UNIVERSITEIT GENT

FACULTEIT ECONOMIE EN BEDRIJFSKUNDE

VAKGROEP MARKETING

Issues in Predictive Modeling of Individual Customer Behavior: Applications in Targeted Marketing and Consumer Credit Scoring

Geert Verstraeten

2005

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

Promotor: Prof. dr. Dirk Van den Poel

Doctoral jury:

Prof. Dr. Dirk Van den Poel
(Ghent University)

Prof. Dr. Patrick Van Kenhove
(Ghent University)

Prof. Dr. Bart Baesens
(University of Southampton)

Prof. Dr. Edward C. Malthouse
(Northwestern University)

Prof. Dr. Stefan Van Aelst
(Ghent University)

Decaan Prof. Dr. Roland Paemeleire
(Ghent University)

Prof. Dr. Eddy Omey
(Ghent University)

ACKNOWLEDGEMENTS

At the end of this PhD thesis, it feels natural to dedicate a word to those people that have contributed – either directly or indirectly – to this work.

First, I would like to thank my advisor, Prof. dr. Dirk Van den Poel, because he introduced me to this fascinating domain when I enrolled into the Master of Marketing Analysis and Planning in September 2000. I am also grateful that – during this year – he encouraged me to join the Marketing Department at Ghent University in September 2001. But more importantly, during the process of my PhD, I have appreciated his focus on scientific research, but equally his interests for industry relevance, and the fact that he allowed and encouraged me to pursue my own research interests. Last but not least, I have appreciated his help regarding the IT infrastructure required in this type of work. I am also grateful to the members of the exam committee for offering their well-appreciated remarks and ideas on this work.

Additionally, I would like to thank several other junior researchers at the Marketing Department for improving the quality of this work. Wouter, I have learned a great deal through our cooperation in two research studies, and the hours we spent discussing our views. During our years in Leuven and our joint PhD adventures, I have greatly learned to appreciate your intelligence, humour, hard-working mentality and sense for team spirit. I am fully confident that we are well equipped for facing our upcoming professional adventures 'in the real world'. Larrie, I have greatly appreciated the thought-provoking discussions we had about our research, and thus I would like to thank you for those learnful interesting interaction moments. Additionally, I am very greatful to Bernd, Jonathan, Larrie and Wouter for sharing their datasets with me, which enabled me to perform the latter studies in this dissertation.

In general, I would like to thank the whole staff at the Marketing Department for making these last years so pleasant. I have greatly enjoyed getting to know you all and spending these few years together, and I cherish the memories of attending conferences, but also playing soccer and pool, throwing snowballs and water balloons, salsa dancing, swimming in Patrick's pool, exploring Ghent's restaurants, eating Nobel-Prize quality cake together while sharing the research blues, and much more. Tine and Larrie, thanks for supporting me and my general insanity in our office for the past four years. At this moment, it is difficult to imagine that the day is near that we will no longer work at the same location, and I am sure I will miss you both. Jonathan, Bernd, Isabel, Marie, Kristof, Tine, Katrien, Nele, Sarah and Leen, I wish you all the best during your own PhD adventures.

I am grateful to friends and family for supporting me, and offering the diversion that was extremely welcome during the process. A special thanks goes to my grandmother, Josie, for being there for me when I need(ed) her help. Thanks to my sister Veerle, for doing all the stuff that older sisters do. I admire your mental power and ambition on a personal as well as a professional level, and I will always look upon your ventures with great enthusiasm. Wouter and Greet, and Yves and Janne, thanks for having me always feel so welcome at your homes in good and in worse times. I would also like to thank the friends of my three favorite soccer teams, Maptiko, d'Anciens, and ZVC Foot, for reminding me weekly in a very pleasant way of the values of being a team player. Finally, a big 'merci' to the handful of good friends who have supported me throughout the process, especially during the hectic last months. I have greatly appreciated your company at lonely PhD times, and I promise that I will return the favor whenever you need it.

Geert Verstraeten, December 22, 2005

TABLE OF CONTENTS

NEDERLANDSTALIGE SAMENVATTING

Dankzij de brede opkomst van klantenkaarten en getrouwheidsprogramma's, zijn bedrijven uit een grote diversiteit van economische sectoren er in de laatste decennia in geslaagd om enorme transactionele klantendatabanken aan te leggen. Deze databanken registreren vaak alle interacties die plaatsvinden tussen het bedrijf en haar klanten, zoals aankoopgedrag, informatieaanvragen, klachten, etc. In het kader van predictieve voorspellingsmodellen wordt juist deze transactionele informatie gebruikt om voorspellingen te maken van toekomstig klantengedrag op het niveau van de individuele klant. Zo kan men o.a. aan de hand van het historische aankoopgedrag van een klant gaan voorspellen of deze nog verdere aankopen zal plegen bij het bedrijf in kwestie, of hij/zij zal reageren op doelgerichte aanbiedingen, of hij/zij geïnteresseerd is in bepaalde producten, of hij/zij in de toekomst meer zal uitgeven, maar ook bv. of hij/zij in staat zal zijn om het verschafte klantenkrediet terug te betalen. Samenvattend kan dus gesteld worden dat de toepassingen van het voorspellen van individueel klantengedrag zich voornamelijk lenen tot de domeinen van *targeted marketing* en de inschatting van het risicoprofiel van de klant in het kader van het aanbieden van commercieel klantenkrediet (*consumer credit scoring*). In dit proefschrift gaan we in op een grote variëteit van toepassingen bij het modelleren van individueel klantengedrag, zijnde het voorspellen van klantentrouw, toekomstige uitgaven, (partiële) verloop van klanten, targeting en credit scoring. Empirische resultaten werden bekomen voor bedrijven o.a. aanwezig in retail, postorder, telecom en de financiële sector. We onderscheiden hiertoe vijf verschillende fasen in het modelleringsproces: (1) projectomschrijving, (2) creatie van de analysetabel, (3) constructie van het voorspellingsmodel, (4) validatie van het model op ongeziene data, en (5) implementatie en praktische toepassingen. Doorheen zes verschillende studies werden bijdragen geleverd in de verschillende fasen van het modelleringsproces.

De twee eerste studies zijn gericht op het vergelijken van de voorspellingskracht van verschillende predictieve technieken uit de domeinen van data mining en statistiek, zijnde lineaire en logistische regressie, lineaire en kwadratische discrimintantanalyse, beslissingsbomen (C4.5 & C4.5 rules), Bayesiaanse netwerken, ARD neurale netwerken en

random forests. Hierbij wordt uitdrukkelijk gekozen voor technieken die naast een goede performantie ook de interpreteerbaarheid van de modellen toelaten. De toepassingen van deze studies liggen respectievelijk in het modelleren van de evolutie van het uitgavenpatroon van klanten en het modelleren van klantentrouw. We besluiten in dit onderdeel o.m. dat de traditionele statistische technieken vaak een goede betrouwbaarheid vertonen voor het voorspellen van individueel aankoopgedrag, gegeven dat ruimschoots aandacht geschonken wordt aan een correcte validatie en het beperken van overfitting en multicollineariteit door de toepassing van variabele selectietechnieken.

In een derde studie onderzoeken we het nut van het belonen van klanten volgens klantentrouw. De mate waarin klanten trouw zijn aan de betrokken winkel is voor veel bedrijven onbekende informatie, bedrijven weten doorgaans enkel hoeveel de klanten bij de winkels van de keten uitgeven, en beschikken slechts uitzonderlijk over uitgaven bij de concurrentie. Op basis van een beperkte steekproef van klanten zijn we erin geslaagd de klantentrouw te voorspellen voor alle klanten van een retailer, waardoor deze informatie bv. kan gebruikt worden als beloningscriterium in een getrouwheidsprogramma. In deze studie bewijzen we dat een dergelijk beloningsprogramma er beter in zou slagen de klanten te bereiken die fluisterreclame veroorzaken, hogere aankoopintenties hebben, en een lagere prijsgevoeligheid vertonen.

Een vierde studie richt zich in het bijzonder op een probleem in customer credit scoring. Indien een nieuwe kredietscore gebouwd wordt, heeft men enkel de uitkomst van bestellingen die in het verleden aanvaard werden, en per definitie moet het model gebouwd worden op een steekproef die niet representatief is voor de toekomstige kredietaanvragen. Door de unieke karakteristieken van de beschikbare data set konden we de impact van deze bias nagaan op de performantie en winstgevendheid van credit scoring modellen. We besluiten in deze studie dat deze impact significant doch klein is.

In de twee laatste studies uit dit doctoraal proefschrift onderzoeken we de validiteit van onze bevindingen overheen verschillende sectoren en applicaties binnen targeted marketing en consumer credit scoring. In een vijfde studie tonen we aan dat het gebruiken van een lukrake opsplitsing in training en validatieset een grote instabiliteit van de resultaten teweeg kan brengen. Hoewel vaak beweerd wordt dat dit effect onbestaand is in grote datasets, merken

we hier het tegendeel. Bovendien bieden we een manier aan die het potentieel heeft om de variabiliteit van de opsplitsing tot 800 keer te reduceren.

In de zesde en laatste studie evalueren we het gebruik van verschillende variabele selectietechnieken, en we onderzoeken de kostprijs (in termen van predictieve performantie) van het opdrijven van de 'face validity' van een predictief model. We besluiten in deze studie dat de voorspelbaarheid niet noodzakelijk daalt wanneer we erop toezien dat de parametertekens in het finale model overeenstemmen met de univariate tekens, zodat het predictieve model een grotere kans heeft aanvaard te worden door managers, werknemers en klanten.

OVERVIEW

1. INTRODUCTION

In the latest decades, marketing has known a remarkable evolution. Whereas previously, product managers had the largest responsibilities when devising marketing actions, nowadays, the focus on products has shifted largely to a focus on customers. To indicate the impact this has had on marketing, different authors relate to this evolution as a paradigm shift (Grönroos, 1997; Kotler, 1991). One of the main triggers of this evolution was undoubtedly the belief that it is several times less demanding – i.e. expensive – to sell an additional product to an existing customer than to sell the same product to a new customer (Rosenberg and Czepiel, 1984). Following this reasoning, companies have increasingly focused on nurturing the customer-company relationships, and the term Customer Relationship Management (CRM) soon became one of the central topics in popular and academic marketing literature (Rigby et al, 2002). Additionally, this focus was confirmed by parallel evolutions in the understanding of the interrelationships between customer satisfaction, trust, commitment (see, e.g. Garbarino and Johnson, 1999), and the ultimate goal in building relationships – customer loyalty, where the latter soon grew to become one of the most important concepts in marketing recently (Reichheld, 1996).

The increasing focus on loyalty has soon led to the adoption of reward programs across a variety of companies and industries. Today, companies such as American Airlines, American

Express, AT&T, Carrefour, Hertz, Hilton Hotels and Shell have adopted reward programs that grant advantages to their customers, proportional to the money spent at their stores. While an assessment of the impact of such programs on customer loyalty remains difficult (see, e.g. Mägi, 2003), an undeniable beneficial effect of these programs lies in the collection of behavioral data on an individual customer level. In a number industries (such as retail banking), companies were already able to collect transactional data for their individual customers. However, since the widespread adoption of reward programs, an increasing number of companies are capable of understanding customer behavior throughout the whole customer-company relationship. Hence, nowadays, a critical mass of companies have reached the possibility to analyze behavior on a relational level as opposed to a transactional level. Complementing these evolutions in the marketing domain, the simultaneous progress in the domain of computer sciences allowed to store such transactional data in large data warehouses, due to a remarkable reduction in the cost of data storage and manipulation. To cite only a few examples, Wal-Mart currently serves 100 million customers weekly, and MasterCard alone processes 15 million transactions a day in 210 countries. Furthermore, it has been estimated that the amount of data stored in the world's databases doubles every 20 months (Witten and Frank, 2005). It must be clear, however, that the mere possession of such vast amounts of data does not offer benefits on itself. Nowadays, the real return on investment of reward programs is only bounded by the quality of the analyses of these transactional data, and the creativity to apply the gained knowledge in (targeted) marketing campaigns. To conclude, in order to follow through on this new path, new competences were required in the marketing arena.

One of the more challenging exercises using the vast amounts of data in real-life commercial transactional databases exists in the construction of predictive models that can be deployed for managing customer relationships. In this dissertation, we will focus on a number of issues in this specific domain. As an introductory example, a given company might be interested in increasing the length of its customer relationships by encouraging customers to continue purchasing at its stores. In such a situation, it may be beneficial to target the customers at risk, i.e. the customers that have a relatively high potential of leaving the company. In this situation, a predictive model will be constructed that determines the probability of leaving the company for every single customer in the transactional database. Hence, the information that resides in the database cannot only be used to describe the current state of affairs, it can be used to predict what customers will do next, which can turn

into very profitable applications for managing customer relationships. As a few examples, Van den Poel and Larivière (2004) show that a one percent increase in customer retention can lead to an increase in total contribution of more than seven percent on a 25-year time frame. In a second example, in a real-life test at a European retailer, Buckinx (2005) has provided evidence that an improvement of the targeting of customers for a promotional leaflet led to an extrapolated yield of over 200 000 euro per two-weekly mailing, indicating an increase of the total company profitability of five per cent. Hence, the bottom-line results of the analysis of transactional customer data can be large, and offer leverage to the vast amounts of information stored in commercial databases. In this doctoral dissertation, we focus on a number of applications in a variety of industries where opportunities of predictive modeling are present, and we attempt to indicate the managerial benefits arising from these applications.

In this dissertation, we will focus on two main applicational domains, namely the domains of targeted marketing and consumer credit scoring. Whereas the use of predictive modeling for targeted marketing is a more recent venture, methods for assessing the credit risk when lending to consumers have been in operations for fifty years (Thomas et al, 2005). Because of the similarities in the techniques and data used for both applications, we focus on both domains jointly in some studies, while other studies will be directed towards either targeted marketing or credit scoring. In the remainder of this introductory chapter, we first sketch the origins of the domain of predictive modeling relative to two of its ancestors, data mining and statistics. Next, we formulate our own conception of five distinct phases in the process of predictive modeling. We continue with a brief overview of the six studies performed in this dissertation, and we end this chapter with a number of general conclusions and directions for further research.

2. DATA MINING OR STATISTICS?

As indicated previously, before the turn of the millennium, our capabilities for collecting and storing data of all kinds had far outpaced the abilities to analyze, summarize, and extract "knowledge" from these data (Fayyad et al, 1996). The development of such analytical competences originated in two distinct scientific communities, namely statistics and data mining. While statistics was the only analytical solution until the second half of the 20th century, computer science, with its subdiscipline of data mining, has since grown into a vast

edifice (Hand, 2004). Whereas the growth of the latter discipline has thrived on the abundance of data, statistics has essentially known a very different beginning. Originally, statistical analysis was performed on data sets restricted to a few (tens of) variables and a few hunderds of observations. Hence, the visualization of data, the use of tedious hypothesis testing procedures and rigourous data collection were of crucial importance in the domain. However, due to the new data availabilities, data visualization has become very difficult at least (a scatter plot turns into a black plane when a million customers are envisionned) and traditional measures of statistical significance are severely influenced by large data sizes. Moreover, the correction of harmful correlations present in the data by performing additional data collection (as suggested in e.g. Rawlings, 1988) is difficult at least. Indeed, instead of working on a sample of the population, in predictive modeling projects, it is not uncommon that the whole population is at one's disposal.

On the other hand, with the development of computer science, the transition from storage and manipulation of large databases to data analysis was hardly a large step (Hand, 2004). Hence, data mining has been defined as "the science of extracting useful information from large data sets or databases" (Hand et al, 2001). In contrast to the rigourous statistical procedures, the field was created to offer solutions to managerial questions *and* to detect business opportunities in a timely, automated manner (Witten and Frank, 2005). The best known marketing example of such an automated data mining solution probably exists in the detection of an assocation in purchasing beer and diapers, which may be used for store layout decision making. To many, the essence of data mining is the possibility of serendipitous discovery of unsuspected but valuable information. This means the process is essentially exploratory. However, statisticians are careful about the ad hoc analysis of a set of data implied by the term data mining because they are aware that an overly intensive search is likely to reveal apparent structures purely on the basis of chance (Hand, 1999). Indeed, databases can contain terabytes (10^{12} bytes) of data, and this abundance of data increases the odds that a data algorithm finds spurious patterns that are not valid in general (Fayyad et al, 1996). However, instead of statistical significance, algorithmic complexity and performance were the main focus of this new domain of data analysis.

As both domains matured, however, the initial differences have faded. In the field of statistics, much recent work has focussed on problems involving large data sets. A good overview of such advances can be found in Elder & Pregibon (1996). In the field of data

mining, it was soon obvious that the statistically well-appreciated concepts of validation, overfitting, and the tradeoff between complexity and predictive performance proved to be of great value (Elder & Pregibon, 1996). In an interesting paper, Breiman (2001) defended his view that the data modeling (i.e. statistics) and the algorithmic modeling (i.e. data mining) cultures can both deliver a contribution to the common goal of predictive modeling, and that sound (statistical) data models do not always deliver the best solution. Where statistical models had been more common in practice, he favors the idea that the problem and the data should guide the tools to be used. While the (lively) discussion of this paper featured both allies and adversaries, currently, an increasing number of authors view both fields rather as complementary than as conflicting. Hand (2004) recently advanced the use of the more neutral term *data analysis* to describe the use of both domains for the same goal of making informed decisions. As another example, (logistic) regression analysis was recently added by Witten and Frank (2005) to their toolbox of data mining techniques, and they state that 'one should not look for a dividing line between data mining and statistics because there is a continuum of data analysis techniques, whereby some derive from a statistical background, while others have arisen out of computer science.' In this dissertation, we comply with this evolution, and we specifically do not refer to either concepts in the title of this work. Instead, we focus on the common goal: building solid predictive models of individual customer behavior. Nevertheless, in different studies, we will use techniques from both backgrounds, and compare their predictive performances. In the following paragraph, we offer an attempt to break up the predictive modeling process into a number of distinct phases.

3. THE PREDICTIVE MODELING PROCESS

In this section, we divide the predictive modeling process into five distinct phases. A similar effort can e.g. be found in Fayyad et al (1996), in their description of the process of Knowledge Discovery in Databases (KDD). Our overview of predictive modeling is similar to the KDD process of Fayyad et al (1996) in the sense that it recognizes that a blind application of techniques without a focus on practicability, validation and interpretation does not lead to solid data analysis. However, the goals are quite different: in the context of KDD, description tends to be more important than prediction (Fayyad et al, 1996), whereas in this work, prediction is the central topic. In Figure 1, we have grouped some key concepts of the predictive modeling process into an overview. The purpose of this effort is twofold: first, it is employed to provide the reader with an insight into some of the more important issues in

the modeling process; second, in a later section, it will prove a useful steppingstone for situating the topics, the variety and the conclusions of the different studies in this dissertation.

In the following, we describe each of the phases presented in Figure 1. Note that we do not claim that we hereby provide the only way to view upon the predictive modeling process. For example, besides the KDD process described above, another fruitful approach can be found in the CRISP-DM (www.crisp-dm.org) data mining process, developed by data mining practitionners, and representing a cross-industry standard process for data mining. Knowing that different authors and different disciplines might regroup the phases differently, we do believe that the scheme suggested in Figure 1 offers a workable methodology for the predictive modeling of customer behavior.

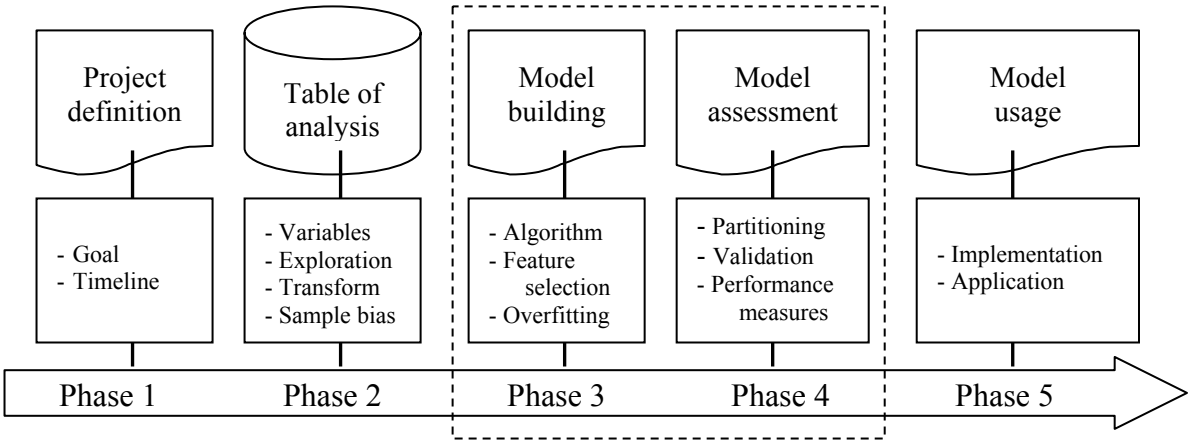


Figure 1: Phases in the predictive modeling process.

Phase 1: Project definition. In this phase, the goals of the analysis are specified. A wide array of predictive modeling problems are currently addressed in academia and industry. For example, companies might engage in predicting (i) which customers will respond to a mailing or an offer, (ii) which products a certain customer will be interested in, (iii) which customers are inclined to discontinue the relationship, (iv) the future evolution of a customer's spending, or a customer's lifetime value, (v), which customers are loyal to the company, (vi) which customers will be able to refund their credit debt, etc. Hence, by defining the goal of the analysis, we define the conception of the target (i.e. outcome, response, dependent) variable. Note that this target variable may be binary (e.g. will the

customer respond), ordinal (e.g. how much products will he/she purchase during the following period), continuous (e.g. how much money will he spend) in nature, etc.
In order to allow for a future prediction of behavior, it is crucial that one is able to reconstruct the past in the exact same way the predictive model will be used in the future. For example, if a model will be used in the future to detect whether a customer will purchase during the next two weeks, based on his purchases in the past year, in order to build the model, it is necessary to mimic this situation at an exact moment in the past, where both the knowledge of dependent and independent variables were available. Hence, a timeline needs to be constructed that clearly defines which part of the historical information will serve to create respectively the dependent and independent variables.

Phase 2: Table of analysis. In this phase, based on the data available in the data warehouse, we construct a table of analysis that complies with the definitions of the goal and the timeline set during the previous phase. This 2-dimensional table usually contains one (or more) target variable(s), a number of independent variables (also features, attributes, explanatory or predictor variables), and one or more identification variables, used to identify the customer. In predictive modeling, these variables are usually computed for a large number of customers, which present the observations (instances) in the table. The creation of this table is optimally followed by a stage of data exploration, which includes an assessment of the data quality and missing values, the computation of useful variable transformations, and the detection of outliers.
Since this table of analysis represents the sample on which the model will be built, it is crucial that this sample is representative for the future population on which it will be applied. Unfortunately, this ambition is not always achievable. For example, in a credit scoring model update, the future population of credit applicants will contain both good and bad applicants, whereas the historic population, and the records on which the credit score will be built, might contain a far lower proportion of bad debtors, because we only have available the outcome of the credit applicants that were accepted by the historical credit assessment. This situation is called sample bias, and will form the topic of one of the research studies.

Phase 3: Model building. The actual construction of the model, together with the next phase of model assessment, represent the most important building blocks of the predictive modeling process. In terms of model building, a wide array of techniques and algorithms have been proposed in either the statistical or data mining communities, and new

developments are still heavily pursued. Some of the widely recognised solutions include linear and logistic regression analysis, neural networks, decision trees, bayesian belief networks, support vector machines, genetic algorithms, etc. While each of these techniques has its own characteristics, we will not engage in a further description of the differences. Instead, we refer to Witten and Frank (2005) for a general description of useful predictive techniques, and to the following chapters in this work, where the techniques that are employed throughout each study are described at length.

As described previously, the predictive modeling of customer behavior often requires an analysis of a large number of observations by a large number of predictive features. This large *dimensionality*, however, causes a number of important issues in predictive modeling. The usage of a very large number of independent variables obviously increases the complexity of the solution. Additionally, the correlations that exist amongst predictive features in predictive modeling applications often imply that their parameters can no longer be interpreted due to the existence of *multicollinearity*. However, the large dimensionality may also have negative consequences on the predictive power, due to the existence of *overfitting* in the model building stage. Described briefly, a model that is built on a given data set may be overly optimistic if the results are evaluated on the same data set. Indeed, and especially when large dimensionality occurs, a model might fit very closely to the specific properties of the data used to build the model, whereas it is no longer representative for the overall pattern that may be present in the data. In a number of studies in this work, we will provide an illustration of the effects of the reduction of the dimensionality on interpretation and predictive performance.

Phase 4: Model assessment. The existence of overfitting, and the possible detection of spurious patterns that are only characteristic for the data set on which the model was built, imply the necessity of a rigourous validation of predictive models. Hence, both the domains of data mining and statistics have long realized the need of validating predictive models on an independent set of data not used for model building (Elder and Pregibon, 1996). The most common way of constructing an objective validation set consists in partitioning the table of analysis into a *training set*, used for building the model, and a *validation set*, which is not used in the model construction process, and can hence serve to build an objective indication of the predictive performance of a model. However, while the use of a single hold-out split still prevails, data-intensive (cross)validation procedures have long been suggested to evaluate (the differences in) predictive model performance.

Additionally, in several occasions, model builders may generate a number of competitive predictive models. In such setttings, it may be useful to split the table of analysis into three partitions, where one is used for model building, one for model selection and one for model validation.

Phase 5: Model use. The ultimate goal of the construction of a predictive model lies in its deployment by the end user, hence, the interpretation and face validity are important issues in the predictive modeling process. Additionally, in this work, we have attempted to focus on suggesting possible applications of the constructed predictive models. In some situations of predictive modeling, the applications are self-explanatory. The main application of a response model lies in better targeting the customers, whereas the main application of a credit scoring model lies in correctly evaluating credit risk. However, in one of the studies, the main focus of the paper will lie in the application of a constructed predictor.

4. DESCRIPTION OF THE STUDIES

In this section, we provide further details about the research studies involved. In Table 1, we present an overview of the title of each research study, and the chapter in which each study is presented in this work.

Table 1. Titles of the different research studies in this dissertation

Chapter	Title
1	Bayesian Network Classifiers for Identifying the Slope of the Customer Lifecycle of Long-Life Customers
2	Predicting Customer Loyalty using the Internal Transactional Database
3	Towards a True Loyalty Program: Investigating the Usefulness and Feasibility of Rewarding Customers According to the Benefits They Deliver
4	The Impact of Sample Bias on Consumer Credit Scoring Performance and Profitability
5	Using Predicted Outcome Stratified Sampling to Reduce the Variability in Predictive Performance of a One-Shot Train-and-Test Split for Individual Customer Predictions
6	Evaluating the Performance Cost of Improved Face Validity: Benchmarking Feature Selection Techniques in Logistic Regression for Individual Customer Predictions

The first three studies of this dissertation are aimed at predicting customer behavior for targeted marketing applications. In a first study, we responded to the recent findings in relationship marketing literature that large differences exist within the group of long-life customers in terms of spending and spending evolution. Thus, we attempt to predict whether a newly acquired customer will increase or decrease his or her future spending from initial purchase information. The focus of this study lies on the predictive performance of different architectures of Bayesian networks when compared to decision trees and discriminant analysis. The most important conclusions of this study are that Bayesian networks succeed in combining acceptable predictive power with the construction of a very parsimonious solution, where the variety of goods purchased, together with the initial purchase volume prove to form the best predictors of future spending evolution.

Also in the second study, we will compare a number of predictive techniques, namely logistic regression, random forests (as an evolution to the well-known decision trees) and automatic relevance determination neural networks. Again, the historical purchasing behavior stemming from transactional data was used, yet here, we aim to predict behavioral loyalty. Indeed, a retailer might know how much a certain customer spends at its stores, but has no information on the purchase behavior of this customer at competitive stores. In this study, we enrich the customer database with a prediction of a customer's behavioral loyalty such that it can be deployed for targeted marketing actions without the necessity to measure the loyalty of every single customer. This study shows that, given the use of variable selection techniques for improving the interpretability and the predictive power of the final model, the linear regression model significantly outperforms the other predictive techniques. Again, the variety of products purchased appears as one of the most important predictors of the target variable.

In the third study, we examine whether the previously constructed prediction of customer loyalty can be efficiently deployed as a reward criterion in the currently used 'loyalty' programs. Whereas customers are currently rewarded based on their historical spending or the length of the customer relationship, the previously described prediction enables companies to reward customers according to their real behavioral loyalty. Using historical purchase data, we show that if customers were rewarded for their predicted behavioral loyalty instead of past spending or length of relationship, the rewards received would better compensate

customers who are spreading positive word-of-mouth, are price insensitive and have high repurchase intentions. In other words, the usage of predicted loyalty may present a more beneficial proxy variable to real loyalty than the currently used criteria. These results were validated in both grocery and general merchandise shopping.

The fourth study is directed towards a specific issue in consumer credit scoring, namely sample bias. For customers who were historically not accepted for purchasing on credit, it is not known whether they would have been able to refund their debt if they would have been accepted. When a new credit score is developed based on the historical data, the problem arises that a specific part of the population (namely those customers previously assessed as bad debtors) will not enter the model building process, while customers with such a profile will appear in the future applicant population. Based on the specific properties of the data set acquired, in this study, it was possible to assess the impact of this bias. In this study, we make use of a logistic regression model, and we argue that the effect of sample bias on predictive performance and profitability in a consumer credit scoring model is significant albeit modest, especially when the cost of correcting for sample bias by accepting bad credit risk orders is accounted for.

In the previous studies, as often in research in the domain of predictive modeling of individual customer behavior, conclusions were drawn based on the application in a single empirical setting. In the fifth and sixth studies of this dissertation, however, analyses are performed based on a collection of data sets, amongst which applications in targeted marketing and consumer credit scoring. In the fifth study, based on a study across six real life predictive modeling applications, we illustrate that the use of a random data partitioning in training and validation set may cause a large instability of the results. In other terms, if the result of a predictive model would be validated based on a different random data partitioning, a very different performance assessment may arise. In this study, we show the usefulness of a different sampling procedure for reducing the variability in the results.

Finally, in the last study, we perform an evaluation of different variable selection techniques in a logistic regression model. Based on an evaluation on nine real-life predictive modeling applications, we evaluate the use of a variable selection technique that ensures that the influence of all variables on the predictive model will be consistent with the univariate relationship between the predictor and the dependent variable. Hence, we envision an

exclusion of all sign violations, implying technically that the signs of all parameters of the final model should correspond to the signs of their univariate counterparts, whereby the interpretability and acceptability of the model is increased for managers, employees and customers. In this study, the predictive performance and benefits of this feature selection technique are carefully compared with the performance of other, frequently used, feature selection techniques. We show that a variable selection technique that excludes sign violations in a predictive model does not generally exhibit reduced predictive performance.

5. GENERAL CONCLUSIONS AND DIRECTIONS FOR FURTHER RESEARCH

In this final section, we will focus on drawing some general conclusions of this work. We will again employ the overview of the predictive modeling process as depicted in Figure 1 in order to structure the main research results.

Phase 1: Project definition. In this phase, the goals and time schedule of the analysis are defined. Study 1 presents a clear example of such a time schedule. However, the goals of the different studies are diverse. In Study 1, we predict the customer's future spending evolution, whereas in Study 2 and 3, the prediction of behavioral loyalty is central. Study 4 focusses on credit scoring, and in Studies 5 and 6, we compare the results across a variety of applications, including loyalty, spending, churn, partial churn, targeting and credit scoring. While we will attempt to predict a continuous variable in Studies 2 and 3 the other studies focus on the prediction of binary target variables.

Phase 2: Table of analysis. In each application, a large number of variables were created, and the exploration of these variables led to the addition of variable transformations, which were added as candidate predictors. Because sample bias can be considered as a threat to predictive performance and profitability in credit scoring models, Study 4 focusses on assessing the effect of this bias in a real-life consumer credit setting. The overall conclusion is that the impact of sample bias is significant albeit modest in this application.

Phase 3: Model building. In this crucial predictive modeling phase, we focused on the use of different algorithms, the application of feature selection techniques, and the existence of overfitting. First, in Studies 1 and 2, we engage in a comparison of a number of carefully selected techniques for predicting a binary and a continuous outcome variable respectively.

In Study 1, while we concluded that Bayesian networks offer a viable alternative predictive technique, the results also indicate that the predictive performance of linear discriminant analysis does not significantly differ from the performance of the best Bayesian network solution. In Study 2, we clearly prove that the proposed linear regression model significantly outperforms the other predictive techniques. From this, we conclude that the predictive performance of statistical techniques, such as linear discriminant analysis and linear regression analysis is not necessarily inferior to the performance of the other techniques stemming from the data mining field. This finding is consistent with other studies that perform such benchmarking studies (see, e.g. Baesens et al, 2003; Dasgupta et al, 1994; Davis et al, 1992). However, the selection of good predictive variables proved to be a key success factor in using linear regression in Study 2. Both Studies 2 and 6 confirm that the different variable selection procedures reduce multicollinearity, whereby the face validity and the interpretability of the final model is increased. Additionally, in Study 6 we propose to alter one of the existing variable selection techniques in order to increase the face validity of the model, and we carefully assess the performance cost related to this alteration.

Phase 4: Model assessment. While the previous phase is crucial, the assessment of the real predictive performance of a model is equally important. Since a single split into training and validation set is still frequently deployed, Study 5 specifically focusses on the variability that may exist due to an often used random split into training and validation set, and illustrates the benefits that may arise by using an alternative splitting procedure. By using the predicted outcome stratified sampling procedure suggested in Study 5, the variation in one specific case was over 800 times lower compared to the use of random sampling. In Study 6, we deploy a more intensive validation procedure, namely 10 times 10-fold cross-validation, in order to detect significant differences in the performance of different variable selection procedures.

Finally, as stated before, model builders may need to choose between alternative models built. In this case, a specific part of the data should be reserved for model selection, next to the parts used for training and validating the model. The use of such procedures is illustrated in Study 2, given the additional complexity of using a sparse data set.

Phase 5: Model use. This dissertation is presented in a domain of applied sciences, and hence our focus lies on testing the application of techniques, algorithms and methodologies on real-life predictive modeling examples. Our interest in the applicability of our work also

implies that, in a number of studies, we attempt to assess the managerial impact of different alternatives. For example, in Study 4, we examine the benefits that may arise in terms of predictive performance and profitability when sample bias can be removed in a real-life consumer credit scoring application. In Study 6, we evaluate the cost in predictive performance lost by increasing the face validity of the final solution in order to make the models more interpretable and acceptable to management. Finally, our interest in applications in this domain encouraged us to focus on generating (and evaluating) useful methods for applying the outcome of the predictive models developed in Studies 1 and 3. Note that the main focus of Study 3 lies in evaluating the usefulness of applying the constructed predictive model.

Finally, in terms of important predictors, we have proven across different settings that the purchase variety can be a very important behavioral indicator of the quality of the customer-company relationship. In Studies 1 and 2, it was a crucial predictor of future spending evolution and behavioral loyalty respectively, and the measure outperformed the most reknown behavioral variables such as recency, frequency and monetary value of previous purchases (see, e.g. Cullinan, 1977). To conclude, because the different studies have different goals, we refer to the later chapters for the more specific conclusions of each research study.

Ample opportunities exist for further research. While the first studies in this dissertation have focused on testing a methodology on a single applicational domain, in the two last studies, we were able to compare the results across different real-life settings. It is our belief, that the field of predictive modeling of individual customer behavior has a higher need for studies of the latter type, where the consistent application of different methodologies is performed on a range of data sets, in order to determine the validity of the proposed procedures in the specific domain of focus. For example, the experimental design and different data sets used in this study can be used to compare the predictive performance of a wide array of competing predictive techniques, stemming either from statistical or data mining backgrounds. Similarily, in the studies in this dissertation, we have not been able to compare all existing variable selection procedures, but we focused on a number of frequently applied wrapper algorithms. Moreover, as several members of the exam committee noted, variable selection is not the only tool that can be used to reduce the variance of the estimates and the multicollinearity that appear in our examples of the predictive modeling of individual customer behavior. Shrinkage approaches, such as ridge regression (see, eg, Frank

and Friedman, 1993), the Lasso (Tibshirani, 1996), the (nonnegative) garrote (Breiman, 1995) and least angle regression (Efron et al, 2004) all serve a similar goal as variable selection, and, while they are currently noticeably less common in industry and academia, they have been introduced successfully into the targeted marketing literature (see, eg, Malthouse, 1999). I fully agree that a comparison of the variable selection techniques used in the last chapter of this dissertation with shrinkage approaches could provide ample material for an interesting future research agenda. Nevertheless, in the last chapter of this work, we believe that we have provided a testbed that can be sequentially expanded for comparing the usefulness of different predictive modeling techniques and methodologies.

Additionally, an increase of the number of settings may be a necessary next step in order to allow for a meta-analysis, through which one could detect in which cases a certain algorithm or modeling methodology should be prefered over competing techniques, and hence deliver a better insight into the question *why* certain methodology should be prefered in a given setting. Finally, different opportunities for research are proposed in each of the following chapters.

REFERENCES

Baesens B., Van Gestel T., Viaene S., Stepanova M., Suykens J., and Vanthienen J. (2003) Benchmarking State of the Art Classification Algorithms for Credit Scoring. Journal of the Operational Research Society **54** pp. 627-635

Breiman L. (1995) Better subset regression using the nonnegative garrote. Technometrics **37** 4 pp. 373-384

Breiman L. (2001) Statistical Modeling: The Two Cultures. Statistical Science **16** (3) pp. 199-231

Buckinx W. (2005) Using Predictive Modeling for Targeted Marketing in a Non-Contractual Retail Setting, PhD thesis, Ghent University

Cullinan G.J. (1977) Picking them by their batting averages recency-frequency-monetary method of controlling circulation. Manual release 2103, Direct Mail/Marketing Association, NY

Dasgupta C.G., Dispensa G.S. and Ghose S. (1994) Comparing the predictive performance of a neural network model with some traditional market response models. International Journal of Forecasting **10** (2) pp. 235-244

Davis R.H., Edelman D.B. and Gammerman A.J. (1992) Machine-Learning Algorithms for Credit-card Applications. IMA Journal of Mathematics Applied in Business and Industry **4** pp. 43-51

Efron B., Hastie T., Johnstone I. and Tibshirani R. (2004) Least angle regression. Annals of Statistics **32** (4) pp. 407-451

Elder J.F. and Pregibon D. (1996) A statistical perspective on knowledge discovery in databases. In Fayyad U.M., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R. (eds.) Advances in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI Press. pp. 83-113

Fayyad U.M., Piatetsky-Shapiro G. and Smyth P. (1996) From Data Mining to Knowledge Discovery: An Overview. In Fayyad U.M., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R. (eds.) Advances in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI Press pp. 1-34

Frank I. and Friedman J. (1993) A statistical view of some chemometrics regression tools. Technometrics **35** 109-148.

Garbarino E. and Johnson, M.S. (1999) The different roles of satisfaction, trust, and commitment in customer relationships. Journal of Marketing **63** (April) pp. 70-87

Grönroos C. (1997) From marketing mix to relationship marketing - towards a paradigm shift in marketing. Management Decision **35** (4) pp. 322-339

Hand D.J. (1999) Statistics and data mining: intersecting disciplines. SIGKDD Explorations **1** pp. 16-19

Hand D.J. (2004) Strength in diversity: the advance of data analysis. In Boulicaut J.-F., Esposito F., Giannotti F, and Pedreshchi D. (eds.) Proceedings of the 15th European Conference on Machine Learning. Pisa, Italy: Springer pp. 18-26

Hand D., Mannila H. and Smyth P. (2001) Principles of Data Mining. MIT Press, Cambridge, MA

Kotler P. (1991) Philip Kotler Explores the New Marketing Paradigm. Review, Marketing Science Institute Newsletter. Cambridge, MA (Spring) pp. 1, 4-5

Mägi A.W. (2003) Share of Wallet in Retailing: the Effects of Customer Satisfaction, Loyalty Cards and Shopper Characteristics. Journal of Retailing **79** pp. 97-106

Malthouse E.C. (1999) Ridge regression and direct marketing scoring models. Journal of Interactive Marketing **13** (4) pp. 10-23

Rawlings J.O. (1988) Applied regression analysis. Brooks/Cole Publishing Company, Pacific Grove, CA

Reichheld F.F. (1996) The Loyalty Effect. Harvard Business School Press, Cambridge, MA

Rigby D.K., Reichheld F.F. and Schefter P. (2002) Avoid the four perils of CRM. Harvard Business Review **80** (2) pp. 101-109

Rosenberg L.J. and Czepiel J.A. (1984) A marketing approach to customer retention. Journal of Consumer Marketing **1** pp. 45-51

Thomas L.C., Oliver R.W. and Hand D.J. (2005) A survey of the issues in consumer credit modelling research. Journal of the Operational Research Society **56** pp. 1006-1015

Tibshirani R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B **58** (1) pp. 267-288

Van den Poel D. and Larivière B. (2004) Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research **157** (1) pp. 196-217

Witten I.A. and Frank E. (2005) Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, CA

CHAPTER I

BAYESIAN NETWORK CLASSIFIERS FOR IDENTIFYING THE SLOPE OF THE CUSTOMER LIFECYCLE OF LONG-LIFE CUSTOMERS[1]

[1] This chapter is based on the following reference: Baesens B., Verstraeten G., Van den Poel D., Egmont-Petersen M., Van Kenhove P., Vanthienen J. (2004) Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers, European Journal of Operational Research, Vol 156 (2), 508-523.

CHAPTER I:

BAYESIAN NETWORK CLASSIFIERS FOR IDENTIFYING THE SLOPE OF THE CUSTOMER LIFECYCLE OF LONG-LIFE CUSTOMERS

ABSTRACT

Undoubtedly, Customer Relationship Management (CRM) has gained its importance through the statement that acquiring a new customer is several times more costly than retaining and selling additional products to existing customers. Consequently, marketing practitioners are currently often focusing on retaining customers for as long as possible. However, recent findings in relationship marketing literature have shown that large differences exist within the group of long-life customers in terms of spending and spending evolution. Therefore, this paper focuses on introducing a measure of a customer's future spending evolution that might improve relationship marketing decision making. In this study, from a marketing point of view, we focus on predicting whether a newly acquired customer will increase or decrease his/her future spending from initial purchase information. This is essentially a classification task. The main contribution of this study lies in comparing and evaluating several Bayesian network classifiers with statistical and other artificial intelligence techniques for the purpose of classifying customers in the binary classification problem at hand. Certain Bayesian network classifiers have been recently proposed in the artificial intelligence literature as probabilistic white box classifiers which give a clear insight into the relationships between the variables of the domain under study. We discuss and evaluate several types of Bayesian network classifiers and their corresponding structure learning algorithms. We contribute to

the literature by providing experimental evidence that: (1) Bayesian network classifiers offer an interesting and viable alternative for our customer lifecycle slope estimation problem; (2) the Markov Blanket concept allows for a natural form of attribute selection that was very effective for the application at hand; (3) the sign of the slope can be predicted with a powerful and parsimonious general, unrestricted Bayesian network classifier; (4) a set of three variables measuring the volume of initial purchases and the degree to which customers originally buy in different categories, are powerful predictors for estimating the sign of the slope, and might therefore provide desirable additional information for relationship marketing decision making.

