ISSUES IN CUSTOMER INTELLIGENCE: DATA AND METHOD CREATIVITY TO IMPROVE MARKETING DECISION MAKING

Kristof Coussement

2008

Dissertation submitted to the Faculty of Economics and Business Administration of Ghent University in fulfillment of the requirements for the degree of Doctor in Applied Economic Sciences

Advisor: Prof. dr. Dirk Van den Poel

ISSUES IN CUSTOMER INTELLIGENCE: DATA AND METHOD CREATIVITY TO IMPROVE MARKETING DECISION MAKING

Kristof Coussement

2008

Dissertation submitted to the Faculty of Economics and Business Administration of Ghent University in fulfillment of the requirements for the degree of Doctor in Applied Economic Sciences

Advisor: Prof. dr. Dirk Van den Poel

"Success is the ability to go from one failure to another with no loss of enthusiasm"
(Sir Winston Churchill).

"Success is a matter of hanging on when others are giving up"
(William Feather)

"The essence of intelligence is the skill of extracting meaning from everyday experience"
(Unknown)

Doctoral committee

Prof. dr. Marc De Clercq
Dean-president, Ghent University

Prof. dr. Patrick Van Kenhove
Academic secretary, Ghent University

Prof. dr. Manfred Krafft
Münster University

Prof dr. Yong Seog Kim
Utah State University

Prof. dr. Stefan Van Aelst
Ghent University

Prof. dr. Mario Vanhoucke
Ghent University

Prof. dr. Dirk Van den Poel
Advisor, Ghent University

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to grab this opportunity to thank all people who contributed – directly or indirectly - to my dissertation.

First, I would like to thank all my PhD committee members for spending the time and effort in reading my manuscript. I really appreciate their input which gave me some high-quality advice for further research and which upgraded my final thesis in the end.

Moreover, I would like to write a few words for my advisor, Prof. dr. Dirk Van den Poel. Despite the fact that we have different characters resulting in other norms and different views on how to do scientific research, I would like to thank him for the scientific freedom he always gave me, for his approval to follow several courses and to attend several interesting conferences, for resolving my computer problems and for linguistically correcting my research papers. Due to his way of 'cooperation', I am fully confident that I am well prepared for tackling my future adventures and knitting a nice piece to my academic career as (assistant) professor in quantitative marketing at the University College Brussels!

Patrick, Maggie, all research members and ex-colleagues of the department of marketing, I would like to thank you for the nice time we spent together. The dynamic atmosphere is unique and I am quite sure this helped me in finishing my PhD so quickly! Moreover, I would like to thank the Old Guard of the party committee! I really appreciated their empathy during the last few years. I wish you all the best for the future!

Karin, la Mama, I would like to thank you for the nice chit-chats we have had, for your help during my administrative storms… I really enjoyed your stay at the department!

Additionally, I would like to thank my parents, my family, my in-laws and all my friends for their interest in my research during my PhD period. Thank you very much for all your love and care and for all your support which for sure helped me the last few years!

Toontje Teletext, I would like to thank you for your interest in my research and my nice stories from time to time! I really appreciate the way you were supporting me! Thank you very much for being the best company I have ever had during the conference season!

Koen, surfer's boy Koen, I admire your spirit and dedication for doing profound scientific research. If I was Ludje (and please don't image that!), I would for sure hire you! Moreover, I really liked your jokes about fancy creations under the armpit of professors and I will never forget the unique functioning of your stomach in thirsty times. I really like to wish you all the best with your ensemble-based research and I really hope we can go for lunch some day or another once you are the president of this boîte!

Grietje, my office mate, I would like to thank you very much for our nice chit-chats, for calming me down in stressful situations, for the interesting methodological conversations we have had… Moreover, I really like to write that I was sometimes really astonished by your motivation to make the world a little bit better by sharing your knowledge to non-profit organizations. I am convinced that you are becoming one of the CI pioneers in the Belgian fundraising scene once you finished your PhD. And as you already know, I'll always be there whenever you need my help!

Dries, Mr. Bayes from the department of marketing, I admire your power and motivation to achieve research goals that are sometimes hard to reach, e.g. Bayesian statistics, social network analysis,… . Moreover, I look up to your atypical cluster-characteristics of sincerity, honesty, openness and sense for team spirit. Furthermore, I would like to thank you for the many orgasms we shared during the past few years. Finally, iek kwamme gelik heel tennen toen ge ui oar liet groeen, moar zoals iek et vandoage zeu zie valt gelik nog mee!

Stefaan, mister VUM, I would like to thank you for your dedication and motivation for making 'OUR' project at VUM a successful one! I am very proud to say that I used your data in all my studies! Muchas gracias for all the efforts you made on the data side! I not only appreciate your help during our joint projects, but I really appreciate the way you are: Fantastic!

Geert and Jones, I really like to thank you for introducing me in the academic world. You were the ones who freely reviewed my research papers, who gave some high quality advice on how to write a research paper and last but not least, you are two fantastic guys which I admire a lot! Geert, I wish you all the best with the expansion of your leading company in CI, Python Predictions! And Jones, I am quite sure that I was not the last man standing!

Bart, I really like to thank you for your empathy, your helpfulness and your criticism on my research papers! I always liked your points of view and your well-considered opinions and remarks. It is clear that I would not be here today in such a short time period without your help! For me and for a lot of people, you will always be my real promoter! Moreover, you are a wonderful friend: someone who supported me in difficult times and shared the happiness in good times! I wish you all the best with your projects in the big marketing journals. I would like to conclude that an act says more than a thousand words: BART, THANK YOU VERY MUCH FOR EVERYTHING!

Schatteke, I really like to thank you for the way you have supported me during my PhD. I admire your enthusiasm for my scientific research, your motivation to participate in my nightly sessions, your thoughts on how to solve my problems… Thank you very much for all your friendship, your love and warmth which guided me through my dissertation!! Thank you very much for setting me back on track and bringing me down to earth over and over again! I love you with everything I am, and more than anyone ever thought possible!!

Kristof Coussement,
17th of October, 2008.

NEDERLANDSTALIGE SAMENVATTING

In de voorbije decennia komen nieuwe uitdagingen voor marketing managers de kop op steken. Zij proberen zoveel mogelijk informatie van hun klanten te verzamelen en te integreren in een klantendatabase. Heden ten dage zijn er reeds heel wat academische en bedrijfsrelevante cases uitgewerkt die de hoge bijkomende waarde aantonen van een klantendatabase. Marketing analysten benadrukken de lange termijn relatie met hun klanten gebruikmakend van een weloverwogen CRM strategie. Bijvoorbeeld, er is aangetoond dat het verzorgen van bestaande klanten minder duur is dan het aantrekken van nieuwe klanten die veelal een lage retentiegraad vertonen. Men kan besluiten dat de transactionele database het startpunt is geworden van heel wat marketing gerelateerde analyses, wat ook het geval is in Customer Intelligence. Dit doctoraat geeft de lezer een introductie tot Customer Intelligence en focust hoe de marketing analyst het marketing beslissingsproces kan verbeteren door middel van data en methode creativiteit.

Customer Intelligence (CI) wordt gedefinieerd als *het process van het ontginnen van informatie uit een klantendatabase door middel van state-of-the-art (statistische) technieken om uiteindelijk te komen tot de creatie van klantenkennis*. Binnen het CI domein vinden we zes belangrijke deeldomeinen terug waaronder *customer acquisition* (i.e. het proces van het aantrekken van prospecten met als doel de klantendatabase uit te breiden), *cross-selling* (i.e. de taak van het verkopen van additionele produkten/services bij bestaande klanten met als doel hun produkt

portfolio uit te breiden), *up-selling* (i.e. de taak van het uitbreiden/vervangen van bestaande produkten binnen het produkten palet door middel van meer kwaliteitsvolle produkten), *customer lifetime value* modelleren (i.e. het proces waarbij de huidige waarde van een klant zijn toekomstige cash flows wordt gemodelleerd), *churn modeling* (i.e. het proces waarbij de marketing analyst tracht te identificeren of een klant al dan niet het bedrijf zal verlaten) en *customer re-activitation modeling* (i.e. het activeringsproces van 'slapende' klanten of het terugwinningsproces van ex-klanten).

Het opzetten van een CI strategie binnen een bedrijf is moeilijk en in veel gevallen moeten dezelfde uitdagingen tot een goed eind worden gebracht. Deze uitdagingen liggen vervat in vier fases: *noise*, *data*, *knowledge* and *genius*.

Fase 1: *Noise*. Marketing analysten erkennen dikwijls dat er heel wat informatie beschikbaar is over de klanten die ze onderzoeken, maar dat deze informatie opgeslagen is in veel verschillende formaten zoals een Oracle database, Excel files, enquêtes... en dit op verschillende plaatsen binnen de organisatie zoals het call center, het analytische marketing departement, het marktonderzoek departement, het IT departement... . Het identificeren van deze verschillende bronnen is vaak een tijdrovende bezigheid.

Fase 2: *Data*. Eens de verschillende bronnen geïdentifieerd zijn, trachten vele marketing analysten deze waaier aan informatie te intergreren in een geaggregeerd raamwerk, die we een datawarehouse noemen. Binnen deze fase is het cruciaal ervoor te zorgen dat de kwaliteit van de data wordt gewaarborgd, daar de data het startpunt is van iedere analyse binnen CI.

Fase 3: *Knowledge*. De klanten analyst kan de data die beschikbaar is via de interne database gaan gebruiken als hefboom met als doel diepere inzichten te verschaffen in bepaalde klanten fenomenen. Met andere woorden, in deze fase wordt informatie geconverteerd naar kennis. Analyses binnen het CI domein kunnen worden opgesplitst in twee groepen: descriptieve en predictieve analyses. Onder descriptieve analyses verstaan we het beantwoorden van adhoc marketing vragen, klanten profileringen, klanten segmentatie..., terwijl de predictieve kant van CI focust op het voorspellen van bepaalde database informatie op basis van vroeger klantengedrag.

Heden ten dage zitten reeds heel wat bedrijven in deze derde fase, i.e. traditionele CI. Maar om u als marketing analyst te onderscheiden en om daadwerkelijk een klantengenie te worden, dient men zich te focussen op creativiteit. Aldus zijn we aanbeland bij Fase 4.

Fase 4: *Genius*. De laatste fase is het topje van de ijsberg. In deze fase is het mogelijk zich te onderscheiden van traditionele marketing analysten via data en methode creativiteit. Hierbij wordt data creativiteit gedefinieerd als het proces van het incorporeren van nieuwe informatie types of het combineren van transactionele database informatie (i.e. 'harde' data) met 'zachte' data, bv. enquête data.

Deze dissertatie bevat vijf studies die bijdragen tot data en/of methode creativiteit in het CI domein.

Studie 1 toont aan dat Support Vector Machines (SVMs) in staat zijn churn te gaan voorspellen in een abonnementscontext. Bijkomend worden er twee parameter-selectie technieken gebaseerd op grid-search en cross-validatie vergeleken. Er wordt aangetoond dat SVMs in combinatie met de juiste parameter-selectie techniek een volwaardig alternatief vormen voor Logistische Regressie. Nochtans dient er vastgesteld te worden dat Random Forests de andere twee classificatie technieken, SVMs en Logistische Regressie, volledig overstijgt in termen van predictieve kracht in deze onderzoekscontext. In termen van meest belangrijke churn indicatoren stellen we vast dat variabelen die het abonnement beschrijven het meeste impact hebben op het churn gedrag. In tegenstelling tot vorige studies, vinden we dat de monetaire waarde en de frequentie van abonnementshernieuwing niet voorkomen in de top tien meest voorspellende variabelen. Bovendien stellen we vast dat verschillende klant/bedrijfsinteractie variabelen een belangrijke rol spelen in het voorspellen van churn. Naast de belangrijkheid van leeftijd, stellen we vast dat socio-demografische variabelen geen rol spelen in het voorspellen van churn.

Studie 2 bewijst dat de integratie van de informatie uit call center emails in een conventioneel churn model hoogst efficient is. Een churn model dat beide types informatie combineert – i.e. traditionele, gestructureerde informatie van de marketing database en ongestructureerde, textuele informatie van de call center emails – superieur is aan het model met alleen marketing gerelateerde informatie. Hierdoor zijn marketing managers in staat de effectiviteit van hun marketing campagnes te verbeteren door de stijgende kracht in predictieve performantie.

Dat een automatisch email-classificatie systeem dat klachten emails onderscheidt van andere emails een robuuste strategie is binnen een call center wordt aangetoond in Studie 3. Het resultaat is dat email classificatie minder tijd in beslag neemt en minder duur is dankzij de lagere arbeidkosten. Bijkomend volgen de klachten van in het begin een apart traject waardoor dit moet resulteren in een meer succesvolle en vluggere afhandeling van de klachten. Bijkomend wordt er aangetoond dat het incorporeren van linguistieke stijl karakteristieken de performantie van het email-classificatie systeem verhoogt. Bijkomend laat deze studie ook toe een grondige analyse te maken van de verschillen in linguistieke stijl tussen klachten en andere klant/bedrijfs interactie emails.

Studie 4 benadrukt dat de keuze van het algoritme en de diversiteit in klanteninformatie belangrijke facetten zijn om de predictieve performantie van een churn model te verbeteren. Deze studie toont aan dat Random Forests, in tegenstelling tot Logistische Regressie en SVMs, een ideale keuze zijn om de predictieve performantie te optimaliseren. Bijkomend wordt aangetoond dat emotie-gerelateerde indicatoren uit call center emails de predictieve kracht verbeteren. Hoe meer emoties – i.e. zowel positieve als negatieve emoties – er worden geuit in een email, hoe positiever de impact op churn. Tot slot wordt er vastgesteld dat er een kleinere kans bestaat dat een specifieke klant zijn abonnement zal opzeggen wanneer deze klant een hoger aandeel klachten heeft in zijn totaal emailpakket.

Studie 5 argumenteert om verschillende misclassificatiekosten te incorporeren in de evaluatie van het predictieve model om zodoende het beslissingsproces te optimaliseren. In een churn context is het immers zo dat het classificeren van churners als niet-churners duidelijk verschillende kosten inhoudt dan niet-churners te classificeren als churners. Deze studie geeft marketing analysten een inzicht welke niet-sample gebaseerde kosten-sensitieve algoritmes (i.e. Herlabelings technieken (met *DMECC* en *Metacost*), *Threshold adjusting* en *Weighting*) te gebruiken. Er wordt aangetoond dat de praktisch geörienteerde cut-off van 0.5 niet opportuun is om churners van niet-churners te onderscheiden wanneer er verschillende misclassificatiekosten zijn, terwijl *DMECC* de meest robuuste techniek is over alle kostenratio's binnen alle churn categorieën. Wanneer men hoge kostenratio's heeft, is er een duidelijke voorkeur voor *Weighting* in termen van totale misclassificatie kost. Wanneer men lage kostenratio's heeft, is *Threshold adjusting* de meest aangeraden techniek wanneer het churn niveau varieert van laag tot medium, terwijl *Metacost* goed presteert wanneer het churn niveau varieert van medium tot hoog.

OVERVIEW

1. INTRODUCTION

In the last decade, marketing professionals are facing new challenges to collect and integrate a maximum of information from their customers into a database [22]. This is caused by the numerous academic and business cases where marketing analysts show the high value of feeding a transactional database with the most relevant information. Nowadays, marketing analysts are stressing the long-term relationships with customers using a well-considered CRM strategy [25] and CRM process measure [21]. For instance, it is shown that selling a product to an existing customer is less expensive than to selling the same product to a new customer [23]. So as well in academics as in business, the transactional database has become the starting point for a lot of marketing related research. This is also the case in Customer Intelligence. This doctoral dissertation introduces the reader to the field of Customer Intelligence and focuses on how the marketing analyst is able to improve marketing decision making by means of data and/or method creativity.

Section 2 defines Customer Intelligence and supplies the reader with a non-exhaustive overview of the main areas within the Customer Intelligence domain. Section 3 gives an overview of the Customer Intelligence pyramid which guides the reader through the different challenges of setting

up a Customer Intelligence strategy. This section leads to the core of this dissertation where one tries to unlock the last layer by stressing data and/or method creativity. The next section describes the research objectives of the different studies, where each of them contributes to data- and/or method creativity to improve traditional marketing analyses. Section 5 summarizes the main findings of the different studies, while the last Section relates to the limitations and suggestions for further research.

2. CUSTOMER INTELLIGENCE

The stream of describing and predicting customer behavior on an individual level based on transactional information has led to a new domain called Customer Intelligence or CI. Customer Intelligence is defined as

the process of exploiting information from a company's customer database by means of state-of-the-art (statistical) techniques in order to create customer knowledge.

Below you will find a non-exhaustive list of six main areas within the CI domain:

- Customer acquisition is the process of managing customer prospects and inquiries in order to extend the current customer base (e.g. [31]).
- Cross-selling is the task of selling additional products or services to existing customers in order to extend their current product portfolio (e.g. [18]).
- Up-selling is the task of extending/replacing existing products in a customer's product palette by an upscale extension/version (e.g. [16]).
- Customer lifetime value modeling is the process of estimating the present value of future cash flows that a customer will generate according to the customer's relationship (e.g. [3]).
- Churn modeling is the process of identifying whether or not a customer will leave the company (e.g. [24]).
- Re-activitation modeling is the process of activating "sleeping" customers or recapturing lost customers out of the database (e.g. [9];[28]).

3. THE CUSTOMER INTELLIGENCE PYRAMID

Nowadays, setting up a CI strategy within an organization is difficult and often one faces the same kind of challenges to tackle. These challenges are summarized into a CI pyramid with four layers of which one is locked (see Figure 1).

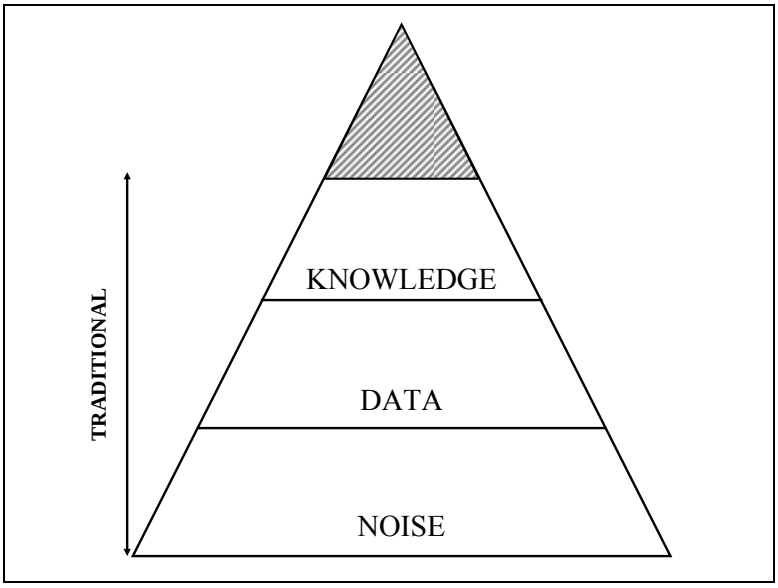


Figure 1: the CI pyramid.

Layer 1: Noise. Marketing analysts often acknowledge the fact that they have diverse and valuable information, but that this information is stored in different formats like Oracle database, excell files, access files, surveys… at very different locations within the organization like the call center, the analytical marketing department, the market research department, the IT department… Identifying the correct source and location of a specific kind of information is a very time-consuming task.

Layer 2: Data. In the second phase of setting up a CI strategy, marketing analysts need to integrate this diverse set of information into an aggregated framework. Several authors (e.g. [11]) stress the advantages of a profound data warehousing system, which integrates the different types of information. Data quality is a very important issue in this phase, because it is the starting point for every profound analysis within CI ([2]).

Layer 3: Knowledge. Based on the data of the internal database which contains detailed customer's information, the customer analyst is able to leverage this information which results in customer knowledge. Analyses within the CI domain are divided in two main categories: descriptive and predictive analyses. Descriptive CI analyses exist of answering ad hoc marketing questions, customer profiling, customer segmentation (e.g. [1])…., while the predictive part of CI focuses on forecasting events based on database information. Examples are predicting whether or not a customer will leave the company, whether or not a customer will buy a certain product…

A lot of marketing departments already passed through these three layers and this is what I call a traditional CI strategy. But to really unlock the top of the CI pyramid and to really become a customer genius, the CI analyst needs to disassociate from traditional CI analyses through creativity (see Figure 2).

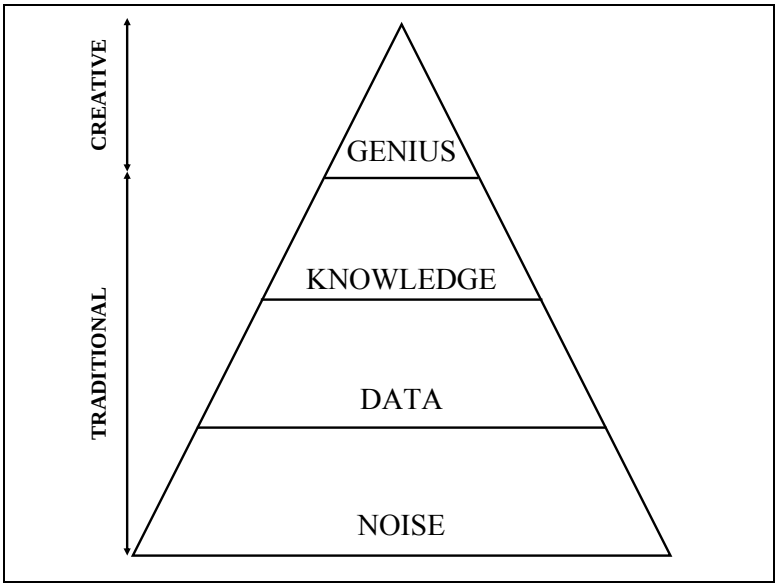


Figure 2: the CI pyramid.

How is one able to become a customer genius? Creativity is stimulated on two facets of the CI analysis: the data and method side. Below a non-exhaustive list is given on how one is able to improve marketing decision making through data and method creativity.

Data creativity is defined as the process of leveraging traditional analyses by incorporating new information types, i.e. *variable creation* or by combining transactional database information with information from surveys, i.e. *data augmentation*. Nowadays, analysts are able to leverage their outcomes via the exploration of new information types (i.e. *variable creation*). For instance, several analysts incorporate new information types like web users navigation information (e.g. [17]) into their daily analysis. A second way in which the data analyst can increase the power of its analyses is via *data augmentation*. 'Hard' data from the customers' database is often combined with 'soft' data from surveys. For instance, [4] were able to predict the loyalty score, which was gathered via customer surveys, using information in the transactional database.

Next to data creativity, analysts can incorporate some creativity in the methodology they apply. They can be creative in the *preprocessing*, the *model building* and the *model evaluation* phase.

During the *preprocessing* phase, analysts can search for other alternatives for missing value imputation. Nowadays, much researchers are using the mean as ideal missing value impute, while more sophisticated imputation methods (e.g. [10]) often yield greater savings. Moreover, opportunities arise to transform variables by mapping categorical variables and standardizing or discretizing continuous variables ([7]) or by taking the log, ln, squared root… in order to allow non-linear effects into a linear model. Finally, it is important that the prediction model is built on a sample which is representative for the true population. Avoiding sample bias often leads to better predictions (e.g. [29]).

During the *model building* phase, algorithms like Logistic Regression, Decision Trees, Neural Networks… are often used in the CI domain. However, advanced algorithms like Random Forests have proven to be excellent algorithms in various applications in the CI domain (e.g. [14]). Besides the algorithm choice, the search for the optimal set of predictors is critical for marketing analysts who want to rely on the key drivers of consumer response for optimizing their marketing decisions ([12]). Standard variable selection techniques like stepwise, forward and backward selection are often applied to finally end up with the ideal set of predictors. Various variable selection techniques building on filters and wrappers often end up with a different set of predictors which leads to increased performance (e.g. [13]).

A last opportunity lies in the *model evaluation* techniques. Nowadays, marketing models are often evaluated in terms of percentage correctly classified, area under the receiver operating curve and top-decile lift (e.g. [6]). However in many real life situations, different costs for false positives and false negatives occur ([30]). The standard evaluation measures do not take into account this cost, so one needs to find other alternatives for evaluating the performance of a binary classification model. This can be done by using the total misclassification cost.

This dissertation adds value to the current literature by focusing on innovative applications which lead to data and/or method creativity in the field of Customer Intelligence.

4. RESEARCH OBJECTIVES

This dissertation contains fives studies which contribute to the data- and/or method creativity into the CI field. Table 1 summarizes the different studies with their contributions to each of them. The following paragraphs describe the general research objectives of these 5 studies.

The main objective of Study 1 is to implement Support Vector Machines (SVMs) in a newspaper subscription context in order to construct a churn model with a higher predictive performance. Due to the fact that the implementation of SVMs highly depends on the parameter choice, we compare two parameter-selection techniques. Both techniques are based on grid-search and cross-validation, but differ because they take a different evaluation measure into account. Both types of SVM models are benchmarked to Logistic Regression and Random Forests. Furthermore, an overview of the most important churn drivers within this subscription context is given. The churn drivers are divided into four categories, namely client/company-interaction variables, renewal-related variables, socio–demographic variables and subscription-describing variables.

The purpose of Study 2 is to optimize the performance of a decision support system for churn prediction. More specifically, this study investigates the beneficial effect of adding the voice of customers through call center emails – i.e. textual information - to a churn prediction system that only uses traditional marketing information. By means of text mining techniques the call center emails are converted into a form which is more suitable for subsequent processing. In summary, this study investigates whether adding unstructured, textual information into a conventional churn prediction model results in a significant increase in predictive performance.

Study	Title	Contributions	
		Data creativity	**Method Creativity**
1	Churn Prediction in Subscription Services: an Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques[1]		• Implementing and contrasting SVMs with Logistic Regression and Random Forests • Comparing two alternative SVMs parameter selection techniques
2	Integrating the Voice of Customers through Call Center Emails into a Decision Support System for Churn Prediction[2]	• Incorporating textual information of client/company interaction emails	
3	Improving Customer Complaint Management by Automatic Email Classification Using Linguistic Style Features as Predictors[3]	• Incorporating linguistic style information of client/company interaction emails	• Generating an automatic email classification system by applying text mining methodology
4	Improving Customer Attrition Prediction by Integrating Emotions from Client/Company Interaction Emails and Evaluating Multiple Classifiers[4]	• Incorporating emotionality information of client/company interaction emails	• Predictive algorithm comparison between Logistic Regression, SVMs and Random Forests
5	Empirical Generalizations in Non-Sampling Based Cost-Sensitive Learning for Optimal Decision Making in a Churn Prediction Context[5]		• Model evaluation based on total misclassification cost using non-sampling based cost-sensitive learners

[1] Published in Expert Systems with Applications (2008)
[2] Published in Information and Management (2008)
[3] Published in Decision Support Systems (2008)
[4] Forthcoming Expert Systems with Applications (2009)
[5] In review Decision Support Systems

Table 1: an overview of the different studies related to their contributions to data- and method creativity.

In Study 3, I introduce a methodology to improve the complaint handling strategy through an automatic email classification system that automatically distinguishes complaints from non-complaints. Nowadays, companies daily receive huge amounts of emails as substitutes for traditional communication methods. We explicitly test if (i) automatic email classification into complaints and non-complaints is a viable strategy within a call center environment and (ii) adding new information about the linguistic style of an email to the traditional text mining variables significantly increases predictive performance.

Study 4 is directed towards the two aspects in which churn prediction models could be improved by (i) focusing on customer information diversity and (ii) choosing the most performing algorithm within the research context. This study examines whether the incorporation of emotionality related variables extracted from call center emails will influence the prediction power of the churn model, while on the other hand a comparison is made between Logistic Regression, SVMs and Random Forests in order to optimize the final prediction model.

In Study 5, I argue to take into account cost asymmetries in the evaluation of a churn prediction system. This study meta-analyses several non-sampling based cost-sensitive learners (i.e. Relabeling techniques (with Direct minimum expected cost criterion (DMECC) and Metacost), Threshold adjusting and Weighting) for optimizing the classification problem. Moreover, a sensitivity analysis on the churn incidence and cost ratio should lead to optimal managerial recommendations.

5. MAIN FINDINGS

Based on the research questions postulated, an overview of the main findings is given per study.

Study 1 indicates that SVMs are able to predict churn in subscription services. Moreover, two parameter-selection techniques based on grid-search and cross-validation are implemented and it shows that SVMs – in combination with the right parameter-selection technique – offer an alternative for Logistic Regression. However, if we compare predictive performance of state-of-the-art Random Forests, this study confirms that Random Forests outperforms SVMs and Logistic Regression in this research setting. Furthermore, this study shows that the most important churn predictors are variables describing the subscription. In contrast to previous findings, monetary value and frequency are not present in the top 10 most important churn

drivers, while on the other hand, several client/company-interaction variables play an important role in predicting churn like variables related to the ability of voluntarily suspending the subscription, the recency of complaining and the purchase motivator "own initiative". In spite of the importance of age, socio-demographics do not play an important role in explaining churn in this study.

Study 2 proves that the integration of the voice of customers by means of call center emails into a conventional churn-prediction system is highly beneficial. A churn-prediction system that combines both types of information - i.e. traditionally-used, structured information from the marketing database and unstructured information from call center emails - highly outperforms the model with only marketing related information. By augmenting the churn model with unstructured information from call center interaction emails, marketing managers may improve the effectiveness of their retention campaigns due to an increase in predictive performance of the churn prediction models applied.

Study 3 shows that an automatic email-classification system that distinguishes complaint emails from non-complaint emails is a feasible and robust strategy within a call center. As a result, email classification becomes less time-consuming and less expensive due to lower labor costs, while a separate complaint track for incoming emails should result in a more successful and faster complaint treatment. Moreover, this study proves that adding linguistic style features of an email increases predictive performance in differentiating between complaints and non-complaints. Furthermore, this study reveals the difference in linguistic style between complaints and non-complaints. It is clear that the likelihood of being a complaint is positively related with the number of words, the time indications, the present and past tense; while it decreases when more future tenses and question marks are used. The probability of classifying an incoming email as a complaint increases when the tone of the email becomes more antagonistic — i.e. it contains more negations, more numbers and more clenched words. Offending the company by using a lot of second person pronouns – e.g. "you are responsible for the misdelivery of the newspaper" – increases the chance of having a complaint.

Study 4 stresses that focusing on optimal classifier choice and customer information type diversity are important facets of improving the predictive performance of a churn model. This study confirms that Random Forests is a viable strategy to improve the predictive power of a churn system over Logistic Regression and Support Vector Machines. Moreover, the impact of emotionality indicators from call center emails on churn is investigated. It is shown that when

more emotional related words – i.e. positive as well as negative emotional words – are used in client/company interaction emails, the impact on churn is positive. Customers using more positive words in complaint emails are intuitively more satisfied, and they often do not have the intention to punish the company for the service failures which result in a positive effect on churn. Customers using more negative emotional words in emails seem to have a positive relation with churn. They tend to be more loyal. These results indicate that customers who write emotional emails (i.e. positive or negative) are of high value for the company. In other words, when customers heavily express their dissatisfaction for the company, this not necessarily means that these customers will churn. Indeed, this study found that the higher the portion of complaint emails, the lower the chance that this specific customer will churn.

Study 5 argues to incorporate different misclassification costs to optimize the decision making process in a churn prediction context. Categorizing churners as non-churners has clearly different costs than classifying non-churners as churners. This study shows evidence when to use which non-sampling based cost-sensitive learner (i.e. Relabeling techniques (with DMECC and Metacost), Threshold adjusting and Weighting). It is shown that the practically-oriented threshold of 0.5 is not opportune to classify customers into churners and non-churners when different misclassification costs occur, while the DMECC is the most robust technique for all cost ratios within all churn incidences. When cost ratios are high, Weighting performs best in terms of total misclassification cost. When cost ratios are low, Threshold adjusting is the most opportune technique in situations with churn level low to medium, while Metacost performs well when churn level ranges from medium to high.

6. SHORTCOMINGS & FURTHER RESEARCH

Despite the fact the five studies significantly contribute to the corresponding literature, several shortcomings and ample opportunities arise for further research. While the first four studies focus on testing a certain methodology in a single application domain, I strongly belief that the need for studies across different real-life settings (like Study 5) is high. Consequently, one will be able to determine the internal validity of the proposed methodology for a specific research problem. Moreover, I propose to test the methodological frameworks in other than the currently applied domains of Customer Intelligence. As a consequence, the analyst will get an idea of the external validity of the proposed framework. Furthermore, I was not able to include all existing algorithms in my research papers (i.e. algorithm comparison constraint). For instance in Study 1, I only focused on the frequently used algorithms in marketing (i.e. Logistic Regression and Random Forests) to compare the SVMs with. In Study 4, it would have been beneficial to adapt ensemble methods for a multi-class setting to a binary classification setting (see [19] or [20]) in order to allow for a more fair comparison between the single classifiers and Random Forests which is an ensemble of decision trees. Another point for further research would be to test the stability of the variable importance measures from Random Forests in Study 1, because these measures are used to rank the importance of the prediction variables. However, recent developments in literature show that Breiman's traditional importance measures become less stable in situations with (i) highly correlated predictors [26] and (ii) varying scale in measurement or number of categories [27]. Moreover, several opportunities arise to include the timing aspect of sending an email to the company in the current text mining framework of Study 2. Consequently, time-related information is available to the data analyst. Additionally, I did not use other than singular value decomposition to reduce the dimension of the term-by-document matrix (i.e. dimension reduction technique constraint). A valuable approach in the text mining methodology should be to test other dimension reduction techniques like non-negative matrix factorization [15] or sparse principal component analysis [32]. Another point for further research would be to embed the current results into a more conceptual framework, e.g. the attribution theory of Fritz Heider [8] which searches for the locus of the problem could be used to explain the differences in writing style between complaints and non-complaints (Study 3) or it would be interesting to have a look at the customer engagement theory to explain the results of Study 4. Furthermore, an in-depth field experiment should be set up to test the proposed methodologies in a real-life setting. Consequently, one will be able to truly evaluate the effectives of the proposed methodologies.

REFERENCES

[1] R.L. Andrews and I.S. Currim, A Comparison of Segment Retention Criteria for Finite Mixture Logit Models, Journal of Marketing Research, (40) 2 (2003), 235-243.

[2] B. Baesens, It's the Data, You Stupid!, Data News (2007).

[3] S. Borle, S.S. Singh and D.C. Jain, Customer Lifetime Value Measurement, Management Science 54 (1) (2008), 100-112.

[4] W. Buckinx, G. Verstraeten and D. Van den Poel, Predicting Customer Loyalty Using the Internal Transactional Database, Expert Systems with Applications, 32 (1) (2007), 125-134.

[5] P. Buhlmann and B. Yu, Boosting with the L2-Loss: Classification and Regression, Journal of the American Statistical Association, 98 (1) (2003), pp 324-339.

[6] J. Burez and D. Van den Poel, CRM at a Pay-TV Company: Using Analytical Models to Reduce Customer Attrition by Targeted Marketing for Subscription Services, Expert Systems with Applications, 32 (2) (2007), 277-288.

[7] S.F. Crone, S. Lessmann and R. Stahlbock, The Impact of Preprocessing on Data Mining: an Evaluation of Classifier Sensitivity in Direct Marketing, European Journal of Operational Research, 173 (3) (2006), 781-800.

[8] F. Heider, The Psychology of Interpersonal Relations, New York: John Wiley & Sons (1958).

[9] C. Homburg , W.D. Hoyer and R.M. Stock, How to Get Lost Customers Back?, Journal of the Academy of Marketing Science, 35 (4) (2007), 461-474.

[10] A. Karmaker and S. Kwek, An Iterative Refinement Approach for Data Cleaning, Intelligent Data Analysis 11 (5) (2007), 547-560.

[11] S. Kelly, Data Warehousing: The Route to Mass Customization, Wiley, New York (1996).

[12] Y.S. Kim, Toward a successful CRM: Variable Selection, Sampling and Ensemble, Decision Support Systems, 41 (2) (2006), 542-553.

[13] R. Kohavi and G.H. John, Wrappers for Feature Subset Selection, Artificial Intelligence, 97 (1–2) (1997), 273–324.

