New insights into churn prediction in the telecommunication sector: A profit driven data mining approach

Article *in* European Journal of Operational Research · April 2012

DOI: 10.1016/j.ejor.2011.09.031 · Source: DBLP

CITATIONS
218

READS
2,556

5 authors, including:

Wouter Verbeke
Vrije Universiteit Brussel
63 PUBLICATIONS **1,216** CITATIONS
SEE PROFILE

Karel Dejaeger
KU Leuven
15 PUBLICATIONS **871** CITATIONS
SEE PROFILE

David Martens
University of Antwerp
99 PUBLICATIONS **3,687** CITATIONS
SEE PROFILE

Joon Hur
Universal Training Solution
3 PUBLICATIONS **239** CITATIONS
SEE PROFILE

European Journal of Operational Research 218 (2012) 211–229

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Computational Intelligence and Information Management

New insights into churn prediction in the telecommunication sector: A profit driven data mining approach

Wouter Verbeke [a,*], Karel Dejaeger [a], David Martens [b], Joon Hur [c], Bart Baesens [a,d,e]

[a] Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[b] Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium
[c] Consulting Department SPSS Korea, Samjoung Building 701-2, Level 3, Yuksam-Dong, Kangnam-Gu, Seoul, Republic of Korea
[d] School of Management, University of Southampton, Highfield Southampton, SO17 1BJ, United Kingdom
[e] Vlerick Leuven Gent Management School, Reep 1, B-9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 6 April 2011
Accepted 12 September 2011
Available online 25 September 2011

Keywords:
Data mining
Churn prediction
Profit
Input selection
Oversampling
Telecommunication sector

ABSTRACT

Customer churn prediction models aim to indicate the customers with the highest propensity to attrite, allowing to improve the efficiency of customer retention campaigns and to reduce the costs associated with churn. Although cost reduction is their prime objective, churn prediction models are typically evaluated using statistically based performance measures, resulting in suboptimal model selection. Therefore, in the first part of this paper, a novel, profit centric performance measure is developed, by calculating the maximum profit that can be generated by including the optimal fraction of customers with the highest predicted probabilities to attrite in a retention campaign. The novel measure selects the optimal model and fraction of customers to include, yielding a significant increase in profits compared to statistical measures.

In the second part an extensive benchmarking experiment is conducted, evaluating various classification techniques applied on eleven real-life data sets from telecom operators worldwide by using both the profit centric and statistically based performance measures. The experimental results show that a small number of variables suffices to predict churn with high accuracy, and that oversampling generally does not improve the performance significantly. Finally, a large group of classifiers is found to yield comparable performance.

1. Introduction

1.1. Customer churn prediction

During the last decade, the number of mobile phone users has increased dramatically. At the end of 2009 the number of mobile phone users worldwide has exceeded four billion,[1] which is over 60% of the world population. Wireless telecommunication markets are getting saturated, particularly in the developed countries, and mobile phone penetration rates are stagnating. Many Western countries already have mobile phone penetration rates above 100%, meaning there are more subscriptions than inhabitants. Therefore, customer retention receives a growing amount of attention from telecom operators. Moreover, it has been shown in the literature that customer retention is profitable to a company because: (1) acquiring new clients costs five to six times more than retaining existing customers (Bhattacharya, 1998; Rasmusson, 1999; Colgate et al., 1996; Athanassopoulos, 2000); (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 (Mizerski, 1982; Stum and Thiry, 1991; Reichheld, 1996; Zeithaml et al., 1996; Paulin et al., 1998; Ganesh et al., 2000); (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 Lariviere, 2004).

Most wireless telecom providers already use a customer churn prediction model that indicates the customers with the highest propensity to attrite. This allows an efficient customer management, and a better allocation of the limited marketing resources for customer retention campaigns. Customer churn prediction models are typically applied in contractual settings, such as the postpaid segment in the wireless telecom industry. For these customers usually more information is at hand than in noncontractual settings, like for instance the prepaid segment which consists mostly of anonymous customers. Various types of information can be used to predict customer attrition, such as for instance

* Corresponding author. Tel.: +32 16 32 68 87; fax: +32 16 32 66 24.
E-mail addresses: wouter.verbeke@econ.kuleuven.be (W. Verbeke), karel.dejaeger@econ.kuleuven.be (K. Dejaeger), david.martens@ua.ac.be (D. Martens), hoh@spss.co.kr (J. Hur), bart.baesens@econ.kuleuven.be (B. Baesens).

