing both tables on their common columns so that hash-joins can be used.

For instance, in a store there is typically one transaction table containing total amounts (sales, tax, discounts) and item counts, and another transaction detail (or line) table containing each individual item scanned at the register. For certain data mining analysis (market basket analysis the detailed purchase information may be required). On the other hand, in a bank there is one table with account summaries (current and by month) and another table with individual banking transactions (deposits, withdrawals, payments, balance inquiry).

Left outer joins for completeness

For analytic purposes it is always best to use as much data as possible. There are strong reasons for this. Statistical models are more reliable, it is easier to deal with missing information, skewed distributions, discover outliers and so on, when there is a large data set at hand. In a large database with tables coming from a normalized database being joined with tables used in the past for analytic purposes may involve joins with records whose foreign keys may not be found in some table. That is, natural joins may discard potentially useful records. The net effect of this issue is that the resulting data set does not include all potential objects (e.g. records, products). The solution is define a universe data set containing all objects gathered with union from all tables and then use such table as the fundamental table to perform outer joins. For simplicity and elegance, left outer joins are preferred. Then left outer joins are propagated everywhere in data preparation and completeness of records is guaranteed. In general such left outer joins have a "star" form on the joining conditions, where the primary key of the master table is left joined with the primary keys of the other tables, instead of joining them with chained conditions (FK of table T1 is joined with PK of table T2, FK of table T2 is joined with PK of T3, and so on).

```
SELECT
 ,record_id
 ,T1.A1
 ,T2.A2
 ..
 ,Tk.Ak
FROM T_UNIVERSE
   JOIN T1 ON T_UNIVERSE.record_id= T1.record_id
   JOIN T2 ON T_UNIVERSE.record_id= T2.record_id
   ..
   JOIN Tk ON T_UNIVERSE.record_id= Tk.record_id;
```

Filtering Records from Data Set

Selection of rows can be done at several stages, in different tables. Such filtering is done to discard outliers [18], to discard records with a significant missing information content (including referential integrity [22], to discard records whose potential contribution to the model provides no insight or sets of records whose characteristics deserve separate analysis. It is well known that pushing selection is the basic strategy to accelerate SPJ (select-project-join) queries [3], but it is not straightforward to apply into multiple queries. A common solution we have used is to perform as much filtering as possible on one data set. This makes code maintenance easier and the query optimizer is able to exploit filtering predicates as much as possible.

Statistics columns

Many times users build data sets with averages, which unfortunately are not distributive. Common examples are computing the mean and standard deviation of some numeric column. A simple solution is to keep sums and counts for every data set, from which it is easy to derive descriptive statistics. In particular, sufficient statistics [26, 2, 17, 19] prove useful for both simple statistics as well as sophisticated multidimensional models. This solution is represented by the sufficient statistics L, Q [17, 19], explained in more detail in Section 3.6. Another particularly useful optimization is to perform aggregations before joins, when data semantics allow it. This optimization has been studied in the database literature [3].

```
SELECT
 ,count(*) AS n
 ,sum(X1) AS L1
 ,sum(X2) AS L2
 ..
 ,sum(Xd) AS Ld
 ,sum(X1*X1) AS Q1
 ,sum(X2*X2) AS Q2
 ..
 ,sum(Xd*Xd) AS Qd
FROM DataSetX;
```

Multiple primary keys

Different sets of tables have different primary keys. This basically, means such tables are not compatible with each other to perform further summarization. The key issue is that at some point large tables with different primary keys must be joined and summarized. Join operations will be slow because indexing involves foreign keys with large cardinalities. Two solutions are common: creating a secondary index on the alternative primary key of the largest table, or creating a denormalized table having both primary keys in order to enable fast join processing.

For instance, consider a data mining project in a bank that requires analysis by customer id, but also account id. One customer may have multiple accounts. An account may have multiple account holders. Joining and manipulating such tables is challenging given their sizes.

Model deployment

Even though many models are built outside the DBMS with statistical packages and data mining tools, in the end the model must be applied in the database [19]. When volumes of data are not large it is feasible to perform model deployment outside: exporting data sets, applying the model and building reports can be done in no more than a few minutes. However, as data volume increases exporting data from the DBMS becomes a bottleneck [19]. This problem gets compounded with results interpretation when it is necessary to relate statistical numbers back to the original tables in the database. Therefore, it is common to build models outside, frequently based on samples, and then once an acceptable model is obtained, then it is imported back into the DBMS. Nowadays, model deployment basically happens in two ways: using SQL queries if the mathematical computations are relatively simple or with UDFs [19], if the computations are more sophisticated. In most cases, such scoring process can work in a single table scan, providing good performance.

2.2 Lessons Learned: Most Common Queries

Data transformation is a time consuming project, but the statistical language syntactic constructs and its comprehensive library of functions make such task easier. In general, users think writing data transformations in SQL is not easy. Despite the abundance of data mining tools users need to understand the basics of query processing. The most common queries needed to create data sets for data mining are (we omit transformations that are typically applied on an already built data set like coding, logarithms, normalization, and so on):

- Left outer joins, which are useful to build a "universal" data set containing all records (observations) with columns from all potential tables. In general, natural joins filter out records with invalid keys which may contain valuable information.

- Aggregations, generally with sums and counts, to build a data set with new columns.

- Denormalization, to gather columns scattered in several tables together.

3. TRANSLATING DATA TRANSFORMATIONS INTO SQL

This section explains the translator, summarizes similarities and differences between the statistical language and SQL and presents common sufficient statistics for several models.

3.1 Definitions

The goal of data manipulation is to build a data set. Let $X = \{x_1, \ldots, x_n\}$ be the data set with n points, where each point has d attributes (variables, features), where each of them can be numeric or categorical. The statistical language is a high-level programming language, based on: data set, observation and variable. We assume SQL is well understood, but we give a summary. In SQL the equivalent terms are table, row and column. A table contains a set of rows having the same columns. The order of rows is immaterial from a semantic point of view. A set of tables is interrelated by means of primary/foreign key relationships.

3.2 General Framework

A statistical language program produces data sets which are basically tables with a set of columns. Columns can be of numeric, string or date data types. On the other hand, SQL manipulates data as tables, with the additional constraint that they must have a primary key. In general, data sets in the statistical language are sequentially manipulated in main memory, loading a number of rows. On the other hand, in SQL relational operations receive tables as input and produce one table as output. There exist two main kinds of statements: (1) Data manipulation (data set transformation). (2) Subroutine calls (functions, procedures, methods).

Translating subroutine calls requires having stack-based calling mechanisms in SQL, which are not generally available, or if they are available parameters are not standardized. That is, the translator produces a "flat" script in SQL.

3.3 Data Manipulation

3.3.1 Importing Data

Importing is a relatively easy procedure which is used mostly to integrate external data sources into the statistical analysis tool. Despite its simplicity it is important to carefully analyze importing statements because they may involve data not stored in the database. Importing data comes in the form of a statement specifying a data set with several columns and an input file.

3.3.2 Arithmetic Equations

Columns in the data set are treated as variables. An assignment creates a new variable or assigns a value to an existing one. If the variable does not exist the type is inferred from the expression. In SQL there is no assignment expression. Therefore, the assignment expression must be converted into SELECT statements with one column to receive the result of each expression.

A sequence of variable assignments creates or updates variables with arithmetic expressions. The first assignment is assumed to use an expression with all instantiated variables. The sequence of assignment statements assumes a variable cannot be used before it has a value. Given the dynamic nature of the sequence of expressions it is necessary for the SQL run-time evaluation algorithm to determine the type of the resulting column. The alternative approach, defining a DDL and then the SQL with expressions would require doing extensive syntactic and semantic analysis when the expression is parsed. Most math functions have one argument and they can be easily translated using a dictionary. String functions are more difficult to translate because besides having different names they may have different arguments and some of them do not have an equivalent in SQL.

Let C be the set of original variables and let V be the set of variables created or updated by assignment. An initial pass on all assigned variables is needed to determine which columns are overwritten by computing $C \cap V$. Each column that belongs to $C \cap V$ is removed from C. Then it is unselected from the original list of variables. Assume then that the input columns become a subset of the original columns: $\mathsf{F} = F_1, F_2, \ldots, F_p$, where $\mathsf{F} \subseteq C$. Then the sequence of expressions can be translated as F, followed by the arithmetic expressions assigned to each variable.

It is important to observe the data types are dynamically inferred by SQL at run-time and that the table is defined as multiset. Performing a static analysis would require a more sophisticated mechanism to infer data types storing variables in a symbol table like a traditional compiler.

3.3.3 IF-THEN and WHERE

A high-level programming language provides great flexibility in controlling assignments. This is more restricted in SQL because only one column can be manipulated in a term. We consider three cases: (1) Chained IF-THEN statements with one assignment per condition; (2) Generalized IF-THEN-DO with IF-THEN-DO nesting and two or more assignments per condition. (3) A WHERE clause. A chained IF-THEN statement gets translated into an SQL CASE statement where each IF condition gets transformed into a WHEN clause. It is important to watch out for new columns when new variables are created. The IF-THEN-DO construct is more complex than the previous case for several reasons: More than one assignment can be done in the IF body; IF statements can be nested several levels. There may be loops with array variables. This case will be analyzed separately. The system uses a stack [1] to keep track of conditions given an arbitrary number of levels of nesting. For every nested IF statement boolean expressions are pushed into the stack. For each assignment each additional boolean expression are popped from the stack and are concatenated using a logical "AND" operator to form an SQL WHEN expression. In other words, nested IF-THEN statements are flattened into "WHEN" substatements in a CASE statement. The WHERE clause translates without changes into SQL. Comparison for numbers and strings use same operators. However, date comparisons are different and therefore special routines. Comparison operators have similar syntax in both languages, whose translation requires a simple equivalence table. Negation (NOT), parenthesis and strings, require similar translation (compilation) techniques.