1. INTRODUCTION

Undoubtedly, Customer Relationship Management (CRM) has gained its importance through the statement that acquiring a new customer is several times more costly than retaining and selling additional products to existing customers [2,20,46]. This simple rule-of-thumb has led to what many authors refer to as 'the paradigm shift in marketing' [4,25], implying that brand strategies are being replaced by customer strategies [3], and more and more voices rise to replace the traditional brand managers by customer (segment) managers [37,47]. Hence, it has become increasingly important to make informed marketing decisions on a customer level, and the customer loyalty of individual consumers has rapidly grown to become the focal point of relationship marketing (see, e.g. [22,31,40,41]).

In order to ensure the success of a CRM strategy, it is crucial that customers remain, at least to a certain extent, loyal to the company in case. However, recent research suggests large heterogeneity in terms of spending and spending evolution within the group of long-life customers [44]. Responding to this finding, in the following section of the paper, we elaborate upon the relevance of an accurate indication of a customer's future spending evolution for improving relationship marketing decision making for long-life customers. Consequently, we try to account for the heterogeneity within the group of long-life customers by adding information about estimated future spending evolutions.

In this study, we limit the focus to estimating whether newly acquired customers will increase or decrease their future spending. Whereas, to the best of our knowledge, no published study has attempted to forecast this variable, we argue in the following section that

the recently evolving literature around the loyalty issue has motivated us to do so. To this end, we will use and compare different recently developed classification techniques for optimally classifying the customers into the two relevant groups (i.e. customers with decreasing versus increasing spending). We hereby focus on techniques that besides yielding good classification accuracy also represent the marginal and conditional independence relations between the variables and how they jointly affect the classification decision.

In recent artificial intelligence literature, Bayesian networks have been suggested as probabilistic white box models that are able to capture even higher-order dependencies between sets of variables. These networks can then also be efficiently adopted for classification purposes. In this paper, we will evaluate and compare several Bayesian network classifiers for the purpose of classifying customers in the binary classification problem at hand. Using the Naive Bayes classifier as a point of origin, we will gradually remove the restrictions put on the network structure and investigate Tree Augmented Naive Bayes classifiers followed by completely unrestricted Bayesian network classifiers. Comparisons will be made with statistical and other artificial intelligence techniques. All classifiers will be evaluated by looking at their classification accuracy and the area under the receiver operating characteristic curve. The latter basically illustrates the behavior of a classifier without regard to class distribution or misclassification cost, so it effectively decouples classification performance from these factors. Furthermore, we will also look at the complexity of the trained classifiers because from a marketing viewpoint, parsimonious, yet accurate and self-explanatory models are to be preferred.

This paper is organized as follows. In the next section, we elaborate on the recent literature on relationship marketing that has provided motivation for investigating the predictability of the customer's spending evolution. To this end, we use Bayesian network classifiers which are discussed in Section 3. The design of the study, including both the data set description and the used performance criteria, are presented in Section 4. Section 5 presents the results of the experiments. Finally, Section 6 concludes the paper.

2. RELEVANCE OF THE ESTIMATION OF A CUSTOMER'S SPENDING EVOLUTION

Advocates of traditional relationship marketing attribute several advantages to loyal customers. Most importantly, these are expected to raise their spending (and contribution to

the company) over their relationship with the company in case [43]. In the most optimistic settings, they are said to generate new customers by their positive word-of-mouth [22], ensure diminished costs to serve [31], exhibit reduced consumer price sensitivities [42] and have a salutary impact on the company's employees [43]. Since, from a database-driven approach, customer tenure (i.e. the length of a customer's relationship with a company) has often been used to approximate the loyalty construct [22,44,45], relationship marketing thrives on the idea that raising the length of the customer-company relationship is the main lever for a company's financial success [43].

Nevertheless, in their recent article, Reinartz and Kumar [45] report a series of studies across industries that challenges most claims of the loyalty advocates. In these studies, they have found no evidence to suggest that long-life customers with steady purchase behavior are necessarily cheaper to serve, less price sensitive, or more effective in bringing new business to the company, such as through word-of-mouth referrals. Additionally, in a previous article, Reinartz and Kumar [44] showed that the contributions of long-life customers were generally declining, although the analysis of this issue was not the focus of their discussion. Finally, the authors pointed out that, at least for a non-contractual setting, short-life but high-revenue customers accounted for a sizeable amount of profits for the mail-order company in case [44].

In the article mentioned above, Reinartz and Kumar clearly illustrate the pitfalls involved with spending a large slice of the marketing budget on customers that have been good customers in the past over a short period of time, yet tend to show a decreasing spending pattern (i.e. customers that have been labelled 'butterflies') [45]. In the example of a mail-order setting, it is generally known that repurchase behavior can – and has – effectively been modeled by using an (often linear) combination of RFM variables, representing the recency of a customer's last purchase, the average frequency of the customer's purchases and the average monetary value spent on the customer's purchase occasions [12,50]. Hence, the group of customers called 'butterflies', being customers with a high historical monetary value, will tend to be over selected for mailing campaigns [45]. An estimation of the future slope of the customer lifecycle (i.e. a customer's spending evolution) would then likely be able to deliver the required insights to the decision-making process and the understanding of the relationship between the slope and other variables, such as customer spending, might generate rich qualitative information for marketers. For instance, for this group of customers,

the company might decide to attempt to improve its return on (direct) marketing investments by shifting its focus from long-term investments to investments or promotions on which a short-term return is possible. Alternatively, the company might even consider abandoning investments in these customers altogether. Thus, in this customer-based view, the a-priori knowledge of the slope of the customer lifecycle would be useful information.

In this research study we limit our attention in terms of marketing contribution to proving that it is possible to predict the slope of the customer lifecycle of long-life customers. Accordingly, due to the limitations that are extensively documented in Section 7 of this paper, it is not within the scope of this paper to devise, implement and test an optimal marketing strategy for a specific company in case, nor for an array of companies in industries with different characteristics. In this attempt, we will compare different techniques for the estimation problem, which can in its essential form be transformed into a binary classification problem: 'Will newly acquired customers increase or decrease their spending after their first purchase experiences?'

In the marketing literature, binary classification problems have typically been tackled by using traditional statistical methods (e.g. discriminant analysis and logistic regression [2,50]), nonparametric statistical models (e.g. k-nearest neighbour [50] and decision trees [49,50]) and neural networks [2,50]. In this paper, we will adopt Bayesian network classifiers which have been recently introduced in the artificial intelligence literature. This is motivated by the fact that Bayesian network classifiers are probabilistic white-box models which facilitate a clear insight into the underlying dependencies pertaining to the domain under study. They are based on solid probabilistic reasoning and offer a great potential for knowledge discovery in data in a marketing context. Unfortunately, despite their attractive properties, their application for business decision making and marketing purposes is still limited. In the following section, we will elaborate on the basic concepts of Bayesian network classifiers and discuss some recently suggested structure learning algorithms.

3. BAYESIAN NETWORKS FOR CLASSIFICATION

A Bayesian network (BN) represents a joint probability distribution over a set of discrete, stochastic variables. It is to be considered as a probabilistic white-box model consisting of a qualitative part specifying the conditional (in)dependencies between the variables and a

quantitative part specifying the conditional probabilities of the data set variables [36]. Formally, a Bayesian network consists of two parts $B = \langle G, \theta \rangle$. The first part G is a directed acyclic graph consisting of nodes and arcs. The nodes are the variables $X_1,\ldots,X_n$ in the data set whereas the arcs indicate direct dependencies between the variables. The graph G then encodes the independence relationships in the domain under investigation. The second part of the network, θ, represents the conditional probability distributions. It contains a parameter $\theta_{x_i|\Pi_{x_i}} = P_B(x_i|\Pi_{x_i})$ for each possible value x_i of X_i, given each combination of the direct parent variables of X_i, Π_{x_i} of Π_{X_i}, where Π_{X_i} denotes the set of direct parents of X_i in G. The network B then represents the following joint probability distribution:

$$P_B(X_1,\ldots,X_n) = \prod_{i=1}^{n} P_B(X_i|\Pi_{X_i}) = \prod_{i=1}^{n} \theta_{X_i|\Pi_{X_i}} \quad . \tag{1}$$

The first task when learning a Bayesian network is to find the structure G of the network. Once we know the network structure G, the parameters θ need to be estimated. In general, these two estimation tasks are performed separately. In this paper, we will use the empirical frequencies from the data D to estimate these parameters: [2]

$$\theta_{x_i|\Pi_{x_i}} = \hat{P}_D(x_i|\Pi_{x_i}). \tag{2}$$

It can be shown that these estimates maximise the log likelihood of the network B given the data D [21]. Note that these estimates might be further improved by a smoothing operation [21].

A Bayesian network is essentially a statistical model that makes it feasible to compute the (joint) posterior probability distribution of any subset of unobserved stochastic variables, given that the variables in the complementary subset are observed. This functionality makes it possible to use a Bayesian network as a statistical classifier by applying the winner-takes-all rule to the posterior probability distribution for the (unobserved) class node [15]. The underlying assumption behind the winner-takes-all rule is that all gains and losses are equal

[2] Note that we hereby assume that the data set is complete, i.e., no missing values.

(for a discussion of this aspect see, e.g., [15]). In what follows, we will discuss several structure learning algorithms for developing Bayesian network classifiers.

3.1 The Naive Bayes classifier

A simple classifier, which in practice often performs surprisingly well, is the Naive Bayes classifier [15,30,33]. This classifier basically learns the class-conditional probabilities $P(X_i = x_i | C = c_l)$ of each variable X_i given the class label c_l. A new test case ($X_i = x_i, \ldots, X_n = x_n$) is then classified by using Bayes' rule to compute the posterior probability of each class c_l given the vector of observed variable values:

$$P(C = c_l | X_1 = x_1, \ldots, X_n = x_n) = \frac{P(C = c_l) P(X_1 = x_1, \ldots, X_n = x_n | C = c_l)}{P(X_1 = x_1, \ldots, X_n = x_n)}. \quad (3)$$

The simplifying assumption behind the Naive Bayes classifier then assumes that the variables are conditionally independent given the class label. Hence,

$$P(X_1 = x_1, \ldots, X_n = x_n | C = c_l) = \prod_{i=1}^{n} P(X_i = x_i | C = c_l). \quad (4)$$

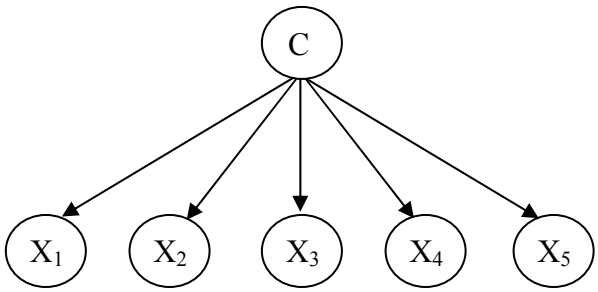


Fig. 1. The Naive Bayes classifier

This assumption simplifies the estimation of the class-conditional probabilities from the training data. Notice that one does not estimate the denominator in expression (3) since it is independent of the class. Instead, one normalises the nominator term $P(C = c_l) P(X_1 = x_1, \ldots, X_n = x_n | C = c_l)$ to 1 over all classes. Naive Bayes classifiers are easy to construct since the structure is given apriori and no structure learning phase is required.

The probabilities $P(X_i = x_i | C = c_l)$ are estimated by using the frequency counts for the discrete variables and a normal or kernel density based method for continuous variables [30]. Fig. 1 provides a graphical representation of a Naive Bayes classifier.

3.2 Tree Augmented Naive Bayes classifiers

In [21] Tree Augmented Naive Bayes classifiers (TANs) were presented as an extension of the Naive Bayes classifier. TANs relax the independence assumption by allowing arcs between the variables. An arc from variable X_i to X_j then implies that the impact of X_i on the class variable also depends on the value of X_j. An example of a TAN is presented in Fig. 2. In a TAN network the class variable has no parents and each variable has as parents the class variable and at most one other variable. The variables are thus only allowed to form a tree structure. In [21], a procedure was presented to learn the optional arrows in the structure that forms a TAN network. This procedure is based on an earlier algorithm suggested by Chow and Liu (CL) [11]. The procedure consists of the following five steps.

1. Compute the conditional mutual information given the class variable C, $I(X_i; X_j | C)$, between each pair of variables, $i \neq j$. $I(X_i; X_j | C)$ is defined as follows:

$$I(X_i; X_j | C) = \sum_{x_i, x_j, c_l} P(X_i = x_i, X_j = x_j, C = c_l) \times \log \frac{P(X_i = x_i, X_j = x_j, C = c_l)}{P(X_i = x_i | C = c_l) P(X_j = x_j | C = c_l)}. \quad (5)$$

   This function is an approximation of the information that X_j provides about X_i (and vice versa) when the value of C is known.
2. Build a complete undirected graph in which the nodes are the variables. Assign to each arc connecting X_i to X_j the weight $I(X_i; X_j | C)$.
3. Build a maximum weighted spanning tree.
4. Transform the resulting undirected tree to a directed one by choosing a root variable and setting the direction of all arcs to be outward from it.
5. Add the classification node C and draw an arc from C to each X_i.

We used Kruskal's algorithm in step 3 to construct the maximum weighted spanning tree [32]. In [21], it was proven that the above procedure builds TANs that maximise the log likelihood of the network given the training data and has time complexity $O(n^2 \cdot N)$ with n the number of variables and N the number of data points. Experimental results indicated that TANs outperform Naive Bayes with the same computational complexity and robustness [21].

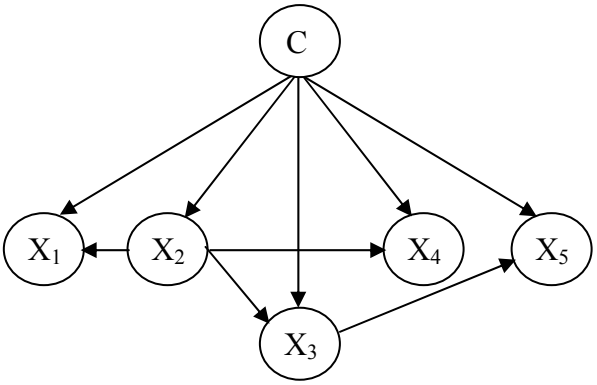


Fig. 2. The Tree Augmented Naive Bayes classifier

3.3 General Bayesian Network classifiers

Many algorithms have been proposed that can learn the structure of a General Bayesian Network (GBN) from a set of (complete) data [5,29]. Some algorithms impose restrictions onto the direction of the arcs that connect the nodes whereas other algorithms omit such restrictions. In this paper, we use the learning algorithm of Cheng et al. [7,9], which assumes an a priori ordering of the variables. Before we discuss the different steps of this algorithm, we first elaborate on the concept of *d*-separation because this plays a pivotal role in the structure learning algorithm.

Let X, Y and Z be mutually disjoint sets of nodes in a directed acyclic graph G. The set Y is said to *d*-separate the sets X and Z in G if for every node $X_i \in X$ and every node $X_j \in Z$, every chain (of any directionality) from X_i to X_j in G is blocked by Y [51]. We say that a chain s is blocked by a set of nodes Y if s contains three consecutive nodes X_1, X_2, X_3, for which one of the following conditions holds [51]:

1. arcs $X_1 \leftarrow X_2$ and $X_2 \rightarrow X_3$ are on the chain s, and $X_2 \in Y$;
2. arcs $X_1 \rightarrow X_2$ and $X_2 \rightarrow X_3$ or $X_1 \leftarrow X_2$ and $X_2 \leftarrow X_3$ are on the chain s, and $X_2 \in Y$;

3. arcs $X_1 \rightarrow X_2$ and $X_2 \leftarrow X_3$ are on the chain s and X_2 and the descendants of X_2 are not in Y.

It can be shown that if sets of variables X and Z are d-separated by Y in a directed acyclic graph G, then X is independent of Z conditional on Y in every distribution compatible with G [24,52]. It is precisely this property that will be exploited in the algorithm of Cheng to learn the Bayesian network structure.

The algorithm consists of four phases. In a first phase, a draft of the network structure is made based on the mutual information between each pair of nodes. The second and third phase then add and remove arcs based on the concept of d-separation and conditional independence tests. Finally, in the fourth phase, the Bayesian network is pruned and its parameters are estimated.

The algorithm proceeds as follows [7,9].

Phase 1: Drafting

1. Initiate a graph $G(X,A)$ where $X = \{X_1,X_2,\ldots,X_n,C\}$ and $A = \{\}$. Initiate two empty ordered sets S and R.
2. Compute the (non-parametric) mutual information $I(X_i; X_j)$ between each pair of variables where $X_i, X_j \in X$, $i \neq j$. $I(X_i; X_j)$ is defined as follows:

$$I(X_i; X_j) = \sum_{x_i, x_j} P(X_i = x_i, X_j = x_j) \times \log \frac{P(X_i = x_i, X_j = x_j)}{P(X_i = x_i) P(X_j = x_j)}. \qquad (6)$$

   The mutual information $I(X_i; X_j)$ is the amount of information gained about X_i when X_j is known, and vice versa $(I(X_i; X_j) = I(X_j; X_i))$. Hence, $I(X_i; X_j)$=0 if and only if X_i and X_j are independent.
3. Sort all pairs of nodes where $I(X_i; X_j)$ is greater than ε from large to small and put them into an ordered set S. In our experiments, we set ε = 0.008 which is an appropriate value for large data sets [9].

4. Add arcs to *A* according to the first two pairs of nodes in *S* and remove them from *S*. The direction of the arcs is decided by the apriori node ordering.
5. Get the first pair of nodes remained in *S* and remove it from *S*. If there is no open path between the two nodes, add the corresponding arc to *A*. Otherwise, add the pair of nodes to the end of an ordered set *R*. Note that an open path is a chain with no collider nodes whereby a collider node is a node having two incoming arcs.
6. Repeat step 5 until *S* is empty.

Phase 2: Thickening

7. Get the first pair of nodes in *R* and remove it from *R*.
8. Find a cut-set that can *d*-separate these two nodes in the current network. Use a conditional independence test (see Eq. (5)) to see if these two nodes are conditionally independent given the cut-set and using a threshold value of 0.008. If so, go to the next step, otherwise, connect the pair of nodes by an arc.
9. Repeat step 7 until *R* is empty.

Phase 3: Thinning

10. For each arc in *A*, if there are other paths besides this arc between the two nodes, remove this arc from *A* temporarily and find a cut-set that can d-separate the two nodes in the current network. Use a conditional independence test to see if the two nodes are conditionally independent given the cut-set and again using a threshold value of 0.008. If so, remove the arc permanently, otherwise add the arc back to the network.

Phase 4: Prune and learn the parameters of the Bayesian network classifier

11. Find the Markov Blanket of the classification node. The Markov Blanket of a node X_i consists of the union of X_i 's parents, X_i 's children and the parents of X_i 's children [36].
12. Delete all the nodes that are outside the Markov Blanket.
13. Learn the parameters of the conditional probability tables and output the Bayesian network classifier.

Note that in steps 8 and 10, it is important to find cut-sets that are as small as possible in order to avoid conditional independence tests with large condition sets. In [1], a correct algorithm is presented to find minimum cut-sets between two nodes. In this paper, we will use the heuristic algorithm suggested by Cheng et al. [7].

It can be shown that when the values of the variables in the Markov Blanket of the classification node are observed, the posterior probability distribution of the classification node is independent of all other variables (nodes) not in the Markov Blanket [34]. Hence, in step 12, all variables outside the Markov Blanket can be safely deleted because they will have no impact on the classification node and thus will not affect the classification accuracy. In this way, the Markov Blanket results in a natural form of variable selection.

Note that this algorithm requires O(N^2) mutual information tests and is linear in the number of cases N. An extension has been presented in [8] in case no node ordering is given. In this paper, we will simply treat the classification node as the first node and order the other nodes based on their correlation with the classification node from large to small.

3.4 Multinet Bayesian Network classifiers

Both TANs and GBNs assume that the relations between the variables are the same for all classes. A multinet Bayesian network allows for more flexibility and is composed of a separate, local network for each class and a prior probability distribution of the class node [10,21,23,28]. Thus, for each value c_i of the classification node C a Bayesian network structure B_i is learned. The multinet M then defines the following joint probability distribution:

$$P_M(C, X_1,..., X_n) = P_C(C) \cdot P_{B_i}(X_1,..., X_n). \quad (7)$$

A new instance is then assigned to the class that maximises the posterior probability $P_M(C|X_1,..., X_n)$ conform the winner-takes-all rule. Since we have

$$P_M\left(C\middle|X_1,...,X_n\right)=\frac{P_M\left(C,X_1,...,X_n\right)}{P_M\left(X_1,...,X_n\right)}, \tag{8}$$

and the denominator is the same for all classes, we can assign the instance to the class that maximises the value of Eq. (7). The term $P_C(C)$ may then be estimated by the empirical frequency of the class variable in the training set $\hat{P}_D(C)$. Note that for multinet classifiers the number of parameters that need to be estimated per training instance inevitably increases. As the parameters are estimated from a limited number of instances, learning a separate multinet structure per class instead of one overall structure results in more unreliable parameter estimates and, hence, a higher probability of overgeneralization. This effect is closely related to the so-called peaking phenomenon, for a discussion see, e.g. [53].

In this paper, we consider both CL multinets and GBN multinets. CL multinets are multinets which are built using the procedure of Chow and Liu [11]. This is essentially the same procedure as the one outlined in Section 3.2 with the exception that step 5 is now omitted and in step 1 the conditional mutual information is replaced by the mutual information (see Eq. (6)). This procedure is then executed separately for each value c_i of the class node C using only the training data D_i whereby D_i contains all instances of D for which $C= c_i$. The resulting multinet then consists of an ensemble of tree structured Bayesian networks. The GBN multinets are trained using the approach of Cheng discussed in Section 3.3 with the exception that the classification node is now omitted in the structure learning phase. Again, the algorithm is executed for each class on the corresponding training data.

4. DESIGN OF THE STUDY

4.1 Data set

We conducted our research on UPC scanner data of a large Belgian DIY (Do-It-Yourself) retail chain. The data we used for our models were all gathered by the customer loyalty cards, which have been in use since January 1995. Due to some restrictions (cf. infra), we were able to use four complete years of information.

Since we are interested in examining the behavior of long-life customers, we imposed three conditions on the data: firstly, we only used customers who started purchasing before

February 1997. Secondly, to ensure the data was not left-censored (i.e. to ensure the customers in our database really started their relationship with the company at the time of our first observation), we only used information of customers who had not purchased before. We thus used the first two years of information in the customer database only to check that the customers in our sample were new customers. Thirdly, using a database containing eight six-month periods of information for all customers of the company, we have selected all customers who purchased in five or more periods. Hence, we arrived at a database containing an approximate sample of the company's long-life customers. In order to assess the quality of our models, we have randomly divided the database into 2 parts. While 2/3 of the observations were used for learning the classifiers, the remaining 1/3 was used as a test set for estimating the generalization behavior of the classifiers. Table 1 displays the characteristics of our data set.

Table 1. Data set characteristics

Data set size	3827 observations
Training set size	2551 observations
Test set size	1276 observations
Number of attributes	15

Table 2. Variables used in the study

1	Total contribution	TotCont
	Total revenues	TotRev
	Total number of articles bought	NumbArt
	Total number of visits to the store (tickets)	NumbTick
2	Number of different categories purchased	DiffCat
	Number of different products purchased	DiffProd
	Maximum percentage of products bought in one product family	MaxPerc
3	Mean margin of articles purchased	MeanMarg
	Mean price of articles purchased	MeanPrice
	Maximum price paid for an article	MaxPrice
	Total value of received discounts / total revenues	PercDisc
	Articles bought in discount / total amount of articles bought	ArtDisc
4	Slope of the 'customer lifecycle' during the first 6 months	Lifec6m
	Contribution in the 6th month	LastCont
	Date the maximum price was paid	DateMaxPrice

By performing a linear regression model on the historical contributions of each customer, we were able to capture the slope of the lifecycle of each individual customer. This slope, after

being discretised into positive or negative to represent increasing or decreasing spending, was henceforth used as the dependent variable in the study (SlopeSign). It is interesting to note that the finding of Reinartz and Kumar that the slope of long-life customers was generally decreasing [44] was validated in our study by the fact that only 28% of those customers in the database exhibited a positive slope. In this case, we have used a set of 15 continuous variables computed on the first 6 months of information, in order to predict the sign of the evolution of the customer's contribution (i.e. the customer lifecycle) for the remaining 42 months of the relationship. While the variables computed are presented in Table 2, the time schedule is given in Fig. 3.

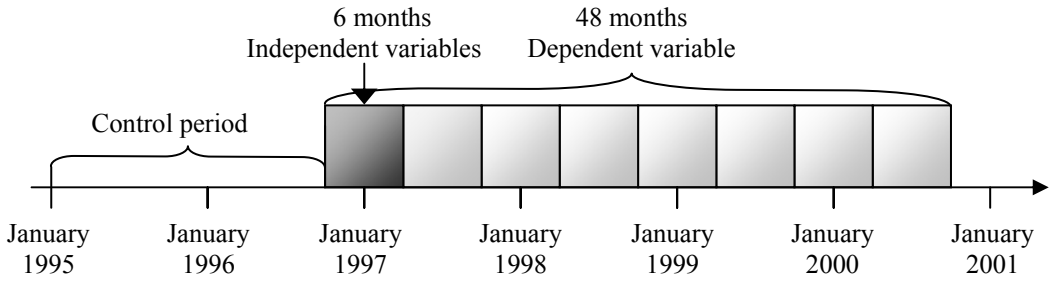


Fig. 3. Time schedule of our empirical study.

The independent variables can be divided into four major logical groups. A first group of variables is constructed to measure the volume of the purchases the subject made during his or her first 6 months as a customer. These contain TotCont, TotRev, NumbArt and NumbTick. Note that the variable TotCont represents the intercept of the customer lifecycle. It is merely the first of the eight data points forming the customer lifecycle. While this first set of attributes can be regarded as the "depth" of the customer purchases, the second group of variables contains the variables that measure the "broadness" of the purchases. These are DiffCat, DiffProd and MaxPerc. The latter variable contains the percentage of products bought in the product category in which the customer has bought most of his or her products. In this way, it can be seen as a skewness indicator, a large indicator meaning that the customer only buys a certain category of products from the company. A third group of variables captures the 'bargaining tendency' and 'price sensitivity' of the customer. The relevant variables here are PercDisc, ArtDisc, MeanMarg, MeanPrice and MaxPrice. Finally, three measures are introduced to value evolutions within the first six months. These are Lifec6m, LastCont and DateMaxPrice.

Table 3. Discretisation of the attributes

Attribute	Values	Encoding
TotCont	1,2,3,4	]-∞;241.74],]241.74;817.92],]817.92;3158.09],]3158.09;∞]
TotRev	1,2,3,4	]-∞;679.53],]679.53;2481.82],]2481.82;7410.12],]7410.12;∞]
NumbArt	1,2,3,4	]-∞;4],]4;13],]13;38],]38;∞]
NumbTick	1,2,3	]-∞;2],]2;5],]5;∞]
DiffCat	1,2,3,4	]-∞;2],]2;6],]6;13],]13;∞]
DiffProd	1,2,3	]-∞;4],]4;11];]11;∞]
MaxPerc	1,2,3,4	]-∞;0.49],]0.49;0.5],]0.5;0.98],]0.98;∞]
MeanMarg	1,2	]-∞;0.53],]0.53;∞]
MeanPrice	1,2	]-∞;118.16],]118.16;∞]
MaxPrice	1,2,3,4	]-∞;165],]165;549],]549;1095],]1095;∞]
PercDisc	1,2	]-∞;0.17],]0.17;∞]
ArtDisc	1,2,3	]-∞;0],]0;0.33],]0.33;∞]
Lifec6m	1,2,3	]-∞;-72.24],]-72.24;87.61],]87.61;∞]
LastCont	1,2,3	]-∞;621.85],]621.85;2010.85],]2010.85;∞]
DateMaxPrice	1,2	]-∞;13544],]13544;∞]

In order to train the Bayesian network classifiers, we discretised all variables by using the discretisation algorithm of Fayyad and Irani with the default options [19]. This algorithm uses an information entropy minimisation heuristic to discretise the range of a continuous-valued attribute into multiple intervals. This discretisation procedure was performed using the Java Weka workbench.[3] Table 3 depicts how the attributes in our data set were discretised into intervals.

4.2 Performance criteria for classification

The performance of all trained classifiers will be quantified using both the classification accuracy and the area under the receiver operating characteristic curve (AUROC). The classification accuracy is undoubtedly the most commonly used measure of performance of a classifier. It simply measures the percentage of correctly classified (PCC) observations. However, it tacitly assumes equal misclassification costs and balanced class distributions [38]. The receiver operating characteristic curve (ROC) is a 2-dimensional graphical illustration of the sensitivity ('true alarms') on the Y-axis versus 1-specificity on the X-axis ('false alarms') for various values of the classification threshold [16,48]. It basically illustrates the behaviour of a classifier without regard to class distribution or misclassification cost. The AUROC then provides a simple figure-of-merit for the performance of the constructed classifier. An intuitive interpretation of the AUROC is that it

[3] http://www.cs.waikato.ac.nz/ml/weka/

provides an estimate of the probability that a randomly chosen instance of class 1 is correctly rated (or ranked) higher than a randomly selected instance of class 0 [26].

We will use McNemar's test to compare the PCCs of different classifiers [17]. This chi-squared test is based upon contingency table analysis to detect statistically significant performance differences between classifiers. In [14], it was shown that this test has acceptable Type I error which is the probability of incorrectly detecting a difference when no difference exists. While Hanley and McNeil described a method for comparing ROC curves derived from the same sample [27], De Long, De Long and Clarke-Pearson [13] developed a nonparametric chi-squared test by using the theory on generalised *U*-statistics and the method of structural components to estimate the covariance matrix of the AUROC. Hence, we will use the latter test to detect statistically significant AUROC differences between classifiers.

5. RESULTS

We compared and contrasted the performance of the Naive Bayes, TAN, CL multinet, GBN, GBN multinet, C4.5, C4.5rules, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) classifiers on our marketing data set. We included the decision tree induction algorithm C4.5 and its rules variant, C4.5rules, because they are also white-box classifiers giving besides a classification decision also a clear explanation why the particular classification is being made [39]. LDA and QDA were included because they are well-known benchmark statistical classifiers. To train the Naive Bayes, TAN, and CL multinet classifiers, we used the Matlab toolbox of Kevin Murphy [35]. For the GBN and

Table 4. Classification accuracy of the Bayesian network classifiers versus C4.5 and discriminant analysis

	Training set	Test set
Naive Bayes	71.0	72.5
TAN	74.9	**74.0**
CL multinet	74.2	72.3
GBN	75.3	**75.0**
GBN multinet	70.6	72.3
C4.5	76.7	**74.1**
C4.5rules	77.8	73.3
LDA	75.5	**74.1**
QDA	72.9	72.7

Table 5. Area under the Receiver Operating curve of the Bayesian network classificiers versus C4.5 and discriminant analysis

	Training set	Test set
Naive Bayes	75.9	**74.3**
TAN	77.8	73.6
CL multinet	77.0	72.6
GBN	77.5	**74.7**
GBN multinet	76.6	74.0
C4.5	76.5	73.8
C4.5rules	77.0	70.9
LDA	77.7	**75.9**
QDA	77.0	72.7

GBN multinet classifiers, we used the PowerPredictor software of Cheng [6]. Table 4 depicts the classification accuracy of all classifiers on both the training and test set. The best test set performance is in bold face and underlined and those not statistically different from it according to McNemar's test (using a significance level of 5%) are in bold face. The GBN classifier achieved the highest classification accuracy on the test set. The classification accuracy of the TAN, C4.5 and LDA classifier was not statistically different from it. Table 5 depicts the area under the receiver operating characteristic curve of all classifiers and has the same setup as Table 4. Note that for the Bayesian network classifiers, the LDA and QDA classifier, the calculation of the AUROC values poses no problems since each of these classifiers yields class probabilities. For C4.5, we use the confidence at the leaves as the class probability. For C4.5rules, we used the confidence of the first rule of the ordered C4.5rules rules set (ordered by class and then by confidence) that matches the instance as its class probability. In [18], it was shown that this is a feasible strategy for computing the AUROC of C4.5rules. Table 5 clearly indicates that the LDA classifier gave the best AUROC performance. However, there is no significant difference with the AUROC performance of the GBN and Naive Bayes classifier according to the test of De Long, De Long, and Clarke-Pearson and again using a significance level of 5%. Observe from Tables 4 and 5 that both multinet classifiers, QDA and C4.5rules never achieved good performance in terms of PCC and AUROC. Note that for all Bayesian network classifiers, we also investigated the impact of smoothing the parameter estimates. However, no significant performance increase in terms of either the PCC or AUROC values were found with parameter smoothing.

Table 6. Complexity of the Bayesian network classifiers and C4.5

Naive Bayes	16 nodes and 15 arcs
TAN	16 nodes and 29 arcs
CL multinet	Net 1: 15 nodes and 14 arcs
	Net 2: 15 nodes and 14 arcs
GBN	4 nodes and 6 arcs
GBN multinet	Net 1: 3 nodes and 2 arcs
	Net 2: 3 nodes and 2 arcs
C4.5	13 internal nodes
	32 leave nodes
C4.5rules	18 rules

Besides looking at the classification performance, we also investigated the complexity of the generated classification models because from a marketing viewpoint, easy to understand, parsimonious models are to be preferred. Table 6 presents the complexity of the generated Bayesian network and C4.5(rules) classifiers. We did not include LDA and QDA because they are basically mathematical models which give a rather limited insight into the relationships and patterns present in the domain under study. The Naive Bayes and TAN network classifiers did not prune any attributes because all attributes remained in the Markov Blanket of the classification node. The TAN added 14 arcs to the Naive Bayes classifier which resulted in a performance increase in terms of PCC (from 72.5 to 74.0) but a performance decrease in terms of AUROC (from 74.3 to 73.6). Hence, the effect of the added complexity was rather marginal in our case. Although the GBN multinet classifier seems attractive because of its simple structure, its performance according to Tables 4 and 5 was rather bad. Also the CL multinet classifier gave bad performance and has on top a complex structure. The tree induced by C4.5 is not easy to handle and interpret because of its large number of internal and leave nodes. Moreover, the C4.5 tree was able to prune only 2 of the 15 attributes. The rule set inferred by C4.5rules contains 18 rules. This might seem interesting but when considering Tables 4 and 5 the performance of C4.5rules in terms of both PCC and AUROC was rather bad. Note that while the C4.5 tree pruned 2 attributes, the C4.5rules rules set still contained all attributes. This can be explained by the fact that C4.5rules starts generating and pruning the rules from the unpruned C4.5 tree. The GBN classifier was able to prune 12 attributes, leaving only 3 attributes in the model. Only 6 arcs where necessary to efficiently model the dependencies between the attributes and the classification node. Furthermore, it gave also a very good performance in terms of PCC and AUROC on the test set. The structure of the GBN classifier is depicted in Fig. 4.

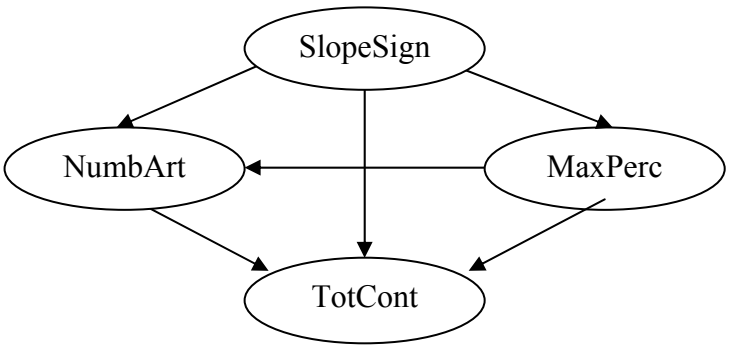


Fig. 4. Unrestricted Bayesian network constructed for marketing case.

This figure clearly illustrates that it is a compact, parsimonious and yet powerful model for decision making. By using only three variables compiled from purchase records of the first 6 months of the customer lifecycle, we have provided evidence that, in our DIY case, the SlopeSign of a lifecycle of 48 months can be predicted with a classification accuracy of 75%. The total contribution of the client (TotCont), the total number of articles bought (NumbArt) and the maximum percentage of products bought in one product family (MaxPerc) proved to be very powerful predictors for the sign of the customer lifecycle slope when using GBN classifiers. While the first two variables present a measure of the volume of the purchases made (the purchase "depth"), the latter variable is an estimator of the variety of product families bought (the purchase "broadness").

The knowledge that these variables are intensely related to the slope's evolution can be useful for marketing decision makers. In this Belgian DIY retail setting, the initial monetary amount spent at the company (TotCont) and the initial number of articles purchased (NumbArt) were found to be negatively related to the SlopeSign, whereas the maximum percentage of products purchased in one category (MaxPerc) was found to be positively related to the SlopeSign. This implies that customers that tend to increase their spending over their lifetime with the company initially spend less money on a lower number of articles, purchasing from a smaller set of product categories. Alternatively, customers spending a lot of money initially on a lot of articles and who purchase products across a lot of different categories tend to decrease their spending in the future. This information may prove valuable – for the company in this case – as a starting point for investigating why high-spending customers generally decrease their spending over time.

To conclude, we can state that Bayesian network classifiers are performing well in predicting the future customer evolution and are able to contribute to an increased understanding of the relationship between the investigated variable and the most relevant explanatory variables. Hence, we have reached our goal to illustrate that Bayesian network classifiers can be considered to be a useful tool in the toolbox of marketing analysts in this application of identifying the slope of the customer lifecycle of long-life customers.

6. CONCLUSIONS

In the theoretical part of this paper, we have argued that long-life/loyal customers have been regularly regarded as a homogeneous group of the most profitable customers of a company. Building on more recent findings, in this study, we have tried to acknowledge the heterogeneity in the group of long-life customers by dividing the group into two subparts, essentially consisting of customers increasing versus decreasing their spending over their relationship with the company in case. Hence, it was the goal of this study to predict the sign of the slope – being the output of the estimation of a linear customer lifecycle – at the individual customer level using Bayesian network classifiers based on information from initial purchase occasions.

Bayesian network classifiers have been recently proposed in the artificial intelligence literature as probabilistic white box models which allow to give a clear insight into the relationships between the variables of the domain under study. Starting from the Naive Bayes classifier, we gradually removed the restrictions put on the network structure and investigated Tree Augmented Naïve Bayes classifiers and general Bayesian network classifiers. The latter were learnt using the algorithm of Cheng et al. We compared the classification accuracy and the area under the receiver operating characteristic curve of all Bayesian network classifiers with discriminant analysis and the widely used C4.5 and C4.5rules algorithms. It was shown that general, unrestricted Bayesian network classifiers have a good performance in terms of both measures. Furthermore, using the Markov Blanket concept allowed us to prune a lot of attributes resulting in a compact, parsimonious, yet powerful Bayesian network classifier for marketing decision making.

In summary, we contribute to the literature by providing experimental evidence that: (1) Bayesian network classifiers offer an interesting and viable alternative for our customer

lifecycle slope estimation problem; (2) the Markov Blanket concept allows for a natural form of attribute selection that was very effective for the case at hand; (3) the sign of the slope can be predicted with a powerful and parsimonious general Bayesian network classifier; (4) a set of three variables measuring the volume of initial purchases and the degree to which customers originally buy in different categories, are a powerful set of predictors for estimating the sign of the slope.

7. PRACTICAL IMPLICATIONS AND ISSUES FOR FURTHER RESEARCH

While it has been the focus of this paper to demonstrate (i) the predictability of the sign of the slope and (ii) the performance of several Bayesian network classifiers versus statistical and other artificial intelligence techniques, here, we elaborate on possible applications of the knowledge of the sign of the slope for relationship marketing decision making. Note, however, that the success of any application requires that such future behavior should be estimable to a reasonable degree, implying that the misclassification costs and estimated error rates should behave in such a way that is beneficial to execute customized treatment (see [54] for an elaborated discussion). Nevertheless, in this section we propose a number of possible applications. Firstly, the sign of the slope might prove to be a useful indicator in the decision upon the type or strength of the marketing investment that can be used vis-à-vis a certain consumer. For example, a company organizing a membership club, with special service offerings, special promotions, etc. might only want to deliver these benefits to consumers that are worthy of such a large investment. In such a case, knowing that certain consumers will decrease their spending might be important for improving the return on the relationship-marketing investment. Alternatively, a company might have two marketing incentives of unequal cost (e.g. a special promotion versus a small gift). Also in this case, it could be useful to assess the future spending of a customer in order to allocate the desired incentive to each customer. Secondly, the estimations may be used in an aggregated way, as a monitor of e.g. customer-acquisition policies. In this way, the percentage of customers that are expected to raise their spending in the future can be compared for different acquisition strategies and campaigns in order to select those target markets with higher potential for establishing enduring relationships. An additional benefit is derived from the fact that it was possible to predict the evolutions very early in the relationships, so acquisition campaigns can be evaluated in a time-effective way. Thirdly, the estimations might be used as a dimension for designing an a-priori segmentation scheme for a company's customer base.

Hence, it might be feasible to delineate a more customized customer strategy per segment. Two possible applications are summarized in Fig. 5. In the first segmentation scheme (a), the sign of the slope is used together with the tenure of customers in order to decide upon the relevant marketing message content and size. Whereas short-life customers only merit investments that can be regained during their limited relationship with the company, long-life customers might effectively be reached through more expensive marketing programs. For customers who are expected to increase the relationship with the company (who are likely to be more satisfied with the company in case) it might be beneficial to offer additional products according to their detected needs (detected e.g. through a cross-selling analysis) or extra value (e.g. through the membership to a club). Alternatively, customers who are expected to decrease their spending might be appropriately managed with a retention program (e.g. focused on complaint detection and complaint handling). In the second segmentation scheme (b), the intercept and the slope of the customer lifecycle are used to delineate the segmentation. Also in this case, argumentation could be found to use specific marketing strategies to target the segments, where it could be argued that not all segments merit targeting (e.g. the segment of customers that starts as low spending customers and are expected to decrease their spending even further).

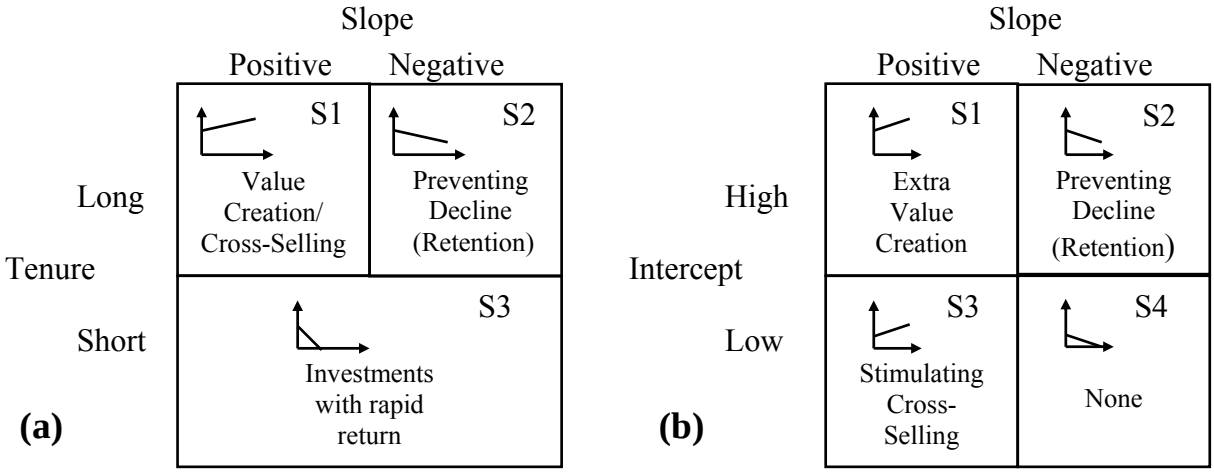


Fig. 5: Summary of possible a-priori segmentation schemes

The desired outcome of this line of research could consist in suggesting an optimal CRM strategy to different segments. However, in order to test the optimality of the proposed strategies, one would have to design and implement an experiment allocating strategies randomly to customers. It can be argued that several important practical problems would arise when attempting to implement such a study.