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

0377-2217/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2011.09.031

socio-demographic data (e.g. sex, age, or zip code) and call behavior statistics (e.g. the number of international calls, billing information, or the number of calls to the customer helpdesk). Alternatively, social network information extracted from call detail records can be explored to predict churn (Dasgupta et al., 2008), which is especially interesting if no other information is available.

1.2. Benchmarking classification techniques for customer churn prediction

In this paper, we study the performance of various state-of-the-art data mining classification algorithms applied to eleven real-life churn prediction data sets from wireless telecom operators around the world. Techniques that are implemented comprise rule based classifiers (Ripper, PART), decision tree approaches (C4.5, CART, Alternating Decision Trees), neural networks (Multilayer Perceptron, Radial Basis Function Network), nearest neighbor (kNN), ensemble methods (Random Forests, Logistic Model Tree, Bagging, Boosting), and classic statistical methods (logistic regression, Naive Bayes, Bayesian Networks). Furthermore, the power and usefulness of the support vector machine (SVM) and the least squares support vector machine (LSSVM) classifiers have not yet been thoroughly investigated in the context of customer churn prediction, and are therefore applied using both linear and radial basis function kernels. Finally, also data preprocessing techniques such as variable selection and oversampling can have a significant impact on the final performance of the model, and will therefore be tested in the benchmarking experiments.

The performance of a classification model is usually evaluated in the literature in terms of the *area under the receiver operating curve* (AUC), which basically represents the behavior of a classifier without regard to class distribution or misclassification costs. However, since only a small fraction of the customers can be included in a retention campaign, a customer churn prediction model is typically evaluated using top decile lift instead of AUC, which only takes into account the performance of the model regarding the top 10% of customers with the highest predicted probabilities to attrite. However, as indicated in Section 3 and demonstrated by the results of the benchmarking experiment in Section 6, from a profit centric point of view using the top decile lift (or the lift at any other cut-off fraction for that matter) results in a suboptimal model selection. Therefore a novel, profit centric, performance measure is introduced, i.e. the maximum profit criterion, which calculates the profit generated by a model when including the optimal fraction of top-ranked customers in a retention campaign. The results of the benchmarking study will be evaluated using both *statistical* performance measures, such as AUC and top decile lift, as well as the newly developed *profit centric* performance measure, which allows to compare both approaches and demonstrate the merits of the newly proposed criterion. Finally, all the experimental results will be rigorously tested using the appropriate test statistics, following a procedure described by Demšar (2006).

The main contributions of this paper lie in the development of a novel, profit centric approach to (1) evaluate and (2) deploy a customer churn prediction model, by (1) calculating the maximum profit that can be generated using the predictions of the model and by (2) including the optimal fraction of customers in a retention campaign. The results of an extensive benchmarking experiment show that both optimizing the included fraction of customers and applying the maximum profit criterion to select a classification model yield significant cost savings. Finally, a number of key recommendations are formulated based on the experimental results regarding both the technical and managerial side of the customer churn prediction modeling process.

The remainder of this paper is structured as follows. The next section provides a brief introduction to customer churn prediction modeling. Then, in Section 3 the maximum profit criterion to evaluate customer churn prediction models is developed, based on a formula to calculate the profits generated by a retention campaign introduced by Neslin et al. (2006). Next, Section 4 defines the experimental design of the benchmarking experiment, and provides an overview of the state-of-the-art classification techniques that are included in the experiment. Also input selection and oversampling for churn prediction are discussed. Section 5 describes the procedure to test the results of the experiments in a statistically sound and appropriate way. Also a brief review is provided of the *statistical* (as opposed to profit centric) performance measures. Section 6 then presents the empirical findings of the experiments, and compares the results of the maximum profit and statistical performance measures. Finally, the last section concludes the paper with a number of managerial and technical recommendations regarding churn prediction modeling, and identifies some interesting issues for future research.