3.3.4 *Looping Constructs*

In the statistical language there may be arrays used to manipulate variables with subscripts. SQL does not provide arrays, but they can be simulated by generating columns whose name has the subscript appended. A FOR loop is straightforward to translate when the subscript range can be determined at translation time; the most common case is a loop where the bounds are static. When an array is indexed with a subscript that has a complex expression (e.g. $a(i*10-j)$) then the translation is more difficult because the target column name cannot be known at translation time.

3.3.5 *Combining Different Data Sets*

We focus on two basic operations: (1) Union of data sets; (2) Merging data sets.

Union: This is the case when the user wants to compute the union of data sets where most or all the variables are equal in each data set, but observations are different. Assume D_i already has observations and variables. The main issue here is that such statement does not guarantee all data sets have the same variables. Therefore, the translation must make sure the result data set includes all variables from all data sets setting to null those variables that are not present for a particular data set. First, we compute the union of all variables. Let p be the cardinality of $R.B = \{B_1, \ldots, B_p\}$. For each data set we need to compute the set of variables from R not present: $U.B - D_i.A$. A total order must be imposed on columns so that each position correspond to one column from U. Such order can be given by the order of appearance of variables in D_i. At the beginning variables are those from $R.A = D_1.A$. If there are new variables, not included in R.A then they are added to R.A. This process gets repeated until D_m is processed. Then we just need to insert nulls in the corresponding variables when the result table is populated. The ith "SELECT" statement has p terms out of which n_i are taken from D_i and the rest are null.

Merging: This is the case where two data sets have a common key, some of the remaining columns are the same and the rest are different. If there are common columns among both data sets columns from the second one take precedence and overwrite the columns from the first data set. In contrast, SQL requires the user to rename columns with a new alias. In general, one of the data sets must be sorted by the matching variables. Let the result columns of M be $B_1, \ldots, B_p$. The dat sets D_1 and D_2 both contain the matching columns $A_1, A_2, \ldots, A_k$. A filtering process must be performed to detect common non-key columns. If there are common columns the column from D_2 takes precedence. The process is similar to the filtering process followed for arithmetic expressions or the union of data sets. This translates into SQL as a full outer join to include unmatched rows from D_1 and D_2.

3.4 Translating Subroutine Procedural Calls

We discuss translation of common procedural calls to equivalent calls using the data mining tool and then we discuss translation of embedded SQL. Translating procedural calls can be done with several alternative mechanisms: A first alternative is to call a data mining tool application programming interface (API) to automatically generate SQL code for univariate statistics or data mining algorithms. This issue is that several intermediate mathematical computations are left outside the final SQL script. Therefore, this alternative does not produce self-contained scripts. A second alternative is to replicate automatically generated SQL code generation in the translator. The benefit of this approach is that we can generate SQL scripts that can manage an entire data mining process, going from basic data transformations and descriptive statistics to actually tuning and testing models. The last alternative is to leave the translation open for manual coding into SQL, including the statistical language code in comments. We include this alternative for completeness because there may be specific statements and mathematical manipulations that cannot be directly translated. User intervention is required in this case.

Interestingly enough, the translation process may be faced with the task of handling embedded SQL statements. Such SQL is in general used to extract data from different DBMSs. We now explain the translation process: The most straightforward translation is a PROC SQL if the SQL corresponds to the same DBMS (in our case Teradata). Care must be taken in creating appropriate temporary tables whose names do not conflict with those used in the code. If data elements are being extracted from a different DBMS then the translation process becomes more complicated: it is necessary to check each referenced table exists in the target DBMS and SQL syntax may have differences. Therefore, it is best to manually verify the translation. If the SQL corresponds to some other DBMS there may be the possibility of finding non-ANSI features. This aspect may be difficult if queries are complex and have several nesting levels combining joins and aggregations. Many script versions need to be maintained.

3.5 Similarities and Differences

In Section 3 we explained how to translate code in the statistical language into SQL statements. Here we provide a summarized description of common features of both languages and where they differ.

Language Similarities

We present similarities going from straightforward to most important. In the statistical language there is no explicit instruction to scan and process the data set by observation: that happens automatically. In SQL the behavior is similar because there is no need to create a loop to process each row. Any SQL statement automatically processes the entire table. However, in the DBMS sometimes it is necessary to perform computations without using SQL to exploit arrays. Then regular looping constructs are required. Processing happens in a sequential fashion for each observation. In the statistical language each variable is created or updated as new assignment expressions are given for each row. In SQL a column is created when a new term in a "SELECT" statement is found. A column cannot be referenced if it has not been previously created with the "AS" keyword or it is projected from some table. Broadly speaking each new procedural call (PROC) reads or creates a new data set. Therefore, this aspect can be taken as a guideline to create temporary tables to store intermediate results.

Language Differences

We discuss differences going from straightforward to those we consider most challenging when making syntactic and semantic analysis.

Macros are different in both languages, being represented by stored procedures in SQL. In the statistical language a data set name or variable name can start with underscore, which may cause conflicts in translation. This can be solved by enclosing the table name or column name in SQL in quotes (e.g. " name"). In the statistical language a missing value is indicated with a dot '.', whereas in SQL it is indicated with the keyword "NULL". A missing value can compared with the equality symbol '=' like any number, whereas SQL requires specialized syntax using the "IS NULL" phrase. Since a number can start with '.' a read-ahead scanner needs to determine if there is a digit after the dot or not. However, equations involving with missing values, in general, return a missing value as well. Both languages have similar semantics for missing information. In the statistical language variable name conflicts are solved in favor of the last reference. In SQL that conflict must be solved by qualifying ambiguous columns or by explicitly removing references to columns with the same name. To store results a table cannot contain columns with the same name. Therefore, for practical purposes duplicate column names must be removed during the SQL code generation. In the statistical language sorting procedures are needed to merge data sets. Sorting is not needed in SQL to join data sets. In fact, there is no pre-defined order among rows. Merging works in a different manner to joins. New variables are added to a given data set for further processing. A data set is always manipulated in memory, but new variables may not necessarily be saved to disk. In SQL a new table must be created for each processing stage. Tables are stored on disk. Some tables are created in permanent storage, whereas the rest of tables have to be created in temporary storage. The statistical language allows creating a data set with up to hundreds of thousands of variables. whereas SQL allows table with up to thousands of columns. This limitation can be solved by counting variables and creating a new table every one thousand variables; this will vertically partition the result table. Joins are required to reference columns from different partitions. The DBMS performs more careful retrieval of rows into main memory processing them in blocks, whereas the statistical language performs most of the processing in main memory based on a single table at one time.

3.6 Sufficient Statistics

We now explain fundamental statistics computed on the data set obtained from the data transformation process introduced in Section 3. These statistics benefit a broad class of statistical and machine learning techniques. Their computation can be considered an intermediate step between preparing data sets and computing statistical models. In general, most statistical data mining tools provide functionality to derive these statistics. In the literature the following matrices are called sufficient statistics [2, 11, 19] because they are enough to substitute the data set being analyzed in mathematical equations. Therefore, it is advantageous they are available for the data set to be analyzed.

Consider the multidimensional (multivariate) data set defined in Section 3.1: $X = \{x_1, \ldots, x_n\}$ with n points, where each point has d dimensions. Some of the matrix manipulations we are about to introduce are well-known in statistics, but we exploit them in a database context. We introduce the following two matrices that are fundamental and common for all the techniques described above. Let L be the linear sum of points, in the sense that each point is taken at power 1. L is a $d \times 1$ matrix, shown below with sum and column-vector notation. $L = \sum_{i=1}^{n} x_i$. Let Q be the quadratic sum of points, in the sense that each point is squared with a cross-product. Q is $d \times d$. $Q = XX^T = \sum_{i=1}^{n} x_i x_i^T$.

The most important property about L and Q is that they are much smaller than X, when n is large (i.e. $d << n$). However, L and Q summarize a lot of properties about X that can be exploited by statistical techniques. In other words, L and Q can be exploited to rewrite equations so that they do not refer to X, which is a large matrix. Techniques that directly benefit from these summary matrices include correlation analysis linear regression [11], principal component analysis [11, 10], factor analysis [11] and clustering [14]. These statistics also partially benefit decision trees [14] and logistic regression [11].

Since SQL does not have general support for arrays these matrices are stored as tables using dimension subscripts as keys. Summary matrices can be efficiently computed in two ways: using SQL queries or using UDFs [19]. SQL queries allow more flexibility, are portable, but incur on higher overhead. On the other hand, UDFs are faster, but they depend on the DBMS architecture and therefore may have specific limitations such as memory size and parameter passing. Having an automated way to compute summary matrices inside the DBMS simplifies the translation process.

4. EXPERIENCE FROM REAL PROJECTS

This section presents a summary of performance comparisons and main outcomes from migrating actual data mining projects into the DBMS. This discussion is a summary of representative successful projects. We first discuss a typical data warehousing environment; this section can be skipped by a reader familiar with the subject. Second, we present a summary of the data mining projects presenting their practical application and the statistical and data mining techniques used. Third, we present time measurements taken from actual projects at each organization, running data mining software on powerful database servers. We conclude with a summary of the main advantages and accomplishments for each project, as well as the main objections or concerns against migrating statistical code into the DBMS.

4.1 Data Warehousing Environment

The environment was a data warehouse, where several databases were already integrated into a large enterprise-wide database. The database server was surrounded by specialized servers performing OLAP and statistical analysis. One of those servers was a statistical server with a fast network connection to the database server.

First of all, an entire set of statistical language programs were translated into SQL using Teradata data mining program, the translator tool and customized SQL code. Second, in every case the data sets were verified to have the same contents in the statistical language and SQL. In most cases, the numeric output from statistical and machine learning models was the same, but sometimes there were slight numeric differences, given variations in algorithmic improvements and advanced parameters (e.g. epsilon for convergence, step-wise regression procedures, pruning method in decision tree and so on).

4.2 Organizations and Data Mining Projects

We now give a brief discussion about the organizations where the statistical code migration took place. We also discuss the specific type of data mining techniques used in each case. Due to privacy concerns we omit discussion of specific information about each organization, their databases and the hardware configuration of their DBMS servers. We can mention all companies had large data warehouses managed by an SMP (Symmetric Multi-Processing) Teradata server having a 32-bit CPU with 4GB of memory on each node and several terabytes of storage. Our projects were conducted on their production systems, concurrently with other users (analysts, managers, DBAs, and so on).

