KU LEUVEN
FACULTEIT ECONOMIE EN
BEDRIJFSWETENSCHAPPEN

Business-Oriented Data Analytics: Theory and Case Studies

Proefschrift voorgedragen tot
het behalen van de graad van
Doctor in de Toegepaste
Economische Wetenschappen

door

Thomas VERBRAKEN

Nummer 427 2013

Committee

Supervisor	Prof. dr. Bart Baesens	KU Leuven
	Prof. dr. Marnik Dekimpe	Tilburg University
	Prof. dr. Theodoros Evgeniou	INSEAD
	Prof. dr. ir. David Martens	Universiteit Antwerpen
	Prof. dr. Martina Vandebroek	KU Leuven
	Dr. Bram Vanschoenwinkel	AE

To Évi and my parents,
For their loving support.

Acknowledgments

First and foremost, I want to thank my promoter, prof. dr. Bart Baesens, for his excellent advice and the opportunities he offered me throughout my PhD. He guided me towards promising research topics, while giving me the freedom to pursue those ideas which looked interesting to me. Furthermore, I know few people who are able to enthuse their employees the way he does, and I very much enjoyed working with him.

Likewise, I am thankful to prof. dr. David Martens for his intelligent and involved advice, for his encouragement, and for the interest he always showed. I also appreciate the amusing and pleasant conversations we had, during a coffee break at work or with a beer after work.

I am grateful to prof. dr. Marnik Dekimpe, prof. dr. Theodoros Evgeniou, prof. dr. Martina Vandebroek, and dr. Bram Vanschoenwinkel, for being part of my PhD committee and for their valuable insights and suggestions during my doctoral seminars. I also would like to thank prof. Ben Immers and prof. dr. Chris Tampère for their role in my early academic career at the Traffic and Infrastructure Centre at KU Leuven. Prof. dr. Frank Goethals, prof. dr. Stefan Lessmann, prof. dr. Richard Weber, prof. dr. Sebastián Maldonado, and dr. Cristián Bravo deserve my gratitude for the fruitful research collaborations which have been an important part of my PhD research.

Although being a PhD student is a rather individual job, the importance of colleagues cannot be underestimated. I would like to thank all of them for the interesting conversations, the occasional beer, and their support, both moral and technical. I especially want

to thank prof. dr. Wouter Verbeke, a former colleague but most of all a very good friend, for his enthusiastic help – without him, my academic career would have looked completely different.

Without question, I very much appreciate the unconditional and everlasting support of my parents, brothers, sister and my soon to be parents-in-law and brother-in-law. Thank you for always being there, in good and bad times. And finally, last but not least, I want to thank the woman who makes me smile every day again, my fiancée I am lucky to marry in nine days, Évi.

Leuven, September 12, 2013.

Contents

Preface

The amount of data stored by human beings has experienced an explosive growth during the past decades. According to Hilbert and López (2011), the worldwide data storage capacity amounted to $2.9 \cdot 10^{20}$ bytes in 2007, and global storage capacity grew at an annual rate of 23% between 1986 and 2007. These astonishing figures illustrate the increasing importance of data in our modern society. However, mere data would not be valuable without the ability to extract information and knowledge from these vast databases. The process of turning raw data into valuable information is known under various names such as knowledge discovery from data (KDD), data mining, or data analytics, and operates on the intersection of disciplines such as artificial intelligence, machine learning, statistics, and database systems.

Several data mining techniques have been developed in order to analyze data in a growing number of application fields. According to Fayyad et al. (1996), there are six classes of data mining tasks: anomaly detection, association rule learning, clustering, summarization, regression, and classification. This PhD dissertation focuses on binary classification, which entails the prediction of a binary target variable, based on a number of dependent variables.

A myriad of classification techniques has been proposed in the academic literature, each with their own merits. According to the *No Free Lunch* theorem from Wolpert and Macready, there exist no single best classifier; the choice of technique rather depends on the characteristics of the data set being analyzed (Ali and Smith, 2006). Essentially, when tackling a specific problem, one always needs to select the optimal classifier for that particular context. Hence, even

though classification techniques have been extensively compared to one another, quality assessment of classification techniques remains important.

A full quality assessment of classification techniques consists of three dimensions. The main requirements for acceptance of a data mining model, as identified by Martens (2008), are (1) the predictive power, (2) its comprehensibility, and (3) its justifiability. The major part of this thesis focuses on the predictive power of classification models, thereby keeping a strong focus on the business context in which the classification model is used. This leads to a new framework for measuring the performance of classification models, the expected maximum profit (EMP) measure. Furthermore, two independent case studies are presented, which demonstrate the usefulness of data mining in a business context. In what follows, the outline of this PhD dissertation and the main contributions are discussed in detail.

Outline and contributions

This dissertation is divided into two main parts: Part I proposes a profit-oriented approach towards classification, whereas Part II presents two case studies in data mining.

Part I – Profit-driven data analytics

Chapter 2 – The EMP framework

This chapter proposes a theoretical profit-based framework for classification performance measurement, the EMP framework.

- A careful and general analysis of the employment of a classification algorithm in a business context is provided, and the involved costs and benefits are outlined.
- The EMP framework is proposed, which is an abstract and general framework for classification performance measurement and can be implemented for different business contexts.

- Attention is being paid to the link with other performance metrics, especially the connection with the receiver operating characteristic (ROC) curve is analyzed in detail.

Chapter 3 – EMP for customer churn prediction

In this chapter, the EMP framework is implemented for application in a customer churn prediction context.

- The classification costs and benefits in the case of customer churn prediction are identified, and the EMP measure for customer churn prediction, EMP^{ccp}, is proposed.
- An estimation procedure – based on empirical ROC curves – is provided. This enables the application of the EMP^{ccp} measure in real life settings.
- The link between the H measure and the EMP^{ccp} measure is investigated, and it is shown, analytically as well as empirically, that the H measure – with appropriately chosen parameters – is an approximation to the EMP^{ccp} measure.
- An extensive case study shows that the EMP^{ccp} measure leads to other rankings than traditional performance measures, and that EMP^{ccp}-based model selection leads to a higher profit than AUC-based model selection.
- It is illustrated that the EMP^{ccp} measure provides information about the fraction of the customer base which should be targeted in a retention campaign, a feature which is not shared with traditional performance measures.
- Finally, a sensitivity analysis reveals that the EMP^{ccp} is robust to variations in the cost and benefit parameters.

Chapter 2 (the general EMP framework) and Chapter 3 (its implementation for customer churn prediction) have been published in:

Verbraken, T., Verbeke, W., Baesens, B., 2013d. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE Transactions on Knowledge and Data Engineering 25 (5), 961–973.

Chapter 4 – EMP for credit scoring

An implementation of the EMP measure for consumer credit scoring is discussed in Chapter 4.

- The costs and benefits in the case of consumer credit scoring are analyzed and EMP^{cs}, an EMP measure tailored to consumer credit scoring, is proposed.
- A procedure for the empirical estimation of EMP^{cs} – based on ROC curves – is established.
- A real life case study illustrates that EMP^{cs}-based parameter tuning leads to other parameter settings than AUC or accuracy-based parameter tuning.
- The EMP^{ccp} measure provides information about the optimal cutoff value, which is needed to operationalize the classification model. It is shown that EMP^{ccp}-based cutoff determination leads to higher overall profitability.

This chapter has been submitted for publication and is currently under review:

Verbraken, T., Bravo, C., Weber, R., Baesens, B., 2013a. Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research (under review).

Chapter 5 – Toward profit-driven model building

Chapter 5 presents an exploratory study on the usefulness of profit-driven model building.

- The EMP framework is combined with ensemble selection (ES), in order to analyze the impact of profit-driven model building.
- A real life case study in a customer churn prediction setting is carried out, in order to assess four research hypotheses on profit-driven model selection and model building.
- The empirical results show that using a profit-based measure in the model building step leads to higher profitability, as compared to traditional model building.

This chapter has been published in:

Verbraken, T., Lessmann, S., Baesens, B., 2012b. Toward profit-driven churn modeling with predictive marketing analytics. In: Cloud computing and analytics: innovations in e-business services. The Eleventh Workshop on E-Business (WEB2012). Orlando (US), 15 December 2012 (accepted).

Part II – Case studies in data mining

Part II presents two standalone case studies on the application of data mining in a business context.

Chapter 7 – Customer churn prediction with Bayesian network classifiers

This study investigates the use of a specific type of classification techniques, i.e. Bayesian network classifiers, for customer churn prediction.

- An overview of the several Bayesian network classification techniques is given, and a related algorithm for feature selection, Markov blanket feature selection, is discussed.
- An extensive benchmarking study of Bayesian network classifiers for customer churn prediction has been carried out. The results indicate that most Bayesian network classifiers are not

significantly better, neither worse, than traditional logistic regression.

- Furthermore, the case study illustrates that Markov blanket feature selection does not negatively impact the classification performance, while reducing the number of variables, and thus increasing interpretability.

- Finally, an example of a compact Bayesian network is given, which performs well in terms of predictive accuracy and is comprehensible, thereby fulfilling the requirements for the acceptance of a data mining model.

This chapter has been accepted for publication:

Verbraken, T., Verbeke, W., Baesens, B., 2013e. Profit optimizing customer churn prediction with Bayesian network classifiers. Intelligent Data Analysis (accepted for publication).

Chapter 8 – Predicting online channel acceptance with social network data

The second case study analyzes the use of social network-based classification techniques to predict e-commerce acceptance.

- Information was gathered on the e-commerce acceptance of people and their social network, through a survey. This information is then used to investigate the use of social network classification techniques for the prediction of e-commerce acceptance.

- The study reveals that knowledge of a person's social network is valuable to predict the person's acceptance of the online channel for buying different products.

- Hereby, no information is needed about the intensity of the relation. This means that a binary connectivity matrix is sufficient to make meaningful predictions.

- Finally, the results indicate that socio-demographic data is not necessarily needed. Hence, social network data only is enough in certain situations, which provides opportunities for companies to identify potential customers for their online shop.

This chapter has been accepted for publication:

Verbraken, T., Goethals, F., Verbeke, W., Baesens, B., 2013b. Predicting online channel acceptance with social network data. Decision Support Systems (accepted for publication).

Part I

Profit-driven data analytics

Chapter 1

Introduction

As a result of the steep growth in computational power and the ever growing amount of data available to companies, predictive analytics has become a popular approach toward managerial decision support. With the increased interest for predictive analytics in the industry, the question arises whether the methods and metrics employed within predictive analytics are adequate for a direct application in a business context. The main thread of Part I deals with this question and advocates the incorporation of profitability into predictive analytics, in order to support companies with one of their important goals, profit maximization.

Section 1.1 first explains the role of predictive analytics in the industry, and how companies can benefit from the adoption of predictive analytics. Section 1.2 briefly introduces the sub domain of predictive analytics we will focus on, i.e. binary classification.

1.1 Predictive analytics in the industry

Data-driven models for decision support are increasingly being employed in the industry. A recent survey among 212 senior executives of Fortune 1000 firms, conducted by Germann et al. (2013), indicates that a one-unit increase (on a scale of one to seven) in the degree of deployment of marketing analytics is, on average, associated with an 8% increase in return on assets. This positive impact is more pronounced for firms which are facing a greater level of competition

within their industry, and for firms dealing with rapidly changing customer preferences (Germann et al., 2013). The beneficial effects of predictive analytics on firms' profitability are further supported by other studies (Hoch and Schkade, 1996; Kannan et al., 2009; Lodish et al., 1988; McIntyre, 1982; Natter et al., 2008; Silva-Risso et al., 1999; Zoltners and Sinha, 2005). However, the survey by Germann et al. (2013) also revealed that few managers are convinced of the benefits of (marketing) analytics. Furthermore, a study of 587 C-level executives of large international companies, carried out by McKinsey & Co (2009), showed that marketing analytics is only regularly used by 10% of the firms. This low prevalence suggests that many decision makers are not convinced about the positive impact.

Lilien et al. confirmed that decision makers relying on a high quality decision support system (DSS) make objectively better decisions than those who use less advanced tools (such as Excel). Nevertheless, despite the positive impact of a DSS on the firms' performance, the users of a DSS do not perceive its usefulness as such. Hence, we should ask the question how we can improve the adoption of predictive analytics in the industry, since there appears to be a mismatch between the perception of its impact, and the effective impact itself. The lack of recognition of the quality of the DSS can be explained by the fit-appropriation model (FAM), proposed by Dennis et al. (2001). FAM states that the impact of a DSS is influenced by the fit between the task and the DSS, i.e. the task-technology fit (appropriation support, i.e. training to incorporate the new DSS into the decision process, is another important factor). A recommendation for an increased acceptance of DSS among decision makers, is to design tools which fit well with the decision context, and which provide specific feedback on the likely (market) outcome.

Generally, within a business context, profitability is one of the main end goals of the company and its managers. Therefore, we argue that profitability should be incorporated into predictive analytics, in the first place to yield better results – in terms of profitability – but also in order to promote the use of these tools among managers. After

all, profitability is the language all managers speak and understand.

1.2 Profit-driven classification

Unlike explanatory modeling, which aims at gaining insight into structural dependencies between variables of interest, the objective of predictive analytics is to construct data-driven models that produce operationally accurate forecasts. Such a predictive analytics tool consists of two components: (1) data-driven models designed to predict future observations and (2) methods to assess the predictive power of such models (Shmueli et al. 2011). This dissertation focuses on a sub domain of predictive analytics: binary classification. Hence, two components are of interest: the classification models themselves, and the classification performance measures. We argue that a focus on profitability should be integrated into both components.

There has already been some attention for cost-sensitive learning. E.g. Domingos (1999) proposed a general method to produce cost-sensitive classifiers, Provost and Fawcett (2001) combined ROC curve analysis with cost distribution information, Bernstein et al. (2005) developed an ontology-based approach for cost-sensitive classification, Zhou and Liu (2006) used over- and undersampling and threshold moving (and an ensemble of these methods) for cost-sensitive learning with neural networks, and more recently, Hand (2009) introduced the H measure, which takes into account misclassification costs. However, as pointed out by Elkan (2001):

> "Although most recent research in machine learning has used the terminology of costs, doing accounting in terms of benefits is generally preferable, because avoiding mistakes is easier, since there is a natural baseline from which to measure all benefits, whether positive or negative."

By focusing on benefits (and profit) rather than costs, we are staying closer to the business reality, and aid the adoption of classification techniques in the industry. Therefore, a profit-based classification

performance measure will be developed and applied to real life business cases. Moreover, an exploratory study on the incorporation of the profitability criterion into the model building step is presented.

The remainder of Part I is organized as follows. First, a profit-oriented framework for classification performance measurement, the Expected Maximum Profit (EMP) measure, will be introduced in Chapter 2. This is a general and abstract framework, which needs to be tailored to specific business environments. Chapter 3 elaborates on the implementation of the EMP framework for a customer churn prediction context, and the benefits hereof are illustrated with a case study. Similarly, the EMP framework is implemented for consumer credit scoring in Chapter 4. Finally, the possibilities of profit-based model building are explored in Chapter 5.

Chapter 2

The EMP framework

Throughout the history of predictive modeling, classification has been an important task, with applications in several industries. For the assessment of classification models, a wide variety of statistically motivated performance measures has been proposed. In this chapter, we will outline the general context of classification and the use of a profit-based performance measure for classification performance measurement.

Section 2.1 first introduces the concept of classification and its related notation. The use of classification models in a business context, and the costs and benefits associated with it, are discussed in Section 2.2. Finally, Section 2.3 outlines a general framework for profit-based classification performance measurement, which is one of the key contributions of this dissertation and will be applied in real life case studies in the following chapters.

2.1 Introduction to classification

In this dissertation, the focus lies on binary classification problems, where each instance has to be assigned a class label, being either 0 or 1. Note that we follow the convention that a 0 corresponds to a case (also called event) whereas a 1 corresponds to a non-case. This convention, opposite to what many textbooks follow, has also been adopted by Hand (2009) because it simplifies notation. The prior probabilities of class 0 and 1 are π_0 and π_1, respectively. Generally,

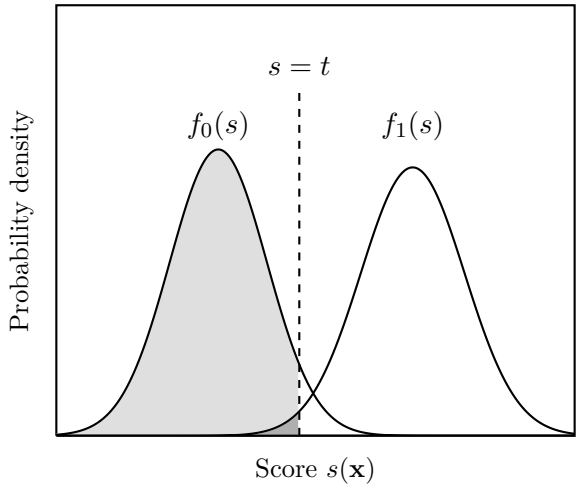


Figure 2.1: Example of score distributions and the classification process.

a classification model provides a continuous score, $s(\mathbf{x})$, which is a function of the attribute vector $\mathbf{x}$ of the respective instance, and discriminates between the two classes. In this dissertation, it is assumed that the instances from class 0 have a lower score than those from class 1 – if not, simply multiply the score by -1. The actual classification, i.e. the assignment of each instance to one of two groups, is achieved by defining a cutoff value t, such that all instances with s less than t are classified as 0, whereas instances for which s is greater than t are classified as 1.

Assume that the function $F_0(s)$ is the cumulative density function of the scores s of the cases, whereas $F_1(s)$ represents the same for the non-cases. Analogously, $f_0(s)$ and $f_1(s)$ are the probability density functions for the cases and the non-cases, respectively. Note that f_k and F_k are connected to a specific classifier, which is producing the scores $s(\mathbf{x})$. The classification process is illustrated by Figure 2.1. Every instance with score $s < t$ is classified as a case (class 0). E.g., *cases* for which $s < t$ (i.e. the shaded area under $f_0(s)$) are correctly predicted. On the other hand, *non-cases* for which $s < t$ (i.e. the shaded area under $f_1(s)$) are incorrectly predicted. Hence, the

	Predicted Label	
True Label	Case	Non-Case
Case	$\pi_0 F_0(t)$ $[c(0\|0)]$	$\pi_0(1 - F_0(t))$ $[c(1\|0)]$
Non-Case	$\pi_1 F_1(t)$ $[c(0\|1)]$	$\pi_1(1 - F_1(t))$ $[c(1\|1)]$

Table 2.1: Confusion matrix and related costs and benefits for a binary classification model.

less overlap there is between the two score distributions, the better predictions the classifier will yield.

The outcome of a classification task can also be summarized in a confusion matrix, as displayed in Table 2.1. The confusion matrix consists of four quadrants, whereby the diagonal contains the correct predictions whereas the off-diagonal elements concern incorrect predictions. Each cell displays the number of instances, expressed as a fraction of the total number of instances, N, and the four cells add up to one. For instance, the upper right cell contains the cases which are incorrectly classified as non-cases. The fraction of instances in this category is given by $\pi_0(1 - F_0(t))$. It is straightforward to see that the fractions of correct and incorrect predictions change when varying the cutoff value t – changing the cutoff t corresponds to translating the dashed line in Figure 2.1 to the left or right.

Furthermore, each cell in the confusion matrix has related costs or benefits. In general, the cost or benefit $c(k|l)$ of classifying an instance from class l into class k (with $k, l \in \{0, 1\}$) can be – and usually is – different for each of the four cells of the matrix. The distribution of these costs and benefits should be taken into account, and has a significant impact on how classification models should be used within a business context, as will be discussed in the next section.

2.2 Business context – costs and benefits

Usually, a classification model serves as input for a business decision which has to be taken, and which requires a certain population to be divided into two groups: cases and non-cases. The main reason for dividing the population into two groups, is to take a certain action towards a subgroup of the entire population, because the payoff of the action is different for the two subgroups. As an illustration, consider credit scoring, where a financial institution wants to reject loan applicants which are going to default. The population is formed by the loan applicants. The cases are the defaulters, whereas the non-cases are *healthy* loan applicants. The action which the financial institution wants to take towards the cases, is to reject their loan application. However, the payoff of rejecting a loan application of a case or a non-case is entirely different. Rejecting a defaulter results in a benefit, since no future losses will be incurred. Rejecting a non-case, however, leads to a cost, as a potential customer and loan is incorrectly rejected.

In order to assess the impact of employing a classification model, these costs and benefits are considered as *incremental* compared to the reference scenario where no classification model is applied. Hence, only the costs and benefits corresponding to *predicted* cases are relevant, since only the predicted cases will experience an impact from the action undertaken. The benefit of a correctly predicted case is b_0, whereas the cost of an incorrectly predicted case is c_1. Also note that the action, undertaken by the company towards an individual case, may come at a cost, which we denote by c^* – it is assumed that $c^* < b_0$, since it does not make sense to take an action at a certain cost, when the expected benefit is smaller. Then, we can define:

$$\begin{aligned} c(0|0) &= b_0 - c^*, \\ c(0|1) &= c_1 + c^*, \\ c(1|0) &= c(1|1) = 0, \end{aligned} \tag{2.1}$$

whereby b_0, c_1, and c^* are assumed positive. Finally, we should

Variable	Range	Description
N	$\mathbb{N}$	Number of instances
i	$\mathbb{N}$	Index for the instances
k	$\{0, 1\}$	Index for the classes
$\mathbf{x}_i$		Attribute vector of instance i
π_0	$[0; 1]$	Fraction of cases in the data set
π_1	$[0; 1]$	Fraction of non-cases in the data set
$s(\mathbf{x})$	$\mathbb{R}$	Score output from a classifier – in general dependent on $\mathbf{x}$
t	$\mathbb{R}$	Cutoff value defined to perform the actual classification
$f_k(s)$	$[0; 1]$	Probability density function of the scores for instances of class k
$F_k(s)$	$[0; 1]$	Cumulative density function of the scores for instances of class k
b_0	$\mathbb{R}^+$	The incremental benefit of correctly identifying a case
c_1	$\mathbb{R}^+$	The incremental cost of incorrectly classifying a non-case as a case
c^*	$\mathbb{R}^+$	The cost of the action undertaken towards a predicted case.
$c(k\|l)$	$\mathbb{R}^+$	Total cost or benefit of classifying an instance from class l into class k

Table 2.2: Overview of the notation used throughout Part I.

mention the fixed cost of building classification models, such as the cost of data collection, data preprocessing, model building and model maintenance. However, these costs are irrelevant for model selection, as they will be approximately the same for all models.

The above definition of classification costs and benefits allows that different models are compared in terms of the incremental profit they generate. Therefore, the average classification profit per customer, generated by employing a classifier, is defined.

Definition 2.1. ***The average classification profit*** *of a classifier, $P(t)$, is the profit generated by the employment of this classifier, and*

is expressed as:

$$\begin{aligned} P(t) = \; & c(0|0) \cdot \pi_0 F_0(t) + c(1|1) \cdot \pi_1 \left(1 - F_1(t)\right) \\ & - c(1|0) \cdot \pi_0 \left(1 - F_0(t)\right) - c(0|1) \cdot \pi_1 F_1(t). \end{aligned} \tag{2.2}$$

This expression can be simplified by taking into account Equation 2.1:

$$P(t; b_0, c_1, c^*) = (b_0 - c^*) \cdot \pi_0 F_0(t) - (c_1 + c^*) \cdot \pi_1 F_1(t). \tag{2.3}$$

Equation 2.3 shows that the average classification profit depends on the following elements:

1. The class distribution, defined by π_0 and π_1,
2. The classification costs and benefits, b_0, c_1, and c^*,
3. The classification output, through $F_0(s)$ and $F_1(s)$,
4. The cutoff t.

Element 1 is determined by the data set which is analyzed, element 2 is related to the business context of the classification problem, and element 3 is connected to the classifier under consideration. These are all fixed for a given data set and classifier. However, element 4, the cutoff t, can be set by the company, and can thus be optimized, which will be discussed in Section 2.4.

2.3 Classification performance measurement

Having defined the context of classification, we now focus on the classification performance measurement itself. In the literature, several performance measures have been proposed. We will briefly discuss commonly used measures and give some more attention to the concept of ROC curves, AUC, and the H measure.

2.3.1 Traditional performance metrics

An overview of performance measures for classification techniques can be found in Baldi et al. (2000). Probably the most well known

performance metric is the accuracy:

$$\text{Accuracy} = \pi_0 F_0(t) + \pi_1(1 - F_1(t)). \tag{2.4}$$

Accuracy measures the fraction of observations correctly classified. However, accuracy may lead to incorrect conclusions with skew data sets or with unbalanced costs (Chawla, 2005). Sensitivity and specificity, however, focus specifically on how effectively the classifier identifies cases and non-cases respectively (Sokolova and Lapalme, 2009):

$$\text{Sensitivity} = F_0(t), \tag{2.5}$$

$$\text{Specificity} = F_1(t). \tag{2.6}$$

A high sensitivity (also known as recall) means that the classifier is able to identify a high fraction of the cases. Precision, on the other hand, looks at how many of the *predicted* cases are correct (Sokolova and Lapalme, 2009):

$$\text{Precision} = \frac{\pi_0 F_0(t)}{\pi_0 F_0(t) + \pi_1 F_1(t)}. \tag{2.7}$$

In other words, precision measures how many true cases there are among the *predicted* cases, whereas recall measures how many of the *total* cases could be identified by the classifier. Precision and recall are combined in the F measure (Sokolova and Lapalme, 2009):

$$\begin{aligned} F_\beta &= \frac{(1+\beta)^2 \cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}, \\ &= \frac{(1+\beta)^2 \pi_0 F_0(t)}{\beta^2 \pi_0 + \pi_0 F_0(t) + \pi_1 F_1(t)}. \end{aligned} \tag{2.8}$$

The F measure is designed in order to measure performance when recall is β times more important than precision (Rijsbergen, 1979). Commonly used F measures are F_2 and $F_{1/2}$.

The measures discussed above have two major limitations. Firstly, they do not take into account the classification costs and benefits

which are inherent to classification in a business context. Secondly, they are usually evaluated at a fixed operational point, i.e. for a fixed cutoff t. Hence, it is not straightforward to identify the optimal cutoff value in a certain situation. In what follows, we will discuss the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC), which deal with the problem of performance evaluation at a fixed cutoff value.

2.3.2 ROC curves and AUC

The receiver operating characteristic (ROC) curve is a concept which has been extensively used in the machine learning community (Fawcett, 2006). A ROC curve is a graphical representation of the classification performance with varying threshold t. It is a plot of the sensitivity versus one minus the specificity, i.e. $F_0(t)$ as a function of $F_1(t)$. An example of a ROC curve is shown in Figure 2.2. The interested reader is referred to Fawcett (2006) or Prati et al. (2011) for an extensive discussion on ROC curves.

Because ROC curves are not convenient to compare classifiers, especially when their ROC curves intersect, the area under the ROC curve (AUC) is often used to quantify the performance. The AUC captures the performance in a single number, but it takes the whole range of cutoff values into account. In terms of the score distributions, AUC is expressed as:

$$\text{AUC} = \int_0^1 F_0(t)\mathrm{d}F_1(t) = \int_{-\infty}^{+\infty} F_0(t)f_1(t)\mathrm{d}t. \tag{2.9}$$

A larger AUC indicates superior performance. The area under the ROC curve, which is closely related to the Gini coefficient and the Kolmogorov-Smirnov statistic, has the following statistical interpretation: *the AUC of a classification method is the probability that a randomly chosen case will have a lower score than a randomly chosen non-case* (Fawcett, 2006).

However, Hand (2009) stated that the AUC is flawed as a performance measure, since it implicitly makes unrealistic assumptions

about the misclassification costs[1]. In the next section, we discuss the H measure, which is Hand's alternative to AUC.

2.3.3 The H measure

Hand (2009) proposed the H measure as a coherent alternative to the area under the ROC curve. We discuss this measure in detail, since it is closely related to the profit-based measure proposed in this dissertation. The main difference is that the H measure only explicitly states the misclassification costs and not the benefits.

The core of the H measure is the average classification *loss*, $Q(t)$, which is equal to the negative of the average classification profit, $P(t)$, with the following assumptions:

$$\begin{aligned} c &= \frac{c(1|0)}{c(1|0) + c(0|1)}, \\ b &= c(1|0) + c(0|1), \\ c(0|0) &= c(1|1) = 0. \end{aligned} \tag{2.10}$$

Hence, the focus is not on the benefits, but on the costs. The average classification loss is then equal to:

$$Q(t; c, b) = b \cdot \left[c\pi_0 \left(1 - F_0(t)\right) + (1 - c)\pi_1 F_1(t) \right]. \tag{2.11}$$

Assume that T is the optimal cutoff, i.e. the cutoff which minimizes $Q(t)$:

$$T = \arg\min_{\forall t} Q(t; c, b). \tag{2.12}$$

It is clear from Equation 2.11 that T depends on c, but not on b.

The next step is to calculate the *expected* minimum loss, by making assumptions regarding the joint probability density function of b and c, i.e. $w(b, c)$. A first assumption is the independence of b and c, namely that $w(b, c) = u(c)v(b)$ holds true, with $u(c)$ and $v(b)$

[1] In a recent paper, however, Hernández-Orallo et al. (2012) refute Hand's statement about the AUC being flawed.

the marginal probability density functions of c and b respectively. The expected minimum loss (L) is then equal to:

$$\mathrm{L} = \mathrm{E}[b] \int_0^1 Q(T(c); b, c) \cdot u(c) \mathrm{d}c, \tag{2.13}$$

with $\mathrm{E}[b] = 1$ for an appropriate choice for the unit in which b is measured. Furthermore, the probability density function of c is supposed to follow a beta distribution with parameters α and β, which is characterized as follows:

$$u_{\alpha,\beta}(p) = \begin{cases} p^{\alpha-1} \cdot (1-p)^{\beta-1} / B(1, \alpha, \beta) & \text{if } p \in [0,1] \\ 0 & \text{else} \end{cases} \tag{2.14}$$

with α and β greater than one, and:

$$B(p, \alpha, \beta) = \int_0^p q^{\alpha-1} \cdot (1-q)^{\beta-1} \mathrm{d}q. \tag{2.15}$$

Finally, to arrive at the H measure, a normalization is performed to obtain a performance measure bounded by zero and one:

$$H = 1 - \frac{\int_0^1 Q(T(c); b, c) \cdot u_{\alpha,\beta}(c) \mathrm{d}c}{\pi_0 \int_0^{\pi_1} c \cdot u_{\alpha,\beta}(c) \mathrm{d}c + \pi_1 \int_{\pi_1}^1 (1-c) \cdot u_{\alpha,\beta}(c) \mathrm{d}c}. \tag{2.16}$$

The denominator gives the misclassification loss for the worst classifier, i.e. a random classifier. Also, the integration over $c = [0; 1]$ corresponds to an integration over the entire ROC curve.

A close inspection of the H measure shows that only a part of the measure changes with varying classification techniques, holding the data set constant:

$$H = 1 + \frac{H_{var} - \pi_0 \int_0^1 c \cdot u_{\alpha,\beta}(c) \mathrm{d}c}{\pi_0 \int_0^{\pi_1} c \cdot u_{\alpha,\beta}(c) \mathrm{d}c + \pi_1 \int_{\pi_1}^1 (1-c) \cdot u_{\alpha,\beta}(c) \mathrm{d}c}, \tag{2.17}$$

with H_{var} defined as:

$$H_{var} = \int_0^1 [\pi_0 c F_0 - \pi_1 (1-c) F_1] \cdot u_{\alpha,\beta}(c) \mathrm{d}c. \tag{2.18}$$

This variable part of H, H_{var}, is an expression which will be useful in Chapter 3.

2.4 A profit-based performance measure

The previous section discussed traditional classification performance measures, and the recently proposed H measure. In this section, we will propose a new general framework for profit-based classification performance measurement, which is founded on the average classification profit given by Definition 2.1. First, a deterministic approach will be discussed after which a probabilistic classification performance measure will be proposed. Finally, we will explore the link of this new performance measure with the ROC curve.

2.4.1 Deterministic approach: MP

Most traditional classification performance measures focus on a statistical criterion to determine classification quality. However, Verbeke et al. (2012) stated that one should focus directly on profitability, which is an important business goal. They argue that the average profit, since it depends on the cutoff value t, should be optimized. This leads to the concept of the maximum profit (Verbeke et al., 2012).

Definition 2.2. ***The maximum profit, MP****, of a classification technique is the profit resulting from the classification outcome when the optimal cutoff, T, is used. The maximum profit is analytically expressed as:*

$$MP = \max_{\forall t} P(t; b_0, c_1, c^*) = P(T; b_0, c_1, , c^*), \tag{2.19}$$

with T the optimal threshold.

The optimal threshold, satisfies the first order condition for the maximization of the average profit P:

$$\frac{f_0(T)}{f_1(T)} = \frac{\pi_1(c_1 + c^*)}{\pi_0(b_0 - c^*)}, \tag{2.20}$$

and thus T depends on the class distribution and on the ratio of costs and benefits, denoted with θ:

$$\theta = \frac{c_1 + c^*}{b_0 - c^*}. \tag{2.21}$$

Hence, T is not dependent on the measurement scale, but only on the ratio of the cost benefit parameters, which is explicitly expressed by writing $T(\theta)$. Since it is assumed that the cost benefit parameters (b_0, c_1, c^*) are positive and that $c^* < b_0$, the value of θ ranges from zero to plus infinity. Equation 2.20 has an appealing graphical interpretation on the ROC curve, as will be shown in Section 2.4.3.

MP in itself can be used as a classification performance measure, as proposed by Verbeke et al. (2012). In fact, doing so allows the practitioner to select the model with the highest incremental profitability. Moreover, contrary to traditional performance measures, the optimal cutoff is clearly defined and the fraction of the customer base towards which the action should be undertaken is equal to:

$$\bar{\eta}_{\text{mp}} = \pi_0 F_0(T) + \pi_1 F_1(T). \tag{2.22}$$

It is clear from this expression that the optimal fraction to be targeted depends on the optimal cutoff, which is determined by three other elements, i.e. the class distribution, the score distributions, and the classification costs and benefits.

2.4.2 Probabilistic approach: EMP

The MP measure is a *deterministic* approach for profit-based classification performance measurement, since the costs and benefits are supposed to be known. In reality however, it is not straightforward to estimate precise values for these parameters. Therefore, a probability distribution is adopted for the cost and benefit parameters, leading to a *probabilistic* profit-based performance measure, the *expected* maximum profit of a classifier.

Definition 2.3. ***The expected maximum profit, EMP****, is the expectation of the maximal profit of a classifier with respect to the distribution of classification costs, and is equal to:*

$$EMP = \int_{b_0} \int_{c_1} \int_{c^*} P(T(\theta); b_0, c_1, c^*) \cdot h(b_0, c_1, c^*) \mathrm{d}c^* \mathrm{d}c_1 \mathrm{d}b_0, \tag{2.23}$$

with $h(b_0, c_1, c^*)$ the joint probability density of the classification costs and benefits.

The expected maximum profit measures the effectiveness of a classification technique in terms of profitability, taking into account the uncertainty of the classification costs and benefits.

Equation 2.23 is the most general expression for the expected maximum profit of a given classifier. Note that for *each* combination of (b_0, c_1, c^*), the optimal threshold T is determined through Equation 2.20. The profit is then calculated for the optimal threshold and weighed with the probability density for this particular combination of classification costs. Integrating over all possible benefits and costs leads to the expected maximum classification profit. When the costs are known with certainty, the function $h(b_0, c_1, c^*)$ is a Dirac impulse at the known values, and EMP simplifies to MP.

Analogously to the deterministic optimal fraction $\bar{\eta}_{\text{mp}}$, the expected profit maximizing fraction, $\bar{\eta}_{\text{emp}}$, is the fraction of the cases towards which an action is undertaken:

$$\bar{\eta}_{\text{emp}} = \int_{b_0} \int_{c_1} \int_{c^*} \left[\pi_0 F_0\left(T(\theta)\right) + \pi_1 F_1(T\left(\theta\right))\right] \cdot h(b_0, c_1, c^*) \mathrm{d}c^* \mathrm{d}c_1 \mathrm{d}b_0, \tag{2.24}$$

Although the previous discussion only focused on the profit related to a classification outcome, there is a link to the ROC curve of the classifier, as will be explored in the next section.

2.4.3 EMP and ROC curves

Also in the context of the EMP framework, the ROC curve has its merit. Since each point on the ROC curve corresponds to a threshold t, the optimal cutoff T, is located somewhere on the curve.

Theorem 1. *If the ROC curve is convex*[2] *and continuously differ-*

[2]It was already pointed out by Hand Hand (2009) that the machine learning community in the context of ROC curves by convention considers a function $g(x)$ convex if $g(\lambda x+(1-\lambda)y) \geq \lambda g(x)+(1-\lambda)g(y)$ for $0 < \lambda < 1$. A twice differentiable function $g(x)$ is convex if $\partial^2 g(x)/\partial x^2 \leq 0$. The mathematical community on the contrary adopts the definition of convexity with the inequality sign reversed.

entiable, the optimal cutoff, for given costs c_1 and c^*, and benefit b_0, corresponds to the point on the ROC curve for which the tangent slope equals $\pi_1\theta/\pi_0 = [\pi_1(c_1 + c^*)]/[\pi_0(b_0 - c^*)]$.