Firstly, a company that has not been performing a broad range of different CRM strategies would have to make large marketing investments in designing an appropriate tactic for each strategic goal (e.g. customer retention through satisfaction research, complaint handling, or other tactics). Secondly, and crucially, for optimally allocating a customer to a strategy, it would be necessary to assign a sizeable part of the customer base randomly to each of the strategies, implying that by definition, customers will be targeted with strategies that are inappropriate for them, implying large marketing expenses with low return on investment, confused and unsatisfied customer responses, especially within the group of high-spending customers that has been proven to expect preferential (or at least reasonable) treatment compared to other customers [45]. While this experimental setting would likely provide rich information to researchers, the costs involved are, especially while marketing management is aware of the long-term potential of customers, of a magnitude that is not acceptable to managers. Thirdly, even if a company would be interested in researching such an optimal segmentation scheme, the generalization capacity would probably be low, considering the specificity of the tactics used. Hence, the scientific outcome of the study might only be reached when validated with several tactics for each strategy, driving the required investments even further. Finally, in order to assess the effect of the approach, the results of the study can only be expected after several years, in order to measure the changes in the slope of the customer lifecycle. The four factors mentioned above all add to the difficulties of funding, designing and implementing an optimal experimental study.

Further research is needed in two major directions. In the domain of marketing, the creation of variables having still better predictive capabilities for predicting the sign of the slope of the linear lifecycle is an interesting research topic. Alternatively, a replication of this study over different customer bases in diverse industries and countries might deliver an insight into the stability of the findings. Eventually, if resources would be available, testing and comparing different strategies (e.g. the frameworks presented in Fig. 5) 'in-the-field' can determine the full potential of the usage of customer spending evolutions for marketing decision making. Considering the Bayesian network classifiers, additional research is needed to investigate the power of other structure learning algorithms. Also the presence of hidden variables in the Bayesian network forms an interesting topic for further research.

REFERENCES

[1] S. Acid and L.M. Campos, An algorithm for finding minimum *d*-separating sets in belief networks, in: Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI), Portland, Oregon, USA, 1996, pp. 3-10.

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

[3] R.C. Blattberg and J. Deighton, Manage marketing by the customer equity test, Harvard Business Review (July-Aug) (1996) 136-144.

[4] R.J. Brodie, N.E. Coviello, R.W. Brookes, and V. Little, Towards a paradigm shift in marketing? an examination of current marketing practices, Journal of Marketing Management 13 (1997) 383-406.

[5] W. Buntine, A guide to the literature on learning probabilistic networks from data, IEEE Transactions On Knowledge And Data Engineering 8 (1996) 195-210.

[6] J. Cheng, Powerpredictor system, 2000. Available from <http://www.cs.ualberta.ca/~jcheng/bnpp.htm>

[7] J. Cheng, D.A. Bell, and W. Liu, An algorithm for bayesian belief network construction from data, in: Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics (AI and STAT), Fort Lauderdale, Florida, USA, 1997, pp. 83-90.

[8] J. Cheng, D.A. Bell, and W. Liu, Learning belief networks from data: an information theory based approach, in: Proceedings of the Sixth ACM Conference on Information and Knowledge Management (CIKM), Las Vegas, Nevada, USA, 1997, pp. 325-331.

[9] J. Cheng and R. Greiner, Comparing bayesian network classifiers, in: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, 1999, pp. 101-108.

[10] J. Cheng and R. Greiner, Learning bayesian belief network classifiers: Algorithms and system, in: Proceedings of the Fourteenth Canadian Conference on Artificial Intelligence (AI), 2001.

[11] C.K. Chow and C.N. Liu, Approximating discrete probability distributions with dependence trees, IEEE Transactions on Information Theory 14 (3) (1968) 462-467

[12] G.J. Cullinan, Picking them by their batting averages' recency-frequency-monetary method of controlling circulation, Manual release 2103, Direct Mail/Marketing Association, NY, 1977.

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

[14] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10 (7) (1998) 1895-1924.

[15] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley, New York, 1973.

[16] J.P. Egan, Signal Detection Theory and ROC analysis. Series in Cognition and Perception, Academic Press, New York, 1975.

[17] B.S. Everitt, The analysis of contingency tables, Chapman and Hall, London, 1977.

[18] T. Fawcett, Using rule sets to maximize roc performance, in: Proceedings of the IEEE International Conference on Data Mining, San Jose, California, USA, 2001.

[19] U.M. Fayyad and K.B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI) , San Francisco, CA, USA, Morgan Kaufmann, 1993, pp. 1022-1029.

[20] C. Fornell and B. Wernerfelt, Defensive marketing strategy by customer complaint management: a theoretical analysis, Journal of Marketing Research 24 (1987) 337-346.

[21] N. Friedman, D. Geiger, and M. Goldszmidt, Bayesian network classifiers, Machine Learning 29 (1997) 131-163.

[22] J. Ganesh, M.J. Arnold, and K.E. Reynolds, Understanding the customer base of service providers: an examination of the differences between switchers and stayers, Journal of Marketing 64 (2000) 65-87.

[23] D. Geiger and D. Heckerman, Knowledge representation and inference in similarity networks and bayesian multinets, Artificial Intelligence 82 (1996) 45-74.

[24] D. Geiger, T.S. Verma, and J. Pearl, Identifying independence in bayesian networks, Networks 20 (5) (1990) 507-534.

[25] C. Grönroos, From marketing mix to relationship marketing - towards a paradigm shift in marketing, Management Decision 35 (4) (1997) 322-339.

[26] 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.

[27] J.A. Hanley and B.J. McNeil, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology 148 (3) (1983) 839-843.

[28] D. Heckerman, Probabilistic Similarity Networks, MIT Press, Cambridge, MA, 1991.

[29] D. Heckerman, A tutorial on learning with bayesian networks, Technical Report MSR-TR-95-06, Microsoft Research, 1995.

[30] G.H. John and P. Langley, Estimating continuous distributions in bayesian classifiers, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI), Montreal, Quebec, Canada, Morgan Kaufmann, San Francisco, CA, 1995, pp. 338-345.

[31] S. Knox, Loyalty-based segmentation and the customer development process, European Management Journal 16 (6) (1998) 729-737.

[32] J.B. Kruskal Jr., On the shortest spanning subtree of a graph and the travelling salesman problem, in: Proceedings of the American Mathematics Society, vol. 7, 1956, pp. 48-50.

[33] P. Langley, W. Iba, and K. Thompson, An analysis of bayesian classifiers, in: Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI), San Jose, CA, USA, AAAI Press, 1992, pp. 223-228.

[34] S.L. Lauritzen, Graphical models, Clarendon Press, Oxford, 1996.

[35] K. Murphy, Bayes net matlab toolbox, 2001. Available from <http://www.cs.berkeley.edu/~murphyk/Bayes/bnt.html>.

[36] J. Pearl, Probabilistic reasoning in Intelligent Systems: networks for plausible inference, Morgan Kaufmann, San Fransico, CA, 1988.

[37] D. Peppers and M. Rogers, Enterprise one to one: tools for competing in the interactive age, Doubleday, New York, USA, 1997.

[38] F. Provost, T. Fawcett, and R. Kohavi, The case against accuracy estimation for comparing classifiers, in: J. Shavlik (Ed.), Proceedings of the Fifteenth International Conference on Machine Learning (ICML), San Francisco, CA, USA, Morgan Kaufmann, San Fransico, CA, 1998, pp. 445-453.

[39] J.R. Quinlan, C4.5 programs for machine learning, Morgan Kaufmann, San Fransico, CA, 1993.

[40] F.F. Reichheld, The Loyalty Effect, Harvard Business School Press, Cambridge, MA, 1996.

[41] F.F. Reichheld, Lead for loyalty, Harvard Business Review (July) (2001) 76-84.

[42] F.F. Reichheld and D.W. Kenny, The hidden advantages of customer retention, Journal of Retail Banking 4 (1990) 19-23.

[43] F.F. Reichheld and W.E. Sasser, Zero defections: quality comes to services, Harvard Business Review (Sept-Okt) (1990) 105-111.

[44] W.J. Reinartz and V. Kumar, On the profitability of long-life customers in a noncontractual setting: an empirical investigation and implications for marketing, Journal of Marketing 64 (2000) 17-35.

[45] W.J. Reinartz and V. Kumar, The mismanagement of customer loyalty, Harvard Business Review (July) (2002) 4-12.

[46] L.J. Rosenberg and J.A. Czepiel, A marketing approach to customer retention, Journal of Consumer Marketing 1 (1984) 45-51.

[47] R.T. Rust, V.A. Zeithaml, and K.N. Lemon, Driving customer equity: how customer lifetime value is reshaping corporate strategy, Free Press, New York, 2000.

[48] J.A. Swets and R.M. Pickett, Evaluation of Diagnostic Systems: Methods from Signal Detection Theory, Academic Press, New York, 1982.

[49] R.P. Thrasher, Cart: a recent advance in tree-structured list segmentation methodology, Journal of Direct Marketing 5 (1) (1991) 35-47.

[50] D. Van den Poel, Response Modeling for Database Marketing using Binary Classification, Ph.D. thesis, K.U. Leuven, 1999.

[51] L.C. Van Der Gaag, Bayesian belief networks: Odds and ends, The Computer Journal 39 (2) (1996) 97-113.

[52] T. Verma and J. Pearl, Causal networks: semantics and expressiveness, in: Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence, Mountain View, CA, USA, 1988, pp. 352-359.

[53] W.G. Waller and A.K. Jain, On the monotonicity of the performance of a bayesian classifier, IEEE Transactions on Information Theory 24 (3) (1978) 392-394.

[54] E.C. Malthouse and R.C. Blattberg, Can we predict customer lifetime value?, Journal of Interactive Marketing 19 (1) (2005) 2-62.

CHAPTER II

PREDICTING CUSTOMER LOYALTY USING THE INTERNAL TRANSACTIONAL DATABASE[4]

[4] This chapter is based on the following reference: Buckinx W., Verstraeten G., Van den Poel D. (2006) Predicting customer loyalty using the internal transactional database, Expert Systems with Applications, Vol 32 (1), forthcoming.

CHAPTER II:

PREDICTING CUSTOMER LOYALTY USING THE INTERNAL TRANSACTIONAL DATABASE

ABSTRACT

Loyalty and targeting are central topics in Customer Relationship Management. Yet, the information that resides in customer databases only records transactions at a single company, whereby customer loyalty is generally unavailable. In this study, we enrich the customer database with a prediction of a customer's behavioral loyalty such that it can be deployed for targeted marketing actions without the necessity to measure the loyalty of every single customer. To this end, we compare multiple linear regression with two state-of-the-art machine learning techniques (random forests and automatic relevance determination neural networks), and we show that (i) a customer's behavioral loyalty can be predicted to a reasonable degree using the transactional database, (ii) given that overfitting is controlled for by the variable-selection procedure we propose in this study, a multiple linear regression model significantly outperforms the other models, (iii) the proposed variable-selection procedure has a beneficial impact on the reduction of multicollinearity, and (iv) the most important indicator of behavioral loyalty consists of the variety of products previously purchased.

1. INTRODUCTION

In the two latest decades, Customer Relationship Management (CRM) has grown to be one of the major trends in marketing, both in academia and in practice. This evolution took form in a dramatic shift in the domain, evolving from transaction-oriented marketing to relationship-oriented marketing (Grönroos, 1997), and builds strongly on the belief that it is several times less demanding – i.e. expensive – to sell an additional product to an existing customer than to sell the product to a new customer (Rosenberg & Czepiel, 1984). Hence, it has been argued that it is particularly beneficial to build solid and fruitful customer relationships, and in this discourse, customer loyalty has been introduced as one of the most important concepts in marketing (Reichheld, 1996).

From an analytical point of view, several tools have emerged in recent years that enable companies to strengthen their relationships with customers. Moreover, the rise of new media such as the World Wide Web, and the continuous technological improvements have further increased the opportunities to communicate in a more direct, one-to-one manner with customers (Van den Poel & Buckinx, 2005). Response modeling – i.e. predicting whether a customer will reply to a specific offer, leaflet or product catalog – represents the most central application in this domain, and serves as a tool to manage customer relationships. Indeed, it would be beneficial for the company-customer relationship that the latter party would receive only information that is relevant to him/her, hence allowing the company to present only those offers for which the individual customer shows a high response probability (Baesens et al, 2002). Related to this, cross-selling analysis is involved with finding the optimal product to offer to a given customer (Chintagunta, 1992; Larivière & Van den Poel, 2004). Additionally, upselling analysis is focused on selling more – or a more expensive version – of the products that are currently purchased by the customer. Both techniques share a similar goal, i.e. to intensify the customer relationship by raising the share of products that is purchased at the focal company, and to prevent that these products would be purchased at competitive vendors. The fear of losing sales to competitors also features in churn analysis, which is focused on detecting customers exhibiting a large potential to abandon the existing relationship. Churn analysis has received great attention in the domain ever since it has been proven that even a small improvement in customer defection can greatly affect a company's future profitability (Reichheld and Sasser, 1990; Van den Poel and Larivière, 2004). Finally, lifetime value (LTV) analysis is a widely used technique to predict the future potential of

customers, in order to target only the most promising customers (Hwang et al, 2004). While these techniques can each serve individually to enhance customer relationships, it should be clear that additional advantages reside in the combination of these analytic techniques. Two recent attempts to integrate such techniques can be found in Baesens et al (2004) and Jonker et al (2004).

2. THE NEED FOR PREDICTING CUSTOMER LOYALTY

Following the previous section, we could state that both the focus on customer loyalty and the analytic tools described above have emerged from the CRM discourse. However, it is very unusual that actual customer loyalty is used to either devise or evaluate a company's targeted marketing strategies. The major cause of this deficiency lies most likely in the unavailability of information. Currently, while companies are maintaining transactional databases that store all details on any of a given customer's contacts with the focal company, these databases cannot capture the amount of products that this customer purchases at competing stores. Indeed, a study by Verhoef et al (2002) showed that only 7.5 % of companies involved in database marketing activities collect such purchase behavior. Hence, the share-of-purchases – or henceforth, the behavioral loyalty – of a certain customer is generally unavailable in the company's records, whereby the full potential of the customer (i.e., the total needs of the customer for products in the relevant category) is unknown to any specific company. However, this information could prove to be extremely valuable in different applications.

First, the knowledge of a customer's loyalty would be useful for improving CRM. We illustrate this with an example from a banking context. It would most likely be more lucrative to offer an additional savings product to a customer who has a high balance at the focal bank and at the same time has large amounts invested at other banking institutions, than to offer the savings product to a customer that has an equally high balance, but where all his/her money is invested at the focal bank. Secondly, a notion of a customer's loyalty could be used for adapting the usefulness of the model-building process. For example, currently, cross-selling models are being built on the total customer database, whereby the users will estimate the probability of purchasing this product *at the focal company*, whereas from a cross-sales point of view, it would be more interesting to estimate whether they are interested in the product category *in general*. To overcome this, it could be interesting to

build a cross-selling model on loyal customers only, because only for these customers, their total product needs are known. In this context, when attempting to model the real – and total – product needs of customers, it might seem suboptimal to include non-loyal customers into the analysis. Thirdly, the knowledge of a customer's loyalty and the evolution therein could be useful for evaluating the results of CRM-related investments, and monitoring whether certain actions lead to the desired results in the relevant customer segments.

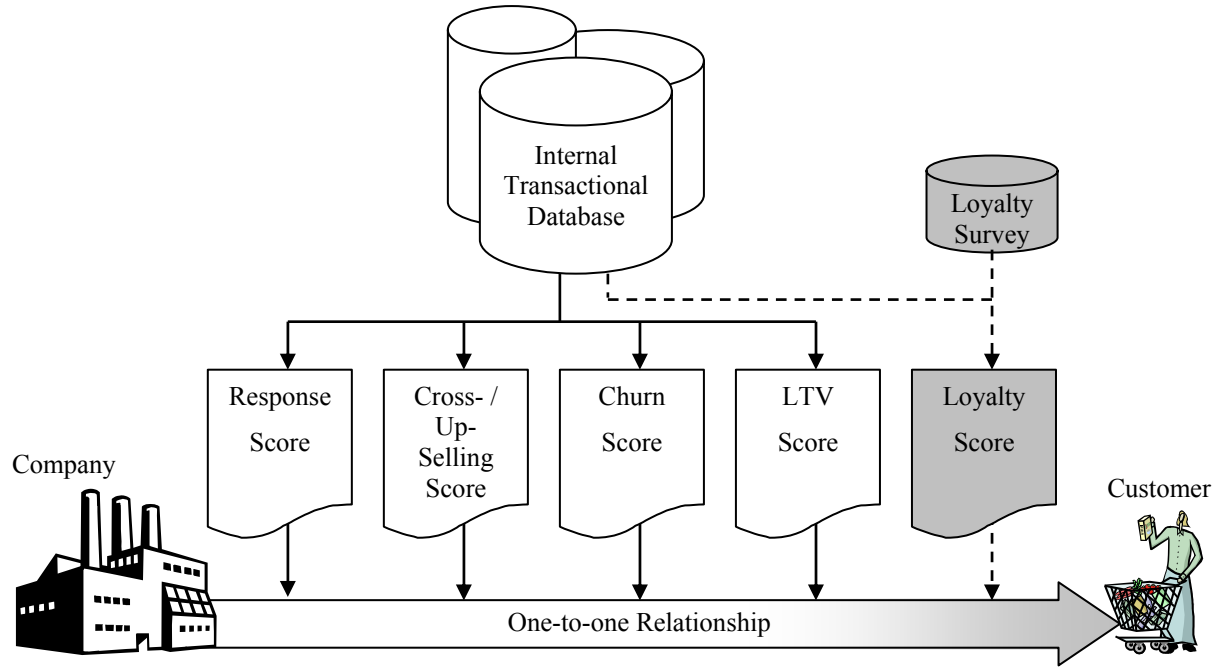


Figure 1: Creating a loyalty score from transactional data and a loyalty survey.

While such loyalty information can be obtained through a questionnaire, it would prove to be financially infeasible to obtain this information for each individual customer, especially when customers would have to be surveyed regularly in order to track changes in their loyalty profile. Consequently, in this paper, we will prove that it is sufficient to survey a sample of the company's customers, since we will combine the information stemming from the survey and the internal transactional database in order to create a loyalty score for all individual customers. Hence, as summarized in Figure 1, this score could provide additional information to the scores based on the transactional data only, and form a valuable expert tool for managing customer relationships.

The remainder of this paper is structured as follows. The next section covers the methodology used, and focuses on a description of the applied predictive techniques, the need for adequate cross-validation, and the variable-selection procedure we propose. Next, we will describe the data used for this study. In a subsequent section, we discuss the results of the proposed predictive modeling study. Finally, we end the paper with a section covering the conclusions and directions for further research.

3. METHODOLOGY

3.1 Predictive techniques

Technically, in this study, we will predict this loyalty for customers that do not belong to the surveyed sample by use of the data that is available for all customers, i.e. the transactional data. In essence this is a problem of predictive modeling. It is not our ambition to compare all possible predictive techniques. Instead, we will compare three techniques that show interesting differences and similarities. Because of the need for an accurate prediction as well as an understanding of the model – in order to explain the findings to management – we only considered models that were expected to (i) deliver adequate predictive performance on a validation set and (ii) provide an insight into the most important variables in the model. As a benchmark predictive technique, we have used a multiple linear regression (MLR) model (Cohen and Cohen, 1983), because of the widespread usage of this statistical technique in industry and academia. In this exercise, basic transformations to the variables were made to account for nonlinearity, and outliers were removed. However, no interaction variables were included into the models. We compared this benchmark with two state-of-the-art techniques from the machine learning and data mining domain. First, given the widespread use of decision trees in prediction problems where the user seeks insight into the predictive process, we have implemented Random Forests (RF, Breiman, 2001). This technique focuses on growing an ensemble of decision trees using a random selection of features to split each node (i.e. the random subspace method), where the final prediction is computed as the average output from the individual trees. RF models have been argued to possess excellent properties for feature selection, and to avoid overfitting given that the number of trees is large (Breiman, 2001). In this approach, we will grow 5000 trees, as in other applications (e.g. Geng et al, 2004). Finally, since Artificial Neural Networks (ANN's) have often been credited for achieving higher predictive performance, we selected MacKay's Automatic

Relevance Determination (ARD, MacKay 1992) neural network because it additionally reveals a Bayesian hyperparameter per input variable, representing the importance of the variable. To this end, the relevance of the features is detected by maximizing the model's marginal likelihood. We respected the author's view that a large number of hidden units should be considered in order to build a reliable model. The use of the ARD model is made possible using Markov Chain Monte Carlo techniques, hence avoiding overfitting due to the use of a Bayesian 'Occam's razor' while allowing an interpretation of the variables' importance (MacKay, 1992).

3.2 Cross-validation

An important early topic in predictive modeling consists in validating the predictive power of a model on a sample of data that is independent of the information used to build the model. In this study, the limited number of observations in each of the two settings and the elaborate number of independent variables make it hard to split our data in an estimation and a hold-out validation set. As a consequence, we prefer a resampling method called leave-one-out cross-validation because it proves to be superior for small data sets (Goutte 1997). Using this procedure, our data are divided into *k* subsets, where *k* is equal to the total number of observations. Next, each of the subsets is left out once from the estimation set and is then used to perform a validation score. To compute the real-life power of the model, the final validation set is built by stacking together the *k* resulting validations and the predictive performance is computed on this stacked set. The performance of the model – on the estimation set as well as on the validation set – is evaluated by computing (i) the correlation between surveyed loyalty and its prediction, (ii) R^2, (iii) Mean Squared Error (MSE) and (iv) the Root of the MSE (RMSE).

3.3 Variable selection

In the current study, it is likely that we can compute a large number of database-related variables in comparison with the number of observations (i.e. the number of respondents of this questionnaire). While both the RF and ARD models claim to avoid overfitting, this effect does provide a reasonable threat to the multiple regression model (Cohen and Cohen, 1983). To overcome this problem, we will make use of a variable-selection technique. Thanks to this method, the dimensionality of the model can be reduced and redundant

variables are removed, which is in favor of the model's performance. Additionally, a variable-selection procedure will allow us to gain insight in selecting the variables with good predictive capacities, and permits us to interpret the parameter estimates due to a plausible reduction of multicollinearity.

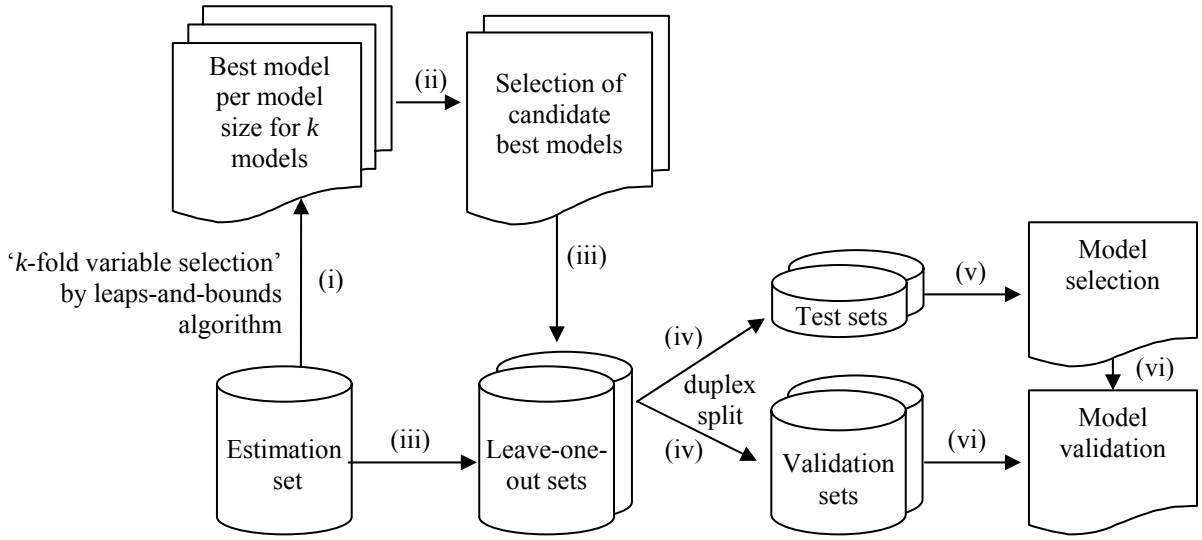


Figure 2: Model selection and validation for the multiple linear regression model.

Figure 2 partitions the variable-selection procedure that was used in this study into six disjoint steps. In step (i), we apply the leaps-and-bounds algorithm proposed by Furnival and Wilson (1974) on the estimation set. Their efficient technique identifies the model with the largest adjusted R^2 for any given model size (i.e. starting from the best model with only one variable to the full model) and at the same time avoids a full search of the variable space. However, because of the leave-one-out procedure described previously, in this case, we cannot simply perform this procedure on the total estimation set. Indeed, in order to allow for a validation of the model, the estimated models should be built when at least one observation is set aside for validation. Since it would be suboptimal to select this observation randomly, in this study we propose an iterative process in which we set aside one observation at a time, such that we create k new estimation sets, where k equals the total number of observations in the original estimation set. Hence, the outcome of this procedure – to which we refer as 'k-fold variable selection' – will consist in a list of k best models per model size. Next, in step (ii) to ensure tractability and to avoid the choice of selecting an unstable model, we reduce this list by selecting, per model size, only those models that were

'winners' in at least 5% of the occasions. In step (iii), we create the leave-one-out predictions for each candidate model using the procedure described in the previous paragraph. In the following steps, we are concerned with selecting the best models, and validating the performance of these models. Because of this dual need, in step (iv) we divide the leave-one-out data set per candidate model into a test set containing 25 % of the observations, that will be used for model selection; and a validation set consisting of the remaining 75 % of the observations, that will be used for detecting the real predictive performance of the model. Considering both the importance of a good split and the low number of observations available, we do not perform a random split, but rather complete the division via the Duplex algorithm (Snee, 1977), which performs best in separating a dataset into two sets covering approximately the factor space. Concretely, here, this factor space is composed of the set of independent variables created for the study. Next, in step (v), based on the leave-one-out test set performance, we select the best-performing model per model size among the selection of candidate models. Additionally, we select the model with the highest overall performance. In the final step (vi), we validate the real predictive performance of the models selected in the previous step on the unseen data.

4. DATA DESCRIPTION

We use data from two retail stores belonging to the same large European chain which were considered, according to management, to be representative for the entire chain. The stores carried a product assortment normally associated with grocery stores (e.g., food and beverages, cosmetics, laundry detergents, household necessities). Detailed purchase records were tracked for a period of 51 months and a summarized customer table was available that tracked basic customer demographics as well as date of first purchase.

4.1 Computation of database-related variables

It is important to mention that all transactions could be linked to customers, as the store requires use of a customer identification card. In total, 35 independent variables are computed, that are related to the following topics: (i) monetary spending, (ii) frequency of purchasing, (iii) recency of last purchase, (iv) length of the customer-company relationship, (v) interpurchase time, (vi) returns of goods, (vii) purchase variety, (viii) promotion

Table 1. Description and predictive performance of variables used.

Variable	Description	MLR Standardized Parameter Estimates		RF Variable Importance	ARD Alpha (Importance)
Spending_1M	Spending during last month.	0.3540	***	0.0086	21.49
Spending_6M	Spending during last six months.	0.4582	***	0.1136	13.00
Spending_1Y	Spending during last year.	0.4789	***	0.2246	15.58
Spending_2Y	Spending during last two years.	0.4742	***	0.0228	23.63
Spending	Spending in total history.	0.4714	***	0	32.08
NumItems	Number of product items bought.	0.4705	***	2.3071	16.27
Spending_Fresh	Spending on fresh food products.	0.4395	***	0.2985	17.52
rSpend_Freq	Average Spending per visit.	0.1785	***	0.0055	7.11
rSpend_Lor	Spending relative to the length of the customer's relationship.	0.4726	***	0.4104	0.16
Frequency_1M	Number of purchases during last month.	0.3477	***	0	2.41
Frequency_6M	Number of purchases during last six months.	0.4356	***	0.035	3.76
Frequency_1Y	Number of purchases during last year.	0.4455	***	0.0544	3.91
Frequency_2Y	Number of purchases during last two years.	0.4494	***	0	2.77
Frequency	Number of purchases in total history.	0.4389	***	0	3.87
Recency	Number of days since last purchase.	–0.2035	***	0	24.44
Ipt	Average number of days between store visits.	–0.2965	***	0.6045	17.23
Std_Ipt	Standard deviation of the number of days between the purchases.	–0.3227	***	0.292	13.20
Lor	Length of customer relationship.	0.0940	***	0	29.92
Numcat_LY	Number of different product categories purchased from during last year.	0.5221	***	0.543	6.32
Numcat_2Y	Number of different product categories purchased from during last two years.	0.4770	***	0.2001	3.09
Numcat_3Y	Number of different product categories purchased from during last three years.	0.4460	***	0.1434	5.27
Numcat	Number of different product categories purchased from during the total history.	0.4805	***	0.2233	10.28
Neg_Inv	Dummy to indicate if the customer ever had a negative invoice (1/0).	0.2919	***	0.1115	2.42
Ret_Item	Dummy to indicate if the customer ever returned an item (1/0).	0.2656	***	0.0293	1.56
Returns	Total value of returned goods.	0.1572	***	0	11.90
NumPromItems	Number of items bought that appeared in company's promotion leaflet.	0.4539	***	0.9065	9.62
SpenPromItems	Money spent on products that appeared in promotion leaflet.	0.4572	***	0.0064	11.79
Visitspromitems	Number of visits on which a product is bought that appeared in the promotion leaflet.	0.4680	***	0.0342	5.35
PercNumPromItems	Percentage of products bought that appeared in leaflet.	0.0139		0.06	8.48
PercResp_Leaf	Percentage of times a purchase is made given that a promotion leaflet was received.	0.4792	***	0	0.22
PercResp_Noleaf	Percentage of times a purchase is made given that no promotion leaflet was received.	0.3098	***	0	1.32
Perc_Noleaf_Freq	PercResp_Noleaf divided by shopping frequency	–0.2235	***	0.1258	2.57
MoreThanOnce	Number of times that a customer visits more than once within the same promotion period.	0.4308	***	0	2.92
PercMoreThanOnce	MoreThanOnce divided by the number of times a customer bought in a promotion period.	0.2940	***	0	0.34
Distance	Distance to the store.	–0.1265	***	0.0457	6.07

*** $p < .01$.

sensitivity, (ix) responsiveness on mailings and (x) distance to the store. The inclusion of these variables was mainly based on previous literature in the domain of predicting the strength of the relationship between a company and its customers (see, e.g., Bult and Wansbeek, 1995; Srinivasan, Anderson and Ponnavolu, 2002; Reinartz and Kumar, 2002; Buckinx and Van den Poel, 2005; Baesens et al, 2004). Table 1 summarizes all these variables, together with a brief description of how they are calculated.

4.2 Loyalty survey

In addition to these transactional data, a self-administered survey was used as a complementary data collection method. Data collection took place in each of the retail stores mentioned previously. Surveys were randomly distributed to customers during their shopping trips, and customer identification numbers were recorded for all customers who received a questionnaire.

Table 2. Wording of the items of the loyalty scale.

Item 1	Buy (much less … much more) grocery products at XYZ than at competing stores.
Item 2	Visit other stores (much less frequently … much more frequently) than XYZ for your grocery shopping (–).
Item 3	Spend (0% … 100%) of your total spending in grocery shopping at XYZ.

A customer's behavioral loyalty was determined as a composite measure by comparing a customer's spending at the retailer with their total spending in the relevant product category. As a first item, and similar to Macintosh and Lockshin (1997), the percentage of purchases made in the focal supermarket chain versus other stores was assessed on an 11-point scale that ranged from 0% to 100% in 10% increments (i.e., 0%, 10%, 20%, and so on). Additionally, two seven-point Likert-type items assessed the shopping frequency of the customers for the focal store when compared to other stores. We pretested the questionnaire and refined it on the basis of pretest results. Table 2 gives the exact wording of the items used. After rescaling the second item (due to its expected negative correlation with both other items), we standardized the 3 loyalty-related questions, and averaged them to represent the behavioral loyalty construct.

5. RESULTS

5.1 Survey response

Of the 1500 distributed questionnaires, we received 878 usable responses (i.e. a ratio of usable response of 58.33%). We successfully tested for nonresponse bias by comparing database variables such as spending, frequency of visiting the store, interpurchase time, length-of-relationship and response behavior towards companies' mailings between respondents and nonrespondents.

A usable response had all fields completed, and the respondent could be successfully linked to his or her transaction behavior in the customer database. We tested construct reliabilities of the loyalty scale by means of Cronbach's coefficient alpha. The resulting coefficient of .871 clearly exceeds the .7 level recommended by Nunnally (1978), which proves it is a reliable scale, especially given the fact that reverse coding was used to measure one item of the 3-item scale.

5.2 Predictive performance

In terms of predictive performance, in Table 3, we compare the results of the different models. Considering the MLR models, we compared the full model with the final model resulting from the variable-selection procedure described previously, which resulted in a selection of just 4 variables. Regarding the results from the RF model, all variables were introduced, yet only 24 variables were selected by the technique. In terms of the ARD model, after extensive trial-and-error testing, we reached an optimal performance by using 24 hidden units. No variables were selected by the latter technique so each variable contributes, to some extent, to the predictive performance.

Different interesting conclusions can be drawn from Table 3. First, it is clear that – as was expected – overfitting prevails in the MLR model, and does not appear in the RF model. This finding is in line with Breiman's (2001) initial claims as well as findings by other authors (e.g. Buckinx and Van den Poel, 2005). Indeed, the R^2 of the full MLR model drops from 0.3208 on the estimation set to 0.2608 on the validation set, which introduces skepticism on the validity of this model. Second, the variable-selection procedure we described previously succeeds in reducing the negative impact related to overfitting. Indeed, the difference

between the R^2 on the estimation set (0.3064) versus the test set performance (0.2962) is sufficiently small. Thirdly, contrarily to what might have been expected using the Bayesian 'Occam's razor' (MacKay, 1992), the ARD model also proves to be sensitive to overfitting, as the performance on the estimation set is substantially higher than the performance after cross-validation. Fourth, given that an efficient variable-selection procedure is performed to the regression model, this model clearly outperforms the other models in terms of predictive performance. Fifth, in order to test whether this result is significant, we tested whether the correlations (R) differ significantly using a test of the difference of dependent samples described in Cohen and Cohen (1983, p. 57). From this test, we can conclude that the MLR model significantly outperforms the RF ($t = 2.57$, $p = 0.01022$) and ARD models ($t = 2.68$, $p = 0.00747$). However, the difference in performance between the RF and ARD models is not significant ($t = 1.39$, $p = 0.16421$).

Table 3. Model performances (highest validated predictive performance indicated in bold face).

	MLR				RF		ARD	
	Full Model (v=35)		**Final Model (v=4)**		Full Model (v=35)		Full Model (v=35)	
	Estimation	Validation	Estimation	**Validation**	Estimation	Validation	Estimation	Validation
R	0.5664	0.5107	0.5535	**0.5442**	0.5186	0.5238	0.5714	0.4935
R^2	0.3208	0.2608	0.3064	**0.2962**	0.2689	0.2744	0.3265	0.2435
MSE	0.5586	0.6107	0.5502	**0.5569**	0.6023	0.5969	0.5586	0.6237
RMSE	0.7474	0.7815	0.7417	**0.7463**	0.7761	0.7726	0.7474	0.7898

To conclude, given that the coefficient of determination of the final MLR model is fairly high (0.2962) for cross-sectional data, and given its significance ($F = 96.39$, $p = <.0001$), we can state that it is possible to predict a customer's loyalty to a reasonable degree from the internal transactional database using a regression model – provided that an elaborate variable-selection procedure is performed. Because of the importance of the latter procedure, we discuss its implications in detail in the following paragraph.

5.3 Usefulness of the variable-selection technique

In Figure 3, we illustrate the effect of the variable-selection technique by plotting the estimation, test and validation performance of the best-performing model per model size. While the error (RMSE) on the estimation data set does not increase substantially as the

number of variables increases, the validity of these models is severely hampered. However, the splitting of the leave-one-out sample into a test and validation set does clearly allow us to select the best-performing model and validate this model, while efficiently exploiting the available observations. Hence, the test set error reached its lowest level with the use of only four variables, whereby overfitting is reduced.

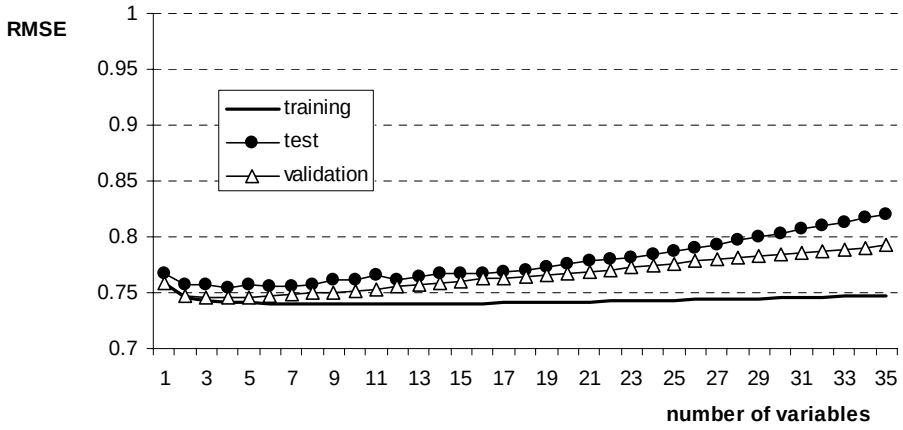


Figure 3: Evidence of overfitting when the number of variables is increased.

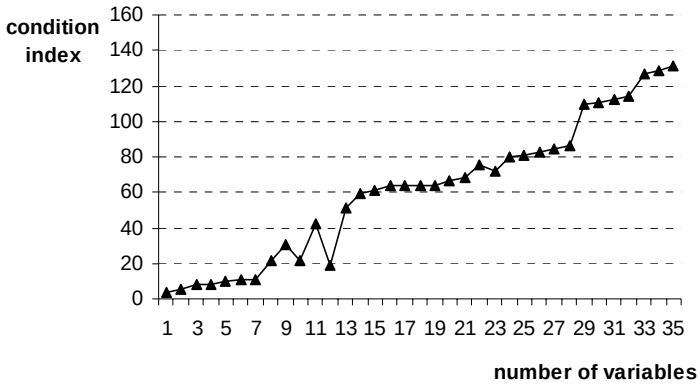


Figure 4: Detecting multicollinearity by the condition index.

While we have focused on the negative impact of using a large set of variables on the predictive performance of the model, an additional threat resides in the occurrence of multicollinearity. Indeed, it is likely that, when using a large number of predictors, several predictors that are jointly used might be severely correlated. Hence, the affected parameter estimates might become unstable and may exhibit high standard errors, reflecting the lack of properly conditioned data (Belsley et al, 1980). In this section, we will illustrate the

existence of multicollinearity graphically. To this goal, we follow the procedure of Belsley et al (1980), and hence we present the evolution of the condition index of the best performing model per model size in Figure 4. Note that the condition indices are the square roots of the ratio of the largest eigenvalue to each individual eigenvalue (as calculated by the *collin* option in *proc reg* in *SAS*). Considering the author's informal suggestion that, at an index larger than 15, weak dependencies may start to affect the regression estimates (Belsley et al, 1980, p. 153), those models incorporating more than 7 variables might exhibit unstable estimates and high standard errors. In order to validate this rule of thumb we have attempted to provide a graphical representation of the stability of the estimates. To this effort, we have computed the parameter estimates of all variables when they are used separately in univariate predictive models. Next, we compared the signs of these parameters – to which we refer as the 'correct' signs – with the signs of the best multiple regression models, and we plotted the percentage of 'correct' signs in Figure 5. The results confirm the previously offered rule-of-thumb, as at least some parameter signs differ in models that contain more than 7 variables. Hence, in these models, the parameter estimates can be considered as unstable.

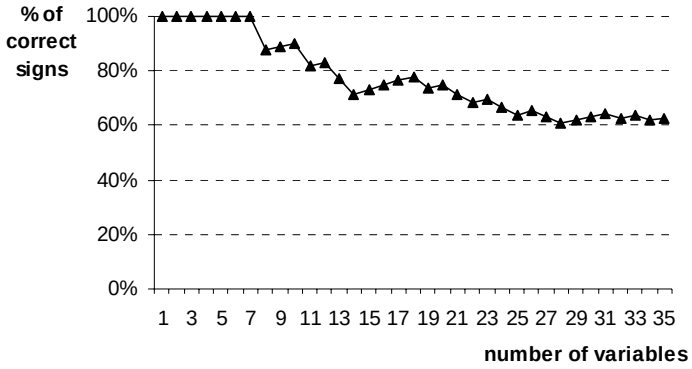


Figure 5: An illustration of the effect of multicollinearity on the parameter signs.

To conclude this section, the full model – containing all variables – shows evidence of multicollinearity that is manifested in a condition index of 131.6 and the fact that only 63% of the parameter signs correspond to their univariate counterparts. However, these problems seem efficiently solved in the final model – containing only the four selected variables – showing a condition index of only 8.5 and a proportion of 100% 'correct' parameter signs. Note, however, that this does not imply that the final set of variables are only weakly correlated. In the Appendix of this chapter, we present the correlations between the four

selected predictors. While on average, the correlation is 0.6808, the condition index remains fairly low, and hence the impact on the proportion of 'correct' parameter signs is low.

Table 4. Parameter estimates of the best predictive models.