The first organization was an insurance company. The data mining goal involved segmenting customers into tiers according to their profitability based on demographic data, billing information and claims. The statistical techniques used to determine segments involved histograms and clustering. The final data set had about $n = 300$k records and $d = 25$ variables. There were four segments, categorizing customers from best to worst.

The second organization was a cellular telephone service provider. The data mining task involved predicting which customers were likely to upgrade their call service package or purchase a new handset. The default technique was logistic regression [11] with stepwise procedure for variable selection. The data set used for scoring had about $n = 10$M records and $d = 120$ variables. The predicted variable was binary.

The third organization was an Internet Service Provider (ISP). The predictive task was to detect which customers were likely to disconnect service within a time window of a few months, based on their demographic data, billing information and service usage. The statistical techniques used in this case were decision trees and logistic regression and the predicted variable was binary. The final data set had $n = 3.5$M records and $d = 50$ variables.

4.3 Performance Comparison

We focus on comparing performance doing statistical analysis inside the DBMS using SQL (with Teradata data mining program) and outside using the statistical server (with existing programs developed by each organization). The comparison is not fair because the DBMS server was in general a powerful parallel computer and the statistical server was a smaller computer. However, the comparison represents a typical enterprise environment where the most powerful computer is precisely the DBMS server.

We now describe the computers in more detail. The DBMS server was, in general, a parallel multiprocessor computer with a large number of CPUs, ample memory per CPU and several terabytes of parallel disk storage in high performance RAID configurations. On the other hand, the statistical server was generally a smaller computer with less than 500 GB of disk space with ample memory space. Statistical and data mining analysis inside the DBMS was performed only with SQL. In general, a workstation connected to each server with appropriate client utilities. The connection to the DBMS was done with ODBC. All time measurements discussed herein were taken on 32-bit CPUs over the course of several years. Therefore, they cannot be compared with each other and they should only be used to understand performance gains within the same organization.

Table 1: Comparing time performance between statistical package and DBMS (time in minutes).

Task	Statistical package (outside DBMS)	DBMS (inside)
Build model:		
Segmentation	2	1
Predict propensity	38	8
Predict churn	120	20
Score data set:		
Segmentation	5	1
Predict propensity	150	2
Predict churn	10	1

Table 2: Time to compute linear models inside the DBMS and time to export X with ODBC (secs).

$n \times 1000$	d	SQL/UDF	ODBC
100	8	4	168
100	16	5	311
100	32	6	615
100	64	8	1204
1000	8	40	1690
1000	16	51	3112
1000	32	62	6160
1000	64	78	12010

We discuss tables from the database in more detail. There were several input tables coming from a large normalized database that were transformed and denormalized to build data sets used by statistical or machine learning techniques. In short, the input were tables and the output were tables as well. No data sets were exported in this case: all processing happened inside the DBMS. On the other hand, analysis on the statistical server relied on SQL queries to extract data from the DBMS, transform the data to produce data sets in the statistical server and then building models or scoring data sets based on a model. In general, data extraction from the DBMS was performed using the fast utilities which exported data records in blocks. Clearly, there exists a bottleneck when exporting data from the DBMS to the statistical server.

Table 1 compares performance between both alternatives: inside and outside the DBMS. We distinguish two phases in each project: building the model and scoring (deploying) the model on large data sets. The times shown in Table 1 include the time to transform the data set with joins and aggregations and the time to compute or apply the actual model. As we can see the DBMS is significantly faster. We must mention that to build the predictive models both approaches exploited samples from a large data set. Then the models were tuned with further samples. To score data sets the gap is wider, highlighting the efficiency of SQL to compute joins and aggregations to build the data set and then compute statistical equations on the data set. In general, the main reason the statistical server was slower was the time to export data from the DBMS and then a secondary reason was its more limited computing power.

Table 3: Project Outcomes.

Outcome	Insurance	Phone	ISP
Advantages:			
Decrease data movement	X		
Prepare data sets more easily	X	X	X
Build models faster	X		
Score data sets faster	X	X	X
Increase security	X	X	
Improve data management		X	
Objections:			
Traditional progr. lang.		X	X
Sampling on large data sets	X	X	
Lack statistical techniques		X	

Table 2 compares time performance to compute a linear model inside the DBMS and the time to export the data set with the ODBC interface. The DBMS runs on a Windows Server with 3.2 GHz, 4 GB on memory and 1 TB on disk, and represents a small database server. The linear models include PCA and linear regression, which can be derived from the correlation matrix of X in a single table scan using SQL and UDFs [19]. Clearly, ODBC is a bottleneck to analyze X outside the DBMS regardless of how fast the statistical package is. Exporting small sample of X may be feasible, but analyzing a large data set without sampling is much faster to do inside the DBMS.

4.4 Data Mining Users Feedback

We summarize main advantages of projects migrated into the DBMS as well as objections from users to do so. Table 3 contains a summary of outcomes. As we can see performance to score data sets and transforming data sets are positive outcomes in every case. Building the models faster turned out not be as important because users relied on sampling to build models and several samples were collected to tune and test models. Since all databases and servers were within the same firewall security was not a major concern. In general, improving data management was not seen as major concern because there existed a data warehouse, but users acknowledge a "distributed" analytic environment could be a potential management issue. We now summarize the main objections, despite the advantages discussed above. We exclude cost as a decisive factor to preserve anonymity of users opinion and give an unbiased discussion. First, many users preferred a traditional programming language like Java or C++ instead of a set-oriented language like SQL. Second, some specialized techniques are not available in the DBMS due to their mathematical complexity; relevant examples include Support Vector Machines, Non-linear regression and time series models. Finally, sampling is a standard mechanism to analyze large data sets.

5. RELATED WORK

There exist many proposals that extend SQL with data mining functionality. Teradata SQL, like other DBMSs, provides advanced aggregate functions to compute linear regression and correlation, but it only does it for two dimensions. Most proposals add syntax to SQL and optimize queries using the proposed extensions. UDFs implementing common vector operations are proposed in [21], which shows UDFs are as efficient as automatically generated SQL queries with arithmetic expressions, proves queries calling scalar UDFs are significantly more efficient than equivalent queries using SQL aggregations and shows scalar UDFs are I/O bound. SQL extensions to define, query and deploy data mining models are proposed in [15]; such extensions provide a friendly language interface to manage data mining models. This proposal focuses on managing models rather than computing them and therefore such extensions are complementary to our UDFs. Query optimization techniques and a simple SQL syntax extension to compute multidimensional histograms are proposed in [12], where a multiple grouping clause is optimized. Computation of sufficient statistics for classification in a relational DBMS is proposed in [9]. Developing data mining algorithms, rather than statistical techniques, using SQL has received moderate attention. Some important approaches include [13, 23] to mine association rules, [20, 18] to cluster data sets using SQL queries, [20, 17] to cluster data sets using SQL queries and [24] to define primitives for decision trees. Sufficient statistics have been generalized and implemented as a primitive function using UDFs benefiting several statistical techniques [19]; this work explains the computation and application of summary matrices in detail for correlation, linear regression, PCA and clustering.

Some related work on exploiting SQL for data manipulation tasks includes the following. Data mining primitive operators are proposed in [4], including an operator to pivot a table and another one for sampling, useful to build data sets. The pivot/unpivot operators are extremely useful to transpose and transform data sets for data mining and OLAP tasks [5], but they have not been standardized. Horizontal aggregations were proposed to create tabular data sets [16], as required by statistical and machine learning techniques, combining pivoting and aggregation in one function. For the most part research work on preparing data sets for analytic purposes in a relational DBMS remains scarce. To the best of our knowledge there has not been research work dealing with the migration of data mining data preparation into a relational DBMS. Also, even though the ideas behind the translator are simple, they illustrate the importance of automating SQL code generation to prepare data sets for analysis.

6. CONCLUSIONS

We presented our experience on the migration of statistical analysis into a DBMS, focusing on data preprocessing (cleaning, transformation, summarization), which is in general the most time consuming, not well planned and error-prone task in a data mining project. Summarization generally has to be done at different granularity levels and such levels are generally not hierarchical. Rows are selected based on a time window, which requires indexes on date columns. Row selection (filtering) with complex predicates happens on many tables, making code maintenance and query optimization difficult. To improve performance it is necessary to create temporary denormalized tables with summarized data. In general, it is necessary to create a "universe" data set to define left outer joins. Model deployment requires importing models as SQL queries or UDFs to deploy a model on large data sets. We also explained how to compute sufficient statistics on a data set, that benefit a broad class

of data mining and machine learning techniques,

including correlation analysis, clustering, principal component analysis and linear regression. We presented a performance comparison and a summary of main advantages when migrating statistical programs into the DBMS by translating them into optimized SQL code. Transforming and scoring data sets is much faster inside the DBMS, whereas building a model is also faster, but less significant because sampling can help analyzing large data sets. In general, data transformation and analysis are faster inside the DBMS and users can enjoy the extensive capabilities of the DBMS (querying, recovery, security and concurrency control).

We presented a prototype tool to translate statistical scripts into SQL, to automate and accelerate the migration of data preparation, which is the most time consuming phase in a data mining project. The tool main goal is to generate SQL code that produces data sets with the same content as those generated by the statistical language: such data sets become the input for statistical or machine learning techniques (a so-called analytical data set).

6. REFERENCES

[1] A. Aho, R. Sethi, and J.D. Ullman. Compilers: Principles, Techniques and Tools. Addison-Wesley, 1986.

[2] P. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In ACM KDD Conference, pages 9–15, 1998.

[3] S. Chaudhuri. An overview of query optimization in relational systems. In ACM PODS Conference, pages 84–93, 1998.

[4] J. Clear, D. Dunn, B. Harvey, M.L. Heytens, and P. Lohman. Non-stop SQL/MX primitives for knowledge discovery. In ACM KDD Conference, pages 425–429, 1999.