Proof of Theorem 1. Taking the derivative of the average classification profit, given by Equation 2.3, with respect to F_1 leads to the first order condition:

$$\frac{\partial F_0}{\partial F_1} = \frac{\pi_1(c_1 + c^*)}{\pi_0(b_0 - c^*)} = \frac{\pi_1\theta}{\pi_0}.$$

Taking the second derivative leads to the second order condition:

$$\frac{\partial^2 P}{\partial F_1^2} = (b_0 - c^*)\pi_0 \frac{\partial^2 F_0}{\partial F_1^2} \leq 0.$$

Hence, due to the assumed convexity of the ROC curve, because the cost benefit parameters are positive, and because $c^* < b_0$, the point with tangent slope $\pi_1\theta/\pi_0$ maximizes the average classification profit. □

When the ROC curve is not convex, and thus the slope is not monotonically decreasing, there may be multiple points on the ROC curve satisfying the first order condition. More specifically, points in the concave region are not optimal for *any* θ, in which case the convex hull needs to be defined.

Theorem 2. *The convex hull of a (non-convex) ROC curve defines a set of points where each point corresponds to the optimal cutoff for a certain value of the cost benefit ratio $\theta \in [0, +\infty)$.*

Proof of Theorem 2. Consider a ROC curve with one concave region, as depicted in Figure 2.2. The tangent slope in R_A, R_B and R_C is equal, and the convex hull is defined by the ROC curve segments $[R_I, R_A]$ and $[R_C, R_E]$, together with the line segment $\overline{R_A R_C}$. It has to be shown that none of the points in the concave region optimize the profit for any θ. Assume that for a certain value of θ both points $R_{(1)}$, somewhere within the concave region, and $R_{(2)}$, somewhere

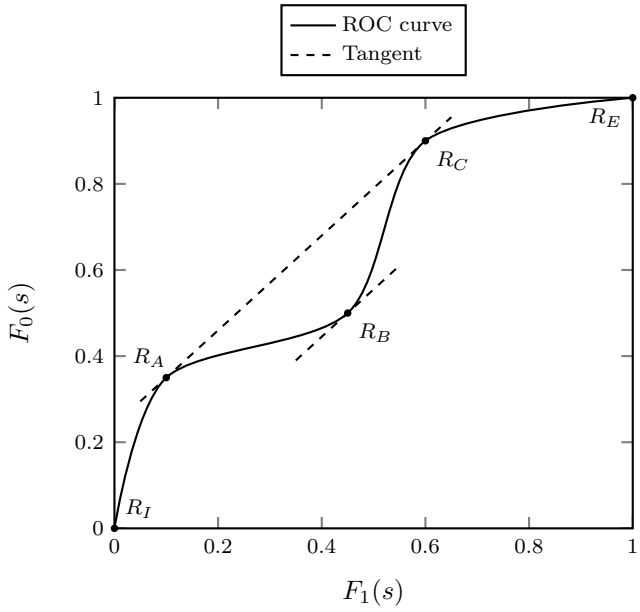


Figure 2.2: A non-convex ROC curve and its convex hull.

outside the concave region, satisfy the first order condition. Then, the difference in profit for both points is given by:

$$\begin{aligned}\Delta P = P_{(2)} - P_{(1)} &= \pi_0(b_0 - c^*)\Delta F_0 - \pi_1(c_1 + c^*)\Delta F_1 \\ &= \pi_0(b_0 - c^*)\Delta F_1 \left[\frac{\Delta F_0}{\Delta F_1} - \frac{\pi_1\theta}{\pi_0}\right].\end{aligned}$$

Since both points satisfy the first order condition, the second term between the squared brackets is equal to the slope of the ROC curve at that particular point.

There are three possibilities for the location of $R_{(1)}$; (1) it lies in the interval (R_A, R_B), (2) it lies in the interval (R_B, R_C) or (3) it coincides with R_B. As will be shown, ΔP will be positive in each of these three situations.

1. In this situation, $R_{(2)}$ has to be situated in the interval $(R_C, R_E]$. Hence, $\Delta F_1 > 0$ and $\Delta F_0/\Delta F_1 > \pi_1\theta/\pi_0$, resulting in $\Delta P > 0$.

2. Now, $R_{(2)}$ is located in the interval $[R_I, R_A)$. Therefore, $\Delta F_1 < 0$ and $\Delta F_0/\Delta F_1 < \pi_1\theta/\pi_0$, resulting in $\Delta P > 0$.

3. In this scenario, $R_{(1)}$ coincides with R_B. Both R_A and R_C qualify for $R_{(2)}$. For the former choice of $R_{(2)}$, $\Delta F_1 < 0$ and $\Delta F_0/\Delta F_1 < \pi_1\theta/\pi_0$. For the latter choice, $\Delta F_1 > 0$ and $\Delta F_0/\Delta F_1 > \pi_1\theta/\pi_0$. In both situations, $\Delta P > 0$, and the profit related to R_A and R_C is equal.

Moreover, every point on the line segment $\overline{R_A R_C}$ will maximize the profit for this θ which gives rise to a slope equal to the slope of the line segment[3]. Consequently, the convex hull consists of a set of points which all maximize the average profit for a certain value of $\theta \in [0, +\infty)$. □

The interesting implication of this fact is that an integration over a range of θ values is equivalent to an integration over the corresponding part of the ROC curve, a characteristic shared with, among others, the AUC and the H measure. The EMP measure thus considers a range of cutoff values.

2.5 Conclusion

In this chapter, we proposed the EMP framework for classification performance measurement. The central concept in this framework is the incorporation of costs and benefits, in order to arrive at the classification profit. Based on the classification profit, a probabilistic profit-based performance measure is proposed, the EMP measure, which takes into account uncertainty about the costs and benefits. Furthermore, the EMP measure provides more guidance about the practical implementation, in that it provides the fraction of the instances which should be targeted with the action, in order to achieve

[3]To obtain a classifier with the characteristics given by a point on the line segment, it is necessary to interpolate between the classifiers corresponding to points R_A and R_C, a procedure which has been described by Fawcett (2006).

optimal profitability. In the following two chapters, the general EMP framework will be implemented for customer churn prediction and consumer credit scoring.

Chapter 3

EMP for customer churn prediction

The previous chapter outlined a general framework for profit-based classification performance measurement, in which the classification costs and benefits do not have a specific meaning yet. This chapter will implement the general EMP framework for customer churn prediction (CCP), which is an important application of classification models within marketing analytics.

Section 3.1 will provide a brief overview of customer churn prediction and its importance in the industry. Section 3.2 will analyze the drivers of the profitability of a customer churn management campaign, which enables us to implement a profit-based performance measure for CCP in Section 3.3. The actual parameter values and the empirical estimation of EMP are discussed in Section 3.4, whereas Section 3.5 explores the link with the H measure and shows that the H measure can be used as an approximation for the EMP measure for customer churn. Finally, the developed performance measures will be tested in an extensive case study, of which the findings are reported in Section 3.6.

3.1 Introduction

Customer churn has become a very important business problem. During the last decade, the number of mobile phone users has increased drastically, with expectations of 5.6 billion mobile phone subscribers in 2011[1], which is around 80% of the world population. Hence, telecommunication markets are getting saturated, particularly in developed countries, and many companies and organizations are confronted with customer churn. For instance, wireless telecom operators report annual churn rates up to 40% of their customer base (Neslin et al., 2006).

Hence, the emphasis is shifting from the attraction of new customers to the retention of the existing customer base. The literature reports that customer retention is profitable because: (1) acquiring a new client is five to six times more costly than retaining an existing customer (Athanassopoulos, 2000; Bhattacharya, 1998; Colgate and Danaher, 2000; Rasmusson, 1999); (2) long-term customers generate higher profits, tend to be less sensitive to competitive marketing activities, become less costly to serve, and may provide new referrals through positive word-of-mouth, while dissatisfied customers might spread negative word-of mouth (Colgate et al., 1996; Ganesh et al., 2000; Mizerski, 1982; Reichheld, 1996; Stum and Thiry, 1991; Zeithaml et al., 1996; Paulin et al., 1998); (3) losing customers leads to opportunity costs because of reduced sales (Rust and Zahorik, 1993). A small improvement in customer retention can therefore lead to a significant increase in profit (Van den Poel and Larivière, 2004). For successful targeted marketing campaigns, it is crucial that operators are able to identify clients with a high probability to churn in the near future.

In this context, customer churn prediction models play a crucial role and they are increasingly being researched. Numerous classification techniques have been applied to predict churn, including traditional statistical methods such as logistic regression (Burez and

[1] www.eito.com

Van den Poel, 2009; Lemmens and Croux, 2006; Neslin et al., 2006), non-parametric statistical models like for instance k-nearest neighbor (Datta et al., 2000), decision trees (Lima et al., 2009; Wei and Chiu, 2002), and neural networks (Au et al., 2003; Hung et al., 2006). An extensive literature review on customer churn prediction modeling can be found in Verbeke et al. (2012). When analyzing and comparing these classification techniques for customer churn prediction, it is imperative to have an adequate performance measure.

As with many data mining applications, and especially for customer churn prediction, the main reason for employing a classification model is profitability enhancement. This is exactly what is measured by the EMP measure. Therefore, the EMP measure will be implemented for CCP. The main advantage of this metric is that it not only unambiguously identifies the classifier which maximizes the profit for a given customer retention campaign. It also determines the fraction of the customer base which should be targeted in this retention campaign in order to maximize the profit. This is a crucial help for practitioners, since deviating from the optimal fraction leads to suboptimal profits.

3.2 Costs and benefits in CCP

In Chapter 2, no assumptions were made about the distribution of the cost and benefit parameters (b_0, c_1, c^*) other than that their values are positive. However, when the EMP framework is to be applied, an interpretation has to be given to these parameters, which requires domain specific knowledge. This section will discuss the drivers of the profitability of a customer churn management campaign.

Figure 3.1 schematically represents the dynamical process of customer churn and retention within a customer base. New customers flow into the customer base by subscribing to a service of an operator, and existing customers flow out of the customer base by churning. When setting up a churn management campaign, the fraction η of the customer base with the highest propensity to churn is contacted

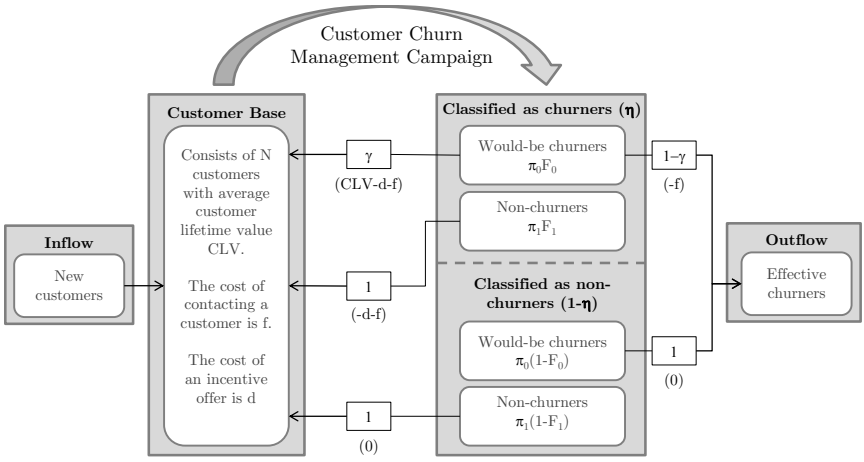


Figure 3.1: Schematic representation of customer churn and retention dynamics within a customer base.

at a cost f per person and is offered an incentive with cost d. In this fraction, there are true would-be churners and false would-be churners. In the latter group everyone accepts the incentive and does not churn, as they never had the intention. From the former group, a fraction γ accepts the offer and thus results in gained remaining customer lifetime value[2] (CLV), whereas the fraction $(1-\gamma)$ effectively churns. In the fraction $(1-\eta)$, which is not targeted, all would-be churners effectively churn, and all non-churners remain with the company. The benefits per customer related to each flow are shown between brackets in Figure 3.1. These are *incremental* benefits, as compared to not undertaking the customer churn management campaign.

This process was described by Neslin et al. (2006), who established the following expression for the total profit of a retention campaign:

$$\text{Profit } = N\eta\left[(\gamma \cdot CLV + d(1-\gamma))\,\pi_0\lambda - d - f\right] - A, \qquad (3.1)$$

with η the fraction of the customer base that is targeted, CLV the remaining customer lifetime value, d the cost of the incentive, f

[2]Not the entire CLV need to be taken into account, but the remaining CLV after the customer would have churned.

the cost of contacting the customer, and A the fixed administrative costs. The lift coefficient, λ, is the percentage of churners within the targeted fraction η of customers, divided by the base churn rate, π_0. Lastly, γ is the fraction of the would-be churners accepting the offer, or alternatively it is interpreted as the probability that a targeted churner accepts the offer and does not churn. It is assumed that CLV, A, f, and d are positive, and that $CLV > d$. Note that η depends on the choice for the threshold t, and thus the company can influence the size of the targeted fraction.

Equation 3.1 can be expressed in terms of the score distributions. Moreover, if the average rather than the total profit is considered, and the fixed cost A, irrelevant for classifier selection, is discarded, it is possible to obtain a functional form equivalent to the expression for the classification profit P in Definition 2.1. To work out the conversion, note that:

$$\begin{aligned}\eta(t) &= \pi_0 F_0(t) + \pi_1 F_1(t),\\ \lambda(t) &= \frac{F_0(t)}{\pi_0 F_0(t) + \pi_1 F_1(t)}.\end{aligned}$$

Moreover, two dimensionless parameters are introduced, being $\delta = d/CLV$ and $\phi = f/CLV$.

Definition 3.1. ***The average classification profit for customer churn****, P^{ccp}, is the interpretation of Definition 2.1 specifically for customer churn:*

$$\begin{aligned}P^{ccp}(t;\gamma,CLV,\delta,\phi) = \; & CLV\,(\gamma(1-\delta)-\phi)\\ & \cdot \pi_0 F_0(t) - CLV(\delta+\phi)\cdot \pi_1 F_1(t). \qquad (3.2)\end{aligned}$$

Comparison with Equation 2.3 shows that:

$$\begin{aligned}b_0 &= \gamma(CLV - d) = \gamma(1-\delta)\cdot CLV,\\ c_1 &= d = \delta \cdot CLV, \qquad (3.3)\\ c^* &= f = \phi \cdot CLV.\end{aligned}$$

Now that the cost and benefit parameters are identified, the profit-based performance measure for CCP can be defined.

3.3 A profit-based measure for CCP

3.3.1 MP for customer churn

When all parameter values in Equation 3.2 are precisely known, the deterministic performance measure, MP, specified by Definition 2.2, can be worked out for a customer churn context.

Definition 3.2. ***The maximum profit measure for customer churn, MP^{ccp}****, is the interpretation of Definition 2.2 in a customer churn setting:*

$$MP^{ccp} = \max_{\forall t} P^{ccp}(t; \gamma, CLV, \delta, \phi). \tag{3.4}$$

As pointed out by Verbeke et al. (2012), the MP measure is preferred over the commonly used top-decile lift. Setting the targeted fraction of customers to e.g. ten percent is a purely arbitrary choice and most likely leads to suboptimal profits and model selection. Since the ultimate goal of a company setting up a customer retention campaign is to minimize the costs associated with customer churn, it is logical to evaluate and select a customer churn prediction model by using the maximum obtainable profit as the performance measure. Moreover, it has another advantage, which is very appealing to practitioners, in the sense that it is possible to determine the optimal fraction, $\bar{\eta}^{ccp}_{\text{mp}}$. This quantity represents how many customers should be targeted for profit maximization, and can be expressed as:

$$\bar{\eta}^{ccp}_{mp} = \pi_0 F_0(T) + \pi_1 F_1(T), \tag{3.5}$$

with:

$$T = \arg\max_{\forall t} P^{ccp}(t; \gamma, CLV, \delta, \phi). \tag{3.6}$$

The combination of the maximum profit MP^{ccp} and the optimal fraction η^{ccp}_{mp} provides telecom operators with a rigorous framework for making operational decisions with profit maximization as main goal.

3.3.2 EMP for customer churn

In the previous section it was assumed that accurate estimates for the parameters in the expression for P^{ccp} are available, in which case a deterministic performance measure can be employed. Often however, there is significant uncertainty involved. Specifically for the problem of customer churn, there are four parameters, of which customer lifetime value (CLV), the cost of the incentive (d) and the contacting cost (f) can be estimated with sufficient reliability. However, about the probability of a churner accepting the retention offer (γ) much less is known. Therefore, this probability is considered a random variable, which introduces uncertainty in the estimation of the maximum profit, leading to a probabilistic performance measure.

Definition 3.3. ***The expected maximum profit measure for customer churn, EMP^{ccp}****, is the interpretation of Definition 2.3 in a customer churn setting:*

$$EMP^{ccp} = \int_{\gamma} P^{ccp}(T(\gamma); \gamma, CLV, \delta, \phi) \cdot h(\gamma) \mathrm{d}\gamma, \tag{3.7}$$

with T, the optimal cutoff for a given γ, and $h(\gamma)$ the probability density function for γ.

Note that T depends on γ through the cost benefit ratio θ^{ccp}. Analogously to the H measure, a beta distribution is proposed, because of its flexibility (see Section 2.3.3). The parameters are named α' and β', where the prime aims to distinguish these parameters from those of the H measure. The cost benefit ratio for customer churn, θ^{ccp}, is equal to:

$$\theta^{ccp} = \frac{c_1 + c^*}{b_0 - c^*} = \frac{\delta + \phi}{\gamma(1-\delta) - \phi}, \tag{3.8}$$

where γ is the parameter introducing the uncertainty.

At this point, an important remark has to be made regarding the relation between the ROC curve of a classifier and its EMP^{ccp}. In Section 2.4.3, it was stated that, theoretically, calculating the

EMP considers all points of the convex hull of a ROC curve. With EMP^{ccp}, the cost and benefit parameters obey certain assumptions, specific for a customer churn context. The value of θ^{ccp} now ranges from $-(\delta+\phi)/\phi$ to $-\infty$ for $\gamma \in [0, \phi/(1-\delta))$, and from $+\infty$ to $(\delta+\phi)/(1-\delta-\phi)$ for $\gamma \in (\phi/(1-\delta), 1]$. Hence, for a certain interval of γ, θ^{ccp} is negative. This corresponds to the situation when the probability of a churner accepting the retention offer is so low that the expected net gained customer lifetime value, i.e. $\gamma(CLV-d)$, does not offset the cost of contacting the customers, f. In this case, the optimal decision is not to contact any customer, which corresponds to the origin of the ROC curve.

A second observation concerns the fact that θ^{ccp} does not extend over the entire positive set of real numbers, $\mathbb{R}^+$, since it only spans $[(\delta+\phi)/(\gamma-\gamma\delta-\phi), +\infty)$. Therefore, points on the convex hull of a ROC curve with a slope smaller than $(\pi_1/\pi_0)\cdot(\delta+\phi)/(1-\delta-\phi)$ are not optimal for *any* value of $\gamma \in [0;1]$. Consequently, when calculating EMP^{ccp}, not all points of the convex hull are taken into account. Moreover, the more skewed the data set towards non churners, the smaller the part of the convex hull which is taken into account. Intuitively this makes sense since, for very low churn rates, the percentage of churners in a large targeted fraction will quickly become insignificant and hence jeopardize the profitability. The optimal targeted fraction can also be determined when dealing with the expected maximal profit. The expected profit maximizing fraction for customer churn, $\bar{\eta}^{ccp}_{\text{emp}}$, is the fraction of the customer base which is targeted when taking into account the uncertainty about γ:

$$\bar{\eta}^{ccp}_{\text{emp}} = \int_{\gamma} \left[\pi_0 F_0\left(T(\theta^{ccp})\right) + \pi_1 F_1(T(\gamma))\right] \cdot h(\gamma) \mathrm{d}\gamma, \tag{3.9}$$

with $T(\gamma)$ being the optimal cutoff, as defined in Equation 3.6.

Hence, there are two frameworks available. One framework, based on MP^{ccp}, takes a deterministic approach and considers all parameters in the profit equation to be known. The second framework is based on EMP^{ccp} and provides practitioners with a probabilistic method to evaluate profits and losses, reflecting the uncertainty about the

response rate γ. Due to its specific costs and benefits, performance measurement in a customer churn setting cannot simply be done by applying the H measure. However, it is possible to approximate the EMP^{ccp} measure to a certain degree, as will be shown in Section 3.5.

3.3.3 Sensitivity of EMP

The EMP^{ccp} measure incorporates uncertainty about the acceptance rate γ, but for the parameters CLV, δ, and ϕ, fixed values are assumed. In this paragraph, the sensitivity of the EMP^{ccp} measure to small variations in these fixed parameters will be assessed. Therefore, the first derivative of Equation 3.7 with respect to CLV is calculated:

$$\frac{\partial \text{EMP}^{ccp}}{\partial CLV} = \int_{\gamma} \frac{\partial P^{ccp}(T(\theta^{ccp}); \gamma, CLV, \delta, \phi)}{\partial CLV} h(\gamma) \text{d}\gamma.$$

Note that the optimal cutoff, $T(\theta^{ccp})$, depends on the value of CLV. In what follows, the partial derivative within the integration is worked out:

$$\begin{aligned}\frac{\partial P^{ccp}(T)}{\partial CLV} =& (\gamma(1-\delta) - \phi) \cdot \pi_0 F_0(T) - (\delta + \phi) \cdot \pi_1 F_1(T) \\ &+ CLV\,(\gamma(1-\delta) - \phi) \cdot \pi_0 f_0(T) \frac{\partial T}{\partial CLV} \\ &- CLV(\delta + \phi) \cdot \pi_1 f_1(T) \frac{\partial T}{\partial CLV},\end{aligned}$$

which, by using Equation 3.2 and Equation 2.20, can be rewritten as follows:

$$\begin{aligned}\frac{\partial P^{ccp}(T)}{\partial CLV} =& \frac{P^{ccp}(T)}{CLV} + CLV \frac{\partial T}{\partial CLV} (\gamma(1-\delta) - \phi) \\ & \pi_1 f_1(T) \cdot \left(\frac{\pi_0 f_0(T)}{\pi_1 f_1(T)} - \theta^{ccp} \right).\end{aligned}$$

Now, Equation 3.8 can be applied, which leads to the last term between brackets being equal to zero, and thus:

$$\frac{\partial \text{EMP}^{ccp}}{\partial CLV} = \int_{\gamma} \frac{P^{ccp}(T)}{CLV} h(\gamma) \text{d}\gamma = \frac{\text{EMP}^{ccp}}{CLV}. \tag{3.10}$$

In other words this means that when CLV is changed, holding δ and ϕ constant, EMP^{ccp} changes proportionally. Note that holding δ and ϕ constant in fact means that the cost of the retention offer *as a percentage of CLV*, and the cost of contacting a customer *as a percentage of CLV* remain constant.

Analogously, for variations in ϕ the following equation can be derived:

$$\begin{aligned}\frac{\partial \text{EMP}^{ccp}}{\partial \phi} &= -CLV \int_{\gamma} \left[\pi_0 F_0(T) + \pi_1 F_1(T)\right] h(\gamma) \mathrm{d}\gamma \\ &= -CLV \cdot \bar{\eta}^{ccp}_{\text{emp}}, \end{aligned} \tag{3.11}$$

where $\bar{\eta}^{ccp}_{\text{empc}}$ is the expected profit maximizing fraction. For changes in δ, the following holds true:

$$\begin{aligned}\frac{\partial \text{EMP}^{ccp}}{\partial \phi} &= -CLV \int_{\gamma} \left[\gamma\pi_0 F_0(T) + \pi_1 F_1(T)\right] h(\gamma) \mathrm{d}\gamma \\ &= -CLV \cdot \bar{\rho}^{ccp}_{\text{emp}}, \end{aligned} \tag{3.12}$$

with $\bar{\rho}^{ccp}_{\text{emp}}$, the expected fraction of the customer base which *accepts the retention offer*. In fact, this implies that the larger the optimal targeted fraction (or the fraction accepting the offer), the more sensitive the expected maximum profit for variations in ϕ (or δ). Also note that an increase in CLV, while holding d and f constant instead of δ and ϕ, corresponds to a parallel decrease in δ and ϕ. In Section 3.6 the sensitivity of the rankings will be analyzed in a real life case study.

3.4 EMP in practice

3.4.1 Parameter values

To employ the MP^{ccp} and EMP^{ccp} measures, it is necessary to obtain values for the parameters in the profit function. The values for these parameters, such as CLV, are industry specific and may even vary from company to company. This chapter focuses on customer churn

prediction in the telecom industry and estimates for the parameters are based on values reported in the literature (Neslin et al., 2006; Burez and Van den Poel, 2007) and information from telecom operators. When a specific company uses MP^{ccp} or EMP^{ccp} for selecting the optimal model, they can plug in their own estimates for the parameters. In this study, the values of the parameters CLV, d, and f are taken as € 200, € 10, and € 1 respectively.

The parameter γ, representing the response rate, is much more difficult to estimate. For the MP^{ccp} criterion, a single point estimate would be required, which corresponds to one degree of freedom. For the EMP^{ccp} measure however, there are two degrees of freedom, i.e. α and β. This enables the practitioner to define an expected value and a standard deviation, where the latter accounts for the uncertainty in the practitioner's estimation. The α and β parameters can be obtained by solving the following system of equations:

$$\begin{cases} \mathrm{E}[\gamma] = \mu = \alpha/(\alpha+\beta) \\ \mathrm{Var}[\gamma] = \sigma^2 = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]. \end{cases}$$

which yields:

$$\begin{cases} \alpha = \mu\left((1-\mu)\mu/\sigma^2 - 1\right) \\ \beta = (1-\mu)\left((1-\mu)\mu/\sigma^2 - 1\right) \end{cases}$$

There is only one restriction, namely that α and β need to be strictly greater than one in order to obtain a unimodal beta distribution.

Neslin et al. (2006) assumed values ranging from 10% to 50%, therefore we propose the expected value and standard deviation of the acceptance rate to be equal to 30% and 10% respectively. This leads to the parameters α' and β' being 6 and 14 respectively, which corresponds to the probability density function plotted in Figure 3.3 (solid line). When MP^{ccp} is calculated, a point estimate for γ is required, in which case the expected value, equal to 30%, is taken.

3.4.2 Empirical estimation of EMP

In the theoretical derivation of the EMP measure in Chapter 2, it has been assumed that the ROC curve is smooth. An empirical ROC

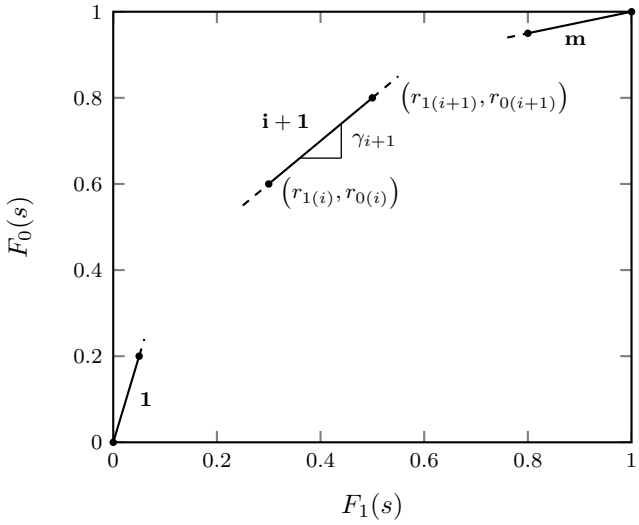


Figure 3.2: The convex hull of an empirical ROC curve.

curve however, is a stepwise function with diagonal elements if there are ties. Hand (2009) describes a procedure to calculate the empirical ROC curve and its convex hull and gives an algorithm to calculate the H measure given an empirical convex hull. A similar procedure will be derived to calculate the EMP^{ccp} measure.

Assume the convex hull consists of m segments, and let (r_{1i}, r_{0i}) be the end points of the segments, with $(r_{10}, r_{00}) = (0, 0)$ and $(r_{1m}, r_{0m}) = (1, 1)$, as illustrated in Figure 3.2. Each end point corresponds to a score s_i. Now, if a score is selected in the interval $[s_i, s_{i+1}]$, it will maximize the profit for the γ which fulfills the following equality:

$$\frac{r_{0(i+1)} - r_{0(i)}}{r_{1(i+1)} - r_{1(i)}} = \frac{\pi_1(\delta + \phi)}{\pi_0\left((1-\delta)\gamma - \phi\right)} \tag{3.13}$$

The γ's will be indexed corresponding to the upper end of the segment. Thus, γ_{i+1} is defined as:

$$\gamma_{i+1} = \frac{\pi_1(\delta + \phi)\left(r_{1(i+1)} - r_{1(i)}\right)}{\pi_0(1-\delta)(r_{0(i+1)} - r_{0(i)})} + \frac{\phi}{1-\delta}, \tag{3.14}$$

and $\gamma_{(0)} = 0$. However, γ is not bounded by one, when running over the ROC curve. When approaching the point $(1, 1)$, γ becomes infinitely large. Therefore, when calculating EMP^{ccp}, one replaces the series $\{\gamma_i | i = 0 \dots m\}$ with $\{\gamma_i | i = 0 \dots k+1\}$, with $k = \max\{i | \gamma_i < 1\}$, and $\gamma_{k+1} = 1$. The EMP^{ccp} measure can then be estimated by:

$$\begin{aligned}\text{EMP}^{ccp} =& CLV \sum_{i=0}^{k} \int_{\gamma_i}^{\gamma_{i+1}} \left[(\gamma(1-\delta) - \phi)\, \pi_0 r_{0i}\right] h(\gamma) \mathrm{d}\gamma \\ & - CLV \sum_{i=0}^{k} \int_{\gamma_i}^{\gamma_{i+1}} \left[(\delta + \phi)\pi_1 r_{1i}\right] h(\gamma) \mathrm{d}\gamma. \end{aligned} \tag{3.15}$$

Since γ is constant over the segments, and r_{0i} and r_{1i} are constant in the nodes, this can be written as:

$$\begin{aligned}\text{EMP}^{ccp} =& CLV \sum_{i=0}^{k} \Bigg\{ \left[(1-\delta)\pi_0 r_{0i}\right] \int_{\gamma_i}^{\gamma_{i+1}} \gamma h_{\alpha',\beta'}(\gamma) \mathrm{d}\gamma \\ & - \left[\phi \pi_0 r_{0i} + (\delta + \phi)\pi_1 r_{1i}\right] \int_{\gamma_i}^{\gamma_{i+1}} h_{\alpha',\beta'}(\gamma) \mathrm{d}\gamma \Bigg\}. \end{aligned} \tag{3.16}$$

From Equation 3.14, it is clear that the calculated γ_i's are all larger than or equal to $\phi/(1-\delta) = \phi/K$. Hence, there is no need to worry about the range of γ resulting in negative slopes. Secondly, there is an upper bound for γ, which is equal to one. Thus, the integration does not cover the entire ROC curve by definition. With the definition of $B(x, \alpha, \beta)$ in Equation 2.15, the sum in the above expression for EMP^{ccp} can be written as:

$$\begin{aligned}\sum_{i=0}^{k} \Bigg\{ & \left[(1-\delta)\pi_0 r_{0i}\right] \frac{B(\gamma_{i+1}, \alpha'+1, \beta') - B(\gamma_i, \alpha'+1, \beta')}{B(1, \alpha', \beta')} \\ & - \left[\phi \pi_0 r_{0i} + (\delta + \phi)\pi_1 r_{1i}\right] \frac{B(\gamma_{i+1}, \alpha', \beta') - B(\gamma_i, \alpha', \beta')}{B(1, \alpha', \beta')} \Bigg\}. \end{aligned} \tag{3.17}$$

In the above formulae, the class zero must represent the churners and the labels are not interchangeable. This is caused by the fact that the integration does not span the entire ROC curve.

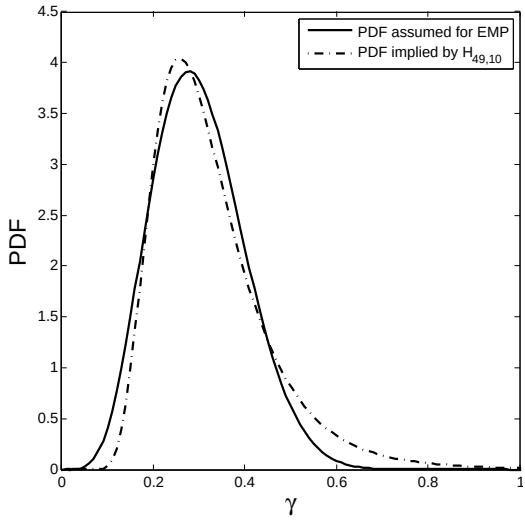


Figure 3.3: The graph shows the proposed density function for the response rate in the EMP^{ccp} measure, and the density function implied by the H measure with optimized parameters.

3.5 EMP and the H Measure

As mentioned in Section 2.3.3, the H measure is an expected minimum loss measure, with certain assumptions about the distribution of the cost and benefit parameters. These assumptions imply that the H measure takes into account all points on the convex hull. The EMP^{ccp} measure however, only takes into account a part of the convex hull. In this section, it will be analyzed whether it is possible to approximate the EMP^{ccp} measure with the H measure. After all, it would be convenient if the EMP^{ccp} measure fits in an existing performance measurement framework.

The discussion will focus on the variable part of the H measure, defined in Equation 2.18. Only this part of the H measure changes with varying classifiers, holding the data set constant, and thus, the ranking of techniques for a data set only depends on H_{var}. At this point, it is assumed that the variable c in H_{var} is a function of γ, δ,

and ϕ:

$$c = \frac{\gamma(1-\delta) - \phi}{\gamma(1-\delta) + \delta}.$$

In what follows, a variable transformation from c to γ will be carried out in the expression for H_{var}. Expressions for $\mathrm{d}c$ and $u_{\alpha,\beta}(c)$ as a function of γ are derived:

$$\mathrm{d}c = \frac{(1-\delta)(\delta+\phi)}{((1-\delta)\gamma+\delta)^2}\mathrm{d}\gamma,$$
$$u_{\alpha,\beta}(c(\gamma)) = \frac{1}{B(1,\alpha,\beta)}\left(\frac{(1-\delta)\gamma-\phi}{(1-\delta)\gamma+\delta}\right)^{\alpha-1}\left(\frac{\delta+\phi}{(1-\delta)\gamma+\delta}\right)^{\beta-1}.$$

Hence, the variable substitution in the expression for H_{var} leads to:

$$\begin{aligned} H_{var} &= \int_0^1 [\pi_0 c F_0 - \pi_1(1-c)F_1] \cdot u_{\alpha,\beta}(c)\mathrm{d}c \\ &= \int_{\phi/(1-\delta)}^{+\infty} [((1-\delta)\gamma-\phi)\pi_0 F_0 - (\delta+\phi)\pi_1 F_1] \cdot \\ &\qquad \underbrace{\frac{(1-\delta)(\phi+\delta)^\beta}{B(1,\alpha,\beta)} \cdot \frac{((1-\delta)\gamma-\phi)^{\alpha-1}}{((1-\delta)\gamma+\delta)^{\alpha+\beta+1}}}_{g(\gamma)} \mathrm{d}\gamma \\ &= \frac{1}{CLV}\int_{\phi/(1-\delta)}^{+\infty} P^{ccp}(T(\gamma); CLV, \gamma, \delta, \phi) \cdot g(\gamma)\mathrm{d}\gamma. \end{aligned}$$

For $\gamma < \phi/(1-\delta)$, the benefit associated with a targeted churner is negative. Since the maximal profit, $P(T(\gamma))$, for this range of γ is equal to zero, H_{var} can be expressed as:

$$H_{var} = \frac{1}{CLV}\int_0^{+\infty} P^{ccp}(T(\gamma); CLV, \delta, \phi) \cdot g(\gamma)\mathrm{d}\gamma. \tag{3.18}$$

Equation 3.18 shows that H_{var} essentially is an EMP^{ccp} measure scaled by CLV, for which the distribution $h(\gamma)$ is replaced with the weight function $g(\gamma)$, and where $\gamma \in \mathbb{R}^+$. It is to be noted that the function $g(\gamma)$ is not a probability density in the strict sense, since generally, the area under g will not be equal to one. However,

by simply multiplying H_{var} with a normalization constant, this is resolved. Let the proper density function $\tilde{h}(\gamma)$ be the normalized function $g(\gamma)$, then H_{var} equals:

$$H_{var} = \frac{\int_0^{+\infty} g(\gamma)\mathrm{d}\gamma}{CLV} \cdot \int_0^{+\infty} P^{ccp}(T(\gamma); CLV, \delta, \phi) \cdot \tilde{h}(\gamma)\mathrm{d}\gamma. \quad (3.19)$$

Because the H measure takes into account all points of the convex hull, this corresponds to integrating over $\gamma \in [0; +\infty)$.

As illustrated by Equation 3.19, the H measure will rank classifiers according to their expected maximum profit, implicitly assuming the distribution $\tilde{h}(\gamma)$, which is specified by the parameters α and β. The previously proposed distribution for γ, $h(\gamma)$, is specified by the parameters α' and β'. It is not possible to analytically find values for α and β so that $\tilde{h}(\gamma)$ equals $h(\gamma)$. Therefore, a distance measure between the two distributions is numerically minimized. Several distance measures have been experimented with, but the *Hellinger distance* leads to the best results. The Hellinger distance between two probability measures, D_h, is defined as the L_2 distance between the square root of their densities (Van der Vaart, 2000). Hence, the optimal parameters for the approximating H measure can be found by solving the following minimization problem:

$$\min_{\forall \alpha,\beta} D_h(h(\gamma), \tilde{h}(\gamma)) = \min_{\forall \alpha,\beta} \int_0^{+\infty} \left(\sqrt{h(\gamma)} - \sqrt{\tilde{h}(\gamma)} \right)^2 \mathrm{d}\gamma. \quad (3.20)$$

Note that $h(\gamma)$ is parameterized by α' and β', which are considered to be known, whereas $\tilde{h}(\gamma)$ is parameterized by α and β, which are to be optimized.