Number of variables	Variable	Standardized Estimate	t Value	Pr > |t|	R^2adj Validation
1	Intercept	0	-15.69	<.0001	0.2678
	Numcat_LY	0.5221	18.12	<.0001	
2	Intercept	0	-14	<.0001	0.2905
	Spending	0.2154	5.56	<.0001	
	Numcat_LY	0.3751	9.67	<.0001	
3	Intercept	0	-14.3	<.0001	0.2934
	Numcat_LY	0.2979	6.16	<.0001	
	PercResp_Leaf	0.1240	2.64	0.0084	
	rSpend_Lor	0.1859	4.59	<.0001	
4	Intercept	0	-13.62	<.0001	0.2919
	Spending_Fresh	0.0887	2.12	0.0343	
	Numcat_LY	0.2741	5.54	<.0001	
	PercResp_Leaf	0.1145	2.43	0.0151	
	rSpend_Lor	0.1468	3.31	0.001	
5	Intercept	0	-11.91	<.0001	0.2926
	Spending_ Fresh	0.0994	2.41	0.0162	
	Numcat_LY	0.2389	4.54	<.0001	
	NumItems	0.1017	2.16	0.031	
	PercResp_Leaf	0.1651	3.07	0.0022	
	rSpend_Freq	0.0739	2.21	0.027	
6	Intercept	0	-8.46	<.0001	0.2911
	Spending_ Fresh	0.1024	2.48	0.0133	
	Numcat_LY	0.2193	4.06	<.0001	
	NumItems	0.1043	2.22	0.0269	
	PercResp_Leaf	0.1487	2.72	0.0066	
	rSpend_Freq	0.0732	2.2	0.0284	
	Std_Ipt	-0.0553	-1.64	0.1007	
7	Intercept	0	-8.53	<.0001	0.2881
	Spending_ Fresh	0.1009	2.44	0.0147	
	Neg_Inv	0.0396	1.2	0.2313	
	Numcat_LY	0.2172	4.03	<.0001	
	NumItems	0.0990	2.09	0.0365	
	PercResp_Leaf	0.1367	2.46	0.0141	
	rSpend_Freq	0.0769	2.3	0.0219	
	Std_Ipt	-0.0520	-1.54	0.1237	

5.4 Variable importance

In order to discuss the importance of the variables to predict behavioral loyalty, we will look both at the univariate performances as well as the inclusion of these variables into the MLR models. First, in terms of the univariate importances, Table 1 illustrates that the different models emphasize different variables. For example, in the ARD model, the length of relationship is considered as the second most important variable, while in the MLR model it features as the second least important variable, and the variable was not selected in the RF model. The difference between the models can be evaluated more formally through the computation of the correlation between the variable importances. The correlation between the MLR model and the RF model is 0.08862 (p=0.6127), between the MLR model and the ARD model -0.16933 (p=0.3308), and between the RF model and the ARD model 0.12051 (p=0.4905), so we conclude that the models really emphasize different predictors. Since the MLR model outperforms the other models, in the remainder of this paragraph, we will focus on the importance of variables according to the MLR model. From the univariate performances, we note that the purchase variety clearly forms the best predictor of loyalty. However, several groups of variables have only a slightly lower performance. Variables related to the spending, frequency, promotion behavior and response on mailings all have a good predictive performance. The other variables, such as recency, interpurchase time, length of relationship, average spending per visit, returns of goods and distance to the store clearly exhibit lower univariate predictive performance. An additional insight can be gained from the inclusion of the variables in the best performing multivariate models. Hence, in Table 4, we present the variables of the selected models that contain up to seven variables. This confirms the fact that purchase variety, spending and a customer's response on mailing folders present the most useful information for predicting behavioral loyalty.

6. CONCLUSIONS AND DIRECTIONS FOR FURTHER RESEARCH

Following the prevalence of the CRM discourse, companies have started to realize the value of loyal customers, and have acquired the competences to manage customer relationships through targeted communications. Intriguingly however, these relationships are currently managed almost unanimously based on transactional data (such as recency, frequency, and monetary value of a customer) while the behavioral loyalty and hence the full potential of a customer is generally unavailable. In this study, we have constructed a reliable three-item scale to measure behavioral loyalty, and we have proven that it is possible to predict a

customer's behavioral loyalty to a reasonable degree based on his/her transactional information. Hence, we have provided a viable methodology for building a loyalty score for all customers, based on a limited sample of customers for which behavioral loyalty was surveyed. This additional customer knowledge can be useful in many marketing applications within the area of customer relationship management, be it direct marketing, model building and customer evaluation.

To this end, we compared three techniques that have been argued to show a good predictive performance and an interpretation of the importance of the predictors. More specifically, we compared multiple linear regression with two state-of-the-art techniques, namely Breiman's regression forests and MacKay's automatic relevance determination. The predictive modeling we propose in this study is different from the general situation of predicting transactional behavior by use of historic transactional behavior in the sense that here, the target variable is only known for a limited set of customers. Because overfitting is more likely to occur when the observations are limited compared to the number of variables, and since overfitting is a well-acknowledged problem in multiple linear regression, the major contribution of this study lies in designing an effective variable-selection procedure. Hence, considering the limited sample size, we propose a model selection and validation procedure that is based on the leaps-and-bounds algorithm using an intelligent split of a leave-one-out cross-validation sample. In a real-life study, we show that this procedure effectively increases the validation performance to an extent that the linear regression model outperforms the other models in terms of predictive accuracy, and that multicollinearity is removed to an adequate degree in the resulting model, allowing for a sound interpretation of the parameters. Hence, we show that purchase variety is the best performing predictor of behavioral loyalty, and that a customer's spending, frequency, promotion behavior, response to mailings and regularity of purchasing all provide useful information to deliver an adequate prediction of a customer's behavioral loyalty.

As any other study, this study has its limitations which may lead to further research. First of all, in this paper it was not our ambition to compare all possible predictive modeling techniques. Hence, it is not excluded that other techniques serve even better to predict behavioral loyalty. Instead, we have confirmed that a proper use of sound statistical techniques is at least able to compete with two state-of-the-art predictive techniques. Second, contrarily to what was expected, we gained evidence of overfitting in the ARD model. While

again it was not the focus of this specific study, this finding seems at least intriguing. Hence, further research might focus on performing a (possibly similar) variable-selection technique for the ARD model to account for the overfitting that was detected. Thirdly, in this case, we have used a leave-one-out cross-validation sample. It is not unlikely, however, that for future usage, the procedure could be applied in a more resource-efficient way by applying a leave-*k*-out cross-validation, where *k* is increased while carefully monitoring the validity of the results. Finally, in this procedure, due to financial constraints, it was not possible to perform an out-of-sample cross-validation to account for any possible model drift. Indeed, a subsequent survey of the behavioral loyalty would prove useful in evaluating the stability of the model for future loyalty predictions.

ACKNOWLEDGEMENTS

The authors are grateful to Leo Breiman (2001) for the public availability of the random forests software, and to Ian T. Nabney (2001) for his implementation of the technique of automatic relevance determination using neural networks.

APPENDIX: CORRELATIONS BETWEEN THE SELECTED MLR VARIABLES

	Spending_Fresh	Numcat_LY	PercResp_Leaf	rSpend_Lor
Spending_Fresh	1	0.65375 <.0001	0.6075 <.0001	0.6951 <.0001
Numcat_LY		1	0.77877 <.0001	0.6868 <.0001
PercResp_Leaf			1	0.66271 <.0001

REFERENCES

Baesens, B., Verstraeten, G., Van den Poel, D., Egmont-Petersen, M., Van Kenhove, P., & Vanthienen, J. (2004). Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. European Journal of Operational Research, 156 (2), 508-523.

Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J. & Dedene, G. (2002). Bayesian Neural Network Learning for Repeat Purchase Modelling in Direct Marketing. European Journal of Operational Research, 138 (1), 191-211.

Belsley, D.A., Kuh, E., & Welsch, R.E. (1980). Regression Diagnostics. New York: John Wiley & Sons, Inc.

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32.

Buckinx, W. & Van den Poel, D. (2005). Customer Base Analysis: Partial Defection of Behaviorally-Loyal Clients in a Non-Contractual FMCG Retail Setting. European Journal of Operational Research, 164 (1), 252-268.

Bult, J.R. & Wansbeek, T. (1995). Optimal Selection for Direct Mail. Marketing Science, 14 (4), 378–394.

Chintagunta, P.K. (1992). Estimating a multinomial probit model of brand choice using the method of simulated moments. Marketing Science, 11 (4), 386-407.

Cohen, J. & Cohen, P. (1983). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (2nd ed). Hillsdale, NJ: Erlbaum.

Furnival, G.M. & Wilson, R.W. (1974). Regressions by Leaps and Bounds. Technometrics, 16, 499–511.

Goutte, C. (1997). Note on Free Lunches and Cross-Validation. Neural Computation, 9, 1245–1249.

Geng, W., Cosman, P., Berry, C.C., Feng, Z. & Schafer, W.R. (2004). Automatic Tracking, Feature Extraction and Classification of C Elegans Phenotypes. IEEE Transactions on Biomedical Engineering, 10 (51), 1811-1820.

Grönroos, C. (1997). From marketing mix to relationship marketing—towards a paradigm shift in marketing. Management Decision, 35 (4), 839-843.

Hwang, H., Jung, T. & Suh, E. (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications, 26 (2), 181-188.

Jonker, J.J., Piersma, N. & Van den Poel, D. (2004). Joint Optimization of Customer Segmentation and Marketing Policy to Maximize Long-Term Profitability. Expert Systems with Applications, 27 (2), 159-168.

Larivière, B. & Van den Poel, D. (2004). 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), 277-285.

Macintosh, G. & Lockshin, L.S. (1997). Retail Relationships and Store Loyalty: A Multi-Level Perspective. International Journal of Research in Marketing, 14 (5), 487–497.

MacKay, D.J. (1992). Bayesian Interpolation. Neural Computation, 4, 415-447.

Nabney, I.T. (2001). Netlab Algorithm for Pattern Recognition, Springer.

Nunnally, J.C. (1978). Psychometric Theory. New York: McGraw-Hill.

Reichheld, F.F. (1996). The Loyalty Effect. Harvard Business School Press, Cambridge, MA.

Reichheld, F.F. & Sasser, W.E. Jr (1990). Zero defections: quality comes to service. Harvard Business Review, 68 (5), 105-111.

Reinartz, W.J. & Kumar, V. (2000). On The Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing. Journal of Marketing, 64 (October), 17–35.

Rosenberg, L.J. & Czepiel, J.A. (1984). A Marketing Approach to Customer Retention. Journal of Consumer Marketing, 1, 45–51.

Snee, R.D. (1977). Validation of regression models: Methods and examples. Technometrics, 19 (4), 415-428.

Srinivasan, S.S., Anderson, R. & Ponnavolu, K. (2002). Customer Loyalty in E-commerce: An Exploration of its Antecedents and Consequences. Journal of Retailing, 78, 41–50.

Van den Poel, D. & Buckinx, W. (2005). Predicting Online Purchasing Behaviour. European Journal of Operational Research, 166 (2), 557-575.

Van den Poel, D. & Larivière, B. (2004). Customer Attrition Analysis for Financial Services Using Proportional Hazard Models. European Journal of Operational Research, 157 (1), 196-217.

Verhoef, P.C., Spring, P.N., Hoekstra, J.C. & Leeflang, P. (2002). The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. Decision Support Systems, 34, 471-481.

CHAPTER III

TOWARDS A TRUE LOYALTY PROGRAM: INVESTIGATING THE USEFULNESS AND FEASIBILITY OF REWARDING CUSTOMERS ACCORDING TO THE BENEFITS THEY DELIVER[5]

[5] This chapter is based on the following reference: Geert Verstraeten, Wouter Buckinx, Dirk Van den Poel, 2005. Towards a True Loyalty Program: Investigating the Usefulness and Feasibility of Rewarding Customers According to the Benefits They Deliver, submitted to Journal of Marketing, 2nd round of review process.

CHAPTER III:

TOWARDS A TRUE LOYALTY PROGRAM: INVESTIGATING THE USEFULNESS AND FEASIBILITY OF REWARDING CUSTOMERS ACCORDING TO THE BENEFITS THEY DELIVER

ABSTRACT

In the discourse of relationship marketing, loyal customers have been proven to deliver a number of valuable benefits to companies. In contrast, for largely practical reasons, the reward criteria for most loyalty programs are not based on customer loyalty. This study examines to what extent the use of an alternative reward system would enable companies to improve the way loyal customers are currently rewarded for the benefits they deliver. Using historical purchase data, we show that if customers were rewarded for their behavioral loyalty instead of past spending or length of relationship, the rewards received would better compensate customers who are spreading positive word-of-mouth, are price insensitive and have high repurchase intentions.

Since our alternative criterion, behavioral loyalty, cannot be easily recorded in customer databases, we provide evidence that it can be predicted from the company's internal data records. Moreover, in a moderated linear regression framework, the constructed predictor promises to provide a more efficient criterion than spending or length of relationship for rewarding those customers who deliver the benefits usually related to loyal customers. Remarkably, our models show that the variety of products purchased and responsiveness to

direct mail are the most valuable predictors of behavioral loyalty. In order to generalize our findings, we validate all results in both grocery and general merchandise shopping.

1. INTRODUCTION

In the previous two decades, marketing has seen a dramatic shift, in which traditional—i.e., product-oriented—marketing has given way to an increasingly customer-oriented view. The best-known theorem underlying this new view states that acquiring a new customer is several times more costly than retaining and selling additional products to existing customers (Rosenberg and Czepiel 1984). In this evolution, to which many authors refer as "the paradigm shift in marketing" (Brodie et al. 1997), the loyalty of individual customers has rapidly grown to become the focal point of relationship marketing (Dick and Basu 1994).

Advocates of traditional relationship marketing attribute several advantages to loyal customers. They are said to increase their spending over the course of their relationship with a company (Reynolds and Arnold 2000), generate new customers by their positive word-of-mouth (Reichheld 2003), require diminished costs to serve (Dowling and Uncles 1997), exhibit reduced customer price sensitivities and have a salutary impact on the company's employees (Reichheld and Sasser 1990). In the remainder of this paper, we will refer to such alleged benefits of loyal customers as 'Loyalty Benefits'. Hence, it is crucial to note that we do not consider these benefits to be attributes of loyal customers, we merely use this term to refer to the benefits often related to customer loyalty. An overview of the main findings with respect to these benefits is shown in the literature review section of this paper. In the development of relationship marketing, different companies have conceived programs, often termed 'Loyalty Programs' or perhaps more accurately 'Reward Programs', in order both to reward and to stimulate such desirable customer behavior (Kivetz and Simonson 2003; Dowling and Uncles 1997). Today, companies ranging from large entities—such as American Airlines[6], American Express, AT&T, Carrefour, Hertz, Hilton Hotels and Shell—to small local merchants, offer reward programs that grant advantages to their customers, proportional to the money spent at their stores. Hence, regardless of the success of relationship marketing, these relationship-building programs are currently focused on rewarding merely repeat-purchase behavior (Nicholls 1989), being just one of the benefits

[6] The Advantage program of American Airlines is often cited as the first example of such a program.

attributed to loyal customers. Conversely, other benefits—which are also considered to be very important for the growth and the continuity of the company—are rewarded to a far lesser degree. Hence, it could be stated that currently, customers are rewarded proportional to a proxy variable of loyalty—spending—instead of loyalty itself.

From a psychological point of view, rewarding customers can have multiple effects. First, the motivating impact of rewards has long been established in well-known experiments where animals have been proven to persist in the rewarded behavior (e.g., Latham and Locke 1991). Again, this underlines the importance of choosing the desired behavior to be rewarded, henceforth called the *reward criterion*. Accordingly, also in human behavior research, people have proven to be highly motivated to deliver efforts directed at achieving future rewards (e.g., Nicholls 1989). For marketing, it has been suggested that the excitement surrounding relationship marketing has created an expectation that customers who deliver benefits for the company will be rewarded for their loyalty (Dowling and Uncles 1997). In the context of loyalty programs, recent research has shown that customers are attracted more to programs if they feel that they are at an advantage to earn rewards when compared to other customers (Kivetz and Simonson 2003), which can again be related to social comparison theory (Festinger 1954). In summary, the design of the current loyalty programs can be seriously questioned. Indeed, loyal customers who deliver benefits to the company, but who are not big spenders, might feel discriminated against by big spenders who reap benefits without being loyal. Hence, companies that are able to compensate customers to alleviate this discrimination might create a competitive advantage.

Intriguingly, to the best of our knowledge, no previous study has focused on evaluating the extent to which a customer's spending or length-of-relationship as proxy variables for loyalty sufficiently reward customers for the benefits often related to loyal customers. In previous research, however, the concept of loyalty itself has been operationalized in different forms. In this study, it is crucial to note that we consider behavioral loyalty: a customer who spends 100% of his purchases in a given store can be seen as 100% loyal. This choice is consistent with previous research on loyalty programs (e.g., De Wulf, Odekerken-Schröder and Iacobucci 2001). This is confirmed in Sharp and Sharp (1997), stating that reward systems attempt to maximize customers' share of wallet and should be evaluated in terms of the behavioral changes they create. Moreover, it is necessary in this study to make a clear distinction between the desired behavior – i.e. a customer who fulfills his full potential at the

desired store – and the benefits often attributed to loyal customers. We refer to section 4.2.1 for an exact description of the measurement of behavioral loyalty in this study.

Consequently, we simulate the use of different reward programs by making use of a moderated linear regression framework. We examine the degree to which a given reward criterion is efficient in compensating customers for the benefits they deliver in terms of word of mouth, price insensitivity and purchase intentions. Furthermore, we propose a viable and feasible solution for each company that administers a customer database, to include our proposed criterion in the architecture of a reward scheme. To be precise, we propose a predictive model, which, at the same time, gives insight into the most important indicators of loyalty available in the database. All results are validated in two different store settings: a grocery shopping environment and a general merchandising shopping setting.

The remainder of this paper is structured as follows. In the following section, we provide an overview of the research surrounding customer loyalty and the currently used reward programs. Next, Section 3 discusses the hypotheses investigated in this study. Section 4 deals with the methodology used to evaluate the hypotheses posited previously. Section 5 provides an overview of the results, and to conclude, Section 6 ends this study with a discussion.

2. LITERATURE REVIEW

2.1 Loyalty benefits

Advocates of traditional relationship marketing attribute several advantages to loyal customers. Table 1 gives an overview of studies focused on evaluating whether loyal customers do exhibit the alleged loyalty benefits. Some studies in this area are restricted to anecdotal discussions. Reichheld and Sasser (1990) were the first to claim that the length of a relationship makes customers more attractive, whereas Dick and Basu (1994) concluded that comparable benefits were dependent upon customers' loyalty level. In contrast, Dowling and Uncles (1997) did not agree and found arguments to dispute all of the proposed benefits.

These contradictions enticed researchers to search for empirical evidence, which only created more ambiguity. Reinartz and Kumar (2000) undermined nearly all of the benefits suggested by Reichheld and Sasser (1990). In contrast, Reynolds and Arnold (2000) supported the

Table 1. Literature Review: Customer Benefits

Author	Target variable	Measurement	Benefits	Relationship	Supported (s) Not supported (ns) Anecdotal (a)	Data
Dick and Basu (1994)	Loyalty	Attitudinal and Behavioral	Word-of-mouth	+	a	General
			Resistance to counter persuasion	+	a	
			Search motivation	–	a	
Dowling and Uncles (1997)	Loyalty	/	Cost of serving	no	a	General
			Price insensitivity	no	a	
			Profitability	no	a	
			Word-of-mouth	no	a	
Reichheld (2003)	Loyalty	Behavioral	Word-of-mouth	+	s	Six industries
Reichheld and Sasser (1990)	Lifetime duration	/	Profitability	+	a	General
			Cost of serving	–	a	
			Price insensitivity	+	a	
			Word-of-mouth	+	a	
Reinartz and Kumar (2000)	Lifetime duration	Lifetime duration model	Profitability	+	ns	Catalog retailer
			Profit increase	+	ns	
			Cost of serving	–	ns	
			Price insensitivity	+	ns	
Reynolds and Arnold (2000)	Loyalty	Attitudinal (4 items)	Word-of-mouth	+	s	Department
			Competitive resistance	+	s	stores
			Share of wallet	+	s	
Srinivasan, Anderson and Ponnavolu (2002)	Loyalty	Attitudinal (7 items)	Word-of-mouth	+	s	Online B2C
			Price insensitivity	+	s	
			Consideration set size	+	ns	
This study	Loyalty	Behavioral (3 items)	Word-of-mouth	+	s	General merchandise
			Price insensitivity	+	s	and
			Purchase intentions	+	s	grocery shopping

existence of beneficial loyalty behavior in a department-store setting, and Srinivasan et al. (2002) came to similar conclusions in an online setting. Finally, Reichheld (2003) confirmed his earlier findings: "Loyal customers talk up a company to their friends and colleagues". The review shows that ambiguity exists in determining whether loyal customers really deliver loyalty benefits. Our analysis will give more insight into this issue. We examine word-of-mouth, price insensitivity and purchase intentions since these are among the items investigated most.

2.2 Current reward programs

As Kivetz and Simonson (2003) note, an important goal of relationship marketing has been the development of customer loyalty. They also mention that loyalty programs have often been used to this end. Hence, while the original design of such programs consisted of rewarding customer loyalty (Dowling and Uncles 1997), in practice, most current reward systems do not use this criterion. Bonus systems like frequent flyer programs and schemes from credit card firms, banks, telephone companies and retailers encourage repeat purchase (Whyte 2004), and are usually rewarding customers for their spending, relationship duration or a combination of both (McMullan and Gilmore 2002). Also in academic research, spending and lifetime are often used to evaluate customers. In their loyalty program evaluation, Dowling and Uncles (1997) only consider reward schemes based on spending level. While Reinartz and Kumar (2000) recommend basing rewards on past spending of customers, in their research they evaluate whether long-life customers exhibit the benefits often attributed to loyal customers. Thus, they clearly evaluate the usefulness of length-of-relationship as an optional reward criterion. Verhoef (2003) makes use of a reward program that gives discounts based on the level of usage and the length of a customer's relationship. Additionally, he suggests that, when the reward structure depends on the length-of-relationship, customers would be less likely to switch, because of the time lag before the same level of rewards can be received by another supplier.

Two main reasons can be found to account for the use of proxy variables such as spending and length-of-relationship. The first reason for companies to make use of behavioral customer information is that such a measure of customer loyalty is not readily available in transactional databases (Jones and Sasser 1995). For a company with many customers, it is impractical to collect the required loyalty data for each of its customers by sending out

questionnaires. In contrast, gaining knowledge about customers' spending behavior and lifetime duration is relatively straightforward because all the required data can be found in customer information files (Verhoef, Franses and Hoekstra 2002). Second, the use of these proxies might be justified because it has been shown that these variables are positively related to customer loyalty. East et al. (1995), for example, proved that highly loyal customers spend 32 percent more than other customers. Recently, Reichheld (2003) confirmed the finding that loyal customers spend more money. To our knowledge, however, the relationship between loyalty and length-of-relationship has not been thoroughly researched and, consequently, will be discussed in this study as well.

3. HYPOTHESES

3.1 Comparison of current and new reward criteria

Our introduction casts doubt on the ability of current reward systems to compensate customers in proportion to the benefits they deliver. Consequently, our next step is to evaluate whether the application of another criterion provides a better solution to this shortcoming. More specifically, for the reasons mentioned before, we expect that (behavioral) loyalty represents a better criterion for rewarding customers for their word-of-mouth, price insensitivity and purchase intentions. Therefore, our first hypotheses make an efficiency comparison between loyalty and the criteria currently used. The resulting hypotheses are as follows.

H1a(**b**) *If customers are rewarded for their behavioral loyalty, the rewards go more to customers who exhibit word-of-mouth, price insensitivity, purchase intentions than if customers are rewarded for their spending (**length-of-relationship**).*

3.2 Rewarding loyals according to their predicted loyalty

Even if rewarding based on customers' loyalty proves to be more efficient, it is not straightforward to implement this in a reward program. Individual loyalty scores are not directly available in a company's database (Keiningham et al. 2003), whereas behavioral proxy variables like spending and lifetime duration are. To avoid the measurement of loyalty for each of its customers, we present a model for predicting actual customer loyalty by using

a set of predictors derived from a company's database. However, in order to validate the usefulness of this new measure, we need to be sure that the efficiency gains attributed to rewarding according to loyalty still hold when rewards are distributed according to these predicted loyalty values. Consequently, both previous hypotheses are repeated, but now predicted loyalty is used instead of actual behavioral loyalty.

H2a(**b**) *If customers are rewarded for their predicted behavioral loyalty, the rewards go more to customers who exhibit word-of-mouth, price insensitivity, purchase intentions than if customers are rewarded for their spending (**length-of-relationship**).*

4. METHOD

4.1 Data

We use data from four retail stores belonging to the same large European chain, in two middle-sized towns. While two of the stores carried a product assortment normally associated with *grocery* stores (e.g., food and beverages, cosmetics, laundry detergents, household necessities), two other stores carried an assortment usually associated with *general merchandise* stores (e.g., apparel, electronics and household appliances, do-it-yourself (DIY) and gardening equipment). In the remainder of the study, *Setting G* indicates the assortment usually associated with grocery stores and *Setting M* indicates the assortment usually associated with stores selling general merchandise. This partitioning is maintained throughout this study, in order to validate our findings across the two different store settings.
Using different store settings within a common store chain ensured comparability because databases were structured similarly, and recorded identical information in different store settings. Detailed purchase records were tracked for a period of 51 months and a summarized customer table was available that tracked basic customer demographics as well as first purchase dates. It is important to mention that all transactions could be linked to customers, as the store requires use of a customer identification card.

In addition to these transactional data, a self-administered survey was used as a complementary data collection method. Data collection took place in each of the four retail stores mentioned previously. Surveys were randomly distributed to customers during their shopping trips, and customer identification numbers were recorded for all customers who

received a questionnaire. Respondents were then asked to complete the questionnaire at home and return the survey in a prepaid envelope. Of the 1500 questionnaires distributed in each setting, we received 875 usable responses in *Setting G*, and 779 usable responses in *Setting M*. A usable response had all fields completed, and the respondent could be successfully linked to his or her transaction behavior in the customer database. Hence, we reached ratios of usable response of 58.33% and 51.93% respectively. Given that customer identification numbers were collected for both respondents and nonrespondents, we tested for nonresponse bias by comparing several database variables between customer groups. We found no significant differences between the groups in terms of their spending, frequency of visiting the store, interpurchase time, length-of-relationship and response behavior towards companies' mailings.

4.2 Measures

In this section, we describe the variables we used, and how they were computed, originating either from our survey or from database records.

4.2.1 Survey-related variables.

We measured word-of-mouth, price insensitivity and purchase intentions, based upon Zeithaml et al. (1996), using seven-point Likert-type items. Consistent with previous research on loyalty programs (e.g., De Wulf, Odekerken-Schröder and Iacobucci 2001), we focus on measuring customer share of wallet to represent customer loyalty. Following Sharp and Sharp (1997), reward systems attempt to maximize customers' share of wallet and should be evaluated in terms of the behavioral changes they create. Hence, in this study, customer loyalty was determined as a composite measure by comparing a customer's spending at the retailer with their total spending in the relevant product category. As a first item, and similar to Macintosh and Lockshin (1997), the percentage of purchases made in the focal supermarket chain versus other stores was assessed on an 11-point scale that ranged from 0% to 100% in 10% increments (i.e., 0%, 10%, 20%, and so on). Additionally, two seven-point Likert-type items assessed the shopping frequency of the customers for the focal store when compared to other stores. We pretested the questionnaire several times and refined it on the basis of pretest results. Table 2(a) gives the exact wording of the items used.

4.2.2 Quality of the measurement model.

We initially performed an exploratory factor analysis using the items of the different scales. Several items were deleted, based on substantial cross-loadings. Because of different cross-loadings in both settings, word-of-mouth (WOM) was represented as a two-item scale in *Setting G*, and as a three-item scale in *Setting M*. The other items had a consistent pattern of cross-loadings, resulting in a three-item scale for Loyalty (LOY), a two-item scale for price insensitivity (PRINS), and a single-item scale to measure purchase intentions (PINT). However, because the two items measuring price insensitivity had a significant yet weak correlation (*Setting G*: $R = 0.2846$, $\alpha = 0.4431$; *Setting M*: $R = 0.3061$, $\alpha = 0.4687$), we decided to reduce this scale to a single-item measure. After deletion of these items, we achieved a four-factor structure in which items loaded on *a priori* dimensions.

Table 2. (a) Wording of the items and **(b)** Factor Loadings and Construct Reliabilities

Construct	Item Label	Item Wording
Word-of-mouth	WOM1	Encourage friends and relatives to do business with XYZ.
	WOM2	Say positive things about XYZ to other people.
	WOM3	Recommend XYZ to someone who seeks your advice.
Purchase Intentions	PINT1	Consider XYZ your first choice to buy groceries / general merchandise.
	PINT2	Do more business with XYZ in the next few weeks.
	PINT3	Do less business with XYZ in the next few months (–).
Price Insensitivity	PRINS1	Pay a higher price than competitors charge for the benefits you currently receive from XYZ.
	PRINS2	Take some of your business to a competitor that offers better prices (–).
Loyalty	LOY1	Buy (much less … much more) grocery / general merchandise products at XYZ than at competing stores.
	LOY2	Visit other stores (much less frequently … much more frequently) than XYZ for your grocery / general merchandise shopping (–).
	LOY3	Spend (0% … 100%) of your total spending in grocery / general merchandise shopping at XYZ.

	SETTING G				**SETTING M**			
	LOY	WOM	PINT	PRINS	LOY	WOM	PRINS	PINT
LOY1	**0.895**	0.299	–0.204	–0.173	**0.889**	0.388	–0.115	–0.208
LOY2	**–0.880**	–0.257	0.187	0.218	**–0.842**	–0.268	0.205	0.198
LOY3	**0.898**	0.327	–0.270	–0.198	**0.838**	0.301	–0.161	–0.165
WOM1	–	–	–	–	0.312	**0.868**	–0.118	–0.119
WOM2	0.229	**0.892**	–0.160	–0.055	0.279	**0.818**	–0.089	–0.143
WOM3	0.367	**0.872**	–0.122	–0.111	0.352	**0.858**	–0.130	–0.130
PINT3	–0.249	–0.161	**0.999**	0.102	–0.223	–0.155	**0.999**	0.106
PRINS2	–0.221	–0.092	0.101	**1.000**	–0.192	–0.136	0.105	**1.000**
Variance Explained	2.680	1.851	1.198	1.142	2.587	2.514	1.129	1.171
Cronbach's α	0.871	0.715	–	–	0.818	0.805	–	–
Correlation	–	0.556	–	–	–	–	–	–

We tested construct reliabilities of the scales by means of Cronbach's coefficient alpha. Coefficients of all measures clearly exceed the .7 level recommended by Nunnally (1978). The output of the exploratory factor analysis, in terms of factor loadings and cross-loadings, the variance explained by each factor, and the reliability of the final scales, can be found in Table 2(b).

Table 3. Model Fit Indexes

	SETTING G		**SETTING M**	
	Initial Solution	**Final Solution**	**Initial Solution**	**Final Solution**
χ^2	111.52	14.07	84.85	16.79
d.f.	38	10	38	16
P (> .05)	**.00**	**.17**	**.00**	**.40**
TLI (NNFI) (> .9)	.98	1.00	.98	1.00
SRMR (< .05)	.035	.013	.031	.015
AGFI (> .9)	.96	.99	.97	.99

Table 4. Correlation Matrix of the Independent Variables

	SETTING G				**SETTING M**			
	LOY	WOM	PINT	PRINS	LOY	WOM	PINT	PRINS
WOM	**0.43**	**1.00**			**0.47**	**1.00**		
	11.59				13.21			
PINT	**–0.31**	**–0.19**	1.00		**–0.27**	**–0.19**	**1.00**	
	–8.24	–4.58			–6.62	–4.42		
PRINS	**–0.26**	**–0.13**	**0.13**	**1.00**	**–0.21**	**–0.17**	**0.13**	**1.00**
	–6.88	–3.10	3.08		–5.13	–4.00	3.04	

In addition, a maximum likelihood confirmatory factor analysis (CFA) was performed in LISREL 8.5 to evaluate the quality of the original measurement models. Since the initial solution did not fit the data well, we proceeded to increase model fit by excluding items until the model fits were acceptable. After several iterations, CFA obtained very satisfactory four-factor models for both settings; and the resulting measurement models were identical to the outcome of the exploratory factor analysis reported above. Since we used single-item scales to assess purchase intentions and price insensitivity, we accounted for the fallibility of such a scale by introducing some error variance (20%) during estimation, a procedure suggested by Jöreskog and Sörbom (1993, p. 37). Considering that the measurement models were not significant ($p > 0.05$), that all regression coefficients were statistically significant (smallest t: 14.21, $p < 0.01$), that the correlation between every item and the corresponding latent variable exceeds .50 (smallest $R = .6325$) and given the sufficient construct reliabilities reported above, we have tested our final models successfully in terms of unidimensionality, convergent validity and reliability (Steenkamp and van Trijp 1991). The model solutions are

presented in Table 3, while the correlation matrices of the independent variables are presented in Table 4.

Finally, discriminant validity was examined by evaluating the decrease in performance when fixing correlations among constructs to 1. All chi-square difference tests (1 degree of freedom) were significant ($p < .01$), which indicates that all pairs of constructs correlated at less than one. For example, the high correlation between word-of-mouth and loyalty corresponds to previous findings in the literature (e.g., Reichheld 2003), yet was found to be statistically different from one (*Setting G*: $\Delta\chi^2 = 235.96$, $df = 1$, $p < 0.01$; *Setting M*: $\Delta\chi^2 = 655.77$, $df = 1$, $p < 0.01$).

4.2.3 Database-related variables.

Spending and length of relationship were measured using the company's purchase transaction records. The former variable was computed as the cumulative amount spent by the customer in any of the stores of the focal supermarket chain since the introduction of the current database system. In comparable studies, the computation of length of relationship was complicated by the fact that researchers had to assess whether the customer was still 'alive' (cf. procedures suggested by Schmittlein and Peterson 1994). However, in this setting, all customers who filled in the questionnaire had visited the store during the weeks in which questionnaires were distributed, meaning that all respondents were active customers. This allowed us to compute the length of relationship by simply subtracting the first purchase date for a given customer in the company records from the date of administration of the questionnaire.

4.3 Model

In order to test our hypotheses, we examined the relationship between loyalty benefits delivered and rewards received by the customer. Based on the combination of survey and database information, we are able to compute per customer (i) to what extent the customer delivers each of the benefits usually related to loyal customers, and (ii) the proportion of the rewards received by the customer if this customer was rewarded according to one of the investigated reward criteria. Hence, in this setting, the relationship between loyalty benefits delivered and rewards received is moderated by the reward criterion deployed. Accordingly,

we will adapt a multiple regression framework with interaction effects to investigate our hypotheses (e.g., Cohen and Cohen 1983, Chapter 8). Graphically, we can sketch an exemplary regression model containing interaction effects as in Figure 1.

The given relationship could be captured in the following regression equation:

$$(1) \qquad Y = B_0^i + B_0^s X + B_1^i d_1 + B_1^s d_1 X + e \ ,$$

where Y represents one of the benefits delivered by the loyal customer, X represents the proportion of rewards received by the same customer, parameters with a superscript *i* indicate intercept parameters, and parameters with a superscript *s* indicate slope parameters. The proportion of rewards received is defined by calculating the rewards allocated to an individual customer as a percentage of the company's total rewards allocated.

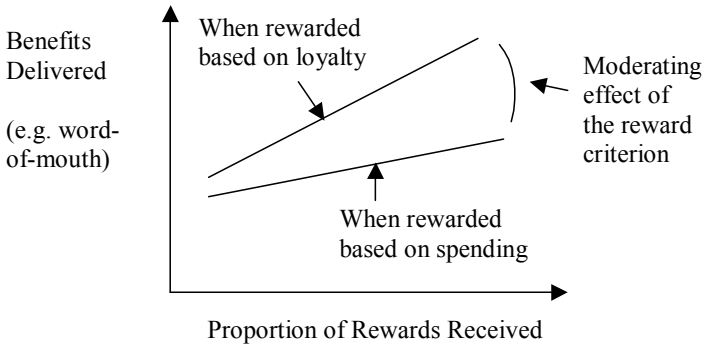


Figure 1: Example of the Moderating Effect of the Reward Criterion on the Relationship between Rewards Received and Benefits Delivered

Furthermore, if we suppose that d_1 represents a dummy variable showing a 0 where customers are rewarded for their spending and a 1 where customers are rewarded for their loyalty, then B_0^s represents the strength of the relationship between the rewards received and the benefits delivered when customers are rewarded for spending, while $B_0^s + B_1^s$ shows the strength of the relationship between the rewards received and the benefits delivered when customers are rewarded for their loyalty. Hence, the test for significance of B_0^s reveals whether customers who deliver benefits (e.g., in terms of word-of-mouth, price insensitivity, or purchase intentions) are rewarded more than others, when all customers are rewarded for

their spending. Accordingly, the test for the significance of B_1^s reveals whether the reward criterion is a significant moderator of the relationship between X and Y, or, in other words, whether the relationship between rewards received and benefits delivered is significantly stronger (or weaker) if customers are rewarded for their loyalty instead of their spending[7]. While the regression equation defined above delivers sufficient information to construct all necessary parameter estimates (and hence the graph given above), not all useful significance tests can be derived from this definition. Indeed, as Cohen and Cohen (1983, p 183) explain, the group that is represented by $d_1 = 0$ functions uniquely as a reference group here, and all the partial coefficients in fact turn upon it, whereby the relationship does not provide us with a test on the significance of the relationship between rewards received and benefits delivered when customers are rewarded for their loyalty. Nevertheless, it is sufficient to adapt the coding scheme, and consider the other possible reward criterion as the reference group, in order to have a different view of the same model. Given this different dummy coding, the significance test of the new parameter B_0^s will reveal whether customers who do deliver benefits are rewarded more than others, when all customers are rewarded for their loyalty.

Supposing that this moderator consists of more than two classes (say, g classes), we will adapt g regression equations to investigate the significance of the g slopes and all interactions between the g groups, where each of the reward criteria serves once as the reference group. Any of these equations—say equation k—can be represented as follows:

$$(2) \qquad Y = B_{0,k}^i + B_{0,k}^s X + \sum_{j=1}^{g-1} (B_{j,k}^i d_{j,k} + B_{j,k}^s d_{j,k} X) + e \ ,$$

where k ranges from 1 to g. Adding to the previous example, supposing we also wish to evaluate the strength of the relationship where customers are rewarded for their length of relationship or their predicted loyalty, then the moderating variable consists of four (or more formally, g) groups, that can be represented by three ($g - 1$) dichotomies, d_1, d_2 and d_3, covering the three possible reward criteria (e.g., $d_1 = d_2 = d_3 = 0$: spending; $d_1 = 1$, $d_2 = d_3 = 0$: loyalty; $d_1 = 0$, $d_2 = 1$, $d_3 = 0$: length of relationship; $d_1 = d_2 = 0$, $d_3 = 1$: predicted loyalty).

[7] Note that interpretation of the intercept parameters is similar, but is of less relevance to our research topic.

This procedure is in accordance with procedures discussed by Cohen and Cohen (1983, chapters 5 and 8) for conducting this type of analysis, and carefully considers the pitfalls indicated by Irwin and McClelland (2001) when interpreting the results of moderated multiple regression models.

4.4 Predicting loyalty

In order to make use of our conceptual model, marketing management needs to be able to define customers' loyalty. Nevertheless, share of purchases cannot be derived directly from the information in a database, so in a real environment, a predictive model is needed. This section describes how the model is built. The variables, predictive technique, validation method and variable-selection procedure are discussed in the following paragraphs.

4.4.1 Variables.

We only used information that is available in the customer database at the individual customer level. These data are collected by the use of a loyalty card. The dependent variable in the model is loyalty, which is measured by a construct of the three above-mentioned questionnaire items. In total, 33 independent variables were compiled to predict loyalty in the general merchandise store setting and 34 independent variables were computed for the grocery setting. Table 5 summarizes all these variables, together with a brief description of how they are calculated. The results of the model are included in this table and discussed in a later section. It shows that we used more or less the same predictors in both shopping environments. We will, therefore, be able to compare the relevant information for the two settings. The following paragraphs give a short overview of the variables that are taken into account.

Reinartz and Kumar (2002) argue extensively for the inclusion of several predictors in their lifetime duration model. Since their variables are also intended to explain the strength of a relationship, our variable list will be similar. As a consequence, we will not discuss the same literature in detail. First, we focus on variables that are commonly used in scoring models for customer relationship management (Bult and Wansbeek, 1995). The level of customer spending and the frequency of customers' visits prove to be efficient behavioral information for the detection of weak or strong relations. Consequently, we include customers' individual

spending and visit frequency derived from data concerning the last month, six months, one year, two years and over our complete data time series. The average spending and customers' spending relative to the length of time since their first purchase are computed to take into account relative figures as well. Related variables in this area are the number of products bought and the amount of money spent on fresh products that need to be weighed by the customers themselves. This last information was only relevant for the grocery setting. Furthermore, we also include the average interpurchase time and the time since the customer's last purchase. All these variables are frequently used to determine loyal customers and to characterize customers who exhibit strong relations with a company (Reinartz and Kumar 2002). Moreover, we include the standard deviation of the interpurchase time as this gives insight into the regularity of customers' visits and turns out to be an important variable for predicting future loyalty (Buckinx and Van den Poel 2005). Some studies support the relation between customers' lifetime and their profitability, while others questioned these results (Reinartz and Kumar 2000). Therefore, we incorporate the length of relationship into our model. Reinartz and Kumar (2002) also incorporate the scope of customers' purchases into their predictive model. Likewise, Baesens et al. (2004) recently showed the variety of products purchased to be a predictor of future spending increases or decreases. Thus, the number of categories from which a customer bought products is included in our model. We summed the same behavior of customers during their previous one, two and three years. Returns of goods can be important information too, though the hypothesis of Reinartz and Kumar (2002) concerning this behavior was not supported. Returns may be a signal for dissatisfaction and consequently a weaker relationship. In contrast, for some products, it is shown that returns signal a positive association with customer loyalty (Buckinx and Van den Poel 2005). We include the total amount of returned goods and two dummies: whether or not a customer ever returned a product or cancelled an order. As earlier in our study, we assume that loyalty is related to price insensitivity (Dowling and Uncles 1997; Srinivasan, Anderson and Ponnavolu 2002). Consequently, we try to derive which customers behave like promotion seekers by computing four promotion-related variables: the number of promoted products bought, the money spent on promotions, the number of visits where at least one promoted product was purchased and, finally, the percentage of products purchased on promotion. The next types of information that we presume to have explanatory power for customer loyalty are variables related to customers' response to mailing actions. Though neither the company from the grocery setting nor the

Table 5. Description and standardized Parameter estimates of Variables Used for Predicting Loyalty

Variable	Description	Grocery Shopping					General Merchandise				
		Multivariate		Univariate		Ranking	Multivariate		Univariate		Ranking
Spending_1M	Spending during last month.			0.35396	***	21			0.12996	***	28
Spending_6M	Spending during last six months.			0.45817	***	11			0.24760	***	18
Spending_1Y	Spending during last year.			0.47892	***	4			0.29260	***	8
Spending_2Y	Spending during last two years.			0.47424	***	6	0.23464	**	0.30156	***	6
Spending	Spending in total history.			0.47144	***	8	–0.20188	*	0.27243	***	13
Frequency_1M	Number of purchases during last month.			0.34773	***	22	–0.07706		0.16882	***	24
Frequency_6M	Number of purchases during last six months.			0.43562	***	19			0.27715	***	12
Frequency_1Y	Number of purchases during last year.			0.44546	***	16	0.26789	*	0.29420	***	7
Frequency_2Y	Number of purchases during last two years.			0.44943	***	14	–0.36123	**	0.28686	***	10
Frequency	Number of purchases in total history.			0.43889	***	18			0.26861	***	15
NumItems	Number of product items bought.	0.09898	**	0.47046	***	9	–0.11047	*	0.16022	***	26
Spending_Weight °	Spending in products that need to be weighted by the customer.	0.10086	**	0.43953	***	17					
rSpend_Freq	Average Spending per visit.	0.07685	**	0.17850	***	30	0.05450		0.00365		33
rSpend_lor	Spending relative to the length of the customer's relationship.			0.47263	***	7			0.28188	***	11
Recency	Number of days since last purchase.			–0.20352	***	29			–0.13204	***	27
Ipt	Average number of days between store visits.			–0.29652	***	25			–0.20427	***	21
Std_Ipt	Standard deviation of the number of days between the purchases.	–0.05203		–0.32269	***	23	–0.10788	***	–0.24934	***	17
Lor	Length of customer relationship.			0.09398	***	33			0.06881	*	29
Numcat_LY	Number of different product categories purchased from during last year.	0.21724	***	0.52211	***	1			0.33452	***	2
Numcat_2Y	Number of different product categories purchased from during last two years.			0.47699	***	5			0.30311	***	4
Numcat_3Y	Number of different product categories purchased from during last three years.			0.44598	***	15			0.25404	***	16
Numcat	Number of different product categories purchased from during the total history.			0.48046	***	2			0.31897	***	3
Neg_Inv	Dummy to indicate if the customer ever had a negative invoice (1/0).	0.03956		0.29186	***	27			0.18993	***	22
Ret_Item	Dummy to indicate if the customer ever returned an item (1/0).			0.26563	***	28			0.17951	***	23
Returns	Total value of returned goods.			0.15719	***	31	–0.06357	*	–0.00814		32
NumPromItems	Number of items bought that appeared in company's promotion leaflet.			0.45390	***	13	0.13010	**	0.22736	***	19
SpenPromItems	Money spent on products that appeared in promotion leaflet.			0.45724	***	12			0.27082	***	14
Visitspromitems	Number of visits on which a product is bought that appeared in the promotion leaflet.			0.46804	***	10	0.11218		0.30203	***	5
PercNumPromItems	Percentage of products bought that appeared in leaflet.			0.01389		34			0.05476		30
PercResp_Leaf	Percentage of times a purchase is made given that a promotion leaflet was received.	0.13667	**	0.47915	***	3	0.26732	***	0.34265	***	1
PercResp_Noleaf	Percentage of times a purchase is made given that no promotion leaflet was received.			0.30984	***	24			0.16728	***	25
MoreThanOnce	Number of times that a customer visits more than once within the same promotion period.			0.43077	***	20			0.29253	***	9
PercMoreThanOnce	MoreThanOnce divided by the number of times a customer bought in a promotion period.			0.29400	***	26	0.06032		0.20873	***	20
Distance	Distance to the store.			–0.12651	***	32			–0.03885		31

° This variable was only included for the grocery setting and not in the general merchandise store setting.