[5] C. Cunningham, G. Graefe, and C.A. Galindo-Legaria. PIVOT and UNPIVOT: Optimization and execution strategies in an rdbms. In VLDB Conference, pages 998–1009, 2004.

[6] L.D. Delwiche and S.J. Slaughter. The SAS little book: a primer. SAS, 4th edition, 2003.

[7] R. Elmasri and S. B. Navathe. Fundamentals of Database Systems. Addison/Wesley, Redwood City, California, 3rd edition, 2000.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11):27–34, November 1996.

[9] G. Graefe, U. Fayyad, and S. Chaudhuri. On the efficient gathering of sufficient statistics for classification from large SQL databases. In ACM KDD Conference, pages 204–208, 1998.

[10] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco, 1st edition, 2001.

[11] T. Hastie, R. Tibshirani, and J.H. Friedman. The Elements of Statistical Learning. Springer, New York, 1st edition, 2001.

[12] A. Hinneburg, D. Habich, and W. Lehner. Combi-operator-database support for data mining applications. In VLDB Conference, pages 429–439, 2003.

[13] R. Meo, G. Psaila, and S. Ceri. An extension to SQL for mining association rules. Data Min. Knowl. Discov, 2(2):195–224, 1998.

[14] T.M. Mitchell. Machine Learning. Mac-Graw Hill, New York, 1997.

[15] A. Netz, S. Chaudhuri, U. Fayyad, and J. Berhardt. Integrating data mining with SQL databases: OLE DB for data mining. In IEEE ICDE Conference, 2001. [16] C. Ordonez. Horizontal aggregations for building tabular data sets. In ACM DMKD Workshop, pages 35–42, 2004.

[17] C. Ordonez. Programming the K-means clustering algorithm in SQL. In ACM KDD Conference, pages 823–828, 2004.

[18] C. Ordonez. Integrating K-means clustering with a relational DBMS using SQL. IEEE Transactions on Knowledge and Data Engineering (TKDE), 18(2):188–201, 2006.

[19] C. Ordonez. Building statistical models and scoring with UDFs. In ACM SIGMOD Conference, pages 1005–1016, 2007.

[20] C. Ordonez and P. Cereghini. SQLEM: Fast clustering in SQL using the EM algorithm. In ACM SIGMOD Conference, pages 559–570, 2000.

[21] C. Ordonez and J. García-García. Vector and matrix operations programmed with UDFs in a relational DBMS. In ACM CIKM Conference, pages 503–512, 2006.

[22] C. Ordonez and J. García-García. Referential integrity quality metrics. Decision Support Systems Journal, 44(2):495–508, 2008.

[23] S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: alternatives and implications. In ACM SIGMOD, pages 343–354, 1998.

[24] K. Sattler and O. Dunemann. SQL database primitives for decision tree classifiers. In ACM CIKM Conference, pages 379–386, 2001.

[25] Teradata. Teradata Warehouse Miner Release Definition Release 5.1. Teradata (NCR), 2006.

[26] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: An efficient data clustering method for very large databases. In ACM SIGMOD Conference, pages 103–114, 1996.

Estimating potential customer value using customer data

Thierry Vallaud and Daniel Larose

Abstraact

This study outlines a method of determining individual customer potential, based solely on data present in the customer database: descriptive information and transaction records. We define *potential* as the incremental turnover that any particular company could do with their present customers.In order to successfully calculate this potential in a large database with multiple variables, we propose grouping together customers who "look like each other" (known as clones), by means of an appropriate clustering technique: Kohonen Networks. This method is applied to actual data sets, and various techniques are employed to check the stability of the clusters obtained. Real potential is then determined by means of an empirical approach: practical application to a major French retailer's database of 5 million customers.

Click Fraud Bot Detection: High Stakes Adversarial Signal Detection

Brendan Kitts, Albert Roux, Jing Ying Zhang, Raj Mahato,
Ron Mills, Wesley Brandi, Kamran Kanany,
Matthew Rice, Gang Wu
Microsoft
One Microsoft Way
Redmond, WA 98116
(US) +1 (425) 722-7587
bkitts@microsoft.com

ABSTRACT

Click Fraud is a challenging problem which some have called the "Achilles Heel" of online advertising. Publishers sign up with Ad networks to display ads on their web pages. The Publishers receive a payout for clicks on these ads. Unfortunately this creates an incentive for the publisher to generate artificial clicks on ads and essentially print their own money. In order to maximize scam effectiveness, Publishers can employ sophisticated methods to cloak their attacks, including the use of distributed networks, hijacked browsers, and click fraud software designed to mimic humans. In this paper we will describe some of the attack vectors ranging from malware to "click fraud penetrators". We will also describe the large-scale data mining technologies employed to detect these programs. We conclude with some reflections on the adversarial nature of the field and some strategies for disrupting attacker evolution.

Correlation Explorations in a Classification Model

Vincent Lemaire
Orange Labs
2 avenue Pierre Marzin
22300 Lannion - France
+33 2 96 05 31 07
vincent.lemaire@orange-ftgroup.com

Carine Hue
GFI Informatique
11 rue Louis Broglie
22300 Lannion - France
chue@gfi.com

Olivier Bernier
Orange Labs
2 avenue Pierre Marzin
22300 Lannion - France
olivier.bernier@orange-ftgroup.com

ABSTRACT

This paper presents a new method to analyze the link between the probabilities produced by a classification model and the variation of its input values. The goal is to increase the predictive probability of a given class by exploring the possible values of the input variables taken independently. The proposed method is presented in a general framework, and then detailed for naive Bayesian classifiers. We also demonstrate the importance of "lever variables", variables which can conceivably be acted upon to obtain specific results as represented by class probabilities, and consequently can be the target of specific policies. The application of the proposed method to several data sets (data proposed in the PAKDD 2007 challenge and in the KDD Cup 2009) shows that such an approach can lead to useful indicators.

Categories and Subject Descriptors

G3 PROBABILITY AND STATISTICS [**Correlation and regression analysis**]: -; I.5 PATTERN RECOGNITION [**Design Methodology**]: Classifier design and evaluation

General Terms

Algorithms, Measurement, Economics, Experimentation.

Keywords

Exploration, Correlation, Classifier.

34. INTRODUCTION

Given a database, one common task in data analysis is to find the relationships or correlations between a set of input or explanatory variables and one target variable. This knowledge extraction often goes through the building of a model which represents these relationships (Han & Kamber, 2006). Faced with a classification problem, a probabilist model allows, for all the instances of the database and given the values of the explanatory variables, the estimation of the probabilities of occurrence of each class target.

These probabilities, or scores, can be used to evaluate existing policies and practices in organizations. They are not always directly usable, however, as they do not give any indication of what action can be decided upon to change this evaluation.

Consequently, it seems useful to propose a methodology which would, for every instance in the database, (i) identify the importance of the explanatory variables; (ii) identify the position of the values of these explanatory variables; and (iii) propose an action in order to change the probability of the desired class. We propose to deal with the third point by exploring the model relationship between each explanatory variable independently from each other and the target variable. The proposed method presented in this paper is completely automatic.

This article is organized as follows: the second section gives the context of the proposed method within Orange. This method is implemented using: (i) a platform for customer analysis, (ii) a tool, named Khiops, to construct classification models and (iii) a tool, named Kawab, to examine the contribution of the input variables and which (iv) allows the exploration of the correlation for these models.

The third section positions the approach in relation to the state of the art in feature importance (or selection), value importance and correlation analysis. The fourth section describes the method at first from a generic point of view and then for the naive Bayes classifier.

Through three illustrative examples, the fifth section allows a discussion and a progressive interpretation of the obtained results. The purpose of the first use case (titanic) is to illustrate the importance of the so-called “lever variables”. The aim of the second use case on the PAKDD challenge 2007 database is to show that our method can suggest useful actions in this case actions to increase the appetency to the product concerned by the challenge. The third use case is on Orange data and shows that with the platform CAP, the software Khiops and the functionalities of the Add-on for Khiops presented in this paper we have all the elements for a success story on “Data Mining Case Studies”. The last section concludes this paper and gives some future trends.

DMCS'09, June 28, Paris, France.
Copyright 2009 ACM 978-1-60558-674-8…$5.00.

35. CORRELATIONS EXPLORATIONS AS AN ELEMENT OF A COMPLETE DATA MINING PROCESS

This section describes the platform named CAP (Customer Analysis Platform) used to classify customers within the Orange information system. This platform (see section 2.1) implements a complete datamining process [10]. A data mining process is constituted of six main steps: business understanding, data understanding, data preparation, modeling, evaluation (interpretation) and deployment. The CAP platform implements in particular two important steps: the data preparation step and the deployment step. This platform also uses the Khiops software (see section 2.2) for the modeling step and uses an "add-on" for Khiops named Kawab for the interpretation or evaluation step (see sections 2.3 and 2.4).

35.1 The CAP platform

A heavy trend since the end of the last century is the exponential increase of the volume stored data. This increase does not automatically translate into richer information because the capacity to process data does not increase as quickly. With the current state of the technology, a difficult compromise must be reached between the implementation cost and the quality of the produced information. An industrial approach has been proposed in [11]1 allowing to increase considerably the capacity to transform data into useful information thanks to the automation of treatments and-the focus on relevant data.

35.2 The Khiops Software

Khiops is a data preparation and modeling tool for supervised and unsupervised learning. It exploits non parametric models to evaluate the correlation between any type of variables in the unsupervised case and the predictive importance of input variables or pairs of input variables in the supervised case. These evaluations are performed by means of discretization models in the numerical case and value grouping models in the categorical case, which correspond to the search for an efficient data representation owing to variable recoding. The tool also produces a scoring model for supervised learning tasks, according to a naive Bayes approach, with variable selection and model averaging. The tool is designed for the management of large datasets, with hundreds of thousands of instances and tens of thousands of variables, and was successfully evaluated in international data mining challenges. This tool is used by more than 60 users in Orange. Example of published applications see [24, 20]. Khiops can be downloaded here: http://www.khiops.com/.