In order to find the optimal parameters for the approximating H measure, the minimization problem in Equation 3.20 has been numerically solved, yielding $\alpha \approx 49$ and $\beta \approx 10$ as best values. This density function is plotted as the dashed line in Figure 3.3, and has an expected value of 28.9% and a standard deviation of 10.4%. Although the density theoretically is non zero for $\gamma > 1$, the accumulated probability for this region is smaller than 10^{-3}. In Section 3.6, a case study will be carried out and it will be investigated how well the

H measure approximates the EMP^{ccp} measure in terms of ranking several classification techniques.

3.6 Case Study

This section presents a case study to illustrate the use of the EMP^{ccp} and H measure for measuring classification performance in a customer churn prediction context. A selection of 21 classification techniques has been tested on 10 different customer churn data sets. The data sets have been preprocessed to prepare them for the analysis. Table 3.1 summarizes the main characteristics of the data sets. The classification techniques used in the benchmarking study are listed in Table 3.2 and include many commonly applied data mining methods. A more extensive discussion regarding the data sets, the preprocessing and the employed techniques can be found in the original benchmarking study, carried out by Verbeke et al. (2012).

Each data set has been split into a training set and a test set, where the former was used to train the classifiers, and the latter to perform an out-of-sample test. Thus, each instance in the test set is assigned a score s, on which classification is based by setting a cutoff value. Through the empirical score distributions, the AUC, H measure, MP^{ccp} measure and EMP^{ccp} measure are then calculated. Per data set, each measure leads to a ranking of the 21 techniques. Since the ranking is the most important output of the classification performance measurement process, the rankings from the different measures will be compared to one another. As the 10 data sets involved in this study are independent, the 10 resulting rankings of techniques are considered independent observations.

3.6.1 EMP versus other performance measures

Thus, each performance measure results in a ranking for every data set. For each single data set, Kendall's τ_b is used as a measure of dependence between the rankings based on different performance

ID	Source	Region	# Obs.	# Att.	%Churn	Reference
O1	Operator	North America	47,761	53	3.69	Mozer et al. (2000); Verbeke et al. (2012); Verbraken et al. (2013e)
O2	Operator	East Asia	11,317	21	1.56	Hur and Kim (2008); Verbeke et al. (2012)
O3	Operator	East Asia	2,904	15	3.20	Hur and Kim (2008); Verbeke et al. (2012)
O5	Operator	East Asia	2,180	15	3.21	Hur and Kim (2008); Verbeke et al. (2012)
O6	Operator	Europe	338,874	727	1.80	Verbeke et al. (2012)
D1	Duke	North America	93,893	197	1.78	Lemmens and Croux (2006); Lima et al. (2009); Neslin et al. (2006); Verbeke et al. (2012); Verbraken et al. (2013e)
D2	Duke	North America	38,924	77	1.99	Verbeke et al. (2012); Verbraken et al. (2013e)
D3	Duke	North America	7,788	19	3.30	Verbeke et al. (2012); Verbraken et al. (2013e)
UCI	UCI	-	5,000	23	14.14	Lima et al. (2009); Verbeke et al. (2012); Verbraken et al. (2013e)
KDD	KDD Cup 2009	Europe	46,933	242	6.98	Verbeke et al. (2012)

Table 3.1: Summary of data set characteristics: ID, source, region, number of observations, number of attributes, percentage churners, and references to previous studies using the data set.

Overview of Classification Techniques	
Decision Tree Approaches	**SVM Based Techniques**
C4.5 Decision Tree (C4.5)	SVM with linear kernel (linSVM)
Classification and Regression Tree (CART)	SVM with radial basis function kernel (rbfSVM)
Alternating Decision Tree (ADT)	LSSVM with linear kernel (linLSSVM)
	LSSVM with radial basis function (rbfLSSVM)
Rule Induction Techniques	Voted Perceptron (VP)
RIPPER (RIPPER)	
PART (PART)	**Ensemble Methods**
	Random Forests (RF)
Nearest Neighbors	Logistic Model Tree (LMT)
k-Nearest neighbors $k = 10$ (KNN10)	Bagging (Bag)
k-Nearest neighbors $k = 100$ (KNN100)	Boosting (Boost)

Table 3.2: O
verview of classification techniques]Overview of the classification techniques. More information regarding these data mining techniques can be found in the original benchmarking study Verbeke et al. (2012) or in a standard data mining handbook (e.g. Tan et al. (2006); Witten and Frank (2000)).

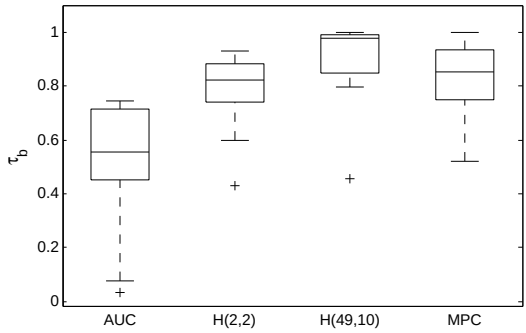


Figure 3.4: Box plot indicating the distribution of correlations between EMP^{ccp} and $H_{2,2}$, $H_{49,10}$, and MP^{ccp} respectively.

measures, yielding ten values for each correlation. The box plot in Figure 3.4 shows the distribution of the values of the correlation between EMP^{ccp} and the four displayed performance metrics. It is clear that the agreement between AUC and EMP^{ccp} based rankings is lowest, and it has the largest variability. The H measure with default parameters ($\alpha = 2$ and $\beta = 2$) shows a higher level of correlation, but still significant variability. The H measure with optimized parameter values however, shows very high correlation and lowest variability, indicating that both rankings agree to a large extent and that this is a reasonable approximation to the EMP^{ccp} measure. Finally, the correlation between MP^{ccp} and EMP^{ccp} is again lower, which can be attributed to the fact that the underlying cost and benefit assumptions are different, i.e. it is a deterministic versus a probabilistic approach. This also suggests that when the operator has accurate estimates of the response rate γ, it is preferable to use MP^{ccp} as criterion. When there is more uncertainty involved with the estimation of γ, a probabilistic approach, and thus EMP^{ccp}, is recommended. Furthermore, the box plot shows an outlier for the correlation between EMP^{ccp} and AUC, $\text{H}_{2,2}$, and $\text{H}_{49,10}$. These outliers correspond to the data set D2, where the expected maximum profit is zero for all techniques. As a result, EMP^{ccp} and MP^{ccp} consider all techniques to be unprofitable, and they all receive the same rank. AUC, $\text{H}_{2,2}$, and

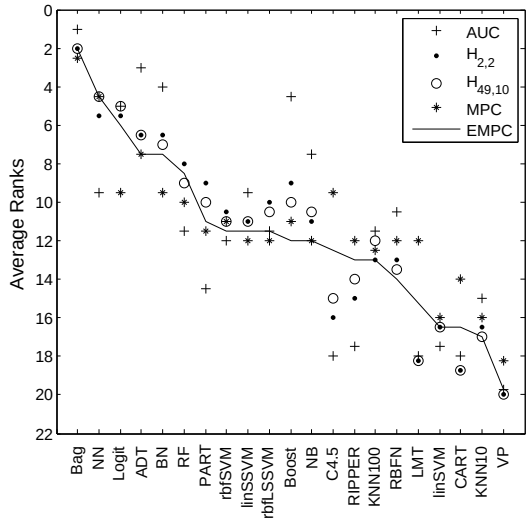


Figure 3.5: The average rank over ten data sets for each technique based on several performance measures (i.e. AUC, $H_{2,2}$, $H_{49,10}$, MP^{ccp}, EMP^{ccp}).

$H_{49,10}$ on the other hand will rank techniques differently. Therefore, the correlation is low for this data set.

A further indication of the agreement in ranking is given in Figure 3.5, which shows the rank of each technique, averaged over the 10 data sets. The full line represents the ranking according to the EMP^{ccp} measure, whereas the other points represent other performance metrics. Hence, a point plotted far from the full line, shows strong disagreement with the EMP^{ccp} based ranking. One can immediately see that the disagreement of AUC with EMP^{ccp} is larger than any other performance measure. The H measure with optimized parameters follows the EMP^{ccp} ranking much closer, again indicating that it is a reasonable approximation.

A second point of view is the profitability of a retention campaign. As explained before, EMP^{ccp} ranks classifiers according to the expected profit based on assumptions about the classification costs and benefits. AUC, as shown by Hand (2009), is also a measure of profit, but with invalid assumptions for the distribution of costs and

	Selection based on EMP^{ccp}			Selection based on AUC			
Data Set	Technique	EMP^{ccp} (€)	η_{emp}(%)	Technique	AUC	EMP^{ccp} (€)	Difference in EMP^{ccp} (€)
O1	ADT	0.129	2.10	Logit	0.747	0.120	0.009
O2	NN	0.051	0.09	Bag	0.644	0.017	0.034
O3	Boost	0.646	3.77	Bag	0.903	0.521	0.125
O5	Bag	1.652	3.44	Bag	0.964	1.652	0
O6	Bag	0.047	0.32	Bag	0.902	0.047	0
D1	KNN10	0.001	0.03	Boost	0.599	0	0.001
D2	PART	0	0	Bag	0.618	0	0
D3	BN	0.351	2.83	NB	0.809	0.214	0.137
UCI	RF	5.734	13.65	Bag	0.915	5.640	0.093
KDD	ADT	0.379	7.82	Bag	0.713	0.375	0.004

Table 3.3: Model selection based on EMP^{ccp} and on AUC, with the expected profit for each method. For model selection with EMP^{ccp}, also the optimal fraction of customers to target is given.

benefits. Therefore, choosing a classifier based on AUC may lead to suboptimal model selection from a profit point of view, which is illustrated in Table 3.3. For each data set, the optimal technique is selected based on EMP^{ccp} and on AUC, and the expected profit for that particular choice of classifier is given. As indicated in the last column, selection based on AUC leads in some cases to suboptimal model selection, with losses up to € 0.137 per customer, a substantial amount for telecom operators with millions of customers. Moreover, for selection based on EMP^{ccp}, it is possible to calculate the fraction of the vast customer base which needs to be targeted to realize the maximal profit, which is also displayed in Table 3.3. This is one of the major advantages of the EMP^{ccp} (and also the MP^{ccp}) measure, as it gives guidance to practitioners about how many customers to include in a retention campaign. When selecting a model with AUC or the H measure, there is no such guidance, and deviating from the optimal fraction may again lead to suboptimal profits.

3.6.2 Sensitivity analysis

Section 3.3.3 discusses the impact of variations in CLV, δ, and ϕ on the estimated expected profit and analytically derives first order approximations for this sensitivity. This yields some straightforward rules of thumb for the sensitivity, such as e.g. the higher the targeted fraction, the more sensitive the profit to changes in ϕ. However, the question arises how the ranking between classification algorithms is affected. Therefore, the results of the case study are used to analyze this impact. First, the techniques have been ranked with the proposed values for CLV, δ, and ϕ. Then, each parameter has been multiplied with a constant (while holding the others equal), and the techniques have been ranked again with this new parameter value. The correlations between the ranking in the base scenario and the ranking in the new scenario have been plotted for varying values of the multiplier, ranging from 1/2 to 2. Figure 3.6 shows these results for all three fixed parameters. The left axis and full lines indicate the correlation between the ranking of the base scenario and the ranking

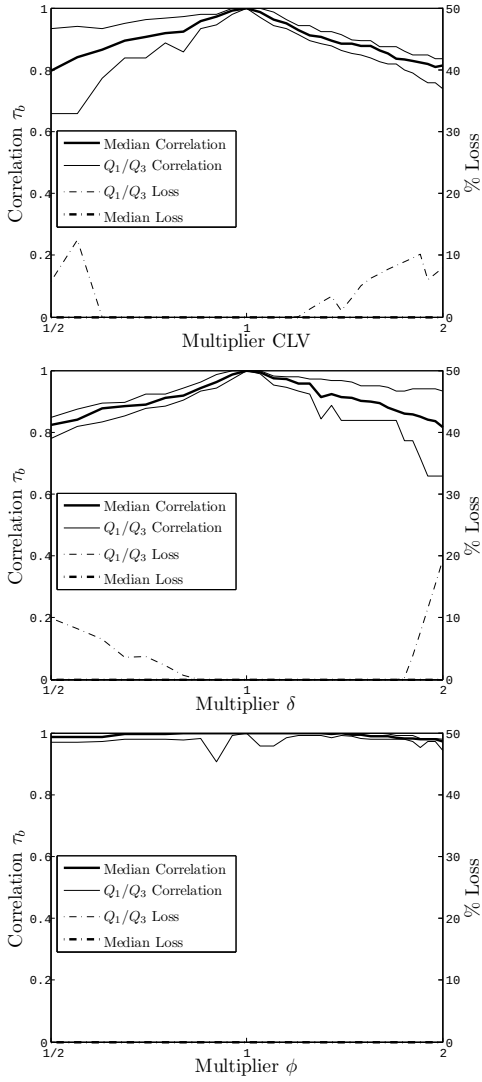


Figure 3.6: Sensitivity of the EMP^{ccp} measure to variations in CLV, δ, and ϕ respectively. The x-axis is plotted on a logarithmic scale with base 2, with the multiplier ranging from 1/2 to 2.

with the changed parameter. The median and the first and third quartile (over the ten data sets) have been plotted. Note that the plot for CLV assumes that CLV is changed while d and f are held constant (not δ and ϕ). It can be seen that variations in both CLV and δ have a similar impact, with median correlations decreasing until 0.8. The impact of ϕ is virtually non-existent. Furthermore, when the ranking changes, also the best performing technique (with the highest expected profit) may change. Therefore, Figure 3.6 also displays the impact of suboptimal classifier selection due to incorrect parameter values on the profitability (the right axis and dashed lines show the percentage loss due to incorrect classifier selection). Again, the impact is most significant for CLV and δ, whereas variations in ϕ do not impact the profitability. But even though there is an impact, it is relatively limited to losses of maximal 20%, and this for substantial variations in the parameters (doubling or halving the CLV or δ). These results, both in terms of correlation between rankings and percentage loss due to incorrect classifier selection, indicate that the EMP^{ccp} measure and corresponding rankings are robust to changes in the fixed parameters. The impact only becomes noticeable for multipliers smaller than 0.75 or larger than 1.5.

3.7 Conclusion

Chapter 2 proposed a theoretical framework which incorporates all gains and losses related to the employment of a data mining technique, and defined a probabilistic performance measure, the expected maximum profit (EMP). However, to apply the EMP measure, it needs to be defined for a specific business context. In this chapter, the EMP measure is implemented for application in a customer churn prediction context, which leads to EMP^{ccp}. Also the link between EMP^{ccp} and the H measure was investigated and it appears that the latter with appropriately chosen distribution parameters is a good approximation of the former.

The EMP measure for customer churn is validated in an extensive

case study. The results clearly indicate that the use of AUC as a performance metric leads to suboptimal profits. The case study also points to one of the major advantages of the EMP^{ccp} measure. It does not only select the classifier which maximizes the profit, it also provides the practitioner with an estimate of the fraction of the customer base which needs to be targeted in the retention campaign. This optimal fraction varies from case to case, and deviating from this fraction again leads to suboptimal profits. Note that the H measure, although it is able to select the most profitable classifier, does not provide guidance on the optimal fraction of the customer base to be included in the retention campaign. Finally, a sensitivity analysis was carried out, to analyze how vulnerable the EMP^{ccp} measure is to incorrect estimation of the fixed parameters CLV, δ, and ϕ. The outcome shows that the EMP^{ccp} measure and its resulting ranking is relatively robust with regard to changes in these parameter values.

Chapter 4

EMP for consumer credit scoring

Whereas the previous chapter turned its attention to the implementation of the EMP framework in a customer churn context, this chapter will implement the EMP framework for consumer credit scoring. The application within a credit scoring context will be illustrated with a real life case study, which shows the advantages of the EMP measure compared to traditional performance metrics.

This chapter is organized as follows. Section 4.1 gives a brief introduction to credit scoring, after which a careful analysis of the involved classification costs and benefits is presented in Section 4.2. Section 4.3 will then show in detail how the general EMP measure can be implemented for a credit scoring context, and an empirical estimation procedure is discussed in Section 4.4. Finally, Section 4.5 will report the results of a case study, illustrating the usefulness of the EMP framework in a credit scoring context.

4.1 Introduction

Credit scoring has become a very important application in statistical modeling, and concerns distinguishing *good* from *bad* loan applicants (Thomas, 2000). The main goal is to estimate the probability of default, i.e. the event of a customer not paying back the loan in

a given time period. For this task, a predictive model is developed which assigns a score to each loan applicant. Such a predictive model is then put to practice, by defining a cutoff value. Each applicant with a score lower than a certain cutoff will be rejected, others will be granted a loan. As pointed out by Hand (2005), it does not matter by how much a score exceeds the cutoff, since the classification and its related action will be the same.

During the past decades, a myriad of classification techniques has been used for credit scoring (Baesens et al., 2003). Hence, performance measurement of classification methods is essential for model selection, i.e. to identify the most suited classification technique as well as to tune the respective parameters (Ali and Smith, 2006). Traditional performance measures such as the Gini coefficient, the KS statistic, and the AUC measure are demonstrated to be inappropriate, since they may lead to incorrect conclusions (Hand, 2005, 2009). The reason for this is that the mentioned statistically motivated measures do not always properly take into account the business reality of credit scoring. Thus a guideline to select the most appropriate classification model as well as to calculate an adequate cutoff value is still missing if it comes to apply credit scoring in a profit-oriented setting, which has already been advocated by e.g. Thomas (2009) and Finlay (2010). This chapter defines an approach which tackles both requirements simultaneously. That is, we propose a classification performance measure, based on the EMP measure, taking into account the business reality of credit scoring and defining the optimal cutoff value (from a profitability perspective).

4.2 Cost and benefits in credit scoring

To apply the general EMP framework, outlined in Chapter 2, the cost and benefit parameters (b_0, c_1, c^*) have to be determined for the credit scoring context. We will use the cost framework developed by Bravo et al. (2012) to calculate each of the parameters needed for the presented method. Note that all costs and benefits are compared

to the base scenario where no credit scoring model is built, and thus all loan applications are accepted.

The first parameter to be defined is b_0, the benefit of correctly identifying a defaulter, which is equal to the loss incurred would the loan have been granted:

$$b_0 = \frac{\text{LGD} \cdot \text{EAD}}{A} = \lambda, \tag{4.1}$$

with λ for notational convenience. A is the principal, LGD is the loss given default, and EAD is the exposure at default (Mays and Nuetzel, 2004). Thus, λ is the fraction of the loan amount which is lost after default, and ranges from zero to one.

The second parameter to be determined is c_1, the cost of incorrectly classifying a good applicant as a defaulter, and is equal to:

$$c_1 = \text{ROI}, \tag{4.2}$$

with ROI being the return on investment of the loan. For any loan, the borrower-requested maturity M (as the number of periods) and the principal amount A can be used to estimate the return, considering the interest rate r typically offered at that term and principal level. Under those assumptions, the ROI of the loan can be estimated using the total interest formulas (Broverman, 2010):

$$ROI = \frac{pM - A}{A} = \frac{rM}{1 - (1 + r)^{-M}} - 1, \tag{4.3}$$

with p being the periodic payment:

$$p = \frac{Ar}{1 - (1 + r)^{-M}}. \tag{4.4}$$

Finally, there is c^*, which is the cost of the action. Since rejecting a customer does not lead to an extra cost, we assume $c^* = 0$. Note that there surely is a cost involved with building the credit scoring model, but this cost is not related to a particular individual (it is not a variable cost) and is similar among models. Therefore, in the long run, it is marginal for large portfolios – as is usually the case in

consumer credit scoring (Edelberg, 2006) – and can be omitted. The average classification profit for credit scoring can then be defined.

Definition 4.1. ***The average classification profit for credit scoring****, P^{cs}, is the interpretation of Definition 2.1 specifically for credit scoring:*

$$P^{cs}(t; \lambda, ROI) = \lambda \cdot \pi_0 F_0(t) - ROI \cdot \pi_1 F_1(t). \tag{4.5}$$

The definition of P^{cs} enables us to specify the MP and EMP measure for a credit scoring context, which is discussed in the next section.

4.3 A profit-based measure for CS

With the definition of the costs and benefits in a credit scoring context, the profit-based performance measures can be implemented for a credit scoring context. Definition 4.1 allows us to define the MP measure for credit scoring, by optimizing P^{cs}.

Definition 4.2. ***The maximum profit measure for credit scoring, MP^{cs}****, is the interpretation of Definition 2.2 in a credit scoring setting:*

$$MP^{cs} = \max_{\forall t} P^{cs}(t; \lambda, ROI). \tag{4.6}$$

Along with the MP measure, the optimal fraction of rejected loan applications can be determined as follows:

$$\bar{\eta}_{mp}^{cs} = \pi_0 F_0(T) + \pi_1 F_1(T), \tag{4.7}$$

with:

$$T = \arg\max_{\forall t} P^{cs}(t; \lambda, \mathrm{ROI}). \tag{4.8}$$

The cutoff $t = T$, for which the maximum profit is achieved, depends on the cost benefit ratio for credit scoring:

$$\theta^{cs} = \frac{c_1 + c^*}{b_0 - c^*} = \frac{\mathrm{ROI}}{\lambda}. \tag{4.9}$$

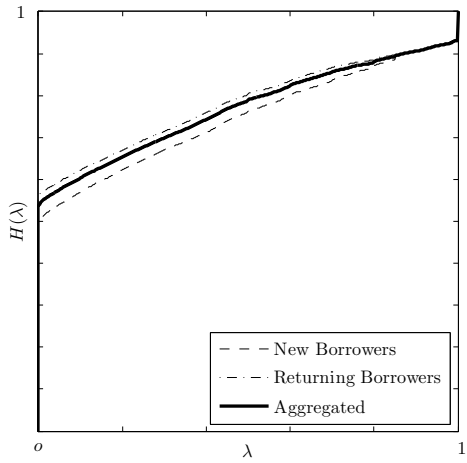


Figure 4.1: Empirical cumulative distribution of λ.

For the MP measure, it is assumed that the cost benefit parameters are precisely known, which in practice is not the case. Hence, the EMP measure adopts a probability distribution for the cost and benefit parameters, $h(b_0, c_1, c^*)$. A first assumption is that the return on investment (c_1) is practically the same for all loans, as is usually the case in consumer credit scoring. From equation (4.3) it can be seen that at fixed terms the ROI depends on the variation of the interest rate. A study by Edelberg (2006) noticed that the interest rates varied between three to five percent per annum on average in over a decade, so the ROI can be considered constant in portfolios with similar terms.

The parameter λ (b_0), however, is much more uncertain, since recovery rates may vary between zero and hundred percent of the total loan amount, and several distributions may arise (Somers and Whittaker, 2007). Figure 4.1 shows the empirical cumulative distribution of λ for the two data sets used in this study. This shows that a large part of the probability mass is situated in $\lambda = 0$, i.e. complete recovery of the loan amount. Another, smaller probability is observed for $\lambda = 1$, and the remainder of the probability is spread

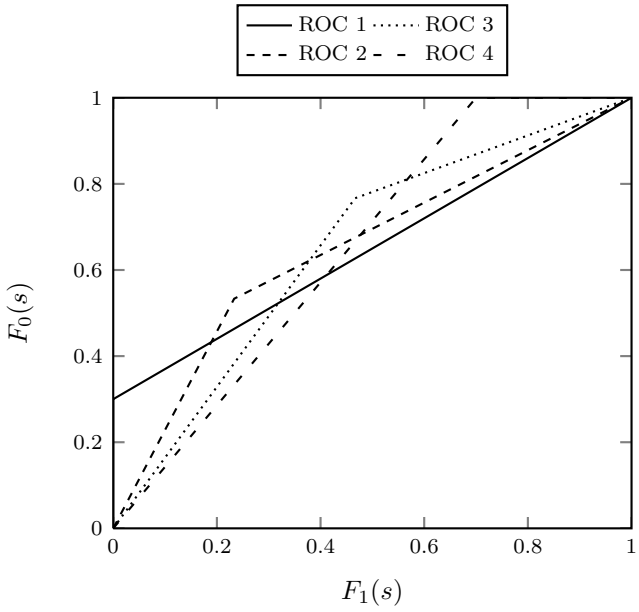


Figure 4.2: Illustration of four ROC curves with the same AUC but different EMP.

out roughly evenly between zero and one. Thus, to calculate EMP, for each defaulter it is assumed that:

- $\lambda = 0$ with probability p_0, implying that the customer eventually pays back the entire loan amount,
- $\lambda = 1$ with probability p_1, i.e. the customer defaults on the entire loan,
- λ follows a uniform distribution in $(0; 1)$, with the probability density equal to $\alpha = 1 - p_0 - p_1$,

with p_0 and p_1 parameters specifying $h(\lambda)$, thereby providing flexibility to adjust it to the specific situation in a certain company. With these elements defined above, the EMP measure for credit scoring can be defined.

Definition 4.3. ***The expected maximum profit measure for credit scoring, EMP^{cs}****, is the interpretation of Definition 2.3 in a*

ROC Curve	AUC	EMP	Fraction rejected
1	0.65	1.65%	2.7%
2	0.65	1.11%	8.45%
3	0.65	0.89%	11.83%
4	0.65	0.76%	14.51%

Table 4.1: Comparison of AUC and EMP for four synthetic ROC curves.

credit scoring setting:

$$EMP^{cs} = \int_{\gamma} P^{cs}(T(\lambda); \lambda, ROI) \cdot h(\lambda) \mathrm{d}\lambda, \tag{4.10}$$

with T the optimal cutoff for a given λ, and $h(\lambda)$ the probability density function for λ.

Note that θ^{cs}, the cost-benefit ratio, ranges from ROI (for $\lambda = 1$) to $+\infty$ (for $\lambda = 0$). This means that the EMP integration does not cover the entire ROC curve, since the slope of the ROC curve varies from $+\infty$ (in the origin) to 0 (in $(1, 1)$). So, even though both AUC and EMP are based on the same ROC curve, they lead to different results. This is illustrated by the four ROC curves shown in Figure 4.2. All four curves have exactly the same AUC, but have different EMP, as displayed in Table 4.1 (for this calculation it is assumed that $\pi_0 = 0.20$, ROI $= 0.2644$, $p_0 = 0.55$, and $p_1 = 0.10$).

Furthermore, the EMP-based fraction of rejected loans can be calculated:

$$\bar{\eta}^{cs}_{\text{emp}} = \int_{\lambda} \left[\pi_0 F_0\left(T(\lambda)\right) + \pi_1 F_1(T\left(\lambda\right))\right] \cdot h(\lambda) \mathrm{d}\lambda, \tag{4.11}$$

with $T(\lambda)$ being the optimal cutoff, as defined in Equation 4.8. In the next section, a procedure for the empirical estimation of EMP^{cs} will be discussed.

4.4 Empirical estimation of EMP

For the theoretical derivation, it is usually assumed that ROC curves are smooth. An empirical ROC curve, however, is a stepwise function

with diagonal elements if there are ties. Furthermore, Fawcett (2006) showed that the points on the convex hull of a ROC curve correspond to the set of optimal operational points. Hence, we follow an approach to calculate the EMP for credit scoring which is based on the convex hull of the ROC curve, analogous to the approach for EMP^{ccp}.

Assume the convex hull of the ROC curve consists of m segments, and let (r_{1i}, r_{0i}) be the end points of the segments with $(r_{00}, r_{10}) = (0, 0)$ and $(r_{0m}, r_{1m}) = (1, 1)$, as illustrated in Figure 4.3. A score, selected in the interval $[s_i, s_{i+1}]$, will be the optimal cutoff for the following value of λ (due to Equation 2.20):

$$\frac{r_{0(i+1)} - r_{0i}}{r_{1(i+1)} - r_{1i}} = \frac{\pi_1}{\pi_0}\theta^{cs} = \frac{\pi_1 \text{ROI}}{\pi_0 \lambda}. \tag{4.12}$$

The λ's are indexed corresponding to the upper end of the segment. Thus, λ_{i+1} is defined as:

$$\lambda_{i+1} = \frac{\pi_1 \left(r_{1(i+1)} - r_{1i}\right)}{\pi_0 \left(r_{0(i+1)} - r_{0i}\right)} \cdot \text{ROI}, \tag{4.13}$$

leading to m values for λ. Additionally we define $\lambda_0 = 0$.

Note that λ is not bounded by one along the ROC curve. When approaching the point $(1, 1)$, λ becomes infinitely large. Therefore, when calculating EMP, one replaces the series $\{\lambda_i | i = 0 \dots m\}$ with $\{\lambda_i | i = 0 \dots k+1\}$, with $k = \max\{i | \lambda_i < 1\}$, and $\lambda_{k+1} = 1$. Based on Equation 4.5 and 4.10, the EMP measure can then be estimated by:

$$\begin{aligned} \text{EMP} = &\left[\lambda_0 \cdot \pi_0 \cdot r_{00} \cdot p_0 - \text{ROI} \cdot \pi_1 \cdot r_{10} \cdot p_0\right] \\ &+ \sum_{i=0}^{k} \int_{\lambda_i}^{\lambda_{i+1}} \left[\lambda \cdot \pi_0 \cdot r_{0i} - \text{ROI} \cdot \pi_1 \cdot r_{1i}\right] h(\lambda) \mathrm{d}\lambda \\ &+ \left[\lambda_{k+1} \cdot \pi_0 \cdot r_{0(k+1)} \cdot p_1 - \text{ROI} \cdot \pi_1 \cdot r_{1(k+1)} \cdot p_1\right], \end{aligned} \tag{4.14}$$

The contributions in the square brackets correspond to the probability mass for $\lambda = 0$ or $\lambda = 1$. Since λ is constant over the segments, r_{0i} and r_{1i} are constant in the end points of the segments, and

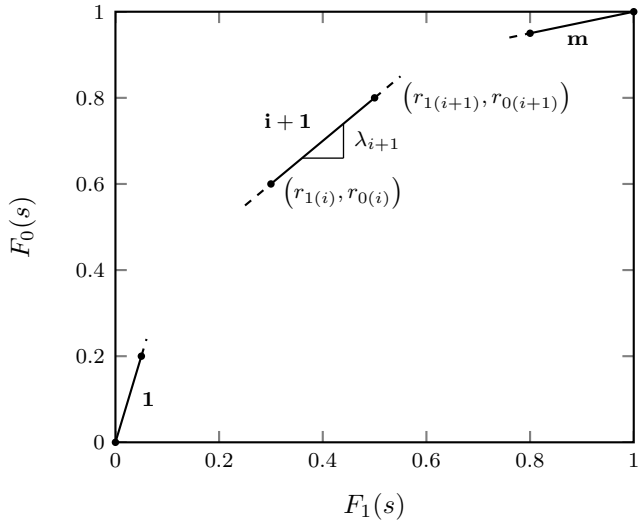


Figure 4.3: Segments of the convex hull of an empirical ROC curve.

$h(\lambda) = 1 - p_0 - p_1$, this can be written as:

$$\begin{aligned}\text{EMP} =& (1 - p_0 - p_1) \sum_{i=0}^{k} \left[\frac{\pi_0 r_{0i}}{2} (\lambda_{i+1}^2 - \lambda_i^2) - \text{ROI} \cdot \pi_1 r_{1i} (\lambda_{i+1} - \lambda_i) \right] \\ &+ \left[\pi_0 \cdot r_{0(k+1)} \cdot p_1 - \text{ROI} \cdot \pi_1 \cdot r_{1(k+1)} \cdot p_1 \right]. \end{aligned} \tag{4.15}$$

Note that the contribution corresponding to λ_0 vanishes – both r_{00} and r_{10} are equal to zero – and that $\lambda_{k+1} = 1$. Due to the upper bound for λ, which is equal to one, the integration does not cover the entire ROC curve by definition. Furthermore, in the above formulas, the class zero must represent the defaulters and the labels are not interchangeable. This is caused by the fact that the integration does not span the entire ROC curve.

4.5 Case study

In this section, we demonstrate the usefulness of the proposed profit-based performance measure for credit scoring with real data. The

EMP measure is an alternative for other existing performance measures. Given their widespread use, we focus on two commonly used performance measures for comparison: AUC and accuracy (ACC). Our experimental procedure compares the use of each of these metrics during the credit scoring development process, while focusing on two important aspects: parameter tuning and cutoff point determination.

First, Section 4.5.1 describes the data set, after which the experimental setup is discussed in Section 4.5.2. Section 4.5.3 and Section 4.5.4 respectively address the results with regards to parameter tuning and cutoff point determination.

4.5.1 Consumer credit data set

For our experiments, we use two data sets composed of loans for micro-entrepreneurs granted by a government organization during the years 1997-2007. The data set characteristics are summarized in Table 4.2 and described below (for more details see Bravo et al. (2012)):

- New Borrowers: The first data set consists of 37,492 loans granted to borrowers with no previous credit history with the organization. Each loan is described by 16 variables, such as socio-demographic descriptors (age, country region, employment, etc.) and an economic profiling (ownership of properties, goods relevant for the application of the loan, etc.). The mean loan value is 1,123 EUR, the mean term is 2.5 years, and the data set presents a default rate of 30.56%.

- Returning Borrowers: The second data set is formed by 103,466 loans granted to borrowers that already have had a loan with the institution, i.e. there was credit history available. The variables presented before are complemented by eight credit history variables, such as the total number of past and concurrent loans, maximum and mean arrears in previous loans, total amount granted, etc. The data set has an average loan value

Table 4.2: Data set characteristics.

	# Observations	# Variables	% Default
New Borrowers	37,492	17	30.56%
Returning Borrowers	103,466	19	20.47%

of 1,150 EUR, with a mean term of 2.4 years and an observed default rate of 20.47%.

Since this data set comes from a previously developed credit scoring model (Bravo et al., 2012), it is known that all predictive variables are significant. Also, besides the predictive variables, additional information was captured while the loan was being repaid[1]. The EAD and LGD of defaulted loans is used to estimate the perceived loss, and the granted amount is used to estimate the profit for each of the loans, as was described in Section 4.2.

4.5.2 Experimental setup

We chose two different models for our experiments: Artificial Neural Networks (ANN) using logistic output transfer functions, with a single hidden layer; and logistic regression. The reasons for these models is that logistic regression is the most commonly used method in credit scoring applications, accounting for more than 90% of all models according to Thomas et al. (2002). ANN, however, were proven to be the most efficient model in a large number of data sets (Baesens et al., 2003). The problem with ANN is that their black-box nature goes against Basel II regulations requiring transparency in the loan granting process, so much more transparent methods are needed (Allen et al., 2004; Martens et al., 2007). Nevertheless, we use ANN as a benchmark to obtain best-case results.

[1]In particular information necessary for estimating the exposure and the loss, that is, repayments made after default, total number of payments, collateral value and recovery percentage at default

During the model building step of ANN, certain parameters, such as the hidden layer size and the number of iterations need to be tuned. Therefore, a grid search is conducted over a set of candidate parameters. In order to select the best set of parameters, a performance measure is needed. We will contrast AUC, accuracy, and EMP. Once the model is obtained, a decision has to be made regarding when any given loan applicant is considered a defaulter, and when it is considered a good customer. That decision is made by setting a cutoff point, which transforms the continuous score into a binary output. According to Bravo et al. (2012), there are two methods to take that decision when no EMP measure is used: (1) focusing on the cost of the operation, or (2) using the accuracy to define the optimal cutoff. The EMP measure, however, gives the fraction of cases that should be rejected, which can then be transformed to a cutoff point. This characteristic is unique among all methods compared, so for the benchmark we took two approaches: if a model was built using accuracy as performance measure, then accuracy is also used to determine the cutoff point (maximum accuracy in training set). For AUC, since there is no way to determine a cutoff from the measure itself, we reproduced the observed default rate, representing the "best bet" in absence of further information.

To assess the quality of the resulting credit scoring model, we compare three different measures: accuracy, total profit, and the average profit per accepted loan.

4.5.3 Parameter tuning

The first aspect analyzed is the parameter tuning of the ANN. Two parameters were tuned during this process: the number of iterations, and the number of neurons in the hidden layer. Each of the data sets was divided into three parts: a parameter tuning data set (20% of observations), used to vary the parameters, a training set (60% of observations) for training the model once the optimal parameters were found, and an independent test set (20% of observations) which is used for reporting results. The test set is the same across all

Table 4.3: Results of parameter tuning for New Borrowers.

Performance Measure	Iters.	Hidden Layer Size	Value of PM	Optimal Fraction
Accuracy	450	29	0.6772	N/A
AUC	150	32	0.6834	N/A
EMP	150	32	0.0301	17.56%

Table 4.4: Results of parameter tuning for Returning Borrowers.

Performance Measure	Iters.	Hidden Layer Size	Value of PM	Optimal Fraction
Accuracy	50	25	0.768	N/A
AUC	250	26	0.827	N/A
EMP	400	21	0.023	10.16%

experiments, so results are comparable throughout the chapter.