2. Customer churn prediction modeling

Customer churn prediction is a management science problem for which typically a data mining approach is adopted. Data mining is the process of automatically discovering useful information in large data repositories (Tan et al., 2006). Data mining is an integral part of knowledge discovery in databases (KDD), which entails the extraction of valuable information from raw data (Fayyad et al., 1996). Based on historical data a model can be trained to classify customers as future churners or non-churners. Numerous classification techniques have been adopted for churn prediction, including traditional statistical methods such as logistic regression (Lemmens and Croux, 2006; Neslin et al., 2006; Burez and Van den Poel, 2009), non-parametric statistical models like for instance k-nearest neighbor (Datta et al., 2000), decision trees (Wei and Chiu, 2002; Lima et al., 2009), and neural networks (Au et al., 2003; Hung et al., 2006). Often conflicts arise when comparing the conclusions of some of these studies. For instance, Mozer et al. (2000) found that neural networks performed significantly better than logistic regression for predicting customer attrition, whereas Hwang et al. (2004) reported that the latter outperforms the former. Furthermore, most of these studies only evaluate a limited number of classification techniques on a single churn prediction data set. Therefore the issue of which classification technique to use for churn prediction remains an open research issue, in which the benchmarking experiment described in this paper aims to provide further insights. For an extensive literature review on customer churn prediction modeling one may refer to Verbeke et al. (2011).

Fig. 1 depicts a process model of the development of a customer churn prediction model. 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. The second step encompasses the actual building of a model. 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, as will be discussed in Sections 3 and 5. In a third step the model is assessed

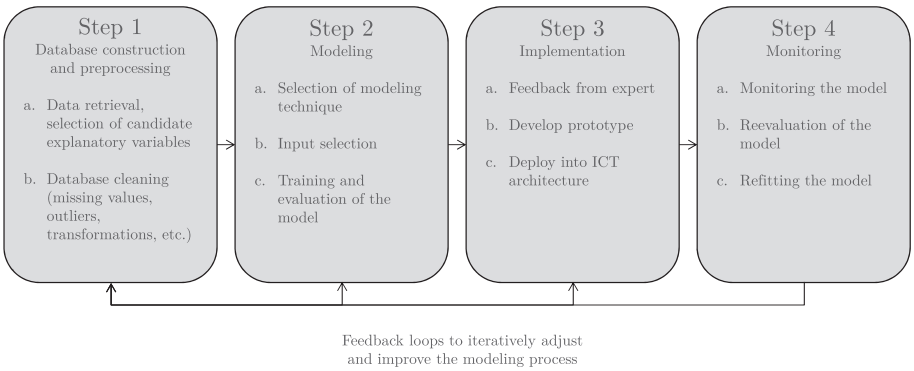

Fig. 1. Average profit per customer using maximum profit (dotted line), lift (dashed line), and AUC (full line).

by a business expert to check whether the model is intuitively correct and in line with business knowledge. 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 telecom sector, a continuous evaluation on newly gathered data is of crucial importance. At the end of each phase the results are evaluated, and if not satisfactory one returns to a previous step in order to adjust the process. As an alternative to the process model depicted in Fig. 1, the global CRISP-DM (Cross Industry Standard Process for Data Mining) methodology could be adopted, which is a well established methodology to develop data mining models (Fayyad et al., 1996). CRISP-DM formally consists of six major phases, from business and data understanding over data preprocessing and modeling to evaluation and deployment, with feedback loops allowing to iterate over these phases.

3. The maximum profit criterion to evaluate customer churn prediction models

In this section a maximum profit criterion is formulated to evaluate the performance of a customer churn prediction model, based on a formula introduced by Neslin et al. (2006) to calculate the profits generated by a retention campaign.