35.3 Variable Contribution

We proposed in [20] a method to interpret the output of a classification (or regression) model. The interpretation is based on two concepts: the variable importance and the value importance of the variable. Unlike most of the state of art interpretation methods, our approach allows the interpretation of the model output for every instance. Understanding the score given by a model for one instance can for example lead to an immediate decision in a Customer Relational Management (CRM) system. Moreover the proposed method does not depend on a particular model and is therefore usable for any model or software used to produce the scores. This method has been sufficiently successful to be adopted by Orange business units which use commercial data-mining software like SAS™, Kxen™ or SPSS™.

For Orange business unit which uses our in house software, Khiops, we have developed an “add-on” to compute contribution indicators especially for the naive Bayes classifier. This add-on implements five importance or contribution indicators for the naive Bayes classifier: two indicators that we proposed (Minimum of variable probabilities difference and Modality Probability) and three other indicators generally found in the state of the art [26]. All these indicators are based on the comparison between the probability of the reference class knowing the value of all the explanatory variables and the probability of the reference class knowing the value of all except on explanatory variable.

35.4 Correlation Exploration

The purpose of this paper is to describe a new method capable of analyzing the correlations in the constructed classification model to propose an action in order to change the customer response. This method is implemented as an add-on for Khiops named Kawab, but as described in this paper, can be used whatever the modeling software used during the datamining process. This method will be downloadable at www.khiops.com in June 2009.

36. BACKGROUND

Machine learning abounds with methods for supervised analysis in regression and/ or classification. Generally these methods propose algorithms to build a model from a training database made up of a finite number of examples. The output vector gives the predicted probability of the occurrence of each class label. In general, however, this probability of occurrence is not sufficient and an interpretation and analysis of the result in terms of correlations or relationships between input and output variables is needed. The interpretation of the model is often based on the parameters and the structure of the model. One can cite, for example: geometrical interpretations [6], interpretations based on rules [30] or fuzzy rules [1], statistical tests on the coefficient's model [23]. Such interpretations are often based on averages for several instances, for a given model, or for a given task (regression or classification).

Another approach, called sensitivity analysis, consists in analyzing the model as a black box by varying its input variables. In such "what if" simulations, the structure and the parameters of the model are important only as far as they allow accurate computations of dependant variables using explanatory variables. Such an approach works whatever the model. A large survey of "what if" methods, often used for artificial neural network, is available in [21, 19].

36.1 Variable importance

Whatever the method and the model, the goal is often to analyze the behavior of the model in the absence of one input variable, or a set of input variables, and to deduce the importance of the input variables, for all examples. The reader can find a large survey in [14]. The measure of the importance of the input variables allows the selection of a subset of relevant variables for a given problem. This selection increases the robustness of models and simplifies the understanding of the results delivered by the model. The variety of supervised learning methods, coming from the statistical or artificial intelligence communities often implies importance indicators specific to each model (linear regression, artificial neural network ...).

Another possibility is to try to study the importance of a variable for a given example and not in average for all the examples. Given a variable and an example, the purpose is to obtain the variable importance only for this example: for additive classifiers see [25], for Probabilistic RBF Classification Network see [27], and for a general methodology see [20]. If the model is restricted to a naive Bayes Classifier, a state of art is presented in [22, 26]. This importance gives a specific piece of information linked to one example instead of an aggregate piece of information for all examples.

36.2 Importance of the value of an input variable

To complete the importance of a variable, the analysis of the value of the considered variable, for a given example, is interesting. For example Féraud et al. [12] propose to cluster examples and then to characterize each cluster using the variables importance and importance of the values inside every cluster. Framling [13] uses a "what if" simulation to place the value of the variable and the associated output of the model among all the potential values of the model outputs. This method which uses extremums and an assumption of monotonous variations of the output model versus the variations of the input variable has been improved in [20].

36.3 Instance correlation between an explanatory variable and the target class

This paper proposes to complete the two aspects presented above, namely the importance of a variable and the importance of the value of a variable. We propose to study the correlation, for one instance and one variable, between the input and the output of the model.

For a given instance, the distinct values of a given input variable can pull up (higher value) or pull down (lower value) the model output. The proposed idea is to analyze the relationship between the values of an input variable and the probability of occurrence of a given target class. The goal is to increase (or decrease) the model output, the target class probability, by exploring the different values taken by the input variable. For instance for medical data one tries to decrease the probability of a disease; in case of cross-selling one tries to increase the appetency to a product; and in government data cases one tries to define a policy to reach specific goals in terms of specific indicators (for example decrease the unemployment rate).

This method does not explore causalities, only correlations, and can be viewed as a method between:

- selective sampling [28] or adaptive sampling [29]: the model observes a restricted part of the universe materialized by examples but can "ask" to explore the variation space of the descriptors one by one separately, to find interesting zones.

- and causality exploration [18, 15]: as example D. Choudat [7] propose the imputability approach to specify the probability of the professional origin of a disease. The causality probability is, for an individual, the probability that his disease arose from exposures to professional elements. The increase of the risk has to be computed versus the respective role of each possible type exposures. In medical applications, the models used are often additive models or multiplicative models.

36.4 Lever variables

In this paper we also advocate the definition of a subset of the explanatory variables, the "lever variables". These lever variables are defined as the explanatory variables for which it is conceivable to change their value. In most cases, changing the values of some explanatory variables (such a sex, age...) is indeed impossible. The exploration of instance correlation between the target class and the explanatory variables can be limited in practice to variables which can effectively be changed.

The definition of these lever variables will allow a faster exploration by reducing the number of variable to explore, and will give more intelligible and relevant results. Lever variables are the natural target for policies and actions designed to induce changes of occurrence of the desired class in the real world.

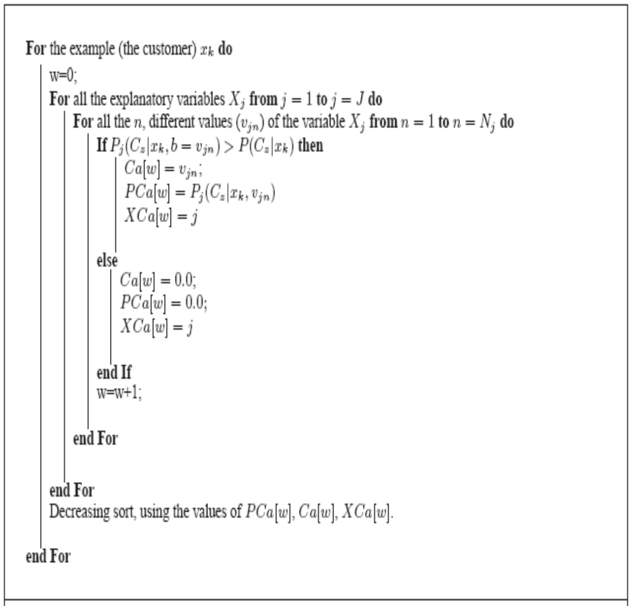
37. CORRELATION EXPLORATION – METHOD DESCRIPTION

In this section, the proposed method is first described in the general case, for any type of predictive model, and then tested on naive Bayes classifiers.

37.1 General case

Let C_z be the target class among T target classes. Let f_z be the function which models the predicted probability of the target class $f_z(X=x) = P(C_z \mid X=x)$, given the equality of the vector X of the J explanatory variables to a given vector x of J values. Let v_{jn} be all the n different possible values of the variable X_j.

The Algorithm 1 describes the proposed method. This algorithm tries to increase the value of $P(C_z \mid X = x_k)$ successively for each of the K examples of the considered sample set using the set of values of all the explanatory variables or lever variables. This method is halfway between selective sampling [28] and adaptive sampling [29]. The model observes a restricted part of the universe materialized by examples but can "ask" to explore the variation space of the descriptors one by one separately, to find interesting zones. The next subsections describe the algorithm in more details.



```
For the example (the customer) $x_k$ do
    w=0;
    For all the explanatory variables $X_j$ from $j = 1$ to $j = J$ do
        For all the $n$, different values $(v_{jn})$ of the variable $X_j$ from $n = 1$ to $n = N_j$ do
            If $P_j(C_z|x_k, b = v_{jn}) > P(C_z|x_k)$ then
                $Ca[w] = v_{jn}$;
                $PCa[w] = P_j(C_z|x_k, v_{jn})$
                $XCa[w] = j$
            else
                $Ca[w] = 0.0$;
                $PCa[w] = 0.0$;
                $XCa[w] = j$
            end If
            w=w+1;
        end For
    end For
    Decreasing sort, using the values of $PCa[w]$, $Ca[w]$, $XCa[w]$.
end For
```

Algorithm 1: Exploration and ranking of the score improvements

37.1.1 Exploration of input values

For the instance x_k, $P(C_z \mid x_k)$ is the "natural" value of the model output. We propose to modify the values of the explanatory variables or lever variables in order to study the variation of the model output for this example. In practice, we propose to explore the values independently for each explanatory variable. Let $P_j(C_z \mid x_k, b)$ be the output model f_z given the example x_k but for which the value of its j^{th} component has been replaced with the value b. For example, the third explanatory variable is modified among five variables: $P_3(C_z \mid x_k, b) = f_z(x_k^1, x_k^2, b, x_k^4, x_k^5)$. By scanning all the variables and for each of them all the set of their possible values, an exploration of "potential" values of the model output is computed for the example x_k.

37.1.2 Domain of exploration

The advantage of choosing the empirical probability distribution of the data as domain of exploration has been showed experimentally in [5, 19, 20]. A theoretical proof is also available for linear regression in [8] and for naive Bayes classifiers in [26]. Consequently the values used for the J explanatory variables will be the values of the K examples available in the training database. This set can also be reduced using only the distinct values: let N_j be the number of distinct values of the variable X_j.

37.1.3 Results ranking

The exploration of the explanatory variables or of the lever variables is done by scanning all the possible values taken by the examples in the training set. When the modification of the value of the variable leads to an improvement of the–probability predicted by the model, three pieces of data are kept (i) the value which leads to this improvement (*Ca*); (ii) the associated improved probability (*PCa*); and (iii) the variable associated to this improvement (*XCa*). These triplets are then sorted according to the improvement obtained on the predicted probability. Note: if no improvement is found, the tables *CA* and *PCa* only contain null values.