The grid in which the search was performed depended on the parameter and the data set. The number of neurons was chosen from the interval $[\frac{V}{2}, 2V]$, with V the number of input variables, while the number of iterations was chosen from the interval $[50, 1000]$ in steps of 50 iterations. A model was trained in each of the grid elements for the parameter set, and the best one for each measure (AUC, ACC, EMP) was then trained using the respective parameter set as input after which the results were measured using the test set. Table 4.3 and Table 4.4 show the results for each of the data sets.

For New Borrowers, both AUC and EMP select the same configuration: 150 iterations and 32 neurons, whereas accuracy expands the training time (450 iterations), but reduces the complexity of the network (29 hidden neurons). For the data set of returning borrowers, EMP selects 400 training iterations, but only 21 neurons, versus 250 iterations and 26 neurons for AUC.

The last two columns of Table 4.3 and Table 4.4 show the value of the respective performance measure and the optimal fraction. The performance in terms of AUC and accuracy is better for the returning borrowers, as one would expect given the richer data. This is not

true, however, for EMP, where performance decreases from 3% to 2.3%. This seems counter-intuitive considering the richer data for Returning Borrowers, but is explained by the fact that the absolute value of EMP cannot be compared over different data sets when the context (e.g. the default rate) is different. In this case, the data set of new borrowers contains more defaulters (30.56%) than the data set of returning borrowers (20.47%). Remember that EMP measures the *incremental* profit as compared to not building a credit scoring model, expressed as a percentage of the total loan amount. The more defaulters there are in a data set, the easier it is to increase the profitability by building a credit scoring model, even with less data available. This also means that it is worthwhile to reject more applicants for New Borrowers (17.56%) as compared to Returning Borrowers (10.16%). Note that AUC and accuracy do not provide information about the profitability, one of the major strengths of EMP besides the optimal fraction, which will be discussed in the next section.

4.5.4 Cutoff point determination

The next step, after having trained a model, is determining the cutoff point which allows the actual classification into two groups. The cutoff value has been determined for the different performance measures, after which the performance of the model has been assessed with the test set. There are four tables of results for this step: two for each model and two for each data set. Table 4.5 and Table 4.6 show the results for the ANN, for which the parameters determined in Section 4.5.3 were used.

From the tables, the advantages of using a profit-driven measure are evident. Considering the total profit, EMP brings the highest value among all combinations, with differences of up to 100,000 EUR as compared to the scenario where no model is used. EMP achieves an excellent trade-off between accuracy and profit, since it also reaches the best accuracy across all data sets, which is mostly driven by the automatic cutoff point determination. The number of granted loans is

highest for the EMP-based model, and at the same time the highest total profit is achieved. One could argue that the EMP-based model ensures better rewards across the distribution of loans, even though that means accepting some riskier loans that might end in a default.

The results for the logistic regression models are shown in Table 4.7 and Table 4.8. Since logistic regression models do not have parameters to be tuned, the variation in the results is entirely attributable to the cutoff value. Hence, these results provide insight in the importance of cutoff point determination. For the model for new borrowers, the difference between EMP-based and ACC-based cutoff value is very small in terms of accuracy, and it is non-existent in the case of returning borrowers. However, there are again differences in the total profit and the average profit per loan. This once again illustrates that the EMP-based cutoff determination is a better overall choice, resulting in the best accuracy and a significant, although lower, improvement in the monetary gain. The results are also consistent regarding the average profit per loan: EMP leads to the lowest one among the three models, and in this case AUC is the one with highest average profit. The reason is that the AUC model is much more restrictive, since we reproduce an already high default rate, with a cutoff of 0.55 for the first data set and 0.67 for the second one, so it takes a much more conservative approach than the other two measures.

As shown in Bravo et al. (2012), a cutoff purely based on accuracy is too lenient to be used on its own, mostly because there usually is a much higher number of good borrowers than bad borrowers in a data set. On the other hand, a cutoff based solely on average cost is too restrictive, since this implies rejecting too many loans as each loan represents a risk. The use of a profit-oriented performance measure such as EMP has the advantage of achieving an acceptable trade-off between both criteria, when just one cutoff is to be determined.

Table 4.5: Cutoff selection for each measure, ANN, New Borrowers.

Model	Cutoff	ACC [%]	Total Profit [EUR]	Profit/Loan [EUR]	Granted Loans
No Model	N/A	69.48	671,712.21	17.92	37,492
ACC-based	0.80	70.32	718,304.08	104.22	6,892
AUC-based	0.56	68.01	679,303.83	131.60	5,162
EMP-based	0.67	70.48	764,680.73	124.84	6,125

Table 4.6: Cutoff selection for each measure, ANN, Returning Borrowers.

Model	Cutoff	ACC [%]	Total Profit [EUR]	Profit/Loan [EUR]	Granted Loans
No Model	N/A	79.83	3,375,666	32.63	103,466
ACC-based	0.80	83.63	3,751,123	209.71	17,887
AUC-based	0.68	82.38	3,602,599	220.44	16,343
EMP-based	0.84	83.74	3,781,266	204.81	18,462

Table 4.7: Cutoff selection for each measure, Logistic Regression, New Borrowers.

Model	Cutoff	ACC [%]	Total Profit [EUR]	Profit/Loan [EUR]	Granted Loans
No Model	N/A	69.48	671,712	17.92	37,492
ACC-based	0.60	69.77	691,468	117.62	5,879
AUC-based	0.55	66.53	603,187.41	116.85	5,162
EMP-based	0.61	69.81	691,485	115.02	6,012

Table 4.8: Cutoff selection for each measure, Logistic Regression, Returning Borrowers.

Model	Cutoff	ACC [%]	Total Profit [EUR]	Profit/Loan [EUR]	Granted Loans
No Model	N/A	79.83	3,375,666	32.63	103,466
ACC-based	0.80	83.20	3,648,778	201.00	18,153
AUC-based	0.67	81.65	3,472,887	212.50	16,343
EMP-based	0.82	83.20	3,687,437	199.81	18,455

4.6 Conclusion

This chapter presents a profit-based performance measure, based on EMP, a recently proposed general classification performance measure. This general approach has been adapted for the specific case of consumer credit scoring. This performance measure accounts for the benefits generated by healthy loans and the costs caused by loan defaults (driven by the loss given default and the exposure at default). As a result, the profit-based measure allows for profit-driven model selection, i.e. it allows practitioners to pick the credit scoring model which increases profitability most. Furthermore, the proposed technique provides the practitioner with the optimal cutoff value, which is required in an operational context where the continuous score of a credit scoring model has to be transformed into a binary decision. This is a feature which other performance measures do not provide, and is a major advantage of the EMP measure.

The results of our experiments indicate that using the EMP measure for model selection leads to more profitable credit scoring models. Moreover, employing the EMP-based cutoff value for rejecting loan applicants further increases the profitability by granting more loans than traditional approaches. Besides, the lender gains insight in the monetary reward of implementing a credit scoring model, which improves the practical use of the model.

Chapter 5

Toward profit-driven model building

In the previous chapters, a framework for profit-based classification performance measurement, the EMP framework, has been proposed. The EMP framework has been applied to customer churn prediction and to credit scoring in order to illustrate its usefulness in a business context. Hereby, the emphasis lied on measuring the performance of traditional classification techniques. In this chapter, however, we will explore the advantages of incorporating the profitability aspect into the model building step itself.

The remainder of this chapter is organized as follows. First we give a brief introduction, after which we outline the decision process and the use of predictive modeling in churn management. We then develop three hypotheses on profit-based modeling and explain our methodology to test these. Subsequently, we report empirical results and then conclude the chapter with a summary and an outlook to future research.

5.1 Introduction

This chapter focuses on resource allocation in a marketing context, and more specifically on selecting a target group for a retention management program. Customer retention is a key driver of firm per-

formance, especially in highly competitive, saturated markets where the acquisition of new customers is difficult and associated with high cost (e.g., Gupta and Zeithaml (2006)). Churn management represents a proactive approach to retain customers. It involves identifying likely churners based on demographic and behavioral customer data, and targeting these customers with a retention campaign (e.g., Neslin et al. (2006)). Hereby, the business objective is to increase firm performance through sustaining profitable customers. Therefore, in order to match managers' requirements and to have a high task-technology fit, churn models should focus on maximizing the profitability of the churn management campaign.

We argue that today's systems are not well aligned with actual business objectives. Whereas some progress on business-oriented assessment criteria has been made (Glady et al., 2009; Verbeke et al., 2012), the model building is carried out in a purely statistical manner, agnostic of business objectives and requirements. Overcoming the partial mismatch between decision makers' needs and the deliverables of predictive models is therefore the overall objective of this study. Predictive models, also known as predictive analytics (PA), consist of two components: (1) data-driven models designed to predict future observations and (2) methods to assess the predictive power of such models (Shmueli and Koppius, 2011). When developing a PA solution for churn management, it is important to recognize that both components contribute to or impede the decision support system (DSS) effectiveness. The specific contribution of this study is to shift attention to the first PA component, the prediction model itself. The first, and maybe most important, step in building predictive models is defining the modeling goal (Shmueli and Koppius, 2011). We argue that, in our application context, this goal should be increasing the profitability of the retention campaign, rather than some statistical measure of model fit. We develop and test three hypotheses related to profit-based model building and the insufficiency of using profit-based performance measures for model assessment only.

5.2 Marketing analytics for customer churn management

A key decision task in churn management is to define the target group of a retention program. A typical campaign planning process consists of two elements:

1. Predicting the likelihood of churn on a customer level and ranking the customers accordingly,
2. Setting a cutoff value and hereby deciding who is included in the campaign.

These two steps then result in a customer churn management campaign, for which the costs and benefits are schematically illustrated in Figure 3.1. Consider that a company with N customers is facing an outflow of π_0 percent of its customer base each period. The average remaining customer lifetime value is CLV. In order to limit losses due to customer churn, the company sets up a churn management campaign, whereby a number of customers, i.e. the target group, will receive a retention offer. The true would-be churners in the target group accept this offer with probability γ, whereas the non-churners certainly accept the offer. Clearly, targeting a non-churner is a waste of marketing budget. The profitability of the campaign is usually assessed through incremental profits, relatively to the base scenario where no campaign is undertaken. In particular, the incremental profit of a churn management campaign is (Neslin et al., 2006):

$$\text{Profit} = N \cdot \eta[(\gamma CLV + d(1-\gamma))\pi_0\lambda - d - f] - A, \tag{5.1}$$

with A the fixed administrative costs, d the cost of the retention offer, and f the cost of contacting a customer. Also important is the probability that a churner accepts the retention offer, γ, the size of the target group (as a fraction of N), η, and the lift coefficient, λ, which is the percentage of churners in the target group divided by the base churn rate, π_0. For example, a lift coefficient equal to 3

means that a model-based targeting identifies 3 times more churners than soliciting customers at random.

PA plays an important role in customer churn management. Demographic and transactional customer data is easily available in company databases and facilitates anticipating future customer behavior through data-driven prediction models (e.g., Padmanabhan and Tuzhilin (2003)). The first step in the campaign planning process, i.e. ranking the customers in terms of their model-estimated probability to churn, is often accomplished by classification methods. Managers can choose from a large set of alternative classifiers (e.g., Hastie et al. (2001)). Given the large variety of available methods, performance assessment and model selection are important issues. Several performance indicators have been proposed (e.g., Baldi et al. (2000)). However, these measures focus on statistical criteria and embody a notion of accuracy that is often less important in a churn context. Therefore, the prevailing performance indicator in the industry is the lift-coefficient (e.g., Neslin et al. (2006)).

However, a disadvantage of the lift coefficient is that it assumes a fixed campaign size. Fixing the size of the target group a priori is not ideal because it leaves out information provided by the churn model. Overall campaign profit often increases when targeting only customers with high model-estimated churn probability (Verbeke et al., 2012). To maximize profits, it may thus be necessary to reduce the size of the target group compared to using the entire budget per default. In order to allow more flexibility with regards to the target group size, Verbeke et al. (2012) propose the maximum profit measure for customer churn (MP^{ccp}) which is tailored to PA in a churn management context. MP^{ccp} directly connects the classification performance of a model with the profitability of the retention campaign. In particular, MP^{ccp} calculates the maximum profit for each model by optimizing the target group size. Besides selecting the most profitable model, MP^{ccp} also provides the marketing analyst with the optimal size of the target group. This is an important improvement compared to the lift coefficient as it further closes the gap between the business objective

and PA-based decision support in churn management. Instead of assuming a fixed marketing budget for a retention campaign, the manager may now adjust the marketing budget in order to optimize the profit of the campaign.

Previous research has mainly focused on incorporating business objectives (such as campaign profit) into the performance assessment component of PA. This ignores the other component, the predictive model itself. Nearly all classifiers described in the literature are general purpose algorithms, well grounded in statistics but unable to account for business objectives or task-specific requirements. This mismatch between the model and the business goal is one reason why we assert that the standard approach toward model-based decision support in churn management is not well aligned with decision makers' needs. We argue that the profitability goal should be taken into account from the very beginning, i.e. in the model building step and not only when selecting a churn model for deployment. In the next section, we develop three hypotheses to test this assertion and to identify further ways to close the gap between the DSS and the manager's goal.

5.3 Research questions

Our research questions focus on both PA components, model assessment and model building. The role of model assessment pertains to identifying a suitable prediction model (e.g., out of a set of alternative models) for deployment. In this sense, profit-oriented assessment criteria can contribute toward the business performance of a churn management DSS if they identify the most appropriate – in business terms – model more accurately than a traditional approach. The main concern from a managerial perspective is whether a churn model emerging from a profit-based model selection truly increases the profitability of a churn management campaign. We therefore test:

***H1**: MP^{ccp}-based model assessment leads to selecting*

more profitable churn models than model assessment based on the lift coefficient.

Profit-oriented model selection is a first step to bridge the gap between managers' needs and PA-based decision support. However, during model construction the individual churn models – from which the best model is selected – optimize some performance criterion which typically is not based on profitability. For example, logistic regression uses maximum likelihood estimation, whereas decision trees use information theoretic criteria (e.g., Hastie et al. (2001)). Our main proposition is that introducing profit-orientation into the model building process further increases models business performance. To confirm this, we test:

***H2**: MP^{ccp}-based model building leads to more profitable churn models than constructing conventional churn models based on conventional statistical assessment criteria.*

Our second hypothesis refers to current churn modeling practices. That is, the practice to predict churn with classification methods. Confirming *H2* would expose the weakness of this approach. To support the effectiveness of a business-oriented model building paradigm and to gain further insight into the relationship between model building and model assessment criteria, our last hypothesis states that:

***H3**: MP^{ccp}-based model building leads to more profitable churn models than model building using the lift coefficient.*

The lift coefficient has up till now only been used for model assessment in marketing. Recall that it is more aligned with the actual decision problem in campaign planning than the statistical accuracy indicators that conventional classifiers embody. However, it does not accurately reflect the true profitability of a retention campaign (Verbeke et al., 2012). The importance of *H3* is thus to show that using business-oriented performance measures for model building leads to decision outcomes which are more closely aligned with actual business requirements.

5.4 Methodology

To test our hypotheses, we first construct a pool of churn models using conventional classification methods. In particular, we choose the methods that have been studied in a recent churn prediction benchmark (Verbeke et al., 2012). This selection spans a wide range of alternative classifiers, from established techniques such as logistic regression to sophisticated boosting machines that represent the state-of-the-art in churn prediction (e.g., Lemmens and Croux (2006)). All techniques provide some parameters that allow tuning the method to a particular prediction task. Using recommendations from the literature (e.g., Caruana et al. (2004)), we define candidate values for these parameters and create models for every parameter setting. This allows us to create a large library of 693 churn models. A brief description of the employed prediction methods and their parameter settings is available online.

We then use MP^{ccp} and the lift coefficient to select one best performing model from the library. To confirm *H1*, we examine the significance of differences in profitability among the two choices. Examining *H2* and *H3* requires a prediction framework that facilitates profit-based model building. To that end, we use the ensemble selection (ES) framework of (Caruana et al., 2004, 2006). The term ensemble refers to a composite-model that integrates several individual prediction models (e.g., Shmueli and Koppius (2011)). ES is a two-stage modeling philosophy. In the first stage a library of candidate models is created. In the second stage, ES iteratively selects candidate models for the ensemble so as to increase the prediction performance of the ensemble. The models included in the ensemble are combined by computing a simple average over their predictions. ES continues to add models to the ensemble as long as this increases the performance of the ensemble prediction (Caruana et al., 2004).

It is important to note that the selection process in the second ES stage does not depend on a particular performance measure. All that matters is whether the inclusion of one additional model improves

the current ensemble. The analyst is free to use any measure s/he deems most appropriate for a given prediction task. ES will gear the construction of the ensemble model towards this measure and thereby account for the task-specific notion of model performance that the chosen measure embodies. This makes ES an appropriate choice for examining the effectiveness of profit-based churn modeling. In particular, to test our second hypothesis, we create an ES-based ensemble model that maximizes the EMP^{ccp} measure during ensemble selection. We then compare this ensemble model to the best – in terms of EMP^{ccp} – churn model from our library. Finally, we test *H3* through a comparison of two ES-based ensemble models. The first is just the same one used to test *H2*. The second model uses the lift-coefficient instead of EMP^{ccp} during member selection and thus gears model building toward achieving high lift.

5.5 Experimental design

We examine our research questions by means of an empirical study in the telecommunication industry. Churn prediction plays an important role in this industry. Due to the drastic increase in mobile phone users, telecommunication markets are getting saturated and the emphasis is shifting from attracting new customers to retaining the existing customer base. Moreover, marketing analytics is most likely to have a positive performance impact when a firm has a substantial amount of data at its disposal, when the industry is characterized by fierce competition, and when the customer preferences change rapidly (Germann et al., 2013). This makes the telecommunication industry a suitable environment to explore the efficacy of predictive marketing analytics. In particular, we use a sample of eight churn data sets from U.S. and European Telco operators. All data sets have been used in previous studies (e.g., Mozer et al. (2000); Verbeke et al. (2012)) and past churn modeling tournaments (e.g., Guyon et al. (2010); Neslin et al. (2006)). Different demographic and behavioral variables associated with, e.g., call detail records, subscription plans,

or billing information are available to describe the customers and to model the binary event variable churn=yes/no. In particular, a churn event involves a customer canceling or not renewing a subscription in a predefined time window. Summary statistics of the data sets are provided in Table 5.1.

To examine the performance of different churn models, we randomly partition all data sets into a training sample (60%) and an out-of-sample test set (40%). This way, all models predict an *unseen* set of data which has not been used during model building. The out-of-sample validation mimics an application of the model in corporate practice and thus facilitates a realistic and robust assessment (e.g., Collopy et al. (1994)). In addition, we require a second set of validation data to test some of our hypotheses. For example, the ES approach employs validation data to select models for the final ensemble. We obtain this validation data by means of cross-validating all churn models on the training sample (Caruana et al., 2006).

5.6 Empirical results

We summarize the results of our experiments in Table 5.1, where we label our data sets D1 – D8. Recall that we built 693 churn models per data set. As an initial exploration, we first assess all models in terms of MP^{ccp} and the lift coefficient and then examine the rank correlation (i.e., Kendall's τ) between the two performance measures (third column of Table 5.1). In several cases, we observe values between 0.4 and 0.6, or above 0.6 indicating moderate or strong positive association, respectively. However, it is important to remember the difference between MP^{ccp} and the lift coefficient. The latter assumes a fixed campaign size whereas MP^{ccp} determines the optimal fraction of customers to target. With this in mind, it is noteworthy that the correlation between the two measures can be as low as 0.12 or 0.17. Thus, we find an indication that a traditional and a profit-based assessment criterion can differ in their ranking of model performance.

	Cases x variables	Consistency	Profit from retention campaign (in EUR) Selection with Lift	Selection with MP^{ccp}	ES maximizing MP^{ccp}	ES maximizing Lift
D1	40,000 x 70	0.17	22.66	22.63	22.65	22.62
D2	93,893 x 196	0.12	22.59	22.58	22.6	22.58
D3	12,410 x 18	0.47	16.54	16.61	16.68	16.66
D4	69,309 x 67	0.6	8.89	8.92	8.94	8.87
D5	21,143 x 384	0.86	0.99	1.02	1.26	1.17
D6	50,000 x 301	0.38	0.05	0.07	0.07	0.06
D7	47,761 x 41	0.59	0.16	0.16	0.18	0.19
D8	5,000 x 18	0.68	6.14	6.19	6.37	6.15

Table 5.1: Empirical results of alternative traditional and profit-based churn models.

Although assessment consistency is a relevant statistical property, only a single churn model is selected to guide campaign targeting. An important managerial concern is thus to ensure that the single, eventually deployed model truly gives the largest profit (i.e., *H1*). To test this, we contrast the profitability (on test data) of the two models that perform best (on validation data) in terms of MP^{ccp} and the lift coefficient, respectively. Specifically, we use MP^{ccp} to determine an optimal target fraction and compute an average per customer profit for an accordingly sized campaign. To that end, we use formula (1) with parameter values as in (Neslin et al., 2006). The resulting profits are shown in columns four and five of Table 5.1. A Wilcoxon signed rank test, the recommended approach for such a comparison (Demšar, 2006), suggests that the null-hypothesis of equal profitability cannot be rejected (p-value: 0.11). In other words, our results do not provide sufficient evidence to conclude that selecting churn models in terms of MP^{ccp} increases eventual campaign profits. This is an intriguing finding. Previous work on profit-based churn modeling has concentrated on developing better assessment criteria (Glady et al., 2009; Verbeke et al., 2012). The implicit assumption of such an approach is that a profit-based model assessment leads to the selection of models with higher business performance. This is the essence of our first hypothesis. Given that our results cast doubt on

H1, the link between model selection and model performance may be weaker than previously assumed. This puts the practice to only focus on the assessment component of PA into perspective and stresses the importance of also considering the model building step.

To test our second hypothesis, we compare two different modeling approaches: (1) selecting the individual churn model with highest MP^{ccp} (on validation data) from the library for deployment, and (2) using ES to create an ensemble model that is designed to maximize MP^{ccp}. Compared to the former approach, ES never performs worse and gives higher per customer profits in seven out of eight cases (column five and six in Table 5.1). Therefore, we accept *H2* (p-value: 0.0156) and conclude that profit-oriented model building facilitates significantly higher profits than simply selecting some standard model in a profit-oriented manner. This suggests that, among the two PA components, model building and model assessment, the former is indeed the more important lever to improve the profitability of churn models. Finally, we note that other factors besides using MP^{ccp} for selecting ensemble members might have influenced the performance of our profit-based modeling approach. To further support the previous conclusion, we create another set of ES-based ensemble models, but now using the lift coefficient instead of MP^{ccp} for member selection (i.e., *H3*). The last column of Table 5.1 reports the profitability of the corresponding churn models. We find that MP^{ccp} is indeed the more appropriate selection criterion. It produces higher profits in seven cases, which represents a significant improvement (p-value: 0.0234). The only difference between the modeling philosophies reported in the two rightmost columns of Table 5.1 is the performance criterion used for selecting ensemble members. Our results evidence that this part of the model building should mimic the business objective as accurately as possible. If decision makers care about profit, predictive decision support models should be built in a profit-driven fashion.

5.7 Conclusion

We have observed that PA-based decision support is often not well aligned with managers' requirements. The suboptimal situation this may lead to has been illustrated in the context of churn management, where the main business concern is to maximize the profit of retention programs. We developed and tested three hypotheses which analyze the weaknesses of current PA practices in churn modeling and how these weaknesses can be overcome. The main implication of our study is that prediction models should be built in awareness of the decision task they are meant to support, and thus should focus on the profitability of a retention campaign in the case of customer churn prediction. In particular, we illustrate that profit-based model building may benefit campaign profits. Another implication is that focusing on business-oriented model assessment alone, a trend we observe in the literature, overlooks the more important part of PA. Eventually, both components should be geared toward business needs. This is an important step to bridge the gap between the manager and the model level and thus to improve the effectiveness of PA-based decision support solutions.

Although this study provides some evidence in favor of profit-based modeling, it has several limitations that need to be addressed in future research. First, the profit model used in this work includes customer profitability in the form of an average remaining CLV. As a consequence, all churners are treated equally and accurately predicting churn becomes the main determinant of profit. Considering profitability on a per-customer-level would help to give better advice with regard to which customers should be contacted in a retention campaign. A closely related issue pertains to the likelihood of actual churners accepting a retention offer. Based on previous research (Neslin et al., 2006), we assumed a fixed acceptance rate. Recently, some progress has been made to remedy this simplification. The expected maximum profit criterion integrates over all possible acceptance rates and is thus a more robust profit indicator (Verbraken

et al., 2013d). Extending our results to this measure is another task for future research. Third, a key challenge in profit-based modeling is to actually implement this paradigm as a prediction model. The ES framework used here is a valuable approach. However, it employs MP^{ccp} only in the second modeling stage. This way, the profit-based modeling is carried out on the basis of the library models, which, in turn, result from standard (statistical) prediction methods. A more direct integration of MP^{ccp} or other business-oriented measures into a prediction method would be useful to augment our findings and test the efficacy of profit-based modeling with higher fidelity. We are currently exploring methods from the realms of heuristic search to construct purely profit-driven churn models.

Part II

Case studies in data mining

Chapter 6

Customer churn prediction with Bayesian network classifiers

Abstract

Customer churn prediction is becoming an increasingly important business analytics problem for telecom operators. In order to increase the efficiency of customer retention campaigns, churn prediction models need to be accurate as well as compact and interpretable. Although a myriad of techniques for churn prediction has been examined, there has been little attention for the use of Bayesian Network classifiers. This chapter investigates the predictive power of a number of Bayesian Network algorithms, ranging from the Naive Bayes classifier to General Bayesian Network classifiers. Furthermore, a feature selection method based on the concept of the Markov Blanket, which is genuinely related to Bayesian Networks, is tested. The performance of the classifiers is evaluated with both the Area under the Receiver Operating Characteristic Curve and the recently introduced Maximum Profit criterion. The Maximum Profit criterion performs an intelligent optimization by targeting this fraction of the customer base which would maximize the profit generated by a retention campaign. The results of the experiments are rigorously tested

and indicate that most of the analyzed techniques have a comparable performance. Some methods, however, are more preferred since they lead to compact networks, which enhances the interpretability and comprehensibility of the churn prediction models.

6.1 Introduction

The number of mobile phone users has increased tremendously during the last decade. At the end of 2010, there will be more than five billion mobile phone users[1], which is over 70% of the world population. As a result, telecommunication markets are getting saturated, particularly in developed countries, and mobile phone penetration rates are stagnating. Therefore, operators are shifting their focus from attracting new customers to retaining the existing customer base. Moreover, the literature reports that customer retention is profitable because: (1) acquiring a new client is five to six times more costly than retaining an existing customer (Athanassopoulos, 2000; Bhattacharya, 1998; Colgate and Danaher, 2000; Rasmusson, 1999); (2) long-term customers generate higher profits, tend to be less sensitive to competitive marketing activities, become less costly to serve, and may provide new referrals through positive word-of-mouth, while dissatisfied customers might spread negative word-of mouth (Colgate et al., 1996; Ganesh et al., 2000; Mizerski, 1982; Reichheld, 1996; Stum and Thiry, 1991; Zeithaml et al., 1996; Paulin et al., 1998); (3) losing customers leads to opportunity costs because of reduced sales (Rust and Zahorik, 1993). A small improvement in customer retention can therefore lead to a significant increase in profit (Van den Poel and Larivière, 2004). For successful targeted marketing campaigns, it is crucial that operators are able to identify clients with a high probability to churn in the near future. Next to correctly classifying the future churners, it is important to gain insight in the reasons for customers to be classified as churners. Therefore, telecom operators prefer compact and interpretable models, as it allows them

[1] http://www.eito.com

to check whether the model is in line with current domain knowledge. Moreover, it enables operators to recognize potential warning signs for customer churn.

Although a myriad of techniques for churn prediction has been examined, there has been little attention to the use of Bayesian Network (BN) classifiers. This chapter will investigate the predictive power of a number of Bayesian Network algorithms, ranging from the Naive Bayes classifier, with very strict independence assumptions, to General Bayesian Network classifiers, which allow more flexibility. The performance of the classifiers is evaluated with both the Area under the Receiver Operating Characteristic Curve and the Maximum Profit criterion and is rigorously statistically tested.

Like most real life data mining problems, also churn prediction involves a large number of attributes. Including irrelevant variables would result in complex and incomprehensible classification models, impeding the interpretability of these classifiers (Martens et al., 2011). Hence, feature selection is commonly applied to withhold only those variables with strong explanatory power. Many feature selection algorithms have been proposed in the literature. However, most of these methods perform the selection from a univariate perspective, i.e. they assess a measure of dependence between the attributes and the target variable for each attribute separately. In this study, a feature selection method based on the concept of the Markov Blanket, which is genuinely related to Bayesian Networks, will be analyzed. This method approaches the input selection problem from a multivariate point of view and relies on the concept of *conditional* independence. Especially in the context of some of the Bayesian Network classifiers, this form of feature selection proves to be useful, as will be shown in Section 6.5. The impact of this variable reduction on classification performance and network complexity is investigated, since, basically, one is looking for the most compact Bayesian network with the highest explanatory power.

For measuring classifier performance and selecting the most appropriate classification method, a variety of performance measures

has been used (Ali and Smith, 2006; Martens et al., 2011). In this study, the well-known Area under the Receiver Operating Characteristic Curve (AUC) is employed, as well as the Maximum Profit (MP) criterion, recently proposed by Verbeke et al. (2012). Instead of measuring performance over the whole output range, as AUC does, the maximum profit criterion performs an intelligent optimization by targeting this fraction of the customer base which would maximize the profit generated by a retention campaign. As such, it is able to indicate the model which maximizes the effectiveness of a retention campaign. The rationale behind this performance measure is that the most important goal for a telecom provider is to optimize its profit.

The remainder of this chapter is organized as follows. In Section 6.2, the general problem of customer churn will be stated. Section 6.3 will give an overview of the main Bayesian Network classifiers and discuss the algorithms briefly. In Section 6.4, the experimental setup is described, the data set characteristics are discussed, and tests for statistical significance are clarified. Finally, the results of the experiments are described in Section 6.5.

6.2 Customer churn prediction

Many companies and organizations are confronted with customer churn. For instance, wireless telecom operators report annual churn rates up to 40% of their customer base (Neslin et al., 2006). Customer churn is associated with a direct loss of income and a diversity of supplementary costs, such as for instance the investments to acquire new customers to maintain the level of the customer base. Therefore, reducing customer churn by directing specifically designed marketing campaigns to the customers with the highest probability to attrite, has proven to be profitable to a company (Van den Poel and Larivière, 2004). To improve the efficiency of customer retention campaigns, a customer churn prediction model is needed to indicate the customers which are the most likely to churn and should be included in the retention campaign.

Customer churn prediction is a problem for which typically a data mining approach is adopted. Data mining entails the overall process of extracting knowledge from data. Based on historical data a model can be trained to classify customers as future churners or non-churners. Numerous classification techniques have been applied to predict churn, including traditional statistical methods such as logistic regression (Burez and Van den Poel, 2009; Lemmens and Croux, 2006; Neslin et al., 2006), non-parametric statistical models like for instance k-nearest neighbor (Datta et al., 2000), decision trees (Lima et al., 2009; Wei and Chiu, 2002), and neural networks (Au et al., 2003; Hung et al., 2006). An extensive literature review on customer churn prediction modeling can be found in Verbeke et al. (2011b).

The process of developing a customer churn prediction model consists of several steps. The first step in this process consists of gathering relevant data and selecting candidate explanatory variables. The resulting data set is then cleaned and preprocessed. In the second step a model is built. A modeling technique is selected based on the requirements of the model and the type of data. Input selection is often applied to reduce the number of variables in order to get a consistent, unbiased, and relevant set of explanatory variables. Depending on the number of observations, which can be small in case of new products, a model is trained by cross validation or by splitting the data set in a separate training and test set. The resulting model is then evaluated, typically by comparing the true values of the target variable with the predicted values, but also, if possible, by interpreting the selected variables and the modeled relation with the target variable. A variety of performance measures to evaluate a classification model have been proposed in the literature (Ali and Smith, 2006). As will be discussed in Section 6.4, in this study both the statistically based area under the receiver operating characteristic curve (Krzanowski and Hand, 2009) and the maximum profit criterion, recently proposed by Verbeke et al. (2012), will be applied. The latter estimates the profit a telecom operator would make when optimally exploiting the results of a particular classifier. Next, in a

third step the model is assessed by a business expert to check whether the model is intuitively correct and in line with business knowledge which requires the induced model to be interpretable (Martens et al., 2011). A prototype of the model is then developed, and deployed in the information and communication technology (ICT) architecture. The final step, once a model is implemented that performs satisfactory, consists of regularly reviewing the model in order to asses whether it still performs well. Surely in a highly technological and volatile environment as the telecommunications sector, a continuous evaluation on newly gathered data is of crucial importance.

Because of the third step, where the model is assessed by a business expert to confirm whether the model is intuitively correct, the comprehensibility aspect of customer churn prediction models is of crucial importance. However, comprehensibility of customer churn models thus far only received limited attention in the literature (Lima et al., 2009; Verbeke et al., 2011b). A specific family of classification techniques, which result in comprehensible models but have not been tested rigorously in a customer churn prediction setting before, are the Bayesian Network Classifiers. Two simple Bayesian Network Classifiers, i.e. Naive Bayes and standard Bayesian Networks, have been included in an extensive benchmarking study by Verbeke et al. (2012), which compares the performance of a variety of state-of-the-art classification techniques applied on an extensive number of data sets. Their results suggest that Bayesian Network Classifiers form a viable alternative modeling approach for customer churn prediction as they are able to produce compact and interpretable models. The next section will discuss a number of Bayesian Network Classification techniques into detail, which are applied to five real-life telecom churn data sets in Sections 6.4 and 6.5.

6.3 Bayesian network classifiers

6.3.1 Bayesian Networks

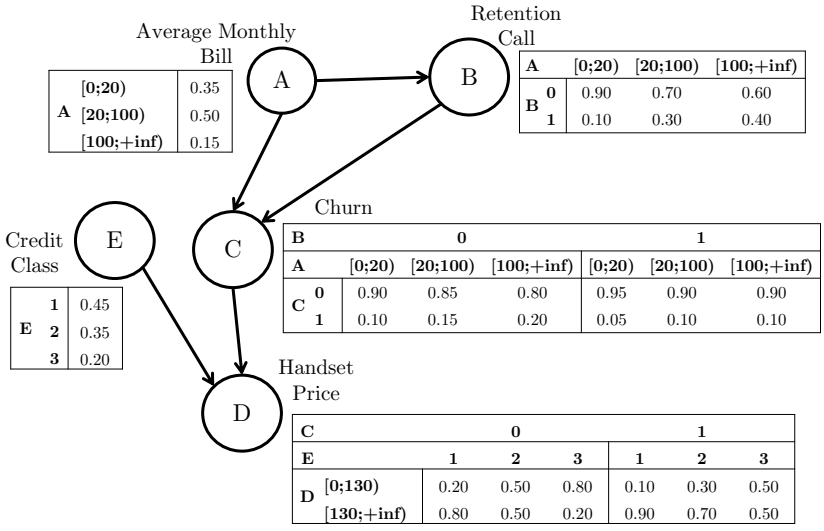
A Bayesian network (BN) represents a joint probability distribution over a set of stochastic variables, either discrete or continuous. It is to be considered as a probabilistic white-box model consisting of a qualitative part specifying the conditional (in)dependencies between the variables and a quantitative part specifying the conditional probabilities of the data set variables (Pearl, 1988). Formally, a Bayesian network consists of two parts $B = \langle G, \Theta \rangle$. The first part G is a directed acyclic graph (DAG) consisting of nodes and arcs. The nodes are the variables X_1 to 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:

$$P_B(X_1, ..., X_n) = \prod_{i=1}^{n} P_B(X_i|\Pi_{X_i}) = \prod_{i=1}^{n} \theta_{X_i|\Pi_{X_i}}. \tag{6.1}$$

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 study, we will use the empirical frequencies from the data D to estimate these parameters:

$$\theta_{x_i|\Pi_{x_i}} = \hat{P}_D(x_i|\Pi_{x_i}) \tag{6.2}$$

It can be shown that these estimates maximize the log likelihood of the network B given the data D. Note that these estimates might be further improved by a smoothing operation (Friedman et al., 1997).



A	
[0;20)	0.35
[20;100)	0.50
[100;+inf)	0.15

A		[0;20)	[20;100)	[100;+inf)
B	0	0.90	0.70	0.60
	1	0.10	0.30	0.40

E	
1	0.45
2	0.35
3	0.20

B		0			1		
A		[0;20)	[20;100)	[100;+inf)	[0;20)	[20;100)	[100;+inf)
C	0	0.90	0.85	0.80	0.95	0.90	0.90
	1	0.10	0.15	0.20	0.05	0.10	0.10

C		0			1		
E		1	2	3	1	2	3
D	[0;130)	0.20	0.50	0.80	0.10	0.30	0.50
	[130;+inf)	0.80	0.50	0.20	0.90	0.70	0.50

Figure 6.1: Classification with a Bayesian network.