3.1. The profit of a retention campaign

Fig. 2 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, a fraction of the customer base is identified correctly by the implemented customer churn prediction model as would-be churners, and offered an incentive to stay. A fraction γ of these customers accepts the offer and is retained, but the remaining fraction $(1-\gamma)$ is not and effectively churns. On the other hand, a fraction of the customer base is incorrectly classified as would-be churners, and also offered an incentive to stay. All of these customers are assumed to accept the offer and none of them will churn. Finally, a fraction of the would-be churners in the customer base is not identified as such, and thus they are not offered an incentive to stay. Hence, all of these customers will effectively churn, and together with the correctly identified would-be churners that are not retained constitute the outflow of the customer base.

Given this dynamical process of customer churn and retention, the profit of a single churn management campaign can be expressed as (Neslin et al., 2006):

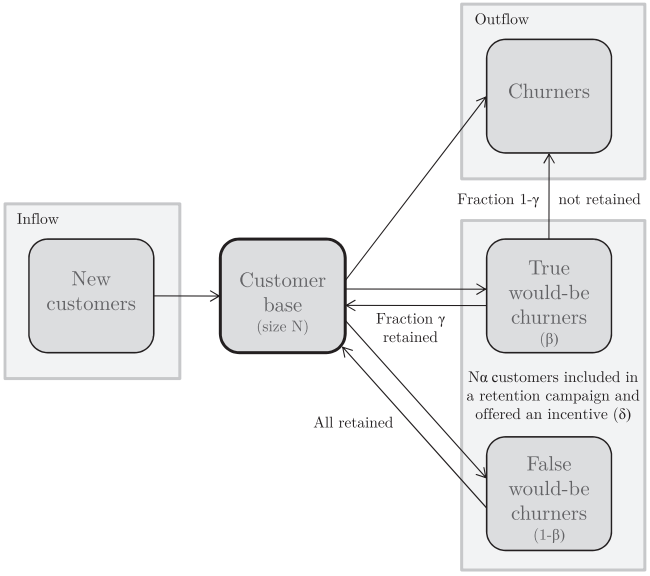

Fig. 2. Schematic representation of customer churn and retention dynamics within a customer base.

$$\Pi = N\alpha[\beta\gamma(CLV - c - \delta) + \beta(1-\gamma)(-c) + (1-\beta)(-c-\delta)] - A \qquad (1)$$

with Π the profit generated by a single 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 the fraction true would-be churners of the customers included in the retention campaign, 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 average customer lifetime value of the retained customers (i.e. the average net present value to the operator of all the revenues a retained customer will generate in the future, Gupta et al. (2006) and Glady et al. (2009)), and A the fixed administrative costs of running the churn management program.

The factor $N\alpha$ in Formula (1) reflects that the costs and profits of a retention campaign are solely related to the customers that are included in the campaign, except for the fixed administrative cost A which reduces the overall profitability of a retention campaign. The term $\beta\gamma(CLV - c - \delta)$ represents the profits generated by the campaign, i.e. the reduction in lost revenues reduced with the costs of the campaign $(CLV - c - \delta)$ due to retaining a fraction γ of the would-be churners of the fraction correctly identified would-be

churners β that are included in the campaign. The costs of the campaign are reflected by the term $\beta(1-\gamma)(-c)$, i.e. the cost of including correctly identified would-be churners that are not retained, and by the term $(1-\beta)(-c-\delta)$, which represents the cost due to including non-churners in the campaign, which are all logically expected to take advantage of the advantageous incentive offered to them in the retention campaign.

The term β reflects the ability of the predictive model to identify would-be churners, and can be expressed as:

$$\beta = \lambda\beta_0 \tag{2}$$

with β_0 the fraction of all the operator's customers that will churn, and λ the 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. Substituting Eq. (2) in Eq. (1), and rearranging the terms, results in:

$$\Pi = N\alpha\{[\gamma CLV + \delta(1-\gamma)]\beta_0\lambda - \delta - c\} - A. \tag{3}$$