It should also be possible (i) to explore jointly two or more explanatory variables; (ii) or to use the value (*Ca*[0]) which best improves the output of the model ($P(C_z \mid X = x)$) (this value *Ca*[0] is available at the end of the Algorithm) and then to repeat again the exploration on the example x_k on its others explanatory variables. These other versions are not presented in this paper but will be the focus of future works.

Algorithm 1: Exploration and ranking of the score improvements

37.1.4 Cases with class changes

When using Algorithm 1, the predicted class can change. Indeed it is customary to use the following formulation to designate the predicted class of the example x_k:

$$\arg\max_{z} P(C_z \mid x_k)$$

Using Algorithm 1 for x_k belonging to the class t $(t \neq z)$ could produce $P(C_z \mid x_k, b) > P(C_t \mid x_k)$. In this case the corresponding value (*Ca*) carries important information which can be exploited.

The use of Algorithm 1 can exhibit three types of values (*Ca*):

- values which do not increase the target class probability;

- values which increase the target class probability but without class change (the probability increase is not sufficient);
- values which increase the target class probability with class change (the probability increase is sufficient).

The examples whose predicted class changes from another class to the target class are the primary target for specific actions or policies designed to increase the occurrence of this class in the real world.

37.2 Case of a naive Bayesian classifier

A naive Bayes classifier assumes that all the explanatory variables are independent knowing the target class. This assumption drastically reduces the necessary computations. Using the Bayes theorem, the expression of the obtained estimator for the conditional probability of a class C_z is:

$$P(C_z|x_k)=\frac{P(C_z)\prod_{j=1}^{J}P(X_j=x_{jk}|C_z)}{\sum_{t=1}^{T}P(C_t)\prod_{j=1}^{J}P(X_j=x_{jk}|C_t)} \quad (1)$$

The predicted class is the one which maximizes the conditional probabilities. Despite the independence assumption, this kind of classifier generally shows satisfactory results [17]. Moreover, its formulation allows an exploration of the values of the variables one by one independently.

The probabilities $P(X_j = v_{jk} \mid C_z)$ ($\forall$ j, k, z) are estimated using counts after discretization for numerical variables or grouping for categorical variables [3]. The denominator of the equation above normalizes the result so that $\sum_z P(C_z \mid x_k)=1$.

The use of the Algorithm 1 requires to compute $P(C_z \mid X =x_k)$, and $P_j(C_z \mid X =x, b)$ which can be written in the form of Equations 2 and 3:

$$P(C_z|x_k)=\frac{\overbrace{P(C_z)\prod_{j=1}^{J}P(X_j=v_{jk}|C_z)}^{e^{L_z}}}{\sum_{t=1}^{T}P(C_t)\prod_{j=1}^{J}P(X_j=v_{jk}|C_t)} \quad (2)$$

$$P_j(C_z|x_k,b)=\frac{\overbrace{P(C_z)\prod_{j=1,j\neq q}^{J}P(X_j=v_{jk}|C_z)\,P(X_q=b|C_z)}^{e^{L_{z'}}}}{\sum_{t=1}^{T}\left[P(C_t)\prod_{j=1,j\neq q}^{J}P(X_j=v_{jk}|C_t)\right]P(X_q=b|C_t)} \quad (3)$$

In Equations 2 and 3 numerators can be written as e^{Lz} and $e^{Lz'}$ with:

$$L_z=\log(P(C_z))+\sum_{j=1}^{J}\log(P(X_j=x_{jk}|C_z))$$

and

$$L_{z'}=\log(P(C_z))+\sum_{j=1,j\neq q}^{J}\log(P(X_j=x_{jk}|C_z))+\log(P(X_q=b|C_z))$$

This formulation will be used below.

37.2.1 Implementation details on very large databases

To measure the reliability of our approach, we tested it on marketing campaigns of France Telecom (results not allowed for publication until now). Tests have been performed using the PAC platform [11] on different databases coming from decision-making applications. The databases used for testing had more than 1 million of customers, each one represented by a vector including several thousands of explanatory variables. These tests raise several implementation points enumerated below:

- To avoid numerical problems when comparing the "true" output model $P(C_z \mid x_k)$ and the "explored" output $P_j(C_z \mid x_k, b), P(C_x \mid x_k)$ is computed as:

  $$P(C_x|x_k)=\frac{1}{\sum_{t=1}^{T}e^{L_t-L_x}}$$

  where

  $$L_t=\log(P(C_t))+\sum_{j=1}^{J}\log(P(X_j=x_{jk}|C_t))$$

- To reduce the computation time: the modified output of the classifier can be computed using only several additions or subtractions since the difference between L_z (used in Equation 2) and $L_{z'}$ (used in Equation 3) is:

  $L_{z'} = L_z - log(P(x_q=v_{jk} \mid C_z)) + log(P(X_q=b \mid C_z))$

- Complexity: For a given example x_k, the computation of tables presented in Algorithm 1 is of complexity $O(\sum_{j=1}^{d} N_j)$.

This implementation is "real-time" and can be used by an operator who asks the application what actions to do, for example to keep a customer.

38. EXPERIMENTATIONS

In this section we describe the application of our proposed method to three illustrative examples. This first example, the Titanic database, illustrates the importance of lever variables. The second example illustrates the results of our method on the dataset used for the PAKDD 2007 challenge. Finally, we present the results obtained by our method on the KDD Cup 2009.

38.1 The Titanic database

38.1.1 Data and experimental conditions

In this first experiment the Titanic (www.ics.uci.edu/~mlearn/) database is used. This database consists of four explanatory variables on 2201 instances (passengers and crew members). The first attribute represents the class trip (status) of the passenger or if he was a crew member, with values: 1st, 2nd, 3rd, crew. The second (age) gives an age indication: adult, child. The third (sex) indicates the sex of the passenger or crew: female or male. The last attribute (survived) is the target class attribute with values: no or yes. Readers can find for each instance the variable importance and the value importance for a naive Bayes classifier in [26].

Among the 2201 examples in this database, a training set of 1100 examples randomly chosen has been extracted to train a naive Bayes classifier using the method presented in [3]. The remaining examples constitute a test set. As the interpretation of a model with low performance would not be consistent, a prerequisite is to check if this naive Bayes classifier is correct. The model used here [16] gives satisfactory results:

- Accuracy on Classification (ACC) on the train set: 77.0%; on the test set: 75.0%;
- Area under the ROC curve (AUC) (Fawcett, 2003) on the train set: 73.0%; on the test set: 72.0%.

The purpose here is to the see another side of the knowledge produced by the classifier: we want to find the characteristics of the instances (people) which would have allowed them to survive.

38.1.2 Input values exploration

Algorithm 1 has been applied on the test set to reinforce the probability to survive. Table 1 shows an abstract of the results: (i) it is not possible to increase the probability for only one passenger or crew; (ii) the last column indicates that, for persons predicted as surviving by the model (343 people), the first explanatory variable (status) is the most important to reinforce the probability to survive for 118 cases; then the second explanatory variable (age) for 125 cases; and at last the third one (sex) for 100 cases. (iii) For people predicted as dead by the model (758) the third explanatory variable (sex) is always the variable which is the most important to reinforce the probability to survive.

Table 1: Ranking of explanatory variables

	Size	Status / Age / Sex
Predicted 'yes'	343	118 / 125 / 100
Predicted 'no'	758	0 / 0 / 758

These 758 cases predicted as dead are men and if they were women their probability to survive would increase sufficiently to survive (in the sense that their probability to survive would be greater than their probability to die). Let us examine then, for these cases, additional results obtained by exploring the others variables using Algorithm 1:

- the second best variable to reinforce the probability to survive is (and in this case they survive):
  - for 82 of them (adult + men + 2nd class) the second explanatory variable (age);
  - for 676 of them (adult + men + (crew or 3rd class)) the first explanatory variable (status);
- the third best variable to reinforce the probability to survive is (and in this case nevertheless they are dead):
  - for 82 of them (adult + men + 2nd class) the first explanatory variable (status);
  - for 676 of them (adult + men + (crew or 3rd class)) the second explanatory variable (age).

Of course, in this case, most explanatory variables are not in fact lever variables, as they cannot be changed (age or sex). The only variable that can be changed is status, and even in this case, only for passengers, not for crew members. The change of status for passengers means in fact buying a first class ticket, which would have allowed them a better chance to survive. The other explanatory variables enable us to interpret the obtained survival probability in terms of priority given to women and first class passengers during the evacuation.

38.2 Application to sale: results on the PAKDD 2007 challenge

38.2.1 Data and experimental conditions

The data of the PAKDD 2007 challenge are used (http://lamda.nju.edu.cn/conf/pakdd07/dmc07/): The data are not on-line any more but data descriptions and analysis results are still available. Thanks to Mingjun Wei (participant referenced P049) for the data (version 3).

The company, which gave the database, has currently a customer base of credit card customers as well as a customer base of home loan (mortgage) customers. Both of these products have been on the market for many years, although for some reasons the overlap between these two customer bases is currently very small. The company would like to make use of this opportunity to cross-sell home loans to its credit card customers, but the small size of the overlap presents a challenge when trying to develop an effective scoring model to predict potential cross-sell take-ups.

A modeling dataset of 40,700 customers with 40 explanatory variables, plus a target variable, had been provided to the participants (the list of the 40 explanatory variables is available at http://perso.rd.francetelecom.fr/lemaire/data_pakdd.zip). This is a sample of customers who opened a new credit card with the company within a specific 2-year period and who did not have an existing home loan with the company. The target categorical variable "Target_Flag" has a value of 1 if the customer then opened a home loan with the company within 12 months after opening the credit card (700 random samples), and has a value of 0 otherwise (40,000 random samples).