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. The underlying assumption behind the winner-takes-all rule is that all gains and losses are equal. For a discussion of this aspect see, e.g., (Duda and Hart, 1973). A simple example of a Bayesian network classifier is given in Figure 6.1. Suppose that, for a particular customer, all variables except C are known and take the following values: $A \in [20; 100)$, $B = 0$, $D \in [0; 130)$ and $E = 1$. The probability that the customer will churn conditional on this information can be calculated as:

$$P\left(C|A, B, D, E\right) = \frac{P\left(C, A, B, D, E\right)}{P\left(A, B, D, E\right)}$$

Thus, reading from Figure 6.1 and using equation 6.1 yields:

$$P\left(C = 0, A \in [20; 100), B = 0, D \in [0; 130), E = 1\right) = 0.85 \cdot 0.5 \cdot 0.7 \cdot 0.2 \cdot 0.45 = 0.0268$$

$$P\left(C = 1, A \in [20; 100), B = 0, D \in [0; 130), E = 1\right) = 0.15 \cdot 0.5 \cdot 0.7 \cdot 0.1 \cdot 0.45 = 0.0024$$

Hence, the conditional probability for churning is:

$$P\left(C = 0 | A \in [20; 100), B = 0, D \in [0; 130), E = 1\right) = \frac{0.0268}{0.0268 + 0.0024} = 0.92$$

$$P\left(C = 1 | A \in [20; 100), B = 0, D \in [0; 130), E = 1\right) = \frac{0.0024}{0.0267 + 0.0024} = 0.08$$

According to the winner-takes-all rule, the customer will be classified as a non-churner. In what follows, several structure learning algorithms for the construction of Bayesian network classifiers will be discussed.

6.3.2 The Naive Bayes Classifier

A simple classifier, which in practice often performs surprisingly well, is the Naive Bayes classifier (Duda and Hart, 1973; John and Langley, 1995; Langley et al., 1992). 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. Assume a new test case $\mathbf{X} = (X_1, ..., X_n)$, which takes the value $\mathbf{x} = (x_1, ..., x_n)$, is presented to the classifier. The test case is then classified by using Bayes' rule to compute the posterior probability of each class c_l given the vector of observed variable values:

$$P(C = c_l | \mathbf{X} = \mathbf{x}) = \frac{P(C = c_l) P(\mathbf{X} = \mathbf{x} | C = c_l)}{P(\mathbf{X} = \mathbf{x})} \tag{6.3}$$

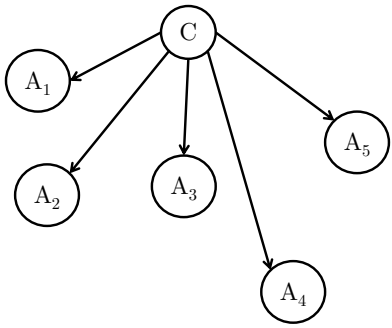


Figure 6.2: Naive Bayes Network

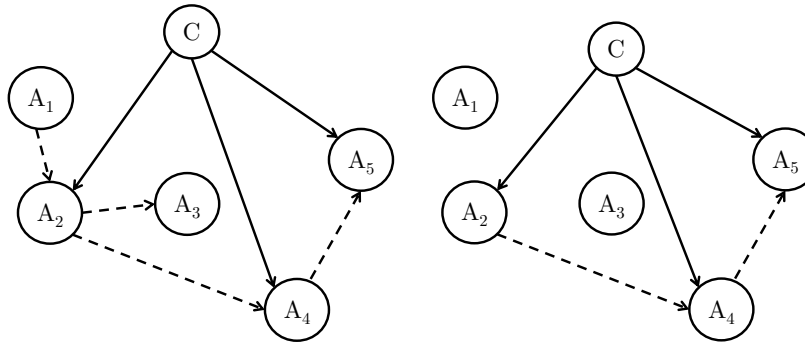
The simplifying assumption behind the Naive Bayes classifier then assumes that the variables are conditionally independent given the class label. Hence,

$$P(X_1 = x_1, ..., X_n = x_n | C = c_l) = \prod_{i=1}^{n} P(X_i = x_i | C = c_l). \quad (6.4)$$

This assumption simplifies the estimation of the class-conditional probabilities from the training data. Notice that one does not estimate the denominator in Equation 6.3 since it is independent of the class. Instead, one normalizes the nominator term $P(C = c_l)P(X_1 = x_1, ..., X_n = x_n | C = c_l)$ to 1 over all classes. Naive Bayes classifiers are easy to construct since the structure is given a priori 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 (John and Langley, 1995). Figure 6.2 provides a graphical representation of a Naive Bayes classifier.

6.3.3 Augmented Naive Bayes Classifiers

The strength of Naive Bayes classifiers inspired several authors to develop *Augmented* Naive Bayes network classifiers. These are methods based on the Naive Bayes classifier while partially relaxing the independence assumption. The *Selective Naive Bayes* classifier omits



(a) Network from SAN operator (b) Network from SAND operator

Figure 6.3: Examples of augmented Bayesian networks.

certain variables to deal with strong correlation among attributes Langley and Sage (1994), whereas the *Semi-Naive Bayesian* classifier clusters correlated attributes into pairwise disjount groups (Kononenko, 1991). Friedman et al. developed *Tree Augmented Naive Bayes (TAN) classifiers*, an algorithm where every attribute has one and only one additional parent next to the class node (Friedman et al., 1997).

In this study, the Augmented Naive Bayes classifiers developed by Sacha (1999b) are used. This is a family of classifiers where the constraints of the TAN approach are further relaxed: not all attributes need to be dependent on the class node and there does not necessarily need to be an undirected path between two attributes. The algorithms exist of a combination of five basic operators, summarized in Table 6.1. The measure of dependency between two attributes, $I(X,Y)$, is defined as follows:

$$\begin{cases} \sum p\,(x,y|c) \log\left(\frac{p(x,y|c)}{p(x|c)p(y|c)}\right) & \text{if X,Y dependent on } C \\ \sum p\,(x,y|c) \log\left(\frac{p(x,y|c)}{p(x|c)p(y)}\right) & \text{if } \textit{only} \text{ X dependent on } C \\ \sum p\,(x,y|c) \log\left(\frac{p(x,y|c)}{p(x)p(y|c)}\right) & \text{if } \textit{only} \text{ Y dependent on } C \\ \sum p\,(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) & \text{if X,Y independent from } C \end{cases} \tag{6.5}$$

Operator	Description
Class Dependency Operator	Connects the class node with *all* attributes. This can be used to construct a naive Bayes network.
Selective Augmented Naive Bayes (SAN)	Connects the class node with the attributes it depends on. A greedy search algorithm is used to seek for attributes dependent on the class node. Starting with an empty set, it adds in each step the attribute to the network that optimizes a quality measure. After adding an attribute, the dependencies among *all* other attributes, whether they are connected with the class node or not, are determined by one of the *augmenter* operators.
Selective Augmented Naive Bayes with Discarding (SAND)	Connects the class node with attributes it depends on, as the SAN operator does. The difference, however, is that the augmenter will only add connections between attributes dependent on the class node, while other attributes are discarded and are no part of the network in that particular iteration. The difference between a network resulting from the SAN operator and SAND operator is illustrated in Figures 6.3(a) and 6.3(b).
Tree-Augmenter	This operator builds the minimum spanning tree among a given set of attributes. The algorithm is based on a method developed by Chow and Liu Chow and Liu (1968), but differs in the way how the mutual information is calculated. Sacha uses the conditional or unconditional probability of X and Y depending on whether there is an arc between the class node and the attribute (see formula 6.5).
Forest-Augmenter	The forest augmenter is also used to create dependencies between the attributes, but allows for more flexibility. The resulting structure can be a forest consisting of a number of disjoint trees, meaning that there does not need to be a path between every attribute.

Table 6.1: Augmented Naive Bayes Approach: different operators.

Combining different class dependency operators with augmenting methods from Table 6.1 yields a number of algorithms, listed below

as per increasing complexity:

- Naive Bayes
- TAN: Tree Augmented Naive Bayes
- FAN: Forest Augmented Naive Bayes
- STAN: Selective Tree Augmented Naive Bayes
- STAND: Selective Tree Augmented Naive Bayes with Discarding
- SFAN: Selective Forest Augmented Naive Bayes
- SFAND: Selective Forest Augmented Naive Bayes with Discarding

The aim of these classifiers is to find a trade-off between the simplicity of the Naive Bayes classifiers (with a limited number of parameters) and the more realistic and complex case of full dependency between the attributes.

Except for Naive Bayes and TAN, all of the above procedures use a search method that requires a quality measure to assess the fitness of a network given the data. In this study, two quality measures proposed by Sacha will be used. The first is the Standard Bayesian (SB) measure, which is proportional to the posterior probability distribution $p(G, \Theta|D)$ and contains a penalty for the network size. This penalty term is a function of the dimension of the network, the latter being defined as the number of free parameters needed to fully specify the joint probability distribution. A derivation of the Standard Bayesian measure can be found in (Sacha, 1999b). The second quality measures is the Local Leave-One-Out Cross Validation. Let V_l be the training set D without instance x_l:

$$V_l = D \backslash \{x_l\} \tag{6.6}$$

The quality measure is then defined as:

$$LOO(G, D) = \sum_{l=1}^{m} \left(p \left(c^{(l)} | \mathbf{a}^{(l)}, V_l, G \right) \right) \tag{6.7}$$

with m instances in the data set and $p\left(c^{(l)}|\mathbf{a}^{(l)}, V_l, G\right)$ being the probability of c^l conditional on the values of the attributes for instance l, the data set V_l and the structure G. In this study, the latter two quality measures were combined with the five algorithms defined above, resulting in ten different classifiers.

It is worthwhile mentioning that the Naive Bayes classifier and the ANB classifiers without node discarding do not have the flexibility to remove variables from the network. Hence, even if a variable is completely independent from the target variable, it will still be used by the classifier. As this is clearly undesirable, a feature selection algorithm is carried out as part of the data preprocessing procedure. Although there are many input selection methods available, in this study the Markov Blanket feature selection will be used, since it tackles the problem from a multivariate perspective (see Section 6.4 for more information). Essentially, it applies the principle of conditional independence, which plays a pivotal role in the theory of Bayesian Networks.

6.3.4 General Bayesian Network Classifiers

All the previously discussed methods restrain the structure of the network in order to limit the complexity of the algorithms. Omitting those restrictions extends the search space of allowed networks to all *General Bayesian Networks (GBN)*. Finding the optimal network in such a solution space is known to be an NP-hard problem since the number of DAGs for n variables is superexponential in n (Chickering, 1996). As described by Kotsiantis (2007), there are two broad categories of structure learning algorithms. The first consists of heuristic methods, searching through the space of all possible DAGs and measuring the fitness of a DAG with a score metric, whereas

the second category comprises constraint based approaches using conditional independence (CI) tests to reveal the structure of the network.

Scoring based methods have been compared with CI-based algorithms by Heckerman et al. (1999), leading to the observation that CI based methods perform better on sparse networks. Search-and-score algorithms, on the other hand, work with a broader variety of probabilistic models even though their heuristic nature may inhibit finding the optimal structure.

Search-and-Score Algorithms

Several algorithms have been proposed in the literature, e.g. (Heckerman et al., 1999) and (Chickering, 2002) show that selecting a single DAG using greedy search often leads to accurate predictions. In this analysis, the well-known *K2* algorithm (Cooper and Herskovits, 1992) is applied. It seeks for the network with the highest posterior probability given a data set and makes the following four assumptions:

- all attributes are discrete,
- the instances occur independently, given one Bayesian network,
- there are no cases that have variables with missing values,
- the prior probability of the conditional probabilities in the conditional probability tables at each node is uniform.

Given these assumptions, a greedy search algorithm will find the network with the highest score, given the database D. For a detailed discussion, one may refer to (Cooper and Herskovits, 1992).

Constraint Based Algorithms

Constraint based algorithms are also known as Conditional Independence (CI) based methods. They do not use a score but employ

conditional independence tests to find the relations between attributes given a data set. In this chapter, the *Three Phase Dependency Analysis (TPDA)* algorithm (Cheng et al., 2002) is used, in which the concept of d-separation plays an essential role. 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 (Geiger et al., 1990; Verma and Pearl, 1988). It is precisely this property that will be exploited in the algorithm of Cheng to learn the Bayesian network structure. The algorithm itself 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 is described in detail by Cheng and Greiner (1999); Cheng et al. (1997).

Hybrid Methods

Also hybrid methods have been developed, combining characteristics of both search-and score and constraint based algorithms. Examples of such techniques are the *Sparse Candidate (SC)* algorithm (Friedman et al., 1999) and the *Max-Min Hill-Climbing learning algorithm (MMHC)* (Tsamardinos et al., 2006).

This latter method finds the parent-children set (**PC**) of each and every node, and thus determines the skeleton of the Bayesian network. In a first phase, also called the forward phase, nodes selected by a heuristic procedure sequentially enter a candidate PC set (**CPC**). The set may contain false positives, which are removed in phase II of the algorithm, i.e. the backward phase. The algorithm tests whether any variable in **CPC** is conditionally independent on the target variable, given a blocking set $\mathbf{S} \subseteq \mathbf{CPC}$. If such variables are found, they are removed from **CPC**. As a measure of conditional (in)dependence, the G^2 measure, as described by Spirtes et al. (2000), is used. This measure is asymptotically following a χ^2 distribution

with appropriate degrees of freedom, which allows to calculate a p-value indicating the probability of falsely rejecting the null hypothesis. Conditional independence is assumed when the p-value is less than the significance level α (0.05 and 0.01 in this study). Once the skeleton is determined, a greedy search method is used to direct the edges between the attributes. This is the second step and the search-and-score part of the algorithm, making it a hybrid method. The *BDeu* score (Heckerman et al., 1995) has been used for our analysis.

6.4 Experimental setup

The aim of the study is twofold. Firstly, the differences in terms of performance between the Bayesian Network classifiers, described in Section 6.3, are investigated. Obviously, the aim is to find the algorithm with the most discriminative power, and to determine whether the differences are significant. Secondly, the experiments need to reveal whether the Markov Blanket feature selection has a deteriorating impact on classification performance. Preferably, the input selection would reduce the number of attributes without affecting the predictive performance substantially, since it is supposed to remove variables which are independent from the target, conditional on the variables which are withheld. The remainder of this section will describe the experimental setup.

6.4.1 Data Sets and Preprocessing

Four real life and one synthetic data set will be used to evaluate the performance of the different classification techniques. Table 6.2 summarizes the most important aspects of these data sets. The first data set was obtained directly from a European telecommunication operator, the next three are available at the website of the Center for Customer Relationship Management at Duke University[2], and

[2] www.fuqua.duke.edu/centers/ccrm/datasets/download.html

ID	Source	# Obs.	# Attr.	%Churn
O1	Operator	47,761	28	3.69
D1	Duke	12,499	12	34.67
D2	Duke	99,986	172 (50)	49.55
D3	Duke	40,000	49	49.98
UCI	UCI	5,000	12	14.14

Table 6.2: Summary of data set characteristics.

the last one is a synthetic data set, available at the UCI repository[3].

The number of observations or instances in a data set is of great importance. The more observations, the more generally applicable the generated model will be. A large data set also allows to split the data set in a separate training and test set. The training set is used to induce a model, while the test set is only used to evaluate the proposed model. This ensures the performance measure is not biased due to overfitting the model to the test data. As can be seen from Table 6.2, the smallest data set has five thousand observations, and the largest up to almost hundred thousand observations. The data sets also differ substantially regarding the number of candidate explanatory variables, in a range from 12 up to 171 (for our analysis this is limited to 50 variables, see below). More attributes do not guarantee a better classification model however. The eventual performance of the model mainly depends on the explanatory power of the variables. Since a large number of attributes heavily increases the computational requirements, and most often only a limited number of variables is effectively valuable to explain the target variable, a feature selection method could be used to limit the available attributes. In the next subsection, a Markov Blanket based algorithm for feature selection is described. Another characteristic of the data set is the class distribution of the target variable, which is usually heavily skewed in a churn prediction setting. Three of the data sets approximately have an even class distribution, which is the result

[3] www.sgi.com/tech/mlc/db

of undersampling (i.e. removing non-churners from the data set). To test the classifiers, the test sets have been adjusted in order to resemble a realistic churn data set.

When working with real life data sets, it is essential to preprocess the data before starting the analysis because of two main reasons (Famili et al., 2010): (1) problems with the data (e.g. missing values, irrelevant data, etc.), and (2) preparation for data analysis (e.g. discretization of continuous variables, etc.). To ensure that the results are comparable, all data sets have been preprocessed in the same manner:

- The data set is split into a training set (66% of the instances) and a test set (34% of the instances).
- The missing values are replaced by the median or the mode for continuous or nominal variables respectively. Note that for none of the variables there were more than 5% missing values.
- Coarse variables (such as e.g. ZIP code) are clustered into smaller groups, in order to have a more meaningful contribution to the prediction of churn behavior.
- Since most Bayesian network algorithms only work with discrete variables, all continuous variables are discretized. This is done according to the algorithm of Fayyad and Irani (1993).
- A large number of attributes demands high computational requirements for the Markov Blanket feature selection algorithm. Therefore, if necessary, the data sets are treated with a χ^2-filter in order to withhold the 50 most predictive attributes as input for the further analysis[4].

[4]The choice for the χ^2 filter is based on a robustness check of several univariate procedures. The outcome suggests that most of the univariate methods lead to similar results. The aim of this step is to remove variables which show very limited dependence on the target (note that independence implies conditional independence, the reverse is not true). The number fifty is chosen to ensure that sufficient variables are withheld. If some variables are still redundant, those will

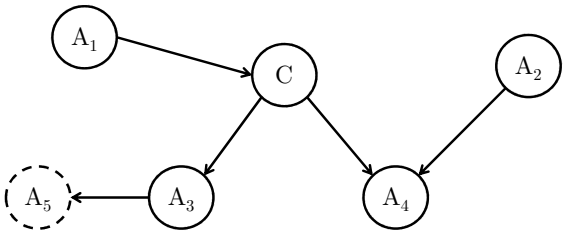


Figure 6.4: The Markov Blanket of a classification node.

Once these preprocessing steps are carried out, the data sets are ready for the Markov Blanket feature selection algorithm and the Bayesian Network classifiers, described in the following subsections.

6.4.2 Markov Blanket Feature Selection

Over the past decades, feature selection has become an essential part in predictive modeling. It aims to deal with three problems (Guyon and Elisseeff, 2003): (1) improving the accuracy of predictive algorithms, (2) developing faster and more cost-effective predictors, and (3) gaining insight in the underlying process that generated the data. Several techniques for input selection have been proposed, but many of these rely on univariate testing procedures, i.e. the variables are assessed one by one in terms of dependency on the target variable. For this study, the Markov Blanket (MB) feature selection algorithm of Aliferis et al. (2010a,b) has been applied. The procedure is based on the concept of the Markov blanket, which also plays a crucial role in Bayesian Networks. As opposed to commonly used univariate methods, this algorithm approaches the task from a multivariate perspective and exploits the concept of conditional independence. The Markov Blanket of a node X, is the union of X's parents, X's children and the parents of X's children. It can be shown that when the values of the variables in the Markov Blanket of the classifica-

be removed by the Markov Blanket feature selection, but the computational cost will have been reduced. Since the MB feature selection does still remove variables from data set D2 (from 50 to 44/42), the number fifty was not set too low.

ID	Source	# Attr. (Orig)	# Attr. (MB.05)	# Attr. (MB.01)
O1	Operator	28	21	12
D1	Duke	12	12	11
D2	Duke	172 (50)	44	42
D3	Duke	49	31	26
UCI	UCI	12	8	8

Table 6.3: Number of attributes before and after feature selection.

tion node are observed, the posterior probability distribution of the classification node is independent of all other variables (nodes) that are not in the Markov Blanket (Lauritzen, 1996). Hence, all variables outside the Markov Blanket can be safely discarded 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. E.g. in Figure 6.4, node A_1 will be part of the Markov Blanket because it is a parent of C, A_3 and A_4 because they are children and A_2 because it is the parent of one of the children of C. A_5 is not part of the Markov Blanket and can be ignored for classification purposes, since it will not influence C for a fixed value of A_3 (which is known in a classification context).

The Markov Blanket feature selection algorithm has been applied to the data sets at a significance level of 1% and 5%, and will be referred to as MB.01 and MB.05. The resulting number of variables are summarized in Table 6.3. Note that if attribute selection is performed, it is applied prior to training and testing the classifiers. The feature selection algorithm has been implemented in the *Causal Explorer* package for Matlab (Aliferis et al., 2003).

6.4.3 Bayesian Network Construction

For all the techniques discussed in Section 6.3, freely available software implementations have been used. The Naive Bayes classifier, TAN-classifier and the Augmented Naive Bayes classifiers have been

Algorithm	Implementation
Markov Blanket Feature Selection	Causal Explorer (Aliferis et al., 2003).
Logistic Regression	Weka Toolbox (Witten and Frank, 2000).
(Augmented) Naive Bayes Classifiers	Bayesian Network Classifier Toolbox (Sacha, 1999a).
K2 Algorithm	Weka Toolbox (Witten and Frank, 2000).
Three Phase Dependency Analysis	Powerpredictor (Cheng, 2000).
Max-Min Hill-Climbing	Causal Explorer (Aliferis et al., 2003) for structure learning and Bayesian Net Toolbox (Murphy, 2001) for inference.

Table 6.4: Implementations used in this study.

implemented by Sacha, and exist in the form of Weka bindings (Sacha, 1999a), allowing to run the software from within the Weka Workbench (Witten and Frank, 2000). In total, there are ten Augmented Naive Bayes classifiers, as a result of combining two quality measures (Local Leave One Out Cross Validation (LCV_LO) and Standard Bayesian measure (SB)) with five algorithms (itemized in Section 6.3). The K2 algorithm is directly available in the Weka Workbench. The constraint based algorithm, also called Three Phase Dependency Analysis, is available in the application *Powerpredictor*, developed by Cheng (2000). The Max-Min Hill-Climbing algorithm is available for Matlab (Aliferis et al., 2003). For inference in the networks generated by the MMHC algorithm, the Bayesian Net Toolbox for Matlab, developed by Murphy (2001), has been used. In Table 6.4, an overview of all software implementations is given. Note that Logistic Regression and the Naive Bayes Classifier are included in this study as benchmarks for the Bayesian Network classifiers.

6.4.4 Measuring Classifier Performance

A variety of performance measures has been used to gauge the strength of different classifiers and to select the appropriate model (Ali

and Smith, 2006; Martens et al., 2011). In this study two measures are reported: the *Area Under the Receiver Operating Characteristic Curve* (AUROC, or briefly AUC), and the *Maximum Profit* (MP) criterion, recently introduced by Verbeke et al. (2012).

We tested 16 algorithms on five data sets, in combination with MB.05, MB.01, or without preceding Markov Blanket feature selection. This results in 240 values for AUC and 240 values for MP. Section 6.4.5 will explain how the statistical significance of differences between methods is analyzed.

Area under the Receiver Operating Characteristic Curve

Most classification techniques result in a continuous output. For instance in a customer churn prediction setting, a probability estimate between zero and one of being a churner is produced by the model. Depending on a threshold probability value, a customer will be classified as a churner or a non-churner. The *receiver operating characteristic* (ROC) curve displays the fraction of the identified churners by the model on the Y-axis as a function of one minus the fraction of identified non-churners. These fractions are dependent on the threshold probability. ROC curves provide an indication of the correctness of the predictions of classification models. In order to compare ROC curves of different classifiers regardless of the threshold value and misclassification costs, one often calculates the *area under the receiver operating characteristic curve* (AUROC or AUC). Assume that a classifier produces a score $s = s(\mathbf{X})$, with $\mathbf{X}$ the vector of attributes. For a BN classifier, the score s is equal to the probability estimate $p(c = 1|\mathbf{X} = \mathbf{x})$. Let $f_l(s)$ be the probability density function of the scores s for the classes $l \in \{0, 1\}$, and $F_l(s)$ the corresponding cumulative distribution function. Then, it can be shown that AUC is defined as follows (Krzanowski and Hand, 2009):

$$AUC = \int_{-\infty}^{\infty} F_0(s) f_1(s) ds \tag{6.8}$$

An intuitive interpretation of the resulting value is that it provides an estimate of the probability that a randomly chosen instance of class one is correctly rated or ranked higher by the classifier than a randomly selected instance of class zero (i.e., the probability that a churner is assigned a higher probability to churn than a non-churner). Note that since a pure random classification model yields an AUC equal to 0.5, a good classifier should result in a value of the AUC much larger than 0.5.

Maximum Profit Criterion

A second performance measure that will be applied in this study is the MP criterion. In order to compare the performance of different classification models, this measure calculates the maximum profit that can be generated with a retention campaign using the output of a classification model. The profit generated by a retention campaign is a function of the discriminatory power of a classification model, and can be calculated as (Neslin et al., 2006):

$$\Pi = N\alpha \left\{ \left[\gamma CLV + \delta(1-\gamma) \right] \beta_0 \lambda - \delta - c \right\} - A \tag{6.9}$$

With:

- Π = profit generated by a customer retention campaign,
- N = the number of customers in the customer base,
- α = the fraction of the customer base that is targeted in the retention campaign and offered an incentive to stay,
- β_0 = the fraction of all the operator's customers that will churn,
- λ = lift, i.e. how much more the fraction of customers included in the retention campaign is likely to churn than all the operator's customers. The lift indicates the predictive power of a classifier, and is a function of the included fraction of customers α with the highest probabilities to attrite, as indicated by the

model. Lift can be calculated as the percentage of churners within the fraction α of customers, divided by β_0. Thus, $\lambda = 1$ means that the model provides essentially no predictive power because the targeted customers are no more likely to churn than the population as a whole.

- δ = the cost of the incentive to the firm when a customer accepts the offer and stays
- γ = the fraction of the targeted would-be churners who decide to remain because of the incentive (i.e. the success rate of the incentive),
- c = the cost of contacting a customer to offer him or her the incentive,
- CLV = the customer lifetime value (i.e., the net present value to the firm if the customer is retained), and
- A = the fixed administrative costs of running the churn management program.

Both the costs and the profits generated by a retention campaign are a function of the fraction α of included customers. Optimizing this fraction leads to the maximum profit that can be determined using the lift curve:

$$\text{MP} = \max_{\alpha}(\Pi) \tag{6.10}$$

Many studies on customer churn prediction modeling calculate the top-decile lift to compare the performance of classification models, i.e. the lift when including the ten percent of customers with the highest predicted probabilities to attrite. However, setting $\alpha = 10\%$ is a purely arbitrary choice, and including ten percent of the customers generally leads to suboptimal profits and model selection, as shown by Verbeke et al. (2012). Since the ultimate goal of a company when setting up a customer retention campaign is to minimize the costs associated with customer churn, it is logical to evaluate and

select a customer churn prediction model by using the maximum profit that can be generated as a performance measure. Whereas the AUC measures the overall performance of a model, the MP criterion evaluates the prediction model at the optimal fraction of clients to include in a retention campaign. To calculate the MP, the values of the parameters CLV, γ, δ, and c in Equation 6.9 are taken equal to respectively 200€, 0.30, 10€ and 1€ based on values reported in the literature (Burez and Van den Poel, 2007; Neslin et al., 2006) and information from telecom operators.

6.4.5 Testing Statistical Significance

The aim of this study is to investigate how classification performance is affected by two specific factors, the type of Bayesian Network classifier and the use of Markov Blanket feature selection. A procedure described in Demšar (2006) is followed to statistically test the results of the benchmarking experiments and contrast the levels of the factors. In a first step of this procedure the non-parametric Friedman test (Friedman, 1940) is performed to check whether differences in performance are due to chance. The Friedman statistic is defined as:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right] \tag{6.11}$$

with R_j the average rank of algorithm $j = 1, 2, \ldots, k$ over N data sets. Under the null hypothesis that no significant differences exist, the Friedman statistic is distributed according to χ_F^2 with $k-1$ degrees of freedom, at least when N and k are big enough (e.g. Lehmann and D'Abrera (2006) give $k \cdot N > 30$ as criterion). This requirement is fulfilled in this study when comparing different classifiers ($N = 5 \cdot 3 = 15$ and $k = 16$) and when analyzing the impact of feature selection ($N = 5 \cdot 16 = 80$ and $k = 3$).

If the null hypothesis is rejected by the Friedman test, we proceed by performing the post-hoc Nemenyi test (Nemenyi, 1963) to compare all classifiers to each other. Two classifiers yield significantly different

results if their average ranks differ by at least the critical difference equal to:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{6.12}$$

with critical values q_α based on the Studentized range statistic divided by $\sqrt{2}$. To compare all classifiers with the best performing classifier the Bonferroni-Dunn test (Dunn, 1961) is applied, which is similar to post-hoc Nemenyi but adjusts the confidence level in order to control the family-wise error for making $k-1$ instead of $k(k-1)/2$ comparisons.

6.5 Discussion of results

The results of the experiments are reported in Table 6.5, in terms of AUC, and in Table 6.6, in terms of MP. The table displays the resulting measure for each classifier-feature selection combination and for the five data sets discussed in Section 6.4. Moreover, the column labeled AR reports the average rank of the classifier over all five data sets. To give an indication about the significance of the results, the following notational convention has been adopted. The score or average rank of the best performing classifier is always underlined. The average rank of a classifier is bold if the performance is not significantly different from the best performing technique at a significance level of 5%. Techniques which are different at a 1% significance level are in italic script, whereas a classifier differing at a 5% but not at a 1% level is in normal script. The Bonferroni-Dunn test (see Section 6.4.5) has been used to compute the statistical significance indicated in Table 6.5 and Table 6.6.

In what follows, the impact of MB feature selection and the performance of the Bayesian Network classifiers will be analyzed. Next to AUC, the Maximum Profit criterion has been used to measure classification performance. The main conclusions will be based on the MP, since it gives a more accurate assessment of classification performance in a practical churn setting. After all, an operator is

	Technique	O1	D1	D2	D3	UCI	AR
No feature selection	Log. Regr.	0.75	0.76	0.66	0.65	0.86	**8.60**
	Naive Bayes	0.72	0.77	0.59	0.61	0.83	12.20
	TAN	0.75	0.75	0.65	0.65	0.88	**8.60**
	FAN-SB	0.75	0.76	0.66	0.65	0.87	**6.60**
	FAN-LCV_LO	0.75	0.76	0.66	0.65	0.88	**6.40**
	SFAN-SB	0.75	0.76	0.65	0.65	0.87	**8.10**
	SFAN-LCV_LO	0.75	0.76	0.66	0.65	0.88	**3.20**
	SFAND-SB	0.75	0.76	0.65	0.65	0.87	**7.90**
	SFAND-LCV_LO	0.75	0.76	0.66	0.65	0.88	**3.20**
	STAN-SB	0.72	0.76	0.65	0.64	0.87	12.60
	STAN-LCV_LO	0.73	0.75	0.66	0.65	0.88	**9.40**
	STAND-SB	0.75	0.75	0.65	0.65	0.87	**8.60**
	STAND-LCV_LO	0.75	0.75	0.66	0.65	0.88	**6.80**
	K2	0.73	0.76	0.66	0.64	0.88	**7.20**
	TPDA	0.67	0.76	0.62	0.63	0.83	*13.60*
	MMHC	0.66	0.76	0.62	0.62	0.85	13.00
MB.05 feature selection	Log. Regr.	0.75	0.76	0.66	0.65	0.86	**8.40**
	Naive Bayes	0.73	0.77	0.60	0.62	0.85	**12.00**
	TAN	0.75	0.75	0.65	0.65	0.88	**8.80**
	FAN-SB	0.75	0.76	0.66	0.65	0.88	**6.40**
	FAN-LCV_LO	0.75	0.76	0.66	0.65	0.88	**6.60**
	SFAN-SB	0.75	0.76	0.65	0.65	0.88	**8.00**
	SFAN-LCV_LO	0.75	0.76	0.66	0.65	0.88	**5.20**
	SFAND-SB	0.75	0.76	0.65	0.65	0.88	**7.60**
	SFAND-LCV_LO	0.75	0.76	0.66	0.65	0.88	**3.60**
	STAN-SB	0.71	0.75	0.65	0.65	0.87	12.60
	STAN-LCV_LO	0.75	0.75	0.66	0.65	0.88	**7.60**
	STAND-SB	0.75	0.75	0.65	0.65	0.88	**8.00**
	STAND-LCV_LO	0.75	0.75	0.66	0.65	0.88	**5.40**
	K2	0.74	0.76	0.66	0.64	0.88	**9.80**
	TPDA	0.67	0.75	0.61	0.64	0.83	*15.20*
	MMHC	0.64	0.76	0.62	0.62	0.88	**10.80**
MB.01 feature selection	Log. Regr.	0.74	0.76	0.66	0.65	0.86	**9.40**
	Naive Bayes	0.73	0.76	0.60	0.63	0.85	**12.00**
	TAN	0.74	0.75	0.66	0.65	0.88	**8.40**
	FAN-SB	0.74	0.76	0.66	0.65	0.88	**6.80**
	FAN-LCV_LO	0.74	0.76	0.66	0.65	0.88	**6.60**
	SFAN-SB	0.74	0.76	0.65	0.65	0.88	**8.40**
	SFAN-LCV_LO	0.75	0.76	0.66	0.65	0.88	**3.40**
	SFAND-SB	0.74	0.76	0.65	0.65	0.88	**8.00**
	SFAND-LCV_LO	0.75	0.76	0.66	0.65	0.88	**3.20**
	STAN-SB	0.72	0.75	0.65	0.64	0.87	12.80
	STAN-LCV_LO	0.74	0.75	0.66	0.65	0.88	**8.00**
	STAND-SB	0.74	0.75	0.65	0.65	0.88	**9.40**
	STAND-LCV_LO	0.75	0.75	0.66	0.65	0.88	**5.40**
	K2	0.74	0.76	0.66	0.64	0.88	**9.40**
	TPDA	0.68	0.76	0.61	0.63	0.83	*14.00*
	MMHC	0.67	0.76	0.61	0.62	0.88	**10.80**

Table 6.5: Results of simulations evaluated with AUC. The average rank (AR) which is highest is underlined and bold, ARs which are not significantly different at a significance level of 5% are in bold. Techniques different at a 1% significance level are in italic script, whereas techniques significantly different at 5% but not at 1% are in normal script.

	Technique	O1	D1	D2	D3	UCI	AR
No feature selection	Log. Regr.	0.12	0.17	0.01	0.00	4.84	**7.60**
	Naive Bayes	0.14	0.17	0.00	0.01	4.22	**8.80**
	TAN	0.16	0.11	0.01	-0.00	4.78	**8.70**
	FAN-SB	0.16	0.15	0.01	-0.00	4.95	**7.60**
	FAN-LCV_LO	0.16	0.14	0.01	0.00	4.78	**6.90**
	SFAN-SB	0.16	0.16	0.01	-0.00	4.93	**5.90**
	SFAN-LCV_LO	0.15	0.16	0.01	0.00	4.76	**8.70**
	SFAND-SB	0.16	0.16	0.01	-0.00	4.93	**6.50**
	SFAND-LCV_LO	0.15	0.16	0.01	0.00	4.76	**7.70**
	STAN-SB	0.12	0.17	0.01	0.00	5.02	**5.60**
	STAN-LCV_LO	0.13	0.11	0.01	0.00	4.76	**10.90**
	STAND-SB	0.16	0.02	0.01	-0.00	4.94	**7.60**
	STAND-LCV_LO	0.15	0.02	0.01	-0.00	4.76	**10.70**
	K2	0.11	0.16	0.01	-0.00	4.90	**8.80**
	TPDA	0.02	0.15	0.00	0.00	4.75	**12.60**
	MMHC	0.02	0.16	0.00	0.00	4.78	**11.40**
MB.05 feature selection	Log. Regr.	0.16	0.17	0.01	0.00	4.81	**6.40**
	Naive Bayes	0.16	0.16	0.00	0.01	4.37	**7.00**
	TAN	0.14	0.11	0.01	-0.00	4.94	**10.20**
	FAN-SB	0.14	0.15	0.01	0.00	5.02	**7.00**
	FAN-LCV_LO	0.14	0.14	0.01	0.00	4.94	**8.40**
	SFAN-SB	0.14	0.16	0.01	0.00	5.02	**6.20**
	SFAN-LCV_LO	0.15	0.16	0.01	0.00	4.94	**6.80**
	SFAND-SB	0.14	0.16	0.01	0.00	5.02	**6.00**
	SFAND-LCV_LO	0.15	0.16	0.01	0.00	4.94	**7.60**
	STAN-SB	0.09	0.16	0.01	-0.00	5.06	**9.00**
	STAN-LCV_LO	0.15	0.11	0.01	-0.00	4.94	**10.20**
	STAND-SB	0.14	0.02	0.01	-0.00	4.94	**10.40**
	STAND-LCV_LO	0.15	0.02	0.01	-0.00	4.94	**9.80**
	K2	0.09	0.16	0.01	-0.00	4.98	**9.00**
	TPDA	0.02	0.16	0.00	0.00	4.75	**12.00**
	MMHC	0.02	0.16	0.00	-0.00	5.04	**10.00**
MB.01 feature selection	Log. Regr.	0.14	0.16	0.01	0.00	4.81	**8.20**
	Naive Bayes	0.13	0.17	0.00	0.01	4.37	**8.60**
	TAN	0.15	0.11	0.01	-0.00	4.94	**9.60**
	FAN-SB	0.15	0.15	0.01	0.00	5.02	**6.80**
	FAN-LCV_LO	0.15	0.14	0.01	0.00	4.94	**7.60**
	SFAN-SB	0.15	0.16	0.01	0.00	5.02	**3.00**
	SFAN-LCV_LO	0.13	0.16	0.01	0.00	4.94	**8.40**
	SFAND-SB	0.15	0.16	0.01	0.00	5.02	**3.40**
	SFAND-LCV_LO	0.13	0.16	0.01	0.00	4.94	**8.80**
	STAN-SB	0.10	0.17	0.01	-0.00	5.06	**8.40**
	STAN-LCV_LO	0.12	0.11	0.01	-0.00	4.94	**11.20**
	STAND-SB	0.15	0.02	0.01	-0.00	4.94	**9.20**
	STAND-LCV_LO	0.13	0.03	0.01	-0.00	4.94	**10.60**
	K2	0.13	0.15	0.01	-0.00	4.98	**8.60**
	TPDA	0.01	0.15	-0.00	0.00	4.75	13.00
	MMHC	0.00	0.16	0.00	-0.00	5.04	**10.60**

Table 6.6: Results of simulations evaluated with the maximum profit criterion. The average rank (AR) which is highest is underlined and bold, ARs which are not significantly different at a significance level of 5% are in bold. Techniques different at a 1% significance level are in italic script, whereas techniques significantly different at 5% but not at 1% are in normal script.