According to Neslin et al. (2006), the direct link between lift and profitability in this equation demonstrates the relevance of using (top decile) lift as a performance criterion in predictive modeling. Lift is also indicated to be the most commonly used prediction criterion in predictive modeling by Neslin et al. (2006), which is confirmed by an extensive literature study on customer churn prediction modeling provided in Verbeke et al. (2011). However, as will be shown in the next section, using lift as a performance measure can lead to suboptimal model selection and as a result to a loss of profit, since the lift of a model is a function of the fraction α of customers that is included in the retention campaign (Provost and Jensen, 1999; Provost, 2005). Piatetsky-Shapiro and Masand (1999) and Mozer et al. (2000) have also formulated expressions to calculate the profit generated by a retention campaign. However, whereas Neslin et al. (2006) correctly discriminates between the cost of including a customer in the retention campaign and the cost of the incentive itself (which is only to be taken into account when a customer accepts the offer and is retained), both Piatetsky-Shapiro and Masand (1999) and Mozer et al. (2000) do not. Furthermore, Neslin et al. (2006) additionally takes into account a fixed administrative cost of running a customer management campaign.

3.2. The maximum profit criterion

As shown by Eq. (3), lift is directly related to profit, and therefore many studies on customer churn prediction use lift as a performance measure to evaluate customer churn prediction models. Since comparing entire lift curves is impractical and moreover rather meaningless, typically the lift at $\alpha = 5\%$ or $\alpha = 10\%$ is calculated.

A first issue regarding the use of the lift criterion is illustrated by Fig. 3. The lift curves of two different customer churn prediction models A and B intersect, resulting in a different model selection using top 5% lift and top 10% lift. In the case of Fig. 3, if a model is selected based on top 10% lift but the effectively included fraction of customers in the retention campaign equals 5%, then the generated profit will be suboptimal due to a suboptimal choice of customer churn prediction model, which results directly from using an inappropriate performance criterion. Therefore, if lift is used to assess and compare the outcomes of different customer churn prediction models, then the lift at the fraction that will effectively be included in the retention campaign should be used.

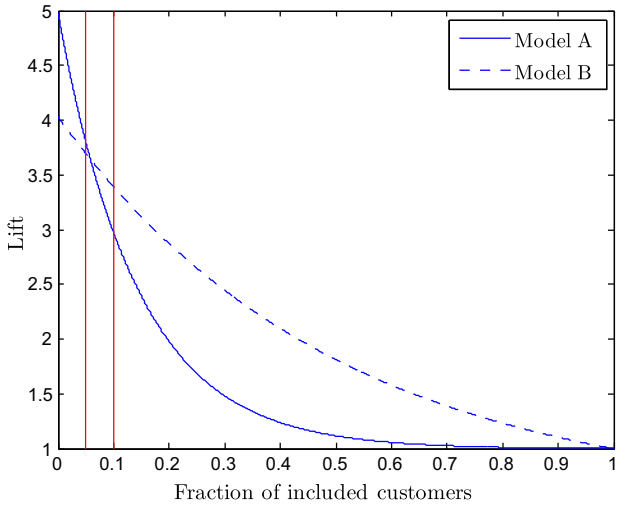

Fig. 3. Lift curves.

Furthermore, as illustrated by Fig. 4(a) and (b), the profit will also be suboptimal if the fraction of customers that is included in the retention campaign is not the optimal fraction. Fig. 4(a) and (b) represent on the *Y*-axis the profit per customer associated with including the fraction α on the *X*-axis of the customers ranked according to their probability to attrite in a retention campaign. The profit curves in Fig. 4(a) and (b) are calculated using Eq. (3) for models A and B and the lift curves shown in Fig. 3. The values of the other parameters are equal to the values provided in Neslin et al. (2006). When using top decile lift or profit, model B would be selected, resulting in a suboptimal profit per customer. When using top 5% lift or profit, model A would be selected, but when including the top 5% of the customers with the highest propensities to attrite, still a suboptimal profit is generated. And thus it is clear that in a practical setting the optimal fraction of customers should be included in a retention campaign to maximize profit, and that the profit or lift for the optimal fraction should be used to assess and compare the performance of customer churn prediction models.