* p <.10, ** p <.05, *** p <.01.

general merchandise store is active in direct marketing, their most important communication channel is a biweekly leaflet. Therefore, for each of the customers, we incorporate the percentage of occasions the customer made a visit to the store after having received the leaflet. Because of limited budgets, not all customers receive a catalogue each week. Therefore, we included the percentage of times a customer came to the store even though he or she had not received a catalogue. We assume a positive relation between the number of times someone visits the store during one and the same promotion period[8] and loyalty. Finally, the strength of a relationship is likely to depend on the costs and benefits experienced. By including the distance between the store and the customers' residence, we test for the influence of living far from or close to the shop.

4.4.2 Predictive technique and leave-one-out procedure.

In order to predict customers' loyalty, we apply a multiple linear regression model. We will evaluate the predictive power of this model on a validation set that is independent of the information used to build the model. However, the limited number of observations in each of the two settings and the elaborate number of independent variables make it hard to split our data in an estimation and a hold-out test set. As a consequence, we prefer a resampling method called leave-one-out cross-validation because it proves to be superior for small data sets (Goutte 1997) and at the same time assures the use of a rigorous predictive validity test. Using this procedure, our data are divided into k subsets, where k is equal to the total number of observations. Next, each of the subsets is left out once from the estimation set and is then used to assess a validation score. To get an idea of the power of the model, the final test set is built by stacking together the k resulting validations. The performance of the model is evaluated by the adjusted R^2 and the MSE—on the estimation set as on the validation set.

4.4.3 Variable selection.

Considering the number of variables and the rather limited number of observations, we make use of a variable-selection technique. Thanks to this method, the dimensionality of the model can be reduced and redundant variables are removed, which is in favor of the performance of the model. In order to guarantee the selection of the best subset, we apply the leaps-and-

[8] A promotion period is the period where the offers of one catalogue are valid.

bounds algorithm proposed by Furnival and Wilson (1974). Their efficient technique identifies the model with the largest adjusted R^2 for each number of variables and at the same time avoids a full search of the variable space. The best subset is chosen based on the adjusted R^2 that can be achieved on the total estimation set.

5. RESULTS

5.1 Introduction

Following Irwin and McClelland (2001), we report the detailed coding scheme used in this research. This coding scheme is represented in Table 6, indicating that loyalty (LOY) was considered as the reference group in the first coding iteration, next Spending (SPEN), Length of Relationship (LOR), and finally Predicted loyalty (PLOY).

Table 6. Coding and Recoding of the Interaction Dummies (Dummy-Variable Coding)

	$r = 1$			$r = 2$			$r = 3$			$r = 4$		
	$d_{1,1}$	$d_{2,1}$	$d_{3,1}$	$d_{1,2}$	$d_{2,2}$	$d_{3,2}$	$d_{1,3}$	$d_{2,3}$	$d_{3,3}$	$d_{1,4}$	$d_{2,4}$	$d_{3,4}$
LOY	0	0	0	0	0	1	0	1	0	1	0	0
SPEN	1	0	0	0	0	0	0	0	1	0	1	0
LOR	0	1	0	1	0	0	0	0	0	0	0	1
PLOY	0	0	1	0	1	0	1	0	0	0	0	0

Since we are interested in the slope parameters in equation (2), they can be summarized as in Table 7(a), where the diagonal represents the slopes of the different relationships, and the off-diagonal figures represent the differences between the slopes. For example, if loyalty is considered as the reference group ($r = 1$), then the relationship between the benefits and the rewards—if customers are rewarded proportionally for their loyalty —can be represented as $B^s_{0,1}$, while the difference between rewarding for spending versus rewarding for loyalty can be represented as $B^s_{1,1}$. The corresponding standard estimate of this parameter allows us to interpret whether this difference is significant. Because these differences are symmetric, all information below the diagonal is redundant and will not be repeated. In Table 7(b), we give an overview of all parameters and their standard errors for the different regression equations. The relationships are also represented graphically in the Appendix.

Besides the use of the results to evaluate our hypotheses, the information in Table 7(b) can be used to perform a validity check on the literature described in the introductory section of this study. In the discourse about loyalty benefits, we posited that loyal customers deliver a number of benefits to the company. In order to validate this crucial finding of previous research, we consider the parameters $B^s_{0,1}$ of the different models. When inspecting the results in Table 7(b), it is clear that these relationships are highly significant.

Table 7. (a) Interpreting (Re)Coded Parameter Estimates and (b) Results of Model Estimation

	B_{LOY}	B_{SPEN}	B_{LOR}	B_{PLOY}
B_{LOY}	$\boldsymbol{B^s_{0,1}}$	$B^s_{1,1} = -B^s_{3,2}$	$B^s_{2,1} = -B^s_{2,3}$	$B^s_{3,1} = -B^s_{1,4}$
B_{SPEN}	$B^s_{3,2} = -B^s_{1,1}$	$\boldsymbol{B^s_{0,2}}$	$B^s_{1,2} = -B^s_{3,3}$	$B^s_{2,2} = -B^s_{2,4}$
B_{LOR}	$B^s_{2,3} = -B^s_{2,1}$	$B^s_{3,3} = -B^s_{1,2}$	$\boldsymbol{B^s_{0,3}}$	$B^s_{1,3} = -B^s_{3,4}$
B_{PLOY}	$B^s_{1,4} = -B^s_{3,1}$	$B^s_{2,4} = -B^s_{2,2}$	$B^s_{3,4} = -B^s_{1,3}$	$\boldsymbol{B^s_{0,4}}$

	Setting G				**Setting M**			
	Parameter Estimates (Standard Error)				**Parameter Estimates (Standard Error)**			
	B_{LOY}	B_{SPEN}	B_{LOR}	B_{PLOY}	B_{LOY}	B_{SPEN}	B_{LOR}	B_{PLOY}
Word-of-mouth								
B_{LOY}	**462.54 (45.07)*** **	–413.18 (49.64)***	–424.46 (61.51)***	–224.14 (74.13)***	**502.55 (47.05)*** **	–455.79 (53.01)***	–532.72 (59.65)***	–262.18 (91.61)***
B_{SPEN}		**49.36 (20.8)** **	–11.28 (46.74)	189.04 (62.42)***		**46.76 (24.43)***	–76.93 (44.06)*	193.61 (82.31)**
B_{LOR}			**38.08 (41.85)**	200.32 (72.21)***			**–30.17 (36.67)**	270.54 (86.73)***
B_{PLOY}				**238.39 (58.85)*** **				**240.37 (78.6)*** **
Price Insensitivity								
B_{LOY}	**343.17 (51.51)*** **	–234.01 (56.81)***	–310.73 (70.26)***	–111.04 (84.79)	**306.12 (56.29)*** **	–286.52 (63.43)***	–242.21 (71.37)***	–204.98 (109.6)*
B_{SPEN}		**109.16 (23.96)*** **	–76.73 (53.45)	122.96 (71.49)*		**19.6 (29.22)**	44.31 (52.71)	81.54 (98.47)
B_{LOR}			**32.43 (47.78)**	199.69 (82.57)**			**63.91 (43.87)**	37.23 (103.77)
B_{PLOY}				**232.13 (67.35)*** **				**101.14 (94.04)**
Purchase Intentions								
B_{LOY}	**397.47 (51)*** **	–272.49 (56.17)***	–342.17 (69.6)***	47.15 (83.87)	**355.3 (55.9)*** **	–249.9 (62.98)***	–381.19 (70.86)***	102.95 (108.83)
B_{SPEN}		**124.97 (23.54)*** **	–69.67 (52.88)	319.64 (70.62)***		**105.41 (29.02)*** **	–131.29 (52.34)**	352.84 (97.78)***
B_{LOR}			**55.3 (47.36)**	389.31 (81.71)***			**–25.89 (43.56)**	484.14 (103.04)***
B_{PLOY}				**444.61 (66.58)*** **				**458.25 (93.38)*** **

For example, the relationship between rewards received if customers would be rewarded based on their behavioral loyalty and word-of-mouth in *Setting G* is positive and significant ($B = 462.54$, $p < 0.0001$). By analogy, we can investigate the other parameters, and we conclude that if customers are rewarded for their loyalty, the rewards would be distributed to customers who engage more in word-of-mouth, are less price sensitive, and exhibit higher purchase intentions, in both settings.

Next, as discussed previously, spending and length-of-relationship are commonly used proxies for behavioral loyalty in general, and because they are more readily available to the company, they are commonly used as reward criteria. Indeed, the analysis of the correlations between loyalty and both proxies suggests a strong significant correlation between loyalty and spending in both settings (*Setting G*: $R = 0.4714$, $p < 0.0001$; *Setting M*: $R = 0.2724$, $p < 0.0001$). The correlation between length-of-relationship and loyalty, however, proves to hold in the setting of grocery shopping ($R = 0.1150$, $p = 0.0006$), but not in the setting related to general merchandise shopping ($R = 0.0393$, $p = 0.2722$).

Likewise, since both spending and length-of-relationship have been used previously as a reward criterion, we examine whether customers who are rewarded for these also deliver the benefits related to loyal customers. Because the results are more ambiguous, we will discuss this relationship for each benefit separately. If customers are rewarded for their spending, the evidence is only moderate that these customers would also deliver more word-of-mouth to the company (*Setting G*: $B = 49.36$, $p = 0.0177$; *Setting M*: $B = 46.76$, $p = 0.0557$). Apparently, this relationship is more pronounced for grocery shopping than general merchandise. This effect is comparable to the effect of the same reward criterion on price sensitivity. If customers are rewarded for their spending, rewards would be distributed significantly more to price insensitive shoppers in the grocery setting ($B = 109.16$, $p < 0.0001$), while no such significant relationship is detected for general merchandise ($B = 19.6$, $p = 0.5024$). Accordingly, those customers rewarded for their previous spending would be customers showing significantly higher purchase intentions towards the store. This effect is consistent in both settings (*Setting G*: $B = 124.97$, $p < 0.0001$; *Setting M*: $B = 105.41$, $p = 0.0003$). If customers are rewarded for their length-of-relationship, the relationships between rewards received and benefits delivered are unambiguous. None of these relationships is significant (significance ranging between $p = 0.1453$ and $p = 0.5524$).

5.2 Hypothesis tests

In order to validate H_{1a} and H_{1b}, we test whether the slope of the curve based on loyalty is significantly higher than the slope of the curves based on spending or length-of-relationship. It is important to notice here that this difference was highly significant in all of the cases ($p < 0.001$ in all cases). Hence, the relationship between the proportion of rewards received and each of the benefits related to loyal customers was significantly higher when customers were rewarded for their loyalty instead of their spending or length-of-relationship.

Finally, in order to test the applicability of a reward scheme based on loyalty, H_{2a} and H_{2b} test the relationship between rewards received and benefits delivered if the reward criterion was predicted loyalty instead of spending or length of relationship. Because the results are again more ambiguous, we will describe the effect per benefit delivered. First, the relationship between rewards received and word-of-mouth delivered by customers is significantly higher if customers are rewarded for their predicted loyalty than if they are rewarded for their spending or length-of-relationship (significance ranging between $p = 0.0187$ and $p = 0.0018$). Second, considering price insensitivity, the results are conditional upon the setting: while there is a marginally significant effect in grocery shopping (PLOY vs SPEN: $B = 122.96$, $p = 0.0855$; PLOY vs LOR: $B = 199.69$, $p = 0.0156$), the effect in general merchandise shopping is clearly insignificant (PLOY vs SPEN: $B = 81.54$, $p = 0.4077$; PLOY vs LOR: $B = 37.23$, $p = 0.7198$). Finally, considering purchase intentions, the results across the two settings are again generally consistent: if customers are rewarded for their predicted loyalty, those customers with higher purchase intentions will be rewarded significantly more than if they were to be rewarded for their spending or length-of-relationship ($p < 0.001$ in all cases).

5.3 Predicting loyalty

In this section, we describe the performance of the multiple linear regression model used to predict loyalty. In Table 8, the performance of the models with all variables—the 'full model'—is compared with the performance of the best performing models in terms of adjusted R^2 and the MSE. We evaluate both the performance of a model where all observations are used for estimation purposes—hence called the 'estimation set'—with a model where the leave-one-out procedure is used to evaluate the real performance of the model. All models are significant considering a significance level smaller than 0.0001.

As could be expected, the leave-one-out performance decreases slightly compared to the estimation set performance. Additionally, the difference between both performance measures decreases when fewer variables are used in the model; indicating that the variable-selection procedure tempers the negative consequences related to overtraining. Finally, predictive performance increases with the use of a variable selection technique, indicating the usefulness of such a procedure for the prediction of loyalty.

Table 8. Model Performance after Variable Selection Procedure

	Setting G				**Setting M**			
	Full Model (v = 35)		Final Model (v = 7)		Full Model (v = 34)		Final Model (v = 13)	
	Estimation Set	Leave-one-out	Estimation Set	Leave-one-out	Estimation Set	Leave-one-out	Estimation Set	Leave-one-out
$R^2_{adjusted}$	0.29256	0.23007	0.30632	0.29416	0.12422	0.04401	0.14119	0.10354
MSE	0.55856	0.61074	0.54770	0.55741	0.63946	0.70856	0.62707	0.65675

Obviously, the most important benefit of the variable-selection procedure lies in detecting a parsimonious subset of database variables that can be used to predict loyalty in both store settings. Remarkably, there is a considerable difference in the number of variables selected in each case. In the grocery setting only 7 of the 34 variables are retained, whereas for general merchandise stores more information is needed: the maximum adjusted R^2 was reached with 13 predictors. Table 5 shows the standardized parameter estimates and the significance levels for the variables that are chosen by the feature selection procedure. We represent the multivariate solutions as well as the univariate results of each individual variable since there is clear evidence of multicollinearity in the multivariate model[9]. For the same reason, we also represent the univariate standardized parameter estimates from variables that were not selected for the final model. While the univariate results should be used for interpretation of the signs and significance of the variables, the multivariate solution delivers the best fit to the data, and hence offers the best prediction of loyalty.

In order to detect whether different variables are important in the different settings, we investigated the Spearman rank-order correlation, which is a nonparametric measure of association based on the rank of the data values. Given the very large and significant correlation of 0.8915 ($p < 0.0001$), we conclude that the importance of the variables does not

[9] For example, several variables that are univariately highly significant are not selected or turn out to be insignificant in the multivariate model.

differ significantly between the two settings. In order to enhance comparability, we included the ranking of the variables in Table 5. However, considering the multicollinearity we discussed previously, the final predictive models in each setting differ considerably in the variables used. As discussed previously, this should not lead the reader to conclude that different variables are needed to predict loyalty in the different settings. The final model for each store setting is shown in Table 5. The importance of each of the variable types for our predictive model is examined in the next section.

6. DISCUSSION

6.1 Loyalty benefits

Previous empirical research, as well as anecdotal evidence, has focused on the relationship between loyal customers and the alleged beneficial characteristics of such loyal customers. However, considering the conflicting results of these studies, decisive conclusions are lacking. Our research, however, confirms the existence of benefits from loyal customers by examining the relationship between loyalty and three different benefits. Customers who spend an important proportion of their total budget in only one company, actively recommend its services to their peers. Besides, these customers are price insensitive and are motivated to repurchase from the focal company in the future. Our findings confirm the results of Reynolds and Arnold (2000) and Srinivasan, Anderson and Ponnavolu (2002), who investigated these associations in an online environment. However, they counter the conclusions of Reinartz and Kumar (2000), who could find no support for any of these benefits although both their and our studies focused on a noncontractual setting. What can be the reason for these mixed results? A credible explanation is the way in which loyalty was approached in each of the studies. When considering all empirical evidence, only Reinartz and Kumar (2000) reject any connection between loyalty and loyalty benefits. Table 1 shows that theirs is the only study to examine lifetime duration, while others took behavioral or attitudinal loyalty into account. This might indicate that the conclusions depend on which criterion is used. Indeed, our study agrees with this reasoning, since a significant relationship between customer lifetime and one of the three benefits examined was not detected. This confirms our assumption that the way in which loyalty is approached drives the studies' conclusions.

6.2 Loyalty outperforms behavioral proxies as reward criterion

This study is the first to question the criteria that are widely used by companies to manage their reward system. Currently, most companies use a reward system where compensations are dependent on customers' spending behavior. Past research concerning human behavior has shown that rewards will motivate customers to do what is necessary to get the related returns (Nicholls 1989). Our results show that if companies want to reward customers for more than only repeat-purchase behavior, they are well advised to take into account customers' (predicted) loyalty rather than relying on spending or customers' lifetime. This implies that companies that stay dedicated to their current reward strategy are neglecting customers who turn out to be beneficial. These customers positively distinguish themselves from other customers because they actively spread positive word-of-mouth about a company, are willing to pay a superior price and have clear positive intentions to visit the store in the future. Current reward schemes do not compensate for these contributions, while these benefits are extremely valuable for growth, profitability and continuity of a company.

Customers' referrals are very influential in decision-making processes since they seem to be reliable sources of information. Reichheld (2003) emphasizes this reasoning in his last study: "The only path to profitability and growth may lie in a company's ability to get its loyal customers to become its marketing department." Customers who recommend a company to their friends and relatives help to avoid leakage from the customer base (Jones and Sasser 1995). In their recent study, Wangenheim and Bayon (forthcoming) provide evidence that positive word-of-mouth referrals can convince up to 16% of the recipients to switch to the 'advertised' company in a consumer market, and as much as 51% in an industrial market, provided that the source is considered experienced and similar to the receiver. Reichheld (2003) warns of a bad mix of promoters and detractors: the percentage of customers who are promoters has a strong relation with a company's growth. The habit of loyal customers of bringing in new customers is particularly valuable, particularly if the company is competing in a mature market. The second benefit of loyal customers can have direct impact on companies' profits: less price-sensitive customers are indifferent about paying more for the same product/service. As a result, it is not necessary to convince these customers by offering them price cuts and discounts. This means that these customers do not come to a store merely to pick all the 'cherries' but buy products that generate higher margins as well. Finally, customers' purchase intentions guarantee companies' continuity. Bolton et al. (2000) found that purchase intentions do have a strong positive relationship with subsequent

repatronage decisions and consequently with retention behavior. This makes them interesting since they assure a steady stream of resources to the company.

The previous paragraph emphasizes the value of the different benefits. In contrast, loyal customers will be discriminated against by companies that apply traditional reward programs. There is a danger that this strategy might motivate loyal customers to leave a company. Feinberg et al. (2002) demonstrate that customers will prefer their favorite firm less when they are put at a disadvantage compared to nonloyal customers—and which company likes to lose customers who deliver substantial benefits? Even worse: promoters of the company can become detractors who will substitute their former recommendations into negative word-of-mouth (Reichheld 2003) that will damage a firm's reputation. Our results suggest that programs that apply (predicted) loyalty as a reward criterion are able to give more rewards to customers with diverse loyalty benefits and less rewards to customers having no loyalty benefits. As such, they would compensate customers more effectively for their beneficial behavior, and consequently, such programs are expected to induce a higher retention rate. Customers who experience appreciation for their contribution and feel recognized in a reward program will weigh comparisons with competitors less heavily in making purchase decisions (Bolton et al. 2000).

Hallberg (2004) reports that the success of companies' reward systems is not only dependent on results that have an immediate financial impact. The extent to which these reward systems attach customers emotionally to a brand or a store is as important. The newly proposed reward criterion in this study will focus management's attention on different types of benefit.

6.3 Effect of reward programs

In addition to marketing research on the profitability of loyal customers, a number of other studies have concentrated on the effects of reward programs on customer behavior. A literature review confirms Dowling and Uncles' (1997) theory that it is hard to influence customer behavior with the current reward schemes. The limited number of studies investigating this topic shows diverse effects of reward programs on behavioral loyalty. Mägi (2003) investigated the effect of loyalty card programs on share of purchases in a grocery shopping environment. Her results confirm the mixed results and suggest that at the

store level, no effect must be expected on the share of purchases. The conclusions of Verhoef (2003) indicated a marginal effect of relationship marketing instruments (RMI) on share development. Even more importantly, the outcomes revealed that loyalty programs' effect was, for the most part, explained by past customer behavior: "Customers with a small (past) customer share are more likely to increase their customer share in the next period." These findings emphasize the need for a reward criterion such as the one we propose in this study. More specifically, Verhoef (2003) investigates the impact of a reward program on the change in share of purchases. However, as for most companies, this study included a reward system that distributed price discounts based on the level of purchases and the length-of-relationship. Such schemes do not take into account a customer's behavioral loyalty, which offers a potential explanation for their marginal effect. Customers exhibiting an already high level of loyalty are not likely to increase their spending, since they already make all their purchases in a particular store. This is supported by the conclusions of Verhoef (2003) on the importance of the initial customer share in explaining the (small) effect (see above). In general, the mixed effects of relationship programs might be explained by this phenomenon. Selection criteria, which define the level of incentives or rewards, should be in accordance with the goals of the marketing program. On that reasoning, spending as a reward criterion to increase customers' behavioral loyalty is not the best option. Instead, making use of (predicted) loyalty to manage reward programs, as suggested in this study, seems a valid solution. Other studies that value customer loyalty for marketing action purposes are those of Dowling and Uncles (1997) and Reinartz and Kumar (2002). Though these last authors examine the value of a lifetime duration framework, their managerial implications emphasize the need for loyalty, measured by share of wallet, to fine tune companies' actions and to deal with different types of customers. Nevertheless, they did not empirically check the advantages related to that proposition, nor did they offer a model to define share of wallet for the total customer base. Therefore, ours is virtually the first study to show empirically the importance of using loyalty in a reward system and to propose a feasible solution that incorporates individual customer loyalty into a relationship-marketing program.

6.4 Model results

The outcomes from the predictive loyalty models lead to the following contributions. First, the significance of the overall predictive models in both settings points to the ability of marketing management to compute a customers' loyalty to an acceptable extent from his or

her transactional data. Without this feature, a company is forced to send out questionnaires to all of its customers in order to know their exact loyalty. Using the method presented above, however, it is sufficient to interrogate a limited number of randomly chosen customers from the database. In this model, we only incorporated data that can be derived directly from the customer database and that is available for all customers thanks to their customer identification cards. This enables companies to create a loyalty score for every customer at any given moment. Given the satisfactory predictive performances of our models, efficiency in rewarding customer benefits validates the usefulness of our new proxy measurement. The results confirm the findings concerning actual customer loyalty: rewarding in accordance with predicted loyalty is significantly better than rewarding in proportion to commonly used proxy variables (see previous paragraph).

Second, the difference in predictive ability between the two store environments is remarkable. Apparently, it is more complex to define loyalty in a general merchandise shopping environment than in a grocery shopping environment. While it is very likely that these differences can be explained by different purchasing patterns in both settings, more research is required to investigate and explain these differences.

As mentioned above, our feature selection procedure proved to be useful for the prediction of loyalty since the multiple regression models achieved an increased performance with fewer predictive variables. In order to draw conclusions on which kind of data explains loyalty, we focus on the univariate models' standardized parameter estimates for each of the predictors. Both store settings are very comparable in terms of the ranking of the explanatory variables, which suggests that our results may be generalizable to different, yet similar, store settings. Nearly all variables feature a significant influence that confirms the findings of Verhoef (2003), that past customer behavior explains most of customer share development. Intriguingly, the most valuable customer information for defining loyalty is the variety of products purchased and responsiveness to direct mails. These variables can be detected within the top three predictors in both settings (the number of different product categories purchased during last year, during the customer's total length-of-relationship and the percentage of times a purchase is made given that a leaflet was received). Our study is the first to show the great importance of this type of customer information when explaining loyalty. In previous research, purchase depth (captured in variables such as the frequency and monetary value of previous purchases) has received more attention than purchase width

(i.e., the purchase variety). However, our findings suggest that the predictive capacity of the latter type of information should not be neglected. Indeed, the more a customer is interested in purchasing a large variety of product categories, the stronger the relationship with the company and hence the more loyal the customer. This conclusion is consistent with the importance of this type of variable for predicting the strength of the customer's relationship and future developments in this relationship (Baesens et al. 2004). Representing another important predictor, the degree of response to leaflets is a signal of loyal customer behavior. This means that the level of past interest someone has shown in a company's communication is related to the fraction of that customer's total household budget that he or she spends at that company. Remarkably, variables related to customer spending or length-of-relationship are not found to be the best predictors, despite these being widely used in companies' reward schemes. The former type of variable shows up in the top 10 importance ranking. Their significance validates much past research that already suggested a relationship between loyalty and customers' spending level (Reichheld 2003). Moreover, buying more promotional products seems to be an indicator of increased loyalty. An explanation for this surprising relation is that these variables correlate highly with the number of items bought and the frequency of visits. Customers who buy more items are expected to exhibit a higher absolute level of promotional purchases as well. Therefore, the parameter estimates of the multiple regression models are biased because of multicollinearity, and the univariate outcomes are driven by the number of items[10] and visit frequency[11] and not by the promotional nature of the products. This is supported by the insignificance of the percentage of promotional products bought (PercNumPromItems) in both settings. Furthermore, information concerning customers' last purchase date and the time between their purchases are significant in our models. The standard deviation of the time between customers' purchases also explains loyalty. The effect suggests that regular customers, who show a low standard deviation, are more loyal to the store. This finding is in line with the loyalty definition of Buckinx and Van den Poel (2005), who incorporated this standard deviation to distinguish loyals from nonloyals. To our knowledge, this is the first study to confirm empirically the value of this behavior for classifying customers in accordance with the

[10] Pearson Correlation Coefficients between 'Numitems' and 'NumPromItems': Grocery shopping .84 ($p < 0.01$); General merchandising .80 ($p < 0.01$).

[11] Pearson Correlation Coefficients between 'Frequency' and 'VisitsPromItems': Grocery shopping .96 ($p < 0.01$); General merchandising .93 ($p < 0.01$).

strength of their relationship. Surprisingly, the length of customers' relationship is ranked at the bottom of the results. Moreover, in the general merchandise store setting, only a marginal effect can be found. This supports the findings of Reinartz and Kumar (2000), who doubt the value of lifetime duration for the characterization of valuable customers. Furthermore, the distance to the store is of minor importance for loyalty.

Perhaps the most notable conclusion from this overview is that customers' spending, frequency and lifetime are not the only sources of information to explain loyalty. This study shows the importance of other behavior when classifying customers according to their loyalty. These findings point to the limited ability of currently used criteria to approximate customer loyalty. The significant explanatory power of just about all variable types explains why our predicted loyalty measure is more efficient in rewarding loyalty benefits than spending and lifetime. The more relevant customer behavior is taken into account, the better loyalty can be approximated and the better the benefits related to loyal customers can be rewarded.

6.5 Limitations and directions for further research

As in any other study, this study has its limitations and encourages further research on the issue and related topics.

First, although we validated this study in two different store settings, we cannot claim that our findings can be generalized to all environments. The results show small differences between the store formats considered: some hypotheses that are supported in the grocery setting are not supported, or only partially supported, in the general merchandise setting. Therefore, further research is needed in order to confirm our results in other industries—not necessarily restricted to consumer markets.

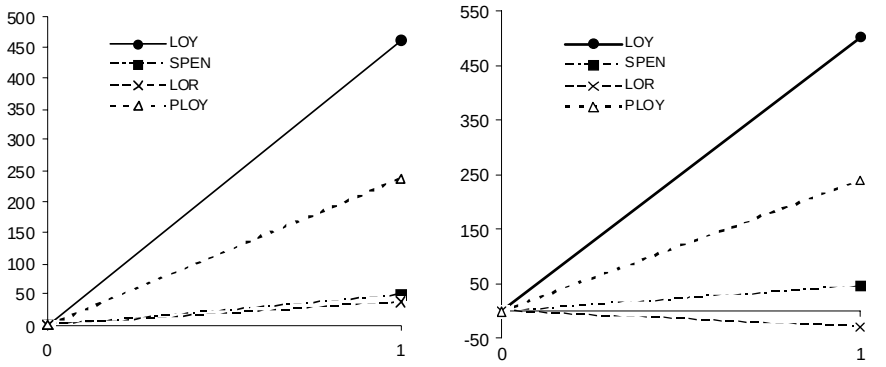
Second, our predictive model included little demographic customer information to explain loyalty. Only the customers' distance to the store was incorporated. Since the European store chain that provided the data does not collect this type of information when customers register, no social demographics were at our disposal for the predictive model. Therefore, the predictive ability of our models might even increase when demographics are available from the company's internal data files.

Third, in this study, we provide evidence that behavioral loyalty can be predicted from the company's internal data records to an extent where it provides a more efficient criterion for rewarding loyalty benefits than spending or length of relationship. Hence, we have only shown that it is feasible to reward customers for their loyalty, and that the currently designed reward schemes do not fully reward loyalty. Indeed, in the present study, we were unable to test the effect of rewarding customers based on different reward criteria in the field. To this end, an economic decision about the most appropriate reward criterion would have to reside on a full cost–benefit analysis, whereby all consequences and benefits related to the reward criteria are quantified. Further increasing complexity, it is not unlikely that the optimal reward program may be constructed by forming a segmented reward criteria approach, using different rewards for different customer groups—based on their scores on different reward criteria. However, considering the involvement of customers in reward programs and the need for clear communication about the reward criterion, companies are extremely reluctant to perform such a real-life test.

Finally, rewarding customers for their predicted loyalty can prove to be difficult to communicate to the total customer base. An operational advantage of the currently used schemes lies in the fact that customers can trust the objectivity of the system: every dollar spent is translated into a certain reward. However, the application of loyalty as a reward criterion does not necessarily imply that successful current systems should be changed. A potential solution would be to maintain the current reward systems and in addition target those customers who are highly loyal but are currently not rewarded for their loyalty, in order to prevent these customers from weakening their relationship owing to a feeling of neglect.

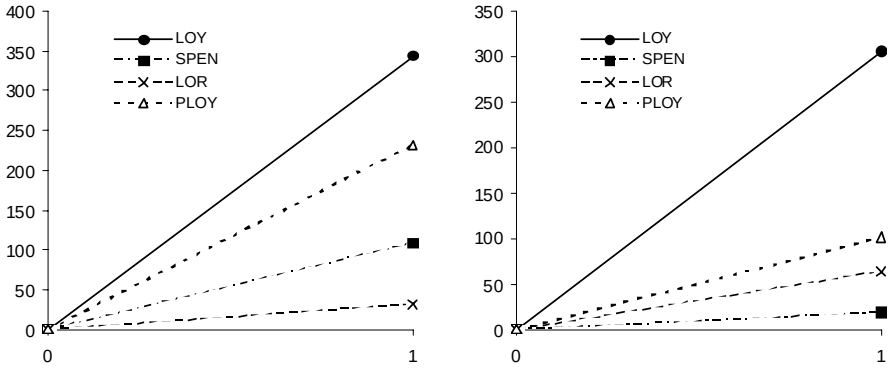
To conclude, a number of further studies can be designed to determine the full potential of using predicted loyalty as a (complementary) reward criterion.

APPENDIX: RELATIONSHIP BETWEEN REWARDS RECEIVED AND BENEFITS DELIVERED



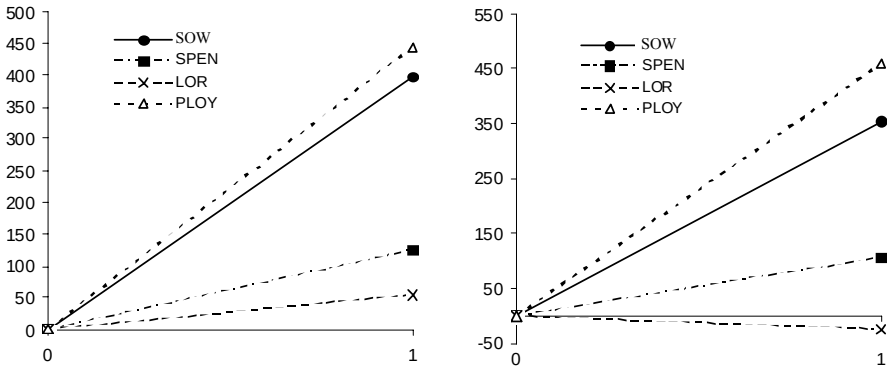
Setting G **Setting M**

A: Word-of-mouth



Setting G **Setting M**

B: Price Insensitivity



Setting G **Setting M**

C: Purchase Intentions

REFERENCES

Baesens Bart, Geert Verstraeten, Dirk Van den Poel, Michael Egmont-Petersen, Patrick Van Kenhove and Jan Vanthienen (2004), "Bayesian Network Classifiers for Identifying the Slope of the Customer Lifecycle of Long-Life Customers," *European Journal of Operational Research*, 156, 508–523.

Bolton, Ruth N., P.K. Kannan and Matthew D. Bramlett (2000), "Implications of Loyalty Program Membership and Service Experiences for Customer Retention and Value," *Journal of the Academy of Marketing Science*, 28 (1), 95–108.

Brodie, Roderick J., Nicole E. Coviello, Richard W. Brookes and Victoria Little (1997), "Towards a Paradigm Shift in Marketing? An Examination of Current Marketing Practices," *Journal of Marketing Management*, 13 (5), 383–406.

Buckinx, Wouter and Dirk Van den Poel (2005), "Customer Base Analysis: Partial Defection of Behaviourally-Loyal Customers in a Non-Contractual FMCG Retail Setting," *European Journal of Operational Research,* 164(1), 252-268.

Bult, Jan Roelf and Tom Wansbeek (1995), "Optimal Selection for Direct Mail," *Marketing Science*, 14 (4), 378–394.

Cohen, Jacob and Patricia Cohen (1983), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum.

De Wulf, Kristof, Gaby Odekerken-Schröder and Dawn Iacobucci (2001), "Investments in Consumer Relationships: A Cross-Country and Cross-Industry Exploration," *Journal of Marketing*, 65 (October), 33–50.

Dick, Alan S. and Kunal Basu (1994), "Customer Loyalty: Toward an Integrated Conceptual Framework," *Journal of the Academy of Marketing Science*, 22 (2), 99–113.

Dowling, Grahame R. and Mark Uncles (1997), "Do Customer Loyalty Programs Really Work?" *Sloan Management Review*, 38 (Summer), 71–82.

East, Robert, Patricia Harris, Gill Willson and Wendy Lomax (1995), "Loyalty to Supermarkets," *International Review of Retail, Distribution and Consumer Research*, 5 (1), 99–109.

Feinberg, Fred F., Aradhna Krishna and John Z. Zhang (2002), "Do We Care What Others Get? A Behaviorist Approach to Targeted Promotions," *Journal of Marketing Research*, 39 (August), 277–291.

Festinger, Leon (1954), "A Theory of Social Comparison Processes," *Human Relations*, 7, 117–140.

Furnival, George M. and Robert W. Wilson (1974), "Regressions by Leaps and Bounds," *Technometrics*, 16, 499–511.

Goutte, Cyril (1997), "Note on Free Lunches and Cross-Validation," *Neural Computation*, 9, 1245–1249.

Hallberg, Garth (2004), "Is Your Loyalty Programme Really Building Loyalty? Why Increasing Emotional Attachment, not Just Repeat Buying, is Key to Maximising Programme Success," *Journal of Targeting, Measurement and Analysis for Marketing*, 12 (3), 231–241.

Irwin, Julie R. and Gary H. McClelland (1991), "Misleading Heuristics and Moderated Multiple Regression Models," *Journal of Marketing Research*, 38 (February), 100–109.

Jones, Thomas O. and Earl W. Sasser (1995), "Why Satisfied Customers Defect," *Harvard Business Review*, November–December, 87–99.

Jöreskog, Karl G. and Dag Sörbom (1995), *LISREL 8: User's Reference Guide*. Chicago: Scientific Software International.

Keiningham, Timothy L., Tiffany Perkins-Munn and Heather Evans (2003), "The Impact of Customer Satisfaction on Share-of-Wallet in a Business-to-Business Environment", *Journal of Service Research*, 6 (1), 37-50.

Kivetz, Ran and Itamar Simonson (2003), "The Idiosyncratic Fit Heuristic: Effort Advantage as a Determinant of Consumer Response to Loyalty Programs," *Journal of Marketing Research*, 40 (4), 454–467.

Latham, Gary P. and Edwin A. Locke (1991), "Self Regulation Through Goal Setting," *Organizational Behavior and Human Decision Processes*, 50 (2), 212–247.

Macintosh, Gerrard and Lawrence S. Lockshin (1997), "Retail Relationships and Store Loyalty: A Multi-Level Perspective," *International Journal of Research in Marketing*, 14 (5), 487–497.

Mägi, Anne W. (2003), "Share of Wallet in Retailing: the Effects of Customer Satisfaction, Loyalty Cards and Shopper Characteristics," *Journal of Retailing*, 79, 97–106.

McMullan, Rosalind and Audrey Gilmore (2002), "The Conceptual Development of Customer Loyalty Measurement: A Proposed Scale," *Journal of Targeting, Measurement and Analysis for Marketing*, 11 (3), 230–243.

Nicholls, John G. (1989), *The Competitive Ethos and Democratic Education*. Cambridge, MA: Harvard University Press.

Nunnally, Jum C. (1978), *Psychometric Theory*. New York: McGraw-Hill.

Reichheld, Frederick F. (2003), "The One Number You Need to Grow," *Harvard Business Review*, December, 46–54.

—— and Earl W. Sasser (1990), "Zero Defections, Quality Comes to Services," *Harvard Business Review*, 68 (September-October), 105–11.

Reinartz, Werner J. and Vikas Kumar (2000), "On The Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing," *Journal of Marketing*, 64 (October), 17–35.

—— and ——(2002), "The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration," *Journal of Marketing*, 67 (January), 77–99.

Reynolds, Kristy E. and Mark J. Arnold (2000), "Customer Loyalty to the Salesperson and the Store: Examining Relationship Customers in an Upscale Retail Context," *Journal of Personal Selling and Sales Management*, 20 (2), 89–98.

Rosenberg, Larry J. and John A. Czepiel (1984), "A Marketing Approach to Customer Retention," *Journal of Consumer Marketing*, 1, 45–51.

Schmittlein, David and Robert A. Peterson (1994), "Customer Base Analysis: An Industrial Purchase Process Application," *Marketing Science*, 13 (1), 41–67.

Sharp, Bryon and Anne Sharp (1997), "Loyalty Programs and Their Impact on Repeat-Purchase Loyalty Patterns," *International Journal of Research in Marketing*, 14 (5), 473–486.

Srinivasan, Srini S., Rolph Anderson and Kishore Ponnavolu (2002), "Customer Loyalty in E-commerce: An Exploration of its Antecedents and Consequences," *Journal of Retailing*, 78, 41–50.

Steenkamp, Jan Benedict E.M. and Hans C.M. van Trijp (1991), "The Use of LISREL in Validating Marketing Constructs," *International Journal of Research in Marketing*, 8, 283–299.

Verhoef, Peter C. (2003), "Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development," *Journal of Marketing*, 67 (October), 30–45.

——, Philip Hans Franses and Janny C. Hoekstra (2002), "The Effect of Relational Constructs on Customer Referrals and Number of Services Purchased from a Multiservice Provider: Does Age of Relationship Matter?" *Journal of the Academy of Marketing Science*, 30 (3), 202–216.

Wangenheim, Florian and Tomás Bayón (2004), "The Contribution of Word-of-Mouth Referrals to Economic Outcomes of Service Quality and Customer Satisfaction," *Journal of the Academy of Marketing Science*, forthcoming.

Whyte, Randall (2004), "Frequent Flyer Programmes: Is it a Relationship, or do the Schemes Create Spurious Loyalty?" *Journal of Targeting, Measurement and Analysis for Marketing*, 12 (3), 269–280.

Zeithaml, Valarie A., Leonard L. Berry and A. Parasuraman (1996), "The Behavioral Consequences of Service Quality," *Journal of Marketing*, 60 (April), 31–46.

CHAPTER IV

THE IMPACT OF SAMPLE BIAS ON CONSUMER CREDIT SCORING PERFORMANCE AND PROFITABILITY[12]

[12] This chapter is based on the following reference: Geert Verstraeten, Dirk Van den Poel, 2005. The impact of sample bias on consumer credit scoring performance and profitability, Journal of the Operational Research Society, Vol 56, pp 981-992.

CHAPTER IV:

THE IMPACT OF SAMPLE BIAS ON CONSUMER CREDIT SCORING PERFORMANCE AND PROFITABILITY

ABSTRACT

This article seeks to gain insight into the influence of sample bias in a consumer credit scoring model. In earlier research, sample bias has been suggested to pose a sizeable threat to predictive performance and profitability due to its implications on either population drainage or biased estimates. Contrary to previous – mainly theoretical – research on sample bias, the unique features of the dataset used in this study provide the opportunity to investigate the issue in an empirical setting. Based on the data of a mail-order company offering short term consumer credit to their consumers, we show that (i) given a certain sample size, sample bias has a significant effect on consumer credit-scoring performance and profitability, (ii) its effect is composed of the inclusion of rejected orders in the scoring model, and – to a lesser extent – the inclusion of these orders into the variable-selection process, and (iii) the impact of the effect of sample bias on consumer credit scoring performance and profitability is modest.