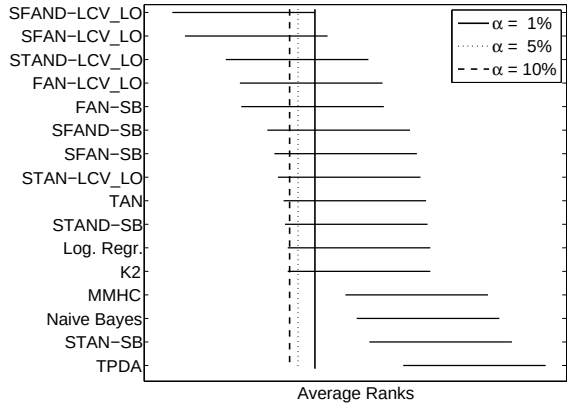
mostly interested in the fraction of the client base which will maximize his profit, i.e. those customers with a high churn probability. Typically, only a small proportion of all customers will be included in a retention campaign.

6.5.1 Classification Performance

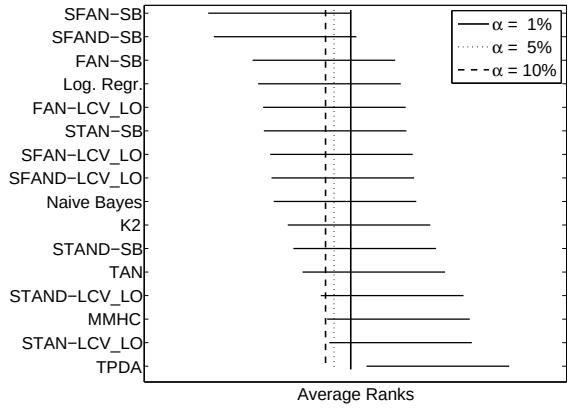
To analyze the impact of feature selection on classification performance, a Friedman test has been employed. When using AUC as performance measure, the p-value is 0.24, for MP it equals 0.85, both unambiguously implying that the feature selection does not significantly affect performance. This is positive, as it gives the opportunity to reduce the number of variables without a significant decrease in the predictive power of the resulting models. The complexity of the resulting models will be discussed in the next subsection.

The Friedman test is also applied to investigate whether the differences among classifiers are significant. The p-value with regards to the AUC measure is $6.53 \cdot 10^{-13}$ and for MP it equals $5.77 \cdot 10^{-4}$, indicating that the differences among classifiers are statistically significant. Hence, a post-hoc Nemenyi test is performed. The outcome of this test is graphically illustrated for both performance metrics in Figure 6.5(a) and Figure 6.5(b). The left end of the line segments indicate the average rank whereas the length is equal to the critical distance at a 1% level of significance, enabling the reader to compare techniques among each other, not only with the best method[5]. The vertical full line gives the cutoff for the 1% level, the dotted and dashed line for the 5% and 10% level whereas the critical distance is equal to 6.75, 5.96, and 5.56 for the the 1%, 5%, and 10% significance levels respectively. One can observe that AUC and MP lead to different rankings, although most of the techniques are not significantly different. AUC is more discriminative as opposed to MP, following from the fact that AUC is measuring classifier performance

[5]Note that the Nemenyi test is designed to allow comparisons between each pair of techniques, i.e. to make $k(k-1)/2$ comparisons



(a) AUC



(b) Maximum profit

Figure 6.5: Post-hoc Nemenyi test at a significance level of 1%. The dotted and dashed line indicate the cutoff at a significance level of 5% and 10% respectively.

over the whole output range, whereas MP only looks at the optimal fraction of clients to include in a retention campaign and assesses prediction performance at this specific point. Thus, when using MP to evaluate the classifiers for practical purposes, i.e. application in a real life churn setting, it is likely that none of the Bayesian Networks significantly outperforms the others, except for TPDA. This is an

interesting conclusion, since it indicates that General Bayesian Network methods (except for TPDA) could be used, which are preferable as they lead to more compact and interpretable networks. It is also remarkable that the BN classifiers are not able to outperform logistic regression, a straightforward and fast technique.

6.5.2 Complexity and Interpretability of the Bayesian Networks

The discussion above focussed on the performance of the classifiers, which is only one aspect of a good churn prediction model. Also the complexity and interpretability of the resulting classifiers are key properties, since comprehensibility of the model is important as discussed in Section 6.2. Bayesian networks are appealing to practitioners, as they give an intuitive insight in the factors driving churn behavior and the dependencies among those factors. This applies less to the Naive Bayes classifier and the Augmented Naive Bayes classifiers, since they include, by nature, all or many variables and prohibit general network structures. General Bayesian networks, on the other hand, are more flexible and typically use less variables. As a result, Markov Blanket feature selection will be more useful in combination with (Augmented) Naive Bayes classifiers, since these do not contain a mechanism to get rid of redundant variables.

Figure 6.6 illustrates the network complexity by showing the network dimension and the number of nodes and arcs for each algorithm-feature selection combination, averaged over the data sets. The dimension of a Bayesian Network is defined as the number of free parameters needed to fully specify the joint probability distribution encoded by the network, and is calculated as:

$$DIM = \sum_{i=1}^{n} (r_i - 1) \cdot q_i \tag{6.13}$$

with r_i being the cardinality of variable X_i and:

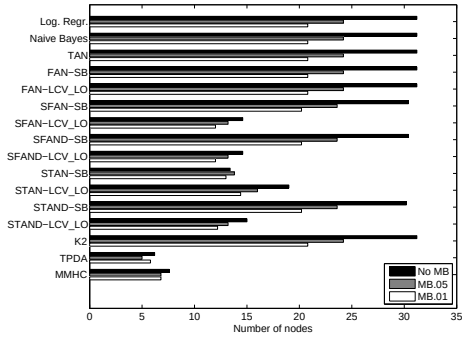
$$q_i = \prod_{X_j \in \Pi_{X_i}} r_j \tag{6.14}$$

with Π_{X_i} the direct parent set for node X_i. For logistic regression, which is included in the study as a benchmark algorithm, the number of nodes is equal to the number of attributes, the dimension (or number of free parameters) equals the number of attributes plus one, and the number of arcs is meaningless and therefore discarded for this algorithm.

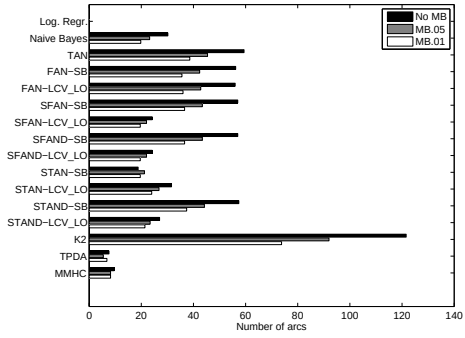
The graphs show that feature selection reduces the number of nodes, arcs and dimension, as one would expect. The reduction of nodes and arcs is more substantial for the (Augmented) Naive Bayes networks, as mentioned before. An exception could be noticed for the TPDA algorithm where the complexity again increases for MB.01 after having dropped for MB.05. This could be attributed to the fact that the MB feature selection algorithm excluded a variable which had large explanatory power within the TPDA framework and in order to offset this loss, the algorithm has to include more variables than before. This case illustrates that one should be careful when applying feature selection prior to the use of a GBN algorithm. This category of methods already reduces the number of parameters by itself, so reducing the variables prior to training the classifier might be redundant at best or even worsen the result.

Moreover, one can observe that the Augmented Naive Bayes classifiers, which relaxed the TAN assumption, are able to reduce the number of nodes and arcs compared to TAN, without a loss in predictive power. Nevertheless, the networks are still too complex to be easily interpretable. The GBN algorithms reduce the complexity even more and contain around 5 to 7 variables, with the only exception of K2, creating very dense networks[6]. Figure 6.7 shows the network created for data set D2 by the MMHC algorithm (without prior input selection). D2 is a real life data set from an operator and contains a high number of variables. Nonetheless, the algorithm is able to withhold only 5 attributes to predict churn behavior. When looking

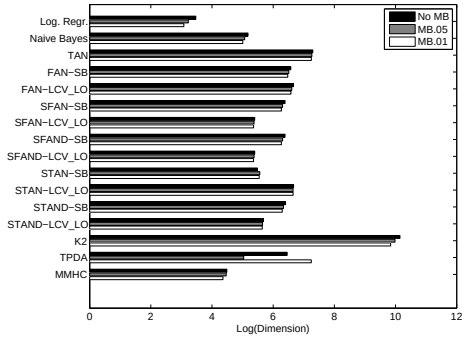
[6]In the K2 algorithm, it is possible to limit the number of parents for each node. As the goal was to test this algorithm as a GBN, this restriction was not imposed, resulting in very dense networks.



(a) Number of nodes



(b) Number of arcs



(c) Logarithm of the network dimension

Figure 6.6: Complexity of the Bayesian networks.

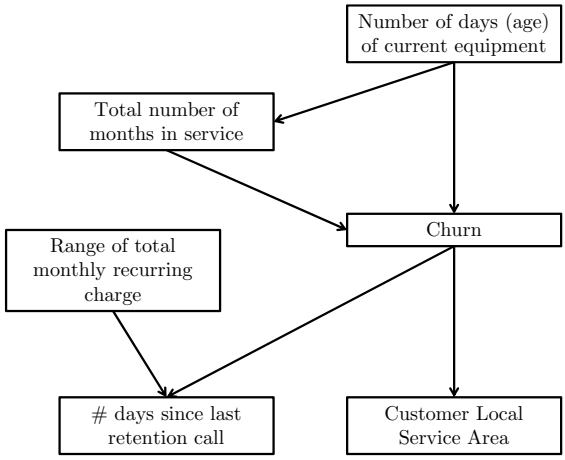


Figure 6.7: Bayesian network for data set D1, created with MMHC without prior feature selection.

at the network, it is very important to realize that the arcs do not necessarily imply causality, but they should rather be interpreted as correlation between the variables. For this network, one can observe that for instance the age of the current handset is correlated with the number of months in service and with churn behavior. The former relation could be explained by the fact that many operators offer a new mobile when signing a contract, whereas the latter could point to a motive for changing mobile phone operator, i.e. a promotional action at another operator. Such relations could be helpful for a company to identify red flags or warning signs for churn behavior. Moreover, it allows to check whether a churn prediction model is in line with current domain knowledge, increasing the credibility and applicability of those models.

6.6 Conclusion

Customer churn is becoming an increasingly important business analytics problem for telecom operators. In order to increase the efficiency of customer retention campaigns, operators employ customer churn

prediction models based on data mining techniques. These prediction models need to be accurate as well as compact and interpretable. This study investigates whether Bayesian Network techniques are appropriate for customer churn prediction.

Classification performance is measured with the Area under the Receiver Operating Characteristic Curve (AUC) and the Maximum Profit (MP) criterion. The results show that both performance measures lead to a different ranking of classification algorithms, even though not always statistically significant, and that AUC is more discriminative than the MP criterion. Whereas AUC measures performance over the whole customer base, the MP criterion only focusses on the optimal fraction of customers in order to maximize the effectiveness of a retention campaign. In a real life context, the MP criterion is preferred as it will maximize the profit for a telecom operator and therefore, the main conclusions will be based on this performance measure.

The results of the experiments show that Bayesian Network classifiers are not able to outperform traditional logistic regression. However, their contribution lies in the fact that they offer a very intuitive insight into the dependencies among the explanatory variables. The study indicates that Augmented Naive Bayes methods do not lead to compact networks, whereas General Bayesian Network algorithms result in simple and interpretable networks. This may aid practitioners in understanding the drivers behind churn behavior and in identifying warning signs for customer churn. Moreover, the Max-Min Hill-Climbing (MMHC) algorithm was able to create a compact and comprehensible model without a statistically significant loss in classification performance, as compared to logistic regression and Augmented Naive Bayes techniques.

Furthermore, the impact of Markov Blanket (MB) feature selection was tested. The outcome indicates that a reduction of variables as a result of MB feature selection does not decrease the performance significantly. Especially for Augmented Naive Bayes networks it proves to be useful, as it decreases the number of attributes sub-

stantially. For General Bayesian Network classifiers on the other hand, MB feature selection is discouraged as the General Bayesian Network algorithms themselves already limit the network complexity, making a prior input selection redundant.

Chapter 7

Predicting online channel acceptance with social network data

Abstract

The goal of this chapter is to identify a new way to predict whether a specific person believes buying online is appropriate for a specific product. By analyzing data that was gathered through a survey, we show that knowledge of a person's social network can be helpful to predict that person's e-commerce acceptance for different products. Our experimental setup is interesting for companies because (1) knowledge about only a small number of connections of potential customers is needed; (2) knowing the intensity of the relation is not necessary, and (3) data concerning variables such as age, gender and whether one likes working with the PC is not needed. Hence, companies can rely on publicly available data on their customers' social ties. Network-based classifiers tend to perform especially well for highly durable goods and for services for which few customers think it is appropriate to reserve them online.

7.1 Introduction

Online sales are on the rise. According to Forrester research (Mulpuru et al., 2012), Americans spent more than $200 billion on online shopping in 2011 and this figure is expected to increase to $327 billion by 2016. Online sales still only make up less than seven percent of overall retail sales. This implies that publicity for an online shop has only a low chance of being shown to a person who thinks that buying that product via an online channel is actually appropriate. The question arises how companies can better target their efforts in order to reach the people who believe that buying a specific product online is appropriate. In fact, prior research (also in this journal, Chiang et al. (2006)) already revealed the importance of being able to accurately predict a consumer's choice for buying via an online channel or a traditional channel.

The channel choice has been related to different elements perceived by consumers. Customers generally prefer traditional markets to web stores but the customer's acceptance of electronic channels depends on the products under consideration. Liang and Huang (1998) tried to relate the acceptance of online buying to the consumer's perception of transaction-costs associated with shopping (which in turn is determined by uncertainty and asset specificity). Other research indicated that online experience is the dominant predictor of whether or not a respondent had ever bought anything online (Bellman et al., 1999). Kwak et al. (2002) confirmed that experience with the Internet is an antecedent of Internet purchasing behavior and they showed that demographics, personality type and attitude towards the Internet are also relevant antecedents. The satisfaction with online shopping was shown to positively correlate with elements such as the consumer's perception of the convenience, product information and financial security of web stores compared to traditional stores (Szymanski and Hise, 2000). Gupta et al. (2004) showed that the loyalty to a channel (offline or online) depends on the risk-averseness of the person; while it is not necessarily so that risk-neutral consumers are more

likely to prefer the electronic channel than risk-averse consumers. All these studies assume that the process of the consumers' channel evaluation is linear compensatory. Chiang et al. (2006) developed neural network models to model noncompensatory decision processes. They found that noncompensatory choice models using neural networks outperform compensatory logit choice models in predicting consumers' channel choice.

All of the studies above found relevant antecedents of channel choice, but it is often difficult for companies to get access to the data about these antecedents if they want to target *potential* customers. In contrast, information on a potential customer's social network is often publicly visible (e.g. via a Facebook account; lists with members of a sports team, etc.) and the question arises whether companies could effectively leverage knowledge about people's social networks. Social networks have been shown to play a role in people's behavior. For example, Burt (1987) showed that connections between physicians have an effect on the adoption of a new drug by the physicians. Sykes et al. (2009) found that an employee's social network characteristics, capturing the employee's position in the network, help in understanding the use of a new company information system. Dierkes et al. (2011) showed that word of mouth in social networks affects churn (i.e., the decision to leave a communication service provider). Goethals (2012) found that knowledge of social links between students is valuable to identify which students will study an online theory video before next class. As a final example, Kiss and Bichler (2008) showed that the choice of network centrality measure used to select an initial set of customers in a viral marketing campaign matters. All in all, we can say that knowledge about a person's social network has been shown to be valuable in several settings.

We are not aware of studies that used knowledge of the consumers' network to predict the consumer's choice of an offline or online channel. Our research takes a complementary approach to the research on e-commerce acceptance mentioned above by using knowledge about social networks in predicting what channel a consumer finds

acceptable for the purchase of different products. More specifically, we suggest the use of network-based classifiers to predict consumers' choice for offline or online channels. Hence, this research addresses the following question: Can partial knowledge about a person's social ties help predicting the person's perceived appropriateness of the online channel to buy different products and services? Rather than trying to investigate this immediately at the level of the entire Internet population, this study tests the new approach in a subset of that population as a kind of proof-of-concept. There are several theories that explain the relevance of considering social networks in predicting people's behavior. People who are close to each other in a social network may behave similarly for several reasons. First, *social contagion* may cause people to choose engaging in similar behaviors. Second, *homophily* may cause people with the same behavior to stick together.

Social contagion arises when people in a social structure use one another to manage the uncertainty of innovation (Burt, 1987). There are two main models explaining social contagion. First, the cohesion model focuses on the effects of frequent and empathetic communication between the parties. The parties are more likely to adopt the same practices because they come to a shared evaluation of the costs and benefits of adoption (Burt, 1987). Second, the structural equivalence model shows that, even among people who are not directly connected but who are connected via a third party, there may be a similar behavior because of social contagion. This model accentuates the competition between individuals: if two individuals share more connections with other parties, there is a more intense feeling of competition between these individuals. If there is an innovation that could make one individual look more attractive than the other, he or she is likely to adopt it rapidly to prevent the other individual from appearing more attractive in the eyes of all shared connections. People then act as they believe people in their position should act (Harkola and Greve, 1995). Examining the importance of cohesion versus structural equivalence, one study found that the adoption

of medication by doctors was strongly determined by structural equivalence, while it was virtually unaffected by cohesion (Harkola and Greve, 1995). Similarly, in their study of perceptions of journal significance among sociological methodologists, Burt and Doreian (1982) found that while both cohesion and structural equivalence influenced expert perceptions of journal significance, the latter was a more accurate predictor. Other studies found that both exerted the same effects. Another study on the diffusion of construction technology suggests that the mechanism that is the most salient is actually contingent upon the diffusion phase (Harkola and Greve, 1995). The term homophily refers to the practice that generally contacts between similar people occur at a higher rate than among dissimilar people. Homophily theory (Lazarsfeld et al., 1954; McPherson et al., 2001) argues that individuals who are similar in terms of demographic and spatial attributes will also be similar in terms of beliefs and values. Homophily has been observed in many kinds of social networks (Blau, 1977; Macskassy and Provost, 2007; McPherson et al., 2001).

To the best of our knowledge, this is the first study that uses social network classifiers to predict the channel choice of an individual. Hence the main contribution of this chapter is that it tests the applicability of social network classification techniques to this particular prediction problem. This exploratory study shows that knowledge about the social network is valuable in this context and that the value depends on the product (group) under consideration. This study thus does not claim that this method outperforms all previous research models (although we did a benchmark, see Section 7.4). Rather, from a practical perspective, it recognizes that it is often easier for companies to get insight into a potential customer's social network (for example through Facebook pages, lists with members of local organizations,etc.), while it may be harder to get information on classic variables (e.g., frequency of Internet use, age, city size, etc.). The latter requires typically direct interaction with the potential customer.

In what follows, we first discuss the way this research was con-

ducted. Next, we present theory on social network classification in Section 7.3. Section 7.4 presents the data analysis approach. Subsequently, the results are discussed in Section 7.5 and conclusions are drawn in Section 7.6.

7.2 Research method

In this section we first justify the choice of our sample and we present respondent characteristics. We also justify the choice of different products that were used in the survey.

7.2.1 Survey procedure

As stated above, our research explores the usefulness of a new approach on one subset of the Internet population. More specifically, we relied primarily on a student sample, as is often done in exploratory Information Systems research. (In fact, 36% of papers in ISR and MIS Quarterly over the period 1990-2010 that used samples of individuals, used student samples (D. Compeau, 2012)). As the goal of our research is to analyze whether knowledge of the choice of a person's friends' channel allows one to predict that person's channel choice, we need to know about the channel choice of a number of interrelated people. As we did not know a priori how important it is to have knowledge on a big part of a person's network, we needed access to a network where a high response rate could be achieved. This is the case in our student population. (As will be clear from the experimental setup in Section 7.4, this allows us to gradually drop information on network nodes in the tests).

Two surveys have been conducted in April 2011. A first survey was meant to gather information about the social network of the respondents. The survey was distributed to students in their third year at a management school. Students were asked to list their closest friends at the school. They also had to grade the intensity of the relation (A: we are together most of the day, B: we meet once a day

for a short talk, C: we meet a few times every week, D: we meet once a week). This survey yields a social network indicating whether there is a link between two individuals, and how strong this link is.

The second survey concerned e-commerce adoption. The students involved in the first survey were explained in class the role of the Internet in doing business and were introduced to the survey. For a bonus mark, students were asked to fill out the survey themselves at home. They were also asked to invite both their parents to fill out the survey. In order to motivate students to collect data rigorously (and of course to have them learn about the issue), they received the additional task to write a paper, formulating findings based on an analysis of the data that was collected.

More specifically, the second survey asked respondents to reveal how appropriate they thought some medium was for buying different products (see Table 7.1 for the product list). They had to reply in line with how they purchased the product, or how they would purchase it if they had not purchased it before. For each product, the respondent had to give a value from one to five to reveal his opinion about the appropriateness of some medium that could be used to achieve the task. The value '1' was used to denote a medium that was considered 'very appropriate'; the value '2' indicated 'appropriate, but less important', etc. Several mediums were included (e.g., the website of the seller; another website (e.g., e-bay), face-to-face contact and telephone) and an option 'other' was included as well. The option 'other' generally received a score of 5 from the respondents, implying that the mediums we mentioned seemed comprehensive.

This survey also asked about several variables that had been shown in prior research to be important antecedents of e-commerce acceptance. For example, Bellman et al. (1999) revealed that online experience was the dominant predictor of whether or not a respondent had ever bought anything online and Kwak et al. (2002) showed the relevance of demographics and attitude towards the Internet. Hence, we included variables in the survey to measure the number of years Internet access, the intensity of Internet use and the intensity of

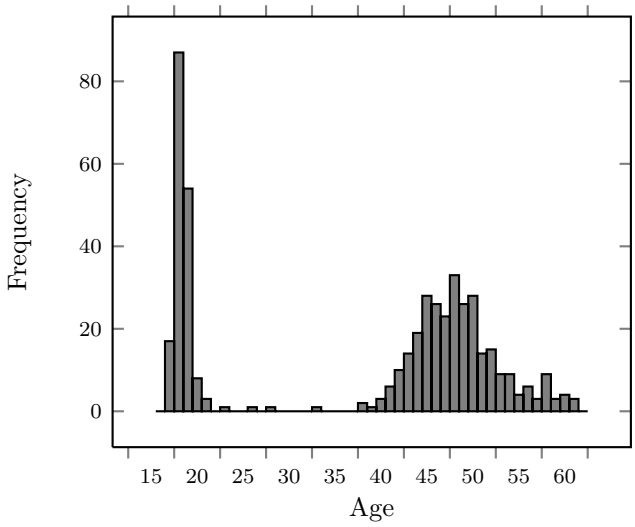


Figure 7.1: Age distribution of the sample.

e-mail use and the city size. Other variables that were included concerned age, gender, whether one likes working with the PC, etc. We refer to Table 7.2 for an overview of the different variables. We note this survey was run several consecutive years and each year the survey instrument was improved, to make sure students and parents would understand the questions correctly. The data used in this research project only stems from the last year this survey was conducted (academic year 2010-2011), as that was the only year also social network data was collected.

7.2.2 Respondent characteristics

We can distinguish two main groups in the sample: the students and their parents. This is clearly reflected in the age distribution, which is shown in Figure 7.1. There is a large group with the age closely distributed around twenty years, and another group where the age varies between 40 and 60 years.

The sample is evenly distributed with regard to the gender, with

approximately as many male (50.6%) as female (49.4%) respondents. Given the fact that the survey was conducted at a French Management School, it is not surprising that the majority (about 80%) of the respondents is French, whereas 8% is Indian and 10% comes from other countries. When looking at the occupation of the respondents, we note that roughly one third of the people are students, 55% is working and 10% is unemployed. Furthermore, most of the respondents use a PC for their occupation, since 66% indicates to use it constantly and 17% to use it sometimes.

7.2.3 Products presented in the survey

Prior research showed that for specific types of products, consumers are more likely to buy them online, while other products are more likely to be bought offline (Kwak et al., 2002). In contrast to such research, we do not try to predict the channel that will be used for different products. Rather, we investigate whether friends that want to buy a specific product are likely to behave alike. Other research showed that the channel attributes that influence a consumer's choice for a channel depend upon the product type (Chiang et al., 2006). While such research investigated whether specific variables play a role for the average customer when buying some product, we implicitly investigate here whether those variables are likely to get a different score from different people. For example, while prior research showed that the question of whether it is 'easy to find product information' is a relevant predictor of the choice of purchasing channel, the answer to that question may get a bigger weight from one person than from another when buying a new computer, depending on the IT skills of the person and the extent to which additional computer system information is directly understandable. On the other hand, when buying a new air-conditioning system, people may be more likely to have the same values that play a role. It is less likely for instance to have 'air-conditioning-savvy' people in our sample. All in all, for some products consumers may be more likely to buy the way their friends do, while for other products consumers may have other

values that play a role and they make up their own mind. Therefore, we considered it to be good practice to ask respondents about the appropriateness of different media to buy different products.

We built on prior research in the choice of products that were included in the survey. For the most part, Liang and Huang (1998) and Chiang et al. (2006) use the same categories of products: an information product, a product that one may want to try, a (deteriorating) product of which one may want to check the quality more intensively, a durable product, a product that may be bought with temporal consideration and a product that may be bought without much thinking. While we respected these categories, we made additional distinctions to get a bigger diversity in products, as shown in Table 7.3. For example, we distinguish products about which risk-averse people may be more or less anxious. Furthermore, we distinguish durable products for which different knowledge levels may be more or less likely. We also distinguish services for which the risk is run at the moment of (online) reservation or after the actual consumption of the service. For services, the practice of 'buying' then took the form of 'making a reservation for' the service. Looking from a different viewpoint, the product list contains products that our respondents were very likely to have bought and products they were less likely to have bought. Similarly, the list contains products and services that are very often bought online and others that are very often bought offline.

Before getting into real network effect tests, it is interesting to visually investigate whether network effects are likely to be found when predicting the channel choice for buying these products. The charts in Figure 7.2 show the relation between the acceptance of online buying by a node and the acceptance of online buying by the node's neighborhood for two products in the list. The dots in the charts show on the Y axis the average adoption rate of the nodes whose neighborhood's adoption rate lies within an interval of $[0-0.1)$; $[0.1-0.2)$; and so on (the dot is put in the middle of the interval on the X-axis). We only show data for some interval if there were

Prod. nr.	Product name	Description
1	2nd hand car	Deteriorating product requiring quality check
2	Books	Information product
3	Pubs and cafetarias	Service that may be acquired with temporal consideration (smaller risk)
4	Computers	Durable product with relatively high cost and requiring maintenance; very different knowledge levels likely
5	Fitness centers	Service bought without much thinking
6	Camera	Durable product with relatively high cost and requiring maintenance; slightly different knowledge levels likely
7	Hotels	Pure service that may be reserved with temporal consideration (financial risk at reservation)
8	Car rental	Product that may be acquired with temporal consideration (bigger risk)
9	Clothing	Product with special needs of physical trial
10	Family doctor	Pure service that may be reserved with temporal consideration (no financial risk at reservation)
11	Airconditioning	Durable product with relatively high cost and requiring maintenance, no different knowledge levels likely

Table 7.1: Products included in the survey

more than five nodes for which the neighborhood's adoption rate was in that interval, so as not to distort the overall picture with scarce cases. A linear trend line was added. Charts for other products are very similar. The charts suggest that for some products we are likely to find network effects (e.g. for product 3) and for other products we are less likely to find such effects (e.g. for product 2).

7.3 Social network classification

Formally, a social network is defined by a graph $\mathbf{G}$, consisting of a set of nodes $v \in \mathbf{V}$, that are connected by a set of links $e \in \mathbf{E}$; and $\mathbf{G} = (\mathbf{V}, \mathbf{E})$. The number of elements in the sets $\mathbf{V}$ and $\mathbf{E}$ are denoted respectively by n and k. The nodes in this study represent

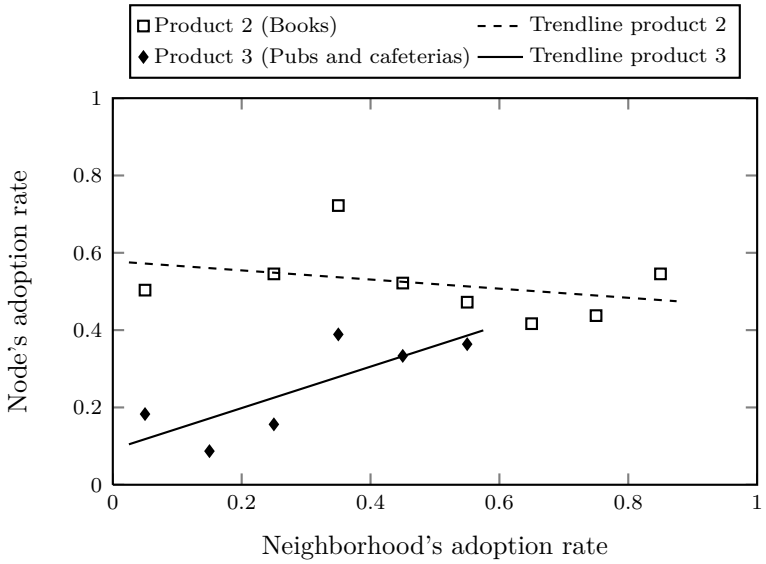


Figure 7.2: Relation between a node's adoption and adoption by the node's neighborhood

persons whereas the links represent social ties. The links in a graph can be directed or undirected. Directed links point from an origin to a destination node and incorporate a direction property (e.g., node v_i gives help to node v_j), whereas undirected links do not (e.g. node v_i and node v_j were present at the same event). In this study, undirected links are adopted, and maximum one link exists between two persons v_i and v_j. A graph can be represented by an adjacency matrix $\mathbf{A} = (a_{ij})$ or a weight matrix $\mathbf{W} = (w_{ij})$ of size $n \times n$. An entry a_{ij} or w_{ij} represents the edge between a vertex v_i and a vertex v_j. The value of a_{ij} in an adjacency matrix (also called connectivity matrix) is equal to one if an edge exists between vertices v_i and v_j, and equal to zero when no connection exists. In principle, the weights $w_{ij} \in \mathbf{W}$ can take any value. Typically, a weight expresses a property or characteristic of a link, in this study the strength of a social tie between two persons. A value of w_{ij} equal to zero means that no relation exists between two vertices. Furthermore, the classifiers

used in this study will focus on the first order neighborhood for classification. The first order neighborhood of a node i is indicated with N_i and contains all nodes which are directly connected with i.

Besides the social network structure, there is further information about the nodes themselves. Assume a data set $\mathbf{D} = (\mathbf{x}_i; l_i)$ with $i = 1 \ldots n$, $\mathbf{x}_i$ an attribute vector representing the information about person i (e.g. age, gender, etc.), and l_i a class label indicating the value for a binary target variable of the customer i (e.g. whether the customer intends to purchase the product online). In the remainder of this study, c will be used to refer to a non-specified class label value, either equal to one or zero. Typically, the class labels are not known for every node, and the set of nodes (persons) with unknown labels is denoted as $\mathbf{V}^U$, whereas $\mathbf{V}^K$ is the set of nodes with known labels. The aim of a classification method is to predict the unknown labels. In general, a classification technique M maps to each attribute vector $\mathbf{x}$ either a class label c or produces a continuous score $s(\mathbf{x})$. Typically, a higher value of the score indicates a higher probability, e.g. to purchase a certain product online, and the continuous score function $s(\mathbf{x})$ can be mapped to class labels by setting a cut-off value.

Traditional (i.e. non-relational) data mining techniques do not take into account information contained in the social network structure, but they use local variables. The difference between a local variable and a network variable is that a local variable can be calculated with information about the individual itself only. To calculate a network variable, information about the social network structure needs to be known. A typical example of a local variable is the age, whereas the average age of someone's friends would be a network variable. In order to incorporate network effects, Macskassy and Provost (2007) introduced a framework for classification in networked data. In this node-centric framework, a full relational classifier comprises a local (non-relational) classifier M_L, a pure relational or network classifier M_R, and a collective inference (CI) procedure. Local classifiers, which solely make use of the attributes that are related to the entity that is to be classified, have been extensively studied by the

data mining community (see above for examples in the e-commerce acceptance domain). Typical examples of local classifiers are decision tree classifiers, neural networks, support vector machines, statistical classifiers such as logistic regression or Bayesian-based methods, and many others. In this chapter, the emphasis lies on the network-based models which have been subject of few research so far, i.e. the pure relational classifier and the collective inference procedure. A pure relational classifier uses information about a node's neighborhood to predict its label or score. However, it is clear that, when two connected nodes have unknown labels, a pure relational classifier will experience problems to predict unknown labels. Collective inference (CI) procedures are developed to assign a label or score to all unknown nodes, by employing an iterative scheme. The relational classifiers and CI procedures applied in this study are briefly discussed in Section 7.3.1 and Section 7.3.2 respectively. For detailed information and formulas, one may refer to the paper of Macskassy and Provost (2007).

7.3.1 Relational classifiers

In this study, two relational classifiers are applied. The first one is the **weighted-vote relational neighbor** (wvrn) classifier, which estimates the probability of a customer to buy a product online as a function of the probabilities of his neighbors. More specifically, it will calculate the weighted sum of the neighbor's probabilities, according to the following formula:

$$P(l_i = c|N_i) = \frac{1}{Z} \sum_{v_j \in N_i} w_{ij} \cdot P(l_j = c|N_j), \tag{7.1}$$

where the linkage strengths w_{ij} serve as weights, and Z is a normalizer in order to yield probabilities between zero and one. The symbol N_i stands for the neighborhood of node i, i.e. it is the set of all the nodes which are connected to i. In other words, this means that the likelihood of someone belonging to class one becomes larger when more of his direct neighbors belong to class one, or when the link

to neighbors belonging to class one becomes stronger. As stated by Macskassy and Provost (2007), the wvrn classifier assumes the concept of homophily, i.e. the propensity for linked entities to belong to the same class.

The **class-distribution relational neighbor** (cdrn) classifier, based on Perlich and Provost (2003) and Rocchio's method (Rocchio, 1971) learns a model based on the global distribution of neighbor class labels, instead of simply using the weighted sum of neighbor class labels as the wvrn classifier does. The class vector $CV(v_i)$ of a node v_i is defined as the vector of summed linkage weights w_{ij} to the various classes:

$$CV(v_i)_k = \sum_{v_j \in N_i} w_{ij} \cdot P(l_j = c_k | N_j), \tag{7.2}$$

whereby the index k refers to the class under consideration. Consider for example a node v_1 which is connected to v_2 with weight 1 ($w_{12} = 1$) and to nodes v_3 and v_4 with weight 5 ($w_{13} = w_{14} = 5$). Furthermore, assume that v_2 is an adopter ($l_2 = 1$) and that v_3 and v_4 are non-adopters ($l_3 = l_4 = 0$). The class vector of v_1 is then equal to:

$$CV(v_1)_0 = 1 \cdot 0 + 5 \cdot 1 + 5 \cdot 1 = 10,$$
$$CV(v_1)_1 = 1 \cdot 1 + 5 \cdot 0 + 5 \cdot 0 = 1.$$

In the example hard labels (0 or 1) were used, but soft labels, i.e. probabilities, can be used during estimation. The reference vector $RV(c)$ of class c on the other hand is defined as the average of all class vectors for nodes known to be of class c:

$$RV(c) = \frac{1}{|\mathbf{V}_c^K|} \sum_{v_i \in \mathbf{V}_c^K} CV(v_i), \tag{7.3}$$

with $\mathbf{V}_c^K = \{v_i | v_i \in \mathbf{V}^K, l_i = c\}$. Then the probability for a customer to have a label equal to one can be calculated as the normalized vector similarity between v_i's class vector and the class reference vector of class one. Note that any vector similarity measure can be used, such as for instance L_1 or L_2 matrix norms, cosine similarity,

etc., normalized to lie in the range $[0; 1]$. In the experimental section of this study, cosine similarity will be applied. So whereas wvrn relies on the concept of homophily, cdrn concentrates less on the local environment of the node as this algorithm learns from the distribution of neighboring class labels in the global picture. We note that both relational classifiers discussed above implicitly make a Markov assumption by only including the neighbors of order one.