[14] B. Lariviere and D. Van den Poel, Predicting Customer Retention and Profitability by Using Random Forest and Regression Forest Techniques, Expert Systems with Applications, 29 (2) (2005), 472-484.

[15] D.D. Lee and H.S. Seung, Learning the Parts of Objects by Non-negative Matrix Factorization, Nature 401 (6755), pp 788-791.

[16] E.M. Okada, Upgrades and New Purchases, Journal of Marketing, 70 (4) (2006), 92-102.

[17] G. Pallis, L. Angelis and A. Vakali, Validation and Interpretation of Web Users' Sessions Clusters, Information Processing and Management, 43 (5) (2007), 1348-1367.

[18] A. Prinzie and D. Van den Poel, Predicting Home-Appliance Acquisition Sequences: Markov/Markov for Discrimination and Survival Analysis for Modeling Sequential Information in NPTB, Decision Support Systems 44 (1) (2007), 28-45.

[19] A. Prinzie and D. Van den Poel, Random Multiclass Classification: Generalizing Random Forests to Random MNL and Random NB. Lecture Notes in Computer Science, 4653 (DEXA 2007), 349-358.

[20] A. Prinzie and D. Van den Poel (2008), Random Forests for Multiclass Classification: Random MultiNomial Logit. Expert Systems with Applications, 34(3) (2008), 1721-1732.

[21] W. Reinartz, M. Krafft, M. and W.D. Hoyer, The Customer Relationship Management Process: Its Measurement and Impact on Performance, Journal of Marketing Research, 41 (3) (2004), 293-305.

[22] V. Nagar and M.V. Rajan, Measuring Customer Relationships: The Case of the Retail Banking Industry, Management Science, 51 (6) (2005).

[23] R.T. Rust and A.J. Zahorik, Customer Satisfaction, Customer Retention, and Market Share, Journal of Retailing, 69 (2) (1993), 193-215.

[24] D.A. Schweidel, P.S. Fader and E.T. Bradlow, Understanding Service Retention Within and Across Cohorts Using Limited Information, Journal of Marketing, 72 (1) (2008), 82-94.

[25] M.J. Shaw, C. Subramaniam, G.W. Tan and M.E. Welge, Knowledge Management and Data Mining for Marketing, Decision Support Systems, 31 (2001), 127-137.

[26] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin and A. Zeileis, Conditional Variable Importance for Random Forests, BMC Bioinformatics 9 (307) (2008).

[27] C. Strobl, A.-L. Boulesteix, A. Zeileis and T. Hothorn, Bias in Random Forests Variable Importance Measures: Illustrations, Sources and a Solution, BMC, Bioinformatics 8 (25) (2007).

[28] J.S. Thomas, R.C. Blattberg and E.J. Fox, Recapturing Lost Customers, Journal of Marketing Research, 41 (1) (2004), 31-45.

[29] G. Verstraeten and D. Van den Poel, The Impact of Sample Bias on Consumer Credit Scoring Performance and Profitability, Journal of the Operational Research Society, 56 (2005), 981-992.

[30] S. Viaene and D. Dedene, Cost-Sensitive Learning and Decision Making Revisited, European Journal of Operational Research, 166 (2005).

[31] J. Villanueva, S. Yoo and D.M. Hanssens, The Impact of Marketing-Induced Versus Word-of-Mouth Customer Acquisition on Customer Equity Growth, Journal of Marketing Research, 45 (1) (2008), 48-59.

[32] H. Zou, T. Hastie and R. Tibshirani, Sparse Principal Component Analysis, Journal of Computational and Graphical Statistics, 15 (2004).

CHAPTER I

CHURN PREDICTION IN SUBSCRIPTION SERVICES: AN APPLICATION OF SUPPORT VECTOR MACHINES WHILE COMPARING TWO PARAMETER-SELECTION TECHNIQUES

This chapter is based on K. Coussement and D. Van den Poel, Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques, Expert systems with applications, 34 (1) (2008), pp 313-327.

CHAPTER I
CHURN PREDICTION IN SUBSCRIPTION SERVICES: AN APPLICATION OF SUPPORT VECTOR MACHINES WHILE COMPARING TWO PARAMETER-SELECTION TECHNIQUES

ABSTRACT

CRM gains increasing importance due to intensive competition and saturated markets. With the purpose of retaining customers, academics as well as practitioners find it crucial to build a churn prediction model that is as accurate as possible. This study applies support vector machines in a newspaper subscription context in order to construct a churn model with a higher predictive performance. Moreover, a comparison is made between two parameter-selection techniques, needed to implement support vector machines. Both techniques are based on grid search and cross-validation. Afterwards, the predictive performance of both kinds of support vector machine models is benchmarked to logistic regression and random forests. Our study shows that support vector machines show good generalization performance when applied to noisy marketing data. Nevertheless, the parameter optimization procedure plays an important role in the predictive performance. We show that only when the optimal parameter selection procedure is applied, support vector machines outperform traditional logistic regression, whereas random forests outperform both kinds of support vector machines. As a substantive contribution, an overview of the most important churn drivers is given. Unlike ample research, monetary value and frequency do not play an important role in explaining churn in this subscription-services application. Even though most important churn predictors belong to the category of variables describing the subscription, the influence of several client/company-interaction variables can not be neglected.

1. INTRODUCTION

Nowadays, more and more companies start to focus on Customer Relationship Management, CRM. Indeed due to saturated markets and intensive competition, a lot of companies do realize that their existing database is their most valuable asset [3,30,47]. This trend is also notable in

subscription services. Companies start to shift away from their traditional, mass marketing strategies, in favor of targeted marketing actions [10]. It is more profitable to keep and satisfy existing customers than to constantly attract new customers who are characterized by a high attrition rate [41]. The idea of identifying those customers most prone to switching carries a high priority [31]. It has been shown that a small change in retention rate can result in significant changes in contribution [48]. In order to effectively manage customer churn within a company, it is crucial to build an effective and accurate customer-churn model. To accomplish this, there are numerous predictive-modeling techniques available. These data-mining techniques can effectively assist with the selection of customers most prone to churn [29]. These techniques vary in terms of statistical technique (e.g. neural nets versus logistic regression), variable-selection method (e.g. theory versus stepwise selection), number of variables included in the model, time spent to build the final model, as well as in terms of allocating the time across the different tasks in the modeling process [39].

This study contributes to the existing literature by investigating the effectiveness of the support vector machines (SVMs) approach in detecting customer churn in subscription services. Ample research focuses on predicting customer churn in different industries, including investment products, insurance, electric utilities, health care providers, credit card providers, banking, internet service providers, telephone service providers, online services, … . Although SVMs have shown excellent generalization performance in a wide range of areas like bioinformatics [16,27,54], beat recognition [1], automatic face authentication [5], evaluation of consumer loans [36], estimating production values [15,40], text categorization [6], medical diagnosis [23], image classification [34] and hand-written digit recognition [11,17], the applications in marketing are rather scarce [18].

To our knowledge only a few implementations of SVMs in a customer churn environment are published [33,53]. This study will extend the use of SVMs in a customer-churn context in two ways: (1) Unlike former studies that implemented SVMs on a very small sample, this study applies SVMs in a more realistic churn setting. Indeed, once a churn model has been built, it must be able to accurately validate a new marketing dataset which contains in practice ten thousands of records and often a lot of noise. This study contributes to the existing literature by using a sufficient sample size for training and validating the SVM models in a subscriber churn framework. These SVMs are benchmarked to logistic regression and state-of-the-art random forests. [39] concluded that logistic modeling may even outperform the more sophisticated techniques (like neural networks), while in a marketing setting random forests already proved to

be superior to other more traditional classification techniques [8,35]. (2) Before SVMs can be implemented, several parameters have to be optimized in order to construct a first-class classifier. Extracting the optimal parameters is crucial when implementing SVMs [28,33]. Consequently, a fine-tuned parameter selection procedure has to be applied. [28] proposed a grid search and a cross-validation to extract the optimal parameters for SVMs. This procedure tries different parameter pairs on the training set using a cross-validation procedure. [28] propose to select that pair of parameters with the best cross-validation accuracy - i.e. percentage of cases correctly classified (PCC). The second contribution of this study lies in extending this principle by selecting one additional parameter pair. Not only the parameters with the best cross-validation accuracy – as proposed by [28] - are selected, also the parameter pair which results in the highest cross-validation area under the receiver operating curve (AUC) is used. In contrast to PCC, AUC takes into account the individual class performance – by use of the sensitivity and specificity - for several thresholds on the classifier's posterior churn probabilities [22,45,44]. In the end, it is possible to compare the predictive performance of these two parameter-selection techniques with that of logistic regression and random forests.

As a substantive contribution, an overview of the most important churn predictors is given within this subscription-services setting. As such, marketing managers gain insight into which predictors are important in identifying churn. Consequently, it may be possible to adapt their marketing strategies based on this newly obtained information.

Following an introduction of the modeling techniques (i.e. SVMs, random forests and logistic regression), Section 3 explains the evaluation measures used in this study. The model-selection procedure for SVMs is presented in Section 4. Section 5 presents the research data, while Section 6 explains the experimental results. Conclusions and directions for future research are given in Section 7.

2. MODELING TECHNIQUES

2.1. Support Vector Machines

The SVM approach is a novel classification technique based on neural network technology using statistical learning theory [49,50]. In a binary classification context, SVMs try to find a linear optimal hyperplane so that the margin of separation between the positive and the negative examples is maximized. This is equivalent to solving a quadratic optimization problem in which

only the support vectors, i.e. the data points closest to the optimal hyperplane, play a crucial role. However, in practice, the data is often not linearly separable. In order to enhance the feasibility of linear separation, one may transform the input space via a non-linear mapping into a higher dimensional feature space. This transformation is done by using a kernel function. There are some advantages in using SVMs [33]: (1) there are only two free parameters to be chosen, namely the upper bound and the kernel parameter, (2) the solution of SVM is unique, optimal and global since the training of a SVM is done by solving a linearly constrained quadratic problem, (3) SVMs are based on the Structural Risk Minimization (SRM) principle, which means that this type of classifier minimizes the upper bound on the actual risk, compared to other classifiers which minimize the empirical risk. This results in a very good generalization performance.

We will give a general overview of a SVM for a binary classification problem. For more details about SVMs, we refer to the tutorial of [12].

Given a set of labeled training examples $\{x_i,y_i\}$ with i= 1, 2, 3, …, N where $y_i \in \{-1,1\}$ and $x_i \in R^n$, and n the dimension of the input space. Suppose that the training data is linearly separable, there exists a weight vector w and a bias b such that the inequalities

$$w \,.\, x_i + b \geq 1 \text{ when } y_i = 1, \tag{1}$$

$$w \,.\, x_i + b \leq -1 \text{ when } y_i = -1, \tag{2}$$

are valid for all elements of the training set. As such, we can rewrite these inequalities in the form:

$$y_i \,(w \,.\, x_i + b) \geq 1 \text{ with } _i = 1, 2, 3, \ldots, N. \tag{3}$$

Eq (3) comes down to find two parallel boundaries,

$$\text{B1: } w \,.\, x_i + b = 1, \tag{4}$$

$$\text{B2: } w \,.\, x_i + b = -1, \tag{5}$$

at the opposite sides of the optimal separating hyperplane,

$$\text{H*: } w \,.\, x + b = 0, \tag{6}$$

with margin width between the two boundaries equal to *2/||w||*. Thus one can find the pair of boundaries which gives the maximum margin by:
minimizing

$$\frac{1}{2}.w^2 \tag{7}$$

subject to

$$y_i\ (w\ .\ x_i + b) \geq 1 \tag{8}$$

This constrained optimization problem can be solved using the characteristics of the Lagrange multipliers (α) by

maximizing

$$W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j \tag{9}$$

subject to

$$\alpha_i \geq 0 \text{ with } i = 1, 2, 3, \ldots, \text{N and} \sum_i \alpha_i y_i = 0 \tag{10}$$

The weight vector could be stated as follows:

$$w = \sum_i \alpha_i y_i x_i \tag{11}$$

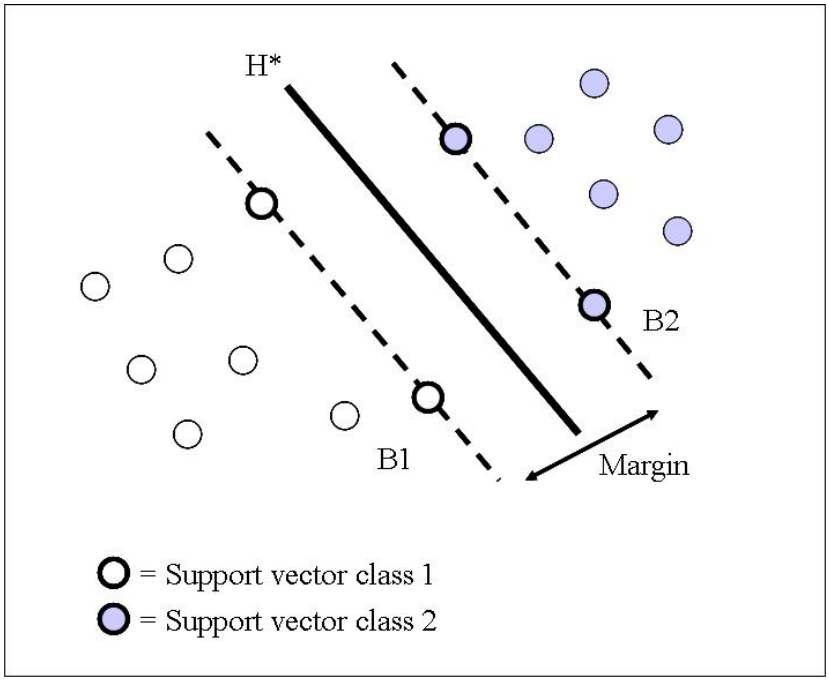


Figure 1: This figure shows the solution for a binary linearly separable classification problem. The boundaries B1 and B2 separate the two classes. Data points on the boundaries are called support vectors. Thus one tries to find the hyperplane H* where the margin is maximal.

The decision function *f(x)* can be written as

$$f(x) = \text{sgn}\ (w \,.\, x + b) \ = \ \text{sgn}\left[\sum_i \alpha_i y_i (x.x_i) + b\right] \tag{12}$$

where sgn is a sign function. In practice, the input data will often not be linearly separable. However, one can still implement a linear model by introducing a higher dimensional feature space to which an input vector is mapped via a non-linear transformation:

$$\Theta : X \rightarrow X' \tag{13}$$

$$x_i \rightarrow \Theta(x_i) \tag{14}$$

where X is the input space, Θ is the non-linear transformation and $\Theta(x_i)$ represents the value of x_i mapped into the higher dimensional feature space X'.

Therefore Equation (9) can be transformed to

$$W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \Theta(x_i)\Theta(x_j) \quad (15)$$

subject to

$$\alpha_i \geq 0 \text{ with } i = 1, 2, 3, \ldots, \text{N and } \sum_i \alpha_i y_i = 0 \quad (16)$$

By mapping the input space into a higher dimensional feature space, the problem of high dimensionality and implementation complexity occurs. One can introduce the concept of inner product kernels. Consequently, there is no more need to know the exact value of $\Theta(x_i)$, only the dot inner product is considered which facilitates the implementation.

$$K(x_i, x_j) = \Theta(x_i).\Theta(x_j) \quad (17)$$

Therefore the decision function becomes

$$f(x) = \text{sgn}\left[\sum_i \alpha_i y_i \Theta(x).\Theta(.x_i) + b\right] = \text{sgn}\left[\sum_i \alpha_i y_i K(x.x_i) + b\right] \quad (18)$$

For resolving this decision function, several types of kernel functions are available as given in Table 1.

Kernel function	Mathematical form*
Linear Kernel	$K(x,x_i) = (x \,.\, x_i)$
Polynomial Kernel of degree d	$K(x,x_i) = (\gamma\, x \,.\, x_i + r)\ \hat{}\ d$
Radial Basis Function	$K(x,x_i) = \exp\{-\gamma \lVert x - x_i \rVert^2\}$
Sigmoid Kernel with r $\in$ N	$K(x,x_i) = \tanh(\gamma\, x \,.\, x_i + r)$
* d, r $\in$ N; $\gamma \in R^+$	

Table 1: Overview of the different kernel functions.

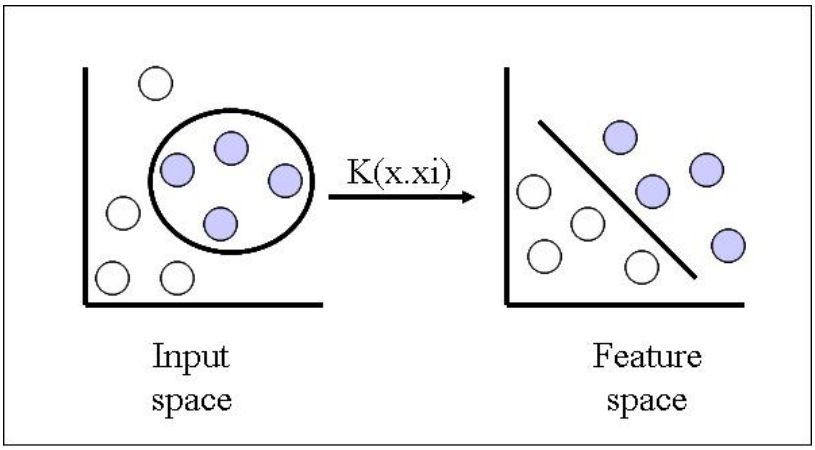


Figure 2: The non-linear boundary in the input space is mapped via a kernel function into higher dimensional feature space. The data becomes linearly separable in the feature space.

It is possible to extend these ideas to handle non-separable data. In this case, the margin will become very small and it will be impossible to separate the data without any misclassification. To solve this problem, we relax the constraints (1) and (2) by introducing positive slack variables (ε) [17].

Equations (1) and (2) become

$$w \,.\, x_i + b \geq 1 - \varepsilon_i \quad \text{when } y_i = 1, \tag{19}$$

$$w \,.\, x_i + b \leq -1 + \varepsilon_i \quad \text{when } y_i = -1, \tag{20}$$

with $\varepsilon_i \geq 0$.

Equations (19) and (20) can be rewritten as

$$y_i \,(w \,.\, x_i + b) \geq 1 - \varepsilon_i \text{ with } i = 1, 2, 3, \ldots, \text{N}. \tag{21}$$

The goal of the optimization process is to find the hyperplane that maximizes the margin and minimizes the probability of misclassification:
minimize

$$\frac{1}{2}.w^2 + C\sum_i \varepsilon_i \tag{22}$$

subject to

$$y_i\ (w\ .\ x_i\ +\ b) \geq 1\ -\ \varepsilon_i \tag{23}$$

with *C*, the cost, the penalty parameter for the error term. The larger *C*, the higher the penalty to errors.

Adapting Equation (15) to the non-separable case, one receives the following optimization problem:
maximizing

$$W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{24}$$

subject to

$$0 \leq \alpha_i \leq C \text{ with } i = 1, 2, 3, \ldots, \text{N and} \sum_i \alpha_i y_i = 0 \tag{25}$$

More details concerning the optimization process can be found in [14].

2.2. Random Forests

In a binary classification context, Decision Trees (DT) became very popular because of their easiness and interpretability [21]. Moreover, DTs have the ability to handle covariates measured at different measurement levels. One major problem with DTs is their high instability [26]. A small change in the data often results in very different series of splits, which is often suboptimal when validating the trained model. In the past, this problem was extensively researched.

It was [7] who introduced a solution to the previously mentioned problem. The new classification technique is called: Random Forests. This technique uses a subset of *m* randomly chosen predictors to grow each tree on a bootstrap sample of the training data. Typically, this number of selected variables – i.e. *m* - is much lower than the total number of variables in the model. After a large number of trees is generated, each tree votes for the most popular class. By aggregating these votes over the different trees, each case is predicted a class label.

Random forests are already applied in several domains like bioinformatics, quantitative criminology, geology, pattern recognition, medicine, … . However, the applications in marketing are rare [8,35]. Random forests are used as benchmark in this study, mainly for five reasons: (1) [38] stated that the predictive performance is among the best of the available techniques. (2) The outcomes of the classifier are very robust to outliers and noise [7]. (3) This classifier outputs useful internal estimates of error, strength, correlation and variable importance [7]. (4) Reasonable computation time is observed by [8]. (5) Random forests are easy to implement because there are only two free parameters to be set, namely *m*, the number of randomly chosen predictors, and the total number of trees to be grown. We follow Breiman's [7] suggestions: *m* is set equal to the square root of the total number of variables - i.e. 9 because 82 explanatory variables are included in the model - and a large number of trees - i.e. 1000 - are chosen.

2.3. Logistic Regression

Logistic regression is a well-known classification technique for predicting a dichotomous dependent variable. In running a logistic regression analysis, the maximum likelihood function is produced and maximized in order to achieve an appropriate fit to the data [2]. This technique is very popular for mainly three reasons: (1) logit modeling is conceptually simple [9]. (2) A closed-form solution for the posterior probabilities is available (in contrary to SVMs); (3) It provides quick and robust results in comparison to other classification techniques [39].

3. EVALUATION CRITERIA

After building a predictive model, marketers want to use these classification models to predict future behavior. It is essential to evaluate the classifier in terms of performance. Firstly, the predictive model is estimated on a training set. Afterwards, this model is validated on an unseen dataset, the test set. It is essential to evaluate the performance on a test set, in order to ensure that the trained model is able to generalize well. For all three modeling techniques, PCC, AUC and the top-decile lift are calculated.

PCC, also known as accuracy, is undoubtedly the most commonly used evaluation metric of a classifier. Practically, the posterior churn probabilities generated by the classifier are ranked from most likely to churn to least likely to churn. All cases above a certain threshold are classified as churners; all cases having a lower churn probability are classified as non-churners.

In sum, PCC computes the ratio of correctly classified cases to the total number of cases to be classified.

It is important to notice that PCC is highly dependent on the chosen threshold because only one threshold is considered. Consequently, it does not give an indication how the performance will vary when the cut-off is varied. Moreover, PCC does not consider the individual class performance of a classifier. For example, within a skewed class distribution, wrong predictions for the underrepresented class are very costly. Nevertheless, a model that predicts always the most common class - thus neglecting the minority class- still provides a relatively good performance when evaluated on PCC.

Unlike PCC, AUC takes into account the individual class performance for all possible thresholds. In other words, AUC will compare the predicted class of an event with the real class of that event, considering all possible cut-off values for the predicted class. The receiver operating curve (ROC) is a graphical plot of the sensitivity - i.e. the number of true positives versus the total number of events - and 1-specificity - i.e. the number of true negatives versus the total number of non-events. The ROC can also be represented by plotting the fraction of true positives versus the fraction of false positives. The area under the receiver operating curve is used to evaluate the performance of a binary classification system [25]. In order to assess whether AUCs of the different classification techniques are significantly different from each other, the non-parametric test of [20] is used.

In marketing applications, one is especially interested in increasing the density of the real events. The top 10% decile is an evaluation measure that only focuses on the 10% cases most likely to churn. Practically, the cases are first sorted from predicted most likely to churn to predicted least likely to churn. Afterwards, the proportion of real events in the top 10% most likely to churn is compared with the proportion of real events in the total dataset. This increase in density is called the top-decile lift. For example, a top-decile lift of two means that the density of churners in the top 10% is twice the density of churners in the total dataset. The higher the top-decile lift, the better the classifier. Potentially this top-decile lift is very interesting to target, because it contains a higher number of real events. In other words, marketing analysts are interested in just 10% of the customer base – i.e. those who are most likely to churn -because marketing budgets are limited and actions to reduce churn would typically involve only 10% of the entire list of customers.

4. MODEL SELECTION FOR THE SUPPORT VECTOR MACHINES

First, we will argue why the radial basis function (RBF) kernel is used as the default kernel function throughout this study. Secondly, the grid-search method and cross-validation procedure for choosing the optimal penalty parameter C and kernel parameter γ is explained. In a third section, two types of parameter selection techniques are described.

4.1. RBF Kernel Function

The RBF kernel function is used as the default kernel function within this study, mainly for four reasons [28]: (1) this type of kernel makes it possible to map the non-linear boundaries of the input space into a higher dimensional feature space. So unlike the linear kernel, the RBF kernel can handle a non-linear relationship between the dependent and the explanatory variables. (2) In terms of performance [32] concluded that the linear kernel with a parameter C has the same performance as the RBF kernel with parameters (C, γ). [37] showed that the sigmoid kernel behaves like the RBF kernel for certain parameters. (3) When looking at the number of hyperparameters, the polynomial kernel has more hyperparameters than the RBF kernel. (4) The RBF kernel has less numerical difficulties because the kernel values lie between zero and one, while the polynomial kernel values may go to infinity or zero while the degree is large. On the basis of these arguments, the RBF kernel is used as the default kernel function.

4.2. Optimal Parameter Selection Using Grid Search and Cross-Validation

The RBF kernel needs two parameters to be set; C and γ, with C the penalty parameter for the error term and γ as the kernel parameter. Both parameters play a crucial role in the performance of SVMs [28,33]. Improper selection of these parameters can be counterproductive. Beforehand it is impossible to know which combination of (C, γ) will result in the highest performance when validating the trained SVM to unseen data. Some kind of parameter selection procedure has to be done. [28] propose a 'grid search' on C and γ and a *v*-fold cross-validation on the training data. The goal of this procedure is to identify the optimal C and γ, so that the classifier can accurately predict unseen data. A common way to accomplish this is 2-fold cross-validation, where the training set is divided into two parts of which one is unseen in training the classifier. This performance better reflects the capabilities of the classifier in validating unknown data. More generally, in a *v*-fold cross-validation, the training data is

split into *v* subsets of equal size. Iteratively, one part is left out for validation, while the other remaining (*v*-1) parts are used for training. Finally, each case in the training set is predicted once. The cross-validation performance will better reflect the 'true' performance as when validating the classifier to unseen data, while the validation set stays untouched. In order to identify which parameter pair performs best, one can repeat this procedure with several pairs of (C, γ). As such it is possible to calculate a cross-validated evaluation measure for every parameter pair. In the end, it is possible to select these parameters based on the best cross-validated performance.

4.3. Two Parameter Selection Techniques

In this study, a grid search on C and γ is performed on the training set using a 5-fold cross-validation. The grid search is realized by evaluating exponential sequences of C and γ (i.e. C = $2^{-5}, 2^{-3}, \ldots, 2^{13}$; $\gamma = 2^{3}, 2^{1}, \ldots, 2^{-15}$). Basically, all combinations of (C, γ) are tried and two pairs of parameters are restrained: (1) the one with the best cross-validated accuracy – as proposed by [28] - and (2) the one with the biggest cross-validated area under the receiver operating curve. This additional parameter pair is selected for the reason that unlike PCC, AUC considers the sensitivity and specificity as individual class performance metrics over all possible thresholds. Once these optimal parameter pairs are obtained, the whole training set is trained again. Both classifiers will be used to validate an unseen dataset. In the end, one can compare and benchmark the performance of both kinds of SVMs.

5. RESEARCH DATA

For the purpose of this study, data from a Belgian newspaper publishing company is used. The subscribers have to pay a fixed amount of money depending on the length of subscription and the promotional offer given. The company doesn't allow ending the subscription prior to the maturity date. The churn-prediction problem in this subscription context comes down to predicting whether the subscription will/will not be renewed within a period of four weeks after the maturity date. During this four-week period, the company still delivers the newspapers to the subscribers. In this way, the company gives the subscribers the opportunity to renew their subscription. Figure 3 graphically traces back the time window of analysis. We use subscription data from January 2002 through September 2005. Using this time frame, it is possible to derive the dependent variable and the explanatory variables. For constructing the dependent variable, the renewal points between July 2004 and July 2005 are considered. Consequently, a customer

is considered as "churner" when his/her subscription is not renewed within four weeks after the expiry date. The explanatory variables contain information covering a 30-month period returning from every individual renewal point. These variables contain information about client/company interactions, renewal-related information, socio-demographics and subscription-describing information (see Appendix A). This variety of information is gathered at two levels: subscription level and subscriber level. At the subscription level, all information from the current subscription is included, while at the subscriber level, all information related to the subscriber is covered. For instance, one can calculate the total number of complaints on the current subscription only - i.e. the subscription level - , while one can also consider the total number of complaints of a subscriber covering all his/her subscriptions - i.e. subscriber level. Finally, one ends up with an individual timeline per subscriber for every renewal point in the time interval.

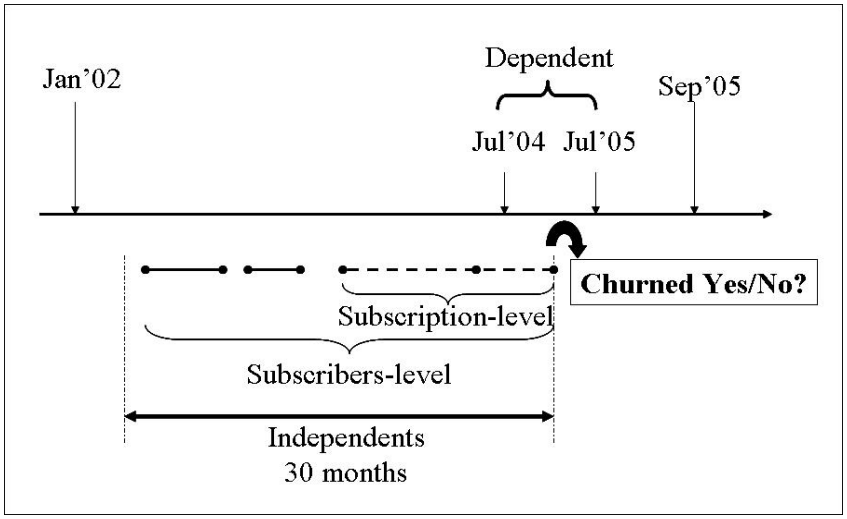


Figure 3: Graphical display of the time window used to build the churn model.

We decided to randomly select two samples of sufficient size; the training set is used to estimate the model, while the test set is used to validate the model. The training set contains as many churners as non-churners because many authors emphasize the need for a balanced training sample in order to reliably differentiate between defectors and non-defectors [19,43,52]. So it is not uncommon to train a model with a non-natural distribution [13,51]. The test set contains a proportion of churners that is representative for the true population in order to approximate the predictive performance in a real-life situation. For both datasets, all variables are constructed in the same way. The explanatory variables are compiled over a 30 month period, while the dependent variable contains information whether the subscription will/will not be renewed.

	Number of observations	Relative percentage
Training set		
Subscriptions not renewed	22500	50%
Subscriptions renewed	22500	50%
Total	*45000*	100%
Test set		
Subscriptions not renewed	5014	11,14%
Subscriptions renewed	39986	88,86%
Total	*45000*	100%

Table 2: Distribution of the training set and test set.

6. EMPIRICAL ANALYSIS

6.1. SVM models

After conducting the grid search on the training data, the optimal (C, γ) is $(2^{13},2^{-7})$ with a cross-validated accuracy of 78.089%. Table 3 summarizes the results of the grid search using the cross-validated accuracy as an evaluation criterion. Furthermore, parameter pair $(2^{7},2^{-7})$ results in the highest cross-validated AUC, 84.702. Table 4 considers the results of the grid-search procedure with the cross-validated AUC as a performance measure.

γ	C								
	2^{-5}	2^{-3}	2^{-1}	2^{1}	2^{3}	2^{5}	2^{7}	2^{9}	2^{13}
2^{3}	56.360	64.351	65.756	66.248	65.469	65.181	64.924	64.519	64.342
2^{1}	68.147	70.525	71.733	71.418	70.453	69.458	68.836	68.314	68.087
2^{-1}	75.353	76.582	77.127	76.262	74.627	72.958	71.947	70.859	70.394
2^{-3}	75.959	77.144	77.649	77.622	77.558	76.440	74.918	73.320	71.842
2^{-5}	74.789	76.164	76.960	77.471	78.039	77.996	77.924	78.056	76.396
2^{-7}	74.367	74.948	75.975	76.440	77.118	77.758	77.719	78.084	78.089
2^{-9}	75.163	74.349	74.827	75.907	76.167	76.693	76.959	77.722	77.726
2^{-11}	74.240	75.209	74.344	74.840	75.856	76.107	76.271	76.517	77.144
2^{-13}	54.767	74.213	75.198	74.403	74.836	75.860	76.093	76.202	76.398
2^{-15}	50.000	64.406	74.103	75.198	74.406	74.829	75.872	76.089	76.182

Table 3: The cross-validated accuracy per (C, γ).

γ	C								
	2^{-5}	2^{-3}	2^{-1}	2^{1}	2^{3}	2^{5}	2^{7}	2^{9}	2^{13}
2^{3}	75.710	76.201	76.083	75.279	74.283	73.669	73.273	72.918	72.610
2^{1}	80.059	80.221	80.092	78.600	77.058	75.878	75.007	74.441	74.085
2^{-1}	82.703	83.552	83.722	82.728	80.951	79.069	77.616	76.386	75.442
2^{-3}	83.865	84.296	84.507	84.472	83.857	82.406	80.500	78.402	76.388
2^{-5}	83.373	83.926	84.172	84.496	84.691	84.592	84.212	83.239	81.745
2^{-7}	82.871	83.188	83.670	83.896	84.172	84.514	84.702	84.699	84.477
2^{-9}	82.232	82.810	83.087	83.506	83.625	83.861	84.173	84.504	84.674
2^{-11}	80.998	82.229	82.790	83.059	83.448	83.462	83.593	83.869	84.190
2^{-13}	72.936	80.996	82.228	82.785	83.052	83.431	83.393	83.436	83.601
2^{-15}	50.000	72.987	80.995	82.228	82.784	83.051	83.427	83.377	83.375

Table 4: The cross-validated performance (AUC) per (C, γ)

These two parameters pairs are used to train a model on the complete training set. Two SVMs are obtained, namely SVMacc[1] and SVMauc[2]. Finally, both models can be validated on a test set.

On the one hand, one can compare the performance among both SVMs, while on the other hand both SVMs can be benchmarked with the performance of the logistic regression and random forests.

6.2. Comparing Predictive Performance among both kinds of SVMs

In this section, a comparison is made between the predictive performance of SVMacc and SVMauc. The evaluation is performed in terms of AUC, PCC and top-decile lift. Both models are trained on a balanced training set, while in the end these classifiers have to be evaluated on a dataset which represents the actual density of churners (see Table 2). In order to assess the sensitivity of the results to the actual proportion of churners in the dataset, we will compare the performance of both SVMs on artificial test sets with different class distributions. More specifically, we compare the 'natural' distribution[3] (11,14% churners) with the artificial ones (50%, 40%, 30%, 20%, 18%, 16%, 14%). These artificial sets are created by randomly undersampling the real test set – i.e. the one with 11,14% churners.

Figures 4 through 6 and Table 5 depict the performance of SVMacc and SVMauc for the different class distributions. As such a comparison can be made between both SVMs. As one

[1] SVMacc = SVM generated using parameters based on the model with the best cross-validated accuracy during grid search
[2] SVMauc = SVM generated using parameters based on the model with the best cross-validated AUC during grid search
[3] i.e. the distribution that contains the proportion of churners that is representative for the true population.

may observe from Figure 4, SVMauc performs better than SVMacc within all class distributions in terms of AUC performance. In order to ensure that the differences in AUC are significant, the test proposed by Delong et al. (1988) is applied. As such one can compare if the AUCs between SVMacc and SVMauc are significantly different within a certain class distribution. Table 5 reveals that on all test sets that contain 30% churners or less, SVMauc significantly outperforms SVMacc on a 90% confidence level [20]. When validated on the 'natural' distribution, SVMauc significantly outperforms SVMacc at the 95% confidence level. Figure 5 shows the performance of both SVMs in terms of PCC. Despite the fact that the differences in PCC are rather small, one may observe that SVMauc does not have an inferior performance compared to SVMacc when coming closer to the 'natural' distribution. Previous findings are confirmed when evaluating both SVMs using the top-decile lift. There is a gap in top-decile lift between SVMacc and SVMauc. SVMauc has a higher top-decile lift compared to SVMacc. This gap increases when deviating from the original training distribution – i.e. the one with 50% churners. On the 'natural' distribution, SVMauc succeeds in retaining more churners within the top 10% customer most likely to churn in comparison to SVMacc.