1. INTRODUCTION

In this paper, the term 'credit scoring' is used as a common denominator for the statistical methods used for classifying applicants for credit into 'good' and 'bad' risk classes. Using various predictive variables from application forms, external data suppliers and own

company records, statistical models, in the industry often termed scorecards, are used to yield estimates of the probability of defaulting. Typically, an accept or reject decision is then taken by comparing the estimated probability of defaulting with a suitable threshold (see e.g. Hand and Henley[1]).

Since the very beginning of credit-scoring techniques, the issue of sample bias has rapidly grown to become an intriguing topic in the credit-scoring domain (see e.g. Heckman[2]). The challenge lies in estimating the default probabilities for all future credit applicants using a model trained on a skewed sample of previously accepted applicants only. Indeed, for the historically rejected applications, we are unable to observe the outcome, being whether or not the applicant was able to refund his debt. 'Reject inference' (see e.g. Hand and Henley[3]) comprises the set of procedures determined to decrease the bias that arises by building scoring models on accepted applicants only, e.g. by imputing the target variable for rejected cases. While the existing literature in the domain has mainly focused on describing and (to a lesser extent) testing different procedures of reject inference, in this paper, due to the special features of the data set used, we focus on the results that could be reached when perfect reject inference would occur. In this way, we hope to shed some new light upon the relevance of reject-inference procedures in a consumer credit-scoring setting.

The remainder of this paper is structured as follows. Section 2 discusses the issue of sample bias and reject inference in the credit-scoring literature. Section 3 covers the methodology used in the empirical part of this paper to research the impact of sample bias on credit-scoring performance and profitability. Section 4 handles the data description of the data set used, and the sample composition needed for the empirical study. Section 5 reports the findings of the different research questions that are defined throughout the paper. Finally, conclusions and limitations and issues for further research are given in Sections 6 and 7 respectively.

2. SAMPLE BIAS IN THE CREDIT-SCORING LITERATURE

Considering the widespread use of the statistical scoring techniques in the consumer credit industry, and considering the longevity of consumer credit scoring research (see, e.g. Myers and Forgy[4] for an early application), the literature surrounding customer credit scoring has been growing steadily (for an overview, see, e.g. Thomas[5]). In this study, we focus on

application scoring (see, e.g. Hand[6] for an overview), i.e., we consider the decision whether or not to grant credit to potential lenders upon application, in contrast to the more recently introduced behavioural scoring (see, e.g. Thomas et al.[7]) where the performance of the customer is assessed for decision-making purposes during the lifetime of the relevant credit opening (e.g. whether the credit limit of a current borrower should be increased). Hence, we focus on the core application within the domain.

An important and fascinating topic that has provided much debate in credit scoring concerns the issue of sample bias: if new credit-scoring models are to be built on previous company records, only the previously accepted orders can be used for building the new credit score. Hence, sample bias may arise when the sample of orders used for model building is not representative of the 'through-the-door' applicant population. Indeed, as mentioned by Hand and Henley[3], the reject region has been so designated precisely because it differs in a non-trivial way from the accept region. Historically, sample bias has been accused of introducing at least one of two major shortcomings into the models, namely population drainage or biased estimates. Technically, the new score should only be applied to the customers that were accepted in the past (see, e.g. Joanes[8]), and those rejected should remain rejected, leading inevitably to a decreasing customer base. Lacking an appropriate term for this in credit scoring, in this paper, we use the term 'population drainage' to cover this phenomenon. Alternatively, should the score be applied to all future orders, biased estimates may result for those credit applicants that would have been rejected by the previous credit score, namely when the ex post probability of default conditional on the covariates differ between accepted and rejected orders (for a clear discussion of the specific conditions for the occurrence of sample bias, see e.g. Banasik et al.[9], p.823). Understandably, both consequences have been proposed to have a negative influence on a company's profitability.
Considering first the intriguing problem at hand, and secondly initial reports of the sizeable influence of sample bias when using discriminant analysis[10], previous research has historically focused on imputation techniques whereby one attempts to predict the (unobserved) outcome of the previously rejected orders, henceforth called 'reject inference'. In fact, research on possible methods of reject inference date back almost as far as the beginning of credit scoring itself (see, e.g. Hsia[11]).

While overviews of such methods are widely available (see, e.g. Joanes[8] and Hand and Henley[1]), it might be useful to cover briefly the ideas behind two of the most renowned

reject inference techniques here. They include (i) augmentation, where the accepted orders are weighted inversely proportional to the probability with which the orders were accepted, in order to increase the impact of orders that are comparable to those orders that were rejected, and (ii) iterative reclassification, where rejected orders are scored and discretized using the classification rule derived from accepted orders, and where the model is re-estimated. After dividing the data again into samples of the same size as the original accept and reject regions, this procedure is repeated iteratively until convergence occurs (see e.g. Joanes[8]). The results of these various attempts of reject inference, however, seem risky at best, leading some authors to conclude that reliable reject inference is impossible (Hand and Henley[3]).

In contrast to studies on reject inference, the possible impact of sample bias itself on customer credit scoring has been covered to a much lesser extent. Additionally, it has been argued recently that the use of discriminant analysis introduces bias (infra), whereby the previous findings concerning the impact of sample bias (in e.g. Eisenbeis[10]) should be reconsidered. While the shortage of a random sample of rejected orders is mentioned frequently, the costs involved with gaining such data are very often mentioned in the same breath (see, e.g. Hand and Henley[1, 3]). Due to the fact that the data set used in this study contains real outcome values for a sizeable group of orders rejected by the statistical scoring process (infra), we attempt to assess the importance of sample bias itself, as an upper limit of the benefits that could result from using models of reject inference. Hence, it lies not within the ambition of this study to test the performance of each of the proposed reject inference methods, but given the fact that reject inference deals with attempting to infer the true creditworthiness status of rejected applicants[3], it would seem beneficial to estimate the maximal improvement that could be reached when the imputation method is 100 % correct.

3. METHODOLOGY

3.1 General outlay of this study

As mentioned earlier, a special feature of the data set used, exists in the fact that we have available the real outcome for a sizeable set of customers that were rejected by the scoring process. This sample of borrowers is henceforth referred to as the 'calibration' set. While the details of the data set will form the main topic of Section 4 of this paper, in this section we

will describe how this calibration set will be used in the current study. First and foremost, it should be clear that the advantages of having such a set are twofold: while it allows the researcher to include the orders into the model-building process, it is equally valuable that these orders can be used in constructing the holdout sample. In this way, it will be possible to mimic the behaviour of the score on a set of orders that is proportional to and hence more representative for the 'through-the-door' applicant population discussed earlier. Note that we fully acknowledge that we only have available a sample representation of an all applicant population, instead of having available the real outcome of an all applicant population. In the section covering the limitations of this study, we elaborate upon the degree to which this can have an impact on the results.

This study will roughly focus on three parts. Using the calibration set, we will first attempt to acknowledge the problems resulting from sample bias. Hence, in our first research question (henceforth called Q1), we will investigate whether sample bias occurs by (i) testing the performance of a classifier built on the orders that were accepted by the score yet applied on the calibration sample only, and (ii), using an extensive variable-selection procedure proposed by Furnival and Wilson[12], we will investigate whether different characteristics would prevail when the calibration sample is included into the variable-selection process. Indeed, in his study, Joanes[8] indicated the common practice of using a variable-selection procedure for detecting a small – yet effective – subset of the total list of potential variables, and uttered that a model derived from previously accepted applicants only may fail to take into account all the relevant risk characteristics.

More crucial to this study, in our second research question (Q2), we will attempt to estimate the gain in performance if the outcome of the rejected orders would be available. In this step, we will compare the performance of a model with sample size n, only containing previously accepted orders, versus a model with an equal sample size n containing a sample of accepted and rejected orders that is proportional to the ratio of accepted and rejected orders in the applicant population – henceforth referred to as 'proportionality'. In this effort, we will clearly distinguish between the benefits of creating a proportional sample through the use of different orders in the training data set, and the benefit arising through the application of the variable-selection procedure on a proportional sample. It should be clear that all models will be tested on a proportional holdout sample. To conclude, in this subsection we will test

whether mimicking the 'proportionality' of the through-the-door population increases predictive performance and profitability, given a certain sample size.

While the general purpose of this study was described here, some methodological decisions were made, resulting in the choice of an iterative resampling procedure using logistic regression analysis, monitored by three different performance indicators. Additionally, we will perform a sensitivity analysis to test the robustness of our findings. The reasons behind these decisions are presented in the sections below.

3.2 Credit-scoring technique

Recent research in credit scoring has been focused on comparing the performance of different credit-scoring techniques, such as neural networks, decision trees, k-nearest neighbour, support vector machines, discriminant analysis, survival analysis and logistic regression (see, e.g. Baesens et al.[13], Stepanova and Thomas[14], Desai et al.[15], Davis et al.[16]). The main conclusions from these efforts are that the different techniques often reach comparable performance levels, whereby traditional statistical methods, such as logistic regression perform very well for credit scoring. Hence, in this paper, we will use the latter method for modelling credit risk. Two other reasons confirm this choice: firstly, several authors consider logistic regression to be one of the main stalwarts of today's scorecard builders (see, e.g. Thomas[5], Hand and Henley[1]), and secondly, discriminant analysis, being another technique that has extensively been used in credit analysis[18], has been proven to introduce bias when used for extrapolation beyond the accept region.[3, 10, 19]
Technically, we can represent logistic regression analysis as a regression technique where the dependent variable is a latent variable, and only a dummy variable y_i can be observed[17]:

$$y_i = \begin{cases} 1 \text{ if the borrower defaults} \\ 0 \text{ if the borrower is able to refund his debt} \end{cases}$$

3.3 Performance measurement

In agreement with recent studies where performance measures for classification are crucial (see, e.g. Baesens et al.[13]), we will not rely on a single performance indicator in reporting the results of our research. The performance measures used are: (i) classification accuracy, or

the percentage of cases correctly classified (PCC), (ii) area under the receiver operating characteristics curve (AUC), and (iii) profitability of the new score. While both the first and the last measure reports a result based on a fixed threshold, the receiver operating characteristic curve illustrates the behaviour of a classifier without regard to one specific threshold, so it effectively decouples classification performance from this factor (see e.g. Egan[20] for more details, or Hand and Henley[1], for an overview of related performance assessment tools in the credit-scoring domain). An intuitive interpretation of the AUC is that it provides an estimate of the probability that a randomly chosen defaulter is correctly rated (i.e. ranked) higher than a randomly selected non-defaulter. Thus, the performance measure is calculated on the total ranking instead of a discrete version of it, so it is clearly independent of any threshold applied ex-post. Note that this probability equals 0.5 when a random ranking is used. Both PCC and AUC have proven their value in related domains for binary classification, such as e.g. direct-mail targeting (see, e.g. Baesens et al.[21]). Additionally, since it was feasible to trace back all revenues and costs to the individual orders, and since profitability is by definition the critical performance measure in a business context, we included it as the third credit-scoring performance measure used in this study.

3.4 Resampling procedure

Throughout the different empirical analyses of this study, a resampling procedure was used to assess the variance of the performance indicators. Considering a low proportion of defaulters in the data set used, in this study we will draw samples of n points without replacement from the n points in the original data set, allocating an equal amount of defaulters to training as to holdout samples. Hence, we will use a stratified resampling procedure. This repartitioning of the data will be performed 100 times, and the differences between the different models will be computed within every iteration (i.e. paired comparisons).

3.5 Sensitivity analysis

One could argue that the value of reject inference is driven by the extent of truncation, being the relative size of the reject region. In order to ensure the validity of our results, using a sensitivity analysis, we will treat previously marginally accepted orders (the orders having a probability of defaulting very close to, but still below the threshold) as rejected orders, to the

extent that 70 % of the total applicants for credit are considered as historically accepted, considering Hand and Henley's[1], expertise of this being a normal acceptance rate in mail-order consumer credit. Hence, given historical scores of orders, we can pretend that some orders were rejected, while they were actually accepted, so we are able to include them in our calibration sample.

3.6 Similarities and differences with previous studies

While the shortage of databases containing default information of the rejected orders prevails, a notable exception was recently put forward by Banasik et al.[9], who had access to a database containing the outcome of all orders that would have been rejected by the current scoring system, but were nonetheless accepted. While they conclude that the scope for model improvement due to a reduction of sample bias is moderate, they clearly state that these results were specific to the acceptance threshold indicated by the data; and hence this paper calls for a validation on other data sets. The present study clearly offers support of the findings on a different dataset, using a different methodology. At least 5 major differences exist between Banasik et al.[9] and this study: (i) for this study, the real outcome of the credit was only available for a sample of the rejected orders; (ii) the present study was performed on a larger, however less balanced dataset; (iii) we had a perfect knowledge of the actual acceptance process, while Banasik et al.[9] had no access to the actual acceptance process, (iv) the particularities of the data did not allow us to create bands of applicants, ranked by predicted creditworthiness; nevertheless, a sensitivity analysis was performed to validate our findings; (v) the particularities of the data allowed us to investigate the impact on profitability and the influence of sample bias on the variable selection process.

4. DATA DESCRIPTION AND SAMPLE COMPOSITION

4.1 Data description

For our research, we used data of a large Belgian direct mail order company offering consumer credit to its customers. Its catalogue offers articles in categories as diverse as furniture, electrical, gardening and DIY equipment and jewellery. We performed the modelling at a moment when the former credit score – constructed by an international company specialized in consumer credit scoring – was about to be updated since it had been

in use for 6 years. For modelling purposes, we will use data of all short-term credit orders placed between July 1st 2000 and February 1st 2002, and their credit repayment information until February 1st 2003. Within this period, all the credits observed had to be refunded, so it was possible to indicate good versus bad credit repayment within 12 months of follow-up.

In the remainder of this section, we try to clarify in which way the data that were used in this study contain several advantages compared to data used in previous research. In order to do this, however, it is crucial to describe the company's order-handling process in detail. We attempt to do so below.

The ordering process at the focal company is bipartite. Firstly, orders are always scored by an automatic scoring procedure, previously called the former credit score. However, in addition to this procedure, an independent manual selection procedure (also called a 'judgmental' procedure) is used for orders with specific characteristics, hence selecting a rather large set of orders that were handled manually, regardless of their score. Therefore, since the scores of all orders were tracked, it is possible to assign each order exclusively to one of the six possible order routes, as given in Table 1. Manual acceptance overrules, yet is not always applied. Hence, we can ex-post define six possible order flows for the orders that were handled.

Table 1. Order flow frequencies

		Judgmental Method		
		Not handled	Accepted	Rejected
Scoring Method	Accepted	A1 32503 obs 528 defaults (1.62 %)	A2 3536 obs 101 defaults (2.86 %)	R3 2844 obs
	Rejected	R1 234 obs	A3 2009 obs 107 defaults (5.33 %)	R2 3228 obs

In order to give a clear overview of the relevance of each of the groups for this study, in Figure 1 we have indicated the positioning of the different groups according to the existing credit score. Note that we have bracketed the traditional terminology concerning overrides, because 'high' and 'low' are conditional on the coding of the dependent variable. In this study we have coded a defaulter as 1 and a non-defaulter as 0, while often in literature

reverse coding is used. Note also that, following most published credit scoring applications, we shall not consider different kinds of defaulters here, yet we will merely distinguish between 'goods' and 'bads'. Additionally, the term 'override' here does not really correspond to its traditional connotations, since the decision to treat an order manually is an autonomous decision, and is by no means based on the scoring of the automatic scoring procedure, yet on a set of disjunctive decision rules: once one of the criteria is met, the order will be processed manually.

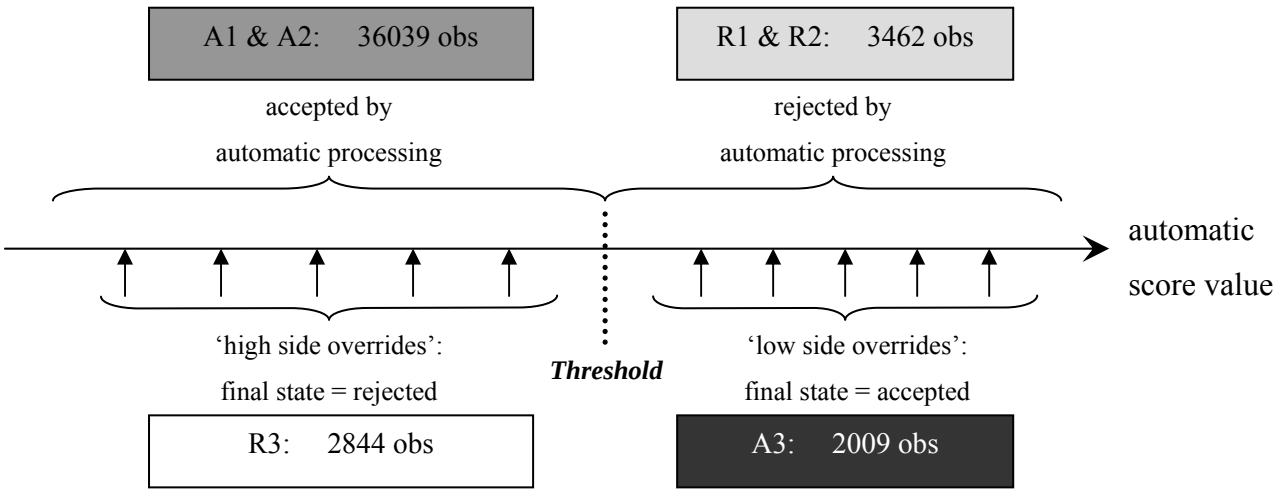


Figure 1: Order flow visualization

Several consequences can be drawn from the knowledge of the order handling process at the company in the case, and are crucial to this study. First, it should be noted that the group of orders that were accepted by the score yet rejected by manual decision (i.e. the orders termed 'R3'), are of little importance to this study: they were mainly rejected for strategic and/or legal reasons (e.g. aged under 18) and thus cannot serve in the modelling process. Indeed, following Hand and Henley[1], high side overrides will not lead to biased samples if the relevant application population is defined exclusively of those eliminated by a high side override. Second, the score accepted 36039 relevant orders, yet rejected a total of 5471 orders, resulting in an acceptance rate of 86.8 %. Third, of the latter set, 2009 orders were overridden (the orders termed 'A3'), and hence these cases provide a sizeable sample of the rejected cases containing default information. Indeed, while there is no objective reason why these orders should be accepted (they have a high probability of defaulting), according to the automatic scoring procedure they were still accepted regardless of any decision rule. It should be mentioned that in more than 95% of the orders rejected by the score, the orders

were handled by the judgmental process (cf. cell 'R1' in Table 1 comprises only 234 orders). Hence, the decision to accept an order that was rejected by the score, was to a very large extent driven by manual decision making, where different company employees rely on different beliefs about credit risk. This sample of orders permits the analyses that are proposed in this study, and has previously been called the calibration sample.

4.2 Sample composition

For our research, each analysis proposed in the methodological section requires the use of different samples. Considering the importance of this sampling for the study, below, we will describe the data used in detail.

4.2.1 Detection of sample bias

Underperformance of the score on rejected orders: In order to detect whether a model trained on the orders accepted by the score is better able to predict similar orders than orders from the calibration set, 50 % of the orders accepted by the score was used for training the model. As holdout samples, the total calibration sample was compared with a sample of the remaining accepted orders. However, in this sampling procedure, an equal fraction of defaulters was guaranteed for both holdout samples. This was indispensable to ensure comparability of the outcome values. We remind the reader that this sampling procedure was repeated 100 times.

Variable-selection process: In order to detect whether the inclusion of the calibration sample influences the variable-selection process, we compared the variables selected on a (training) sample of 50 % of the accepted orders with a sample constituted by 50 % of the calibration sample (i.e. 1005 rejected orders), and a sample of accepted orders (6617 accepted orders), ensuring the proportionality of the 'through-the-door' applicant population (i.e. 36039 accepted versus 5471 rejected orders).

4.2.2 Influence of sample bias

In this research question, we test the effect of ensuring proportionality, given a certain sample size. Hence, the holdout sample that will be used will consist of 50 % of the orders in

the calibration sample (i.e. 1004 rejected orders), and a sample of accepted orders (6617 accepted orders), ensuring the proportionality of the 'through-the-door' applicant population. In terms of the training data, we will compare a model built on the remaining 1005 rejected orders and a proportional sample of accepted orders (6617 accepted orders) with a sample of 7622 accepted orders only. A visualization of this sampling procedure can be found in Appendix A.

Considering the possible impact of the ratio of accepted versus rejected orders on the results, following Hand and Henley's[1] suggestion, a sensitivity analysis will be performed whereby only 70 % of the orders will be treated as historically accepted orders. While the real threshold was defined on a probability of defaulting of 0.04, in this analysis, we have reconstructed the situation where orders would have been rejected if they had a default probability larger than 0.01743. It is important to note that, since the company tracked the historical scores, we were able to mimic the original acceptance procedure faultlessly, contrarily to Banasik et al.[9]. Hence, a sample of 6833 orders was appended to the calibration sample, ensuring that half of this sample was used for model building, while the other half was used for model testing. In this case, the holdout sample consisted of 50% of the orders in the calibration sample (4421 'rejected' orders), and a sample of accepted orders (10494 accepted orders), ensuring the 70 % proportionality. In order to check the impact of sample bias on credit-scoring performance, a model built on a proportional sample of 4421 rejected and 10494 accepted orders was compared with a model built on a sample of 14915 orders only. Again, we remind the reader that this sampling procedure was repeated 100 times, and that a stratified sampling procedure was used to ensure an equal ratio of defaulters in training versus holdout samples.

4.3 Variable creation

Both in the conception of the dependent (classification) variable as in terms of the independent variables (characteristics), the authors have relied heavily on the experience by the managers of the company as well as on findings in previous research. In terms of dependent variable, we have used the fact whether a third reminder was sent to the customer, because (i) this is the moment that the customer will be charged for his delay in refunding, (ii) this reminder really urges the customer to repay and (iii) this variable had historically been used by the company to investigate defaulting behaviour. While information was also

available about whether the individual order was eventually profitable to the company, the use of the third reminder was preferred over the profit information, as the number of unprofitable orders was low, which degraded the performance of all models severely.

The customer characteristics that were used in the study belong to the traditional set of characteristics used for credit-scoring purposes. However, due to the fact that this is a consumer credit setting, it is a strategic decision of the company to limit the information inquired upon application and to rely more on own-company credit records. Nevertheless, we were also able to include characteristics covering demographic information and occupation, financial information (e.g. number of credits still open) and default information (e.g. ratings by credit bureau reports, own company default information). The list and description of the 45 characteristics used in this study can be found in Appendix B.

5. RESULTS

5.1 Detection of sample bias

5.1.1 Underperformance of the score on rejected orders

When considering sample bias, a basic assumption is that a scoring model built on accepted orders only, performs significantly worse on orders that would have been rejected by the previous score, than on orders that would have been accepted by the previous score. The results of this first step confirm the hypothesis. A model trained on orders accepted by the score was tested on similar orders, and compared with a model tested on orders rejected by the score, yet accepted by the judgmental procedure (i.e. the calibration sample). After repeating a resampling procedure 100 times, the mean AUC of the (holdout) sample containing a sample of previously accepted orders was 0.7533, while the mean AUC of the (holdout) sample containing the calibration sample was 0.6782, hence the AUC of the latter set was on average 7.51 percentage points lower than the AUC of the first set, a difference that was significant at $p < 0.0001$ (t Value of 44.37). Hence, we can indeed conclude that the extrapolation of the score towards a range of customers that was not used for training does prove to be more erroneous than applying the score towards similar orders as those in the training set. The degree to which this poses a problem for the scoring performance as a whole, will be discussed at length after the following section.

5.1.2 Variable-selection process

In terms of the variable selection procedure, we have used the leap-and-bound algorithm of Furnival and Wilson[12] to detect the best model for all possible model sizes (number of variables), requiring a minimum of arithmetic, and the possibility for finding the best subsets without examining all possible subsets. However, considering the initial problems that arose due to multicollinearity, following Cohen et al.[22] (2003, p. 428), we first created principal components from the characteristics, whereby a set of independent dimensions can be used as variables feeding into the variable selection procedure. Since the leap-and-bound algorithm only provides a likelihood score (chi-square) statistic without significance tests, we have used the algorithm proposed by De Long et al.[23] in order to investigate whether the AUC of a model with a given model size differs significantly from the AUC of the single model containing all of the explanatory effects. Starting from the full model containing 45 principal components and reducing model size, we have selected the last model that does not differ significantly from the model using all components at a 0.05 significance level. This procedure was performed twice: once in the actual setting, where 86.8 % of the orders were accepted, and once in the sensitivity analysis, where the situation where only 70 % of the orders would have been accepted was mimicked. In the actual setting, of the 42 components that were selected in either of both models, 27 components occurred in both models (i.e. 64.29 %). In the sensitivity analysis, we counted 29 matching components out of the 40 that appeared in any model (i.e. 72.5 %). Due to the presence of multicollinearity, it is hard to conclude from this analysis whether the characteristics needed for the prediction of creditworthiness of previously rejected orders are different from the characteristics needed to evaluate accepted orders. However, the degree to which the selection of different variables has an influence on credit-scoring performance and profitability is an important topic in the following section.

5.2 Influence of sample bias

In this section, we will give an overview of the results of our main research topic, namely the impact of sample bias on consumer credit-scoring performance and profitability for a given sample size. In this set of analyses, we have attempted to assess the confidence interval around each performance indicator by resampling the data 100 times, where the reported

differences in (test set) performance between different models were always computed within one iteration of the resampling procedure. Hence the degrees of freedom used for the tests were 99. Following Hand and Henley[1], who state that, in mail-order purchasing, a figure of 70 % accepted orders is quite usual, all analyses that were performed here are validated in a sensitivity analysis, where we reuse the data to reconstruct the case if only 70 % of all applicants had been accepted. To enhance comparability, the results of the real setting (acceptance rate of 86.8 %) and the sensitivity analysis (70 %) are always presented side by side. To enable the reader to compare the results with the performance of the previous credit score, we included its performance and labelled it "old" performance. Model 1 represents a model where only orders were used that were accepted by the scoring procedure, both in terms of inclusion of the orders in the training sample, as in terms of inclusion of the orders in the variable selection procedure (labelled 'AA'). Model 2 uses a sample of orders accepted by the score and/or accepted by the manual selection process, in such a way that the proportionality of accepted versus rejected orders is respected. However, in model 2, the variable-selection procedure is still performed on a sample of accepted orders only ('PA'). Finally, in model 3, the same orders were used as model 2, but the variables of this model were selected on a proportional sample of accepted and rejected orders ('PP'). We first start by covering credit-scoring performance, and then discuss profitability issues.

Table 2. PCC performance when reducing sample bias at a given sample size

	Actual setting (86.8 % accepted)			**Sensitivity analysis (70 % accepted)**		
Morrison	0.8520			0.6950		
	Mean	**Std Dev**		**Mean**	**Std Dev**	
PCC old	0.8601	0		0.7069	0	
PCC model 1: 'AA'	0.8643	0.0016		0.7111	0.0010	
PCC model 2: 'PA'	0.8648	0.0014		0.7112	0.0009	
PCC model 3: 'PP'	0.8648	0.0016		0.7113	0.0010	
	Mean	**t Value**	**P > |t|**	**Mean**	**T Value**	**P > |t|**
PCC 2 – PCC 1	0.0005	3.90	0.0002	55E-6	0.66	0.5087
PCC 3 – PCC 1	0.0005	3.00	0.0034	0.0002	1.59	0.1153
PCC 3 – PCC 2	262E-8	0.02	0.9833	0.0001	1.26	0.2093

5.2.1 Predictive Performance

The results in terms of PCC performance can be found in Table 2. In order to compute these results, the probability of defaulting was discretized into two classes, new accepts and new rejects, in a way that the proportionality of the new model was ensured. Hence, we report the performance if the same acceptance rate would be applied for the use of the new model than during the use of the previous model. In terms of testing the classification accuracy versus a random model, we have used the formula proposed by Morrison[24] which states that the accuracy of a random model is defined by:

$$p\,\alpha + (1 - p)\,(1 - \alpha),$$

where p represents the true proportion of refunded orders and α represents the proportion of applicants that will be accepted for credit. All models performed significantly better than this random-model benchmark ($p < .0001$). Hence, we are confident that all models (also the previous credit score) perform reasonably well. Additionally, the PCC of all new models was significantly higher than the PCC of the previous model ($p < .0001$), while the mean differences ranged between 0.41 and 0.46 %, indicating a clear improvement by the new model.

The impact of sample bias on PCC performance is illustrated by the three lower rows of Table 2, where the differences between the new models are tested. From these tests, it is clear that (i) in the actual setting, the second model – containing a proportional sample of accepted and rejected orders - performs significantly better than the first model, built on accepted orders only. Henceforth, we consider significance at the 0.05 level. Furthermore, (ii) again in the actual setting, the third model – containing the variables selected on a proportional sample of rejected and accepted orders – performs significantly better compared to the first, but not when compared to the second model. Additionally, (iii) in the sensitivity analysis, none of the three new models show significant differences in terms of PCC performance. To conclude, we can state that in both settings, in terms of PCC, performing the variable selection procedure on the proportional sample does not increase predictive performance, and that the impact on PCC that can be reached by including the orders of calibration sample into the modelling process in a proportional way seems to be low (0.0005), especially when compared to the difference resulting from the update of the model (0.0042).

Table 3. AUC performance when reducing sample bias at a given sample size

	Actual setting (86.8 % accepted)			Sensitivity analysis (70 % accepted)		
	Mean		**Std Dev**	**Mean**		**Std Dev**
AUC old	0.7131		0.0138	0.7041		0.0054
AUC model 1: 'AA'	0.7464		0.0196	0.7457		0.0137
AUC model 2: 'PA'	0.7522		0.0186	0.7483		0.0116
AUC model 3: 'PP'	0.7537		0.0189	0.7561		0.0116
	Mean	**t Value**	**P > |t|**	**Mean**	**t Value**	**P > |t|**
AUC 2 – AUC 1	0.0057	6.80	<.0001	0.0027	2.77	0.0067
AUC 3 – AUC 1	0.0072	5.39	<.0001	0.0104	10.09	<.0001
AUC 3 – AUC 2	0.0015	1.41	0.1610	0.0078	13.52	<.0001

A main drawback of PCC performance is that it requires the user to discretize the probability of defaulting, such that the model will only be evaluated for a given threshold. This, however, does not give the user any indication of how the model performs if other threshold levels were to be used. Since AUC does give an evaluation of a score across the total range of default probabilities, we report the AUC performance in Table 3.

The results of this analysis are analogous to the PCC results. Hence, in the actual setting, (i) all new models perform significantly better than the previous model ($p < 0.0001$), (ii) model 2 performs significantly better than model 1, (iii) there is no statistical difference between models 2 and 3, and (iv) the improvement of performance between model 2 and model 1 seems relatively small compared to the difference resulting from the update of the model. However, in the sensitivity analysis, the impact of using the variables selected on the proportional sample does prove useful. Indeed, when 70 % of all orders would have been accepted, the highest predictive performance would have been reached by using model 3. Yet again, the improvements of performance between the new models seem relatively small compared to the difference resulting from the update of the model.

5.2.2 Profitability

Since order profitability could be computed at the individual order level in the database, in this section, we will review the impact of sample bias on consumer credit scoring profitability, given a certain sample size. Considering the confidentiality of the data, the

authors were unable to reveal absolute profit information. Therefore, Table 4 only represents the relative profit changes that could be reached by introducing the information stemming from rejected orders. This difference is again computed per resampling iteration, and the average of the 100 resamples is represented and tested against the null hypothesis that this difference is zero.

Finally, also in terms of profitability we reach similar conclusions as before. Hence, the effect of bias in terms of including the orders in the modelling does also play a part when considering credit-scoring profitability. However, the effect of including the calibration sample into the variable-selection process again seems less lucrative. Indeed, as model 2 again performs significantly better than model 1, model 3 adds no further significant improvement (on a 0.05 confidence level) to credit scoring profitability. Again, while sample bias has proven to have a significant impact, this impact seems to be low. For example, in the current setting of 86.8 % accepted orders, the maximum profit gain that could be reached by gaining the knowledge of the outcome for all rejected orders (assuming the price for gaining this knowledge to be zero), would be 0.4 %. Hence, any procedure of reject inference that results in perfect imputations of the defaulting behaviour of rejected orders would maximally reach this improvement. In conclusion, the profit that can be realized by introducing information from the rejected orders into the model seems modest, especially when it should be able to cover the defaulting cost of including a random sample or the time cost involved in applying any reject-inference procedure.

Table 4. Profit implications when reducing sample bias at a given sample size

	Actual setting (86.8 % accepted)			**Sensitivity analysis (70 % accepted)**		
	Mean	**t Value**	**P > |t|**	**Mean**	**t Value**	**P > |t|**
Profit Difference 2 vs. 1	0.0040	3.27	<.0001	0.0102	3.56	0.0006
Profit Difference 3 vs. 1	0.0034	2.77	0.0067	0.0127	4.33	<.0001
Profit Difference 3 vs. 2	-59E-5	-1.06	0.2926	0.0025	1.83	0.0705

To conclude this section, we detect from the comparison of the profit implications of the actual setting with the sensitivity analysis that the profitability from including the calibration sample rises as the proportion of rejected orders grows larger. While this effect was not

tested statistically, it seems only logical that the impact of the reduction of sample bias rises when sample bias itself grows in size.

6. CONCLUSIONS

In this study, the authors attempted to indicate and quantify the impact of sample bias on consumer credit scoring performance and profitability. Historically, sample bias has been suggested to pose a sizeable threat to profitability due to its implications on either population drainage or biased estimates. While previous research has mainly been focused on offering various attempts to reduce this bias, mainly due to lack of appropriate credit scoring data sets, the impact of sample bias itself has been largely unexplored. By means of the properties of the data available for this study, however, the authors were able to assess the existence and the impact of sample bias in an empirical setting. In the remainder of this section, we summarize and discuss the results of the analyses performed.

The results of the study indicate that sample bias does appear to have a negative influence on credit-scoring performance and profitability. Indeed, first, a model trained on accepted orders only reveals an important and significant underperformance on the rejected orders compared to its performance on other accepted orders. Secondly, the variable selection procedure used in this paper, showed that (at least some) other characteristics were selected for predicting the creditworthiness of previously rejected or previously accepted orders. Thirdly, and most importantly, in order to quantify the effect of sample bias on credit scoring performance and profitability, we compared model performance of a model built on accepted orders only, with an equally-sized sample of accepted and rejected orders proportional to the applicant population. The results from this analysis indicate that sample bias does prove to have a significant, albeit modest effect on consumer credit scoring performance and profitability. Note that these results confirm previous findings by Banasik et al.[9], where different data sets and methodologies were used to investigate the research question. Additionally, in terms of predictive performance as well as profitability, the negative impact occurs mainly through the inclusion of rejected orders in the training data set, as the inclusion of the rejected orders in the variable-selection procedure clearly proved to be of lesser importance. Despite the fact that different information was selected in the variable selection process, the use of different variables this did not always result in a significant difference on different indicators and in different settings. It should be mentioned, however,

that both in the actual setting as during a sensitivity analysis, the impact of knowing the outcome of all the orders rejected by the score is limited, especially when the costs of gaining this knowledge must be accounted for. To conclude, on a theoretical level, the effect of proportionality prevails, i.e. given that the outcome is known for (a subset of) the orders rejected by the automatic scoring procedure, it can be used effectively to construct a sample that is more proportional to the through-the-door population. Thus, it has been proven that enhancing proportionality can result in improvements in classification accuracy and profitability. However, it should be clear that, at least in this consumer credit setting, the resulting benefits from determining the true outcome values of the rejected cases are low, and company resources could be spent more efficiently by handling other topics relevant to consumer credit scoring.

7. LIMITATIONS AND ISSUES FOR FURTHER RESEARCH

This study was only feasible because of the specific properties of the data available: it is quite exceptional to academics to have available information on the outcome of a sizeable range of the reject region. However, in this study, we did not obtain the outcome of all rejected orders, but only of those orders accepted by a manual scoring procedure. This raises the question to which extent the orders in the calibration sample are representative for the complete reject region. While the overrides make up 36.72% of the rejected cases, in the sensitivity analysis, by treating the orders that were historically accepted yet close to the cut-off as rejected orders, we construct a situation in which we have available the real outcome of 71.86% of the cases that would be rejected if 70% of all orders would have been accepted by the scoring system. Additionally, a plausibly sizeable group of the orders rejected by both the automatic as the manual procedure are rejected for strategic and/or legal reasons (e.g. aged under 18) and given that these reasons are not altered, they should not be included in the calculations, implying that the calibration sample constitutes an even larger proportion of the relevant reject region. Hence, including the overrides does allow one to observe the performance of a reasonable proportion of the cases normally rejected by a scoring model. Nevertheless, the mere fact that different information can and will be used in the manual acceptance procedure than in the automated acceptance procedure, will lead to an omitted-variable problem, which has been proven to impart sample bias despite the use of a sizeable sample representation as an approximation of the all applicant population[3]. To conclude, however, a more complete evaluation of the impact of sample bias is only possible if the real

outcome was available for all rejected cases, cfr. the data from Banasik et al.[9], but it is generally accepted that such data are virtually unique.

Additionally, while the specificity of the data was attempted to be minimized through a sensitivity analysis and the use of different performance measures, this study was executed on the data of a direct-mail company. Unfortunately, the results cannot be extrapolated without reflection towards non-consumer credit scoring, considering the specific properties of the data set used, being (i) a rather large percentage of accepted orders (86.8 %), (ii) a rather low percentage of defaulters (1.94 %), (iii) a rather balanced structure of the misclassification costs (the cost involved with a defaulter was only 2.58 times higher than the profit gained from a non-defaulter). Consequently, it would be useful to replicate the analyses performed here on the data of other credit-offering institutions. Nevertheless, in this paper, we have offered a workable methodology towards analyzing the impact of sample bias in any credit scoring environment. More specifically, during the sensitivity analysis, we offered a procedure that can be implemented to investigate the impact of sample bias whenever historical score values were recorded. Further research largely depends on the availability of other credit scoring datasets.

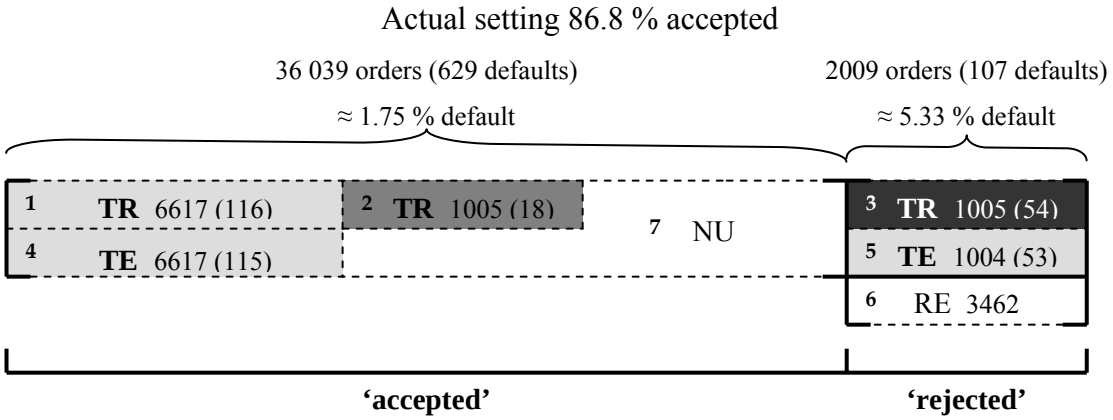
ACKNOWLEDGEMENTS

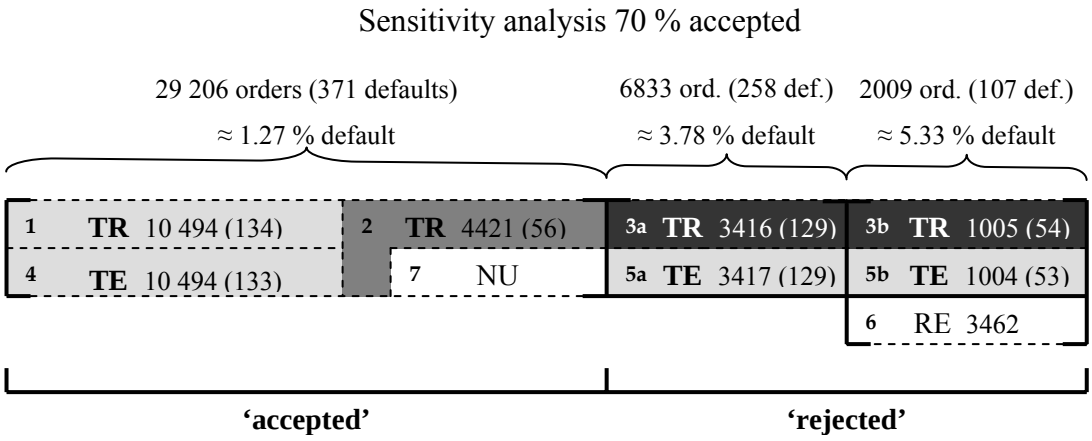
The authors wish to thank the participants of the BANFF Credit Risk Conference 2003 for their comments on an earlier draft of this paper. Additional gratitude is expressed to the anonymous reviewers for their constructive comments and suggestions, and to Jonathan Burez, Joanna Halajko and Bernd Vindevogel, graduates from the Master of Marketing Analysis at Ghent University, for their useful assistance in terms of data preparation.

APPENDIX A: SAMPLE COMPOSITION RESEARCH QUESTION 2

TR = training set **TE** = testset RE = rejects NU = not used

Number of observations (number of defaults)





APPENDIX B: LIST OF VARIABLES USED

	Name	Type	Description
	Home_Own1	Binary	Home ownership: 1=owner, 0=renter
	Home_Own2	Binary	Home ownership: 1=missing , 0=non missing
	Occup1	Binary	1=full-time, 0=other cases
	Occup2	Binary	1=retired, 0=other cases
	Occup3	Binary	1=housewife, 0=other cases
	Occup4	Binary	1=living on social welfare, 0=other cases
	Occup5	Binary	1=student, 0=other cases
	Occup6	Binary	1=without profession, 0 = other cases
	Occup7	Binary	1=missing, 0=other cases
	Bank_Acc	Continuous	Number of bank accounts
	Debt1*	Binary	1=current debt <10.000, 0=other cases
	Debt2*	Binary	1=(10.000<=current debt<25.000), 0=other cases
	Debt3*	Binary	1=(25.000<=current debt <60.000), 0=other cases
	Debt4*	Binary	1=(current debt >=60.000), 0=other cases
	Open_credit*	Continuous	Number of credits that are still due
	Amount_open*	Continuous	Amount that is still open for those credits
	Open_credit_old*	Continuous	Number of credits that are older than 120 days
	Amount_open_old*	Continuous	Amount open for those credits
	Amount_all*	Continuous	Amount of all the credits of a customer
	Amount_paid*	Continuous	Amount of the credits that were refunded
	Ratio_paid*	Continuous	Amount_paid / Amount_all
	Blacklist1	Binary	1=listed on the 'black list' VKC, 0=other cases
	Blacklist2	Binary	1=missing value 'black list' VKC, 0=other cases
	Blacklist3	Binary	1=listed on the 'black list' UPC, 0=other cases
	Blacklist4	Binary	1=missing value 'black list' UPC, 0=other cases
	Remind1*	Binary	1=reminder history not known, 0=other cases
	Remind2*	Binary	1=received 2nd reminder, 0=other cases
	Remind3*	Binary	1=received 3rd reminder, 0=other cases
	Remind4*	Continuous	Number of 1st reminders over customer relationship
	Remind5*	Continuous	Number of 2nd reminders over customer relationship
	Remind6*	Continuous	Number of 3rd reminders over customer relationship
	Remind_install4*	Continuous	Number of 1st reminders per installment
	Remind_install5*	Continuous	Number of 2nd reminders per installment
	Remind_install6*	Continuous	Number of 3rd reminders per installment
	Remind7*	Continuous	Number of 1st reminders on short term consumer credit
	Remind8*	Continuous	Number of 2nd reminders on short term consumer credit
	Remind9*	Continuous	Number of 3rd reminders on short term consumer credit
	Remind_install7*	Continuous	Number of 1st reminders on short term credit per installment
	Remind_install8*	Continuous	Number of 2nd reminders on short term credit per installment
	Remind_install9*	Continuous	Number of 3rd reminders on short term credit per installment
	Default1*	Binary	Client has defaulted on his credit during the last two years
	Default2*	Binary	Client has defaulted on his credit during the last 15 years
	Remind_last*	Continuous	Summary score for the reminders on the last order
	Remind_1butlast*	Continuous	The same summary score for the one but last order
	Increase_remind*	Continuous	Increase/decrease in the summary score for reminders

* These variables were computed on internal company records about previous credit applications. All information used stemmed from before the date of application of the orders analyzed in the predictive model.