7.3.2 Collective inference procedures

A relational classifier cannot assign labels to all nodes with unknown labels in one single step, since two connected nodes may both have unknown labels. Therefore, a collective inference procedure is needed in combination with the relational classifiers discussed above. We have experimented with several collective inference procedures proposed in (Macskassy and Provost, 2007), such as Relaxation Labeling and Spreading Activation Collective Inference (spaCI), and found that these collective inference procedures performed very similar. Hence, since the methods do not make fundamentally different assumptions, we will only report the results from tests with Relaxation Labeling in our study.

Relaxation Labeling (RL) is based on a method introduced by Chakrabarti et al. (1998), and adopts the following procedure (Macskassy and Provost, 2007). The unknown nodes in $\mathbf{V}^U$ are initialized using a local classifier model M_L on the data in $\mathbf{D}$, whereby so called soft labels are used. This means that, rather than sampling a label value $l_i = \{0, 1\}$, the predicted probability $p_i \in [0, 1]$ is used as a label. Another option, which we used in this project, is not to use a local classifier. Then, the unknown nodes are initially left unknown. In a next step, an iterative procedure is followed, whereby for each unknown node a score (probability) is predicted, based on a relational classifier M_R. The iterations are done pseudo-simultaneously, based on the state of the network after iteration t. As opposed to the original method, with a pre-defined fixed number of iterations, we used an adapted version where the iterative procedure stops one step

after all the unknown labels are filled in. This is done in order to prevent that the collective inference procedure leads to a smoothing over the network where all persons receive very similar scores.

7.3.3 Alternative network classification methods

The network classifiers described until now, mainly focus on the network structure and the known target labels. The local attributes, such as socio-demographic information, are only taken into account through the local model, M_L, which is used to initialize the node labels. We have experimented with alternative methods to incorporate the network information into a local model, e.g. into a logistic regression, through the creation of extra variables containing information regarding the network structure. A first straightforward approach is to simply create one extra variable which is equal to the output score of the network classifier. A logistic regression then uses local attributes and the score given by the network classifier in order to estimate class membership.

Another way to take into account the network structure is by looking at someone's neighborhood and creating extra variables based on this neighborhood. Consider that an individual is characterized by a feature vector $\mathbf{x}$, consisting of m variables such as age, gender, etc. For each of the variables, one could look how an individual's neighborhood is characterized. For example, instead of only looking at the influence of someone's age on e-commerce adoption, this would also take into account the average age of the friends. Hereby, the average is calculated for continuous variables, and the mode for categorical variables. Such an approach is called propositionalization or featurization, and allows to incorporate social network effects within non-relational models.

We have experimented with these alternative network classification procedures in this study. The results were, however, not better than when applying the network classifier or logistic regression separately. Therefore, details about these results will not be reported and we only mention the alternative network classification procedures

here for reasons of completeness.

7.4 Empirical case study

In what follows, the social network classifiers introduced in Section 7.3 will be used to analyze e-commerce adoption, based on the survey data presented in Section 7.2. First, the necessary data preprocessing is discussed. Then we briefly give an overview of the experimental setup, after which the results are shown.

7.4.1 Data preprocessing

The first survey collected information about students' social ties by asking students to list their friends. Processing the survey data results in a network graph with directed links, whereas the methods described in Section 3 work with undirected links. Therefore, the directed links have been replaced with undirected links. Three different representations of the network were created that way: one with binary information about the existence of links, one with the maximum link strength of both directed links, and one with the average link strength of both directed links. This resulted in social networks with 681 nodes and 1102 undirected links.

In the second survey the respondents were asked to provide information about how they purchased each of the products, or how they would purchase it if they had not purchased it before. No distinction was made in the tests between an actual purchase and the reflection of how a customer would make a purchase, following the reasoning of Jackson et al. (1997). The respondents indicated the appropriateness of two online media on a scale from 1 to 5. Given the fact that a value of '1' means that the medium is considered very appropriate, the minimum of the scores on the two media was taken to have a score for the appropriateness of online media. Next, this score (still having possible values from 1 to 5) has been transformed into a binary target variable where 1 corresponds to appropriate

(1 or 2) and 0 to less appropriate (3 to 5). This binary target variable was to be predicted by the social network classifiers. As respondents could mention friends at school who did not fill out the survey (but nevertheless connected several students), we did not have data on all nodes in the network. Overall, data on 38% of the nodes was missing.

Besides the target variables, further information from the survey is available, displayed in Table 7.2. These variables would typically be used by a traditional data mining approach (by a local model, as defined in Section 7.3), whereas the social network information is used by the relational classifiers. In this study, a logistic regression model will be used as local model for benchmarking. The relevance of many variables in Table 7.2 has been shown in prior research (as illustrated with references).

7.4.2 Experimental setup

The network-based classifiers employed in this study only consist of a relational classifier M_R and a collective inference procedure, and do not incorporate a local model. The two relational classifiers introduced in Section 7.3.1 have been combined with the CI procedure from Section 7.3.2, leading to two networked classification techniques, which only take into account the network structure and the known labels. Remind that these techniques have been used on three types of network structure: a simple connectivity matrix, a weight matrix taking the average link strength, and a weight matrix with the maximum link strength. Furthermore, a logistic regression model (LR), which is a local classifier, has been included in order to compare the performance of the networked models too. However, it is important to note that the nature of the data required by both models is significantly different. The logistic regression model is a local model and hence requires local data. The network classifiers on the other hand only need a connectivity/weight matrix, which reveals the connections between people. This information is much more readily available on online social network platforms, such as Facebook for instance. The networked learners have been implemented in Matlab,

Variable Description	Type	Range	References
Age	Continuous	19-73	(Bellman et al., 1999; Lieber and Syverson, 2011; Richa, 2012)
Gender	Categorical	Male / Female	(Hannah and Lybecker, 2010; Lieber and Syverson, 2011; Richa, 2012)
Education	Categorical	Primary / Secondary / Polytechnic / Univ.	(Hannah and Lybecker, 2010; Lieber and Syverson, 2011)
City Size	Categorical	Big / Medium / Small	(Lieber and Syverson, 2011)
Nr of persons in family	Integer	0-9	(Richa, 2012)
Intensity of Internet use	Continuous	0-10	(Hannah and Lybecker, 2010)
Intensity of e-mail use	Continuous	0-10	(Hannah and Lybecker, 2010)
Nationality	Categorical	French / Indian / Other	(Bellman et al., 1999)
Occupation	Categorical	Student / Working / Unemployed	(Hannah and Lybecker, 2010)
PC use at school / work	Categorical	Constantly / Sometimes / Never / na	(Bellman et al., 1999)
Level of Internet access at school / work	Categorical	Full / Restricted / LittleOrNone / na	(Bellman et al., 1999)
Internet access at school / work since	Continuous	0-19 years	(Bellman et al., 1999)
I like working with the PC	Continuous	0-10	(Bellman et al., 1999)
If I would buy online, I would only buy on sites in my own language	Continuous	0-10	(Nazir et al., 2012)
Response concerns a purchase I really did (control variable)	Categorical	Yes / No	
Purchasing online is considered appropriate	Binary	0/1	

Table 7.2: Description of the variables used in the logistic regression (the last variable is the target variable).

whereas for logistic regression the Weka toolbox was used.

To gauge the strength of classification techniques, a variety of performance measures has been proposed (Ali and Smith, 2006). A very commonly used tool in performance measurement of classifiers is Receiver Operating Characteristic (ROC) analysis (Fawcett, 2006). Typically, a classifier assigns a score $s(\mathbf{x})$ to each instance, based on the attribute vector. Classification is then based on the score by defining a cutoff t, whereby instances with scores higher (lower) than t are classified as cases (non-cases). A ROC curve shows the fraction of the correctly classified cases (i.e. instances with label one), versus one minus the fraction of the correctly classified non-cases (i.e. instances with label zero), for a varying cutoff. A classifier which ROC curve lies above the ROC curve of a second classifier is superior, and the point $(0; 1)$ corresponds to perfect classification. Although ROC curves are a powerful tool for comparing classifiers, practitioners prefer a single number indicating the performance over the visual comparison of ROC curves. Therefore, the area under the ROC curve (AUC) has been widely adopted in the data mining community for its simplicity of use. Assume that $f_l(s)$ is the probability density function of the scores s for the classes $l \in \{0, 1\}$, and $F_l(s)$ the corresponding cumulative distribution function. Then, it can be shown that AUC is defined as follows:

$$\text{AUC} = \int_{-\infty}^{+\infty} F_0(s) \cdot f_1(s) ds. \tag{7.4}$$

If the AUC of a classifier is larger than 0.5, then the classifier is able to draw meaningful conclusions with regard to the target variable (Fawcett, 2006).

Besides AUC, which is an important tool to compare classifiers, we also used the lift coefficient in this study to measure performance. This measure is commonly used in marketing literature (see e.g. (Neslin et al., 2006)) and focuses on the prediction performance in the targeted fraction, rather than the whole sample. The lift at T% is equal to the percentage adopters in the top T% ranked customers (ranked by the network model), divided by the base rate

of adopters. For instance: assume that a predictive model is used to rank a population (with 25% of adopters) in terms of their likelihood to adopt a certain technology. The 10%-lift then only looks at the percentage of adopters in the 10% highly ranked customers. If there are 75% of adopters in this top fraction, the 10% lift is equal to 3, i.e. the model does 3 times better than a random selection. Since the lift coefficient is considered in this domain to be more relevant than other measures (such as accuracy) (Piatetsky-Shapiro and Masand, 1999), predictive performance will also be reported in terms of the top 10% lift (Lift10).

For classification performance measurement in this study, the data set has been split into a test set and a training set. Three scenarios have been experimented with: a test set consisting of 30%, 50% and 70%, drawn as a stratified random sample from the set of students with known labels. We used three scenarios to assess the practical relevance of our results. While companies may have easy access to networks of people, they do not necessarily have knowledge about the behavior of these people. By increasing the test set size, we simulate the fact that a company would only know the label of very few connected nodes in the social network. By assuming the labels in the test set to be unknown, the percentage of customers (nodes) with unknown labels in the network increases from 38% (i.e., friends that were mentioned by students in the first survey but which did not fill out the second survey) to respectively 55%, 68%, and 81%. The reason why the instances with unknown labels, other than those in the test set, have not been removed from the data set is because they may contain relevant information about the social network structure. Note that for the relational classifier, there is no need to train a classifier. For the logistic regression, however, the nodes which are not in the test set, and which have a known label, are used to train the model.

As can be clear from the presentation of the network classifiers in Section 3, the calculations typically take weights of relations into account. However, a company that analyses social network data

in practice often only knows about the existence of a relationship, without knowing the 'weights' of the relation. The network classifiers described in Sectio 7.3 are able to work with a connectivity matrix and with a weight matrix. Therefore, we ran all calculations three times: twice using a weight matrix (maximum weight and average weight) and once using an adjacency matrix (i.e. with binary values rather than weights).

The complete tables of results are reported in the appendix in Tables 7.4 to 7.9. A discussion of the main research findings follows in Section 7.5.

7.5 Discussion

7.5.1 Practical relevance

The results are interesting in many respects. Let us return to our research question: whether social network data could be helpful to predict a person's perceived appropriateness of the online channel to buy different products and services. To answer this question, we basically only need to investigate the results of the smallest test set, which are summarized in Figure 7.3. This graph plots the performance of the relational classifiers and the logistic regression for each product, in terms of AUC (top curves, left axis) and Lift10 (bottom curves, right axis).

A first observation is that for all products the AUC of the network model is considerably higher than 0.50 (the critical value) and almost always above 0.60; with a highest score 0.67 for fitness centers (product 5). This implies that knowledge about social networks is helpful to predict a consumer's channel choice. While benchmarking with logistic regression models is not the main goal of this study, we observe that the network model mostly outperforms the logistic regression model in terms of AUC (with rather big differences especially for products 1, 3, 5, 10 and 11). When it does not outperform the logistic regression model, the network model performance is often

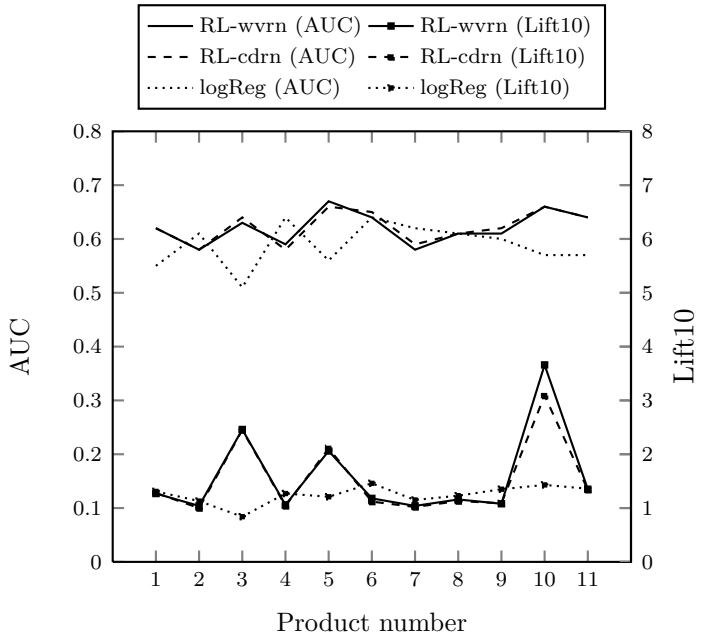


Figure 7.3: Comparison of social network classifiers vs. logistic regression in terms of AUC and Lift10 (test set size: 30%; link type: average).

close to that of the logistic regression model.

A look at the results in terms of Lift10 confirms that the network model often is very valuable. This is especially the case for products 3, 5 and 10. Interestingly, these are the items for which fewest people say that online buying is appropriate. The value of studying the network dynamics then becomes apparent when considering the percentage of people that say online buying is appropriate. The correlation between the percentage of the respondents that accept online buying and Lift10 (or AUC) is very negative: -0.83 (for both). (Correlation coefficient calculated using the results of the RL-wvrn network model using average weights and a 30% test set, as shown in Table 7.3.)

For companies that are in an industry where many people think buying online is appropriate (such as the hotel industry), it is less relevant to predict for unknown people whether they think buying

online is appropriate (as the chance is very high for a random person to accept online buying), and consequently there is little useful to learn from a network model. On the other hand, for companies in industries where few people think online buying (or making online reservations) is appropriate, it is important to identify those people. It is exactly in those cases that the network model gives a (very) good performance, with a Lift10 going as high as 3.66. We notice that – in contrast to what we see with the relational models – the correlation between the AUC of the logistic regression and the percentage of people accepting online buying is highly positive (0.72), while the Lift10 is hardly correlated with that percentage (-0.05). Hence, for companies that are in an industry where many people think buying online is appropriate (such as the hotel industry) that still want to find out who (not) to target with their online reservation system, a more fully fledged logistic regression may be more appropriate (if the necessary data to run such a regression is available). That is, the network model and the logistic regression could be applied in different industries.

7.5.2 Differences across product groups

Turning to differences between different products, theory suggests that under conditions of higher uncertainty people would behave more like their friends (Burt, 1987). Above we already mentioned the fact that our results are relevant because the relational classifier scores particularly well when few people accept to buy the product online (while this is not the case for the logistic regression). E.g., for services like hotels many people have made reservations before. They have lived their own experiences, which influence them (possibly together with the influence of other people). The products in Table 7.3 have been grouped in 5 categories, so that we can formulate findings for different types of products. We distinguish durable goods (split in highly durable and moderately durable goods) from soft goods and services (split in services that are typically not reserved online and services that are commonly reserved online). These groups are often

	AUC (network model)	Lift10 (network model)	AUC (LogReg)	Lift10 (LogReg)	% of nodes accepting online buying	nr. of nodes having bought product already
Family doctor (prod. 10)	0.66	3.66	0.57	1.43	10	376
Pubs and cafetarias (prod. 3)	0.63	2.46	0.51	0.84	21	369
Fitness center (prod. 5)	0.67	2.06	0.56	1.21	23	198
G1: Services typically reserved offline	0.65	2.73	0.55	1.16	18	314
Airconditioning (prod. 11)	0.64	1.35	0.57	1.36	33	110
2nd hand car (prod. 1)	0.62	1.27	0.55	1.31	33	234
G2: Highly durable goods	0.63	1.31	0.56	1.34	33	172
Camera (prod. 6)	0.64	1.18	0.64	1.46	43	361
Computers (prod. 4)	0.59	1.06	0.64	1.27	49	406
G3: Moderately durable goods	0.62	1.12	0.64	1.37	46	384
Clothing (prod. 9)	0.61	1.08	0.6	1.35	45	455
Books (prod. 2)	0.58	1.04	0.61	1.13	51	454
G4: Soft goods	0.6	1.06	0.61	1.24	48	455
Car rental (prod. 8)	0.61	1.16	0.61	1.23	58	268
Hotels (prod. 7)	0.58	1.04	0.62	1.15	70	393
G5: Services often reserved online	0.60	1.10	0.62	1.19	64	331

Table 7.3: Data used to calculate the correlation between Lift10, AUC (of the network model) and % of nodes accepting online buying and number of nodes having bought product/service already, online or offline (data from the RL-wvrn model using average weights with a 30% test set).

mentioned in marketing literature and – although there are only 2 or 3 items per group – we can get initial insights into the role of social network models in predicting the acceptance of online buying of some product group. To do so, we make averages of the values of the different items in the group (shown in gray in Table 7.3). The correlation between the AUC (or Lift10) and the percentage of nodes accepting online buying for these groups is very high: -0.93

(or -0.83). Our respondents clearly think that buying highly durable goods online is less appropriate than buying moderately durable or soft goods online. In line with what we said above, the network model then performs better in the context of highly durable goods. Similarly, the network model performs better for the services that are seldom reserved online than for services that are very often reserved online.

The five groups of products in Table 7.3 have different scores in terms of Lift10 and AUC, which can be explained by the uncertainty associated with respect to buying these products/services online (as mentioned above). There are also some within-group differences. While the differences in Lift10 within groups 1, 3, 4 and 5 correlate with the percentage of nodes that accept online buying, other dynamics are creating the difference within group 2: for both products 33% of respondents said that online buying is appropriate but the AUC and Lift10 are different. This difference could be related to another element of uncertainty: whether the person already bought the product before. Only 110 respondents said they had bought airconditioning before while more than double (234) said they had bought a 2nd hand car. If a person never bought airconditioning before, (s)he has no personal experience in this and friends will act more like role models. In general, people are more uncertain about products and services they have not bought/reserved before than for those they have reserved before. The correlation coefficient between the number of respondents that actually bought a product and the AUC score of the relational classifier for that product is negative (-.55) (data used on individual products as in Table 7.3). In conclusion, while we suggest further research to investigate specific product groups in more detail by considering more items for a group, our findings suggest that network models are particularly valuable in the context of highly durable goods and services that are seldom bought online.

7.5.3 Impact of relational classifier and link type

The choice of the relational classifier seems relevant, as demonstrated by Figure 7.4. This graph compares the two types of relational

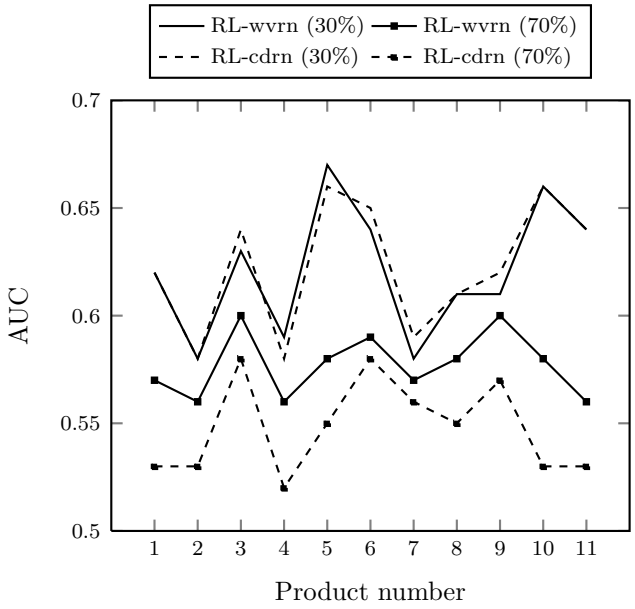


Figure 7.4: Comparison of relational classifiers for different test set sizes in terms of AUC (link type: average).

classifiers (wvrn vs. cdrn) for two test set sizes (30% vs. 70%) in terms of AUC. More specifically, while wvrn is not outperforming cdrn significantly in case of a test set of 30%, it is always the best performing relational classifier when a test set of 70% is used. This implies that in practice, where companies only know the labels of very few nodes in the network, companies can rely on wvrn, which happens to be the easiest of the two relational classifiers to implement. We note that this finding suggests our approach is viable in a company. Using a test set of 70% implies 81% of the labels were unknown (see above). Given that each student mentioned only few links in the survey, the absolute number of labels that were known for people in the neighborhood was very low. By extension, if a company studies the social network (e.g. on Facebook) of its online customers, it has a reasonably good chance of decently predicting at least the e-commerce acceptance of common friends.

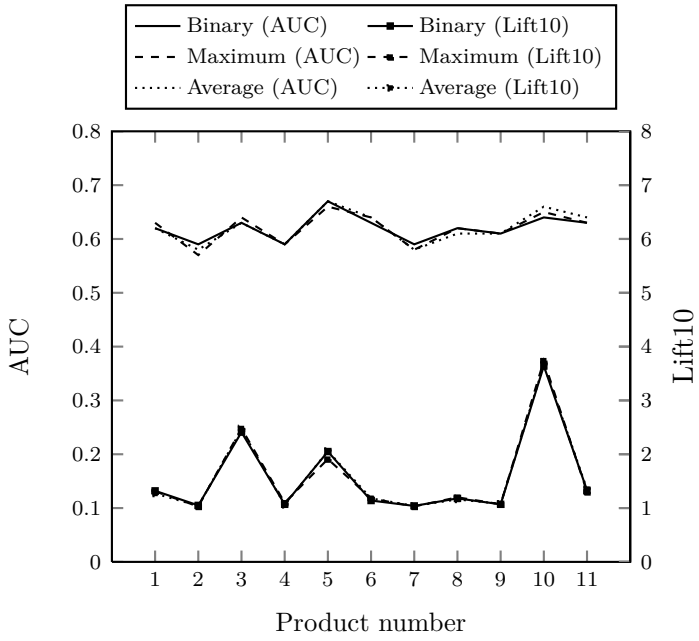


Figure 7.5: Impact of link type in terms of AUC and Lift10 (RL-wvrn classifier and test set size: 30%).

A second important observation (for practical reasons) is that the results of the calculations are valuable even though only information on undirected links was used. The algorithms of Macskassy and Provost (2007) were implemented using undirected links and further research can investigate the value of having different link weights in different directions. Figure 7.5 also enables us to analyze the impact of the different types of connectivity matrix. It compares the performance of the best performing relational classifier (wvrn) in terms of AUC (top curves, left axis) and Lift10 (bottom curves, right axis), and this for three link types: binary (Bin), average link strength (Avg), and maximum link strength (Max). It is clear from the graph that neither AUC nor Lift10 is influenced by the choice of link type. This implies that companies do not need data on the weights of relations but can use basic information on the presence of social ties.

Such binary information can often be easily retrieved, be it via online social media (such as Facebook) or other media (such as membership lists of organizations; purchases done together, for example for events, etcetera). Moreover, the social network algorithms do not need data on variables such as age, gender, intensity of Internet use, etcetera, so that our approach can really be operationalized in companies.

7.5.4 Limitations and suggestions for future research

This chapter's new approach to predicting channel choice opens up several avenues for future research. First, the fact that a significant AUC is achieved is promising, given that we only consider a low number of friends for every node. In our dataset, there are 1102 (undirected) edges for 681 nodes, implying there are on average only 3.24 connections per node. Further research could investigate the social network at a bigger scale, including more friends per node. Given that our research results show – as one could expect – that the AUC goes down if less labels are known, the AUC presumably will go up even further if we would have knowledge on labels of more connections. Further research could investigate the value of adding knowledge on more links.

Secondly, and as an alternative to the first suggestion, future research can consider different subsets from the Internet population. One limitation of our research is that we only cover specific age ranges. While student samples have been used in similar research before and a student sample was interesting to get a high response rate to obtain sufficient data to test whether considering social networks could be valuable, future research can start from our finding that social networks matter, at least in our subset of the population. Interestingly, our research suggests that future research can even start from little knowledge about links in the social network, so that more sparse networks from non-student samples are justified in future research. Such future studies may want to use a sample from different age categories than the ages considered in this study, so as to test whether our findings apply to other age groups.

Thirdly, we used products in the survey that are likely to differ in terms of risk that is associated with the purchase, product knowledge levels that are possible, the moment financial risk is taken when reserving a service and so on (see Section 7.2). However, we did not measure the respondents' perception of these factors. Future research could try to relate the performance of the network model to the variability in risk perception, knowledge level, etc. among connected nodes in the network. Another possible extension of the research involves investigating other steps of the procurement process, such as the step where product information is gathered and the step where one is searching for a supplier. Again one could investigate what media are considered appropriate to find product information or to find a supplier and check whether network effects play a role in that step.

Furthermore, given the different practical consequences of needing social network data rather than data on other variables (such as age, gender, intensity of Internet use, etc.), our research project did not aim at comparing network model results with results given by other models. The first benchmark results, comparing network models with logistic regression models, are promising. Future research is needed to benchmark the network model results with more advanced models such as models based on neural networks(Chiang et al., 2006), but our research suggests that network models and logistic regression models could be valuable in different industries. Finally, we only consider few products per product category. In order to draw valid conclusions about entire product categories, it would be good to focus future studies on many products within individual product groups. In line with the way recommender systems are often built (Huang et al., 2004), such a system could also predict the acceptance of online buying by node v_i for product X by looking at node v_i's acceptance of online buying for related products A, B and C and the way other nodes accept online buying for products A, B, C and X.

7.6 Conclusion

The main contribution of this chapter is that it shows that knowledge of a person's social network can be valuable to help predict the person's acceptance of the online channel for buying different products. For our sample of the Internet population, consisting of students and their parents, we found that knowledge on a small number of connections is sufficient to get a better-than-random prediction of the person's e-commerce acceptance. Secondly, our research results suggest that knowing the intensity of the relation is not required. Thirdly, with our approach companies do not need to gather data concerning variables such as age, gender or whether one likes working with the PC. With our new approach companies can try to use publicly available data on their customers' social ties to predict the e-commerce acceptance of the customers' connections. All in all, it seems interesting for companies to build a decision support system which analyses the social networks of existing customers, especially when selling services/products that are reserved/bought online by few people. It allows companies to identify potential consumers for their online shop, so that online advertisements can for example be targeted at only those people that are likely to buy online.

Appendix – Tables of results

Prod. nr.	Binary		Maximum		Average		LogReg
	RL-wvrn	RL-cdrn	RL-wvrn	RL-cdrn	RL-wvrn	RL-cdrn	
1	0.62	0.62	0.63	0.63	0.62	0.62	0.55
	(0.05)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)
2	0.59	0.58	0.57	0.58	0.58	0.58	0.61
	(0.04)	(0.06)	(0.04)	(0.04)	(0.04)	(0.06)	(0.04)
3	0.63	0.63	0.64	0.64	0.63	0.64	0.51
	(0.05)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.04)
4	0.59	0.58	0.59	0.59	0.59	0.58	0.64
	(0.04)	(0.06)	(0.04)	(0.05)	(0.04)	(0.06)	(0.04)
5	0.67	0.65	0.66	0.65	0.67	0.66	0.56
	(0.05)	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)
6	0.63	0.64	0.64	0.65	0.64	0.65	0.64
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
7	0.59	0.60	0.58	0.60	0.58	0.59	0.62
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)
8	0.62	0.62	0.62	0.62	0.61	0.61	0.61
	(0.04)	(0.05)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)
9	0.61	0.61	0.61	0.61	0.61	0.62	0.60
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
10	0.64	0.64	0.65	0.65	0.66	0.66	0.57
	(0.09)	(0.09)	(0.08)	(0.09)	(0.08)	(0.08)	(0.07)
11	0.63	0.63	0.63	0.63	0.64	0.64	0.57
	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.06)	(0.04)

Table 7.4: Average AUC of the classification methods over 100 iterations, with the standard deviation between brackets, for the test set size of 30%

Prod. nr.	Binary		Maximum		Average		LogReg
	RL-wvrn	RL-cdrn	RL-wvrn	RL-cdrn	RL-wvrn	RL-cdrn	
1	0.59	0.58	0.60	0.59	0.60	0.59	0.56
	(0.04)	(0.06)	(0.04)	(0.06)	(0.03)	(0.05)	(0.03)
2	0.57	0.56	0.57	0.57	0.57	0.56	0.60
	(0.03)	(0.06)	(0.03)	(0.05)	(0.04)	(0.06)	(0.03)
3	0.62	0.62	0.63	0.63	0.62	0.61	0.52
	(0.04)	(0.05)	(0.04)	(0.06)	(0.05)	(0.07)	(0.04)
4	0.57	0.56	0.58	0.56	0.58	0.55	0.61
	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.07)	(0.03)
5	0.64	0.61	0.63	0.60	0.64	0.62	0.54
	(0.04)	(0.05)	(0.04)	(0.07)	(0.04)	(0.06)	(0.03)
6	0.61	0.62	0.62	0.63	0.62	0.63	0.62
	(0.03)	(0.05)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)
7	0.58	0.58	0.57	0.58	0.58	0.59	0.60
	(0.04)	(0.07)	(0.04)	(0.06)	(0.04)	(0.05)	(0.04)
8	0.60	0.58	0.61	0.59	0.61	0.60	0.59
	(0.03)	(0.06)	(0.03)	(0.06)	(0.03)	(0.05)	(0.03)
9	0.61	0.60	0.61	0.60	0.61	0.60	0.59
	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.06)	(0.04)
10	0.63	0.59	0.61	0.60	0.63	0.60	0.53
	(0.07)	(0.10)	(0.07)	(0.09)	(0.06)	(0.10)	(0.06)
11	0.60	0.58	0.60	0.59	0.60	0.59	0.56
	(0.04)	(0.07)	(0.03)	(0.06)	(0.03)	(0.06)	(0.03)

Table 7.5: Average AUC of the classification methods over 100 iterations, with the standard deviation between brackets, for the test set size of 50%.

Prod. nr.	Binary RL-wvrn	Binary RL-cdrn	Maximum RL-wvrn	Maximum RL-cdrn	Average RL-wvrn	Average RL-cdrn	LogReg
1	0.56 (0.03)	0.54 (0.06)	0.56 (0.03)	0.53 (0.07)	0.57 (0.03)	0.53 (0.07)	0.54 (0.03)
2	0.56 (0.03)	0.54 (0.05)	0.56 (0.03)	0.53 (0.06)	0.56 (0.03)	0.53 (0.06)	0.57 (0.03)
3	0.60 (0.04)	0.57 (0.08)	0.60 (0.04)	0.58 (0.08)	0.60 (0.04)	0.58 (0.07)	0.51 (0.03)
4	0.55 (0.03)	0.52 (0.06)	0.54 (0.03)	0.52 (0.06)	0.56 (0.03)	0.52 (0.06)	0.57 (0.03)
5	0.58 (0.04)	0.55 (0.07)	0.59 (0.04)	0.56 (0.07)	0.58 (0.04)	0.55 (0.07)	0.54 (0.04)
6	0.59 (0.03)	0.58 (0.06)	0.59 (0.03)	0.58 (0.06)	0.59 (0.03)	0.58 (0.07)	0.59 (0.03)
7	0.56 (0.03)	0.54 (0.07)	0.57 (0.04)	0.55 (0.08)	0.57 (0.03)	0.56 (0.07)	0.57 (0.04)
8	0.59 (0.03)	0.55 (0.07)	0.59 (0.03)	0.55 (0.07)	0.58 (0.03)	0.55 (0.06)	0.57 (0.03)
9	0.59 (0.03)	0.56 (0.06)	0.58 (0.03)	0.56 (0.06)	0.60 (0.03)	0.57 (0.07)	0.57 (0.03)
10	0.57 (0.06)	0.55 (0.11)	0.59 (0.06)	0.56 (0.09)	0.58 (0.06)	0.53 (0.10)	0.54 (0.06)
11	0.57 (0.03)	0.53 (0.06)	0.56 (0.03)	0.53 (0.06)	0.56 (0.03)	0.53 (0.07)	0.54 (0.03)

Table 7.6: Average AUC of the classification methods over 100 iterations, with the standard deviation between brackets, for the test set size of 70%.

Prod. nr.	Binary RL-wvrn	Binary RL-cdrn	Maximum RL-wvrn	Maximum RL-cdrn	Average RL-wvrn	Average RL-cdrn	LogReg
1	1.32 (0.26)	1.33 (0.26)	1.31 (0.25)	1.31 (0.25)	1.27 (0.25)	1.28 (0.25)	1.31 (0.35)
2	1.05 (0.13)	0.99 (0.18)	1.03 (0.14)	1.01 (0.14)	1.04 (0.15)	1.00 (0.15)	1.13 (0.19)
3	2.41 (0.54)	2.41 (0.53)	2.48 (0.58)	2.41 (0.59)	2.46 (0.61)	2.45 (0.59)	0.84 (0.41)
4	1.07 (0.14)	1.03 (0.18)	1.10 (0.15)	1.09 (0.16)	1.06 (0.14)	1.04 (0.18)	1.27 (0.23)
5	2.05 (0.53)	2.10 (0.54)	1.91 (0.54)	1.97 (0.51)	2.06 (0.50)	2.11 (0.52)	1.21 (0.45)
6	1.14 (0.16)	1.11 (0.16)	1.18 (0.18)	1.14 (0.17)	1.18 (0.17)	1.12 (0.16)	1.46 (0.28)
7	1.04 (0.08)	1.03 (0.08)	1.03 (0.08)	1.02 (0.08)	1.04 (0.08)	1.02 (0.08)	1.15 (0.14)
8	1.18 (0.10)	1.14 (0.11)	1.20 (0.10)	1.15 (0.10)	1.16 (0.09)	1.13 (0.10)	1.23 (0.19)
9	1.07 (0.14)	1.06 (0.14)	1.07 (0.16)	1.05 (0.15)	1.08 (0.17)	1.09 (0.16)	1.35 (0.26)
10	3.64 (1.17)	2.94 (1.00)	3.74 (1.09)	3.05 (0.99)	3.66 (1.19)	3.09 (1.16)	1.43 (0.84)
11	1.33 (0.21)	1.34 (0.22)	1.30 (0.27)	1.29 (0.30)	1.35 (0.24)	1.34 (0.25)	1.36 (0.37)

Table 7.7: Average top 10% lift of the classification methods over 100 iterations, with the standard deviation between brackets, for the test set size of 30%.

Prod. nr.	Binary RL-wvrn	Binary RL-cdrn	Maximum RL-wvrn	Maximum RL-cdrn	Average RL-wvrn	Average RL-cdrn	LogReg
1	1.29 (0.23)	1.24 (0.30)	1.32 (0.25)	1.26 (0.30)	1.34 (0.24)	1.32 (0.28)	1.30 (0.29)
2	1.03 (0.12)	0.96 (0.18)	1.04 (0.12)	1.00 (0.14)	1.03 (0.12)	0.96 (0.20)	1.10 (0.17)
3	2.22 (0.43)	2.24 (0.54)	2.32 (0.48)	2.31 (0.53)	2.18 (0.51)	2.14 (0.57)	0.90 (0.36)
4	1.06 (0.12)	1.03 (0.18)	1.12 (0.14)	1.06 (0.21)	1.08 (0.12)	0.99 (0.24)	1.25 (0.19)
5	1.87 (0.42)	2.01 (0.44)	1.82 (0.43)	1.87 (0.51)	1.85 (0.44)	1.95 (0.47)	1.09 (0.39)
6	1.15 (0.15)	1.11 (0.19)	1.13 (0.14)	1.12 (0.15)	1.15 (0.15)	1.13 (0.17)	1.36 (0.22)
7	1.05 (0.08)	1.01 (0.13)	1.02 (0.08)	1.00 (0.11)	1.03 (0.08)	1.01 (0.10)	1.13 (0.13)
8	1.15 (0.10)	1.08 (0.16)	1.17 (0.10)	1.10 (0.17)	1.16 (0.12)	1.11 (0.17)	1.19 (0.15)
9	1.07 (0.15)	1.03 (0.20)	1.08 (0.17)	1.04 (0.24)	1.10 (0.14)	1.06 (0.21)	1.29 (0.19)
10	3.08 (0.90)	2.22 (1.02)	3.02 (0.78)	2.28 (1.02)	3.16 (0.90)	2.29 (0.93)	1.32 (0.66)
11	1.31 (0.25)	1.24 (0.36)	1.34 (0.25)	1.29 (0.30)	1.31 (0.22)	1.27 (0.30)	1.29 (0.29)

Table 7.8: Average top 10% lift of the classification methods over 100 iterations, with the standard deviation between brackets, for the test set size of 50%.