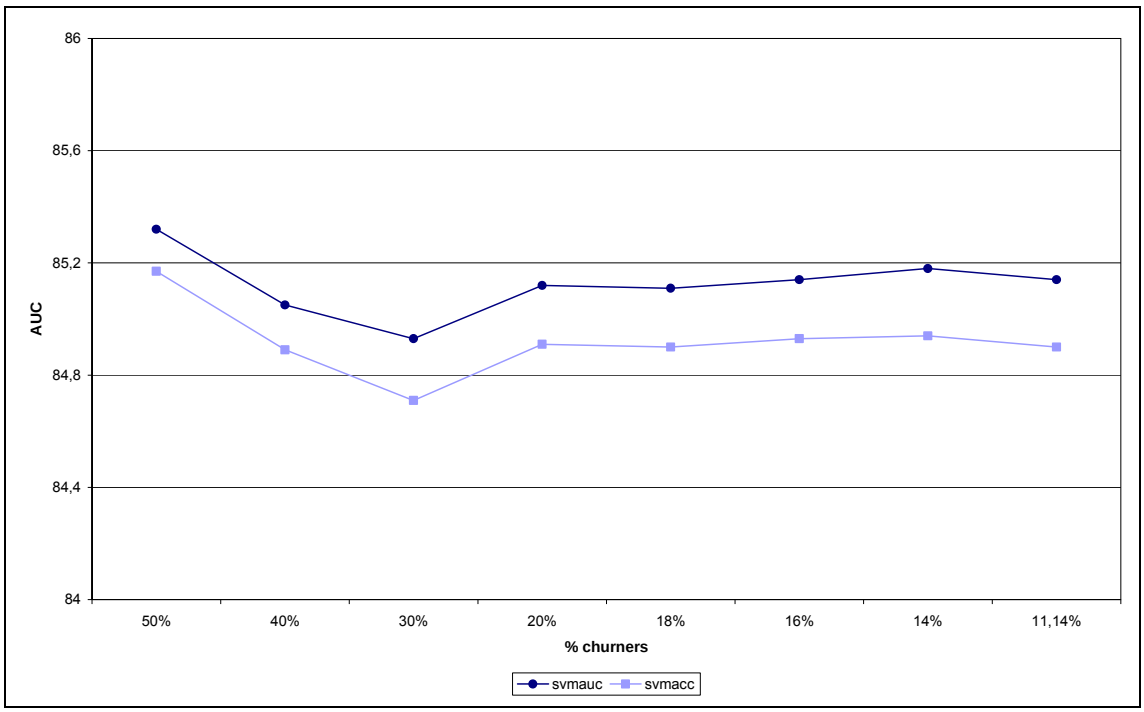


Figure 4: Area under the receiver operating curve for SVMacc and SVMauc applied to several test sets with different class distributions

Number of churners	SVMacc – SVMauc
50%	1,16 (1); 0,281
40%	1,57 (1); 0,210
30%	3,65 (1); 0,056[a]
20%	3,78 (1); 0,052[a]
18%	4,35 (1); 0,037[a,b]
16%	4,27 (1); 0,039[a,b]
14%	5,96 (1); 0,014[a,b]
11,14%	6,04 (1); 0,014[a,b]

Chi^2 (df); p-value
(a) significantly different on 90% confidence level
(b) significantly different on 95% confidence level

Table 5: Pairwise comparison of performance (AUC) among several test sets using different class distributions

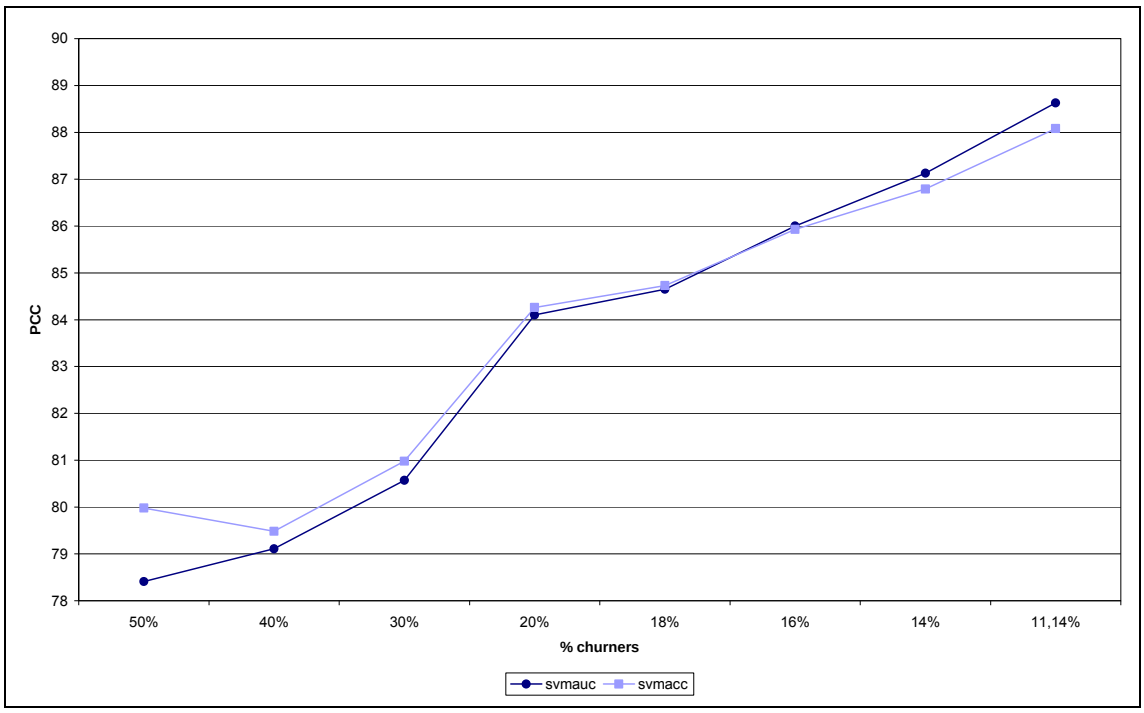


Figure 5: Percentage correctly classified for SVMacc and SVMauc applied to several test sets with different class distributions

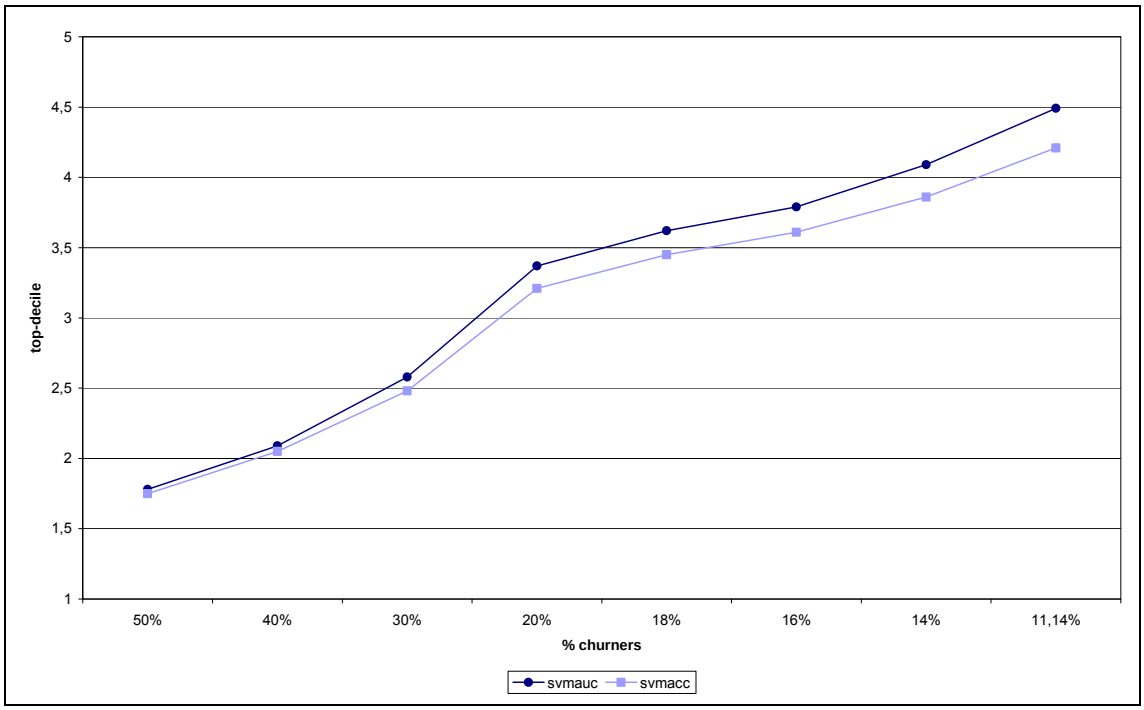


Figure 6: Top-decile lift for SVMacc and SVMauc applied to several test sets with different class distributions

Table 6 compares the predictive capabilities between SVMacc and SVMauc on the real test set (see Table 2). One can clearly see the gap in performance. SVMauc exhibits better predictive performance than SVMacc when both models are evaluated on the real test set. In terms of PCC, the increase is 0.55 percent points. There is also a significant improvement in AUC of 0.24 [20]. With respect to the top-decile lift, an increase from 4.209 to 4.492 is achieved.

	PCC	AUC	Top-Decile lift
SVMacc	88,08	84,90[a]	4,209
SVMauc	88,63	85,14[a]	4,492
(a) significantly different on 95% confidence level			

Table 6: The performance of SVMacc and SVMauc: PCC and AUC on the real test set

In sum, when a SVM is trained with a non-natural distribution, it may be better to select its parameters during the grid search based on the cross-validated AUC. The new parameter-selection technique significantly improves the AUC and the top-decile lift of the model, while accuracy is certainly not decreased.

In the following part, we compare the performance of both kinds of SVMs with logistic regression and random forests.

6.3. Comparing Predictive Performance of SVMs, Logit and Random Forests

The evaluation measures on the real test set (see Table 2) for all models are represented in Tables 7, 8 and 9. Table 7 compares the predictive performance of logit, random forests, SVMacc and SVMauc in terms of PCC and AUC. Table 8 shows the results from the test of [20] which investigates if the AUCs of two models are significantly different. One can find the top-decile lift for all models in Table 9.

Model	PCC	AUC
Logit	88,47	84,60
Random Forests	89,14	87,21
SVMacc	88,08	84,90
SVMauc	88,63	85,14

Table 7: The performance of the different algorithms: PCC and AUC on the real test set

	Random forests	SVMacc	SVMauc
Logit	219,52 (1)[a]	2,53 (1)[b,c]	12,56 (1)[a]
Random forests		190,44 (1)[a]	166,58 (1)[a]
SVMacc			6,04 (1)[a]

Chi^2 (df)
(a) significantly different on 95% confidence level
(b) equal on a 95% confidence level
(c) equal on a 90% confidence level

Table 8: Pairwise comparison of performance (AUC) on the real test set

Logit	Random forests	SVMacc	SVMauc
4,478	4,754	4,209	4,492

Table 9: The performance of the different algorithms: Top-decile lift on the real test

Additionally, Tables 7, 8 and 9 give information concerning the performance of SVMacc and SVMauc benchmarked to logistic regression. Only SVMauc differs significantly in terms of predictive performance when compared to logistic regression. In contrast to SVMauc, SVMacc classifies fewer cases correctly than logistic regression. Moreover the test of [20] confirms that the AUC of SVMacc is not significantly different from that of the logistic regression. The need to select the right parameter-selection technique is confirmed when looking at the top-decile lift criterion. SVMauc identifies more churners than logistic regression, while the top-decile lift of SVMacc is lower than that of logit.

From Tables 7, 8 and 9, one can also compare the performance of both SVMs with the performance of the random forests. It is clear that despite the parameter selection technique, SVMs are surpassed by random forests.

In sum, it is shown that the parameter-selection technique influences the predictive performance of SVMs. Consequently, when a SVM is trained on a balanced distribution, it may be viable and preferable to consider other than the traditional parameter-selection methods. Each improvement in predictive performance will result in a better return on investment of subscriber-retention actions based on these prediction models. In this study, SVMs are trained on a non-natural distribution; it is shown that selecting the parameters based on the best cross-validated AUC results in a better performance than when selecting them based on the highest cross-validated accuracy as was suggested in [28]. In sum, one may say that choosing the right parameter-selection technique is vital for optimizing a SVM application.

In the end, it would also be counterproductive to simply rely on traditional techniques like logistic regression. SVMs - in combination with the correct parameter selection technique - and random forests, both outperform logistic regression. Nevertheless, in this study random forests are better in predicting churn in the subscription services than SVMs.

6.4. Variable importance

In this section, an overview of the most important variables is given. This is done based on the outcome of the random forest importance measures for mainly two reasons: (i) Random forests give the best predictive performance compared to logistic regression and SVM. (ii) Unlike random forests, the SVM software does not produce an internal ranking of variable importance. Moreover, we do not report any measures for logistic regression - e.g. standardized estimates - because most measures are prone to multicollinearity. However, this is not a problem when the focus lies mainly on prediction. In this study, we will elaborate the top-10 most important churn predictors.

It is clear from Appendix B that the length of the subscription and recency – i.e. elapsed time since last renewal – which both belong to the category of variables describing a subscription[4] are ranked on top. Furthermore, another variable from the same category – i.e. the month of contract expiration - is part of the top-10 most explaining churn variables. In contrast to extant research (e.g. [4]), monetary value and frequency – i.e. the number of renewal points – are not present within the top-10 list of most important churn predictors in this study.

[4] see Appendix A

Although most important churn predictors are variables that belong to the group of variables describing a subscription, the impact of some client/company-interaction variables cannot be neglected when investigating the top-10 list of most important variables: (i) Variables related to the ability of voluntarily suspending the subscription – during holiday, during a business trip, … - are present in the top-10. (ii) Recency of complaining – i.e. the elapsed time since the last complaint - is also present in the top-10 most important churn predictors. Consequently, efficient-complaint handling strategies are important. [46] already stated that companies do not deal successfully with service failures because most companies underestimate the impact of efficient complaint handling. (iii) Moreover, this study shows that the variable which indicates whether or not a subscription started from own initiative belongs to the top-10 list in contrast to similar variables related to other purchase motivators like direct mailing campaigns, tele-marketing actions, face-to-face promotions, … .

In spite of the importance of age, one can conclude that socio-demographics do not play an important role in explaining churn in this study which confirms the finding of [24] and more recently, [42].

7. CONCLUSIONS AND FUTURE RESEARCH

In this study, we show that SVMs are able to predict churn in subscription services. By mapping non-linear inputs into a high-dimensional feature space, SVMs break down complex problems into simpler discriminant functions. Because SVMs are based on the Structural Risk Minimization principle that minimizes the upper bound on the actual risk, they show a very good performance when applied to a new, noisy marketing dataset. To validate the performance of this novel technique, we statistically compare its predictive performance with those of logistic regression and random forests. It is shown that a SVM – which is trained on a balanced distribution - outperforms a logistic regression only when the appropriate parameter selection technique is applied. However, when comparing the predictive capabilities of these SVMs with state-of-the-art random forests, our study indicates that SVMs are surpassed by the random forests.

Particularly in this study, we implement a grid search using a 5-fold cross-validation for obtaining the optimal upper bound C and kernel parameter γ that are the most important when implementing a SVM. This study offers an alternative parameter selection technique that outperforms the previously used technique by [28]. The way in which the optimal parameters

are selected, can have significant influences on the performance of a SVM. Taking into account alternative parameter-selection techniques is crucial because even the smallest change in predictive performance can have significantly increases in the return on investment of the marketing-retention actions based on these prediction models [48].

In addition, one can say that academics as well as practitioners don't have to simply rely on traditional techniques like logistic regression. SVMs – in combination with the right parameter-selection technique – and random forests offer some alternatives. Nevertheless, a trade-off has to be made between the time allocated to the modeling procedure and the performance achieved.

In this study, most important churn predictors are part of the group of variables describing the subscription. Unlike ample research, monetary value and frequency are not present in the top-10 most important churn drivers. On the other hand, several client/company-interaction variables play an important role in predicting churn. In spite of the importance of age, socio-demographics do not play an important role in explaining churn in this study.

Directions for future research are given by the fact that nowadays there is no complete working meta-theory to assist with the selection of the correct kernel function and SVM parameters. Deriving a procedure to select the proper kernel function and correct parameter values according to a specific type of classification problem is an interesting topic for further research. Furthermore, applying SVMs using a sufficient sample size can be very time-consuming due to the long computational time and often requires specific software. Before SVMs can be widely adopted, easy-to-use computer software should be available in the traditional data mining packages.

ACKNOWLEDGEMENTS

We would like to thank the anonymous Belgian publishing company for disposing their data. Next, we also like to thank (1) Ghent University for funding the PhD project of Kristof Coussement (BOF 01D26705) and (2) the Flemish government and Ghent University (BOF equipment 011B5901) for funding our computing resources during this project. Also special thanks to L. Breiman (†) for freely distributing the random forest software, as well as C.-C. Chang and C.-J Lin for sharing their SVM-toolbox, LIBSVM.

APPENDIX A: EXPLANATORY VARIABLES INCLUDED IN THE CHURN-PREDICTION MODEL

Client/company-interaction variables: variables describing the client/company relationship:

- The number of complaints,
- Elapsed time since the last complaint,
- The average cost of a complaint (in terms of compensation newspapers),
- The average positioning of the complaints in the current subscription,
- The purchase motivator of the subscription,
- How the newspaper is delivered,
- The conversions made in distribution channel, payment method & edition,
- Elapsed time since last conversion in distribution channel, payment method & edition,
- The number of responses on direct marketing actions,
- The number of suspensions,
- The average suspension length (in number of days),
- Elapsed time since last suspension,
- Elapsed time since last response on a direct marketing action,
- The number of free newspapers.

Renewal-related variable: variables containing renewal-specific information:

- Whether the previous subscription was renewed before the expiry date,
- How many days before the expiry date, the previous subscription was renewed,
- The average number of days the previous subscriptions are renewed before expiry date,
- The variance in the number of days the previous subscriptions are renewed before expiry date,
- Elapsed time since last step in renewal procedure,
- The number of times the churner did not renew a subscription.

Socio–demographic variables: variables describing the subscriber:

- Age,
- Whether the age is known,
- Gender,
- Physical person (is the subscriber a company or a physical person),
- Whether contact information (telephone, mobile number, email) is available.

Subscription-describing variables: group of variables describing the subscription:

- Elapsed time since last renewal,
- Monetary value,
- The number of renewal points,
- The length of the current subscription,
- The number of days a week the newspaper is delivered (intensity indication),
- What product the subscriber has,
- The month of contract expiration.

APPENDIX B: VARIABLE IMPORTANCE MEASURES

No.	AvgNormImp	Variable Name	Level[5]	Relative variable[6]
1	73.946	The length of the current subscription	Subscription	
2	65.335	Elapsed time since last renewal	Subscription	
3	59.460	Elapsed time since last suspension	Subscriber	
4	54.764	Elapsed time since last suspension	Subscription	
5	54.035	The month of contract expiration	Subscription	
6	52.705	Age	Subscriber	
7	51.467	Elapsed time since last complaint	Subscriber	
8	51.056	The average suspension length (in number of days)	Subscriber	X
9	50.251	The purchase motivator of the subscription: own initiative	Subscription	
10	48.560	The average suspension length (in number of days)	Subscriber	
11	48.073	Elapsed time since last complaint	Subscription	
12	47.330	Monetary value	Subscription	
13	46.882	Elapsed time since last step in renewal procedure	Subscription	
14	46.520	Physical person: physical personYES/NO	Subscriber	
15	44.811	The variance in the number of days the previous subscriptions are renewed before expiry date	Subscriber	
16	44.357	The average number of days the previous subscriptions are renewed before expiry date	Subscriber	
17	43.337	Elapsed time since last response on a direct marketing action	Subscriber	
18	42.310	The average number of days the previous subscriptions are renewed before expiry date	Subscription	
19	40.011	The number of renewal points	Subscription	
20	38.448	The number of suspensions	Subscriber	X
21	37.295	The average suspension length (in number of days)	Subscription	X
22	37.158	The purchase motivator of the subscription: direct marketing action	Subscription	
23	36.536	The number of suspensions	Subscription	X
24	35.519	How many days before the expiry date, the previous subscription was renewed	Subscription	
25	35.279	Elapsed time since last conversion in payment method	Subscriber	
26	33.802	Elapsed time since last conversion in payment method	Subscription	
27	33.396	The number of complaints	Subscriber	X

[5] see section 5: Research Data.
[6] correction of the variable by using the length of subscription

28	33.146	The average positioning of the complaints in the current subscription	Subscription	
29	32.520	The conversions made in payment method	Subscription	
30	32.481	The average suspension length (in number of days)	Subscription	
31	32.107	The conversions made in payment method	Subscription	X
32	31.637	The number of responses on direct marketing actions	Subscriber	X
33	31.144	The variance in the number of days the previous subscriptions are renewed before expiry date	Subscription	
34	29.640	The conversions made in payment method	Subscriber	
35	28.116	What product the subscriber has: edition X	Subscription	
36	28.027	The purchase motivator of the subscription: tele marketing action	Subscription	
37	27.860	What product the subscriber has: edition Y	Subscription	
38	27.584	The conversions made in payment method	Subscriber	X
39	26.390	Elapsed time since last conversion in edition	Subscriber	
40	25.442	The number of responses on direct marketing actions	Subscriber	
41	24.942	Elapsed time since last conversion in distribution channel	Subscription	
42	24.802	The number of suspensions	Subscriber	
43	24.237	The number of complaints	Subscription	X
44	24.193	Whether the previous subscription was renewed before the expiry date	Subscription	
45	23.993	Elapsed time since last conversion in edition	Subscription	
46	23.545	The purchase motivator of the subscription: promotional offer	Subscription	
47	23.008	The number of suspensions	Subscription	
48	22.991	Elapsed time since last conversion in distribution channel	Subscriber	
49	22.486	The number of complaints	Subscriber	
50	21.466	How the newspaper is delivered: private distribution channel	Subscription	
51	20.917	Gender: female YES/NO	Subscriber	
52	20.087	The number of complaints	Subscription	
53	19.624	Physical person: company YES/NO	Subscriber	
54	19.600	How the newspaper is delivered: individual newsboy	Subscription	
55	18.930	Whether the age is known	Subscriber	
56	18.906	The number of times the subscriber did not renew a subscription	Subscriber	
57	18.426	The conversions made in distribution channel	Subscriber	X
58	17.802	The conversions made in edition	Subscriber	X
59	17.718	The conversions made in distribution channel	Subscriber	

60	17.289	The purchase motivator of the subscription: direct marketing action	Subscription	
61	17.249	The purchase motivator of the subscription: face-to-face marketing	Subscription	
62	16.996	The conversions made in edition	Subscriber	
63	16.534	The conversions made in distribution channel	Subscription	
64	16.095	The conversions made in edition	Subscription	X
65	15.446	The conversions made in distribution channel	Subscription	X
66	15.406	What product the subscriber has: edition Z	Subscription	
67	15.222	The conversions made in edition	Subscription	
68	14.531	The average cost of a complaint (in terms of compensation newspapers)	Subscriber	X
69	13.995	The average cost of a complaint (in terms of compensation newspapers)	Subscription	X
70	13.602	Gender: male YES/NO	Subscriber	
71	12.587	The average cost of a complaint (in terms of compensation newspapers)	Subscriber	
72	12.005	How the newspaper is delivered: public distribution channel	Subscription	
73	11.830	Gender: private company YES/NO	Subscriber	
74	11.550	The purchase motivator of the subscription: direct marketing mailing action	Subscription	
75	11.059	How the newspaper is delivered: pick up newspaper at shop	Subscription	
76	10.651	The average cost of a complaint (in terms of compensation newspapers)	Subscription	
77	7.601	Gender: public company YES/NO	Subscriber	
78	7.027	The number of free newspapers	Subscription	
79	5.190	The number of days a week the newspaper is delivered (intensity indication)	Subscription	
80	4.979	Whether contact information (telephone, mobile number, email) is available	Subscriber	
81	2.991	How the newspaper is delivered: delivered abroad via courier	Subscription	
82	2.093	What product the subscriber has: edition W	Subscription	

REFERENCES

[1] N. Acir, A support vector machine classifier algorithm based on a perturbation method and its application to ECG beat recognition systems, Expert Systems with Applications, 31 (1) (2006) pp. 150-158.

[2] P.D. Allison, Logistic regression using the SAS sytem: theory and application, Cary, NC: SAS Institute Inc. (1999).

[3] A.D. Athanassopoulos, Customer satisfaction cues to support market segmentation and explain switching behavior, Journal of Business Research 47 (3) (2000) pp. 191-207.

[4] C.L. Bauer, A direct mail customer purchase model, Journal of Direct Marketing, 2, (3), (1988) pp. 16-24.

[5] M. Bicego, E. Grosso and M. Tistarelli, Face authentication using one-class support vector machines, Lecture Notes in Computer Science 3781 (2005) pp. 15-22.

[6] A. Bratko and B. Filipic, Exploiting structural information for semi-structured document categorization, Information Processing & Management 42 (3) (2006) 679-694.

[7] L. Breiman, Random forests, Machine Learning 45 (1) (2001) pp. 5-32.

[8] W. Buckinx and D. Van den Poel, Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting, European Journal Of Operational Research 164 (1) (2005) pp. 252-268.

[9] R.E. Bucklin and S. Gupta, Brand choice, purchase incidence and segmentation: an integrated modeling approach, Journal of Marketing Research, 29 (1992) pp. 201-215.

[10] J. Burez and D. Van den Poel, CRM at Canal+ Belgique: reducing customer attrition through targeted marketing, forthcoming in Expert Systems with Applications.

[11] C.J.C. Burges and B. Scholkopf, Improving the accuracy and speed of support vector machines, in Mozer, M., Jordan, M., Petche, T., Advances in Neural Information Processing Systems, Cambridge, M.A. MIT Press (1997).

[12] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (2) (1998) pp. 121-167.

[13] P.K. Chan and S.J. Stolfo, Learning with non-uniform class and cost distributions: a case study in credit card fraud detection, Proceedings Fourth Intl. Conf. On Knowledge Discovery and Data Mining (1998), pp. 164-168.

[14] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University (2004).

[15] K.-Y. Chen and C.-H. Wang, A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan, forthcoming in Expert Systems with Applications.

[16] X.J. Chen, R. Harrison and Y.Q. Zhang, Multi-SVM fuzzy classification and fusion method and applications in bioinformatics, Journal of Computational and Theoretical Nanoscience 2 (4) (2005) pp. 534-542.

[17] C. Cortes and V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) pp. 273-297

[18] D. Cui and D. Curry, Predictions in marketing using the support vector machine, Marketing Science 24 (4) (2005) pp. 595-615.

[19] M.G. Dekimpe and Z. Degraeve, The attrition of volunteers, European Journal of Operational Research 98 (1) (1997) pp. 37-51.

[20] E.R. DeLong, D.M. DeLong and D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics, 44 (3) (1988) pp. 837-845.

[21] R.O. Duda, P.E. Hart and D.G. Stork, Pattern classification, Wiley, New York (2001).

[22] J.P. Egan, Signal detection theory and roc analysis, Series in Cognition and Perception, Academic Press, New York (1975).

[23] D. Glotsos, J. Tohka and P. Ravazoula, Automated diagnosis of brain tumours astrocytomas using probabilistic neural network clustering and support vector machines, International Journal of Neural Systems 15 (1-2) (2005) pp. 1-11.

[24] P.M. Guadagni and J.D.C Little, A logit model of brand choice calibrated on scanner data, Marketing Science 2(3) (1983) pp. 203–238.

[25] J.A. Hanley and B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve, Radiology, 143 (1) (1982) pp. 29-36.

[26] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning: data mining, inference and prediction, Springer-Verlag (2001).

[27] J.Y. He, H.J. Hu and R. Harrison, Understanding protein structure prediction using SVM_DT, Lecture Notes in Computer Science 3759 (2005) pp. 203-212.

[28] C.-W. Hsu, C.-C. Chang and C.-J. Lin, A practical guide to support vector classification, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University (2004).

[29] S.-Y. Hung, D.C. Yen and H.-Y. Wang, Applying data mining to telecom churn management, forthcoming in Expert Systems with Applications.

[30] M.A. Jones, D.L. Mothersbaugh and S.E. Beatty, Switching barriers and repurchase intentions in services, Journal of Retailing 76 (2) (2000) pp. 259-374.

[31] S. Keaveney and M. Parthasarathy, Customer switching behavior in online services: an exploratory study of the role of selected attitudinal, behavioral and demographic factors, Journal of the Academy of Marketing Science 29 (4) (2001) pp. 374-390.

[32] S.S. Keerthi and C.-J. Lin, Asymptotic behaviours of support vector machines with gaussian kernel, Neural Computation 15 (7) (2003) pp. 1667-1689.

[33] S. Kim, K.S. Shin and K. Park, An application of support vector machines for customer churn analysis: credit card case, Lecture Notes in Computer Science 3611 (2005) pp. 636-647.

[34] S.K. Kim, S. Yang and K.S. Seo, Home photo categorization based on photographic region templates, Lecture Notes In Computer Science 3689 (2005) pp. 328-338.

[35] B. Larivière and D. Van den Poel, Predicting customer retention and profitability by using random forests and regression forests techniques, Expert Systems With Applications 29 (2) (2005) pp. 472-484.

[36] S.-T. Li, W. Shiue and M.-H. Huang, The evaluation of consumer loans using support vector machines, Expert Systems with Applications, 30 (4)(2006) pp. 772-782.

[37] H.-T. Lin and C.-J. Lin, A study on sigmoid kernels for SVM and the training of non-psd kernels by SMO-type methods, Technical report, Department of Computer Science and Information Engineering, National Taiwan University (2003).

[38] T. Luo, K. Kramer, D.B. Goldgof, L.O. Hall, S. Samson, A. Remsen and T. Hopkins, Recognizing plankton images from the shadow image particle profiling evaluation recorder, IEEE Transactions on Systems Man and Cybernetics Part B – Cybernetics, 34 (4) (2004) pp. 1753-1762.

[39] S.A. Neslin, S. Gupta, W. Kamakura, J. Lu and C. Mason, Defection detection: improving predictive accuracy of customer churn models, Working Paper (2004).

[40] P.F. Pai and C.S. Lin, Using support vector machines to forecast the production values of the machinery industry in Taiwan, International Journal of Advanced Manufacturing Technology 27 (1-2) (2005) pp. 205-210.

[41] W. Reinartz and V. Kumar, The impact of customer relationship characteristics on profitable lifetime duration, Journal of Marketing 67 (1) (2003) pp. 77-99.

[42] P.E. Rossi, R.E. McCulloch and G.M. Allenby, Value of household information in target marketing, Marketing Science 15 (1996) pp. 321–340.

[43] R.T. Rust and R. Metters, Mathematical models of service, European Journal of Operational Research 91 (3) (1996) pp. 427-439.

[44] J.A. Swets, Roc analysis applied to the evaluation of medical imaging techniques, Investigative Radiology, 14 (1989), pp 109-121.

[45] J.A. Swets and R.M. Pickett, Evaluation of diagnostic systems: methods from signal detection theory, Academic Press, New York (1982).

[46] S.S. Tax, S.W. Brown and M. Chandrashekaran, Customer Evaluations of Service Complaint Experiences: Implications for Relationship Marketing, Journal of Marketing 62 (April) (1998) pp. 60-76.

[47] J.S. Thomas, A methodology for linking customer acquisition to customer retention, Journal of Marketing Research 38 (2) (2001) pp. 262-268.

[48] D. Van den Poel and B. Larivière, Customer attrition analysis for financial services using proportional hazard models, European Journal of Operational Research 157 (2004) pp. 196–217.

[49] V. Vapnik, Statistical learning theory, Wiley, New York (1998).

[50] V. Vapnik, The nature of statistical learning theory, Springer, New York (1995).

[51] G. Weiss and F. Provost, The effect of class distribution on classifier learning, Technical Report ML-TR-43, Department of Computer Science, Rutgers University (2001).

[52] K. Yamaguchi, Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of 'permanent employment' in Japan, Journal of the American Statistical Association 87 (No. 418) (1992) pp. 284-292.

[53] Y. Zhao, B. Li and X. Li, Customer churn prediction using improved one-class support vector machine, Lecture Notes in Artificial Intelligence 3584 (2005) pp. 300-306.

[54] W. Zhong, J. He, R. Harrison, P.C. Tai. and Y. Pan, Clustering support vector machines for protein local structure prediction, forthcoming in Expert Systems with Applications (2006).

CHAPTER II

INTEGRATING THE VOICE OF CUSTOMERS THROUGH CALL CENTER EMAILS INTO A DECISION SUPPORT SYSTEM FOR CHURN PREDICTION

This chapter is based on K. Coussement and D. Van den Poel, Integrating the voice of customers through call center emails into a decision support system for churn prediction, Information and Management, 45 (3) (2008), pp 164-174.

CHAPTER II

INTEGRATING THE VOICE OF CUSTOMERS THROUGH CALL CENTER EMAILS INTO A DECISION SUPPORT SYSTEM FOR CHURN PREDICTION

ABSTRACT

We studied the problem of optimizing the performance of a DSS for churn prediction. In particular, we investigated the beneficial effect of adding the voice of customers through call center emails – i.e. textual information - to a churn prediction system that only uses traditional marketing information. We found that adding unstructured, textual information into a conventional churn prediction model resulted in a significant increase in predictive performance. From a managerial point of view, this integrated framework helps marketing-decision makers to identify customers most prone to switch. Consequently, their customer retention campaigns can be targeted effectively because the prediction method is better at detecting those customers who are likely to leave.

1. INTRODUCTION

In the past, companies focused on selling products and services with little knowledge or strategy concerning the customers who bought the products. Today business is evolving from this 'product-centered' to a 'customer-centered' environment. Companies need to find ways to capture and enhance market share while reducing costs [7]. Consequently, existing companies must reconsider the business relationships with their customers [24].

Customer relationship management (CRM) is becoming a critical success factor in today's business environment [2,16]. Data mining is being implemented to gain customer knowledge from organizational data warehouses [35]. A way to manage customer churn is to predict which customers are most likely to leave and then target them with incentives to stay. Consequently, these IS support marketing-decision makers to generate marketing campaigns for the right customers. A field experiment by [9] has already shown that companies can boost profitability by shifting from mass to focused marketing strategies. It is more profitable to keep and satisfy existing customers than to attract new ones with a high attrition rate [26]. Identifying customers most prone to switch, is thus important

[17]. In order to develop an effective customer retention program, the company must build a model that is as accurate as possible; indeed Van den Poel and Larivière [36] showed that a small change in retention rate can result in a significant change in profitability.

We decided it was necessary to incorporate the voice of customers (VOC) through call center emails into a traditional churn-prediction model in order to provide a better model: one with a higher predictive performance. The rapid development of IT and the internet has made it easier for customers to communicate with the company. Call centers are expanding rapidly in scope, number and size [1], because many firms rely on them to address customer concerns and provide product information [25]. However, marketing managers tend to neglect this valuable information because (i) it is not directly applicable in a traditional marketing context, (ii) there is seldom in-house knowledge on how to convert this (textual) information into an analyzable form, and (iii) no ready-to-use framework is available to integrate the information. We developed a DSS for churn prediction; it integrates free-formatted, textual information from customer emails with information derived from the marketing database. Although previous research used the VOC in understanding customers' needs and behavior (e.g. [10, 11, 21]), no prior work has used VOC in a churn-prediction model.

2. METHODOLOGY

Figure 1 shows how the integration of information types in a churn-modeling system was achieved.