A prediction dataset (8,000 sampled cases) has also been provided to the participants with similar variables but withholding the target variable. The data mining task is to produce a score for each customer in the prediction dataset, indicating a credit card customer's propensity to take up a home loan with the company (the higher the score, the higher the propensity).

The challenge being ended it was not possible to evaluate our classifier on the prediction dataset (the submission site is closed). Therefore we decide to elaborate a model using the 40 000 samples in a 5-fold cross validation process. In this case each 'test' fold contains approximately the same number of samples as the initial prediction dataset. The model used is again a naive Bayes classifier (Boullé, 2008; Guyon, Saffari, et al., 2007). The results obtained on the test sets are:

- Accuracy on Classification (ACC): 98.29% ± 0.01% on the train sets and 98.20% ± 0.06% on the test sets.
- Area under the ROC curve (AUC): 67.98% ± 0.74% on the train sets and 67.79% ± 2.18% on the test sets.
- Best results obtained on one of the folds: Train set AUC=68.82%, Test set AUC=70.11%.

Table 2: PAKDD 2007 challenge: the first three best results

id participant	AUC for test set	Rank	Modeling Technique
P049	70.01%	1	TreeNet + Logistic Regression
P085	69.99%	2	Probit Regression
P212	69.62%	3	MLP + n-Tuple Classifier

Table 2 shows the first three best results and corresponding method of winners of the challenge. Results obtained here by our model are coherent with those of the participants of the challenge.

38.2.2 Input values exploration

The best classifier obtained on the test sets in the previous section is used. This naive Bayes classifier (Boullé, 2007) uses 8 variables out of 40 (the naïve Bayes classifier takes into account only input variables which have been discretized (or grouped) in more than one interval (or group) see (Boullé, 2006)). These 8 variables and their intervals of discretization (or groups) are presented in Table 3. All variable are numerical except for the variable "RENT_BUY_CODE" which is symbolic with possible values of 'O' (Owner), 'P' (Parents), 'M' (Mortgage), 'R' (Rent), 'B' (Board), 'X' (Other).

The lever variables were chosen using their specification (see http://lamda.nju.edu.cn/conf/pakdd07/dmc07/ or the appendix A). These lever variables are those for which a commercial offer to a customer can change the value. We define another type of variable which we will explore using our algorithm, the observable variables. These variables are susceptible to change during a life of a customer and this change may augment the probability of the target class, the propensity to take up a home loan. In this case, the customers for which this variable has changed can be the target of a specific campaign. For example the variable "RENT_BUY_CODE" can not be changed by any offer but is still observable. The customer can move from the group of values [O,P] ('O' Owner, 'P' Parents) to [M,R,B,X] ('M' Mortgage, 'R' Rent, 'B' Board, 'X' Other). Among the eight variables (see Table 3) chosen by the training method of the naive Bayes classifier, two are not considered as 'lever' variables or observable variables ("AGE_AT_APPLICATION" and "PREV_RES_MTHS") and will not be explored.

Table 3: Selected explanatory variables (there is no reason in [2] to have two intervals for each variable, it is here blind chance).

Explanatory Variables	Interval 1 or Group 1	Interval 2 or Group 2
RENT_BUY_CODE	M,R,B,X	O,P
PREV_RES_MTHS	]-∞,3.5[	[3.5,+∞ [
CURR_RES_MTHS	]-∞,40.5[	[40.5,+∞ [
B_ENQ_L6M_GR3	]-∞,0.5[	[0.5,+∞ [
B_ENQ_L3M	]-∞,0.5[	[3.5,+∞ [
B_ENQ_L12M_GR3	]-∞,1.5[	[1.5,+∞ [
B_ENQ_L12M_GR2	]-∞,0.5[	[0.5,+∞ [
AGE_AT_APPLICATION	]-∞,45.5[	[45.5,+∞ [

Algorithm 1 has been applied on the 40700 instances in the modeling data set. The 'yes' class of the target variable is chosen as target class (C_z = 'yes'). This class is very weakly represented (700 positive instances out of 40700). The AUC values presented in Table 2 or on the challenge website does not show if customers are classified as 'yes' by the classifier. Exploration of lever variables does not allow in this case a modification of the predicted class. Nevertheless Table 4 and Figure 1 show that a large improvement of the 'yes' probability (the probability of cross-selling) is possible.

In Table 4 the second column (C2) presents the best $P_j(C_z \mid x_k, b)$ obtained, the third column (C3) the initial corresponding $P(C_z \mid x_k, b)$, the fourth column (C4) the initial interval used in the naive Bayes formulation (used to compute $P(C_z \mid x_k, b)$) and the last column (C5) the interval which gives the best improvement (used to compute $P_j(C_z \mid x_k, b)$). This table shows that:

- for all lever or observable variables, there exists a value change that increases the posterior probability of occurrences of the target class;
- the variable that leads to the greatest probability improvement is B_ENQ_L3M (The number of Bureau Enquiries in the last 3 months), for a value in [1.5,+∞[rather than in]- ∞,1.5[; This variable is an observable variable, not a lever variable, and means that a marketing campaign should be focused on customers who contacted the bureau more than once in the last three months;
- nevertheless, none of those changes leads to a class change as the obtained probability ($P_j(C_z \mid x_k, b)$) stays smaller than $P(C_z \mid x_k)$.

Table 4: Best *$P(C_z)$='yes'* obtained

C1: explored variable	C2	C3	C4	C5
RENT_BUY_CODE	0.6	0.26	[O,P]	[M,R,B,X]
CURR_RES_MTHS	0.36	0.21	[40.5,+ ∞ [	]-∞,40.5[
B_ENQ_L6M_GR3	0.25	0.10	]-∞,0.5[	[0.5,+∞[
B_ENQ_L3M	0.12	0.12	]-∞,1.5[	[1.5,+ ∞[
B_ENQ_L12M_GR3	0.36	0.16	]-∞,0.5[	[1.5,+∞[
B_ENQ_L12M_GR2	0.36	0.24	[0.5,+∞[	]- ∞,0.5[

In Figure 1 the six dotted vertical axis represent the six lever or observable variables as indicated on top or bottom axis. On the left hand size of each vertical axis, the distribution of $P(C_z \mid x_k)$ is plotted (□) and on the right hand size the distribution of $P_j(C_z \mid x_k, b)$ is plotted (■). Probability values are indicated on the y-axis. In this Figure only the best $P_j(C_z \mid x_k, b)$ (*PCa*[0] in Algorithm 1) is plotted. This figure illustrates in more details the same conclusions as given above.

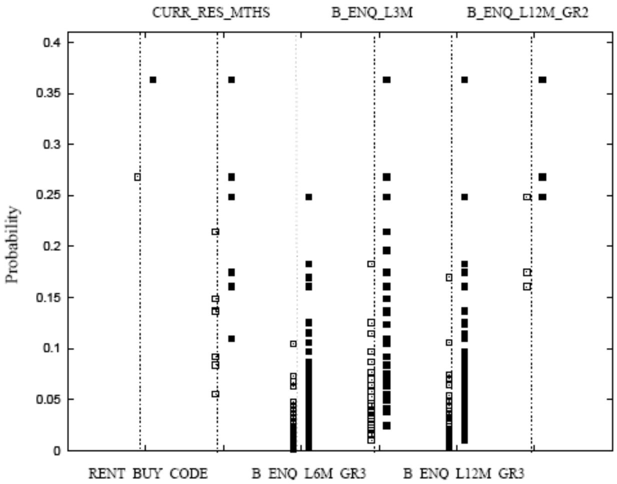


Fig 1: Obtained results on $P_j(C_z \mid x_k, b)$.

38.3 Test on the KDD Cup 2009

38.3.1 Task description

The KDD Cup 2009 offers the opportunity to work on large marketing databases from the French Telecom company Orange. The goal is to predict the propensity of customers to switch provider (churn), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable (up-selling).

In this section we consider only the problem of churn. The churn rate is also sometimes called the attrition rate. In its broadest sense, the churn rate is a measure of the number of individuals or items moving into or out of a collection over a specific period of time. The term is used in many contexts, but is most widely applied in business. For instance, it is an important factor for-mobile telephone networks and pay TV operators.

In this study the moment of churn is the moment when the client cancels (“closes”) his Orange product or service. A churner is a client having a product or service at time t_n and having no product at time t_{n+1} For more details see the presentation at: http://perso.rd.francetelecom.fr/lemaire/kddcup/ChallengePresentation_03192009.pdf. Churn has high cost as to conquer a customer is more expensive than to keep a customer.

38.3.2 Data and experimental conditions

In this paper we consider the small dataset available at the end of the fast challenge. Both training and test sets contain 50,000 examples. This real life dataset has numerical and categorical variables: the first 190 variables are numerical and the last 40 are categorical. These 230 variables are currently used by the marketing teams.

We used the Khiops software to elaborate a naive Bayes classifier. The performance of this classifier, on the small dataset, can be found on the challenge website with the name “reference”. The AUC obtained on the training set is 0.6791 and 0.6827 on 10% of the test set. Results on 100% of the test set will be available only at the end of the challenge and can not be divulged for the time being.

38.3.3 Input values exploration

Algorithm 1 has been applied on the training set to reinforce the probability to stay loyal (the probability of not churning, the reference class C_z). Only the variable which reinforces the most the probability of not churning is kept.

Table 5 presents the list of variables which can reinforce the probability of not churning (note that sometimes it is not possible to reinforce this probability, therefore the sum of the second column is not equal to 50000).

In this table the first column gives the identifier of the input variable, the second column (C2) gives the number of client who see their probability of not churning reinforced using the variable indicates in column one, the third column (C3) indicates for the corresponding line the number of customers for whom the reinforcement leads to a change of class for the reference class, the fourth column (C4) the number of customers for whom the initially predicted class is the reference class and the last column (C5) the number of customers for whom the reinforcement does not change the predicted class for the reference class.

For the challenge, the meaning of the variables is not revealed so it is not possible to see if these variables are lever variables. But the results of the table 5 indicate a high potential of the proposed method for Orange. Even using an “action” which does not lead to a class change, as for example when using the input variable 189, clients are pushed far from the churn “boundary” (see Figure 2).