Prod. nr.	Binary RL-wvrn	Binary RL-cdrn	Maximum RL-wvrn	Maximum RL-cdrn	Average RL-wvrn	Average RL-cdrn	LogReg
1	1.31 (0.27)	1.24 (0.43)	1.30 (0.26)	1.15 (0.44)	1.35 (0.27)	1.12 (0.46)	1.25 (0.26)
2	1.00 (0.16)	0.92 (0.21)	1.02 (0.15)	0.96 (0.17)	1.00 (0.19)	0.92 (0.21)	1.10 (0.14)
3	1.96 (0.45)	1.81 (0.66)	1.80 (0.45)	1.73 (0.59)	1.86 (0.47)	1.82 (0.61)	0.88 (0.28)
4	1.09 (0.14)	0.96 (0.29)	1.05 (0.15)	0.90 (0.27)	1.07 (0.14)	0.93 (0.27)	1.15 (0.16)
5	1.64 (0.41)	1.58 (0.58)	1.61 (0.34)	1.60 (0.60)	1.57 (0.43)	1.54 (0.59)	1.12 (0.34)
6	1.15 (0.17)	1.11 (0.26)	1.13 (0.17)	1.09 (0.28)	1.15 (0.20)	1.09 (0.29)	1.21 (0.21)
7	1.01 (0.09)	0.94 (0.15)	1.04 (0.09)	0.97 (0.13)	1.04 (0.10)	0.99 (0.14)	1.13 (0.11)
8	1.16 (0.12)	1.04 (0.23)	1.14 (0.13)	1.02 (0.23)	1.14 (0.12)	1.06 (0.21)	1.15 (0.13)
9	1.04 (0.17)	0.98 (0.28)	1.08 (0.18)	1.01 (0.27)	1.11 (0.18)	1.02 (0.28)	1.18 (0.19)
10	2.38 (0.80)	1.63 (1.00)	2.40 (0.81)	1.73 (0.85)	2.52 (0.70)	1.55 (0.92)	1.27 (0.54)
11	1.29 (0.23)	1.18 (0.42)	1.32 (0.23)	1.21 (0.40)	1.33 (0.24)	1.17 (0.46)	1.20 (0.26)

Table 7.9: Average top 10% lift of the classification methods over 100 iterations, with the standard deviation between brackets, for the test set size of 70%.

Chapter 8

Conclusions and future research

This dissertation consists of two main parts. The first part proposes a new theoretical framework for profit-driven classification performance measurement. The second part presents two case studies which illustrate the usefulness of data mining in the industry. Section 8.1 first discusses the main conclusions of this dissertation, after which Section 8.2 will provide some interesting opportunities for future research.

8.1 Conclusions

8.1.1 Profit-driven data analytics

The implications of the findings presented in Part I are straightforward but essential. Companies rely more than ever on data mining techniques to support their decision making processes. When evaluating a classification technique which is to be used in a business context, it is imperative to base any evaluation criterion on the goal of the end user. Since companies strive for profit maximization, a performance measure evidently should take this into account. The very commonly used area under the ROC curve (AUC) does have its merits and an interesting interpretation in the sense that the AUC of a classification method is the probability that a randomly

chosen case will be assigned a lower score than a randomly chosen non-case. However, as Hand (2009) has shown, AUC makes incorrect implicit assumptions about the misclassification costs, and the use of this performance metric in a business environment leads to suboptimal profits. This dissertation outlines a theoretical framework which incorporates all gains and losses related to the employment of a data mining technique, and defines a probabilistic performance measure, the expected maximum profit (EMP). We have also shown the interesting link between EMP and the ROC curve, whereby EMP corresponds to an integration along the ROC curve of a classifier.

As each corporate environment has its own specificities, the framework is defined on a general level. To be applicable to a certain business problem, the particularities of its cost and benefit structure need to be incorporated. This process is worked out in detail for the problem of customer churn and an EMP measure for customer churn, EMP^{ccp}, is derived. Also the link between EMP^{ccp} and the H measure is investigated and it appears that the latter with appropriately chosen distribution parameters is a good approximation of the former. The performance measure for customer churn is validated in an extensive case study. The results clearly indicate that the use of AUC as a performance metric leads to suboptimal profits. The case study also points to one of the major advantages of the EMP^{ccp} measure. It does not only select the classifier which maximizes the profit, it also provides the practitioner with an estimate of the fraction of the customer base which needs to be targeted in the retention campaign. This optimal fraction varies from case to case, and deviating from this fraction again leads to suboptimal profits. Note that the H measure, although it is able to select the most profitable classifier, does not provide guidance on the optimal fraction of the customer base to be included in the retention campaign. Finally, a sensitivity analysis was carried out, to analyze how vulnerable the EMP^{ccp} measure is to incorrect estimation of the fixed parameters CLV, δ, and ϕ. The outcome shows that the EMP^{ccp} measure and its resulting ranking is relatively robust with regard to changes in these parameter values.

Besides the implementation for customer churn prediction, Chapter 4 proposes an EMP measure for consumer credit scoring, EMP^{cs}. This performance measure accounts for the benefits generated by healthy loans and the costs caused by loan defaults (driven by the loss given default and the exposure at default). The EMP^{cs} measure allows for profit-driven model selection and parameter tuning, and thus enables practitioners to pick the credit scoring model which increases profitability most. Furthermore, the EMP^{cs} measure provides information about the optimal fraction of loan applications which should be rejected, from which the optimal cutoff value can be derived. The cutoff value is of crucial importance in an operational context, since it is required to transform the ranking, provided by a classification technique, into an actual division into two groups. By applying the proposed cutoff value, companies are maximizing their profitability. This is illustrated with the case study, which shows that employing the EMP-based cutoff value leads to a higher overall profit, while less loan applicants are rejected, than with traditional approaches (based on AUC and accuracy).

Finally, we have explored the incorporation of the profitability criterion into the model building step in Chapter 5. In this exploratory study, four research hypotheses concerning profit-based model selection and prediction have been tested with a case study in customer churn prediction. The results indicate that a focus on profitability during the model building step – and not only for model selection – provides opportunities for a further increase in profitability. Although this study is based on one type of classification techniques, i.e. the ensemble selection (ES) framework, it shows promising results and encourages a continued research into profit-driven classification models.

8.1.2 Case studies in data mining

The case study concerning the use of Bayesian network (BN) classifiers for customer churn prediction benchmarked a whole range of BN classifiers with traditional logistic regression. The study shows

that BN classifiers perform similarly to logistic regression, but are more comprehensible due to the visualization the Bayesian network provides. This comprehensibility feature enables practitioners to interpret results easier, and encourages the adoption of data mining techniques in the industry.

The case study about social network-based classification algorithms for the prediction of e-commerce adoption also illustrates the usefulness of data mining in a business context. The study shows that companies are able to make meaningful predictions based on social network data only, which is more easily available than the traditional socio-demographic and behavioral data. This opens opportunities for companies to identify potential customers, and to target those customers with tailored marketing campaigns.

8.2 Future research

8.2.1 Profit-driven data analytics

The EMP framework presented in this dissertation provides many opportunities for future research on profit-driven classification. A first very interesting but challenging direction has been suggested in Chapter 5. Part I mainly focuses on profit-based classification *performance measurement*, whereas the exploratory study in Chapter 5 provides evidence that profit-driven *model building* provides further opportunities for profitability enhancement. Currently, the majority of models are trained by optimizing a statistical criterion (such as e.g. the maximum likelihood), whereas we argue that this criterion should be profit-oriented. There has been research into the matter of incorporating profit (or rather cost) into the model training step. E.g. Dmochowski et al. (2010) showed that the weighted maximum likelihood estimation (MLE) is an upper bound of the empirical loss – and it is convex, which makes estimation easier. They argue that the weighted MLE is the preferred cost-sensitive technique. Masnadi-Shirazi et al. (2012), on the other hand, propose a novel

method for training cost-sensitive support vector machines, which are able to take into account example dependent cost. However, even though these techniques are cost-sensitive, they would benefit from a general framework, such as the EMP framework, which enables transferability between different application domains and cost structures. To fulfill this goal, the EMP framework should be linked to the model estimation step of several established techniques. This shift from statistical-based model building to a more profit-oriented approach poses many challenges, but is worthwhile to be more deeply investigated.

Another track for future research concerns the EMP framework and its cost benefit parameters itself. At this stage, the cost and benefit parameters are considered independent and constant over all customers, which is a simplified view. However, there may be several sources of heterogeneity (e.g. customers have different CLV values) and there may exist dependencies among the cost benefit parameters (e.g. the acceptance rate in a retention campaign depends on the nature of the retention offer). Future research should focus on how to deal with these sources of heterogeneity and dependencies, preferably in an abstract and generic way so that it is applicable to all industry settings. Potentially, copulas may be an interesting mathematical construct to introduce these dependencies in a generic way, since they allow to easily model joint distributions by estimating the marginal distribution and the copula separately (Nelsen, 1999). The nature of dependencies between cost benefit parameters can then be modeled in a later stage, when the industry-specific dependencies are known.

A third obvious though not straightforward opportunity for further research consists of the extension of the framework to multi-class problems. There is already an extension of the area under the ROC curve for multi-class problems (Hand and Till, 2001), which could provide inspiration to develop the EMP framework for multi-class problems. Furthermore, we could go a step further and extend the EMP framework to other tasks in predictive analytics, such as e.g. regression, association rule learning or clustering. Although there is

no direct applicability of the EMP framework to these fields, the observed importance of profitability remains equally important for these other types of predictive analytics within a business context.

Finally, besides the application of the EMP framework to customer churn prediction and consumer credit scoring, there remain many business problems where a profit-driven performance measure would add value. Among others, data mining techniques are employed for direct marketing response models (Cui et al., 2006), fraud detection (Viaene et al., 2002), and viral marketing in social networks (Domingos, 2005). Basically, the theoretical EMP framework provides guidance for any application of classification models within a business context, as long as the involved costs and benefits are carefully identified.

8.2.2 Case studies in data mining

The study on social network-based classification for the prediction of e-commerce also provides interesting opportunities for future research. The current study is based on a survey among students, and consequently does not reflect the entire population. A follow up study could investigate whether the same conclusions can be drawn for other subsets of the population. Likewise, we only consider few products per product category. In order to draw valid conclusions about entire product categories, it would be good to focus future studies on many products within individual product groups.

In addition, we used products in the survey that are likely to differ in terms of risk that is associated with the purchase, product knowledge levels that are possible, the moment financial risk is taken when reserving a service, and so on. However, we did not measure the respondents' perception of these factors. Future research could try to relate the performance of the network model to the variability in risk perception, knowledge level, etc. among connected nodes in the network.

Another possible extension of the research involves investigating other steps of the procurement process, such as the step where

product information is gathered and the step where one is searching for a supplier. Again one could investigate what media are considered appropriate to find product information or to find a supplier and check whether network effects play a role in that step.

List of Figures

List of Tables

Bibliography

Ali, S., Smith, K., 2006. On learning algorithm selection for classification. Applied Soft Computing 6 (2), 119–138.

Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X., 2010a. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. The Journal of Machine Learning Research 11, 171–234.

Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X., 2010b. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part II: Analysis and Extensions. The Journal of Machine Learning Research 11, 235–284.

Aliferis, C. F., Tsamardinos, I., Statnikov, A., 2003. Causal explorer: Causal probabilistic network learning toolkit for biomedical discovery.
URL `http://www.dsl-lab.org/causal_explorer`

Allen, L., DeLong, G., Saunders, A., 2004. Issues in the credit risk modeling of retail markets. Journal of Banking & Finance 28 (4), 727–752.

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

Au, W., Chan, K., Yao, X., 2003. A novel evolutionary data min-

ing algorithm with applications to churn prediction. IEEE Transactions on Evolutionary Computation 7 (6), 532–545.

Baesens, B., Setiono, R., Mues, C., Vanthienen, J., 2003. Using neural network rule extraction and decision tables for credit-risk evaluation. Management Science 49 (3), pp. 312–329.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16 (5), 412.

Bellman, S., Lohse, G. L., Johnson, E. J., 1999. Predictors of online buying behavior. Communications of the ACM 42 (12), 32–38.

Bernstein, A., Provost, F., Hill, S., 2005. Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. Knowledge and Data Engineering, IEEE Transactions on 17 (4), 503–518.

Bhattacharya, C., 1998. When customers are members: Customer retention in paid membership contexts. Journal of the Academy of Marketing Science 26 (1), 31–44.

Blau, P. M., 1977. Inequality and heterogeneity: A primitive theory of social structure. Free Press New York.

Bravo, C., Maldonado, S., Weber, R., 2012. Granting and managing loans for micro-entrepreneurs: New developments and practical experiences. European Journal of Operational Research Available online, 10 November 2012.

Broverman, S. A., 2010. Mathematics of investment and credit. Actex Publications.

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

Burez, J., Van den Poel, D., 2009. Handling class imbalance in customer churn prediction. Expert Systems with Applications 36 (3), 4626–4636.

Burt, R. S., 1987. Social contagion and innovation: Cohesion versus structural equivalence. American journal of Sociology 92, 1287–1335.

Burt, R. S., Doreian, P., 1982. Testing a structural model of perception: conformity and deviance with respect to journal norms in elite sociological methodology. Quality & Quantity 16 (2), 109–150.

Caruana, R., Munson, A., Niculescu-Mizil, A., 2006. Getting the most out of ensemble selection. In: Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE, pp. 828–833.

Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A., 2004. Ensemble selection from libraries of models. In: Proceedings of the twenty-first international conference on Machine learning. ACM, p. 18.

Chakrabarti, S., Dom, B., Indyk, P., 1998. Enhanced hypertext categorization using hyperlinks. In: ACM SIGMOD Record. Vol. 27. ACM, pp. 307–318.

Chawla, N., 2005. The Data Mining and Knowledge Discovery Handbook. Springer, Ch. Data mining for imbalanced datasets: An overview, pp. 853–867.

Cheng, J., 2000. Powerpredictor system.
URL `http://www.cs.ualberta.ca/~jcheng/bnpp.htm`

Cheng, J., Bell, D., 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, U.S., pp. 83–90.

Cheng, J., 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., Greiner, R., Kelly, J., D. Bell, W. L., 2002. Learning Bayesian networks from data: An information-theory based approach. Artificial Intelligence 137, 43–90.

Chiang, W. K., Zhang, D., Zhou, L., 2006. Predicting and explaining patronage behavior toward web and traditional stores using neural networks: a comparative analysis with logistic regression. Decision Support Systems 41 (2), 514–531.

Chickering, D. M., 1996. Learning Bayesian networks is NP-complete. In: Learning from Data: Artificial Intelligence and Statistics. Springer-Verlag, pp. 121–130.

Chickering, D. M., 2002. Optimal structure identification with greedy search. Journal of Machine Learning Research 3, 507–554.

Chow, C., Liu, C., 1968. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory 14, 462–467.

Colgate, M., Danaher, P., 2000. Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution. Journal of the Academy of Marketing Science 28 (3), 375–387.

Colgate, M., Stewart, K., Kinsella, R., 1996. Customer defection: A study of the student market in ireland. International Journal of Bank Marketing 14 (3), 23–29.

Collopy, F., Adya, M., Armstrong, J. S., 1994. Research report principles for examining predictive validity: The case of information systems spending forecasts. Information Systems Research 5 (2), 170–179.

Cooper, G., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9, 309–347.

Cui, G., Wong, M., Lui, H., 2006. Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. Management Science 52 (4), 597.

D. Compeau, B. Marcolin, H. K. C. H., 2012. Generalizability of information systems research using student subjects. Information Systems Research 23 (4), 1093–1109.

Datta, P., Masand, B., Mani, D., Li, B., 2000. Automated cellular modeling and prediction on a large scale. Artificial Intelligence Review 14, 485–502.

Dejaeger, K., Verbraken, T., Baesens, B., 2012. Assessing Bayesian network classifiers for software defect prediction. In: EURO 2012 Conference. Vilnius (Latvia) 9-11 July 2012.

Dejaeger, K., Verbraken, T., Baesens, B., 2013. Towards comprehensible software fault prediction models using Bayesian network classifiers. IEEE Transactions on Software Engineering 39 (2), 237–257.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30.

Dennis, A. R., Wixom, B. H., Vandenberg, R. J., 2001. Understanding fit and appropriation effects in group support systems via meta-analysis. MIS Quarterly, 167–193.

Dierkes, T., Bichler, M., Krishnan, R., 2011. Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with Markov logic networks. Decision Support Systems 51 (3), 361–371.

Dmochowski, J. P., Sajda, P., Parra, L. C., 2010. Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds. The Journal of Machine Learning Research 9999, 3313–3332.

Domingos, P., 1999. Metacost: A general method for making classifiers cost-sensitive. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 155–164.

Domingos, P., 2005. Mining social networks for viral marketing. IEEE Intelligent Systems 20 (1), 80–82.

Duda, R., Hart, P., 1973. Pattern classification and scene analysis. John Wiley, New York.

Dunn, O., 1961. Multiple comparisons among means. Journal of the American Statistical Association 56, 52–64.

Edelberg, W., 2006. Risk-based pricing of interest rates for consumer loans. Journal of Monetary Economics 53 (8), 2283 – 2298.

Elkan, C., 2001. The foundations of cost-sensitive learning. In: International Joint Conference on Artificial Intelligence. Vol. 17. pp. 973–978.

Famili, A., Shen, W., Weber, R., Simoudis, E., 2010. Data Preprocessing and Intelligent Data Analysis. International Journal on Intelligent Data Analysis 1 (1), 1–2.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters 27 (8), 861–874.

Fayyad, U., Irani, K., 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI). Morgan Kaufmann, Chambéry, France, pp. 1022–1029.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery in databases. AI magazine 17 (3), 37.

Finlay, S., 2010. Credit scoring for profitability objectives. European Journal of Operational Research 202, 528–537.

Friedman, M., 1940. A comparison of alternative tests of significance for the problem of m rankings. Annals of Mathematical Statistics 11, 86–92.

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Machine Learning 29, 131–163.

Friedman, N., Nachman, I., Peer, D., 1999. Learning Bayesian network structure from massive datasets: The sparse candidate algorithm. In: Fifteenth Conference on Uncertainty in Artificial Intelligence.

Ganesh, J., Arnold, M., Reynolds, K., 2000. Understanding the customer base of service providers: An examination of the differences between switchers and stayers. Journal of Marketing 64 (3), 65–87.

Geiger, D., Verma, T., Pearl, J., 1990. Identifying independence in Bayesian networks. Networks 20 (5), 507–534.

Germann, F., Lilien, G. L., Rangaswamy, A., 2013. Performance implications of deploying marketing analytics. International Journal of Research in Marketing 30 (2), 114 – 128.

Glady, N., Baesens, B., Croux, C., 2009. Modeling churn using customer lifetime value. European Journal of Operational Research 197 (1), 402–411.

Goethals, F., 2012. Perceived social influence in watching online theory presentations. In: Workshops on Business Informatics Research. Springer, pp. 130–142.

Gupta, A., Su, B.-C., Walter, Z., 2004. Risk profile and consumer shopping behavior in electronic and traditional channels. Decision Support Systems 38 (3), 347–367.

Gupta, S., Zeithaml, V., 2006. Customer metrics and their impact on financial performance. Marketing Science 25 (6), 718–739.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. The Journal of Machine Learning Research 3, 1157–1182.

Guyon, I., Lemaire, V., Boullé, M., Dror, G., Vogel, D., 2010. Design and analysis of the KDD cup 2009: fast scoring on a large orange customer database. ACM SIGKDD Explorations Newsletter 11 (2), 68–76.

Hand, D., 2005. Good practice in retail credit scorecard assessment. Journal of the Operational Research Society 56 (9), 1109–1117.

Hand, D., 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning 77 (1), 103–123.

Hand, D. J., Till, R. J., 2001. A simple generalisation of the area under the roc curve for multiple class classification problems. Machine Learning 45 (2), 171–186.

Hannah, B., Lybecker, K. M., 2010. Determinants of online purchasing and the percentage of income spent online. International Business Research 3 (4), 60–71.

Harkola, J., Greve, A., 1995. Diffusion of technology: Cohesion or structural equivalence? In: Academy of Management Proceedings. Vol. 1995. Academy of Management, pp. 422–426.

Hastie, T., Tibshirani, R., Friedman, J. J. H., 2001. The elements of statistical learning. Vol. 1. Springer New York.

Heckerman, D., Geiger, D., Chickering, D., 1995. Learning bayesian networks: The combination of knowledge and statitical data. Machine Learning 20, 194–243.

Heckerman, D., Meek, C., Cooper, G., 1999. A Bayesian approach to causal discovery. In: Computation, Causation and Discovery. MIT Press, pp. 141–165.

Hernández-Orallo, J., Flach, P., Ferri, C., 2012. A unified view of performance metrics: Translating threshold choice into expected classification loss. Journal of Machine Learning Research 13, 2813–2869.

Hilbert, M., López, P., 2011. The world's technological capacity to store, communicate, and compute information. Science 332 (6025), 60–65.

Hoch, S. J., Schkade, D. A., 1996. A psychological approach to decision support systems. Management Science 42 (1), 51–64.

Huang, Z., Chung, W., Chen, H., 2004. A graph model for e-commerce recommender systems. Journal of the American Society for information science and technology 55 (3), 259–274.

Hung, S., Yen, D., Wang, H., 2006. Applying data mining to telecom churn management. Expert Systems with Applications 31, 515–524.

Hur, J., Kim, J., 2008. A hybrid classification method using error pattern modeling. Expert Systems with Applications 34 (1), 231–241.

Jackson, C. M., Chow, S., Leitch, R. A., 1997. Toward an understanding of the behavioral intention to use an information system. Decision sciences 28 (2), 357–389.

John, G., Langley, P., 1995. Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the Eleventh Confer-

ence on Uncertainty in Artificial Intelligence (UAI). Morgan Kaufmann, Montreal, Québec, Canada, pp. 338–345.

Kannan, P., Pope, B. K., Jain, S., 2009. Practice prize winner – pricing digital content product lines: A model and application for the national academies press. Marketing Science 28 (4), 620–636.

Kiss, C., Bichler, M., 2008. Identification of influencers – measuring influence in customer networks. Decision Support Systems 46 (1), 233–253.

Kononenko, I., 1991. Semi-naive Bayesian classifier. In: Kodratoff, Y. (Ed.), Proceedings Sixth European Working Session on Learning. Berlin: Springer Verlag, pp. 206–219.

Kotsiantis, S. B., 2007. Supervised machine learning: A review of classification techniques. Informatica 31 (3), 249 – 268.

Krzanowski, W., Hand, D., 2009. ROC curves for continuous data. CRC/Chapman and Hall.

Kwak, H., Fox, R. J., Zinkhan, G. M., 2002. What products can be successfully promoted and sold via the internet? Journal of Advertising Research 42 (1), 23–38.

Langley, P., Iba, W., Thompson, K., 1992. An analysis of Bayesian classifiers. In: Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI'92). AAAI Press, San Jose, CA, U.S., pp. 223–228.

Langley, P., Sage, S., 1994. Induction of selective Bayesian classifiers. In: R. Lopez de Mantaras, D. P. (Ed.), Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence. Advances in Data Mining in Marketing. San Francisco CA: Morgan Kaufmann, pp. 399–406.

Lauritzen, S., 1996. Graphical models. Oxford: Clarendon Press.

Lazarsfeld, P. F., Merton, R. K., et al., 1954. Friendship as a social process: A substantive and methodological analysis. Freedom and control in modern society 18 (1), 18–66.

Lehmann, E. L., D'Abrera, H. J., 2006. Nonparametrics: Statistical methods based on ranks. Springer New York.

Lemmens, A., Croux, C., 2006. Bagging and boosting classification trees to predict churn. Journal of Marketing Research 43 (2), 276–286.

Liang, T.-P., Huang, J.-S., 1998. An empirical study on consumer acceptance of products in electronic markets: A transaction cost model. Decision support systems 24 (1), 29–43.

Lieber, E., Syverson, C., 2011. Online versus offline competition. The Oxford Handbook of the Digital Economy, 189.

Lima, E., Mues, C., Baesens, B., 2009. Domain knowledge integration in data mining using decision tables: Case studies in churn prediction. Journal of the Operational Research Society 60 (8), 1096–1106.

Lodish, L. M., Curtis, E., Ness, M., Simpson, M. K., 1988. Sales force sizing and deployment using a decision calculus model at syntex laboratories. Interfaces 18 (1), 5–20.

Macskassy, S. A., Provost, F., 2007. Classification in networked data: A toolkit and a univariate case study. The Journal of Machine Learning Research 8, 935–983.

Martens, D., 2008. Building acceptable classification models for financial engineering applications. Ph.D. thesis, KU Leuven.

Martens, D., Baesens, B., Van Gestel, T., Vanthienen, J., 2007. Comprehensible credit scoring models using rule extraction from support vector machines. European Journal of Operational Research 183 (3), 1466 – 1476.

Martens, D., Vanthienen, J., Verbeke, W., Baesens, B., 2011. Performance of classification models from a user perspective. Decision Support Systems 51 (4), 782–793.

Masnadi-Shirazi, H., Vasconcelos, N., Iranmehr, A., 2012. Cost-sensitive support vector machines.

Mays, E., Nuetzel, P., 2004. Credit Scoring for Risk Managers: The Handbook for Lenders. South-Western Publishing, Mason, OH, Ch. Scorecard Monitoring Reports, pp. 201–217.

McIntyre, S. H., 1982. An experimental study of the impact of judgment-based marketing models. Management Science 28 (1), 17–33.

McKinsey & Co, 2009. Mckinsey global survey results: Measuring marketing.

McPherson, M., Smith-Lovin, L., Cook, J. M., 2001. Birds of a feather: Homophily in social networks. Annual review of sociology, 415–444.

Mizerski, R., 1982. An attribution explanation of the disproportionate influence of unfavourable information. Journal of Consumer Research 9 (December), 301–310.

Mozer, M., Wolniewicz, R., Grimes, D., Johnson, E., Kaushansky, H., 2000. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. IEEE Transactions on Neural Networks 11 (3), 690–696.

Mulpuru, S., Sehgal, V., Evans, P. F., Hoar, A., Roberge, D., 2012. US online retail forecast, 2011 to 2016. Forrester Research, February.

Murphy, K., 2001. The Bayes net matlab toolbox.
URL `http://code.google.com/p/bnt/`

Natter, M., Mild, A., Wagner, U., Taudes, A., 2008. Practice prize report – planning new tariffs at tele. ring: The application and impact of an integrated segmentation, targeting, and positioning tool. Marketing Science 27 (4), 600–609.

Nazir, S., Tayyab, A., Sajid, A., ur Rashid, H., Javed, I., 2012. How online shopping is affecting consumers buying behavior in Pakistan? International Journal of Computer Science Issues 9 (1), 1694–0814.

Nelsen, R. B., 1999. An introduction to copulas. Springer.

Nemenyi, P. B., 1963. Distribution-free multiple comparisons. Ph.D. thesis, Princeton University.

Neslin, S., Gupta, S., Kamakura, W., Lu, J., Mason, C., 2006. Detection defection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research 43 (2), 204–211.

Padmanabhan, B., Tuzhilin, A., 2003. On the use of optimization for data mining: Theoretical interactions and eCRM opportunities. Management Science 49 (10), 1327–1343.

Paulin, M., Perrien, J., Ferguson, R., Salazar, A., Seruya, L., 1998. Relational norms and client retention: External effectiveness of commercial banking in Canada and Mexico. International Journal of Bank Marketing 16 (1), 24–31.

Pearl, J., 1988. Probabilistic reasoning in Intelligent Systems: Networks for Plausible Inference. Morgan Kaufmann.

Perlich, C., Provost, F., 2003. Aggregation-based feature invention and relational concept classes. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 167–176.

Piatetsky-Shapiro, G., Masand, B., 1999. Estimating campaign benefits and modeling lift. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 185–193.

Prati, R. C., Batista, G., Monard, M. C., 2011. A survey on graphical methods for classification predictive performance evaluation. Knowledge and Data Engineering, IEEE Transactions on 23 (11), 1601–1618.

Provost, F., Fawcett, T., 2001. Robust classification for imprecise environments. Machine Learning 42 (3), 203–231.

Rasmusson, E., 1999. Complaints can build relationships. Sales and Marketing Management 151 (9), 89–90.

Reichheld, F., 1996. Learning from customer defections. Harvard Business Review 74 (2), 56–69.

Richa, D., 2012. Impact of demographic factors of consumers on online shopping behavior. International Journal of Engineering and Management Sciences 3 (1), 43–52.

Rijsbergen, C. J. V., 1979. Information Retrieval, 2nd Edition. Butterworth-Heinemann, Newton, MA, USA.

Rocchio, J. J., 1971. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, Englewood Cliffs NJ, Ch. Relevance feedback in information retrieval, pp. 313–323.

Rust, R., Zahorik, A., 1993. Customer satisfaction, customer retention, and market share. Journal of Retailing 69 (2), 193–215.

Sacha, J., 1999a. Bayesian network classifier toolbox. URL `http://jbnc.sourceforge.net/`

Sacha, J., 1999b. New synthesis of Bayesian network classifiers and cardiac spect image interpretation. Ph.D. thesis, University of Toledo.

Seret, A., Verbraken, T., Baesens, B., 2013. A new knowledge-based constrained clustering approach: Theory and application in direct marketing. Applied Soft Computing (under review).

Seret, A., Verbraken, T., Versailles, S., Baesens, B., 2012. A new SOM-based method for profile generation: Theory and an application in direct marketing. European Journal of Operational Research 220 (1), 199–209.

Shmueli, G., Koppius, O. R., 2011. Predictive analytics in information systems research. MIS Quarterly 35 (3), 553–572.

Silva-Risso, J. M., Bucklin, R. E., Morrison, D. G., 1999. A decision support system for planning manufacturers' sales promotion calendars. Marketing Science 18 (3), 274–300.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45 (4), 427–437.

Somers, M., Whittaker, J., 2007. Quantile regression for modelling distributions of profit and loss. European journal of operational research 183 (3), 1477–1487.

Spirtes, P., Glymour, C., Scheines, R., 2000. Causation, prediction, and search. The MIT Press.

Stum, D., Thiry, A., 1991. Building customer loyalty. Training and Development Journal 45 (4), 34–36.

Sykes, T. A., Venkatesh, V., Gosain, S., 2009. Model of acceptance with peer support: A social network perspective to understand employees' system use. MIS Quarterly 33 (2), 371.

Szymanski, D. M., Hise, R. T., 2000. E-satisfaction: An initial examination. Journal of retailing 76 (3), 309–322.

Tan, P., Steinbach, M., Kumar, V., 2006. Introduction to Data Mining. Pearson Education, Boston, MA.

Thomas, L. C., 2000. A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting 16(2), 149–172.

Thomas, L. C., 2009. Consumer Credit Models: Pricing, Profit and Portfolios. Oxford University Press, New York.

Thomas, L. C., Crook, J. N., Edelman, D. B., 2002. Credit Scoring and its Applications. SIAM.

Tsamardinos, I., Brown, L. E., Aliferis, C. F., 2006. The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning 65 (1), 31–78.

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.

Van der Vaart, A., 2000. Asymptotic Statistics. Cambridge Univ Pr.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B., 2012. New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach. European Journal of Operational Research 218 (1), 211–229.

Verbeke, W., Dejaeger, K., Verbraken, T., Martens, D., Baesens, B., 2011a. Mining social networks for customer churn prediction. In: Interdisciplinary Workshop on Information and Decision in Social Networks. Cambridge (US), 31 May - 1 June 2011.

Verbeke, W., Martens, D., Mues, C., Baesens, B., 2011b. Building comprehensible customer churn prediction models with advanced

rule induction techniques. Expert Systems with Applications 38, 2354–2364.

Verbeke, W., Verbraken, T., Martens, D., Baesens, B., 2011c. Relational learning for customer churn prediction: The complementarity of networked and non-networked classifiers. In: Conference on the Analysis of Mobile Phone Datasets and Networks. Cambridge (US), 10–11 October 2011.

Verbraken, T., Bravo, C., Weber, R., Baesens, B., 2013a. Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research (under review).

Verbraken, T., Goethals, F., Verbeke, W., Baesens, B., 2012a. Using social network classifiers for predicting e-commerce adoption. In: E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life. The Tenth Workshop on E-Business (WEB2011). Springer, Shanghai (China), 4 December 2011, pp. 9–21.

Verbraken, T., Goethals, F., Verbeke, W., Baesens, B., 2013b. Predicting online channel acceptance with social network data. Decision Support Systems (accepted for publication).

Verbraken, T., Lessmann, S., Baesens, B., 2012b. Toward profit-driven churn modeling with predictive marketing analytics. In: Cloud computing and analytics: innovations in e-business services. The Eleventh Workshop on E-Business (WEB2012). Orlando (US), 15 December 2012 (accepted).

Verbraken, T., Van Vlasselaer, V., Verbeke, W., Martens, D., Baesens, B., 2013c. Advanced Database Marketing: Innovative Methodologies and Applications for Managing Customer Relationships. Gower Publishing, Ch. Advanced rule base learning: Active learning, rule extraction, and incorporating domain knowledge, pp. 145–163.

Verbraken, T., Verbeke, W., Baesens, B., 2013d. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE Transactions on Knowledge and Data Engineering 25 (5), 961–973.

Verbraken, T., Verbeke, W., Baesens, B., 2013e. Profit optimizing customer churn prediction with Bayesian network classifiers. Intelligent Data Analysis (accepted for publication).

Verma, T., Pearl, J., 1988. Causal networks: Semantics and expressiveness. In: Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence. Mountain View, CA, U.S., pp. 352–359.

Viaene, S., Derrig, R. A., Baesens, B., Dedene, G., 2002. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. The Journal of Risk and Insurance 69 (3), pp. 373–421.

Wei, C., Chiu, I., 2002. Turning telecommunications call details to churn prediction: A data mining approach. Expert Systems with Applications 23, 103–112.

Witten, I. H., Frank, E., 2000. Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Zeithaml, V., Berry, L., Parasuraman, A., 1996. The behavioural consequences of service quality. Journal of Marketing 60 (2), 31–46.

Zhou, Z.-H., Liu, X.-Y., 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. Knowledge and Data Engineering, IEEE Transactions on 18 (1), 63–77.

Zoltners, A. A., Sinha, P., 2005. The 2004 ISMS practice prize winner – sales territory design: Thirty years of modeling and implementation. Marketing Science 24 (3), 313–331.

Publication list

Articles in internationally reviewed scientific journals

- Verbraken, T., Verbeke, W., Baesens, B., 2013d. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE Transactions on Knowledge and Data Engineering 25 (5), 961–973

- Verbraken, T., Verbeke, W., Baesens, B., 2013e. Profit optimizing customer churn prediction with Bayesian network classifiers. Intelligent Data Analysis (accepted for publication)

- Verbraken, T., Goethals, F., Verbeke, W., Baesens, B., 2013b. Predicting online channel acceptance with social network data. Decision Support Systems (accepted for publication)

- Dejaeger, K., Verbraken, T., Baesens, B., 2013. Towards comprehensible software fault prediction models using Bayesian network classifiers. IEEE Transactions on Software Engineering 39 (2), 237–257

- Seret, A., Verbraken, T., Versailles, S., Baesens, B., 2012. A new SOM-based method for profile generation: Theory and an application in direct marketing. European Journal of Operational Research 220 (1), 199–209

Articles in academic book, internationally recognized scientific publisher

- Verbraken, T., Van Vlasselaer, V., Verbeke, W., Martens, D., Baesens, B., 2013c. Advanced Database Marketing: Innovative Methodologies and Applications for Managing Customer Relationships. Gower Publishing, Ch. Advanced rule base learning: Active learning, rule extraction, and incorporating domain knowledge, pp. 145–163

Papers at international conferences and symposia, published in full in proceedings

- Verbraken, T., Lessmann, S., Baesens, B., 2012b. Toward profit-driven churn modeling with predictive marketing analytics. In: Cloud computing and analytics: innovations in e-business services. The Eleventh Workshop on E-Business (WEB2012). Orlando (US), 15 December 2012 (accepted)

- Verbraken, T., Goethals, F., Verbeke, W., Baesens, B., 2012a. Using social network classifiers for predicting e-commerce adoption. In: E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life. The Tenth Workshop on E-Business (WEB2011). Springer, Shanghai (China), 4 December 2011, pp. 9–21

Meeting abstracts, presented at international conferences and symposia, published or not published in proceedings or journals

- Dejaeger, K., Verbraken, T., Baesens, B., 2012. Assessing Bayesian network classifiers for software defect prediction. In: EURO 2012 Conference. Vilnius (Latvia) 9-11 July 2012

- Verbeke, W., Verbraken, T., Martens, D., Baesens, B., 2011c. Relational learning for customer churn prediction: The complementarity of networked and non-networked classifiers. In: Conference on the Analysis of Mobile Phone Datasets and Networks. Cambridge (US), 10–11 October 2011

- Verbeke, W., Dejaeger, K., Verbraken, T., Martens, D., Baesens, B., 2011a. Mining social networks for customer churn prediction. In: Interdisciplinary Workshop on Information and Decision in Social Networks. Cambridge (US), 31 May - 1 June 2011

Articles submitted for publication in internationally reviewed scientific journals

- Verbraken, T., Bravo, C., Weber, R., Baesens, B., 2013a. Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research (under review)

- Seret, A., Verbraken, T., Baesens, B., 2013. A new knowledge-based constrained clustering approach: Theory and application in direct marketing. Applied Soft Computing (under review)

Doctoral dissertations from the Faculty of Business and Economics

A full list of the doctoral dissertations from the Faculty of Business and Economics can be found at:

www.kuleuven.ac.be/doctoraatsverdediging/archief.htm.