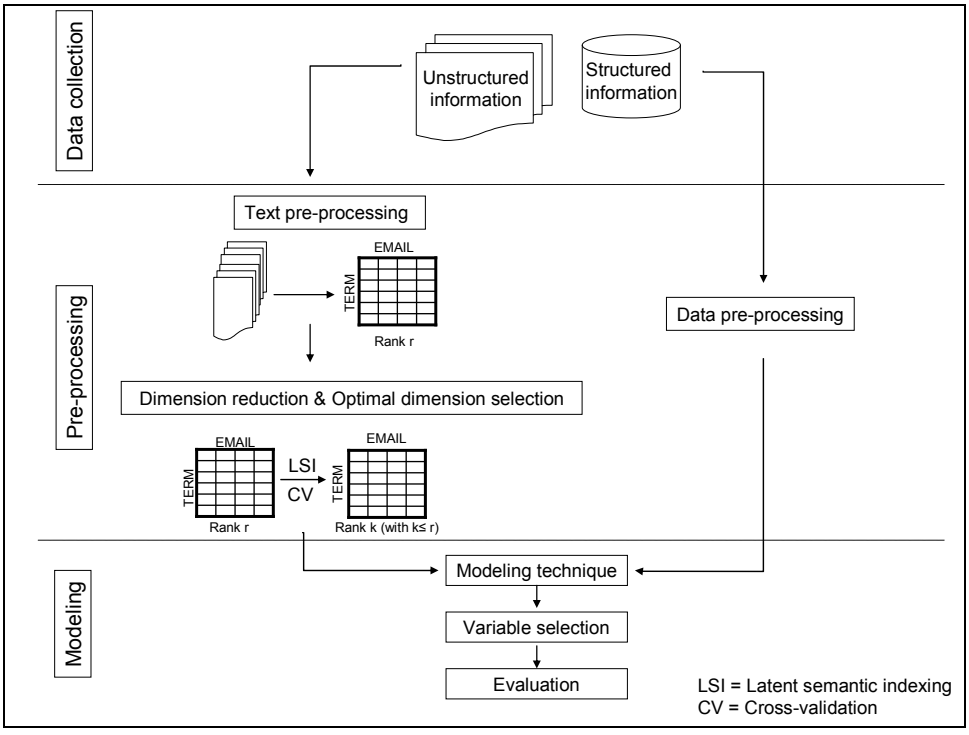


Figure 1: An integrated churn-modeling system that uses structured, database-related information and free-formatted, textual information.

2.1. Data collection

Structured marketing information can be extracted from a common marketing database in which all transactional and marketing-related information has been stored. In contrast, call-center emails are highly unstructured. Thus, extracting information from emails requires meticulous pre-processing to capture the relevant details for inclusion in a churn detection/prediction DSS.

2.2. Pre-processing

Data and Text pre-processing

The structured information is internally available at a very low cost and available for pre-processing and integration into our model. However, the original emails are unformatted by nature. They are converted into a structured representation using the vector-space of Salton's SMART [31]: an email is represented as a vector of weighted frequencies of designated words. Thus emails are *n*-dimensional vectors, with *n* the number of distinct terms in the dictionary. Each vector component reflects the importance of the corresponding term with respect to the semantics of the email [6] and each component has a weight if the term is present or zero otherwise. Thus a collection of emails is represented as a term-by-email matrix. Figure 2 shows the steps in this pre-processing phase whereby raw emails become a term-by-email matrix.

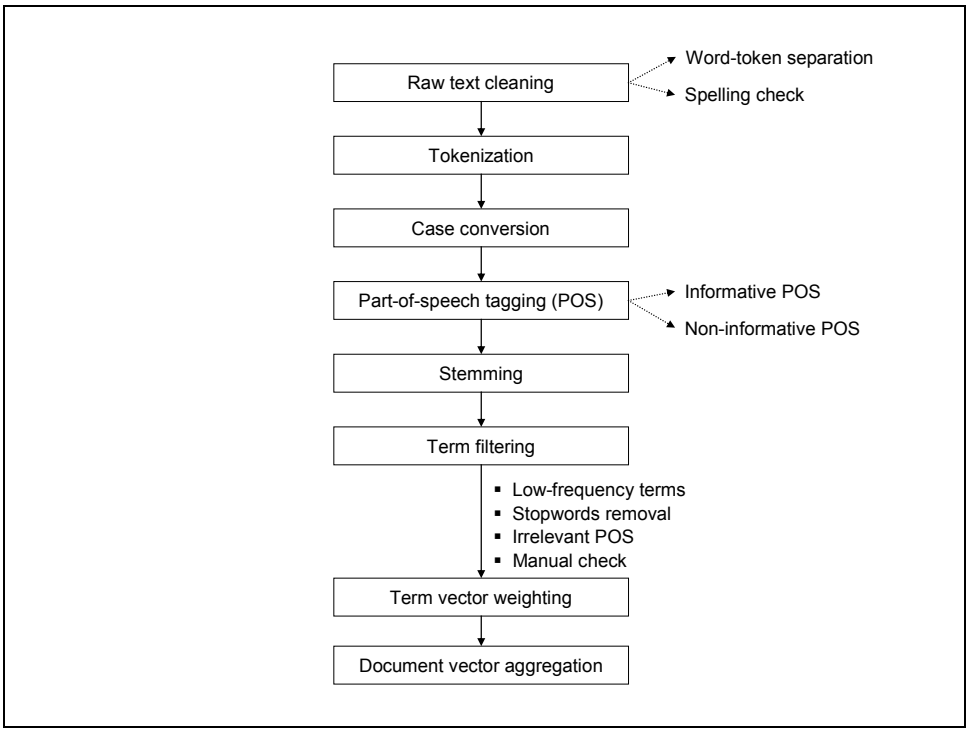


Figure 2: Different steps in the text pre-processing phase.

In the first step, *raw text cleaning*, special characters and punctuation are removed from words and spelling errors are corrected by comparing with words in a reference dictionary using a synonym data set. *Tokenization* converts the input stream into tokens and words. It uses blanks as delimiters for words which are then converted to lower case (*case conversion*). *Part-of-speech tagging* gives words their syntactic category: informative (nouns, verbs, adjectives, and adverbs) or non-informative.

Next, terms are replaced by their stem; e.g., *connect* is the stem for *connected*, *connecting*, *connection*, *etc*., *Stemming* reduces the number of terms significantly [5] and increases retrieval performance [19]. A dictionary-based stemmer is used. When a term is unrecognizable, standard decision rules are applied to give the term a correct stem.

The result of this process is a high-dimensional term-by-email matrix having many distinct terms. This matrix is reduced by applying *term filtering:* rare words are eliminated because they seldom help in future classifications. Word frequencies follow a Zipf distribution [37]: thus half of them appear only once or twice. Eliminating words under those thresholds often yield great savings [22]. Stopwords, (e.g. ‘the’ or ‘a’) are also removed. Next, the non-informative parts of speech are left out from the analysis. A last step in the term-filtering phase is removing irrelevant terms by manually checking the temporary dictionary.

In the *term vector weighting* phase, a weighted term vector for every email is constructed. By now, the values in the term-by-email matrix are simply the raw frequencies of appearance for a term in an email. Spark Jones [33] showed significant improvements in retrieval performance when using weighted term vectors. Term weighting is often done by determining the product of the term frequency (*tf*) and the inverse email frequency (*idf*) [27, 28, 29, 34]. The result is a high-dimensional, weighted term-by-email matrix. Appendix 1 describes the term vector weighting phase in detail.

In the final step, an aggregated term-by-email matrix is generated (i.e. *email vector aggregation*). The aim is to make an aggregation of the email vectors that belong to the same customer. This is necessary because a customer can send more than one email during the observation period, while from a prediction point of view, a prediction is made per individual customer. As such an aggregation of the information for all emails from the same customer is needed. The aggregated weight of term *i* for all emails belonging to subscription *j* (Aw_{ij}) is

$$Aw_{ij} = \sum_{k=1}^{r} w_{ik} \tag{4}$$

with w_{ik} equal to the weight of term *i* in email *k* and *r* equal to the number of emails belonging to the same observation.

Using each distinct term as a feature in the churn-modeling phase would lead to an unmanageable number of explanatory variables. Moreover, due to the high dimensionality of the feature space, most weights are zero for a single email. Thus, using a large and sparse term-by-email matrix would be counterproductive in the predictive-modeling context.

Dimension reduction

The dimension of the aggregated (weighted) term-by-email matrix is reduced by using Latent Semantic Indexing (LSI). It reduces the dimensionality of the feature space by grouping together related terms [12]. Deerwester et al. [12] used singular value decomposition (SVD) to form semantic generalizations from emails. It uses the fact that certain terms appear in similar emails to establish relationships between the terms. Consequently, SVD projects emails from the high-dimensional term space to an orthonormal, semantic, latent subspace by grouping together similar terms into concepts. As such, each concept can be described using many different keywords as it has a high discriminatory power to other concepts in the reduced feature space. See Appendix 2 for more detailed information about LSI using SVD.

Optimal dimension selection

The intensity of dimension reduction during the SVD phase is critical. Ideally, the number of concepts, *k*, must be large enough to fit all the underlying, relevant concepts in the email collection, but small enough to prevent the model from fitting sampling errors and unimportant details. Moreover, the obtained optimal *k* must be workable from a prediction point of view. In the factor-analytic literature, such choices are still an unanswered question. Deerwester et al. [12] propose using an operational criterion, i.e. a value of *k* that yields good performance. In our application, we are especially interested in the predictive performance of the SVD output.

It is not possible to know what value of *k* will lead to an optimal solution when validating the predictive model initially. As such, improper selection of the parameter *k* is ineffective if too few concepts are included or computational expensive if too many irrelevant concepts are incorporated. Consequently, a parameter-selection procedure is needed. We construct several rank-*k* models and the most favorable rank-*k* model (based on the cross-validated performance) is retained for further analysis. As such, the optimal value of *k* is obtained in a 5-fold cross-validation on the training set. The training set is divided into five subsets of equal size. Iteratively, each part is used for validation, while the other parts are used for training. So finally, each case in the training set is predicted once. The cross-validation performance better reflects the real performance when validating the classifier for unseen data. In the end, it is possible to select the optimal value of *k* based on the most favorable cross-validated model. Kim [18] stated that it is very important for data analysts to consider the relationship between the amount of information and the complexity of predictive models because compact information models show great improvement in terms of predictive performance and robustness.

2.3. Modeling

Modeling technique and variable selection

Logistic regression is used. In applying it, a maximum-likelihood function is produced and maximized in order to become an appropriate fit to the data [3]. With a training set of $T = \{(\mathbf{x}_i, y_i)\}$ and $i=\{1,2,\ldots,N\}$ and input data $\mathbf{x}_i \in R^n$ and corresponding binary target labels $y_i \in \{0,1\}$, logistic regression is used to estimate the probability $P(y=1|\mathbf{x})$ given by

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(w_0 + \mathbf{w}\ \mathbf{x}))} \tag{9}$$

with $\mathbf{x} \in R^n$ an n-dimensional input vector, $\mathbf{w}$ the parameter vector and w_0 the intercept.

This technique is used because it is conceptually simple [8], a closed-form solution for the posterior probabilities is available and Neslin et al. [23] stated that it provides quick and robust results in a churn prediction context.

Variable selection is the process of choosing a subset of the original variables by eliminating some variables based on their predictive performance. Kim [18] stated that there are three main reasons for using a variable-selection technique: saving computational time and cost by extracting as much information with the smallest number of variables, improving the comprehensibility of the resulting models and making the model generalize better.

Our study employs a forward-selection procedure: the algorithm added one variable at a time. The first variable to enter the model is that with the highest chi-square statistic. At each step, the remaining variables are considered for inclusion in the final model. Forward selection adds variables until a stopping rule is satisfied. The choice of this standard variable-selection technique makes it easy for implementation, while more sophisticated algorithms are computationally more expensive and require additional parameter settings.

Evaluation criteria

To evaluate the performance of classification models, two commonly used criteria are used: the lift and the area under the receiving operating curve (AUC).

Lift is the most commonly used performance measure for evaluating classification models. It reflects the increase in density of the churn event relative to the density of churners in the total database. The higher the lift, the better the predictive model. In marketing applications, it is interesting to increase the density of churners, especially in the top 10% cases most likely to churn. Marketing-decision makers are typically interested in only 10% of the entire marketing database because budgets are often limited and actions to reduce churn typically involve only 10% of the entire customer database. Practically, all cases are sorted from most likely to churn to least likely to churn. Afterwards, the density of churners from the top 10% cases most likely to churn is compared with the density of churners in the entire customer collection. This increase in density is called the top-decile lift. Intuitively, a top-decile lift of two means that the density of churners in the top 10% cases most likely to churn is twice the density of churners in the entire database.

The *AUC* takes into account the predicted class of an event with the real class of that event, considering all possible cut-off values. Consequently, the AUC takes into account the individual class performance for a range of possible thresholds. If TP (true positives) are the number of positives that are correctly identified, FP (false positives) are the number of negatives that are classified as positives, FN (false negatives) are the number of positive cases that are identified as negatives, and TN (true negatives) are the number of negative cases that are classified as negatives, then

- the sensitivity is (TP/(TP+FN)): the proportion of positive cases that are predicted to be positive
- the specificity is (TN/(TN+FP)): the proportion of negative cases that are predicted to be negative.

These vary when the threshold value is varied. The receiver operation characteristics curve (ROC) is a two-dimensional plot of the sensitivity versus (1-specificity). In order to compare the performance of two or more classification models, the area under the receiver operating characteristics curve is calculated. This measure is used to evaluate the performance of a binary classification system [15]. In order to test if two AUCs are significantly different, one can apply the non-parametric test of Delong et al. [13].

3. EMPIRICAL VERIFICATION

3.1. Research Data

In our study, we used data obtained from a large Belgian newspaper publishing company. Subscribers have to pay a fixed price for their newspapers, depending on the length of subscription and the promotional offer given. The company does not allow subscribers to end their subscription before the expiry date. The churn-prediction problem therefore involves predicting whether a subscription will be renewed during the four-week period after maturity. During this period, the newspaper publishing company still delivers the newspapers in order to allow subscribers time to renew their subscription. The company has a structured, marketing database where transactional and subscription related information is stored and they save all customer emails sent to the call center. Figure 3 shows the time window of analysis in our study.

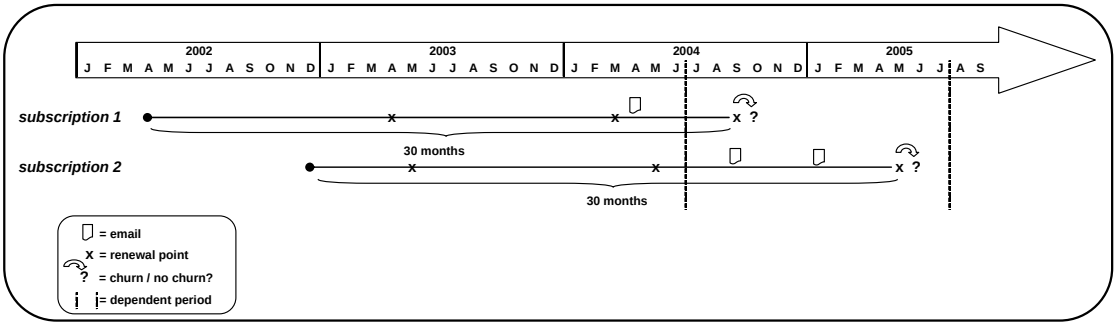


Figure 3: Time window of analysis.

Subscription data from January 2002 through September 2005 was analyzed. Consequently, it is possible to define the dependent and the explanatory variables. All renewal points between July 2004 and July 2005 were considered. A customer was seen as a 'churner' when the subscription was not renewed in a four-week period following the maturity date. The explanatory variables were constructed from the two available types of information. These were used to predict whether a subscription would be renewed.

The first type of variables contained information from the structured, marketing database. These variables contained information on a 30-month period. They were subdivided into four categories (see Appendix 3);

- client/company interaction variables,
- subscription related variables,
- renewal specific variables and
- socio-demographics.

The second type of information consisted of all information sent by the subscriber via email during the last period of his/her subscription. Because this information is highly unstructured, the emails were pre-processed to represent them in our churn-prediction model.

In order to compare the beneficial effect of unstructured information from call center emails in a churn-prediction model, subscriptions with at least one email sent during the last term of the subscription were considered.

Table 1 and 2 summarize the data characteristics for the randomly split training and test set. The training set was used to obtain the optimal SVD dimension and the model estimates, while the test set is used to validate and compare the different models.

	Number of subscriptions	Relative percentage
Training set		
Subscriptions not renewed	1777	18.50%
Subscriptions renewed	7826	81.50%
Total	*9603*	100%
Test set		
Subscriptions not renewed	593	18.76%
Subscriptions renewed	2568	81.24%
Total	*3161*	100%

Table 1: Overview of the marketing data characteristics.

	Number of emails	Average number of mails per subscription	Average number of words per email	Average number of words per sentence	Average number of unique words per email
Training set	14083	1.47	113.27	28.98	72.54
Test set	4694	1.48	116.33	29.33	72.49

Table 2: Overview of the call center emails characteristics.

3.2. Optimal dimension selection

The text pre-processing phase resulted in a high-dimensional term-by-email matrix. This was unworkable from a prediction point of view. Its optimal reduced rank was obtained by applying a cross-validation procedure on the training data. Figure 4 shows the results of this cross-validation; the X-axis has the number of concepts and the Y-axis represents the cross-validated AUC. It is clear that in the range of 1 to 100 concepts, the cross-validated performance was increasing rapidly. From 100 concepts on, the cross-validated AUC was growing less rapidly, while in the region around 170 concepts, the cross-validated performance was stabilizing. Including more than 170 concepts resulted in a more complex churn model, while the predictive performance hardly increased. Thus 170 concepts was chosen as the optimal number for representing the textual information in our study. At this point, a good balance was achieved between the number of concepts and the predictive performance.

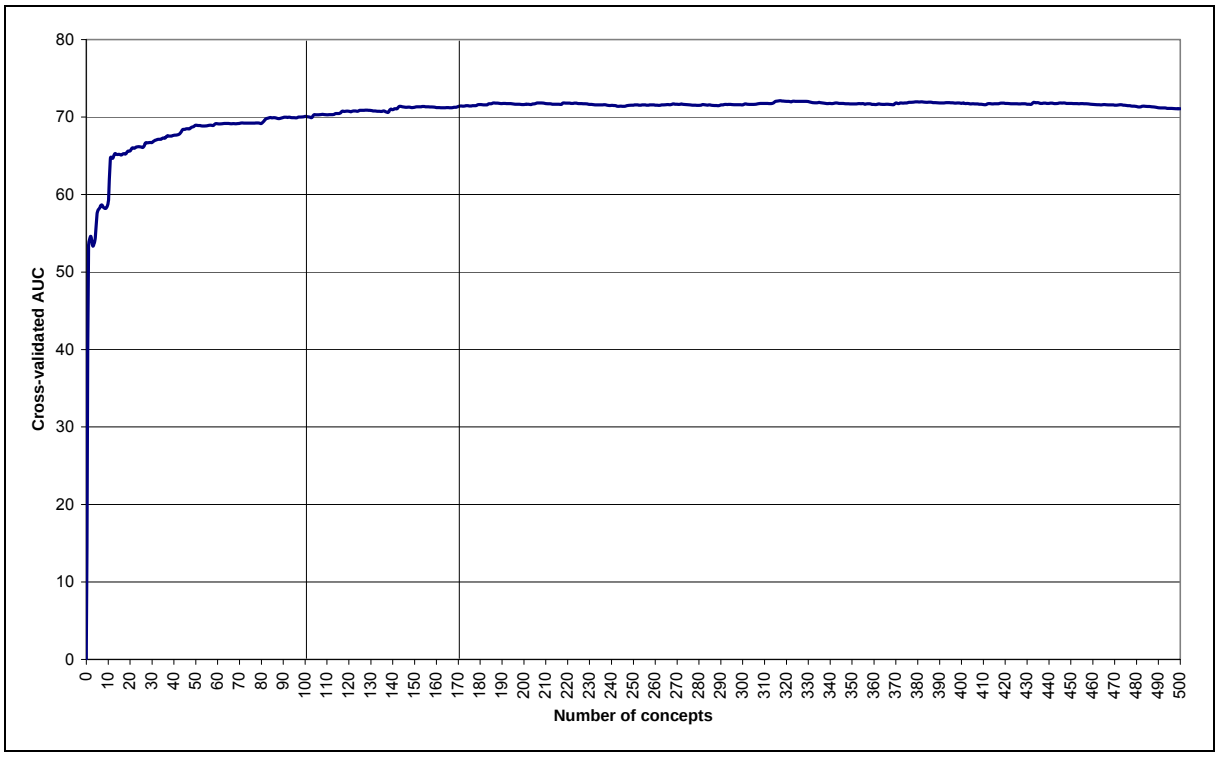


Figure 4: The cross-validated AUC during the optimal dimension selection phase.

3.3 Defining the best subset of the structured marketing variables

Before comparing the predictive performance of the model with structured marketing information only (ModStruc) to the performance of the model that combined the structured marketing information and textual information (ModStruc-Unstruc), the optimal set of structured marketing variables was found by employing the forward selection procedure. It resulted in a best subset of 20 marketing variables (see Table 3)

Step	Variable name
1	Elapsed time since the last complaint
2	Monetary value
3	Elapsed time since last suspension
4	The length of the current subscription
5	The average positioning of complaints in the current subscription (with 0=start of the subscription and 1=end of subscription)
6	Whether the previous subscription was renewed before the expiry date
7	Whether the subscriber is a woman
8	The variance in the number of days the previous subscriptions are renewed before expiry date
9	The number of renewal points
10	Whether the newspaper edition is 'X1'
11	Whether the subscriber is a public institution
12	How many days before the expiry date, the previous subscription was renewed
13	The number of suspensions[x]

14	The average suspension length (in number of days)[x]
15	The number of suspensions
16	The average suspension length (in number of days)
17	Whether the purchase motivator is a direct marketing campaign
18	Whether the newspaper is picked up at the shop
19	Elapsed time since last conversion in payment method
20	The conversions made in payment method[x]

x = variable corrected for the length of subscription

Table 3: Best subset of marketing variables employed by the forward selection procedure.

Modstruc was built using the 20 marketing variables, while ModStruc-Unstruc was a combination of those 20 marketing variables with those variables representing the textual information – i.e. 170 additional variables.

3.4. Comparing predictive performance

	AUC	Top-decile lift
ModStruc	73.80	2.69
ModStruc-Unstruc	77.75	3.07

Table 4: The performance of ModStruc and ModStruc-Unstruc: AUC and top-decile lift on the test set.

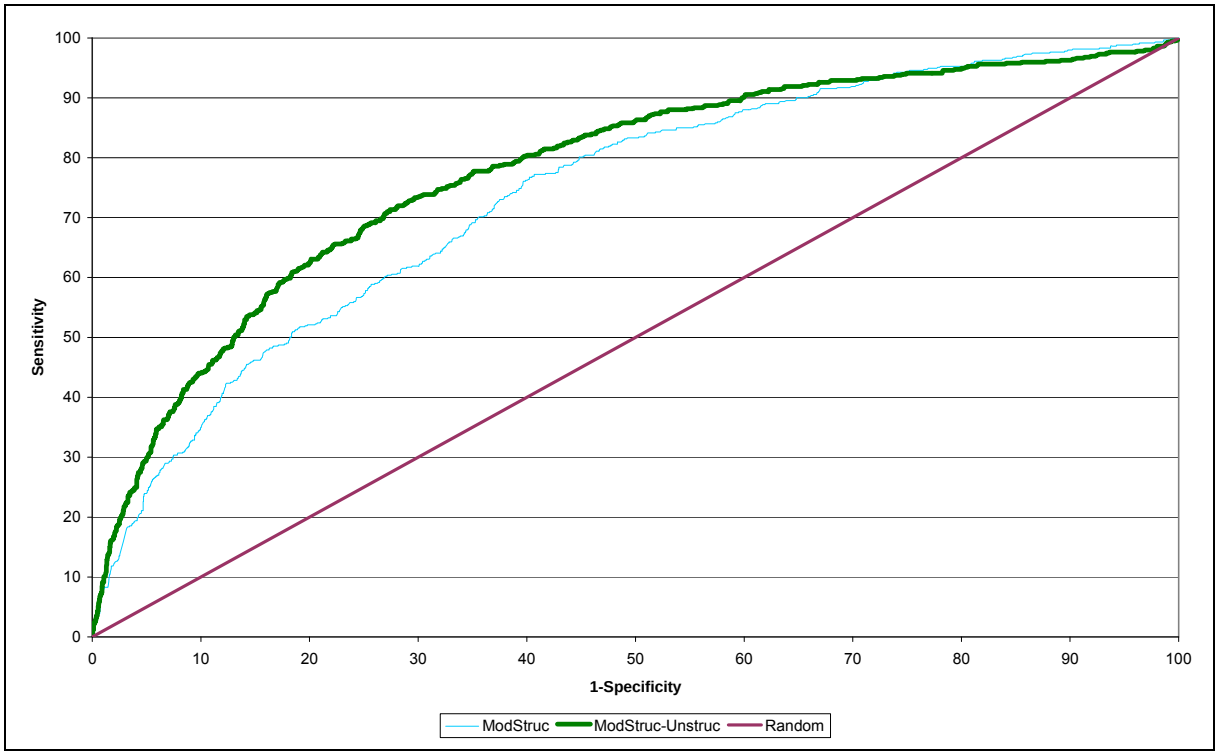


Figure 5: The ROC curves for ModStruc, ModStruc-Unstruc and the Random model (or the zero-information model).

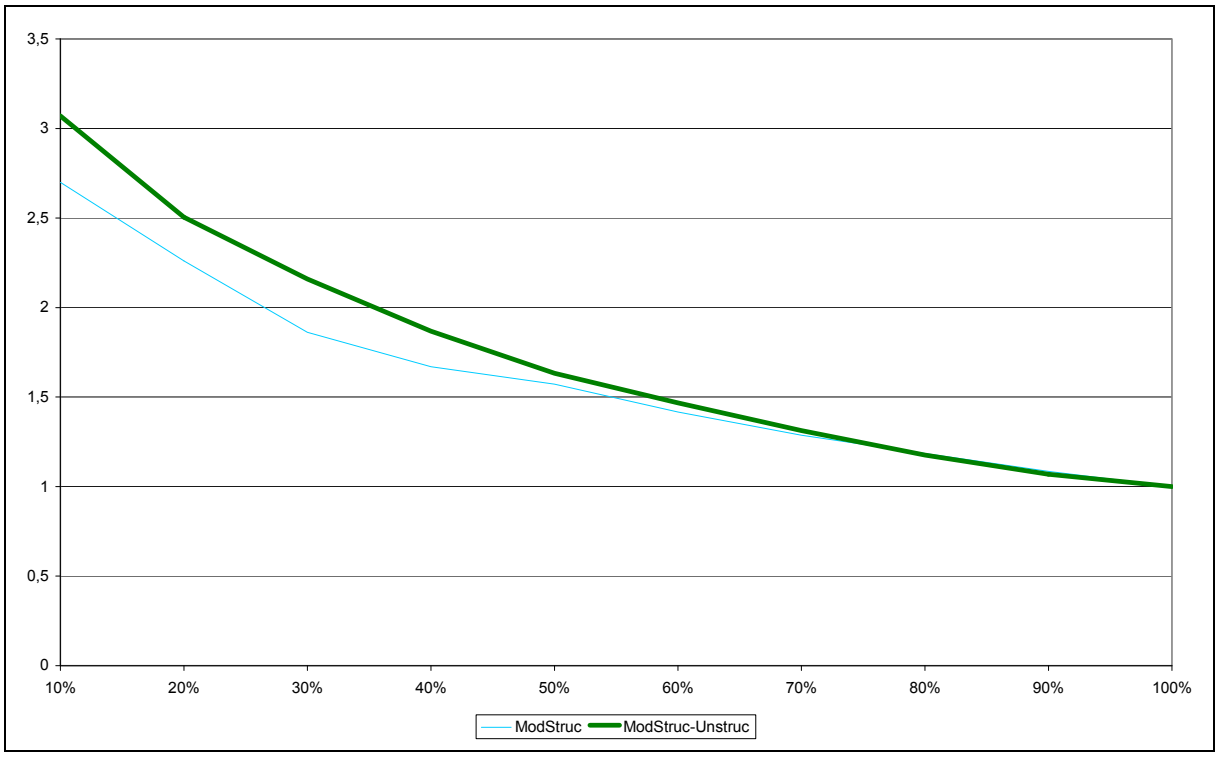


Figure 6: The cumulative lift charts of ModStruc and ModStruc-Unstruc.

Table 4, Figure 5 and Figure 6 show that the predictive performance of ModStruc-Unstruc significantly outperformed that of ModStruc. The AUC increased from 73.80 to 77.75 by adding textual information to a traditional churn-prediction model. This improvement was significant (χ^2=23.1,df=1,p<0.001). The ROC curve of ModStruc-Unstruc is located further from the Random model than that of ModStruc, thus the area under the ROC of ModStruc-Unstruc is larger than that of ModStruc. ModStruc-Unstruc was thus able to better distinguish churners from non-churners. Moreover, the beneficial effect of textual information on predictive performance was confirmed in terms of top-decile lift. The cumulative lift curve of ModStruc-Unstuc laid above that of ModStruc. ModStruc-Unstruc is able to identify more customers truly at risk than ModStruc within a specific decile. Lift in the first decile or the 10% top-decile – i.e. the 10% point - increased from 2.69 to 3.07.

Our study provided a realistic framework that increased the predictive performance of a churn model for subscribers whose textual information is available. Since ModStruc and ModStruc-Unstruc were built on a selective sample of subscribers who contacted the company at least once per email, one may suggest including more subscribers – i.e. those who did not send an email. One should verify whether a separate churn model of subscribers who sent at least one email is the best strategy in obtaining optimal predictive performance. Practically, the current training set of subscribers was extended by randomly selecting subscriptions of customers who had not send any email (ModStruc-k, with *k* the number of randomly selected subscriptions whereby *k*={0,5000,10000,…, 100000}) with the intent of

building a churn model with better predictive performance on the current test set. Figure 7 graphically indicates the results. The horizontal lines indicating the performance of ModStruc and ModStruc-Unstruc are included for reasons of comparability, despite the fact that they were independent of *k*.

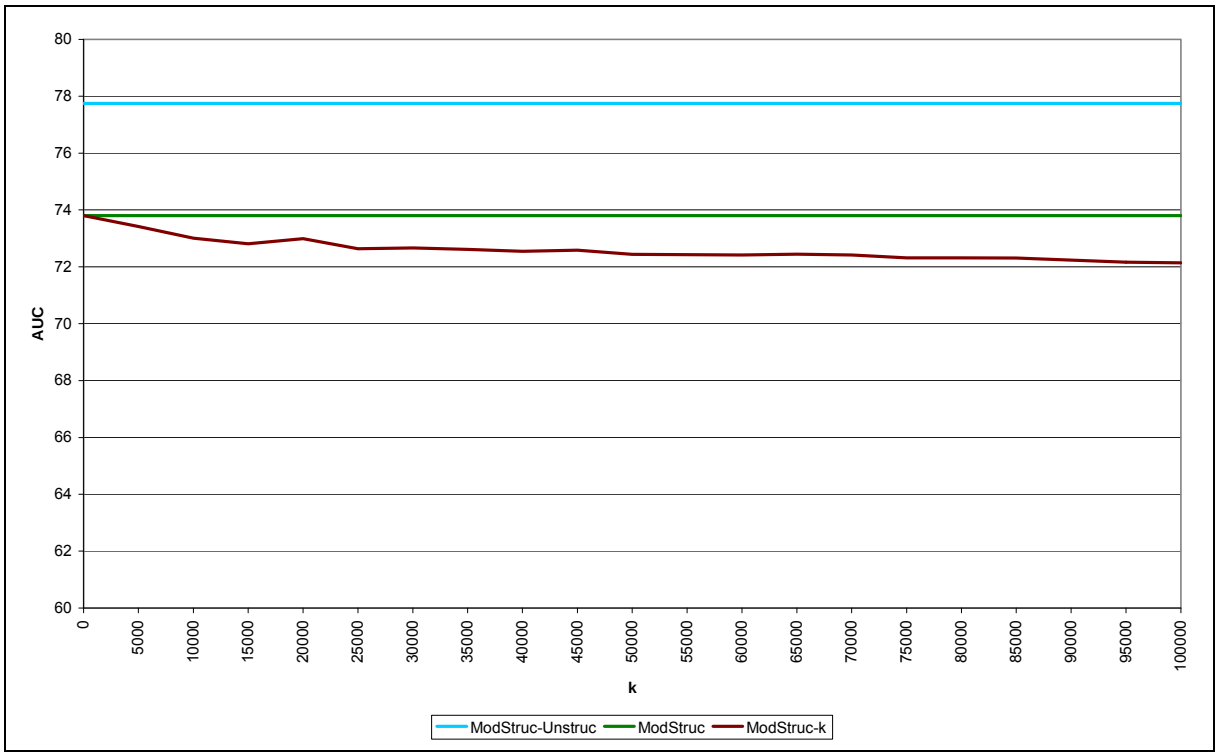


Figure 7: The AUC performance of ModStruc-Unstruc, ModStruc and ModStruc-k

As one observes from Figure 7, it was indeed better to build a separate model for subscribers who sent at least one email. The predictive performance of ModStruc was always higher than ModStruc-k. This clearly pointed out that subscribers from whom textual information was available have a unique churn pattern. The performance of ModStruc-Unstruc dominated those of ModStruc and ModStruc-k.

4. CONCLUSION

Adding the VOC by means of call center emails into a standard churn-prediction system helps marketing-decision makers to identify with a higher precision those customers most prone to switch. Consequently, retention campaigns to these customers can become more targeted. The framework integrated textual information from call center emails with traditionally-used marketing information. Converting the unstructured call center emails into a structured form suitable for churn prediction, required specialized pre-processing and dimension reduction steps.

Moreover, our study confirmed the importance of a well-considered email handling strategy. It provided a methodology that may increase the profitability of the call center by offering a model for

marketing-decision makers using customers of whom textual information is available. By enriching the churn model with this unstructured information from call center emails, marketing managers may improve the effectiveness of their retention campaigns.

ACKNOWLEDGMENTS

We would like to thank the anonymous Belgian company for their efforts in providing us with their data. Moreover, we also like to thank (1) BOF (01D26705) for funding the PhD project of Kristof Coussement, (2) BOF (011B5901) for funding the computing infrastructure, (3) Jonathan Burez, Bart Larivière and Ilse Bellinck for their insights and suggestions during this project. This project was realized using SAS v9.1.3, SAS Text Miner v5.2 and Matlab v7.0.4.

APPENDIX 1: TERM VECTOR WEIGHTING PHASE

The *tf* measures the frequency of occurrence of an index term in the email text. The more a term is present, the more important this term is in characterizing the content of that email. As such the frequency of occurrence of a content word is used to indicate term importance for content representation [4, 20, 30]. In our study, the *tf* was obtained by taking a logarithmic transformation of the original term frequency. Taking the logarithmic transformation reduced the importance of the raw *tf*, which was important for email collections of varying length.

The *idf* was incorporated so that the more rare a term occured in the collection of emails, the more discriminating it was. Therefore, the weight of a term was inversely related to the number of emails in which the term occured – i.e. the frequency of the term [14, 32] . The logarithm of the *idf* was taken to decrease the effect of the raw *idf*-factor.

Finally the weight of term *i* in an email *j* (w_{ij}) was given by

$$w_{ij} = tf_{ij}\, idf_i$$

with tf_{ij} equal to the term frequency of term *i* in email *j*, idf_i is equal to the inverse email frequency of term *i*.

Mathematically,

$$tf_{ij} = \log_2(n_{ij} + 1)$$

with n_{ij} equal to the frequency of term *i* in email *j* and

$$idf_i = \log_2\left(\frac{n}{df_i}\right) + 1$$

with *n* equal to the total number of emails in the entire email collection and df_i equal to the number of emails where term *i* was present.

APPENDIX 2: DIMENSIONS REDUCTION USING LSI VIA SVD

A high-dimensional term-by-email matrix A was constructed so that location (i,j) indicated w_{ij} the weight of term i for email j. SVD factorized A into three distinct matrices by

$$A = U \Sigma V^t \tag{5}$$

with Σ equal to a diagonal matrix containing the singular values of matrix A, U equal to the term-concept similarity matrix and V equal to the concept-email similarity matrix.