Two types of action are possible:

- Preventive (or Push) Action: to prevent a customer from churning an operator can propose an offer, a service, to a customer and in this case his corresponding attribute changes to decrease (or to increase) a probability output of the model;
- Reactive Action: a modification of an attribute of customer is observed and detected since with this value this customer goes near the churn boundary.

Table 5: List of best variables

Variable	C2	C3	C4	C5
6	16	0	16	0
7	4	0	4	0
73	3885	371	3465	49
74	775	0	775	0
81	122	47	68	7
113	884	52	828	4
126	42148	5627	34949	1572
189	446	51	390	5
193	61	0	61	0
205	96	0	96	0
206	372	0	372	0
210	2	0	2	0
212	168	0	168	0
213	8	0	8	0
218	377	62	275	40
227	254	0	254	0
228	382	0	382	0

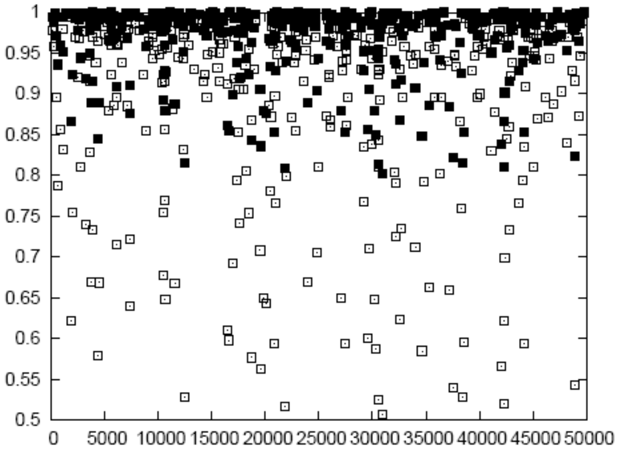


Figure 2: Obtained results on $P_j(C_z|x_k, b)$ using the input variable 189. The horizontal axis indicates the number of the client in the training dataset. On the vertical axis: the distribution of $P(C_z|x_k)$ is plotted using a □ and the distribution $P_j(C_z|x_k, b)$ is plotted using a ■. In this Figure only the best $P_j(C_z|x_k, b)$ (PCa[0] in Algorithm 1) is plotted and only for clients for who a reinforcement is possible but who were already classified as “loyal” (column C4 in Table 5).

39. CONCLUSION

In this paper we proposed a method to study the influence of the input values on the output scores of a probabilistic model. This method is a part of a complete data mining process adopted by several Orange business units.

The method has first been defined in a general case valid for any model, and then been detailed for naive Bayes classifier. We also demonstrate the importance of "lever variables", explanatory variables which can conceivably be changed. Our method has first been illustrated on the simple Titanic database in order to show the need to define lever variables. Then, on the PAKDD 2007 challenge databases, a difficult problem of cross-selling, the results obtained show that it is possible to create efficient indicators that could increase sells. Finally we demonstrated the applicability of our method to the KDD Cup 2009.

The case study presented on the Titanic dataset illustrates the point of applying the proposed method to accident research. It could be used for example to analyze road accidents or air accidents. In the case of the air accidents any new plane crash is thoroughly analyzed to improve the security of air flights. Despite the increasing number of plane crashes, the relative frequency of those in relation to the volume of traffic is decreasing and air security is globally improving. Analyzing the correlations between the occurrence of a crash and several explanatory variables could lead to a new approach to the prevention of plane crashes.

This type of relationship analysis method has also great potential for medicine applications, in particular to analyze the link between vaccination and mortality. The estimated 50% reduced overall mortality currently associated with influenza vaccination among the elderly is based on studies neither fully taking into account systematic differences between individuals who accept or decline vaccination nor encompassing the entire general population. The proposed method in this paper could find interesting data for infectious diseases research units. Another potential area of application is the analysis of the factors causing a disease, by investigating the link between the occurrence of the disease and the potential factors.

Three main future works are also under consideration:

- the study of the temporal evolution of predicted scores when the values of the explicative variables are likely to change;
- the possibility to learn iteratively a new predictive model after having modified the data according to the best action found after correlation exploration;
- performing a controlled test on the 'lever variables' to see if the action of moving from one set of values to another affects churn/response/survival etc - i.e. if there is in fact an underlying causality in the lever variable (this causality can not be concluded from correlations).

The proposed method is very simple but efficient. It is now implemented in an add-on of the Khiops software—(see http://www.khiops.com), and its user guide (including how to obtain the software) is available at: http://perso.rd.francetelecom.fr/lemaire/understanding/Guide.pdf

This tool could be useful for companies or research centers who want to analyze classification results with input values exploration.

40. REFERENCES

[1] J. M. Benitez, J. L. Castro, and I. Requena. Are artificial neural networks black boxes. IEEE Transactions on Neural Networks, 8(5):1156–1164, 1997. Septembre.

[2] M. Boullé. Compression-based averaging of selective naive Bayes classifiers. Journal of Machine Learning Research, 8:1659–1685, 2007.

[3] M. Boullé. Khiops: outil de préparation et modélisation des données pour la fouille des grandes bases de données. In Extraction et gestion des connaissances (EGC'2008), pages 229–230, 2008.

[4] M. Boullé. Modl: a bayes optimal discretization method for continuous attributes. Machine Learning, 65(1):131–165, 2006.

[5] L. Breiman. Random forest. Machine Learning, 45, 2001.stat-www.berkeley.edu/users/breiman/Breiman.

[6] J. J. Brennan and L. M. Seiford. Linear programming and l1 regression: A geometric interpretation. Computational Statistics & Data Analysis, 1987.

[7] D. Choudat. Risque, fraction étiologique et probabilité de causalité en cas d'expositions multiples, i : l'approche théorique. Archives des Maladies Professionnelles et de l'Environnement, 64(3):129–140, 2003.

[8] G. Diagne. Sélection de variables et méthodes d'interprétation des résultats obtenus par un modèle boite noire. Master's thesis, UVSQ-TRIED, 2006.

[9] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Labs, 2003., 2003.

[10] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. A*dvances in Knowledge Discovery and Data Mining*, chapter From data mining to knowledge discovery : An overview. AAAI/MIT Press, 1996.

[11] R. Féraud, M. Boullé, F. Clérot, and F. Fessant. Vers l'exploitation de grandes masses de données. In *Extraction et Gestion des Connaissances (EGC)*, pages 241–252, 2008.

[12] R. Féraud and F. Clérot. A methodology to explain neural network classification. *Neural Networks*, 15(2):237–246, 2002.

[13] K. Främling. *Modélisation et apprentissage des preferences par réseaux de neurones pour l'aide à la decision multicritère*. PhD thesis, Institut National des Sciences Appliquées de Lyon, 1996.

[14] I. Guyon. *Feature extraction, foundations and applications*. Elsevier, 2005.

[15] I. Guyon, C. Constantin Aliferis, and A. Elisseeff. *Computational Methods of Feature Selection*, chapter Causal Feature Selection, pages 63–86. Chapman and Hall/CRC Data Mining and Knowledge Discovery Ser., 2007.

[16] I. Guyon, A. Saffari, G. Dror, and J. Bumann. Report on preliminary experiments with data grid models in the agnostic learning vs. prior knowledge challenge. In *IJCNN: International Joint Conference on Neural Networks*, 2007.

[17] D. Hand and K. Yu. Idiot's Bayes - not so stupid after all? *International Statistical Review*, 69(3):385–399, 2001.

[18] M. S. Kramer, J. M. Leventhal, T. A. Hutchinson, and A. R. Feinstein. An algorithm for the operational assessment of adverse drug reactions. i. background, description, and instructions for use. *Journal of the American Medical Association*, 242(7):623–632, 1979.

[19] V. Lemaire and R. Féraud. Driven forward features selection: a comparative study on neural networks. In *International Conference on Neural Information Processing*, 2006.

[20] V. Lemaire, R. Féraud, and N. Voisine. Contact personalization using a score understanding method. In

International Joint Conference on Neural Network, Hong-Kong, October 2008.

[21] P. Leray and P. Gallinari. Variable selection. Technical Report ENV4-CT96-0314, University Paris 6, 1998.

[22] M. Možina, J. Demšar, M. Kattan, and B. Zupan. Nomograms for visualization of naive Bayesian classifier. In *PKDD '04: Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 337–348, New York, USA, 2004. Springer-Verlag New York, Inc.

[23] J. Nakache and J. Confais. *Statistique explicative appliquée*. TECHNIP, 2003.

[24] P. Poirier, C. Bothorel, E. Guimier De Neef, and M. Boullé. Automating opinion analysis in film reviews : the case of statistic versus linguistic approach. In *LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology*, pages 94–101, 2008.

[25] B. Poulin, R. Eisner, D. Szafron, P. Lu, R. Greiner, D. S. Wishart, A. Fyshe, O. Pearcy, C. Macdonell, and J. Anvik. Visual explanation of evidence with additive classifiers. In *IAAI*, 2006.

[26] M. Robnik-Sikonja and I. Kononenko. Explaining classifications for individual instances. *IEEE TKDE*, 20(5):589–600, 2008.

[27] M. Robnik-Sikonja, A. Likas, C. Constantinopoulos, and I. Kononenko. An efficient method for explaining the decisions of the probabilistic rbf classification network. currently under review, partialy available as TR, `http://lkm.fri.uni-lj.si/rmarko`, 2009.

[28] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. 18th International Conf. on Machine Learning*, pages 441–448.Morgan Kaufmann, San Francisco, CA, 2001.

[29] A. Singh, R. Nowak, and P. Ramanathan. Active learning for adaptive mobile sensing networks. In *IPSN '06: Proceedings of the fifth international conference on Information processing in sensor networks*, pages 60–68, New York, NY, USA, 2006. ACM Press.

[30] S. Thrun. Extracting rules from artifcial neural networks with distributed representations. In M. Press, editor, *Advances in Neural Information Processing Systems*, volume 7, Cambridge, MA, 1995. G. Tesauro, D. Touretzky, T. Leen