REFERENCES

1 Hand DJ and Henley WE (1997). Statistical classification methods in consumer credit scoring: a review. J Roy Stat Soc A 160: 523-541.

2 Heckman J (1979). Sample selection bias as a specification error. Econometrica 47: 153-161.

3 Hand DJ and Henley WE (1994). Can reject inference ever work? IMA J Math Appl Bus Ind 5: 45-55.

4 Myers JH and Forgy EW (1963). The development of numerical credit evaluation systems. J Am Stat Assoc 58: 799-806.

5 Thomas LC (2000). A survey of credit and behavioral scoring: forecasting financial risk of lending to customers. Int J Forecast 16: 149-172.

6 Hand DJ (2001). Modelling consumer credit risk. IMA J Manag Math 12: 139-155.

7 Thomas LC, Ho J and Scherer WT (2001). Time will tell: behavioural scoring and the dynamics of consumer credit assessment. IMA J Manag Math 12: 89-103.

8 Joanes DN (1994). Reject inference applied to logistic regression for credit scoring. IMA J Math Appl Bus Ind 5: 35-43.

9 Banasik J, Crook J, and Thomas L (2003). Sample selection bias in credit scoring models. J Oper Res Soc 54: 822-832.

10 Eisenbeis RA (1977). Pitfalls in the application of discriminant analysis in business, finance and economics. J Financ 32: 875-900.

11 Hsia DC (1978). Credit scoring and the equal credit opportunity act. Hastings Law J 30: 371-448.

12 Furnival GM and Wilson RW (1974). Regressions by leaps and bounds, Technometrics 16: 499-511.

13 Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J, and Vanthienen J (2003). Benchmarking state of the art classification algorithms for credit scoring. J Oper Res Soc 54: 627-635.

14 Stepanova M and Thomas L (2002). Survival analysis methods for personal loan data. Oper Res 50: 277-289.

15 Desai VS, Crook JN and Overstreet GA Jr. (1996). A comparison of neural networks and linear scoring models in the credit union environment. Eur J Oper Res 95: 24-37.

16 Davis RH, Edelman DB and AJ Gammerman (1992). Machine-learning algorithms for credit-card applications. IMA J Math Appl Bus Ind 4: 43-51.

17 Maddala GS (1992). Introduction to Econometrics Maxwell MacMillan Int. Editions: New York.

18 Rosenberg E and Gleit A (1994). Quantitative methods in credit management: a survey. Oper Res 42: 589-613.

19 Feelders AJ (2000). Credit scoring and reject inference with mixture models. Int J Intell Syst Account Financ Manag 8: 271-279.

20 Egan JP (1975). Signal detection theory and ROC analysis Academic Press: New York.

21 Baesens B, Viaene S, Van den Poel D, Vanthienen J and Dedene G (2002). Using bayesian neural networks for repeat purchase modelling in direct marketing. Eur J Oper Res 138: 191-211.

22 Cohen J, Cohen P, West SG, and Aiken LS (2003). Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences (3rd ed) Lawrence Erlbaum Associates: Mahwah, New Jersey.

23 De Long ER, De Long DM and Clarke-Pearson DL (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44: 837–845.

24 Morrison DG (1969). On the interpretation of discriminant analysis. J Mark Res 6: 156-163.

CHAPTER V

USING PREDICTED OUTCOME STRATIFIED SAMPLING TO REDUCE THE VARIABILITY IN PREDICTIVE PERFORMANCE OF A ONE-SHOT TRAIN-AND-TEST SPLIT FOR INDIVIDUAL CUSTOMER PREDICTIONS[13]

[13] This chapter is based on the following reference: Geert Verstraeten, Dirk Van den Poel, 2005. Using Predicted Outcome Stratified Sampling to Reduce the Variability in Predictive Performance of a One-Shot Train-and-Test Split for Individual Customer Predictions, submitted to ICDM'2006, where accepted papers will be published by Springer Verlag in the book 'Advances in Data Mining', as a volume of Lecture Notes in Artificial Intelligence (LNAI).

CHAPTER V:

USING PREDICTED OUTCOME STRATIFIED SAMPLING TO REDUCE THE VARIABILITY IN PREDICTIVE PERFORMANCE OF A ONE-SHOT TRAIN-AND-TEST SPLIT FOR INDIVIDUAL CUSTOMER PREDICTIONS

ABSTRACT

Since it is generally recognised that models evaluated on the data that was used for constructing them are overly optimistic, in predictive modeling practice, the assessment of a model's predictive performance frequently relies on a one-shot train-and-test split between observations used for estimating a model, and those used for validating it. Previous research has indicated the usefulness of stratified sampling for reducing the variation in predictive performance in a linear regression application. In this paper, we validate the previous findings on six real-life European predictive modeling applications for marketing and credit scoring using a dichotomous outcome variable. We find confirmation for the reduction in variability using a procedure we describe as predicted outcome stratified sampling in a logistic regression model, and we find that the gain in variation reduction is – also in large data sets – almost always significant, and in certain applications markedly high.

1. INTRODUCTION

In the latest decades, due to the increasing usage of customer identification cards and loyalty programs, companies in very diverse industries have been able to proceed in building large

transactional databases, recording all detailed interactions on an individual customer basis. Such interactions often include purchasing behavior, information requests, complaint behavior and subsequent complaint handling, survey information, etc. While this information serves for a large number of applications, in this study, we focus on the use of the transactional database for the predictive modeling of individual customer behavior, i.e. *individual customer predictions*. Indeed, ample previous research has proven that the historical information that resides in customer databases can aid in predicting future customer behavior on an individual level. For example, using the purchasing history of a given customer, companies have tried to assess e.g. whether this customer will (i) cease purchasing, (ii) respond to folders, (iii) be interested in certain products, (iv) increase his/her spending over their lifetime, (v) be able to refund granted credit, etc. Summarized, individual customer predictions mainly serve for targeted marketing and consumer credit scoring applications.

An intriguing concept in predictive modeling lies in the existence of overfitting. It is well established that predictive models have the tendency to be overly optimistic when their performance is measured on the same data used to build the models. Hence, adequate validation of such models require – at least – the usage of an independent holdout sample, a sample of data unseen by the classifier, that can be used to evaluate the true performance of the classifier [1]. As a practical solution to this, practitioners and researchers often start from a table of analysis, which is then split into two partitions: one used for estimating the model, and one used for validation. Very frequently, this split is performed using a random data partitioning. In his research, however, Malthouse [2] provided evidence that, in this approach, the results are highly dependent on the particular split of the data used. Accordingly, replicating the test using a different random partitioning might produce very different performances of the particular estimation and validation sets. Additionally, he examined the use of Winsorization and stratified sampling to a multiple linear regression problem in an attempt to reduce the variability of the results. While Winsorization focuses on imposing boundaries for outliers in the target variable, stratified sampling ensures that the data is split in such a way that the distribution of the target variable of the estimation and validation sets is as similar as possible.

Building on Malthouse's study, we note that a large number of applications in the domain of individual customer predictions do not imply the use of a continuous variable, but instead

attempt to predict a binary output variable. For example, we assess whether or not a customer will respond to an offer, will leave the company in a given time period, will purchase a certain product, will repay his credit, etc. In this study, we assess the usefulness of adapting the ideas in [2] to accommodate the use of a binary target variable, and we evaluate the benefits in terms of variance reduction on six real-life predictive modeling data sets.

The remainder of this paper is structured as follows. In the following section we describe the methodology that will be used in this paper, and defend the choices we make to perform the analyses. The next section covers a description of the data sets used in this study. Next, the results of the study are discussed, and in the last sections, the reader is offered conclusions, limitations, and suggestions for further research.

2. METHODOLOGY

2.1 One-shot train-and-test validation

Recently, the literature surrounding the assessment of a predictive model's performance has evolved drastically. To the best of our knowledge, current state-of-the-art domain knowledge prescribes that – in order to compare the predictive performance of different models – ten iterations of tenfold cross-validation should be applied. In tenfold cross-validation, the data is randomly split into ten subsamples. Subsequently, each sample serves iteratively as the holdout sample, while the other samples are used for model estimation. In order to compute the accuracy of the model, model performances are then averaged over the validation sets. In 10 x 10 cross-validation, the previously described procedure is performed ten times using a different random partitioning, and [3] have proven that this test shows a high degree of replicability of the test besides acceptable Type I and Type II errors. However, because the samples used in this test are not independent, in order to correct for the resulting increased Type I error, they apply the 'corrected resampled t-test' suggested by [4].

However, for a variety of reasons, the widespread use of the one-shot train-and-test validation for predictive modeling is not without merit. The fact that the 10 x 10 cross-validation test requires 100 models to be built and validated might be responsible for the fact that only few applications involve in such rigourous testing. Indeed, in a number of

situations, model builders require a more straightforward insight into the absolute performance of their models, while they would not necessarily proceed in testing whether significant differences occur between different model architectures. For example, a company that realizes that its customers are leaving will want to apply a predictive model in a timely manner in order to address the customers at risk. Hence, this company might continue to lose a lot of customers during a very extensive validation procedure, so time efficiency translates seamlessly into cost efficiency, and the company might choose to adopt a more straightforward validation procedure. Additionally, also in scientific readings, the use of the one-shot train-and-test validation is still popular. For example, many recent well-appreciated predictive modeling studies in *Marketing Science* report the use of a single split (see, e.g. [5, 6, 7]) for model validation. However, since it has been proven that the results of such a validation procedure are highly dependent on the particular split of the data used [2], in this study, we will consider the use of stratified sampling in an attempt to reduce this variability.

2.2 Predictive modeling technique

In the domain of individual customer predictions, given the variety of data available, it is possible to generate a large number of predictors that can serve in the model. It is not uncommon that such analyses are based on several hunderds of thousands of observations using several hunderds of candidate predictive variables. While at the advent of statistical theory, data sets of such magnitude were most likely beyond imagination, statistical techniques such as linear and logistic regression have been proven to show adequate predictive performance in such settings when benchmarked to other classifiers such as neural networks, decision trees, k-nearest neighbour, discriminant analysis and support vector machines [8, 9, 10]. They have become the standard method of analysing data with a discrete outcome variable in many fields in the eighties [1], and plausibly due to their ease-of-interpretation, regression models are still one of the main stalwarts of today's predictive model builders in industry [11]. Hence, in this study, we will use multiple logistic regression to predict customer behavior.

However, the use of a large number of candidate predictor variables implies that caution should be used when applying such models. First, the fact that the predictor variables are often closely related – often described as multicollinearity – has often been accused of influencing parameter signs and greatly boosting the variance of the parameter estimates, rendering them uninterpretable [1]. Still, it has been well documented that this phenomenon

need not hamper predictive performance on the condition that the multicollinearity persists in the validation set, and in the future population at large [12]. A second result of the large dimensionality is given by the existence of overfitting, implying that the inclusion of a large number of predictor features might lead to increases in the performance on the data used for calibrating the model, whereas *real* predictive performance – as measured when the model is applied to unseen data - does not increase, or even decreases. While we previously focused on the necessity of adequate model validation, feature selection can serve as a tool to reduce overfitting [1] and hence improve the predictive performance while at the same time reducing unnecessary or even unwanted complexity. In this study, we will apply a stepwise variable selection procedure, implying that features are entered iteratively according to the maximal contribution to the chi-square statistic, but the effects entered do not necessarily remain in the model. Each introduction of a new feature is followed by any possible removal of insignificant features, according to the Wald test for individual parameters [1]. In the analysis of the effect of stratification on the variance of predictive performance, we will compare the results of a model using all parameters, henceforth the *full* model, with the results of a model using only those parameters selected during a stepwise variable selection procedure.

2.3 Stratified sampling

In his recent study, [2] described the use of stratified sampling to reduce the variance of the estimates in a linear multiple regression problem. In this procedure, the author first sorts the data set according to the dependent variable. Next, strata are created by grouping consecutive observations, e.g. stratum one groups the first two observations, stratum two the following two observations, etc. Finally, the split between estimation and validation is performed by randomly assigning one of the observations in each stratum to the estimation set, and the remaining observation to the validation set. Hence, they use stratification to ensure that the distribution of the dependent variable is similar in both the training and test sets.

However, as already indicated, the domain of predictive modeling for targeted marketing and consumer credit scoring contains a number of core applications where the target variable is dichotomous. In its minimal form, in such cases, stratification implies that the sampling should ensure that the proportion of cases where the signal occurs (i.e. the incidence) is

equal in both training and validation sets. It should be clear that this stratification procedure is far less stringent that the procedure offered by [2], and will not necessarily imply that the variation is adequately reduced.

In their study on variable selection in logistic regression models, [13] illustrate a convienient way of transforming a logistic regression problem into a linear regression problem. Several steps suffice in this procedure. First, the logistic procedure is performed, and the predicted probabilities (which we label *pred*) are registered. Next, the binary outcome variable (*y*) is transformed via the equation $z = \log(pred / (1 - pred)) + ((y - pred) / (pred * (1 - pred)))$. Let the observations be weighed by a variable defined as $w = pred * (1 - pred)$. A regression procedure that uses the same predictors, yet using *z* instead of *y* as a dependent variable, and uses the weight variable *w*, will then obtain the same least squares estimates as the logistic regression. Interestingly, while designed for a very different application, this procedure does result in the creation of a new dependent variable, *z*, that is continuous, and that can serve to adapt the stratification procedure to resemble the procedure for multiple linear regression described in [2]. In the remainder of this paper, we will call this procedure *predicted outcome stratified sampling*, or shorter, POS sampling.

In this study, we will compare the variability of the predictive performance of a random partitioning into training and validation set with a stratified splitting as described in the procedure above. To this end, we will perform both the random and the stratified splitting 100 times using a different random number sequence. Note, however, that in both procedures, we ensure to control for the incidence, so that for example in a credit scoring problem where 1% of the customers fail to repay their debts, the defaulters are proportionally distributed across estimation and validation sets, so that the percentage of defaulters is constant over the different sets. Another difference in comparison with [2] is that, in our study, the estimation set will be twice as large as the validation set, since it is more common that a smaller amount of the observations are held out for validation purposes. To summarize, this implies that strata of three consecutive observations will be created based on a ranking according to the *z* values described in [13], whereby two observations of each stratum are randomly chosen for estimating the model, while the remaining observation is used in validating the model.

For model evaluation purposes, because we do not always possess profit information in every application, we will not use the gains chart used by [2], but instead we report the area under the receiver operator characteristics curve (AUC), since this measure evaluates the performance of a given classifier regardless of the choice of a particular discretisation cutoff. An intuitive interpretation of the AUC is that it provides an estimate that a randomly chosen instance of class 1 is correctly rated higher than a randomly selected instance of class 0 [14].

3. DATA

In this study, we make use of six real-life proprietary European predictive modeling data sets. All data sets were constructed for company-driven applications, and hence represent a sizeable test bed for comparing alternative predictive models. All cases are binary classification cases, and applications lie in the domains of targeted marketing and credit scoring. In Table 1, we present some descriptive statistics about the datasets used, namely (i) the case description, (ii) the industry of the application, (iii) the incidence of the target feature, e.g. the percentage of churners, buyers, defaulters, etc present in the data set, (iv) the number of observations, (v) the condition index, representing the degree of multicollinearity present in the data set. All data sets involved show high degrees of multicollinearity, considering the fact that condition indexes of 100 or more appear to be large, causing substantial variance inflation and great potential harm to regression estimates [12], and (vi) the number of predictive features in the data set.

Table 1. Descriptive statistics of the data sets used

(i) Case	(ii) Industry	(iii) Incid	(iv) Obs	(v) C.I.	(vi) Features
Loyalty	Retail	0.4738	878	111	35
Spending	DIY retail	0.2814	3 827	1 442	15
Partial Churn	Retail	0.2515	32 371	241	45
Churn	Subscription services	0.1307	143 198	767	167
Targeting	Retail	0.3082	741 234	4030	100
Credit Scoring	Mailorder	0.0089	38 064	114 593	137

4. RESULTS

Table 2 presents some descriptive statistics of the predictive performance of the different models. In this table, we distinguish between the predictive performance on the estimation

Table 2. Overview of descriptives of the variability in the predictive performance of the different models

Loyalty			Estimation					Validation					Overfitting		
Model	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev
Full, Random	0.7852	0.7558	0.8142	0.0584	0.0120	0.7223	0.6645	0.7869	0.1224	0.0253	0.0629	-0.0285	0.1497	0.1782	0.0367
Full, Stratified	0.7827	0.7758	0.7934	0.0176	0.0033	0.7282	0.6895	0.7532	0.0637	0.0119	0.0545	0.0226	0.1019	0.0793	0.0143
Stepwise, Random	0.7610	0.7302	0.7918	0.0616	0.0122	0.7413	0.6765	0.7976	0.1212	0.0252	0.0197	-0.0674	0.1154	0.1828	0.0371
Stepwise, Stratified	0.7584	0.7473	0.7691	0.0218	0.0047	0.7477	0.7162	0.7720	0.0559	0.0105	0.0107	-0.0218	0.0511	0.0729	0.0143
Spending			Estimation					Validation					Overfitting		
Model	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev
Full, Random	0.7621	0.7483	0.7795	0.0312	0.0064	0.7587	0.7260	0.7865	0.0604	0.0126	0.0035	-0.0376	0.0535	0.0911	0.0189
Full, Stratified	0.7635	0.7612	0.7679	0.0067	0.0011	0.7562	0.7425	0.7611	0.0186	0.0030	0.0072	0.0014	0.0218	0.0204	0.0034
Stepwise, Random	0.7620	0.7464	0.7764	0.0300	0.0063	0.7619	0.7303	0.7928	0.0625	0.0131	0.0001	-0.0439	0.0461	0.0899	0.0190
Stepwise, Stratified	0.7635	0.7593	0.7673	0.0079	0.0016	0.7600	0.7454	0.7656	0.0202	0.0033	0.0035	-0.0046	0.0171	0.0217	0.0036
Partial Churn			Estimation					Validation					Overfitting		
Model	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev
Full, Random	0.8197	0.8147	0.8251	0.0104	0.0019	0.8159	0.8052	0.8244	0.0191	0.0038	0.0038	-0.0096	0.0199	0.0295	0.0057
Full, Stratified	0.8191	0.8187	0.8196	0.0010	0.0002	0.8171	0.8158	0.8179	0.0020	0.0004	0.0020	0.0009	0.0034	0.0025	0.0005
Stepwise, Random	0.8190	0.8136	0.8245	0.0108	0.0019	0.8157	0.8047	0.8246	0.0198	0.0038	0.0033	-0.0109	0.0197	0.0307	0.0057
Stepwise, Stratified	0.8184	0.8179	0.8192	0.0012	0.0002	0.8170	0.8160	0.8179	0.0019	0.0004	0.0014	0.0004	0.0031	0.0028	0.0005
Churn			Estimation					Validation					Overfitting		
Model	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev
Full, Random	0.7759	0.7717	0.7803	0.0086	0.0016	0.7711	0.7618	0.7805	0.0187	0.0032	0.0048	-0.0085	0.0185	0.0270	0.0047
Full, Stratified	0.7758	0.7751	0.7766	0.0015	0.0003	0.7714	0.7697	0.7727	0.0030	0.0006	0.0044	0.0029	0.0063	0.0033	0.0006
Stepwise, Random	0.7737	0.7673	0.7777	0.0103	0.0018	0.7700	0.7601	0.7785	0.0184	0.0032	0.0036	-0.0091	0.0169	0.0260	0.0048
Stepwise, Stratified	0.7737	0.7684	0.7751	0.0067	0.0008	0.7701	0.7672	0.7719	0.0047	0.0009	0.0035	-0.0001	0.0060	0.0061	0.0011
Targeting			Estimation					Validation					Overfitting		
Model	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev
Full, Random	0.7400	0.7387	0.7414	0.0027	0.0005	0.7401	0.7372	0.7426	0.0054	0.0009	-0.0001	-0.0039	0.0042	0.0081	0.0014
Full, Stratified	0.7401	0.7400	0.7402	0.0002	2.8E-05	0.7398	0.7397	0.7399	0.0003	4.7E-05	0.0003	0.0002	0.0004	0.0003	4.8E-05
Stepwise, Random	0.7399	0.7386	0.7413	0.0027	0.0005	0.7401	0.7372	0.7426	0.0054	0.0009	-0.0002	-0.0040	0.0040	0.0080	0.0014
Stepwise, Stratified	0.7400	0.7399	0.7401	0.0002	3.7E-05	0.7398	0.7397	0.7399	0.0003	5E-05	0.0002	4.9E-05	0.0003	0.0003	5.5E-05
Credit Scoring			Estimation					Validation					Overfitting		
Model	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev	Mean	Min	Max	Range	Stddev
Full, Random	0.9031	0.8823	0.9179	0.0356	0.0065	0.8456	0.7889	0.8948	0.1058	0.0193	0.0575	-0.0044	0.1274	0.1318	0.0243
Full, Stratified	0.9026	0.8889	0.9121	0.0232	0.0048	0.8480	0.7974	0.8745	0.0771	0.0138	0.0546	0.0233	0.1146	0.0914	0.0161
Stepwise, Random	0.8554	0.6709	0.9004	0.2295	0.0567	0.8297	0.6052	0.8983	0.2931	0.0600	0.0257	-0.0302	0.0942	0.1243	0.0235
Stepwise, Stratified	0.8582	0.6658	0.8901	0.2243	0.0541	0.8333	0.6234	0.8780	0.2545	0.0583	0.0249	-0.0140	0.0712	0.0852	0.0172

sample, the validation sample, and overfitting, which is defined as the difference between estimation and validation sample within a single split. For each of these samples, we report the mean, minimal, maximal, the range (being the difference between the maximal and the minimal), and the standard deviation of the AUC performance measure. The most important conclusion from this table is that the range as well as the standard deviation of the AUC is *in all cases* reduced by performing the stratified sampling procedure. We also note that these findings are consistent across the estimation and validation samples, and are also reflected in the variation of overfitting. Additionnally, no large differences can be found when a full model is computed versus a model that uses a stepwise variable selection procedure, implying that the results are not sensitive to the particular variables used in the different models. It is clear, however, that the improvements vary in size across the different data sets. Hence, Table 3 presents a summarized overview of the reduction in variance that can be reached by using POS sampling in a logistic regression model.

Table 3. Factor with which the variance decreases by using stratification.

	Full Model			Stepwise Model		
	Estimation	Validation	Overfitting	Estimation	Validation	Overfitting
Loyalty	**13.03**	**4.54**	**6.56**	**6.83**	**5.75**	**6.74**
Spending	**35.01**	**17.43**	**29.87**	**14.94**	**15.87**	**27.97**
Partial Churn	**102.53**	**81.39**	**130.49**	**80.00**	**81.64**	**114.21**
Churn	**22.35**	**29.17**	**52.49**	**4.75**	**11.73**	**19.79**
Targeting	**262.71**	**373.60**	**813.55**	**150.78**	**322.24**	**605.13**
Credit Scoring	**1.87**	**1.95**	**2.29**	1.10	1.06	**1.86**

The numbers in Table 3 should be interpreted as follows. The upper left figure is reached by dividing the variance in the predictive performance of the estimation set of the full model of the 'Loyalty' application when a random partitioning is used, by the corresponding variance when POS sampling is used. Hence, in that particular situation, the variation of the random partitioning is over 13 times as large as the variation of the POS sampling. Levene's test for the homogeneity of variance [15] was applied in order to analyse the significance of the differences in variation. Significant drops in the variance (at $p < 0.01$) are indicated in bold face. We conclude that, in all models but the stepwise model of the 'Credit scoring 1' data set, the drop in variance is statistically highly significant.

5. CONCLUSIONS

In business as well as academia, the use of a single shot train-and-test split to perform model assessment is not uncommon. Surprisingly, even when large data sets are used, the results of the models can vary strongly when data is partitionned into an estimation and a validation sample on a random basis. The ongoing use of a single split as a validation procedure implies that model builders may benefit from a reduction of variability in model performance. In this study, we provide evidence that the insights of [13] regarding the similarities between linear and logistic regression can be used to adapt the stratification procedure suggested in [2] in order to apply a variance reduction heuristic that can accommodate predictive models with a dichotomous outcome variable. In this study, we have computed the reduction in variance of the predictive performance on six real-life European predictive modeling applications for marketing and credit scoring. The predicted outcome stratified (POS) sampling used consistently succeeds in reducing the variance of the predictive performance in the estimation and validation samples, but also in the overfitting, and this effect was confirmed across a model containing all variables, and a model containing only those variables selected by a stepwise variable selection procedure. However, across the different applications, the gains that can be used in terms of variance reduction vary. In the least successful case, the gains are non-significant, whereas in the most successful application, the variation in overfitting is over 800 times lower when POS sampling is used instead of random sampling.

This has important implications. In those situations that time is only available to compute one (or a limited amount) of validation iterations, the use of a random split seems never more justified than the use of the stratified split procedure suggested here. Indeed, the only requirement to perform a stratified split is the outcome of a run of the logistic regression model on the total data set. The gain exists in the fact that it is (sometimes far) more likely that the resulting performance assessment will be more accurate.

6. LIMITATIONS AND ISSUES FOR FURTHER RESEARCH

This study has a number of limitations. In contrast to the paper of [2], in this study, our focus lies on the absolute question instead of the relative question. Indeed, the center of attention in this study lies in model assessment, whereas [2] focusses on model selection, i.e. deciding

which model offers the best predictive qualities. Due to the data complexities involved in an analysis on a test bed of large data sets, we did not compare different classifiers, variable selection techiques, etc. Hence, future research might be directed towards assessing the usefulness of POS sampling in logistic regression in order to compare the differences in performance of alternative predictive models.

Additionally, in this study, we have used the score $z = \log(pred / (1 - pred)) + ((y - pred) / (pred * (1 - pred)))$ in order to create the continuous version of the response variable. As mentionned by several members of the exam committee of this dissertation, a more direct approach would exist in computing the log-odds of the predicted success probabilities as a transformed response variable. In future research, it would be interesting to compare the performance of both approaches.

ACKNOWLEDGEMENTS

The authors would like to express their highest gratitude to Wouter Buckinx, Jonathan Burez, Bart Larivière and Bernd Vindevogel, who contributed the data sets they gathered during their PhDs in order to enable the experiments performed in this study.

REFERENCES

1. Hosmer D.W. and Lemeshow S.: Applied Logistic Regression, John Wiley & Sons, New York (1989)
2. Malthouse E.C.: Assessing the performance of direct marketing scoring models. Journal of Interactive Marketing 15(1) (2001) 49-62
3. Bouckaert R. and Frank E.: Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In: Dai H., Srikant R. and Zhang C. (eds.): Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'2004). Springer (2004) 3-12
4. Nadeau C. and Bengio Y.: Inference for the generalization error. Machine Learning 52 (2003) 239-281
5. Montgomery A.L., Li S., Srinivasan K. and Liechty J.C.: Modeling Online Browsing and Path Analysis Using Clickstream Data. Marketing Science 23(4) (2004) 579-595
6. Park Y.-H. and Fader P.S.: Modeling Browsing Behavior at Multiple Websites. Marketing Science 23(3) (2004) 280-303
7. Swait J. and Andrews R.L.: Enriching Scanner Panel Models with Choice Experiments. Marketing Science 22(4) (2003) 442-460
8. Baesens B., Van Gestel T., Viaene S., Stepanova M., Suykens J., and Vanthienen J.: Benchmarking State of the Art Classification Algorithms for Credit Scoring. Journal of the Operational Research Society 54 (2003) 627-635
9. Davis R.H., Edelman D.B. and Gammerman A.J.: Machine-Learning Algorithms for Credit-card Applications. IMA Journal of Mathematics Applied in Business and Industry 4 (1992) 43-51
10. Dasgupta C.G., Dispensa G.S. and Ghose S.: Comparing the predictive performance of a neural network model with some traditional market response models. International Journal of Forecasting 10(2) (1994) 235-244
11. Thomas L.C.: A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Customers. International Journal of Forecasting 16 (2000) 149-172
12. Belsley D.A.: Conditioning diagnostics, collinearity and weak data in regression. John Wiley & Sons, New York (1991)
13. Hosmer D.W., Jovanovic B., & Lemeshow S.: Best subsets logistic regression. Biometrics 45 (1989) 1265-1270

14. Hanley J.A. and NcNeil B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve, Radiology 148 (1983) 839-843
15. Levene, H.: Robust Tests for the Equality of Variance. In Olkin I. (ed): Contributions to Probability and Statistics. Stanford University Press, Palo Alto, CA (1960) 278-292

CHAPTER VI

EVALUATING THE PERFORMANCE COST OF IMPROVED FACE VALIDITY: BENCHMARKING FEATURE SELECTION TECHNIQUES IN LOGISTIC REGRESSION FOR INDIVIDUAL CUSTOMER PREDICTIONS[14]

[14] This chapter is based on the following reference: Geert Verstraeten, Dirk Van den Poel, 2005. Evaluating the Performance Cost of Improved Face Validity: Benchmarking Feature Selection Techniques in Logistic Regression for Individual Customer Predictions, submitted to PAKDD'2006, where accepted papers will be published by Springer Verlag as a volume of Lecture Notes in Artificial Intelligence (LNAI).

CHAPTER VI:

EVALUATING THE PERFORMANCE COST OF IMPROVED FACE VALIDITY: BENCHMARKING FEATURE SELECTION TECHNIQUES IN LOGISTIC REGRESSION FOR INDIVIDUAL CUSTOMER PREDICTIONS

ABSTRACT

The usefulness of predictive models depends on several factors, but it could be argued that predictive performance and acceptability are the most important criteria. However, intuitively, a trade-off seems to exist between these two concepts, as highly complex predictive models do not always offer good interpretability. In this study, we evaluate the use of 'Forward selection with Sign Restrictions' (FSR) as a feature selection technique for multiple logistic regression. Using this procedure, we ensure a selection technique where all parameter signs in the final model are equal to their univariate parameter signs, resulting in models that are much more likely to be understood and accepted by managers, employees and customers. In an empirical study on nine real-life European predictive modeling data sets, we quantify the loss in performance due to this restriction, and we provide evidence that the inclusion of such a restriction does not generally hamper performance.

1. INTRODUCTION

Following the evolutions in computational power and data storage, and the ongoing shift in marketing from a product-oriented approach to a customer-oriented approach [1], companies

have increasingly engaged in tracking and understanding individual customer behavior. Moreover, many have started to acquire the competences needed to predict future customer behavior on an individual customer level. Such 'individual customer predictions' are currently mainly used in two distinct domains in predictive modeling: targeted marketing and consumer credit scoring. In terms of marketing applications, companies currently attempt to predict which customers will cease purchasing, respond to folders, be interested in certain products, increase their spending over their lifetime, etc. whereas credit scoring is focused on predicting the creditworthiness of customers – or more in general loan applicants.

In his invited paper for the PAKDD conference in 2003, Bradley [2] distinguishes two main factors influencing the likelihood of obtaining a high quality, useful predictive model. Clearly, the foremost indicator of its quality lies in the predictive performance, and the robustness of the model with respect to small changes. For example, the raison d'être of a targeting model lies in the capability of distinguishing between responders and nonresponders. As a second main factor, he considers the ease at which the insight gained by the model can be communicated to the model's users [2]. Indeed, besides proving predictive accuracy, it is not surprising that the proposed models should be acceptable. Suppose management of a bank knows that younger customers clearly exhibit a higher risk of churning (i.e., closing their account). A good predictive model, however, may contain a large number of parameters. Due to several reasons that will be explained below, it is not excluded that the parameter sign of the feature 'age' will be reversed in the model with the best predictive capabilities. Not surprisingly, model builders may find it difficult in this situation to convince management of the usefulness of the proposed model, even if the predictive performance is adequately validated. Moreover, management is not the only actor that needs to be convinced of the validity of a model. As a second example, consider a credit scoring model for a catalogue retailer. For this retailer, it might be beneficial that all employees as well as customers understand the motivations why a customer is not accepted for credit according to the predictive model used. Note also that in some countries, it is even a legal requirement to disclose the credit scoring formula used. In these cases, it is also important that the features in the model contribute in an intuitive way to the outcome. It may seem very unrational to customers that they would be rejected because they have a *good* credit repayment history, so also here, parameter signs can be of crucial importance to the acceptance of a predictive model.

In an interesting experiment, Pazzani and Bay [3] questionned undergraduate students on whether they would use a set of proposed models for determining the salaries of baseball players. In their research, they have proven that, if the signs of the parameters in the multiple model correspond to the univariate parameter signs, the acceptance rate of the model will rise. Additionally, they show that a feature selection technique that restricts the parameter signs offered acceptable predictive accuracy. However, the results of their study cannot directly be extrapolated towards the domain of individual customer predictions, as (i) they did not make use of significance tests required to evaluate differences in predictive model performances, and (ii) they made use of machine learning data sets, which are essentially remarkably smaller in size, and hence differ substantially from real-life predictive modeling applications.

A number of similarities exist between the decisions of the students and the decisions of the managers and the customers of the previous examples. The manager's belief about the relationship between age and the churn probability, and the customer's belief about the impact of his good credit repayment history on his credit score can be considered equivalent to the knowledge of the univariate parameter sign. Hence, predictive model builders could include a set of restrictions on the parameter signs of the multiple regression model. However, it is not unlikely that the inclusion of such a restriction will hamper predictive accuracy, as the model that is optimal in terms of predictive preformance will only rarely fully comply with the parameter restrictions (infra). In other words, a trade-off may be seen to exist between predictive accuracy and the acceptability of the model. In this paper, it is our goal to assess the size of this trade-off: how much of the predictive performance is lost due to the restriction that all parameter signs of the multiple regression should correspond to their univariate counterparts. We will report the empirical results of applying the proposed feature selection technique on nine real-life datasets used for different predictive modeling applications across different European industries, in order to give theoreticians and practitionners an idea of the plausible performance cost of improving the face validity of their predictive models.

The remainder of this paper is structured as follows. In the following section we describe the methodology that will be used in this paper. We explain why we focus on comparing different feature selection techniques for logistic regression models, and defend the choices we make to perform the analyses. The next section covers a description of the data sets that

are used in this study. Next, the results of the study are discussed in depth, and in the last sections, the reader is offered conclusions, limitations, and suggestions for further research.

2. METHODOLOGY

2.1 Predictive modeling technique

In the predictive modeling literature, and especially in the domain of credit scoring, evidence exists that traditional statistical methods, such a logistic regression, show comparable predictive results when benchmarked to other classifiers such as neural networks, decision trees, k-nearest neighbour, discriminant analysis and support vector machines [4,5]. Additionally, plausibly due to their ease-of-interpretation, regression models are still one of the main stalwarts of today's predictive model builders in industry [6]. While logistic regression does not strictly belong to the toolset of the data mining community, it can surely be considered an important tool for data analysis [7], which is essentially the underlying goal of this study. Hence, in this study, we will use multiple logistic regression to predict customer behavior. However, as with several other tools of data analysis, it is well documented that regression models suffer from overfitting [8], implying that the inclusion of a large number of predictor features might lead to increases in the performance on the data used for calibrating the model, whereas *real* predictive performance – as measured when the model is applied to unseen data - does not increase, or even decreases. In these cases, feature selection can serve as a tool to reduce overfitting and hence improve the predictive performance while at the same time reducing unnecessary or even unwanted complexity.

2.2 Feature selection techniques

In this paragraph, we will discuss the different feature selection techniques used in this study. The first four techniques are commonly available in most statistical software packages, and are often used as benchmarking techniques [9,10]. The latter technique involves an alteration of the standard techniques.

Forward selection (FWD) is clearly the most straightforward selection technique. Starting from an intercept-only model, iteratively, the feature that offers the largest chi-square improvement is added to the model provided that the parameter is significant. The procedure

is stopped when no significant parameters can be added to the model. Note that effects entered in the model are never again removed.

Backward selection (BWD) can be considered the opposite procedure of forward selection. Starting from a complete model, iteratively, the least significant feature is removed, based on the Wald test for individual parameters. This procedure ends when no insignificant features remain in the final model. However, once an effect is removed, it never re-enters the model.

Stepwise selection (SW) can be considered a combination of the previous techniques. It is comparable to forward selection, in the sense that features are entered iteratively according to the chi-square statistic, but the effects entered do not necessarily remain in the model. Each introduction of a new feature is followed by any possible removal of insignificant features.

Best subset selection (BS) differs substantially from the previous techniques. It uses the leap-and-bound algorithm of [11] to detect the best model for all possible model sizes (number of features), requiring a minimum of arithmetic, and the possibility for finding the best subsets without examining all possible subsets. While being substantially more elaborate than previous techniques, this is the only technique that ensures that the optimal model be detected. Nevertheless, the computing time needed for this procedure increases exponentially as the number of features in the data set increases.

None of the previous techniques take into account the sign of the parameter estimates. Therefore, it is not unlikely that the final models include parameter signs that do not correspond to their univariate counterparts. According to [12], at least three reasons can be given that explain the occurence of sign violations. First, computational error might exist when the magnitudes of features differ drastically, due to a lack of precision of some computational procedures. Second, as in forward selection, coefficients that do not significantly differ from zero might remain in the model, and the sign of these parameters might be erratic. And third, due to multicollinearity, it is possible that a certain feature compensates for another highly correlated feature, so that the parameter signs of these positively correlated features might be opposite. The following feature selection technique rules out the problems mentioned above.

Forward selection with Sign Restrictions (FSR) is very similar to forward selection, discussed above, and only differs in the fact that features are only included when the signs of all parameters in the suggested model correspond to their univariate counterparts. Hence, the procedure ends when no additional parameters can be added to the model that (i) are significant *and* (ii) result in a full model where all parameter signs correspond to the

univariate parameter signs. As with forward selection, effects entered are never again removed from the model.

2.3 Significance testing

In this study, it is our goal to evaluate the statistical significance of the differences in performance of the examined feature selection techniques. In his experimental tests, Dietterich [13] has provided evidence that the resampled *t* test and the 10-fold cross-validated *t* test show an excessive Type I error, indicating that these tests incorrectly detect a difference when no difference exists. As an alternative, he proposed a test based on five iterations of twofold cross-validation, the 5 x 2 cv test, a heuristic test that has proven to show acceptable Type I and Type II errors. In subsequent work, however, [14] argue that, besides Type I and Type II error, also the replicability of a test is of importance. In other words, if the outcome of a test is strongly dependent on the particular random partitioning used to perform it, this test is inferior when compared to a test showing a high replicability, ceteris paribus. Based on their experiments on machine learning problems, they suggest to expand the test suggested by [13] to a similar test based on ten iterations of tenfold cross-validation, i.e. the 10 x 10 cv test. However, in such a cross-validation procedure, the independence assumption on the data sets is highly flawed, and hence the variance of this test is underestimated because the samples are no longer independent. In order to correct for the resulting increased Type I error, they apply the 'corrected resampled t-test' suggested by [15]. Hence, they use the following statistic, and describe it as the 'corrected repeated *k*-fold cv test', where the *t* test has (*r* x *k)* - 1 degrees of freedom:

$$t = \frac{\frac{1}{k.r}\sum_{i=1}^{k}\sum_{j=1}^{r} x_{ij}}{\sqrt{\left(\frac{1}{k.r} + \frac{n_2}{n_1}\right)\hat{\sigma}^2}}$$

where n_1 and n_2 represent respectively the size of the training and test set, in an *r*-times *k*-fold cross-validation procedure, and where x_{ij} represents the difference in predictive accuracy, and $\hat{\sigma}^2$ represents the variance of the accuracy differences across the *r* x *k* evaluations.

2.4 Predictive accuracy

In the results, we will use two measures of predictive accuracy. First and foremost, we compute the percentage of cases correctly classified (PCC). In this measure, the outcome of the predictive models are discretized in proportion to the incidence of the data set. The significance tests described above were applied in order to detect statistical significance, but discretisation levels were varied to evaluate the sensitivity of the results to the chosen cutoff value. In addition to the PCC measure, we also report the area under the receiver operator characteristics curve (AUC, or AUROC), since this measure evaluates the performance of a given classifier regardless of the choice of a particular discretisation cutoff.

3. DATA

In this study, we make use of nine real-life proprietary European predictive modeling data sets. All data sets were constructed for company-driven applications, and hence represent a sizeable test bed for comparing alternative predictive models. All cases are binary classification cases, and applications lie in the domains of marketing and credit scoring. In order to ensure tractability of the cases, in light of the extensive significance tests described supra, a sampling procedure was used to ensure that a maximum of 50 000 observations was used per data set. In these cases, the sampling was performed according to the stratification procedure suggested in [16] to ensure representativeness of the subsamples. In Table 1, we present some descriptive statistics about the datasets used, namely (i) the case description, (ii) the industry of the application, (iii) the incidence of the target feature, e.g. the percentage of churners, buyers, defaulters, etc present in the data set, (iv) the number of observations, (v) the condition index, representing the degree of multicollinearity present in the data set. All data sets involved show high degrees of multicollinearity, considering the observation that condition indexes of 100 or more appear to be large, causing substantial variance inflation and great potential harm to regression estimates [8], (vi) the number of predictive features in the data set and (vii) the number of features where the sign in the full model *does not* correspond to the sign in the univariate model, representing the number of sign violations.

Table 1. Descriptive statistics of the data sets used

(i) Case	(ii) Industry	(iii) Incid	(iv) Obs	(v) C.I.	(vi) Features	(vii) Sign Violations
Loyalty	Retail	0.4738	878	111	35	15
Spending	DIY retail	0.2814	3 827	1 442	15	5
Partial Churn 1	Retail	0.2515	32 371	241	45	18
Partial Churn 2	Banking	0.0609	527 549	57 449	178	71
Churn 1	Subscription	0.1307	143 198	767	167	69
Churn 2	Telecom	0.0276	237 001	583	424	187
Targeting	Retail	0.3082	741 234	4030	100	51
Credit Scoring 1	Mailorder	0.0089	38 064	114 593	137	47
Credit Scoring 2	Mailorder	0.0176	1 732 214	1135	239	88

These descriptive statistics clearly indicate the relevance of the problem. In the full models, on average only 60 % of the parameter signs correspond to the relevant univariate parameter signs. In one of the applications, no less than 187 parameters have an influence that is exactly opposite to the influence that may be expected based on the univariate relationships. It is not unlikely that the face validity of the models is severely hampered by such an obvious presence of counter-intuitive parameter influences. In the following section, we analyse the relationship between model size and face validity and we evaluate the performance costs that can be attributed to complying to the parameter restrictions described previously.