Mathematically, $\Sigma = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$ was the singular-values matrix where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r$. U and V were column-orthonormal matrices. The weights of the original matrix depended on the latent concepts by

$$w_{ij} = \sum_{x=1}^{r} U_{ix} \Sigma_x d_{jx} \tag{6}$$

LSI based on SVD allowed a simple strategy to approximate the original matrix A with rank r by $\hat{A}$ with rank k where $k \leq r$. Therefore LSI ignored the smaller lambda values in Σ by retaining only the first predetermined singular values equal to or greater than k, i.e. $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_k$, while only the first k columns of U and V were retained.

$$\hat{A}_k = U_k \Sigma_k V_k^t \tag{7}$$

with U_k, Σ_k and V_k were equal to the k-rank approximation of respectively U, Σ and V.

Matrix V_k is the approximated k-rank concept-email similarity matrix. A cell in the matrix V_k represented the loading for a specific email on one of the k concepts. This matrix contained information on how well a certain email loads on the different k concepts. The concepts reflected the hidden patterns in the textual data. Consequently, these concepts were used as explanatory variables in the churn-prediction model because they represented the latent semantic patterns of the textual information.

It is important that the concept loadings from the training vectors were comparable with those from the test vectors. The meaning of the concepts during testing should stay the same as those during training.

Consequently, emails of the test set were projected into the same semantic latent subspace as created during training.

In order to compare a test email *d* with the training emails, its term vector A_d was derived using the same pre-processing steps Deerwester et al. [12] proposed projecting each new term vector into the same latent semantic subspace as that created during training by

$$V_d = A_d' U_k \Sigma_k^{-1} \tag{8}$$

with U_k the *k*-rank concept-term similarity matrix and Σ_k the diagonal singular value matrix in rank *k*, both of the original SVD. V_d was the new concept-email vector which was comparable to the concept-email vectors of the matrix V_k.

However the choice of *k* was critical for optimal predictive performance.

APPENDIX 3: OVERVIEW OF STRUCTURED MARKETING INFORMATION

Client/company-interaction variables: variables describing the client/company relationship:

- The number of complaints,
- Elapsed time since the last complaint,
- The average cost of a complaint (in terms of compensation newspapers),
- The average positioning of the complaints in the current subscription,
- The purchase motivator of the subscription,
- How the newspaper is delivered,
- The number of conversions made in distribution channel, payment method & edition,
- Elapsed time since last conversion in distribution channel, payment method & edition,
- The number of responses on direct marketing actions,
- The number of suspensions,
- The average suspension length (in number of days),
- Elapsed time since last suspension,
- Elapsed time since last response on a direct marketing action,
- The number of free newspapers.

Renewal-related variable: variables containing renewal-specific information:

- Whether the previous subscription was renewed before the expiry date,
- How many days before the expiry date, the previous subscription was renewed,
- The average number of days the previous subscriptions are renewed before expiry date,
- The variance in the number of days the previous subscriptions are renewed before expiry date,
- Elapsed time since last step in company retention procedure,
- The number of times the customer did not renew a subscription.

Socio–demographic variables: variables describing the subscriber:

- Age,
- Whether the age is known,
- Gender,
- Physical person (is the subscriber a company or a physical person),
- Whether contact information (telephone, mobile number, email) is available.

Subscription-describing variables: group of variables describing the subscription:

- Elapsed time since last renewal,
- Monetary value,
- The number of renewal points,
- The length of the current subscription,
- The number of days a week the newspaper is delivered (intensity indication),
- Which edition the subscriber has (X1,X2,X3),
- The month of contract expiration.

REFERENCES

[1] M. Adria and S.D. Chowdhury, Centralization as a Design Consideration for the Management of Call Centers, Information and Management 41 (4) (2004), 497-507.

[2] K. Alajoutsijarvi, K. Mannermaaa and H. Tikkanen, Customer Relationships and the Small Software Firm: a Framework for Understanding Challenges Faced in Marketing, Information and Management 37 (3) (2000), 153-159.

[3] P.D. Allison, Logistic Regression using the SAS System: Theory and Application, Cary, NC: SAS Institute Inc. (1999).

[4] P.B. Baxendale, Machine-made Index for Technical Literature - an Experiment, IBM Journal of Research and Development 2 (4) (1958), 354- 361.

[5] C. Bell and K.P. Jones, Toward Everyday Language Information Retrieval Systems via Minicomputers, Journal of the American Society for Information Sciences 30 (1979), 334-338.

[6] M.W. Berry, Z. Drmac, E. Jessup, Matrices, Vector Spaces, and Information Retrieval, SIAM Review 41 (1999), 335-362.

[7] I. Bose and R.K. Mahapatra, Business Data Mining – a Machine Learning Approach, Information and Management 39 (3) (2001), 211-225.

[8] R.E. Bucklin and S. Gupta, Brand Choice, Purchase Incidence and Segmentation: an Integrated Modeling Approach, Journal of Marketing Research 29 (2) (1992), 201-215.

[9] J. Burez and D. Van den Poel, CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services Expert Systems with Applications, 32 (2) (2007), 277-288.

[10] R.G. Cooper and E.J. Kleinschmidt, Determinants of Timeliness in Product Development, Journal of Product Innovation Management 11 (5) (1994), 381–396.

[11] J.J. Cristiano, J.K. Liker and C.C. White, Customer-driven Product Development through Quality Function Deployment in the US and Japan, Journal of Product Innovation Management 17 (4) (2000), 286–308.

[12] S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science 41 (6) (1990), 391- 407.

[13] E.R. DeLong, D.M. DeLong and D.L. Clarke-Pearson, Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach, Biometrics 44 (3) (1988), 837-845.

[14] W.R. Greiff, A Theory of Term Weighting Based on Exploratory Data Analysis, in W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson and J. Zobel (eds.), Proceedings of the 21st SIGIR Conference New York:ACM (1998), 11-19.

[15] J.A. Hanley and B.J McNeil, The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve, Radiology 143 (1) (1982), 29-36.

[16] B. Karakostas, D. Kardaras and E. Papathanassiou, The State of CRM Adoption by the Financial Services in the UK: an Emperical Investigation, Information and Management 42 (6) (2005), 853-863

[17] S. Keaveney and M. Parthasarathy, Customer Switching Behavior in Online Services: an Exploratory Study of the Role of Selected Attitudinal, Behavioral and Demographic Factors, Journal of the Academy of Marketing Science 29 (4) (2001), 374-390.

[18] Y.S. Kim, Toward a Successful CRM: Variable Selection, Sampling and Ensemble, Decision Support Systems 41 (2) (2006), 542-553.

[19] W. Kraaij and R. Pohlmann, Viewing Stemming as Recall Enhancement, in Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland (1996), 40-48.

[20] H.P. Luhn, A Statistical Approach to Mechanized Encoding and Searching of Literary Information, IBM Journal of Research and Development 4 (4) (1957), 600-605.

[21] K. Matzler and H.H. Hinterhuber, How to Make Product Development Projects More Successful by Integrating Kano's Model of Customer Satisfaction into Quality Function Deployment, Technovation 18 (1) (1998), 25–38.

[22] G.A. Miller and E.B. Newman, Tests of a Statistical Explanation of the Rank-frequency Relation for Words in Written English, American Journal of Psychology 71 (23) (1958), 209-218.

[23] S.A. Neslin, S. Gupta, W. Kamakura , J. Lu and C. Mason, Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models, Journal of Marketing Research 43 (2) (2006), 204-211.

[24] S.L. Pan and J.N. Lee, Using e-CRM for a Unified View of the Customer, Communications of ACM 46 (4) (2003), 95-99.

[25] M. Pontes and C. Kelly, The Identification of Inbound Call Center Agents' Competencies that are Related to Callers' Repurchase Intentions, Journal of Interactive Marketing 14 (3) (2000), 41–49.

[26] W. Reinartz and V. Kumar, The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration, Journal of Marketing 67 (1) (2003), 77-99.

[27] G. Salton, A Theory of Indexing, Bristol, UK: J.W. Arrowsmith (1975).

[28] G. Salton, Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer, Reading MA: Addison-Wesley (1989).

[29] G. Salton and C. Buckley, Term-weighting Approaches in Automatic Text Retrieval, Information Processing & Management 24 (5) (1988), 513-523.

[30] G. Salton and M.J. Mcgill, Introduction to Modern Information Retrieval, New York:Mcgraw-Hill (1983).

[31] G. Salton, The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice Hall, Englewood Cliffs, NJ (1971).

[32] G. Salton and C.S. Yang, Specification of Term Values in Automatic Indexing, Journal of Documentation 29 (4) (1973), 351-372.

[33] K. Sparck Jones, A Statistical Interpretation of Term Specificity and its Application in Retrieval, Journal of Documentation 28 (1) (1972), 11-21.

[34] K. Sparck Jones, Index term weighting, Information Storage and Retrieval 9 (11) (1973), 619-633.

[35] I. Spiegler, Technology and Knowledge: Bridging a "Generating" Gap, Information and Management 40 (6) (2003), 533-539.

[36] D. Van den Poel and B. Larivière, Customer attrition analysis for financial services using proportional hazard models, European Journal of Operational Research 157 (1) (2004), 196–217.

[37] G.K. Zipf, Human Behaviour and the Principle of Least Effort, Cambridge, MA:Addison-Wesley (1949).

CHAPTER III

IMPROVING CUSTOMER COMPLAINT MANAGEMENT BY AUTOMATIC EMAIL CLASSIFICATION USING LINGUISTIC STYLE FEATURES AS PREDICTORS

This chapter is based on K. Coussement and D. Van den Poel, Improving Customer Complaint Management by Automatic Email Classification Using Linguistic Style Features as Predictors, Decision Support Systems, 44 (4) (2008), pp 370-382.

CHAPTER III

IMPROVING CUSTOMER COMPLAINT MANAGEMENT BY AUTOMATIC EMAIL CLASSIFICATION USING LINGUISTIC STYLE FEATURES AS PREDICTORS

ABSTRACT

Customer complaint management is becoming a critical key success factor in today's business environment. This study introduces a methodology to improve complaint handling strategies through an automatic email classification system that distinguishes complaints from non-complaints. As such, complaint handling becomes less time-consuming and more successful. The classification system combines traditional text information with new information about the linguistic style of an email. The empirical results show that adding linguistic style information into a classification model with conventional text-classification variables results in a significant increase in predictive performance. In addition, this study reveals linguistic style differences between complaint emails and others.

1. INTRODUCTION

Due to the rapid development of information technology and Internet, new opportunities arise for marketing analysts nowadays. For instance, companies can easily advertise through the email channel [10] or offer products in an electronic commerce [14]. This study focuses on the usefulness of client/company interactions through email as the basis for improved customer complaint management. Nowadays, companies receive daily huge amounts of emails due to the fact that their clients become more used to sending emails as a substitute for traditional communication methods [29]. Growingly, efficient email handling is becoming a critical key success factor in today's business environment. Recently, companies start to outsource their customer-email management by relying on customer-call centers to address the voice of customers - i.e. customer complaints and service-information requests [20].

Indeed, Internet enables customers to easily express their problems with a product or a service. Consequently, customer complaint management and service recovery are becoming key drivers for improved customer relationships. Several studies have shown the positive financial impact of

investments in efficient complaint handling in a wide range of industries [13][28]. It is crucial that service-recovery efforts are forceful and effective [4], because ample research has shown that failed service-recovery actions have a significant influence on customer-switching behaviour [26].

A tool to support efficient processing of customer-complaint emails is the use of an automatic email-classification system. Automatic text classification labels incoming emails into predefined categories – i.e. complaints versus non-complaints in this study. As a consequence, customer-complaint management becomes more successful in mainly two ways: (i) In contrast to manual text classification, automatic text classification is time-saving and thus less expensive in terms of labor costs. It makes the email-handling process more efficient. (ii) By classifying incoming emails into complaints and non-complaints, one can optimize the complaint-handling process. By making a distinction between complaint emails and other email types (e.g. information requests on promotional deals), the company is able to set up a separate complaint handling department with specially-trained complaint handlers. One can create a separate treatment procedure for complaint emails. Consequently, call centers can react more helpful on occurring problems or service failures. In general, the consistency in the way complaint emails are handled increases due to the fact that not every employee in the call center needs to be trained for all email types. In summary, building an automatic email classification system that distinguishes complaints from non-complaints is necessary for optimizing the complaint-handling process within a call center.

Automatic text-classification systems are typically built using the conventional vector-space approach proposed by [24] (e.g. [16], [6] and [1]). This means that every email is converted into a vector that contains a stream of words or terms. This is often a very high-dimensional vector due to the many distinct terms in the email corpus. This study employs Latent Semantic Indexing by means of Singular Value Decomposition as proposed by [7] to reduce the dimensionality. Consequently, textual information is represented as k distinct concepts or explanatory variables. Within this study, linguistic style information is introduced as a new type of textual information. Moreover, linguistic style differences between complaint emails and other emails are explored. Furthermore, the beneficial effect of adding linguistic style characteristics to a traditional complaint-classification system is investigated.

This study offers marketing managers a valuable system for automatic email classification in order to enable them to optimize the client/company relationship through efficient and effective complaint handling. Moreover, it introduces linguistic style characteristics of an email as a new type of textual information in a customer complaint setting. Accordingly, this study contributes to the existing literature in two ways: (i) it shows that adding these linguistic style features into a conventional email-classification model results in an additional increase in predictive performance in distinguishing

complaints from non-complaints. (ii) Moreover, this study proves that linguistic style differences exist between complaints and others.

This paper is organized as follows. Section 2 describes the methodology used throughout this study. Section 3 describes how the proposed framework is applied and evaluated within a real life call-center setting. In a last Section, the findings of this study are summarized, while also some shortcomings and directions for further research are given.

2. METHODOLOGY

This Section describes the methodology used throughout this study. In Section 2.1 the content-based approach which is often used in traditional text-classification problems is explained. In this study, a new type of information is used to predict whether the incoming email is a complaint or not – i.e. linguistic style information of an email. Section 2.2 gives an overview how the linguistic style features are extracted from the email corpus. Section 2.3 gives an overview of the classification technique. In order to compare the performance of the different classification models, some objective evaluation criteria are needed. The evaluation criteria used throughout this study are covered in Section 2.4.

2.1. Vector-space approach

This Section gives an overview of the conventional text-classification approach using the vector-space approach proposed by [24]. Original documents are converted into a vector in a feature space based on the weighted term frequencies. Each vector component reflects the importance of the corresponding term by giving it a weight if the term is present or zero otherwise. The purpose is to select the most informative terms from the number of distinct terms in the corpus dictionary. All documents are traditionally converted from the original format to word vectors following the steps as shown in Figure 1.

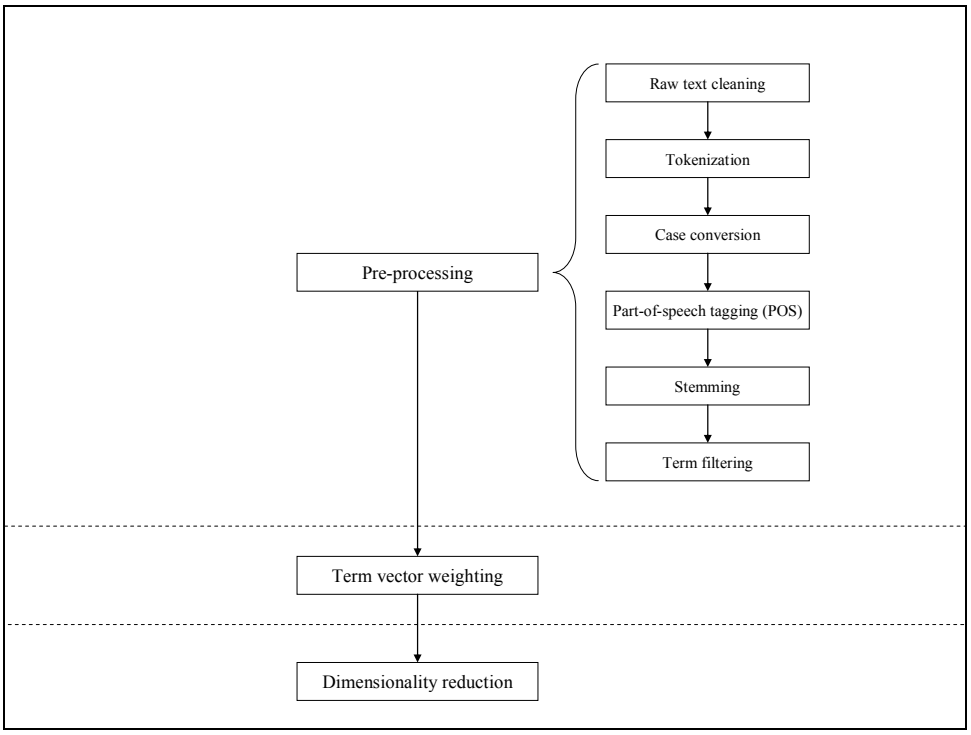


Figure 1: An overview of the conventional vector-space approach for text classification.

2.1.1. Pre-processing

Raw text cleaning converts documents into a form which is more suitable for subsequent processing. In this step, special characters and punctuations are separated from words, while spelling errors are handled by comparing all words in the document with a reference dictionary.

During the *tokenization* step, documents are divided into tokens or words, by which white space characters are used as separators. Once the text field is divided into words, words are converted to lower case – i.e. *case conversion*.

All words are tagged a *part of speech* based on their syntactic category. All words are summarized into informative and non-informative parts of speech. The non-informative parts of speech contain determiners, conjunctions, auxiliaries, prepositions, pronouns, negative articles or possessive markers, interjections, proper nouns, abbreviations and numbers. On the other hand, words can be part of an informative part of speech like nouns, verbs, adjectives and adverbs.

A next step in the text-preprocessing phase is *stemming* or lemmatization. Word variations are conflated into a single representative form, called the stem. A typical example of a stem is the word ‘connect’ which is the stem for the variants ‘connected’, ‘connecting’, ‘connection’ and ‘connections’. Stemming has two advantages: it reduces the corpus dictionary enormously [3] and it increases the

retrieval performance significantly [15]. A dictionary-based stemmer is used throughout this study. The huge advantage is that all morphological variations are treated naturally by comparing them with a reference dictionary. When a corpus term is unrecognizable, the stemmer applies some standard decision rules to give the term the correct stem.

In order to reduce the number of irrelevant terms in the corpus dictionary, a number of *term filtering* tasks are performed. Firstly, rare words are left out from further analysis because they are unable to aid in future classification. Consequently, all words appearing less then three times over the entire document corpus are eliminated for further analysis. Additionally, overly common words like for instance 'a' or 'the' are also removed from the corpus dictionary. These type of words or stopwords appear so often that they are not discriminative anymore. A stoplist is language and domain specific, as a consequence a standard stoplist is often manually adapted to avoid the risk of removing relevant words. Furthermore, only words that are part of an informative part of speech are included, because these words contain relevant information to aid in future classification. In the end, the temporary dictionary is manually checked and irrelevant words are removed from the dictionary.

The result is a high-dimensional term-by-document matrix where each cell in the matrix represents the raw frequency of appearance of a term in a document. To correct for the importance of a term in a document and its importance in the corpus dictionary, the term vectors in the term-by-document matrix are weighted.

2.1.2. Term-vector weighting

In the term vector weighting phase, a weighted term vector for every document in the document collection is constructed. Right now, the values in the term-by-document matrix are simply the raw frequencies of appearance for a term in a document. Term-vector weighting is often done by determining the product of the term frequency (*tf*) and the inverse document frequency (*idf*) [27][21][23][22].

The *tf* measures the frequency of occurrence of an index term in the document text [23]. The more a term is present in a document, the more important this term is in characterizing the content of that document. As such the frequency of occurrence of a content word is used to indicate term importance for content representation e.g. [21] and [22]. In this study, the *tf* is obtained by taking a logarithmic transformation of the original term frequency. Taking the logarithmic transformation reduces the importance of the raw *tf*, which is important for document collections with a varying length. The term frequency of term i in document j (tf_{ij}) is given by

$$tf_{ij} = \log_2(n_{ij} + 1) \quad (1)$$

with n_{ij} equal to the frequency of term *i* in document *j*

The *idf* takes into account that the more rarely a term occurs in a document collection, the more discriminating that term is. Therefore, the weight of a term is inversely related to the number of documents in which the term occurs – i.e. the document frequency of the term [21],[22] and [27]. The logarithm of the *idf* is taken to decrease the effect of the raw *idf*-factor. The inverse document frequency of term *i* (idf_i) is given by

$$idf_i = \log_2\left(\frac{n}{df_i}\right) + 1 \quad (2)$$

with *n* equal to the total number of documents in the entire document collection and df_i equal to the number of documents where term *i* is present.

Finally, the weight of term *i* in document *j* (w_{ij}) is given by

$$w_{ij} = tf_{ij}\ idf_i \quad (3)$$

with tf_{ij} equal to the term frequency of term *i* in document *j*, idf_i equal to the inverse document frequency of term *i* .

2.1.3. Dimensionality reduction

This weighted term-by-document matrix is a high-dimensional matrix due to the many distinct corpus terms. Moreover, this matrix is very sparse – i.e. it contains a lot of zeros -, since not all documents contain all corpus terms. In order to reduce the dimensionality of the feature space, this study employs Latent Semantic Indexing by using Singular Value Decomposition (SVD) as proposed by [7]. Latent Semantic Indexing projects documents from the high-dimensional term space to an orthonormal, semantic latent subspace by grouping together similar terms into several distinct concepts *k*. All textual information can be summarized into *k* concepts. Furthermore, these *k* concepts or SVD variables are often used as explanatory variables in a traditional text-classification model. In summary, one concludes that Latent Semantic Indexing approximates the original weighted term-by-document matrix in a smaller rank *k* – i.e. *k* concepts or variables that summarizes the emails content - which

makes it workable from a prediction point of view. Factor-analytic literature proposes an operational criterion to find the optimal value for *k* [7].

2.2. Linguistic style features

The linguistic style features are introduced as a new set of text classification predictors. These variables are created using Linguistic Inquiry and Word Count [19][30]. This program searches individual text files, while it computes the percentage of words that were earlier judged to reflect the linguistic categories. These categories are described using an extensive dictionary. The percentage of the total words in the email on a specific category is used as an additional feature in the text-classification system. In sum, a detailed overview of the linguistic style categories is given in Table 1.

Abbreviation	Dimension	Examples	Number of words
WC	Word Count		
WPS	Words per sentence		
Qmarks	Sentences ending with ?		
Unique	Unique words (type/token ratio)		
Sixltr	% words longer than 6 letters		
Pronoun	Total pronouns	I, our, they, you're	38
I	1st person singular	I, my, me	7
We	1st person plural	we, our, us	6
Self	Total first person	I, we, me	13
You	Total second person	you, you'll	7
Other	Total third person	she, their, them	12
Negate	Negations	no, never, not	36
Assent	Assents	yes, OK, mmhmm	50
Article	Articles	a, an, the	3
Preps	Prepositions	on, to, from	48
Number	Numbers	one, thirty, million	107
Time	Time indication	hour, day, oclock	269
Past	Past tense verb	walked, were, had	1773
Present	Present tense verb	walk, is, be	1886
Future	Future tense verb	will, might, shall	19

Table 1: Overview of the linguistic style features extracted from the call-center emails.

2.3. Classification technique

Boosting is used as the main classification technique for discriminating complaints from non-complaints throughout this study (see Section 3.3 and 3.4). It is a relatively young, yet extremely powerful machine learning technique. The main idea behind boosting algorithms is to combine the outputs of many "weak" classifiers to produce a powerful "committee" or ensemble of classifiers [12]. Several studies show that ensembles generally achieve a significantly lower error rate than the best

single model (e.g. [18]). Although being refined subsequently, the main idea of all boosting algorithms can be traced back to the first practical boosting algorithm, Adaboost [9]. Adaboost and related algorithms produce extremely competitive results to other classification algorithms in many settings, most notably for text classification (e.g. [25]).

This study concisely describes Adaboost for a two class classification problem. For more details about Adaboost, we refer to [12]. Consider a training set T={(x_i,y_i)} with i={1,2,…,N}; the input data $x_i \in R^n$ and corresponding binary target labels coded as $y_i \in$ {-1,1}. The final classifier of the Adaboost procedure is given by

$$F(x)=\sum_{m=1}^{M} c_m f_m(x) \qquad (4)$$

with *m* the number of iterations, $f_m(x)$ the classifier predicting values +/-1 during the m^{th} round and c_m the weight of the contribution of each $f_m(x)$ in the final classifier. The purpose of Adaboost is to train classifiers $f_m(x)$ on weighted versions of the training data obtained by modifying the data at each boosting step by applying weights $w_1, w_2, \ldots, w_N$ to each of the training observations (x_i,y_i) with i={1,2,…,N}. Initially all the weights are set to $w_i=\frac{1}{N}$, so the first step simply trains the classifier on the data in the usual manner. For each successive iteration *m*=2,3,…M, (i) the weights are individually modified giving a higher weight to cases that are currently misclassified and (ii) the classifier is reapplied to the weighted observations. Thus as the iterations proceed, observations that are difficult to classify receive ever-increasing influence due to the higher weight assigned. So each successive classifier is thereby forced to concentrate on those training observations that are missed by previous ones in the sequence. In the end when the maximum number of iterations M is reached, predictions from all classifiers are combined through a weighted majority vote to produce the final classifier F(x).

In order to give the reader more insights into the linguistic style differences between complaints and other email types (see Section 3.2), a stepwise logistic regression that differentiates between complaint emails and other email types is run using the linguistic style variables as described in Table 1. This technique is used because it is a conceptually simple binary classifier [5], while it provides standardized parameter estimates for the explanatory variables.

2.4. Evaluation criteria

In order to evaluate the performance of different predictive models, two criteria are used: the percentage correctly classified (PCC) and the area under the receiving operating curve (AUC). The PCC compares the a posteriori probability of being a complaint email with the true type of the email. If TP, FP, TN and FN are respectively the number of complaints that are correctly predicted (True Positives), the number of non-complaints that are predicted as complaints (False Positives), the number of non-complaints that are classified correctly (True Negatives) and the number of complaints that are predicted as non-complaints (False Negatives), the PCC is defined as (TP+TN)/(TP+FP+TN+FN). The PCC should be benchmarked to the proportional chance criterion ($=\text{percentage}_{event}^2 + (1-\text{percentage}_{event})^2$) in order to confirm the predictive capabilities of a classifier [17]. A disadvantage of the PCC is that it is not very robust concerning the chosen cut-off value on the a posteriori probabilities [2]. In order to equally compare different classification models on PCC, the cut-off value for classifying emails into complaints or non-complaints is chosen so that the a posteriori incidence equals the a priori occurrence of complaints. For instance, 35% of the emails having the highest complaint probability will be classified as complaints when the a priori frequency of complaints is 35%. In contrast to PCC, AUC takes into account all possible thresholds on the a posteriori probabilities. For all the different levels, it considers the sensitivity (TP/(TP+FN)) and 1 minus the specificity (TN/(TN+FP)) in a two-dimensional graph, named the receiving operating curve (ROC). The area under the ROC curve is used to evaluate the performance of a binary classifier [11]. [8] propose a non-parametric test to compare the performance of different classification models.

3. EMPIRICAL RESULTS

This Section applies the proposed framework in a real-life call center setting. In a first Section, detailed information concerning the call-center setting is given. Section 3.2 explores the linguistic style differences between complaint emails and other email types, while Section 3.3 investigates the beneficial effect of including linguistic predictors into a traditional text-classification setting. In a last Section, the robustness of the proposed methodology is investigated.

3.1. Corpus construction

In this study, emails sent to the call center of a large Belgian newspaper company are used. Subscribers of this newspaper have the possibility to send their concerns or complaints and information requests to the call center via email. When an email message comes in, the message is manually encoded into a complaint or another email type. The former email type reports all different service failures (e.g. newspaper not delivered, financial complaints …), while the latter type consists

of information requests like subscription related questions or information on promotional actions. Manually encoding email messages is a very time-consuming and very inefficient task. Moreover, all types of emails are treated equally within the same department, while in fact complaint emails need a different treatment by specialized people during email handling. The email classification problem in this context comes down to predicting whether the incoming email is a complaint or not. Consequently, complaint handling becomes more efficient due to a faster detection of the emails at risk.

All emails from July 2004 till December 2004 are used within this study. Consequently, it is possible to derive the dependent and the explanatory variables. Because historical data is used, all email messages are already manually encoded by the staff of the call center. The dependent variable is encoded as '1' when the email is a complaint and '-1' otherwise. There are two types of independent variables. The first type of explanatory variables is extracted using the methodology of the vector-space approach. Email messages are converted into a high-dimensional term-by-document matrix. However, this matrix is unworkable from a prediction point of view due to the large number of terms or variables. Consequently, several reduced rank-k models (with k={10,20,…,200}) are obtained by applying Latent Semantic Indexing using SVD. The second type of independent variables is derived by processing all emails through Linguistic Inquiry and Word Count. These independent variables represent word counts for the different categories derived from the linguistic program. These variables which contain information about the linguistic styles are used to explore their beneficial effect on top of the traditionally-used SVD variables.

Intended to methodologically correctly predict whether an email is complaint or not, the data set is divided into training, test and validation set. Emails between July 2004 and November 2004 are randomly assigned using a 70-30 split to the training and test set. The former one is used to generate and train the classifiers, while the test set is used to test the classifier on an unseen data sample. In this study, all emails of December 2004 are assigned to the validation set or out-of-period set. This dataset is used to verify the robustness of the proposed methodology. Table 2 summarizes the characteristics of the different data sets.

	Number of emails	Relative percentage
Training set		
Complaint emails	1890	36.37%
Others	3306	63.63%
Total	5196	100%
Test set		
Complaint emails	838	37.61%
Others	1390	62.39%
Total	2228	100%
Validation set		
Complaint emails	571	32.59%
Others	1181	67.41%
Total	1752	100%

Table 2: Overview of the data characteristics.

As such, one is able to explain differences in linguistic style between complaint emails and other email types using a traditional stepwise logistic regression (see Section 3.2). Furthermore, a comparison is made in predictive performance between the models built with the k concepts (with k={10,20,…,200}) extracted using the SVD procedure (i.e. Adaboost model with SVD variables or ADA_SVD), the model using only the linguistic style predictors (i.e. Adaboost model with linguistic style feature or ADA_LS) and the model using both types of information or ADA_SVD_LS (see Section 3.3).

3.2. Linguistic style differences between complaint emails and other email types

This Section explores if linguistic style differences exist between a complaint email and another email type. Table 3 shows the standardized parameter estimates of the linguistic style variables kept during the stepwise logistic regression whereby one tries to predict whether the received email involves a complaint or not using only the linguistic style variables as described in Table 1.

Linguistic style		Standardized parameter estimates*
Abbreviation	**Dimension**	
Article	Articles	0,1438
Future	Future tense verb	-0,0897
I	1st person singular	-0,1225
Negate	Negations	0,4886
Number	Numbers	0,1078
Other	Total third person	-0,0548
Past	Past tense verb	0,1009
Preps	Prepositions	-0,4739
Present	Present tense verb	0,0628
Qmarks	Sentences ending with ?	-0,0833
Sixltr	% words longer than 6 letters	-0,1295
Time	Time indication	0,2122
WC	Word Count	0,1758
You	Total second person	0,0759

* = all parameter estimates are significant at 95% confidence level.

Table 3: Standardized parameter estimates.

Table 3 clearly indicates that the more words (WC), the more articles (Article) and the less prepositions (Preps) are found in an email; the more likely, the email is classified as a complaint. In contrast to other types of emails (e.g. an information request), the probability of being a complaint increases when more time indications (Time) are found in an email. Moreover, the likelihood of being a complaint is positively related with the present tense (Present) – *e.g. Hopefully, you can fix the misdelivery of my newspaper today (Present, Time)* – and the past tense (Past) - *e.g. Moreover, the newspaper was not delivered last week either (Past, Time, Time)* - , while the possibility of classifying an email as an information request increases when more future tenses (Future) and questions (Qmarks) are used – *e.g. Will there be a reduction on my next subscription? (Future, Qmarks)*. When the tone of an email is more 'aggressive' – i.e. it contains more negations (Negate), more numbers (Numbers) and more clenched words (Sixltr) - the chance of having a complaint increases. Furthermore, complainants often directly blame the company for the service failure (I, Other, You) – *e.g Dear, the newspaper is not delivered today. It is already the sixth time this month. You must have noticed already some delivery problems. You have to fix this problem as soon as possible.(Negate, Numbers, You, You)*. These results indicate that differences in linguistic style exist between complaints and non-complaints.

3.3. Comparing predictive performance of ADA_SVD, ADA_LS and ADA_SVD_LS

This Section compares the predictive performance of ADA_SVD, ADA_SVD_LS and ADA_LS in terms of AUC and PCC. Figure 2 shows the predictive performance of the different models on the test set in terms of AUC, while a comparison in terms of PCC is shown in Figure 3. The number of SVD concepts is represented on the X-axis, while on the Y-axis, the performance measure is shown. As a remark; (i) ADA_LS is a horizontal line because its performance is independent of the number of SVD concepts, but it is incorporated in the Figures for comparison reasons only, (ii) Appendix A includes ROCs on the test set for the models with k = {50,100,150,200} in order to provide the reader with more in-depth information.

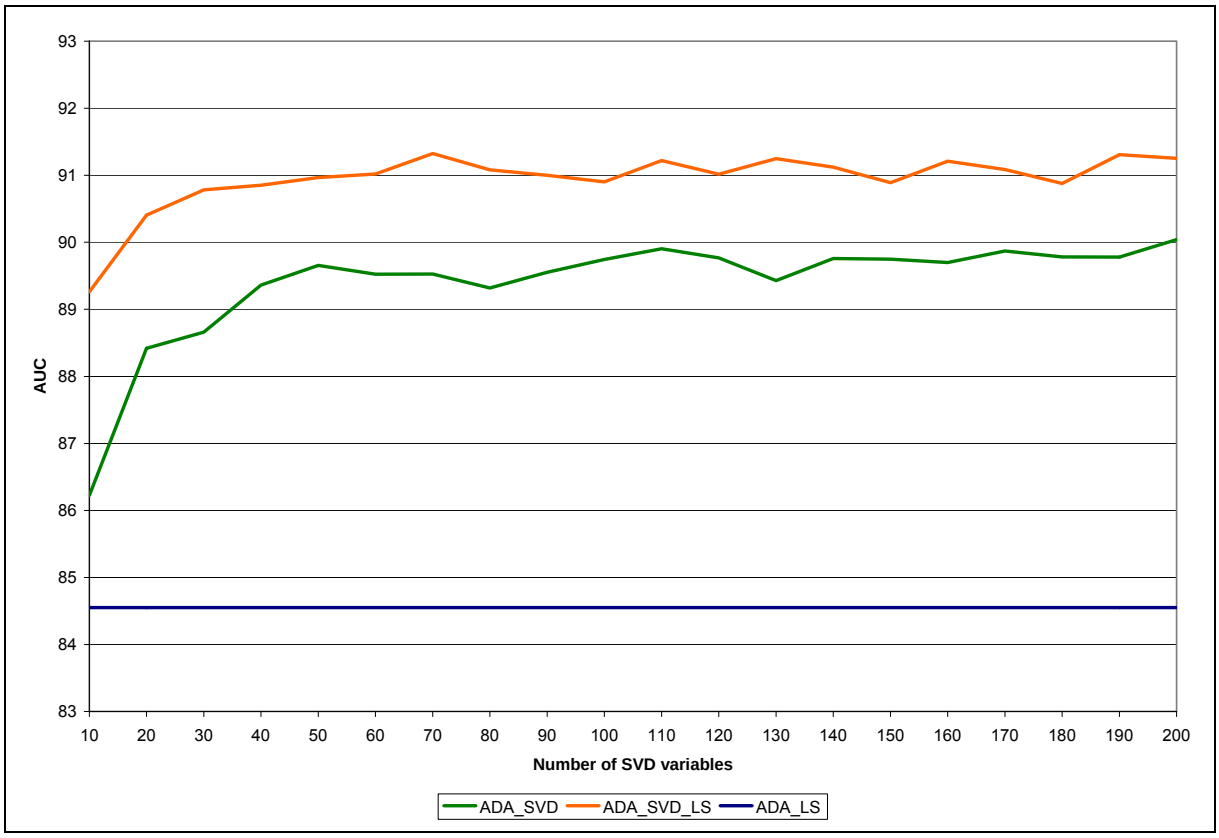


Figure 2: The AUC performance on the test set of ADA_SVD, ADA_SVD_LS and ADA_LS.

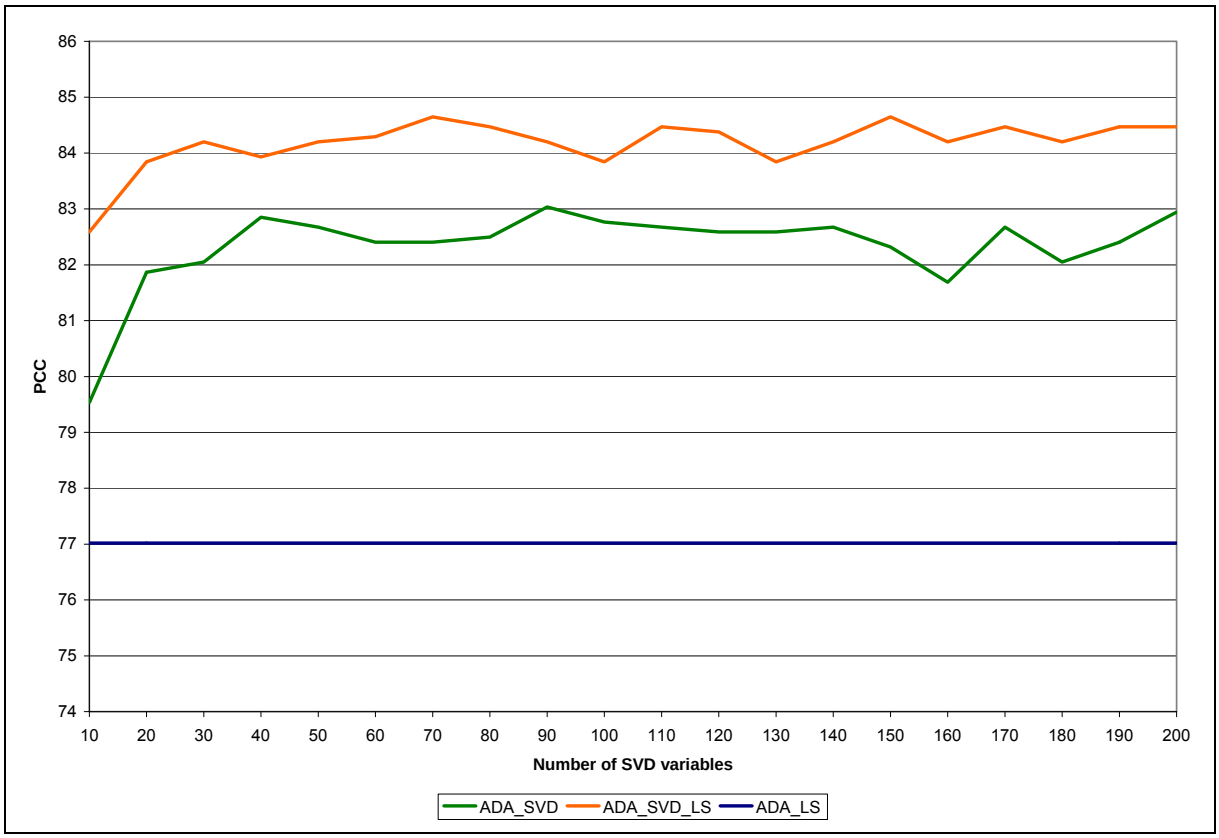


Figure 3: The PCC on the test set of ADA_SVD, ADA_SVD_LS and ADA_LS.

As one observes from Figure 2 and Figure 3, all models perform enormously well in distinguishing complaint emails from non-complaint emails. Indeed, the AUC performance of all models lies between 84.55 and 91.32, while the PCC lies in the range of 77.02 and 84.65 which clearly outperforms the proportional chance criterion of 53.07 ($= 0.3761^2 + (1-0.3761)^2$) [17]. These results clearly indicate that all models – i.e. ADA_SVD, ADA_SVD_LS and ADA_LS – have predictive capabilities in distinguishing complaints from other emails.

Figure 2, Figure 3 and Table 4 give an answer to the question if adding additional linguistic style predictors to the traditional SVD dimensions is beneficial from a text classification point of view. In other words, does ADA_SVD_LS significantly outperform ADA_SVD and ADA_LS?

ADA_SVD_LS-ADA_SVD																				
SVD	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*100*	*110*	*120*	*130*	*140*	*150*	*160*	*170*	*180*	*190*	*200*
AUC difference	*3.038*	*1.988*	*2.122*	*1.490*	*1.310*	*1.495*	*1.797*	*1.762*	*1.448*	*1.159*	*1.317*	*1.250*	*1.820*	*1.365*	*1.143*	*1.512*	*1.214*	*1.093*	*1.528*	*1.217*
$\chi 2$	*42.508*	*27.385*	*31.337*	*19.307*	*15.103*	*20.315*	*26.283*	*27.261*	*18.328*	*12.139*	*13.376*	*14.104*	*29.678*	*15.290*	*11.127*	*19.811*	*12.028*	*10.442*	*18.294*	*12.796*
df	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
p	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

ADA_SVD_LS-ADA_LS																				
SVD	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*100*	*110*	*120*	*130*	*140*	*150*	*160*	*170*	*180*	*190*	*200*
AUC difference	*4.710*	*5.857*	*6.233*	*6.301*	*6.417*	*6.469*	*6.773*	*6.530*	*6.451*	*6.353*	*6.670*	*6.466*	*6.699*	*6.571*	*6.341*	*6.659*	*6.537*	*6.326*	*6.756*	*6.705*
$\chi 2$	*73.6034*	*98.172*	*111.069*	*109.766*	*111.647*	*112.022*	*126.516*	*113.330*	*107.168*	*108.200*	*118.390*	*109.018*	*116.827*	*114.625*	*102.791*	*115.494*	*112.497*	*106.401*	*119.710*	*118.243*
df	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
p	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Table 4: AUC significance test statistics of Delong et al. (1988) for ADA_SVD_LS – ADA_SVD and ADA_SVD_LS – ADA_LS.

Table 4 indicates that ADA_SVD_LS significantly outperforms ADA_LS and ADA_SVD on all SVD dimensions [8]. Figure 2 confirms these results graphically in terms of AUC. Furthermore, the PCC of ADA_SVD_LS is always higher then ADA_SVD and ADA_LS as can be seen in Figure 3. These results indicate the highly beneficial impact of combining traditional SVD predictors with linguistic style indicators into one text-classification model. As such, predictive modelers are able to build better email-classification models by incorporating this new type of information.

3.4. Out-of-period validation

In order to verify the robustness of the proposed methodology, all models are scored on an out-of-period validation set. This is necessary because in a realistic call-center environment, the incoming emails lie by definition in another timeframe than the ones used during model training. As such, one is able to verify if the models built during training are still valid when validating them on another timeframe. If the proposed methodology is robust, the performance on the test set and validation set has to be stable. Figure 4 and Figure 5 illustrate the performance stability between the test and validation set. On the X-axis, the number of SVD variables is shown, while the Y-axis indicates the performance. The solid line represents the test-set performance, while the dashed line represents the validation-set performance. Additionally, ROCs on the validation set for the classification models with k = {50,100,150,200} are presented in Appendix B.

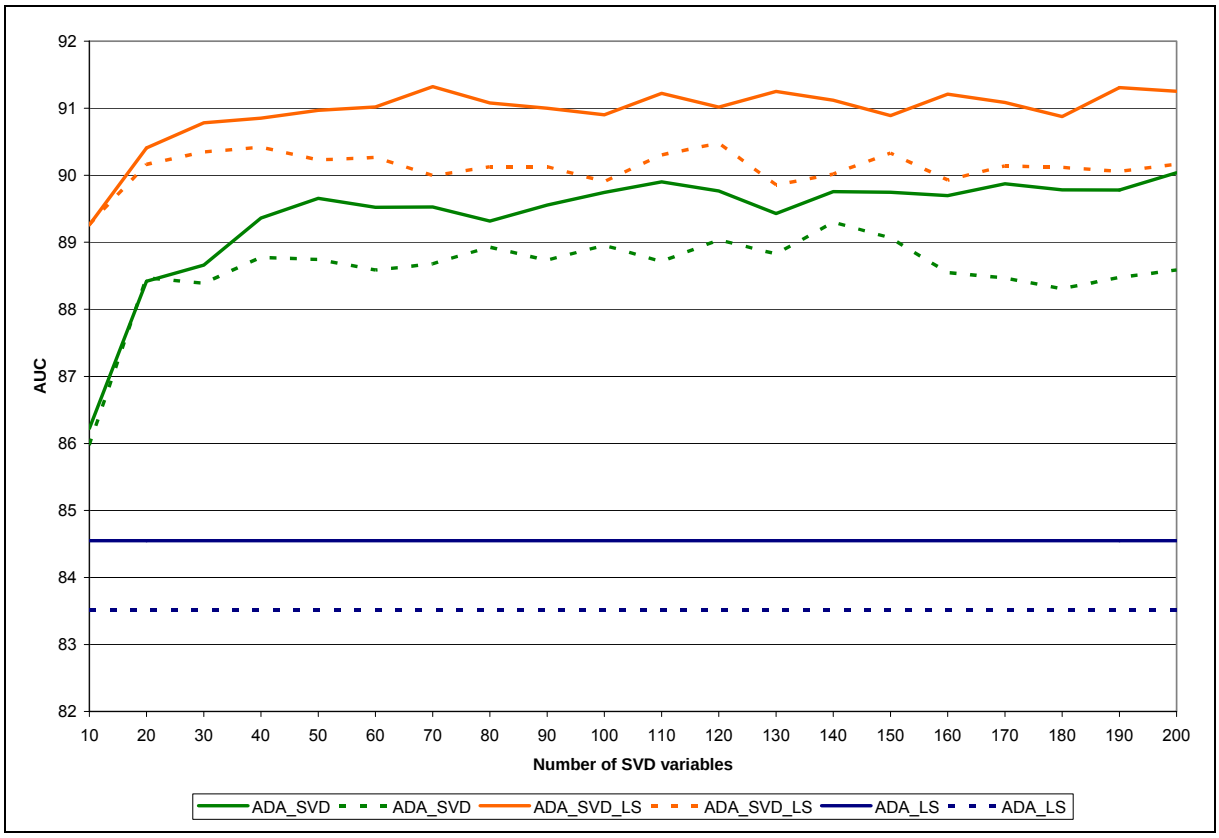


Figure 4: The AUC performance on the test (solid line) and validation set (dashed line) of ADA_SVD, ADA_SVD_LS and ADA_LS.

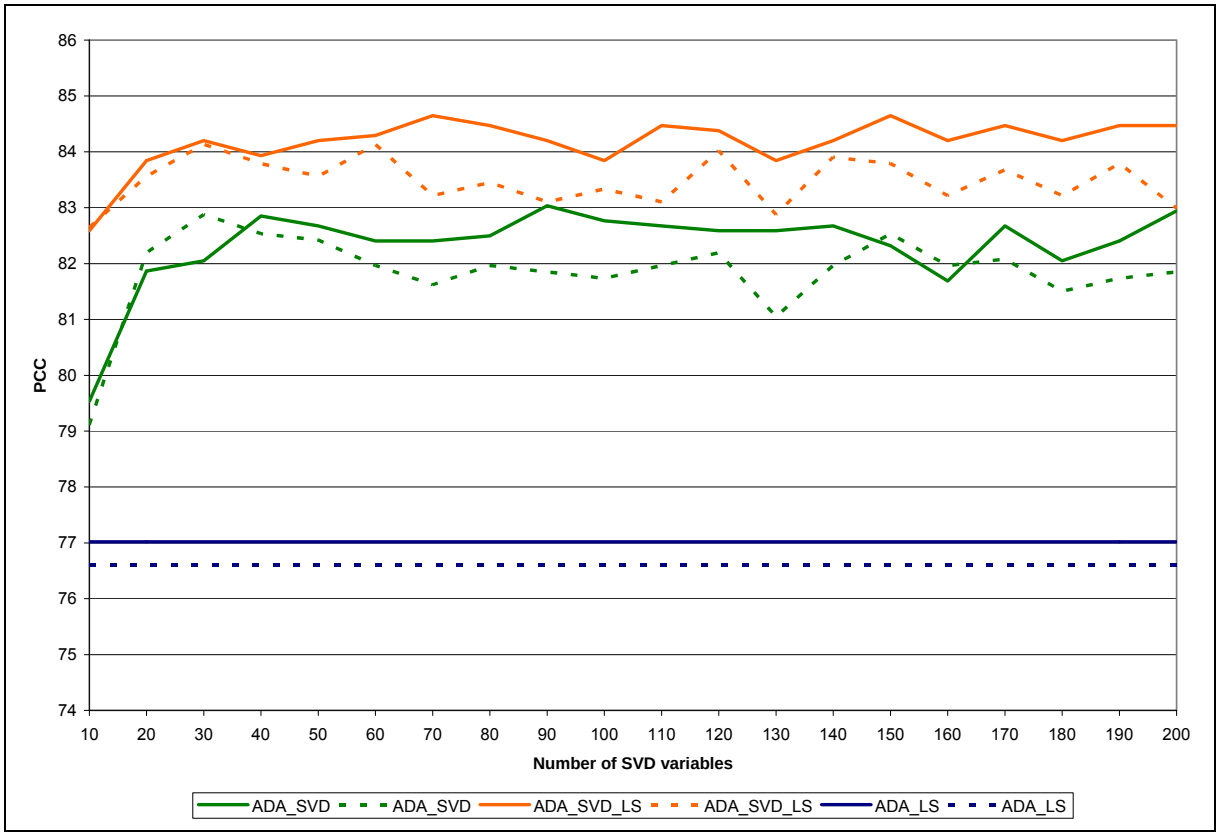


Figure 5: The PCC on the test (solid line) and validation set (dashed line) of ADA_SVD, ADA_SVD_LS and ADA_LS.

Figure 4 and Figure 5 prove that the proposed methodology is valuable, stable and extendible to other timeframes. The conclusions as drawn in Section 3.3 are still valid on the out-of-period validation set. ADA_SVD, ADA_LS and ADA_SVD_LS perform also very well on the validation set. The AUC lies in the range of 83.51 and 90.48, while the PCC of all models exceeds the proportional chance criterion of 56.06 ($= 0.3259^2 + (1-0.3259)^2$) [17]. Figure 4 and Figure 5 show that ADA_SVD_LS significantly outperforms ADA_SVD and ADA_LS on the validation set in terms of AUC and PCC.

Furthermore, the robustness of this complaint management system is confirmed by comparing the performance on the test set with that on the out-of-period validation set. The absolute value of the AUC difference between the test set and the out-of-period validation set fluctuates between 0.008 and 1.476 AUC points over all different models (see Figure 4). This indicates that the models are robust over the proposed time period. These results are confirmed when having a look at the absolute value of the difference in PCC (see Figure 5). This difference fluctuates between 0.063 and 1.535 PCC points.

In summary, applying these classifiers to a new timeframe does not result in a drastic drop in performance. Contrary, implementing the proposed methodology within a call center environment is a valuable strategy to improve customer complaint management.

4. CONCLUSIONS AND DIRECTION FOR FURTHER RESEARCH

Due to the strong increase in Internet penetration, a lot of customers write an email as a substitute for traditional communication methods as for instance a letter or a telephone call. As a consequence, companies receive daily a huge amount of emails. Nowadays, companies outsource their internal email management to a specialized call center environment. Efficient email handling becomes one of the major key challenges in business. This study focuses on how a company can optimize its complaint handling strategies through an automatic email-classification system. Indeed, practitioners and academics feel the need for an efficient and successful complaint handling strategy, because recovering service failures as quick as possible results in additional benefits.

This study offers an automatic email classification system that distinguishes complaints from non-complaints. In contrast to manually encoding the incoming emails, this study offers a feasible methodology to automate this process. As a result, email handling becomes less time-consuming and less expensive due to the lower labor costs. Table 5 indicates that implementing the current methodology in a real-life call center setting saves a lot of resources within this specific case.

Call Center Setting	Discounted Cost per Year (in Euro)					Total Cost after 5 year (in Euro)	Additional Savings over Manual Labeling after 5 year (in Euro)
	Year						
	1	2	3	4	5		
Manual Labeling	7,250	6,971	6,703	6,445	6,197	33,567	
Automatic Email Classification	1,813	1,743	1,676	1,611	1,549	8,392	25,175

Table 5: Real-life call center example.

Suppose that in the current situation, an employee labels an incoming email at a realistic 45 sec per email or 80 email messages per hour, whereas the explicit cost of re-labeling falsely called complaints as non-complaints or visa versa is similar. Moreover, the call center receives about 20,000 emails a year. Table 5 shows that the total costs over 5 years with 20,000 email messages a year for manually labeling is about 33,567 Euros having a discount rate of 4% and a gross employee cost of 29 Euros per hour. If the call center implements the proposed methodology, resources are saved. Knowing that the email classification system easily succeeds in classifying 83% of the email messages correctly, the call center saves approximately 25,175 Euros over a 5-year period supposing a labeling efficiency gain of 75%. Indeed, the employee's time-consuming labeling task gives way to a less expensive rearranging task of the correctly classified email set.
Furthermore, there is a possibility to treat incoming complaints in a separate way than other email types during the email-handling process. Transferring complaints to a separate complaint-handling department with well-trained complaint handlers should result in a more successful and faster complaint treatment which results in an overall increase of customer satisfaction.

Moreover, this study explores the differences in linguistic style between complaints and non-complaints. The probability of having a complaint email increases when more words and time indications are used. In contrast to for instance an information request, the likelihood of being a complaint is positively related with the present and past tense, while it decreases when more future tenses and question marks are used. Furthermore, the possibility of classifying an incoming email as a complaint increases when the tone of email becomes more antagonistic – i.e. it contains more negations, more numbers and more clenched words. Offending the company by using a lot of second person pronouns– e.g. y*ou* *are responsible for the misdelivery of the newspaper* -, increases the chance of having a complaint.

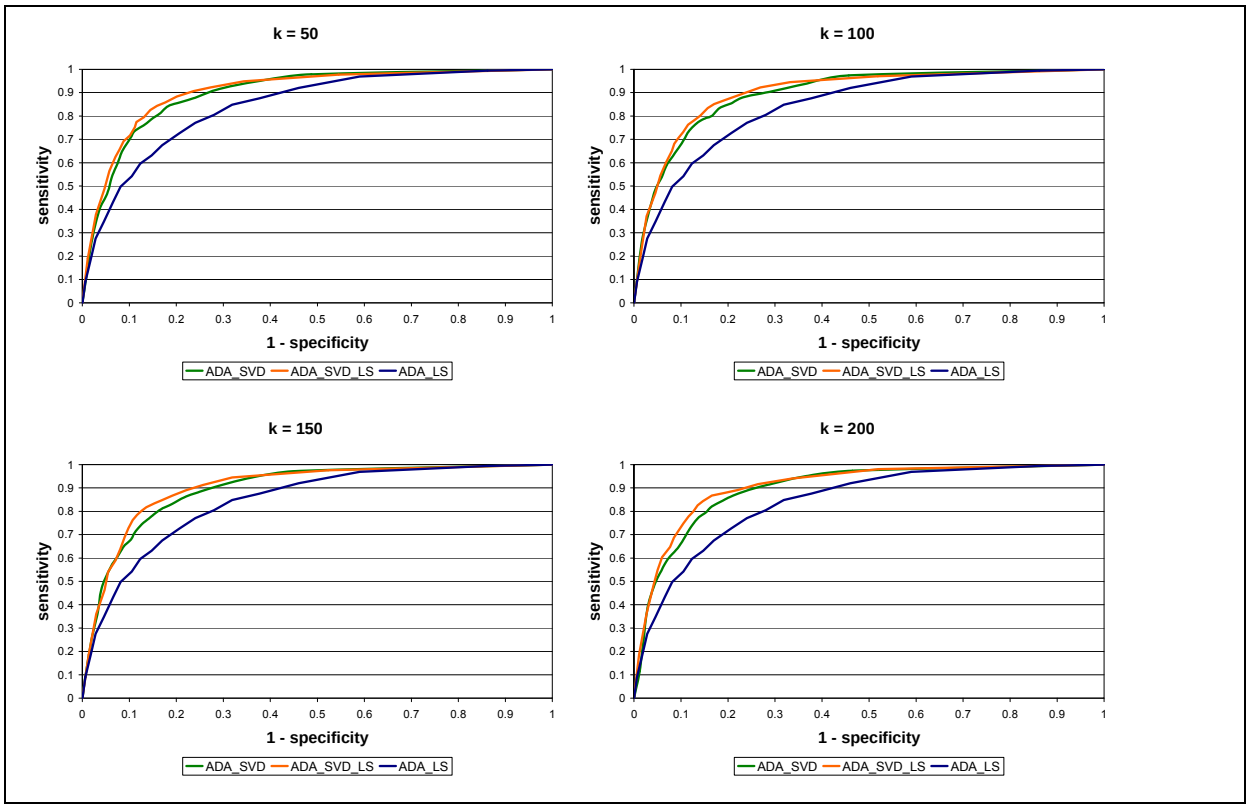
Furthermore, this study proves that adding linguistic style features as an additional set of predictors in a traditional text-classification model significantly increases the predictive performance. In addition, the robustness of the proposed methodology is confirmed by validating the text-classification models on an out-of-period dataset.

While we believe that this study contributes to today's literature, some shortcomings and directions for future research are given. First of all, it is not clear whether adding linguistic style predictors in other text-classification tasks will result in the same highly beneficial increase in predictive performance. Additional experiments need to be done to answer this question. Moreover, additional efforts can be done to refine the proposed methodology. For instance, automatically detecting different types of complaints (e.g. delivery problems, financial problems …) would give us valuable and in-depth information on the occurring problems and service failures that customer encounter.

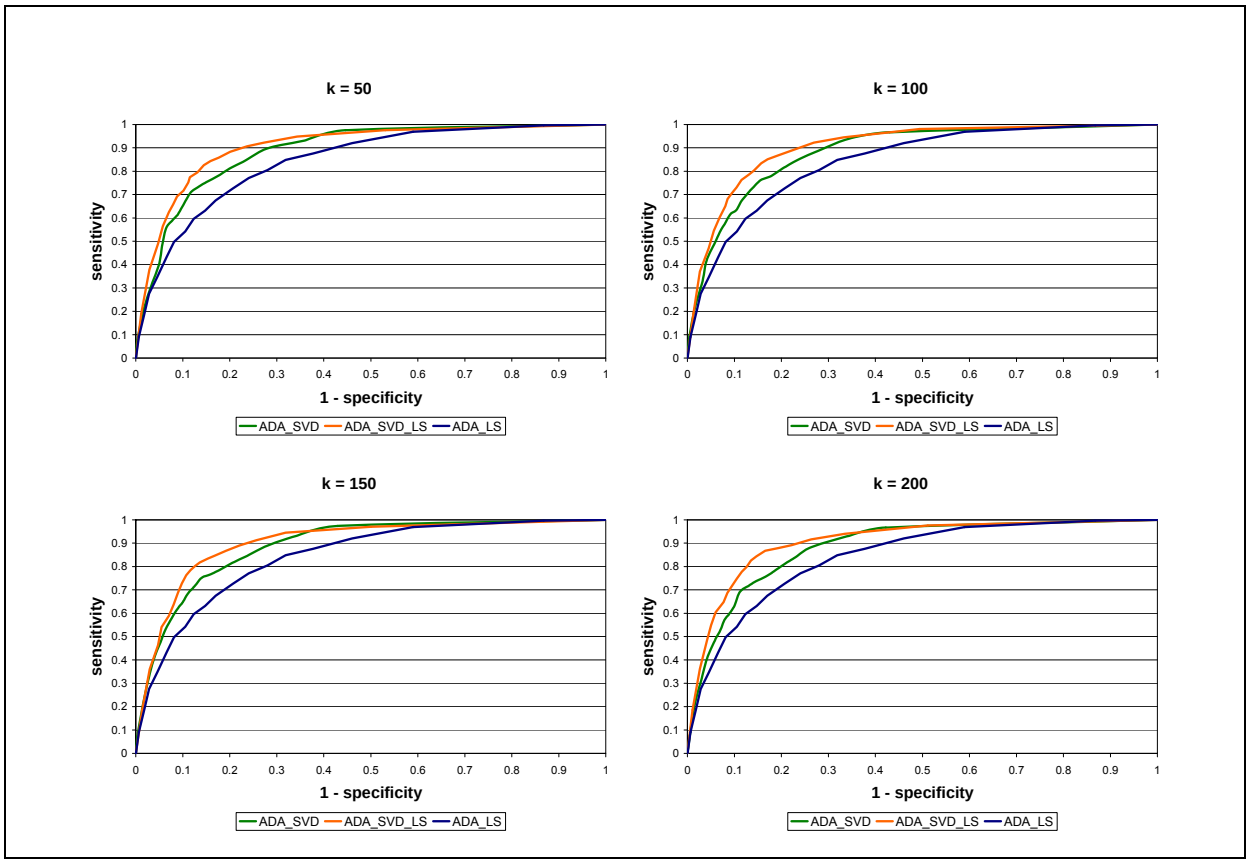
ACKNOWLEDGMENTS

We would like to thank (1) the anonymous Belgian company for providing us with data for testing our research questions, (2) Ghent University for funding the PhD project of Kristof Coussement (BOF 01D26705), (3) Bart Larivière, Jonathan Burez and Ilse Bellinck for their insights during this project. This project was realized using SAS v9.1.3, SAS Text Miner v2.3 and Matlab v7.0.4.

Appendix A: ROCs on test set for ADA_SVD, ADA_SVD_LS and ADA_LS for k={50,100,150,200}



Appendix B ROCs on validation set for ADA_SVD, ADA_SVD_LS and ADA_LS for k={50,100,150,200}



REFERENCES

[1] C. Aasheim and G.J. Koehler, Scanning World Wide Web Documents with the Vector Space Model, Decision Support Systems 42 (2) (2006).

[2] B. Baesens, S. Viaene, D. Van den Poel, J. Vanthienen and G. Dedene, Bayesian Neural Network Learning for Repeat Purchase Modeling in Direct Marketing, European Journal of Operational Research 138 (1) (2002).

[3] C. Bell and K.P. Jones, Toward Everyday Language Information Retrieval Systems via Minicomputers, Journal of the American Society for Information Sciences 30 (1979).

[4] R. Bougie, R. Pieters and M. Zeelenberg, Angry Customers Don't Come Back, They Get Back: the Experience and Behavioural Implications of Anger and Dissatisfaction in Services, Journal of the Academy of Marketing Science 31 (4) (2003).

[5] R.E. Bucklin and S. Gupta, Brand Choice, Purchase Incidence and Segmentation: an Integrated Modeling Approach, Journal of Marketing Research 29 (1992).

[6] R.C. Chen and C.H. Hsieh, Web Page Classification Based on a Support Vector Machine Using a Weighted Vote Schema, Expert Systems with Applications 31 (2) (2006).

[7] S. Deerweester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science 41 (6) (1990).

[8] E.R. DeLong, D.M. DeLong and D.L. Clarke-Pearson, Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach, Biometrics 44 (3) (1988).

[9] Y. Freund Y and R.E. Schapire, A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting, Journal of Computer and System Sciences 55 (1) (1997).

[10] R.D. Gopal, A.K. Tripathi and Z.D. Walter, Economics of First-Contact Email Advertising, Decision Support Systems 42 (3) (2006).

[11] J.A. Hanley and B.J. McNeil, The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve, Radiology 143 (1) (1982).

[12] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction (Springer Series in Statistics, Springer-Verlag New York, 2003)

[13] J.L. Heskett, T.O. Jones, G.W. Loveman, W.E. Sasser and L.A. Schlesinger, Putting the Service-Profit Chain to Work, Harvard Business Review 72 (March–April) (1994).

[14] M.Y. Kiang, T.S. Raghu and K.H.M. Shang, Marketing on the Internet - Who Can Benefit From an Online Marketing Approach?, Decision Support Systems 27 (4) (2000).

[15] W. Kraaij and R. Pohlmann, Viewing Stemming as Recall Enhancement, Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Zurich Switzerland).

[16] C.Y. Liang, L. Guo, Z.H. Xia, F.G. Nie, X.X. Li, L.A. Su and Z.Y. Yang, Dictionary-Based Text Categorization of Chemical Web Pages, Information Processing & Management 42 (4) (2006).

[17] D.G. Morrison, On the Interpretation of Discriminant Analysis, Journal of Marketing Research 6 (1969).

[18] P. Mangiameli, D. West and R. Rampal, Model Selection for Medical Diagnosis Decision Support Systems, Decision Support Systems 36 (3) (2004).

[19] J.W. Pennebaker, M.E. Francis and R.J. Booth, Linguistic Inquiry and Word Count (LIWC) (Mahwah NJ: Erlbaum Publishers, 2001).

[20] M. Pontes and C. Kelly, The Identification of Inbound Call Center Agents' Competencies that are Related to Callers' Repurchase Intentions, Journal of Interactive Marketing 14 (2000).

[21] G. Salton, A Theory of Indexing (Bristol, UK: J.W. Arrowsmith, 1975).

[22] G. Salton, Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer (Reading MA: Addison-Wesley, 1989).

[23] G. Salton and C. Buckley, Term-weighting Approaches in Automatic Text Retrieval, Information Processing & Management 24 (5) (1988).

[24] G. Salton, The SMART Retrieval System: Experiments in Automatic Document Processing (Prentice Hall Englewood Cliffs, NJ, 1971).

[25] R.E. Schapire and Y. Singer, BoosTexter: A Boosting-based System for Text Categorization, Machine Learning 39 (2-3) (2000).

[26] A.K. Smith and R.N. Bolton, The Effect of Customers' Emotional Responses to Service Failures on Their Recovery Effort Evaluations and Satisfaction Judgments, Journal of the Academy of Marketing Science 30 (2002).

[27] K. Sparck Jones, Index Term Weighting, Information Storage and Retrieval 9 (11) (1973).

[28] S.S. Tax and S.W. Brown, Recovering and Learning from Service Failures, Sloan Management Review 40 (1) (1998).

[29] S.S. Weng and C.K. Liu, Using Text Classification and Multiple Concepts to Answer Emails, Expert Systems with Applications 26 (4) (2004).

[30] H. Zijlstra, T. van Meerveld, H. van Middendorp, J.W. Pennebaker and R. Geenen, Dutch Version of the Linguistic Inquiry and Word Count (LIWC); a Computerized Text Analysis Program, Behaviour and Health (Dutch journal) 32 (2004).

CHAPTER IV

IMPROVING CUSTOMER ATTRITION PREDICTION BY INTEGRATING EMOTIONS FROM CLIENT/COMPANY INTERACTION EMAILS AND EVALUATING MULTIPLE CLASSIFIERS

This chapter is based on K. Coussement and D. Van den Poel, Improving Customer Attrition Prediction by Integrating Emotions from Client/company Interaction Emails and Evaluating Multiple Classifiers, Expert Systems with Applications (Forthcoming, 2009).

CHAPTER IV

IMPROVING CUSTOMER ATTRITION PREDICTION BY INTEGRATING EMOTIONS FROM CLIENT/COMPANY INTERACTION EMAILS AND EVALUATING MULTIPLE CLASSIFIERS

ABSTRACT

Predicting customer churn with the purpose of retaining customers is a hot topic in academy as well as in today's business environment. Targeting the right customers for a specific retention campaign carries a high priority. This study focuses on two aspects in which churn prediction models could be improved by (i) relying on customer information type diversity and (ii) choosing the best performing classification technique. (i) With the upcoming interest in new media (e.g. blogs, emails, ...), client/company interactions are facilitated. Consequently, new types of information are available which generate new opportunities to increase the prediction power of a churn model. This study contributes to the literature by finding evidence that adding emotions expressed in client/company emails increases the predictive performance of an extended RFM churn model. As a substantive contribution, an in-depth study of the impact of the emotionality indicators on churn behaviour is done. (ii) This study compares three classification techniques – i.e. Logistic Regression, Support Vector Machines and Random Forests – to distinguish churners from non-churners. This paper shows that Random Forests is a viable opportunity to improve predictive performance compared to Support Vector Machines and Logistic Regression which both exhibit an equal performance.

1. INTRODUCTION

As markets become increasingly saturated, academic researchers and companies have acknowledged that focussing on identifying customer most likely to churn is of crucial importance [27]. [43] remark that organizations are realizing that not all customers generate the same economic value to the company. Establishing valuable relationships with existing customers produces higher revenues and margins than attracting new customers [41]. Consequently, investments for retention strategies have a higher net return than for acquisitions. So, it is supported that companies first spend their marketing resources to keep existing customers rather than to attract new ones [44]. In order to preserve the

existing customer base, marketing consultants try to proactively target those customers most likely to churn. Indeed, companies are moving away from traditional mass marketing strategies in favour of a customer-focussed strategy [7]. As such, Customer Intelligence researchers are increasingly investigating the underlying break-through of improving customer attrition modelling in several research settings (e.g. [30,8]).

In recent years, data mining techniques explore and analyse huge amounts of available data in order to assist with the selection of customers most prone to switch [24]. Many academics and practitioners have built different model types that attempt to predict customers' future behaviour. The research on improving the methodology of churn prediction models is still growing because of (a) it is worthwhile to target customers proactively based on a churn prediction model by sending them churn prevention actions. For instance, a field experiment by Burez and Van den Poel (2007) has shown that companies can double their profits by sending prevention actions to their most likely churners based on a prediction model, (b) keeping existing customers is less expensive then acquiring new ones which are often characterized by a high attrition rate [43] and (c) the churn prediction system has to be as accurate as possible, because [47] have shown that even a small change in retention rate can result in significant changes in contribution.

This study focuses on two aspects on how to optimize a classification model, i.e. (i) incorporated information and (ii) classification technique. (i) Different types of information are available for churn prediction. [18] conclude that customer's past behaviour is an important predictor for one's future behaviour. In the direct marketing literature, it is common practice to summarize customers' past behaviour in terms of their Recency (i.e. the elapsed time since last purchase or renewal), Frequency (i.e. the number of prior purchases or renewals) and Monetary value (i.e. the total amount of purchases) or their RFM characteristics. However, attrition models are often more complex because several other variables, such as socio-demographics and other transactional data, are included on top of the RFM characteristics. We will refer to such a model as an extended RFM model (or eRFM model).

Nowadays, new opportunities arise to increase the prediction power of an eRFM model by incorporating information from new media (e.g. websites, blogs and emails). New valuable information is available to the data analyst because customers interact more frequently with the company through these media types. Indeed, [49] conclude that clients become more familiar with sending emails as a substitute for traditional communication. Moreover, the customer relationship literature states that interactions between client and company are potentially important [19]. As a consequence, academics and marketing consultants try to improve customer relationships by incorporating client/company information in their analysis. For instance, online feedback mechanisms

experience growing popularity and have important implications for a wide range of management activities including customer acquisition and retention [16]. Moreover, call centres which handle telephone calls as well as emails are often seen as the link between customers and the organization. They are potentially important because they offer clients the possibility (a) to report service failures or (b) to ask product related information. (a) Indeed, complaint handling is an important tool to win competitive advantage [4] and it provides a good way to enhance the retention of customers who experience service problems (e.g. [21]). (b) Besides complaints, information requests are another way in which clients interact with the company. For instance, a lot of customers ask information about promotional deals and subscription related aspects. However, it is remarkable that the impact of information requests on customer churn is underinvestigated. All client/company interaction information offers the marketing manager a unique opportunity to explore the client/company relationship and to improve the performance of their churn model. More specifically, [33] focus on the fact that client/company interactions express several emotions via the words they contain. Commonly, one argues that these interactions are homogeneous in terms of expressed emotions, but this is certainly not the case. A common typology to classify emotions is to consider positive as well as negative emotions [12]. This study investigates whether these distinct emotions from call centre emails increase the prediction performance of a churn system. Furthermore, it deepens out the impact of these emotions expressed in client/company emails on one's attrition behaviour.