4. RESULTS

4.1 The inverse relationship between model size and face validity

In the previous section, we already indicated the lack of face validity of the full model. In Fig. 1, we deepen this understanding by plotting the evolution of the percentage of 'correct' signs as model size increases. In an iterative process, we introduce the feature that offers the largest chi-square improvement until the full model is reached. In essence, this embodies a forward feature selection technique without the customary stopping rule based on feature signficance. The results of this analysis are unambiguous: whereas small models fully comply with the univariate signs, a negative relationship exists between model size and face validity: the more features are introduced, the lower the percentage of 'correct' parameters will be. All relationships were tested to be significantly negative, at a significance level of $p < 0.0001$. Additionally, in these same graphs, a black vertical line was used to indicate the

selected model size if a stopping rule would be introduced to prevent the inclusion of nonsignificant features. Note that this line presents the result of the forward feature selection technique described in a previous section. The graphs show that, in two distinct examples – i.e. loyalty and spending – the forward selection technique succeeds in selecting models where all parameter signs in the selected model correspond to their univariate counterparts. However, in all other applications, a number of counterintuitive parameter signs remain in the selected model.

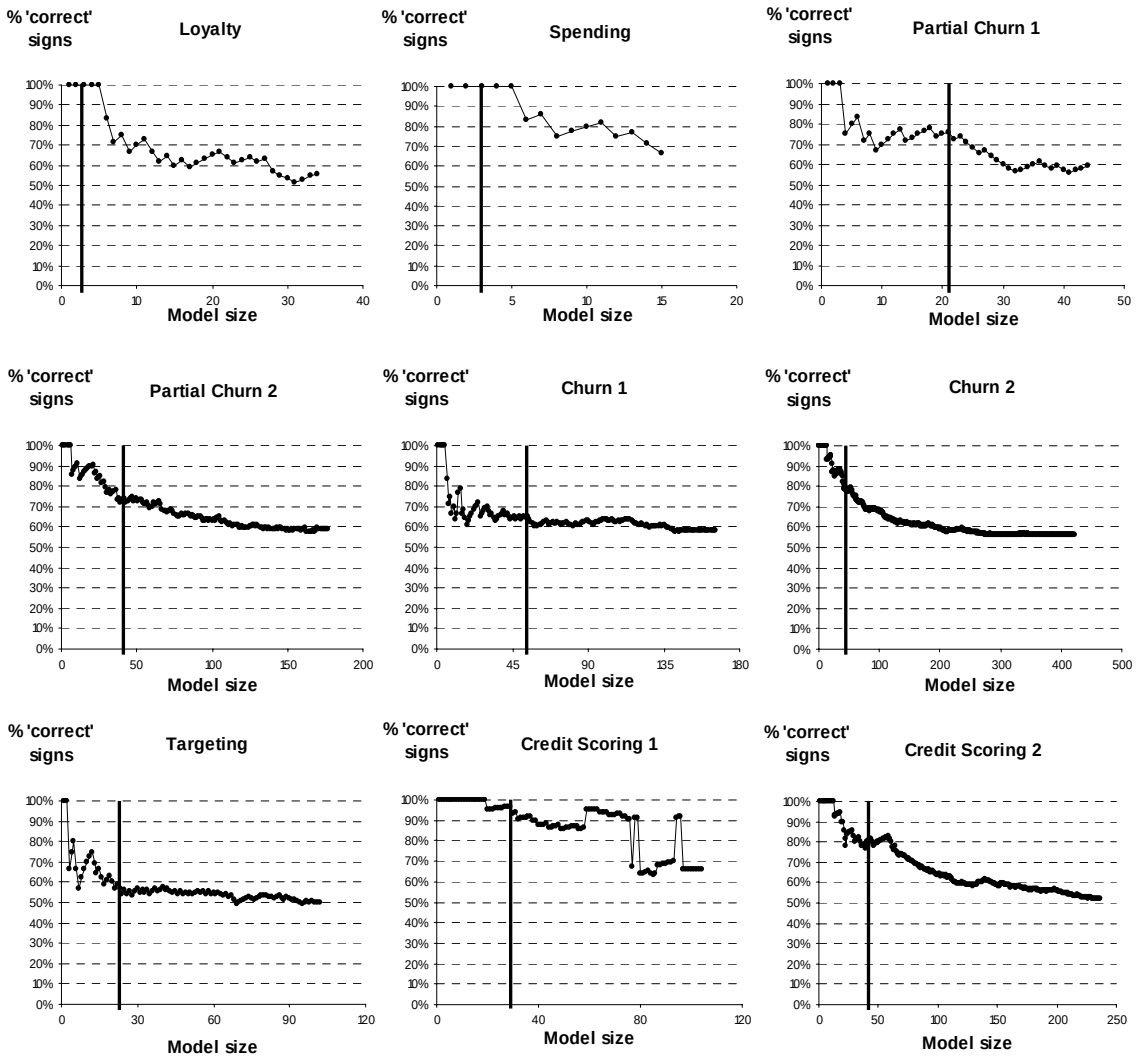


Fig. 1. The inverse relationship between model size and face validity

Considering the negative relationship between model size and face validity, the use of a forward feature selection technique implies that the face validity will increase. Note that this finding is consistent with [3], who also found that FWD mitigates, but does not eliminate, the problem of sign violations. In Table 2, we present the percentage of 'correct' signs of the

different feature selection techniques. From this table, it is clear that feature selection techniques consistently reduce the sign violations when compared to the full model. One noticeable exception can be found in the BWD selection of the 'Churn 1' case, exhibiting a slightly lower percentage of corresponding signs.

To conclude, the feature selection techniques proposed in Table 2 do not necessarily solve sign violations. In the following subsection, we will evaluate the performance cost of using FSR to ensure all parameter signs in the final model correspond to the univariate parameter signs.

Table 2. Percentage of correct signs per feature selection technique

	FWD	BWD	SW	BS	FULL
Loyalty	100 %	100 %	100 %	77.78 %	57.14 %
Spending	100 %	100 %	100 %	85.71 %	66.67 %
Partial Churn 1	76.19 %	76.19 %	76.19 %	80.00 %	60.00 %
Partial Churn 2	73.91 %	63.64 %	78.26 %	60.00 %	60.11 %
Churn 1	63.64 %	54.84 %	64.15 %	60.78 %	58.68 %
Churn 2	79.55 %	65.28 %	81.08 %	56.00 %	55.90 %
Targeting	54.17 %	51.35 %	61.91 %	66.67 %	49.00 %
Credit Scoring 1	83.33 %	82.76 %	88.00 %	73.17 %	65.69 %
Credit Scoring 2	81.82 %	70.37 %	90.00 %	65.79 %	63.18 %

4.2 The performance cost of increasing face validity

As mentioned previously, in this study, we adapt forward selection in such a way that a parameter is only included into the model if all parameter signs correspond to their univariate couterparts, and we have described this technique previously as Forward selection with Sign Restrictions (FSR). In this crucial section, it is our goal to evaluate the performance cost of increasing face validity by imposing parameter restrictions. To this end, we will compare the predictive performance of FSR with the performance of some of the most frequently used feature selection techniques, being forward (FWD), backward (BWD), stepwise (SW) selection, and best subset selection (BS) introduced by [11]. Note that, in order to ensure tractability of the cases, the best subset technique was applied on a preselection of the 80 univariately best performing features per application. In all other situations, all features belonging to the full model were introduced into the selection techniques.

Table 3 represents the results of these analyses. While the left-hand side of the table represents the model sizes and the (cross-validated) predictive performance of the different feature selection techniques averaged over the 10 x 10 cross-validation runs, the right-hand side represents the results of the significance tests described above. The values on the right-hand side should be interpreted as the probability that the feature selection techniques are performing equally well. If this value lies below 0.05, we consider the differences in PCC predictive accuracy between two techniques to be significant, and we conclude that one technique significantly outperforms the other. Visually, significant values are indicated in bold face.

Table 3. Predictive performance of the different feature selection techniques

Loyalty								
model	size	AUC	PCC	BWD	SW	BS	FSR	FULL
FWD	3	0.7559	0.6941	0.9668	1	0.6578	1	0.2198
BWD	3	0.7571	0.6946		0.9668	0.7011	0.9668	0.2040
SW	3	0.7559	0.6941			0.6578	1	0.2198
BS	9	0.7556	0.6984				0.6578	0.0948
FSR	3	0.7559	0.6941					0.2198
FULL	35	0.7333	0.6805					
Spending								
model	size	AUC	PCC	BWD	SW	BS	FSR	FULL
FWD	3	0.7624	0.7326	1	1	0.5538	0.7494	0.3928
BWD	3	0.7624	0.7326		1	0.5538	0.7494	0.3928
SW	3	0.7624	0.7326			0.5538	0.7494	0.3928
BS	7	0.7502	0.7301				0.6819	0.9095
FSR	8	0.7591	0.7316				.	0.5509
FULL	15	0.7569	0.7296					
Partial Churn 1								
model	size	AUC	PCC	BWD	SW	BS	FSR	FULL
FWD	21	0.8177	0.7905	1	1	**<.0001**	**<.0001**	0.6954
BWD	21	0.8177	0.7905		1	**<.0001**	**<.0001**	0.6954
SW	21	0.8177	0.7905			**<.0001**	**<.0001**	0.6954
BS	5	0.8099	0.7844				0.9123	**<.0001**
FSR	17	0.8082	0.7843					**<.0001**
FULL	45	0.8172	0.7907					
Partial Churn 2								
model	size	AUC	PCC	BWD	SW	BS	FSR	FULL
FWD	46	0.7066	0.9030	0.1025	0.1080	**0.0043**	0.5288	0.9643
BWD	66	0.7084	0.9037		0.5891	**0.0001**	0.0867	0.1146
SW	46	0.7081	0.9034			**0.0004**	0.1715	0.3626
BS	35	0.6980	0.9012				**0.0275**	**0.0140**
FSR	32	0.6939	0.9026					0.5832
FULL	178	0.7006	0.9029					

Table 3 (continued). Predictive performance of the different feature selection techniques

Churn 1								
model	size	AUC	PCC	BWD	SW	BS	FSR	FULL
FWD	55	0.7699	0.8620	0.8605	0.8387	**0.0010**	**<.0001**	0.4227
BWD	62	0.7703	0.8619		0.9407	**0.0002**	**<.0001**	0.4394
SW	53	0.7699	0.8620			**0.0012**	**<.0001**	0.4710
BS	51	0.7600	0.8591				**<.0001**	**0.0025**
FSR	28	0.7434	0.8451					**<.0001**
FULL	167	0.7688	0.8615					
Churn 2								
model	size	AUC	PCC	BWD	SW	BS	FSR	FULL
FWD	44	0.7061	0.9523	0.7600	0.5553	**0.0056**	0.2124	**0.0140**
BWD	72	0.7111	0.9524		0.4108	**0.0123**	0.2157	**0.0019**
SW	37	0.7048	0.9521			**0.0149**	0.4190	**0.0229**
BS	25	0.6634	0.9512				0.0731	0.9425
FSR	30	0.7006	0.9519					0.1356
FULL	424	0.6745	0.9512					
Targeting								
model	size	AUC	PCC	BWD	SW	BS	FSR	FULL
FWD	24	0.7400	0.7235	0.5252	0.2471	0.3301	**0.0313**	0.1372
BWD	37	0.7399	0.7231		0.9559	0.7666	0.1277	0.3264
SW	21	0.7399	0.7231			0.7944	0.0959	0.3965
BS	48	0.7387	0.7229				0.1758	0.5076
FSR	16	0.7375	0.7218					0.3982
FULL	100	0.7387	0.7225					
Credit Scoring 1								
model	size	AUC	PCC	BWD	SW	BS	FSR	FULL
FWD	30	0.8645	0.9846	0.9759	0.8195	0.1833	0.8615	0.8733
BWD	29	0.8727	0.9846		0.8304	0.2009	0.8777	0.8446
SW	25	0.8651	0.9846			0.2183	0.9504	0.9999
BS	41	0.8483	0.9843				0.1905	0.2411
FSR	26	0.8617	0.9846					0.9643
FULL	137	0.8534	0.9846					
Credit scoring 2								
model	size	AUC	PCC	BWD	SW	BS	FSR	FULL
FWD	44	0.8815	0.9720	0.5872	**0.0280**	**0.0053**	0.2029	**0.0267**
BWD	54	0.8814	0.9721		**0.0235**	**0.0024**	0.1395	**0.0132**
SW	20	0.8764	0.9713			0.2192	0.7441	0.9766
BS	38	0.8587	0.9708				0.1891	0.3543
FSR	17	0.8621	0.9714					0.7308
FULL	239	0.8628	0.9713					

In six out of the nine cases, no other feature selection technique was found that significantly outperforms FSR. Hence, in the majority of the applications, the restrictions posed upon the parameter signs do not significantly hamper predictive performance. In one application, labeled 'targeting', FWD significantly outperforms the FSR model, yet the average performance differences are negligible (AUC difference = .0025; PCC difference = 0.0017).

However, in the remaining two applications, the FWD, BWD and SW selection techniques all significantly outperform the proposed FSR technique. While in 'Partial Churn 1', the differences remain rather low (AUC difference = .0095; PCC difference = 0.0062), in the 'Churn 1' case, they become more substantial (AUC difference = .0269; PCC difference = 0.0169). In general, however, it is clear that the sign restrictions imposed in the FSR technique do not necessarily hamper predictive performance. Furthermore, in all applications but one, the FSR model is more parsimonious than the unrestricted FWD model. Another remarkable result lies in the observation that best subset (BS) selection, the feature selection technique that should theoretically provide the best model, does not live up to its promises. Indeed, in none of the cases, BS outperforms FWD, BWD or SW selection, while the runtime of BS can be substantially longer (in one of the cases, the runtime of BS was 8257 times longer than FWD). Even stronger, in no less than five applications, the FWD or BWD selection techniques significantly outperform the BS technique. As a plausible explanation, it must be noted that the latter technique ensures the detection of the best model on the training set, and thus, as noted by [10], does not make any allowance for the overfitting which occurs. This is consistent with [17], who claim that the use of a fixed path estimator, i.e. a model based on training performance only, such as BS, turns out to be highly biased for submodel selection. In the next section, we offer a more general overview of the conclusions of our empirical work.

5. CONCLUSIONS

In his invited paper for the PAKDD conference in 2003, [2] distinguishes two main factors influencing the likelihood of obtaining a high quality, useful model, namely (i) the predictive capabilities and robustness of the model, and (ii) the interpretability of this model. A feature selection technique that restricts parameters signs to correspond to their univariate counterparts offers an improvement on both crucial issues. Indeed, due to the use of feature selection, overfitting is reduced, leading to more robust predictive models [10], while the acceptability of the model can be improved by excluding sign violations [3]. In this study, we have empirically evaluated the usefullness of Forward selection with Sign Restrictions (FSR) as an alternative to the standard feature selection techniques on a collection of nine real-life European predictive modeling data sets. In this evaluation, we have proven that (i) a negative relationship exists between model size and the percentage of parameters that correspond to their univariate signs, (ii) all tested feature selection techniques also reduce the

percentage of sign violations, but do not necessarily ensure models where *all* signs correspond to their univariate conterparts (iii) most feature selection techniques show comparable results in terms of predictive performance, (iv) in none of the cases, the theoretically preferred best subset technique significantly outperforms other techniques, and most importantly (v) the cost of excluding sign violations in the feature selection process is non-existant or very low in all cases but one.

Following the results of our empirical analysis, we conclude that FSR has the ability of delivering a predictive model of reasonable performance that is more intuitively acceptable. However, since this conclusion does not hold in all nine settings, we note that the choice of the most suitable feature selection technique should be based on a careful comparison of the results of different techniques. Also, different settings require a different focus on either predictive performance or interpretability. Nonetheless, we are convinced that it is useful to consider the application of FSR as a feature selection technique whenever well-performing models need to be built that can be interpreted and accepted by management, employees and / or customers.

6. LIMITATIONS AND ISSUES FOR FURTHER RESEARCH

This study has a number of limitations. In this paper, we do not attempt an exhaustive comparison of all existing feature selection techniques nor of all existing classification algorithms. Instead, we have focused on a solid comparison of a limited number of wrappers for feature selection in a logistic regression framework. Additionally, shrinkage methods may deliver another alternative method for increasing predictive accuracy while reducing multicollinearity, and their comparison with the variable selection techniques used in this study can form an interesting future research topic. Also, although we disposed of a reasonable range and variety of data sets, the nine cases do not allow a useful meta-analysis to explain the differences in performance across the applications. In an attempt to perform such an analysis, we computed 105 descriptive features about each case, including the features presented in Table 1, covering incidence, number of observations and features, overfitting, number of factors in the data, multicollinearity, predictive performances of the full model and univariate models, etc. However, we found no consistency about significant influences on both performance measures (AUC and PCC). Possibly, this is due to not only

the low number of observations in the meta analysis, but also the fact that in only two cases the differences in performance were considered significant.

ACKNOWLEDGEMENTS

The authors would like to express their highest gratitude to Wouter Buckinx, Jonathan Burez, Bart Larivière and Bernd Vindevogel, who contributed the data sets they gathered during their PhDs in order to enable the experiments performed in this study.

REFERENCES

1. Grönroos C.: From Marketing Mix to Relationship Marketing—Towards a Paradigm Shift in Marketing, Management Decision 35(4) (1997) 322–339
2. Bradley P.S.: Data Mining as an Automated Service. In: K.-Y. Whang, J. Jeon, K. Shim, J. Srivatava (eds.): Advances in Knowledge Discovery and Data Mining (PAKDD'2003). Springer (2003) 1-13
3. Pazzani M.J. and Bay S.D.: The Independent Sign Bias: Gaining Insight from Multiple Linear Regression. In: Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society (1999) 525-530
4. Baesens B., Van Gestel T., Viaene S., Stepanova M., Suykens J., and Vanthienen J.: Benchmarking State of the Art Classification Algorithms for Credit Scoring. Journal of the Operational Research Society 54 (2003) 627-635
5. Davis R.H., Edelman D.B. and Gammerman A.J.: Machine-Learning Algorithms for Credit-card Applications. IMA Journal of Mathematics Applied in Business and Industry 4 (1992) 43-51
6. Thomas L.C.: A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Customers. International Journal of Forecasting 16 (2000) 149-172
7. Hand D.J.: Strength in Diversity: The advance of data analysis. In: Boulicaut J.-F., Esposito F., Giannotti F, and Pedreschi D. (eds.): Knowledge Discovery in Databases (PKDD'2004). Springer (2004) 18-26
8. Belsley D.A., Kuh E. and Welsch R.E.: Regression diagnostics: identifying influential data and sources of collinearity. John Wiley & Sons, New York (1980)
9. Tibshirani R.: Regression Shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B 58(1) (1996) 267-288
10. Miller A.: Subset selection in regression. Second Edition. Chapman & Hall/CRC, New York (2002)
11. Furnival G.M. and Wilson R.W.: Regressions by Leaps and Bounds. Technometrics 16 (1974) 499-511
12. Mullet, G.: Why Regression Coefficients Have the Wrong Sign. Journal of Quality Technology 8(3) (1976) 121-126
13. Dietterich T.G.: Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation 10(7) (1998) 1895-1923

14. Bouckaert R. and Frank E.: Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In: Dai H., Srikant R. and Zhang C. (eds.): Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'2004). Springer (2004) 3-12
15. Nadeau C. and Bengio Y.: Inference for the generalization error. Machine Learning 52 (2003) 239-281
16. Malthouse E.C.: Assessing the performance of direct marketing scoring models. Journal of Interactive Marketing 15(1) (2001) 49-62.
17. Breiman L.: The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error. Journal of the American Statistical Association 87 (419) (1992) 738-754

BIBLIOGRAPHY

Acid S. and Campos L.M. (1996) An algorithm for finding minimum *d*-separating sets in belief networks. In: Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI), Portland, Oregon, USA, pp. 3-10

Baesens B., Van Gestel T., Viaene S., Stepanova M., Suykens J. and Vanthienen J. (2003) Benchmarking State of the Art Classification Algorithms for Credit Scoring. Journal of the Operational Research Society **54** pp. 627-635

Baesens B., Verstraeten G., Van den Poel D., Egmont-Petersen M., Van Kenhove P. and Vanthienen J. (2004) Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. European Journal of Operational Research **156** (2) pp. 508-523

Baesens B., Viaene S., Van den Poel D., Vanthienen J. and Dedene G. (2002) Bayesian Neural Network Learning for Repeat Purchase Modelling in Direct Marketing. European Journal of Operational Research **138** (1) pp. 191-211

Banasik J., Crook J. and Thomas L. (2003) Sample selection bias in credit scoring models. Journal of the Operational Research Society **54** pp. 822-832

Belsley D.A. (1991) Conditioning diagnostics, collinearity and weak data in regression. John Wiley and Sons, New York

Belsley D.A., Kuh E. and Welsch R.E. (1980) Regression diagnostics: identifying influential data and sources of collinearity. John Wiley and Sons, New York

Blattberg R.C. and Deighton J. (1996) Manage marketing by the customer equity test. Harvard Business Review (July-Aug) pp. 136-144

Bolton R.N., Kannan P.K. and Bramlett M.D. (2000) Implications of Loyalty Program Membership and Service Experiences for Customer Retention and Value. Journal of the Academy of Marketing Science **28** (1) pp. 95–108

Bouckaert R. and Frank E. (2004) Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In: Dai H., Srikant R. and Zhang C. (eds.): Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'2004). Springer pp. 3-12

Bradley P.S. (2003) Data Mining as an Automated Service. In: K.-Y. Whang, J. Jeon, K. Shim, J. Srivatava (eds.): Advances in Knowledge Discovery and Data Mining (PAKDD'2003). Springer pp. 1-13

Breiman L. (1992) The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error. Journal of the American Statistical Association **87** (419) pp. 738-754

Breiman L. (1995) Better subset regression using the nonnegative garrote. Technometrics **37** 4 pp. 373-384

Breiman L. (2001) Random Forests. Machine Learning **45** pp. 5-32

Breiman L. (2001) Statistical Modeling: The Two Cultures. Statistical Science **16** (3) pp. 199-231

Brodie R.J., Coviello N.E., Brookes R.W. and Little V. (1997) Towards a paradigm shift in marketing? an examination of current marketing practices. Journal of Marketing Management **13** pp. 383-406

Buckinx W. (2005) Using Predictive Modeling for Targeted Marketing in a Non-Contractual Retail Setting. PhD thesis, Ghent University

Buckinx W. and Van den Poel D. (2005) Customer Base Analysis: Partial Defection of Behaviorally-Loyal Clients in a Non-Contractual FMCG Retail Setting. European Journal of Operational Research **164** (1) pp. 252-268

Buckinx W., Verstraeten G. and Van den Poel D. (2006) Predicting customer loyalty using the internal transactional database. Expert Systems with Applications **32** (1) forthcoming

Bult J.R. and Wansbeek T. (1995) Optimal Selection for Direct Mail. Marketing Science **14** (4) pp. 378–394

Buntine W. (1996) A guide to the literature on learning probabilistic networks from data. IEEE Transactions On Knowledge And Data Engineering **8** pp. 195-210

Cheng J. (2000) Powerpredictor system. Available from <http://www.cs.ualberta.ca/~jcheng/bnpp.htm>

Cheng J., Bell D.A., and Liu W. (1997) An algorithm for bayesian belief network construction from data. In: Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics (AI and STAT), Fort Lauderdale, Florida, USA, pp. 83-90

Cheng J., Bell D.A., and Liu W. (1997) Learning belief networks from data: an information theory based approach. In: Proceedings of the Sixth ACM Conference on Information and Knowledge Management (CIKM), Las Vegas, Nevada, USA, pp. 325-331

Cheng J. and Greiner R. (1999) Comparing bayesian network classifiers. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, pp. 101-108

Cheng J. and Greiner R. (2001) Learning bayesian belief network classifiers: Algorithms and system. In: Proceedings of the Fourteenth Canadian Conference on Artificial Intelligence (AI)

Chintagunta P.K. (1992) Estimating a multinomial probit model of brand choice using the method of simulated moments. Marketing Science **11** (4) pp. 386-407

Chow C.K. and Liu C.N. (1968) Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory **14** (3) pp. 462-467

Cohen J. and Cohen P. (1983) Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (2nd ed). Hillsdale, NJ: Erlbaum

Cohen J., Cohen P., West S.G. and Aiken L.S. (2003) Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences (3rd ed) Lawrence Erlbaum Associates: Mahwah, New Jersey

Cullinan G.J. (1977) Picking them by their batting averages' recency-frequency-monetary method of controlling circulation. Manual release 2103, Direct Mail/Marketing Association, NY

Dasgupta C.G., Dispensa G.S. and Ghose S. (1994) Comparing the predictive performance of a neural network model with some traditional market response models. International Journal of Forecasting **10** (2) pp. 235-244

Davis R.H., Edelman D.B. and Gammerman A.J. (1992) Machine-Learning Algorithms for Credit-card Applications. IMA Journal of Mathematics Applied in Business and Industry **4** pp. 43-51

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

De Wulf K., Odekerken-Schröder G. and Iacobucci D. (2001) Investments in Consumer Relationships: A Cross-Country and Cross-Industry Exploration. Journal of Marketing **65** (October) pp. 33–50

Desai V.S., Crook J.N. and Overstreet G.A. Jr. (1996) A comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research **95** pp. 24-37

Dick A.S. and Basu K. (1994) Customer Loyalty: Toward an Integrated Conceptual Framework. Journal of the Academy of Marketing Science **22** (2) pp. 99–113

Dietterich T.G. (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation **10** (7) pp. 1895-1924

Dowling G.R. and Uncles M. (1997) Do Customer Loyalty Programs Really Work? Sloan Management Review **38** (Summer) pp. 71–82

Duda R.O. and Hart P.E. (1973) Pattern Classification and Scene Analysis. John Wiley, New York.

East R., Harris P., Willson G. and Lomax W. (1995) Loyalty to Supermarkets. International Review of Retail, Distribution and Consumer Research **5** (1) pp. 99-109

Efron B., Hastie T., Johnstone I. and Tibshirani R. (2004) Least angle regression. Annals of Statistics **32** (4) pp. 407-451

Egan J.P. (1975) Signal Detection Theory and ROC analysis. Series in Cognition and Perception. Academic Press, New York

Eisenbeis R.A. (1977) Pitfalls in the application of discriminant analysis in business, finance and economics. Journal of Finance **32** pp. 875-900

Elder J.F. and Pregibon D. (1996) A statistical perspective on knowledge discovery in databases. In: Fayyad U.M., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R. (eds.)

Advances in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI Press. pp. 83-113

Everitt B.S. (1977) The analysis of contingency tables. Chapman and Hall, London.

Fawcett T. (2001) Using rule sets to maximize roc performance. In: Proceedings of the IEEE International Conference on Data Mining, San Jose, California, USA

Fayyad U.M. and Irani K.B. (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI), San Francisco, CA, USA, Morgan Kaufmann, pp. 1022-1029

Fayyad U.M., Piatetsky-Shapiro G. and Smyth P. (1996) From Data Mining to Knowledge Discovery: An Overview. In: Fayyad U.M., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R. (eds.) Advances in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI Press pp. 1-34

Feelders A.J. (2000) Credit scoring and reject inference with mixture models. International Journal of Intelligent Systems in Accounting, Finance & Management **8** pp. 271-279

Feinberg F.F., Krishna A. and Zhang J.Z. (2002) Do We Care What Others Get? A Behaviorist Approach to Targeted Promotions. Journal of Marketing Research **39** (August) pp. 277–291

Festinger L. (1954) A Theory of Social Comparison Processes. Human Relations **7** pp. 117-140

Fornell C. and Wernerfelt B. (1987) Defensive marketing strategy by customer complaint management: a theoretical analysis. Journal of Marketing Research **24** pp. 337-346

Frank I. and Friedman J. (1993) A statistical view of some chemometrics regression tools. Technometrics **35** 109-148

Friedman N., Geiger D., and Goldszmidt M. (1997) Bayesian network classifiers. Machine Learning **29** pp. 131-163

Furnival G.M. and Wilson R.W. (1974) Regressions by Leaps and Bounds. Technometrics **16** pp. 499–511

Ganesh J., Arnold M.J., and Reynolds K.E. (2000) Understanding the customer base of service providers: an examination of the differences between switchers and stayers. Journal of Marketing **64** pp. 65-87

Garbarino E. and Johnson, M.S. (1999) The different roles of satisfaction, trust, and commitment in customer relationships. Journal of Marketing **63** (April) pp. 70-87

Geiger D. and Heckerman D. (1996) Knowledge representation and inference in similarity networks and bayesian multinets. Artificial Intelligence **82** pp. 45-74

Geiger D., Verma T.S. and Pearl J. (1990) Identifying independence in bayesian networks. Networks **20** (5) pp. 507-534

Geng W., Cosman P., Berry C.C., Feng Z. and Schafer W.R. (2004) Automatic Tracking, Feature Extraction and Classification of C Elegans Phenotypes. IEEE Transactions on Biomedical Engineering **10** (51) pp. 1811-1820

Goutte C. (1997) Note on Free Lunches and Cross-Validation. Neural Computation **9** pp. 1245–1249

Grönroos C. (1997) From Marketing Mix to Relationship Marketing—Towards a Paradigm Shift in Marketing, Management Decision **35** (4) pp. 322–339

Hallberg G. (2004) Is Your Loyalty Programme Really Building Loyalty? Why Increasing Emotional Attachment, not Just Repeat Buying, is Key to Maximising Programme Success. Journal of Targeting, Measurement and Analysis for Marketing **12** (3) pp. 231–241

Hand D., Mannila H. and Smyth P. (2001) Principles of Data Mining. MIT Press, Cambridge, MA

Hand D.J. (1999) Statistics and data mining: intersecting disciplines. SIGKDD Explorations **1** pp. 16-19

Hand D.J. (2001) Modelling consumer credit risk. IMA Journal of Management Mathematics **12** pp. 139-155

Hand D.J. (2004) Strength in diversity: the advance of data analysis. In Boulicaut J.-F., Esposito F., Giannotti F, and Pedreshchi D. (eds.) Proceedings of the 15th European Conference on Machine Learning. Pisa, Italy: Springer pp. 18-26

Hand D.J. and Henley W.E. (1994) Can reject inference ever work? IMA Journal of Mathematics Applied in Business and Industry **5** pp. 45-55

Hand D.J. and Henley W.E. (1997) Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society Series A **160** pp. 523-541

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

Hanley J.A. and McNeil B.J. (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology **148** (3) pp. 839-843

Heckerman D. (1991) Probabilistic Similarity Networks. MIT Press, Cambridge, MA

Heckerman D. (1995) A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research

Heckman J. (1979) Sample selection bias as a specification error. Econometrica **47** pp. 153-161

Hosmer D.W. and Lemeshow S. (1989) Applied Logistic Regression, John Wiley and Sons, New York

Hosmer D.W., Jovanovic B., and Lemeshow S. (1989) Best subsets logistic regression. Biometrics **45** pp. 1265-1270

Hsia D.C. (1978) Credit scoring and the equal credit opportunity act. Hastings Law Journal **30** pp. 371-448

Hwang H., Jung T. and Suh E. (2004) An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications **26** (2) pp. 181-188

Irwin J.R. and McClelland G.H. (1991) Misleading Heuristics and Moderated Multiple Regression Models. Journal of Marketing Research **38** (February) pp. 100–109

Joanes D.N. (1994) Reject inference applied to logistic regression for credit scoring. IMA Journal of Mathematics Applied in Business and Industry **5** pp. 35-43

John G.H. and Langley P. (1995) Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI), Montreal, Quebec, Canada, Morgan Kaufmann, San Francisco, CA, pp. 338-345

Jones T.O. and Sasser E.W. (1995) Why Satisfied Customers Defect. Harvard Business Review (November–December) pp. 87–99

Jonker J.J., Piersma N. and Van den Poel D. (2004) Joint Optimization of Customer Segmentation and Marketing Policy to Maximize Long-Term Profitability. Expert Systems with Applications **27** (2) pp. 159-168

Jöreskog K.G. and Sörbom D. (1995) LISREL 8: User's Reference Guide. Chicago: Scientific Software International

Keiningham T.L., Perkins-Munn T. and Evans H. (2003) The Impact of Customer Satisfaction on Share-of-Wallet in a Business-to-Business Environment. Journal of Service Research **6** (1) pp. 37-50

Kivetz R. and Simonson I. (2003) The Idiosyncratic Fit Heuristic: Effort Advantage as a Determinant of Consumer Response to Loyalty Programs. Journal of Marketing Research **40** (4) pp. 454-467

Knox S. (1998) Loyalty-based segmentation and the customer development process. European Management Journal **16** (6) pp. 729-737

Kotler P. (1991) Philip Kotler Explores the New Marketing Paradigm. Review, Marketing Science Institute Newsletter. Cambridge, MA (Spring) pp. 1, 4-5

Kruskal J.B. Jr. (1956) On the shortest spanning subtree of a graph and the travelling salesman problem. In: Proceedings of the American Mathematics Society **7** pp. 48-50

Langley P., Iba W. and Thompson K. (1992) An analysis of bayesian classifiers. In: Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI), San Jose, CA, USA, AAAI Press, pp. 223-228

Larivière B. and Van den Poel D. (2004) 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) pp. 277-285

Latham, G.P. and Locke E.A. (1991) Self Regulation Through Goal Setting. Organizational Behavior and Human Decision Processes **50** (2) pp. 212–247

Lauritzen S.L. (1996) Graphical models. Clarendon Press, Oxford

Levene, H. (1960) Robust Tests for the Equality of Variance. In: Olkin I. (ed): Contributions to Probability and Statistics. Stanford University Press, Palo Alto, CA pp. 278-292

Macintosh G. and Lockshin L.S. (1997) Retail Relationships and Store Loyalty: A Multi-Level Perspective. International Journal of Research in Marketing **14** (5) pp. 487-497

MacKay D.J. (1992) Bayesian Interpolation. Neural Computation **4** pp. 415-447

Maddala G.S. (1992) Introduction to Econometrics Maxwell MacMillan Int. Editions: New York

Mägi A.W. (2003) Share of Wallet in Retailing: the Effects of Customer Satisfaction, Loyalty Cards and Shopper Characteristics. Journal of Retailing **79** pp. 97-106

Malthouse E.C. (1999) Ridge regression and direct marketing scoring models. Journal of Interactive Marketing **13** (4) pp. 10-23

Malthouse E.C (2001) Assessing the performance of direct marketing scoring models. Journal of Interactive Marketing **15** (1) pp. 49-62

Malthouse E.C. and Blattberg R.C. (2005) Can we predict customer lifetime value? Journal of Interactive Marketing **19** (1) pp. 2-62

McMullan R. and Gilmore A. (2002) The Conceptual Development of Customer Loyalty Measurement: A Proposed Scale. Journal of Targeting, Measurement and Analysis for Marketing **11** (3) pp. 230–243

Miller A. (2002) Subset selection in regression. Second Edition. Chapman and Hall/CRC, New York

Montgomery A.L., Li S., Srinivasan K. and Liechty J.C. (2004) Modeling Online Browsing and Path Analysis Using Clickstream Data. Marketing Science **23** (4) pp. 579-595

Morrison D.G. (1969) On the interpretation of discriminant analysis. Journal of Marketing Research **6** pp. 156-163

Mullet G. (1976) Why Regression Coefficients Have the Wrong Sign. Journal of Quality Technology **8** (3) pp. 121-126

Murphy K. (2001) Bayes net matlab toolbox. Available from <http://www.cs.berkeley. edu/~murphyk/Bayes/bnt.html>

Myers J.H. and Forgy E.W. (1963) The development of numerical credit evaluation systems. Journal of the American Statistical Association **58** pp. 799-806

Nabney I.T. (2001) Netlab Algorithm for Pattern Recognition, Springer

Nadeau C. and Bengio Y. (2003) Inference for the generalization error. Machine Learning **52** pp. 239-281

Nicholls J.G. (1989) The Competitive Ethos and Democratic Education. Cambridge, MA: Harvard University Press

Nunnally J.C. (1978) Psychometric Theory. New York: McGraw-Hill

Park Y.-H. and Fader P.S. (2004) Modeling Browsing Behavior at Multiple Websites. Marketing Science **23** (3) pp. 280-303

Pazzani M.J. and Bay S.D. (1999) The Independent Sign Bias: Gaining Insight from Multiple Linear Regression. In: Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society pp. 525-530

Pearl J. (1988) Probabilistic reasoning in Intelligent Systems: networks for plausible inference. Morgan Kaufmann, San Fransico, CA

Peppers D. and Rogers M. (1997) Enterprise one to one: tools for competing in the interactive age. Doubleday, New York, USA

Provost F., Fawcett T., and Kohavi R. (1998) The case against accuracy estimation for comparing classifiers. In: J. Shavlik (Ed.), Proceedings of the Fifteenth International Conference on Machine Learning (ICML), San Francisco, CA, USA, Morgan Kaufmann, San Fransico, CA, pp. 445-453

Quinlan J.R. (1993) C4.5 programs for machine learning. Morgan Kaufmann, San Fransico, CA

Rawlings J.O. (1988) Applied regression analysis. Brooks/Cole Publishing Company, Pacific Grove, CA

Reichheld F.F. (1996) The Loyalty Effect. Harvard Business School Press, Cambridge, MA

Reichheld F.F. (2001) Lead for loyalty. Harvard Business Review (July) pp. 76-84

Reichheld F.F. (2003) The One Number You Need to Grow. Harvard Business Review (Dec) pp. 46–54

Reichheld F.F. and Kenny D.W. (1990) The hidden advantages of customer retention. Journal of Retail Banking **4** pp. 19-23

Reichheld F.F. and Sasser W.E. (1990) Zero defections: quality comes to services. Harvard Business Review (Sept-Okt) pp. 105-111

Reinartz W.J. and Kumar V. (2000) On the profitability of long-life customers in a noncontractual setting: an empirical investigation and implications for marketing. Journal of Marketing **64** pp. 17-35

Reinartz W.J. and Kumar V. (2002) The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. Journal of Marketing **67** (January) pp. 77–99

Reinartz W.J. and Kumar V. (2002) The mismanagement of customer loyalty. Harvard Business Review (July) pp. 4-12

Reynolds K.E. and Arnold M.J. (2000) Customer Loyalty to the Salesperson and the Store: Examining Relationship Customers in an Upscale Retail Context. Journal of Personal Selling and Sales Management **20** (2) pp. 89-98

Rigby D.K., Reichheld F.F. and Schefter P. (2002) Avoid the four perils of CRM. Harvard Business Review **80** (2) pp. 101-109

Rosenberg E. and Gleit A. (1994) Quantitative methods in credit management: a survey. Operations Research **42** pp. 589-613

Rosenberg L.J. and Czepiel J.A. (1984) A marketing approach to customer retention. Journal of Consumer Marketing **1** pp. 45-51

Rust R.T., Zeithaml V.A. and Lemon K.N. (2000) Driving customer equity: how customer lifetime value is reshaping corporate strategy. Free Press, New York

Schmittlein D. and Peterson R.A. (1994) Customer Base Analysis: An Industrial Purchase Process Application. Marketing Science **13** (1) pp. 41–67

Sharp B. and Sharp A. (1997) Loyalty Programs and Their Impact on Repeat-Purchase Loyalty Patterns. International Journal of Research in Marketing **14** (5) pp. 473-486

Snee R.D. (1977) Validation of regression models: Methods and examples. Technometrics **19** (4) pp. 415-428

Srinivasan S.S., Anderson R. and Ponnavolu K. (2002) Customer Loyalty in E-commerce: An Exploration of its Antecedents and Consequences. Journal of Retailing **78** pp. 41-50

Steenkamp J.B.E.M. and van Trijp H.C.M. (1991) The Use of LISREL in Validating Marketing Constructs. International Journal of Research in Marketing **8** pp. 283–299

Stepanova M. and Thomas L. (2002) Survival analysis methods for personal loan data. Operations Research **50** pp. 277-289

Swait J. and Andrews R.L. (2003) Enriching Scanner Panel Models with Choice Experiments. Marketing Science **22** (4) pp. 442-460

Swets J.A. and Pickett R.M. (1982) Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic Press, New York

Thomas L.C. (2000) A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Customers. International Journal of Forecasting **16** pp. 149-172

Thomas L.C., Ho J. and Scherer W.T. (2001) Time will tell: behavioural scoring and the dynamics of consumer credit assessment. IMA Journal of Management Mathematics **12** pp. 89-103

Thomas L.C., Oliver R.W. and Hand D.J. (2005) A survey of the issues in consumer credit modelling research. Journal of the Operational Research Society **56** pp. 1006-1015

Thrasher R.P. (1991) CART: a recent advance in tree-structured list segmentation methodology. Journal of Direct Marketing **5** (1) pp. 35-47

Tibshirani R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B **58 (**1) pp. 267-288

Van den Poel D. (1999) Response Modeling for Database Marketing using Binary Classification. Ph.D. thesis, K.U. Leuven

Van den Poel D. and Buckinx W. (2005) Predicting Online Purchasing Behaviour. European Journal of Operational Research **166** (2) pp. 557-575

Van den Poel D. and Larivière B. (2004) Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research **157** (1) pp. 196-217

Van Der Gaag L.C. (1996) Bayesian belief networks: Odds and ends. The Computer Journal **39** (2) pp. 97-113

Verhoef P.C., Spring P.N., Hoekstra J.C. and Leeflang P. (2002) The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. Decision Support Systems **34** pp. 471-481

Verhoef P.C. (2003) Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development. Journal of Marketing **67** (October) pp. 30-45

Verhoef P.C., Franses P.H. and Hoekstra J.C. (2002) The Effect of Relational Constructs on Customer Referrals and Number of Services Purchased from a Multiservice Provider: Does Age of Relationship Matter. Journal of the Academy of Marketing Science **30** (3) pp. 202–216

Verma T. and Pearl J. (1988) Causal networks: semantics and expressiveness. In: Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence, Mountain View, CA, USA, pp. 352-359

Verstraeten G. and Van den Poel D. (2005) The impact of sample bias on consumer credit scoring performance and profitability. Journal of the Operational Research Society **56** pp. 981-992

Waller W.G. and Jain A.K. (1978) On the monotonicity of the performance of a bayesian classifier. IEEE Transactions on Information Theory **24** (3) pp. 392-394

Wangenheim F. and Bayón T. (2004) The Contribution of Word-of-Mouth Referrals to Economic Outcomes of Service Quality and Customer Satisfaction. Journal of the Academy of Marketing Science, forthcoming.

Whyte R. (2004) Frequent Flyer Programmes: Is it a Relationship, or do the Schemes Create Spurious Loyalty? Journal of Targeting, Measurement and Analysis for Marketing **12** (3) pp. 269–280

Witten I.A. and Frank E. (2005) Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, CA

Zeithaml V.A., Berry L.L. and Parasuraman A. (1996) The Behavioral Consequences of Service Quality. Journal of Marketing **60** (April) pp. 31–46