(ii) Next to the incorporation of new information types, several classification techniques are at the disposal of a data analyst. This study benchmarks the predictive performance of two state-of-the-art classifiers, Support Vector Machines and Random Forests, with the base classifier in marketing, Logistic Regression. It was [36] who conclude that Logistic Regression is a well-known and robust classification technique in marketing, while [46] confirms that it is most often used by predictive model builders in industry. As a consequence, Logistic Regression is used to benchmark Support Vector Machines and Random Forests, two state-of-the-art classifiers within this study. Support Vector Machines have already proven his excellence performance in a wide range of industries like bioinformatics, beat recognition, image classification, ... while over the years it is gaining popularity in churn prediction too (e.g. [28,50,15]). Moreover, Random Forests have also shown their good predictive capabilities in a lot of industries including customer churn prediction (e.g. [31]). The purpose of this research study is to find evidence which state-of-the-art classification technique performs best in optimizing the performance of a churn system.

In conclusion, consolidating customer relationships by avoiding attrition is an important issue for marketing managers and CRM consultants. A first step in addressing this issue is finding who to target in retention actions. The choice of the most appropriate classification technique is an important issue in improving the performance of a churn model. This study compares the predictive performance of

Logistic Regression, Support Vector Machines and Random Forests in distinguishing churners from non-churners. Moreover, customer's past behaviour is an important predictor for one's future behaviour by which RFM models are typically built. Such models are often extended with other transactional and socio-demographic variables. Due to the rapid development of internet and information technology, new client/company information is available. This study investigates whether the emotions expressed in customer emails have an impact on subsequent churn behaviour. Therefore, this paper contributes to the existing literature by investigating (i) whether the predictive performance of an eRFM model increases when emotionality indicators of client/company interaction variables are included?, (ii) which classification technique – i.e. Logistic Regression, Support Vector Machines or Random Forests - performs best in distinguishing churners from loyal customers? and (iii) how do emotions expressed in written client/company interactions through call centre emails (i.e. service failures and information requests) relate to one's churn behaviour?

This paper is organised as follows. Section 2 gives an overview of the classification methods (i.e. Logistic Regression, Support Vector Machines and Random Forests) used throughout this study, while Section 3 gives an overview of the evaluation criteria for the different classification models. Section 4 describes how the emotionality indicators are extracted from the call center emails. In a next Section, the research setting is explained. The empirical results are given in Section 6, while in a last Section conclusions are presented.

2. CLASSIFICATION METHODS

This paragraph introduces the three classification techniques used throughout this study, i.e. Logistic Regression, Support Vector Machines and Random Forests.

2.1. Logistic Regression

Logistic Regression is a well-known technique that is often used in traditional marketing applications [36]. Moreover, it is a simple technique [6], while it provides quick and robust results. Furthermore, a closed-form solution for the 'a posteriori' probabilities is available. Logistic Regression tries to maximize the log-likelihood function in order to become an appropriate fit to the data [1]. Including all predictors into one regression model often results in overfitting and poor predictions, in settings where many variables have little to add to the prediction model. [29] states that variable selection improves the comprehensibility of the resulting model and makes the resulting models generalize better. As done by many researcher and consultants (e.g. [36]), this study employs a stepwise Logistic Regression for churn prediction.

2.2 Support Vector Machines

Support Vector Machines (SVMs) were introduced by Vapnik and his colleagues for solving binary classification problems ([14,48]). The purpose of SVMs in a binary classification context comes down to finding an optimal hyperplane that maximizes the margin between positive and negative examples. We refer to the tutorial of [9] for more details about SVMs, while more information on the optimization process is provided by [11].

In order to implement SVMs, a decision on the kernel function is needed. This study uses the RBF kernel (instead of the linear kernel, the sigmoid kernel or the polynomial kernel) as the default kernel function. In contrast to the *linear kernel* function, RBF kernel makes it possible to map non-linear boundaries of the input-space into a higher dimensional feature space [23]. [32] found in their research that the *sigmoid kernel* behaves like the RBF kernel for certain parameters. When looking at the number of hyperparameters, the *polynomial kernel* has more hyperparameters than the RBF kernel, which makes the optimization process more complex. Moreover, the RBF kernel has less numerical difficulties because the kernel values lie between zero and one, while the polynomial kernel values may go to infinity or zero while the degree is large. Considering these arguments, the RBF kernel function is used as the default kernel function throughout this study.

In order to obtain an optimal performance using a RBF kernel function, one needs to optimize two parameters, namely C and γ with C the penalty parameter for the error term and γ the kernel parameter. Both parameters play a crucial role in the performance of SVMs (e.g. [23]). This study applies a 'grid-search' on C and γ with a two-fold cross-validation as described by [23] for optimizing these parameters. The best parameter pair (C, γ) based on the highest cross-validated AUC is used for further analysis.

2.3. Random Forests

Random Forests is a classification technique that was introduced by [3]. It is an ensemble technique that grows many classification trees in order to overcome the instability of a traditional decision tree [22]. To classify a new object from an input vector, the input vector is run through each of the trees in the forest. Each tree gives a classification and votes for the most popular class. The forest chooses to classify the case according to the label with the most votes over all the trees in the forest.

We follow Breiman's [3] suggestions where the number of variables for growing the tree is set equal to the square root of the total number of variables and a large number of trees - i.e. 1,000 - is chosen.

3. EVALUATION CRITERIA

In order to evaluate the performance of the classification techniques, two criteria are used throughout this study, namely the Percentage Correctly Classified (PCC) and the Area Under the receiving operating Curve (AUC). Both measures are often used as performance criteria in different retention studies (e.g. [5]). The PCC compares the 'a posteriori' probability of being a churner with the true status of that customer. If TP, FP, TN and FN are the True Positives, False Positives, True Negatives and False Negatives in the confusion matrix, then the PCC is defined as (TP+TN)/(TP+FP+TN+FN). The disadvantage of PCC is that it is not very robust concerning the chosen cut-off value on the 'a posteriori' churn probability ([2]). In contrast to PCC, AUC performance takes into account all possible cut-off levels on the 'a posteriori' probabilities. For these cut-off points, it considers the sensitivity (i.e. the number of True Positives versus the total number of churners) and the specificity (i.e. the number of True Negatives versus the total number of non-churners) of the confusion matrix in a two-dimensional graph, resulting in the Receiving Operating Curve or ROC curve. The area under the ROC curve is used to evaluate the performance of a classification model ([20]). In order to compare if two classification models are significantly different in terms of AUC, the non-parametric test of [17] is used.

4. EXTRACTING EMOTIONS FROM CLIENT/COMPANY INTERACTION EMAILS

The ways in which individuals write provide windows into their emotive and cognitive worlds. For instance, [39] investigated cognitive, emotional and language processes in disclosure, while [37] predict deception from language characteristics. Text files are analyzed using the computerized text analysis program Linguistic Inquiry and Word Count, LIWC ([40,51]). As such, it is possible to measure the amount of emotions in written communication.

The scientific program consists of a predefined set of words categorized by psychometric experts into positive and negative emotions. The category of positive emotions is summarized by 690 target words (e.g. happy, good,…), while 1347 target words are used to categorize negative emotions (e.g. hate, sad, ...). The externally-validated program searches the individual text files on a word-by-word basis. Each word in the text is compared against the predefined set of words. After counting the words in each category, the program computes the percentage of the total words for that text. Besides the fact that computerized word count approaches are typically blind to context in which the words are used, they have shown promising and reliable results in personality, social and clinical psychology (e.g. [35,40]). In short, the program makes it possible to compile positive as well as negative emotionality

indicators from call center emails. These figures can be further employed to investigate their impact of emotions on subsequent churn behaviour.

5. RESEARCH SETTING

For this study, data is collected from the largest Belgian newspaper company. The subscribers have to pay a fixed amount of money for their newspaper. Subscription data is used from January 2002 through September 2005. Figure 1 visualizes the window of analysis.

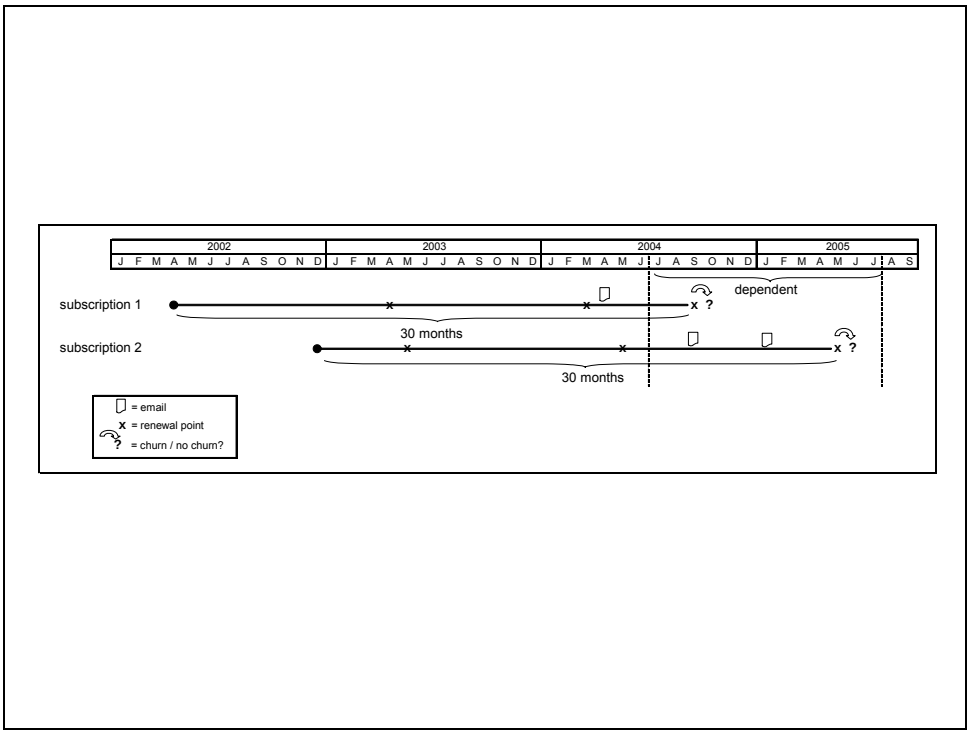


Figure 1: Window of analysis.

Using this time frame, it is possible to define the dependent variable and the explanatory variables. For defining whether a subscriber is a churner or not, all renewal points between July 2004 and July 2005 are considered within this study. Moreover, only subscriptions having at least one email sent during the last subscription term are included in subsequent analysis. There are 11,836 subscriptions included of which 9,600 subscribers (81.11%) renew their subscription and are considered to be behavioural loyal, while 2,236 (18.89%) do not renew their product and are considered as churners. Someone is considered a churner when he/she does not renew his/her subscription within a four week period after the renewal date. During this four week period, the newspaper company still delivers the newspapers to the customers in order to give them the opportunity to renew their subscription. Besides the dependent variable, several explanatory variables need to be constructed in order to predict one's churn behaviour. The explanatory variables, such as RFM and other transactional information, contain

information covering a 30-month period returning from every individual renewal point. This information is stored in a large transactional database. Moreover, subscribers have the possibility to report service failures or to ask questions via email. All emails are manually labelled by the staff of the call center by which a distinction between information requests and complaint emails is available. All emails from the last term of a subscription are considered for further analysis. There are 18,331 emails of which 6,560 complaints (35.79%) and 11,771 information requests (64.21%). Since subscribers can send more than one email during the last term of their subscription, the emotionality indicators extracted using the methodology as explained in Section 4 are averaged per subscription type. In other words, the emotionality variables represent the general positive/negative tone in which subscribers interact with the company by means of written complaints or information requests to the call center.

In order to correctly assess the predictive capabilities of different classification models, the dataset is divided into a training and test set. The former one is composed by randomly assigning 70 percent of the subscriptions, while the other 30 percent are assigned to the test set. The training set is used to estimate the different classification models, while the test set is used for assessing the performance of the different models to an unseen data sample. The data set characteristics are given in Table 1.

	Number of subscriptions	**Relative percentage**
Training set		
Subscriptions not renewed	*1,792*	*18.93%*
Subscriptions renewed	*7,676*	*81.07%*
Total	*9,468*	*100%*
Test set		
Subscriptions not renewed	*444*	*18.75%*
Subscriptions renewed	*1,924*	*81.25%*
Total	*2,368*	*100%*

Table 1: Overview of the data characteristics.

6. EMPIRICAL RESULTS

6.1. Churn Predictors

In the CRM literature, it is a common use to summarize customers' future behaviour based on their past behaviour. The available data are often stored in large databases and consist of RFM figures, other transactional data and socio-demographic information (see Appendix 1 for an in-depth overview of the variety of churn predictors). However, new media (e.g. email) can assist with the improvement of client/company relations. As a consequence, additional explanatory variables are extracted from this new information type. Using the methodology as described in Section 4, several features indicating the emotionality in client/company interactions are extracted from the call center emails (see Table 2).

Variable name	Description
Posemo_complaint	Positive emotions in complaint emails
Posemo_contact	Positive emotions in information requests
Negemo_complaint	Negative emotions in complaint emails
Negemo_contact	Negative emotions in information requests
Percentage_complaint	Percentage of complaints from total client/company interactions

Table 2: Emotionality indicators extracted from client/company interactions.

Consequently, several measurement models are built to correctly discriminate churners from loyal customers. For validating the hypothesis of additional predictive performance of the emotionality indicators on top of the churn variables (see Section 6.2.1.) and for comparing the predictive performance of the different classifiers (see Section 6.2.2.), two different churn models are built. The first one is an extended RFM model (hereafter abbreviated as eRFM) that uses the traditional RFM figures, while also adding other customer information like other transactional data and customer socio-demographics. A second prediction model adds the emotionality related variables from Table 2 to eRFM (hereafter abbreviated as eRFM-EMO). As such these models are used to assess the impact of emotionality from client/company interactions to customer's churn behaviour, while also the performance of the different classification models is compared. In order to correctly verify the impact of emotions in client/company emails on churn behaviour, univariate Logistic Regression models are built using only the emotionality indicators (see Section 6.3).

6.2. Predicting Churn Using Emotionality Indicators

6.2.1 Impact of Emotionality Indicators on Predictive Performance

This paragraph compares the predictive performance between an eRFM model and that of an eRFM-EMO model. Table 3 gives an overview of the predictive performance of all models in terms of AUC and PCC, while Table 4 contains the significance tests of [17] for comparing similar classification techniques between an eRFM and eRFM-EMO context. In summary, this paragraph investigates whether adding emotionality indicators to an attrition model, results in an additional increase in predictive performance in distinguishing churners from non-churners. In other words, does an eRFM-EMO model perform better than an eRFM model for a given classifier?

Model		eRFM		eRFM-EMO	
		AUC	**PCC**	**AUC**	**PCC**
Logit	*Training set*	73.59	77.76	74.89	77.99
	Test set	73.24	77.28	74.16	77.70
Random Forests	*Training set*	74.84	78.26	75.65	78.37
	Test set	75.12	78.29	76.02	78.97
SVMs	*Training set*	72.99	77.90	74.51	78.43
	Test set	72.53	77.53	73.83	77.87

Table 3: The predictive performance of Logit, Random Forests and SVMs.

Model	AUC		Significantly different on 95% confidence level?
	eRFM	eRFM-EMO	
Logit	73.24	74.16	YES
Random Forests	75.12	76.02	YES
SVMs	72.53	73.83	YES

Table 4: Pairwise comparison of AUC performance between eRFM models and eRFM-EMO models on test set.

Table 3 and Table 4 let us conclude that incorporating emotions from client/company interactions is a viable strategy for improving predictive performance of an eRFM churn model. The models including the additional client/company interaction variables perform always better in distinguishing churners from non-churners. Indeed, an eRFM-EMO model always has a higher predictive performance than the corresponding eRFM model in terms of PCC, while also the differences in terms of AUC are significant between similar classifiers [17].

In conclusion, incorporating emotions from client/company emails into a traditional customer attrition model is beneficial from a prediction point of view.

6.2.2. Comparing Predictive Performance of Logit, SVMs and Random Forests

This paragraph compares the predictive performance of Logit, SVMs and Random Forests within this churn context. Table 5 and Table 6 give an overview of the results from the significance test of [17].

Model 1	Model 2	AUC		Significantly different on 95% confidence level?
		Model 1	Model 2	
Logit	Random Forests	73.24	75.12	YES
SVMs	Random Forests	72.53	75.12	YES
Logit	SVMs	73.24	72.53	NO

Table 5: Pairwise comparison of AUC performance on the eRFM test set.

Model 1	Model 2	AUC		Significantly different on 95% confidence level?
		Model 1	Model 2	
Logit	Random Forests	74.16	76.02	YES
SVMs	Random Forests	73.83	76.02	YES
Logit	SVMs	74.16	73.83	NO

Table 6: Pairwise comparison of AUC performance on the eRFM-EMO test set.

From Table 5 and Table 6, it is clear that Random Forests performs best in distinguishing churners from non-churners. Its performance is always significantly higher than the performance of Logit and SVMs in terms of PCC, as well as in terms of AUC [17].

Besides the excellent performance of Random Forests, one notices that Logit and SVMs perform equally well in this churn context. As one observes from Table 5 and Table 6, the AUC performance of Logit and SVMs are not significantly different within the different research contexts.

In summary, implementing Random Forests within this churn context is a viable opportunity to improve predictive performance in comparison to the performance of Logit and SVMs which both have an equal performance [17].

6.3. Impact of Emotionality Indicators on Churn Behaviour

This Section explores the impact of emotions in client/company emails on customer attrition through univariate logistic regression. In contrast to Random Forests, Logistic Regression is able to show the directional impact of the predictors. Table 7 shows the individual standardized parameter estimates, the Wald significance tests and the odds ratios as estimated during a univariate Logistic Regression using the emotionality variables. In other words, this paragraph measures the impact of every single variable from Table 2 on churn behaviour.

Variable name	Standardized estimate	Wald significance tests	Odds ratio
Posemo_complaint	-0.1801**	98.1339	0.674
Posemo_contact	-0.0006 ns	0.00240	0.999
Negemo_complaint	-0.1940**	82.2778	0.561
Negemo_contact	-0.1231**	53.3812	0.752
Percentage_complaint	-0.3561**	255.994	0.418
**: $p < 0.001$;*ns*: *not significant*			

Table 7: Univariate Standardized Parameter Estimates, Wald Significance Tests & Odds Ratio.

A primary finding of this research is that there is a significant relationship between positive expressed emotions in complaint emails and customer attrition (posemo_complaint:β=-0.1801,p<0.001). This result confirms the findings of [33] who also found that positive emotions expressed during complaining reduce the chance of churning. The positive emotions expressed during the reporting of a service failure are assumed to counteract the eagerness to punish the company for the failure and to indicate the good prior relationship with the company. Moreover, this study shows evidence that a significant relationship exists between negative emotions expressed in complaint emails and customer churn (negemo_complaint:β=-0.1940,p<0.001). Contrary to the results of [33], we find that the more negative emotional words are used in complaint emails, the lower the risk that the customer will leave the company. Considering that the company has a good failure recovery system, this result is in line with the satisfaction framework of [38]. He states that when the gap between the outcome – in this case the service failure recovery – and the expectations – in this case the service recovery expectations – increases, the satisfaction of that particular customer increases. In other words, the more negative

emotional words one uses during complaining, the more disillusioned the customer is. However, when the company decently recovers the failure, the customer will be very satisfied.

This study found no significant relationship between positive expressed emotions in information requests and someone's churn behaviour (posemo_contact:β=-0.0006,ns). Moreover, negative expressed emotions in information requests seem to have a significant influence on customers' churn behaviour (negemo_contact:β=-0.1231,$p<0.001$). One can say that the more negative emotional words are used in emails other than complaints, the lower the chance that the customer will churn. In intensive markets like the newspaper industry, customer satisfaction level becomes an important issue [25]. So increasing customer satisfaction is the starting point of the email handling process. For instance, customers are often comparing alternatives for their current product by the end of the subscription period. As such they often come in touch with new promotional deals, they want to capture. These information requests via email are often more negatively connoted than others because they express the disillusion of not having a promotional offer on their current subscription. However, the company tries to handle all questions as efficiently as possible by offering them a comparable subscription deal. This act increases customer satisfaction, because [38] states that satisfaction is increased when the final outcome – i.e. here the proposal for a new subscription - exceed the rather bad expectations of fulfilling the request.

Moreover, this study found a strong relationship between the percentage of complaints of all client/company interactions and one's churn behaviour (percentage_complaint:β=-0.3561,$p<0.001$). The more complaints a customer has in his/her portfolio of emails sent to the company, the more certain he/she stays with the company. As such, complaining does not necessarily mean that the customer will leave the company. For instance, [34] and [26] state that complaint behaviour results in a favourable behaviour towards the company when service failure recovery is satisfactory. Moreover, [10] state that complainers are less vulnerable to switch because they have a higher commitment and trust towards the company (e.g. [45]). Also [13] state that customer who complain and receive a proper response to their service failures are more likely to stay.

7. CONCLUSION & FURTHER RESEARCH

As a conclusion, one can say that the predictive performance of a churn model can be optimised by (i) exploring and adding new types of customer information into a conventional churn model and (ii) choosing the right classification technique. (i) In the last decade, there is a rapid development of the internet and information technology. As such, new information types are available to the data analyst. Nowadays, emails are seen as a valid alternative (next to letters and telephone calls) for customers to interact with the company. Most companies store these huge amounts of textual information in large

databases, but hardly use them in their day-to-day analysis. As a consequence, unique opportunities arise to extract information from these emails to enrich churn models. This study shows the beneficial effect of including emotionality indicators extracted from call center emails into a customer attrition model. Indeed, the predictive performance significantly increases when these emotionality indicators are included into a eRFM attrition model. (ii) The predictive performance of two state-of-the-art classifiers SVMs and Random Forests is benchmarked with that of a base classifier, Logistic Regression. It is shown that Random Forests significantly outperforms the other two classification techniques, while the predictive performance of Logistic Regression and Support Vector Machines is not significantly different within this research setting. Compared to Logistic Regression and SVMs, Random Forests' predictive dominance lies in the ability to discover hidden patterns in the complex data structure by (a) combining several outputs of 'weak' classifiers into a strong ensemble of classifiers and (b) by doing a random feature selection to split each node in the individual decision trees ([3]).

Despite the fact that differences in performance may not seem large, the impact on the retention rate can be significant when targeting the right customers accordingly. Table 8 shows that increasing the retention rate with only 1% already has large implications for the long-term increase in profitability within this specific case.

Retention rate (= 100%-churn%)	No of customers					Average Contribution per subscriber per year (in Euro)	Total Contribution after 5 year per 1,000 subscribers (in Euro)	Additional profit over 82% after 5 year per 1,000 subscribers (in Euro)
	Year							
	1	2	3	4	5			
82%	1,000	820	672	551	452	200	657,355	
83%	1,000	830	689	572	475	200	669,799	12,445
100%	1,000	1,000	1,000	1,000	1,000	200	925,979	256,180

Table 8: Real-life retention example.

In the current situation, about 18% of the clients defect. As such, we expect to keep each year about 82% of the subscribers. Additionally, the ideal situation of 100% retention rate is also included into the analysis for comparison reason only. This situation is utopian because there will always be people who will churn, e.g. due to natural defection. Suppose further that due to an increase in predictive performance and targeting the right customers accordingly, the company can increase the retention rate with 1% - i.e. from 82% to 83% retention rate. Table 8 indicates that an additional increase in retention rate with 1% results in a boost of total contribution over 5 year per 1,000 customers from 657,355 Euro to 669,799 Euro having an average contribution of 200 Euro and a discount rate of 4%. If the company succeeds in increasing the retention rate with only 1%, an additional contribution of 12,445 Euro per 1,000 customers is gained.

As a substantive contribution, the impact of emotionality indicators on churn is investigated. It is shown that the impact on churn is positive when more emotional related words – i.e. positive emotions or negative emotions - are used. Customers that use more positive emotions in complaint emails are by nature more satisfied and certainly do not want to punish the company for the service failures which decreases of course the fact that one will attrite. The use of negative emotions in emails seems to have a positive influence on one's churn behaviour. This means that people who use more negative emotional connoted words tend to be more loyal. These results have large managerial implications for all managers dealing with call center management and email handling. Emotional emails require special treatment, because these people tend to more loyal than others. Even in the case of a very negative complaint email, the call center agent must be aware of the fact that this customer is of high value for the company. In fact, when customers express their dissatisfaction with the service by writing complaint emails, this does not necessarily mean that these customers will churn. Contrary, this study found that the higher the portion of complaint emails, the lower the chance that this specific customer will churn.

While we strongly believe that this research paper adds value to the current literature, there is still scope for some suggestions for further research. In this study, service recovery data was unavailable to the data analyst. As such, additional efforts could be made in collecting service recovery measures for every client/company interaction. Consequently, additional analysis using this type of data can be introduced in the current framework. Moreover, the proposed framework of customer churn prediction is applied in a newspaper subscription business. There is a wide variety of data mining applications, e.g. cross- and up-sell applications, customer acquisition… in wide range of industries, e.g. retail, financial services, e-commerce… where the incorporation of this type of 'soft' data can improve the predictive performance. Additional analysis need to be done to validate the findings proposed in this research paper.

ACKNOWLEDGEMENTS

We would like to thank the anonymous Belgian company for providing us with data for testing our research questions and Ghent University for funding the PhD project of Kristof Coussement (BOF 01D26705). Also special thanks to L. Breiman (†) for freely distributing the Random Forests software, as well as C.-C. Chang and C.-J. Lin for sharing their SVM-toolbox, LIBSVM. Moreover, we would like to thank Bart Larivière and Ilse Bellinck for their insights during this project.

APPENDIX 1: OVERVIEW OF THE CHURN PREDICTORS

	Variable type		Description
eRFM	***RFM***		Elapsed time since last renewal (Recency) The number of renewal points (Frequency) Monetary value (Monetary value)
	Other transactional information	*Client/company characteristics*	The purchase motivator of the subscription How the newspaper is delivered The number of complaints Elapsed time since the last complaint The average cost of a complaint (in terms of compensation newspapers) The conversions made in distribution channel, payment method & edition Elapsed time since last conversion in distribution channel, payment method & edition The number of responses on direct marketing actions Elapsed time since last response on a direct marketing action The number of free newspapers
		Renewal-related variables	Whether the previous subscription was renewed before the expiry date How many days before the expiry date, the previous subscription was renewed The average number of days the previous subscriptions are renewed before expiry date The variance in the number of days the previous subscriptions are renewed before expiry date Elapsed time since last step in renewal procedure The number of times the churner did not renew a subscription
		Subscription-describing variables	The length of the current subscription The number of days a week the newspaper is delivered (intensity indication) What product the subscriber has The month of contract expiration
	Socio-demographics		Age Whether the age is known Gender Physical person (is the subscriber a company or a physical person) Whether contact information (telephone, mobile number, email) is available

REFERENCES

[1] P.D. Allison, Logistic Regression Using the SAS System: Theory and Application, SAS Institute Inc.: Cary, NC (1999).

[2] B. Baesens, S. Viaene, D. Van den Poel, J. Vanthienen and G. Dedene, Bayesian neural network learning for repeat purchase modeling in direct marketing, European Journal of Operational Research, 138 (1) (2002), 191-211.

[3] L. Breiman, Random forests, Machine Learning, 45 (1) (2001), 5-32.

[4] S.W. Brown, Service recovery through IT: complaint handling will differentiate firms in the future, Marketing Management, 6 (3) (1997), 25-27.

[5] W. Buckinx and D. Van den Poel, D., Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting, European Journal of Operational Research, 164 (1) (2005), 252-268.

[6] R.E. Bucklin and S. Gupta, Brand Choice, Purchase Incidence and Segmentation: an Integrated Modeling Approach, Journal of Marketing Research, 29 (2) (1992), 201-215.

[7] J. Burez and D. Van den Poel, CRM at Canal+ Belgique: reducing customer attrition through targeted marketing, Expert Systems with Applications, 32 (2) (1997), 277-288.

[8] J. Burez and D. Van den Poel, Handling Class Imbalance in Customer Churn Prediction, Expert Systems With Applications, under review.

[9] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, 2 (2) (1998), 121-167.

[10] R. Bougie, R. Pieters and M. Zeelenberg, Angry Customers Don't Come Back, They Get Back: The Experience and Behavioral Implications of Anger and Dissatisfaction in Services, Journal of the Academy of Marketing Science, 31 (4) (2003), 377-393.

[11] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. Technical Report, Department of Computer Science and Information Engineering; National Taiwan University (2004).

[12] A. Chaudhuri, Product class effects on perceived risk: the role of emotion, International Journal of Research in Marketing, 15 (1998), 157–168.

[13] D.E. Conlon and N.M. Murray, Customer Perceptions of Corporate Responses to Product Complaints: The Role of Explanations, Academy of Management Journal, 39 (4) (1996), 1040-1056.

[14] C. Cortes and V. Vapnik, Support-vector network, Machine Learning, 20 (1995), 273–297.

[15] K. Coussement and D. Van den Poel, Churn Prediction in Subscription Services: an Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques, Expert Systems with Applications, 34 (1) (2008), 313-327.

[16] C. Dellarocas, The digitization of word of mouth: promise and challenges for online feedback mechanisms, Management science, 49 (10) (2003), 1407-1424.

[17] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson, Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach, Biometrics, 44 (3) (1988), 837-845.

[18] P.S. Fader, B.G.S. Hardie, and K.L. Lee, RFM and CLV: Using Iso-Value Curves for Customer Base Analysis, Journal of Marketing Research, 62 (2005), 415-430.

[19] S. Ganesan, Determinants of long-term orientation in buyer-seller relationships, Journal of Marketing, 58 (2) (1994), 1-19.

[20] J.A. Hanley, and B.J. McNeil, The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve, Radiology, 143 (1) (1982), 29-36

[21] C.W.L. Hart, J.L. Heskett and Jr. W.E. Sasser, The profitable art of service recovery, Harvard Business Review, July-August (1990), 146-156.

[22] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning: data mining, inference and prediction, Springer-Verlag (2001).

[23] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, A practical guide to support vector classification. Technical Report, Department of Computer Science and Information Engineering; National Taiwan University (2004).

[24] S.-Y. Hung, D.C. Yen, and H.-Y. Wang, Applying data mining to telecom churn management, Expert Systems with Applications, 31 (3) (2006), 515-524.

[25] T.O. Jones and Jr. W.E. Sasser, Why Satisfied Customer Defect, Harvard Business Review, 73 (1995), 88-99.

[26] S. Keaveney, Customer Switching Behavior in Service Industries: An Exploratory Study, Journal of Marketing, 59 (April) (1995), 71-82.

[27] S. Keaveney and M. Parthasarathy, Customer switching behavior in online services: an exploratory study of the role of selected attitudinal, behavioral and demographic factors, Journal of the Academy of Marketing Science, 29 (4) (2001), 374-390.

[28] S. Kim, K.S. Shin and K. Park, An application of support vector machines for customer churn analysis: credit card case, Lecture Notes in Computer Science, 3611 (2005), 636-647.

[29] Y.S. Kim, Toward a Successful CRM: Variable Selection, Sampling and Ensemble, Decision Support Systems, 41 (2) (2006), 542-553.

[30] B. Larivière and D. Van den Poel, Investigating the role of product features in preventing customer churn,by using survival analysis and choice modeling: The case of financial services, Expert Systems with Applications, 27 (2) (2004), 277-285.

[31] B. Larivière and D. Van den Poel, Predicting customer retention and profitability by using random forests and regression forests techniques, Expert Systems with Applications, 29 (2) (2005), 472-484.

[32] H.-T. Lin and C.-J. Lin, A study on sigmoid kernels for SVM and the training of non-psd kernels by SMO-type methods, Technical report, Department of Computer Science and Information Engineering; National Taiwan University (2003).

[33] J. Mattsson, J. Lemmink and R. McColl, The Effect of Verbalized Emotions on Loyalty in Written Complaints, Total Quality Management & Business Excellence, 15 (7) (2004), 941-958.

[34] J.G. Maxham, Service recovery's influence on consumer satisfaction, positive word-of-mouth, and purchase intentions, Journal of Business Research, 54 (1) (2001), 11-24.

[35] E. Mergenthaler, Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes, Journal of Consulting and Clinical Psychology, 64 (1996), 1306-1315.

[36] S.A. Neslin, S. Gupta, W. Kamakura, J. Lu and C. Mason, Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models, Journal of Marketing Research, 43 (2) (2006), 204-211.

[37] M.L. Newman, J.W. Pennebaker, D.S. Berry and J.M. Richards, Lying words: Predicting Deception from Linguistic Style, Personality and Social Psychology Bulletin, 29 (2003), 665-675.

[38] R.L. Oliver, A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions, Journal of Marketing Research, 17 (4) (1980), 460-469.

[39] J.W. Pennebaker and M.E. Francis, Cognitive, emotional and language processes in disclosure: physical health and adjustment, Cognition and Emotion, 10 (1996), 601-626.

[40] J.W. Pennebaker, M.E. Francis and R.J. Booth, Linguistic Inquiry and Word Count (LIWC), Erlbaum Publishers; Mahwah, NJ (2001).

[41] F.F. Reichheld and W.E. Sasser, Zero Defections: Quality Comes to Services, Harvard Business Review, 68 (5) (1990), 105-111.

[42] W. Reinartz, M. Krafft and W.D. Hoyer, The Customer Relationship Management Process: Its Measurement and Impact on Performance, Journal of Marketing Research, 41 (3) (2004), 293-305.

[43] W. Reinartz and V. Kumar, The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration, Journal of Marketing, 67 (1) (2003), 77-99.

[44] R.T. Rust and A.J. Zahorik, Customer Satisfaction, Customer Retention, and Market Share, Journal of Retailing, 69 (2) (1993), 193-215.

[45] S.S. Tax, S.W. Brown and M. Chandrashekaran, Customer Evaluations of Service Complaint Experiences: Implications for Relationship Marketing, Journal of Marketing, 62 (April) (1998), 60-76.

[46] L.C. Thomas, A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Customers, International Journal of Forecasting, 16 (2) (2000), 149-172.

[47] D. Van den Poel and B. Larivière, Customer attrition analysis for financial services using proportional hazard models, European Journal of Operational Research, 157 (1) (2004), 196–217.

[48] V. Vapnik, Statistical Learning Theory. Wiley; New York (1998).

[49] S.S. Weng and C.K. Liu, Using Text Classification and Multiple Concepts to Answer Emails, Expert Systems with Applications, 26 (4) (2004), 529-543.

[50] Y. Zhao, B. Li and X. Li, Customer churn prediction using improved one-class support vector machine, Lecture Notes in Artificial Intelligence, 3584 (2005), 300-306.

[51] H. Zijlstra, T. van Meerveld, H. van Middendorp, J.W. Pennebaker and R. Geenen, Dutch Version of the Linguistic Inquiry and Word Count (LIWC); a Computerized Text Analysis Program, Behaviour and Health (Dutch journal), 32 (2004).

CHAPTER V

COMPARING NON-SAMPLING BASED COST-SENSITIVE LEARNERS IN A CHURN PREDICTION CONTEXT

This chapter is based on K. Coussement and D. Van den Poel, Empirical Generalizations in Non-Sampling Based Cost-Sensitive Learning for Optimal Decision Making in a Churn Prediction Context, Decision Support Systems (in review).

CHAPTER V

COMPARING NON-SAMPLING BASED COST-SENSITIVE LEARNERS IN A CHURN PREDICTION CONTEXT

ABSTRACT

Marketing decision makers realize that information extracted from the customer database is a valuable tool to leverage their customer knowledge. One of the facets of a well-considered CRM system is customer churn prediction where one tries to predict whether or not a customer will stay with the company. This study argues to consider different misclassification costs when one evaluates the performance of a customer churn prediction system. Moreover, this study meta-analyses several non-sampling based cost-sensitive learners (Direct minimum expected cost criterion, Metacost, Thresholding and Weighting) which take into account different costs asymmetries. A sensitivity analysis based on the cost ratio and the churn incidence shows that Direct minimum expected cost criterion is the most robust alternative, while the arbitrary classification threshold of 0.5 performs worst. Moreover, it is shown that Weighting performs best when cost ratios are high. When cost ratios are low, Thresholding performs well with churn levels from low to medium, while Metacost shows to have a good performance when churn levels range from medium to high.

1. INTRODUCTION

In recent years, the availability of large volumes of data on customers has created new opportunities to leverage customer knowledge and gain competitive advantage. On the one hand, many organizations realize that knowledge extracted from these large databases is supporting marketing decisions, while on the other hand, the intense competition and increased choices available to customers create new pressures on marketing decision makers to manage their

customers in a long-term relationship [23]. This is called customer relationship management (CRM). Nowadays, many organizations are turning to CRM to better serve customers and facilitate closer relationships with them in many different industries (e.g. [25] [18] [27]). One of the aspects of CRM is customer churn prediction. Marketing decision makers confirm that identifying those customers most likely to churn becomes increasingly important, especially when markets become saturated [13]. As such, establishing valuable relationships with existing customers rather than attracting new customers which are characterized by a high attrition rate is of crucial importance [21].

The use of information technology and the ever increasing computer performance sheds new light on customer churn prediction through the concept of data mining. Specifically, data mining techniques explore and analyze the huge amounts of available data in order to assist with the selection of those customers at risk [12]. Developments in database processing [29], data warehousing [14], machine learning [11][22] and knowledge management [1], have made data mining an attractive and practical tool to uncover hidden patterns in customer data. An important component of data mining is called classification [10]. Classification is the procedure in which data instances or predictor variables are placed into predefined instance classes based on a training set of previously labeled instances. This is no different for customer churn prediction where one tries to predict whether or not a customer will leave the company (e.g. [2] [6]).

Most of the currently available algorithms for classification are designed to minimize the error rate or zero-one loss rather than the total cost of misclassification [7]. This has several complications, since in many real-world decision making situations the assumption of equal misclassification costs, the default operating mode for many learners, is most likely violated [28]. Medical diagnosis is a classic. In this case, failing to detect a disease (false negative) may have fatal consequences, whereas diagnosing a disease for a patient that does not actually have it (false positive) may be less serious. A comparable situation arises for customer churn detection. Here, failing to identify a real churner (false negative) has larger consequences than categorizing a non-churner as a churner (false positive). In case of a false negative prediction, the company loses the foregone profits, the opportunities to cross-sell and up-sell… while there is only a 'pampering' cost for false positive classifications.

In this study, we revisit several cost-sensitive learners for making classifiers taking into account cost asymmetries in order to optimize decision making in a customer churn context. Two main

categories of cost-sensitive learners exist: non-sampling and sampling (i.e. under- and oversampling) [24]. This study focuses on non-sampling based cost-sensitive learners, because sampling based cost-sensitive techniques have several drawbacks. (i) They distort the distribution of examples, which may seriously affect the performance of some algorithms. (ii) It reduces the data available for training, if sampling is carried out by undersampling, while (iii) oversampling drastically increases the learning time of classifiers.

Non-sampling based cost-sensitive learners are divided in three categories: Relabeling, Threshold adjusting and Weighting [24]. The first two categories are algorithms that are able to make a broad variety of classifiers cost sensitive, i.e. methods that are designed to make learners minimize total misclassification cost rather than minimizing the error rate. The latter category, Weighting, makes individual algorithms cost sensitive by emphasizing instances by assigning a weight depending on the class labels.

Relabeling is based on reassigning the classes of the data instances by applying the direct minimum expected cost criterion, i.e. an instance is assigned to the class with the lowest misclassification cost. Relabeling can be done in the *post-learning* phase, this is called Direct minimum expected cost classification [8] or in the *pre-learning* phase on the training instances as proposed by [7] via the Metacost algorithm. Threshold adjusting searches for the best probability on the training set as a threshold for optimal future decision making in terms of misclassification cost [24]. Weighting assigns a certain weight to each instance depending on its class, such that the learning algorithm is in favor of the class with the highest weight/cost.

This study presents a meta-analysis of different cost-sensitive algorithms in a customer churn prediction context. Section 2 summarizes the traditionally used performance measures in customer churn classification and highlights their weaknesses by neglecting the cost asymmetries. In a next Section, the methodological issues are discussed by giving an overview of the base learner and cost-sensitive learning techniques, i.e. Relabeling (with Direct minimum expected cost classification and Metacost), Threshold adjusting and Weighting. The data sets, the practical threshold and the empirical results are discussed in Section 4. In a last Section, we discuss the results and their managerial implications, while we identify some suggestions for further research.

2. MISCLASSIFICATION COST AS PERFORMANCE MEASURE

Despite the fact that the error rate, the (area under the) receiver operating curve and the top-decile lift are often used to evaluate the performance of traditional churn prediction models (e.g. [15] [6]), they totally neglect the asymmetry of misclassification costs needed for optimal decision making.

The error rate (or accuracy defined as 1-error rate) is undoubtedly the most common performance measure used in churn prediction applications (e.g. [6]). Practically, all data instances are ranked from most likely to churn to least likely to churn. All instances with a churn probability above a certain threshold are classified as churners, while all others are seen as stayers. In sum, the error rate computes the ratio of incorrectly classified instances to the total number of instances to be classified. Besides its popularity, the error rate as performance measure is often criticized by several researchers (e.g. [20]), as it assumes equal misclassification costs and relatively balanced class distributions. Within a naturally skewed class distribution, wrong predictions for the underrepresented class are very costly. A model minimizing the error rate often results in a useless model that always predicts the most frequent class.

Several authors (e.g. [3]) came up with the concept of the receiver operating curve to evaluate the performance of binary classification systems. In contrast to the error rate, the receiver operating curve takes into account the individual class performance of a classifier. The receiver operating curve is a two dimensional visualization of the sensitivity, i.e. the number of true positives versus the total number of events, and 1-specificity, i.e. the number of true negatives versus the total number of non-events. It illustrates the decision making behavior for various values of the classification threshold. In other words, the receiver operating curve allows us to visualize and evaluate the quality of the rankings on the posterior probabilities. However, the receiver operating curve suffers from two main drawbacks from a decision making perspective. Firstly, it does not show us how to actually set the classification threshold to optimize decision making [17]. Secondly, the accurate ranking of data instances that underlies the receiver operating curve semantics is not equivalent to the accurate posterior class membership probability estimation task [28]. For making optimal decisions, a model that perfectly ranks the data instances based on the posterior probabilities, while failing to estimate those probabilities accurately, is insufficient.

Another performance measure that is frequently used in marketing applications is the top-decile lift (e.g. [15]). It measures the increase in density of the real churners within the 10% cases most likely to churn. Practically, all data instances are sorted from least likely to churn to most likely to churn. Within the top 10% cases most likely to churn, the proportion of real churners is compared with the proportion of real churners in the total database. This increase in density is called the top-decile lift. From an optimal decision-making perspective, several problems arise. Firstly, the threshold for classifying a data instance as a churner is very arbitrarily chosen, as it completely neglects the minimization of the total misclassification cost. Secondly, by focusing on maximizing the relative percentage of churners in the top 10%, the top-decile lift does not take into account the misclassification of data instances.

Furthermore, improving marketing decision making by minimizing the overall misclassification cost is supported in many practical situations. For instance, suppose that customer X enters a bank office. Based on his/her historical information, the company's churn model indicates that client X has a probability of seventy percent churning within the next six months. In order to take the appropriate churn prevention action, the company needs an indication whether customer X will be classified as leaving or staying. Marketing decision makers must take an optimal decision in correspondence to the minimization of the total misclassification cost.

In sum, it is clear that the error rate, the (area under the) receiver operating curve and the top-decile lift do not lead to an optimal evaluation of the classification algorithm when different misclassification costs occur. The extra element needed for optimal decision making is an evaluation in terms of total misclassification costs.

3. METHODOLOGY

3.1. Base learner

In cost-sensitive learning, [7] suggests using a classifier that is well-suited to the domain. As such, Logistic Regression is used as the base classifier throughout this research paper, because [19] state that it is a well-known classification technique in traditional marketing applications like customer churn prediction. Moreover, it is a simple technique [5] providing quick and robust results.

Logistic Regression is able to estimate $p(j=1|x)$, while it makes the assumption that the difference between the natural logarithms of the class-conditional data density functions is linear in the predictors via

$$\ln\left(\frac{p(x \mid j=1)}{p(x \mid j=0)}\right) = b + w^t x \tag{1}$$

with $w \in \Re^n$ as coefficient vector and $b \in \Re$ as the intercept. As such, the class membership probability $p(j=1|x)$ is obtained from (1) by

$$\mathrm{p}(j=1|x) = \frac{\exp(b'+w^t x)}{1+\exp(b'+w^t x)} \tag{2}$$

with $b'=b+\ln\left(\frac{p(j=1)}{p(j=0)}\right)$, $p(j=1)$ and $p(j=0)$ as the class priors. The class membership probabilities are used to obtain the maximum likelihood estimates for w and b'.

3.2. Cost-sensitive learning techniques

3.2.1. Relabeling

3.2.1.1. Direct minimum expected cost classification (DMECC)

Many classification techniques produce a posteriori probabilities that can be used to classify predictor vectors into predefined classes. [28] revisited DMECC as a valuable tool to improve decision making when misclassification costs are different. In a binary classification context, optimal Bayes decision making criterion dictates that a predictor vector $x \in \Re^n$ should be assigned to the class $t \in \{0,1\}$ associated with the minimum expected cost [8]. Optimal Bayes decision criterion for an example x is the class t where

$$\arg\min_{t \in \{0,1\}} \sum_{j=0}^{1} p(j \mid x) c_{tj} \tag{3}$$

where p($j|x$) is the conditional probability of class j given predictor vector x and c_{tj} is the cost of classifying a data instance with predictor vector x and actual class j as class t. If $t=j$ then the prediction is correct, while if $t\neq j$ the prediction is incorrect.

Typically, the cost matrix C has the following structure in a binary classification context (see Table 1).

		Actual class	
		Positive	Negative
Predicted class	Positive	c_{11}	c_{10}
	Negative	c_{01}	c_{00}

Table 1: overview cost matrix for binary classification.

By following the convention of recent papers, the cost matrix rows correspond to alternative predicted classes, while the columns correspond to actual classes, i.e. row/column = t/j = predicted/actual class.

The cost for a true positive is denoted as c_{11}, the cost for a true negative is denoted as c_{00}, the cost for a false positive is denoted as c_{10} and the cost for a false negative is denoted as c_{01}. Note that if $c_{..}$ larger than 0, it represents an actual cost, whereas if $c_{..}$ is smaller than 0, it represents a benefit.

Cost matrices often implicitly meet two reasonableness conditions formulated by [9]. The first reasonableness condition implies that neither row in the matrix dominates the other. [16] has pointed out that for some cost matrices, class labels are never predicted by the optimal policy given by equation (3). A simple, intuitive criterion for when this happens: say that row 1 dominates row 2 in a cost matrix C in Table 1 if for all $j \in \{0,1\}$, $c_{1j} \geq c_{0j}$. In this case the cost of predicting *negative* is no greater than the cost of predicting *positive*, regardless of what the true class j is. So, it is optimal never to predict *positive*. The second reasonableness condition implies that the cost of labeling an instance correctly is always lower than the cost of labeling an instance incorrectly, i.e. $c_{tj} \leq c_{jj}$.

Given the reasonableness conditions and for a binary classification system, the prediction for the positive class is optimal if the expected cost of this prediction is less than the expected cost of predicting the negative class, i.e. if

$$p(j=0|x)\, c_{10} + p(j=1|x)\, c_{11} < p(j=1|x)\, c_{01} + p(j=0|x)\, c_{00} \quad (4)$$

which is equivalent to

$$(1\text{-}p)\, c_{10} + p\, c_{11} < p\, c_{01} + (1\text{-}p)\, c_{00} \tag{5}$$

with p=p(j=1|x). If this inequality is in fact an equality, then predicting either class is optimal. The threshold for making optimal decisions is p* such that

$$(1\text{-}p^*)\, c_{10} + p^*\, c_{11} = p^*\, c_{01} + (1\text{-}p^*)\, c_{00}. \tag{6}$$

Rearranging the equation for p* leads to the following solution

$$p^* = \frac{c_{10} - c_{00}}{c_{10} - c_{11} + c_{01} - c_{11}}. \tag{7}$$

In other words, the classification rule that assigns the positive class is

$$\text{p}(j=1|x) > \frac{c_{10} - c_{00}}{c_{10} - c_{11} + c_{01} - c_{11}} \tag{8}$$

and the negative class otherwise.

The costing method, DMECC, relabels the posterior probabilities according to the direct minimum expected cost criterion. The advantage of using this algorithm is that (i) any error-based learner that produces an estimate of $p(t|x)$ can make use of equation (8) to determine the optimal class and (ii) the models do not need to be retrained when the cost structure changes, because relabeling is only introduced in the post-learning phase.

3.2.1.2. Metacost

Metacost is aimed at making an error-based classifier cost-sensitive by manipulation of the training data instance target labels ([7] [28]). It is based on wrapping a "meta-learning" stage around the error-based classifier, in such a way that the classifier effectively minimizes cost. In other words, Metacost treats the underlying classifier as a black box, requiring no knowledge of its functioning or change to it. This study implements Domingos' Metacost algorithm which

relabels the training instances according to equation (3) and then learns the final classification model using these relabeled instances.

More specifically, [7] estimates the $p(j|x)$ by using a variant of Breiman's [4] bagging (or bootstrap aggregation). In other words, multiple bootstrap replicates of the training set are formed and a classifier is learned on each of the bootstrap samples. The final $p(j|x)$ for an instance is obtained by taking the unweighted average of its probabilities given the models and the instance. Using optimal Bayes decision criterion as formulated by equation (3), each training instance is relabeled with its optimal class. In a final stage, the classifier is reapplied to the relabeled training instances to obtain the final model. More information on the Metacost algorithm can be found in the paper of [7].

3.2.3. Threshold adjusting

Threshold adjusting [24] tries to find to optimal threshold that minimizes the total misclassification cost. Practically, the total misclassification cost for every possible threshold on the posterior probabilities on the training set is calculated and the one that is minimal is used as the optimal threshold to predict the class labels of the test instances. A test instance with predicted probability above or equal to this optimal cut-off point is predicted as positive, otherwise as negative. In sum, the posterior probability that yields the lowest misclassification cost on the training set is retained and used as the optimal threshold to classify the test instances into predefined classes.

3.2.4. Weighting

Weighting assigns a certain weight to each training instance according to its class [26]. As such, the learning algorithm is in favor of the class with the highest weight. As Logistic Regression is the base learner within this research paper, a weighted Logistic Regression is applied. [9] proposes to weight the majority class by $\frac{c_{10}}{c_{01}}$ for classifiers with a probability threshold of 0.5 and $c_{00}=c_{11}=0$. In other words, the weights assigned to training instances resemble the cost ratios between false positive and false negative classifications. The weights are linearly incorporated into the calculation of the likelihood function. As such, the likelihood function ℓ for a given beta vector β and n instances is given by

$$\ell(\beta) = \prod_{i=1}^{n} w_i \{\pi(x_i)\}^{y_i} \{1 - \pi(x_i)\}^{1-y_i}$$

with w_i the weight for instance i, $\pi(x_i)$ is the conditional probability that y is equal to 1 for a given x_i or p$(y=1|x_i)$, while $y_i = 1$ for churners and $y_i = 0$ otherwise.

4. EMPIRICAL EVALUATION

4.1. Data sets

In this study, six real-world customer churn datasets are used to meta-analyze the performance of the non-sampling based cost-sensitive learners. All datasets are constructed for real-world churn applications in order by large companies. Consequently, these datasets are an ideal test bed to compare the different algorithms in a churn prediction context. Table 2 gives a description of the different datasets.

Dataset	Description
Newspaper	Defection of newspaper subscriptions
Pay TV	Defection of Pay TV subscriptions
Retail	Partial defection of customers on a 5 month period
Financial institution	Partial defection of current account holders on a 3 month period
Retail	Partial defection of customers on a 4 month period
Financial institution	Partial defection of current account holders on a 12 month period

Table 2: description of the churn prediction datasets.

As seen in Table 2, two types of churn prediction settings occur, namely contractual versus non-contractual. In the former setting, it is easy to observe whether a customer will end his/her current subscription. One is able to determine the exact point in time in which customers can interrupt their relationship with the company (i.e. the Newspaper and Pay TV dataset). In a non-contractual setting, it is often more complex to determine when a customer will end the relationship with the company. However, these customers do not tend to terminate their relationship with the company all of a sudden. They switch some of their purchases to another store which is called partial defection (i.e. the Retail and the Financial institution datasets).

Moreover, Table 3 gives an overview of the different datasets in terms of (i) industry, (ii) number of instances, (iii) churn incidence and (iv) churn level.

Dataset	Number of examples	Churn incidence	Churn level	Number of predictors
Newspaper	134,084	11.95%	Medium	39
Pay TV	143,198	13.07%	Medium	64
Retail	100,000	30.86%	High	26
Financial institution	117,808	6.29%	Low	54
Retail	32,371	25.14%	High	20
Financial institution	102,279	5.98%	Low	48

Table 3: descriptive statistics of the churn datasets.

The datasets are collected from companies in a rich variety of industries like one newspaper publisher, one pay TV distributor, two financial institutions and two retailers. In order to make a sensitivity analysis of the different algorithms in terms of churn incidence, three categories (*low* (less than ten percent churners), *medium* (between ten and twenty percent churners) and *high* (more than twenty percent churners)) are created.

In this study, experiments are conducted using the following cost model with $c_{00}=c_{11}=0$, $c_{10}=1$ and $c_{01}=r$ where r is set alternately to 2, 3, 5, 10 and 20. Note that the absolute values for c_{10} en c_{01} are irrelevant for algorithm comparison purposes, because only their ratio 1:r is significant.

The datasets are randomly divided into training and a test set using a 70-30 split. The training set is used to learn the model, while the test set is used to validate the model. Both datasets contain a proportion of churners that is representative for the true population in order to approximate a real-life situation. The total misclassification cost on the test set is used as performance measure to evaluate the different decision making alternatives.

4.2. Practical threshold

The threshold in many practical decision-making situations is based on the arbitrary probability of 0.5. Many marketing managers implicitly decide to classify a customer as a churner when the churn probability exceeds 0.5. Moreover, standard learning algorithms are often designed to yield classifiers that minimize error-rate, while their classification decisions are based on the probability threshold of 0.5 [9]. This practically-oriented threshold, hereafter abbreviated as P50, is benchmarked to the other more sophisticated ones in terms of total misclassification cost.

4.3. Empirical results

This Section reviews the empirical results of the meta-analysis for the different cost-sensitive learners. Paragraph 4.3.1. summarizes the results based on a pair wise comparison of the different techniques in terms of wins and losses. At this point, the reader will have a general view on the performance of the different cost-sensitive learners neglecting the cost ratios and the churn incidence. Furthermore, Paragraph 4.3.2. will unravel the general results by taking into account to cost ratios and the churn incidence.

4.3.1. General Results

The empirical results are summarized in Table 4. An entry w/l means that the approach at the corresponding row wins w runs and loses l runs compared to the approach at the corresponding column. Thirty runs, five cost ratios (that is 1:2, 1:3, 1:5, 1:10 and 1:20) on each of the six datasets, are available to compare the different algorithms.

	Metacost	Thresholding	Weighting	DMECC
P50	1/29	1/29	1/29	1/29
Metacost		18/12	13/17	16/14
Thresholding			11/19	12/18
Weighting				17/13

Table 4: summary of the empirical findings in terms of wins/losses table based on minimizing total misclassication cost.

Table 4 clearly shows that P50 performs always worse than the other cost-sensitive approaches. Incorporating unequal misclassification costs, while classifying your instances based on the arbitrary threshold of 0.50 results in a poor performance in terms of total misclassification cost. Moreover, incorporating costs/weights into the algorithm itself (Weighting) performs always better compared to the Relabeling techniques (DMECC and Metacost) and Threshold adjusting. On the other hand, Thresholding performs worse than Weighting and the Relabeling approaches (DMECC and Metacost). Within the Relabeling approaches, there is a slight advantage for Metacost compared to DMECC. A translation of the mutual challenges to a ranking based on the wins/losses is found in Table 5.

Low	Weighting ↔ Metacost ↔ DMECC ↔ Thresholding ↔ P50	**High**

Table 5: ranking of the approaches based on wins/losses (low=best, high=worst).

Table 5 shows the rank of the different approaches based on the wins/losses. The lower the rank, the better the approach. Weighting has clearly more wins than losses compared to the other algorithms, while P50 has always more losses than wins. Thresholding beats P50, but is beaten by Weighting and the Relabeling techniques. Comparing the Relabeling techniques, there is a slight advantage for Metacost compared to DMECC.

4.3.2. Sensitivity to cost ratio & churn incidence

In this Section, the sensitivity of the different cost-sensitive learners in terms of cost ratio and churn incidence is evaluated. Table 6 represents the results using the single best criterion. In other words, the figures in Table 6 represent the number of times that an approach has the lowest misclassification cost. For instance, a figure ‘8’ in the matrix means that the cost-sensitive learner won eight times in terms of lowest misclassification cost.

		Cost ratio		Churn level			Cost ratio *Low* Churn level			Cost ratio *High* Churn level		
	All	*Low*	*High*	*Low*	*Medium*	*High*	*Low*	*Medium*	*High*	*Low*	*Medium*	*High*
Metacost	8	8	0	1	3	4	1	3	4	0	0	0
Weighting	12	3	9	5	3	4	2	0	1	3	3	3
Thresholding	8	6	2	3	4	1	2	3	1	1	1	0
DMECC	1	0	1	0	0	1	0	0	0	0	0	1
P50	1	1	0	1	0	0	1	0	0	0	0	0
Total	30	18	12	10	10	10	6	6	6	4	4	4

Table 6: overview of the single best approach (cells represent # wins).

Table 6 clearly indicates that Weighting is the winner (twelve out of thirty) followed by Metacost and Thresholding (each eight wins out of thirty), while P50 only wins once over all the runs. These results are in line with the rank given in Table 5. Due to the fact that the performance of DMECC seems contradictory with the ranking in Table 5 as it is only once the winner, this discussion is postponed to the end of this Section.

In order to evaluate the sensitivity of the different approaches in terms of the cost ratio, two categories are created: the *low* cost ratio category contains cost ratios 1:2, 1:3 and 1:5, while the cost ratios 1:10 and 1:20 are categorized under *high* cost ratio. Table 6 let us conclude that Weighting outperforms the other techniques within the *high* cost ratio group. It performs best in nine out of twelve runs. However, Metacost and Thresholding are preferred techniques for *low* cost ratios. In detail, Metacost wins eight and Thresholding wins six out of the eighteen runs.

To evaluate how the cost-sensitive learners perform under circumstances with different churn levels, the datasets are divided based on their churn incidence into *low*, *medium* or *high*. A dataset belongs to *low* churn level when the churn incidence is less than ten percent, *medium* churn level when the churn incidence lies between ten and twenty percent, *high* churn level when the churn incidence is higher than twenty percent. Table 6 shows that Metacost performs better when churn level ranges from *medium* to *high*, while Thresholding performs well in cases with a *low* to *medium* churn incidence. Weighting performs similar over the different churn levels, while its performance is competitive to Metacost and Thresholding.

When one cross-tabulates the cost ratios with the different churn levels, several interesting conclusions are drawn. In case of *high* cost ratios, Weighting is the absolute winner in terms of misclassification cost. It is the best in three out of four runs for each churn level. When cost ratios are *low*, there is a dominance of Thresholding and Metacost. Thresholding has a good performance within *low* and *medium* churn level, while Metacost has a good performance when churn level is *medium* or *high*.

The DMECC is only once the absolute winner over all thirty runs based on the results in Table 6. This result seems contradictory with Table 5, where based on the mutual comparisons of the wins/losses, DMECC is ranked third. Based on Figure 1, one is able to clarify the results of DMECC. Figure 1 shows the absolute ranking of the cost-sensitive learners for the different datasets. For every cost ratio (1:2, 1:3, 1:5, 1:10 and 1:20) of a given dataset, a ranking of the different techniques based on the misclassification cost is done. A rank 1 means that for the given

cost ratio, this cost-sensitive learner has the lowest misclassification cost and thus the best approach, while a rank 5 indicates that the approach has the highest misclassification cost compared to the other techniques.

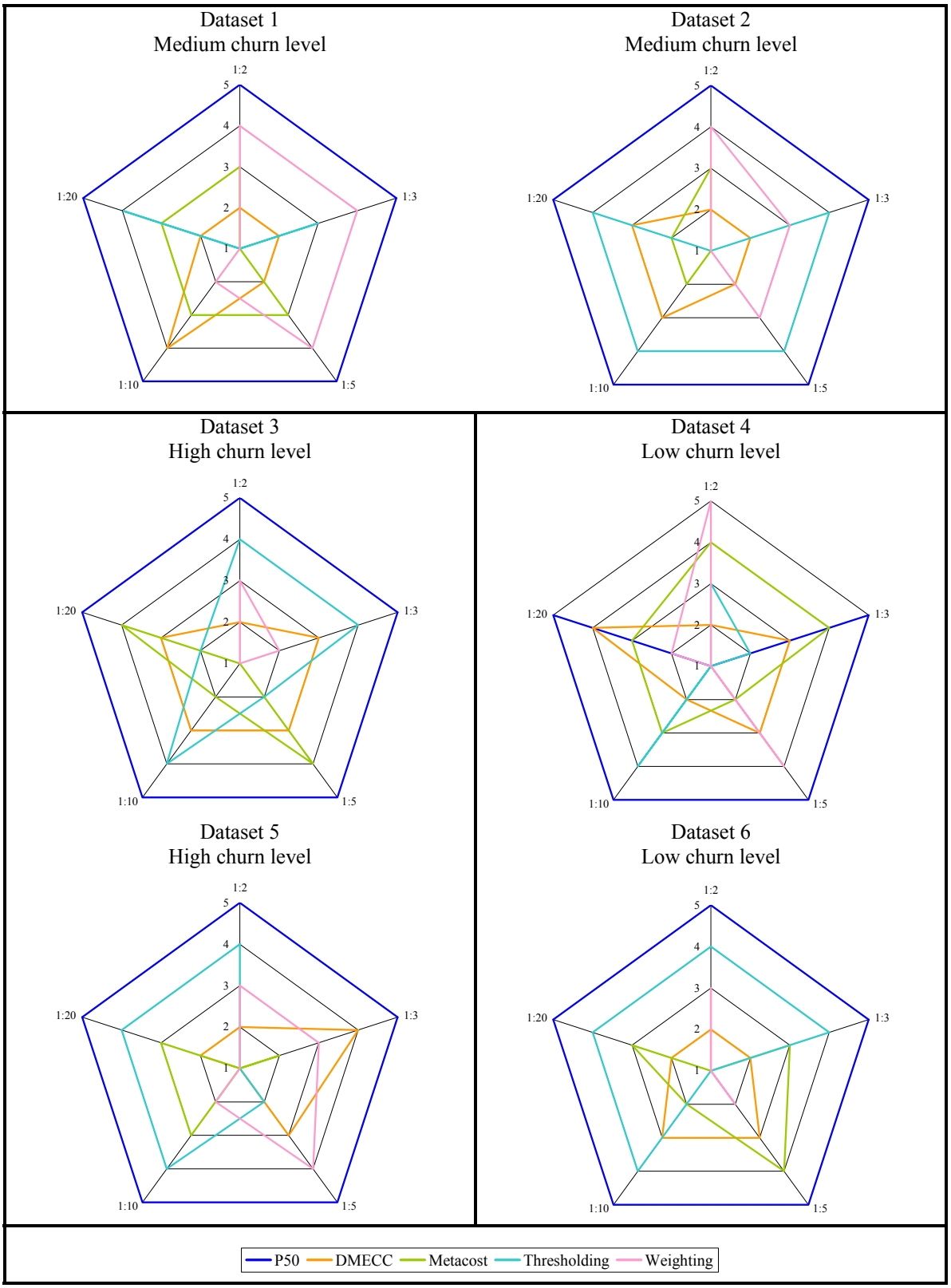


Figure 1: absolute ranking of the cost-sensitive learners per dataset (with rank 1 = lowest misclassification cost to rank 5 = highest misclassification cost).

Figure 1 shows that DMECC is the most robust technique in terms of cost ratio and churn incidence. The ranking of DMECC (orange) is in all datasets very close to the best technique or rank 1, while it is only once the best approach (in dataset 5 - churn level high - cost ratio 1:10). DMECC gives robust performance over the various churn levels and cost ratios. Figure 1 draws the same conclusions for Weighting, Metacost, Thresholding and P50 as stated above. For instance, P50 (blue) performs worse than the other techniques and is always but once ranked as '5' or Weighting (pink) dominates when cost ratios are high and is then closer situated to the center.

5. CONCLUSION AND DIRECTIONS FOR FURTHER RESEARCH

Nowadays, marketing decision makers realize that the information extracted from the customer database is a valuable tool to leverage their customer knowledge. Consequently, they are forced to take into account the long-term relationships with their customers which leads to a well-considered CRM strategy. One of the dimensions of CRM is customer churn prediction where one tries to predict whether or not a customer will stay with the company based on its historical information.

This study argues that incorporating different misclassification costs is of crucial importance for optimal decision making in a churn-prediction context. Indeed, categorizing a churner as a non-churner (false negative) is more costly than classifying a non-churner as a churner (false positive). In case of a false negative, the company loses the foregone profits, opportunities to cross and up sell,… while only a 'pampering' cost is lost when a false positive prediction occurs.

This study focuses on meta-analyzing various non-sampling based cost-sensitive learners that incorporate different misclassification costs. These non-sampling based cost-sensitive learners are categorized in three main categories; Relabeling (DMECC and Metacost), Threshold adjusting and Weighting. This study leads to managerial recommendations for optimizing the decision making process in a customer churn context (see Table 7).

		Cost ratio *Low*			Cost ratio *High*		
Cost ratio		Churn level			Churn level		
Low	*High*	*Low*	*Medium*	*High*	*Low*	*Medium*	*High*	*Low*	*Medium*	*High*
Metacost	Weighting		Metacost			Metacost			Weighting	
Thresholding			Weighting			Thresholding				
			Thresholding							

Table 7: managerial recommendations.

When cost ratios are high, Weighting is the most appropriate technique over all different churn levels. When cost ratios are low, Thresholding performs well in situations with churn level low to medium, while Metacost shows to perform well when churn level ranges from medium to high. Moreover, it is clear that P50 and DMECC are not included into Table 7. The practically-oriented classification threshold of 0.5 (P50) is not a viable alternative when cost asymmetries occur. Marketing managers need to search for alternatives to optimize decision making. Moreover, DMECC is not included into Table 7, because it is found that DMECC is the most robust alternative for the other techniques in any given situation.

Using the guidelines as presented in Table 7, it is possible to optimize the decision making process in a customer churn context.

While we strongly believe that this research paper contributes to the decision making literature, several suggestions for furthers research are given. This research paper uses Logistic Regression as the base learner. However, other base learners can be used and benchmarked to Logistic Regression. Moreover, the focus of this research paper lies on the non-sampling based cost-sensitive learning techniques. A suggestion for further research lies in adding sampling based techniques like oversampling, undersampling,… to the experiment. Furthermore, the research paper takes only single cost models into consideration. An additional experiment could be to contrast the performance of these techniques with Adacost, a variant of Adaboost that captures different misclassification costs. The concept of this research paper can also be extended to other than churn contexts.

ACKNOWLEDGMENTS

We would like to thank Ghent University for funding the PhD project of Kristof Coussement (BOF 01D26705). Moreover, we would like to thank Bart Larivière and Ilse Bellinck for their insights during this project and all PhDs who gathered research datasets on which this meta-analysis study is done.

REFERENCES

[1] D.M. Amidon, Blueprint for 21st Century Innovation Management, Journal of Knowledge Management 2 (1) (1998).

[2] W.H. Au, K.C.C. Chan and X. Yao, A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction, IEEE Transactions on Evolutionary Computation 7 (6) (2003).

[3] A.P. Bradley, The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms, Pattern Recognition 30 (7) (1997).

[4] L. Breiman, Bagging Predictors, Machine Learning 24 (1996).

[5] R.E. Bucklin and S. Gupta, Brand Choice, Purchase Incidence and Segmentation: an Integrated Modeling Approach, Journal of Marketing Research 29 (2) (1992).

[6] K. Coussement and D. Van den Poel, Churn Prediction in Subscription Services: an Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques, Expert Systems with Applications 34 (1) (2008).

[7] P. Domingos, Metacost: A General Method for Making Classifiers Cost-Sensitive, Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Diego, CA, USA, 1999).

[8] R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification (Wiley, New York, USA, 2000).

[9] C. Elkan, The Foundations of Cost-sensitive Learning, Seventeenth International Joint Conference on Artificial Intelligence (Seattle, WA, USA, 2001).

[10] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, From Data Mining to Knowledge Discovery: An Overview, Advances in Knowledge Discovery and Data Mining (AAAI press, Menlo Park, CA, 1996).

[11] C.W. Holsapple, R. Pakath, V.S. Jacob and J.S. Zaveri, Learning by Problem Processors: Adaptive Decision Support Systems, Decision Support Systems 10 (2) (1993).

[12] S.-Y. Hung, D.C. Yen and H.-Y. Wang, Applying Data Mining to Telecom Churn Management, Expert Systems with Applications, 31 (3) (2006).

[13] S. Keaveney and M. Parthasarathy, Customer Switching Behavior in Online Services: an Exploratory Study of the Role of Selected Attitudinal, Behavioral and Demographic Factors, Journal of the Academy of Marketing Science 29 (4) (2001).

[14] S. Kelly, Data Warehousing: The Route to Mass Customization (Wiley, New York, 1996).

[15] A. Lemmens and C. Croux, Bagging and Boosting Classification Trees to Predict Churn, Journal of Marketing Research, 43 (2) (2006).

[16] D. Margineantu, On Class Probability Estimates and Cost-sensitive Evaluation of Classifiers, In Workshop Notes: Workshop on Cost-sensitive Learning, International Conference on Machine Learning (2000).

[17] D. Margineantu, Methods for Cost-sensitive Learning, PhD thesis: Department of Computer Science, Oregon State University, Corvallis, OR, USA (2001).

[18] V. Nagar and M.V. Rajan, Measuring Customer Relationships: The Case of the Retail Banking Industry, Management Science 51 (6) (2005).

[19] S.A. Neslin, S. Gupta, W. Kamakura, J. Lu and C. Mason, Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models, Journal of Marketing Research 43 (2) (2006).

[20] F. Provost and T. Fawcett, Robust Classification for Imprecise Environments, Machine Learning 42 (3) (2001).

[21] F.F. Reichheld and W.E. Sasser, Zero Defections: Quality Comes to Services, Harvard Business Review 68 (5) (1990).

[22] M.J. Shaw, Machine Learning Methods for Intelligent Decision Support: an Introduction, Decision Support Systems 10 (2) (1993).

[23] M.J. Shaw, C. Subramaniam, G.W. Tan and M.E. Welge, Knowledge Management and Data Mining for Marketing, Decision Support Systems 31 (2001).

[24] V.S. Sheng and C.X. Ling, Thresholding for Making Classifiers Cost-Sensitive, In Proceedings of the National Conference on Artificial Intelligence 21 (1) (2006).

[25] T.S.H. Teo, P. Devadoss and S.L. Pan, Towards a Holistic Perspective of Customer Relationship Management (CRM) Implementation: A Case Study of the Housing and Development Board, Singapore, Decision Support Systems 42 (2006).

[26] K.M. Ting, An Instance-weighting Method to Induce Cost-sensitive Trees, IEEE Transactions on Knowledge and Data Engineering 14 (2002).

[27] G. Torkzadeh, J.C.J. Chang and G.W. Hansen, Identifying Issues in Customer Relationship Management at Merck-Medco, Decision Support Systems 42 (2006).

[28] S. Viaene and G. Dedene, Cost-Sensitive Learning and Decision Making Revisited, European Journal of Operational Research 166 (2005).

[29] W. Ziarko. The Discovery, Analysis and Representation of Data Dependencies in Databases. In: PieteskyShapiro G. & Frawley, W.J. (Eds.). Knowledge Discovery in Databases. MIT press, Massachusetts, Chapter 11 (1991).