Since the ultimate goal of a company by setting up a customer retention campaign is to minimize the costs associated with customer churn, it is logical to evaluate customer churn prediction models by using the maximum profit they can generate as a performance measure. In the remainder of this paper we will refer to this performance measure as the maximum profit (MP) criterion, which is formally defined as:

$$\text{MP} = \max_{\alpha}(\Pi). \tag{4}$$

In order to calculate the maximum profit measure a pragmatic approach is adopted, making two assumptions; (1) the retention rate γ is independent of the included fraction of customers α, and (2) the average CLV is independent of the included fraction of customers α. These assumptions allow to use a constant value for both γ and *CLV* in Eq. (3), and given the lift curve of the classification model which represents the relation between the lift and α, the maximum of Eq. (3) over α can be calculated in a straightforward manner.

In a realistic customer churn prediction setting, a retention campaign will only be profitable when including a rather small top-fraction of customers, with high predicted probabilities and with a relatively large subfraction of true would-be churners (i.e. with high lift). Hence, the optimal fraction α to maximize the returns of a retention campaign will lie within a rather small interval,

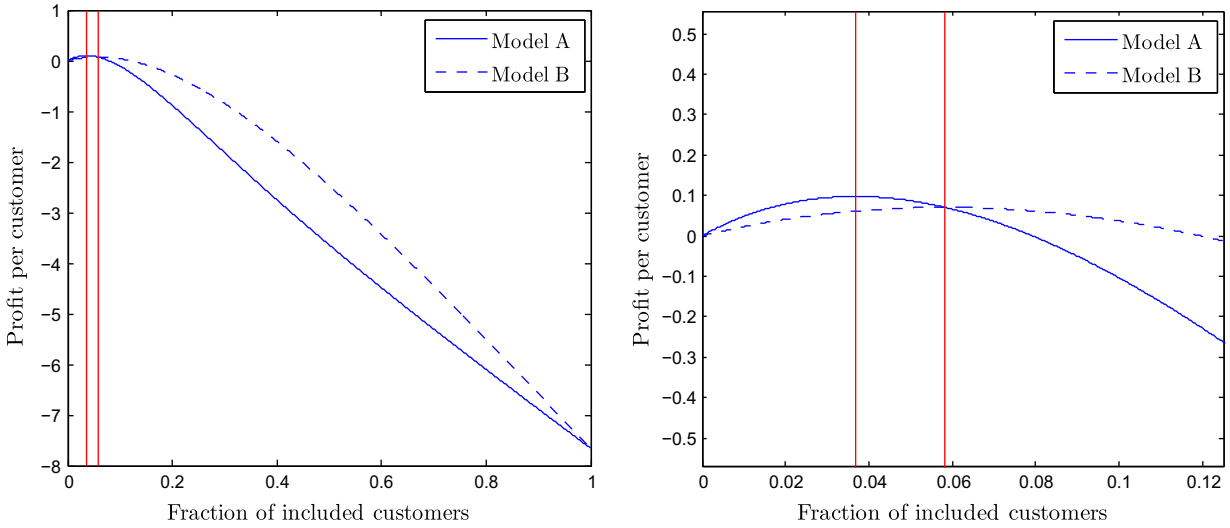

Fig. 4. Profit function with vertical lines indicating the maximum profit per customer, and detail of top decile results.

with the lower bound equal to 0%. The assumptions that are made therefore relax to independency of both γ and CLV of α within this limited interval of α. In other words, the churners within the top-fraction of customers that are detected by classification models are assumed to be randomly distributed over this interval with respect to their CLV and probability to be retained. Since classification models induced by different techniques yield different probability scores to churn, and consequently result in a different distribution of the detected churners within the top ranked customers, this seems to be a reasonable assumption, which has not been contradicted to our knowledge by any empirical study in the literature.

The independency assumptions discussed above are strongly related to uplift modeling for churn prediction, i.e. adding a second stage model to a customer churn prediction model to identify the customers with the highest predicted probabilities to be retained, or with the highest expected return on investment, within the group of customers with the highest predicted probabilities to churn (as indicated by the first stage customer churn prediction model). Uplift modeling, the existence of a correlation between the probability to be retained and the predicted probability to churn, and between the tendency to churn and the CLV, are marked as prime topics for future research, given the limited amount of studies that are available on this subject and the major relevance to customer relationship management and the development of retention strategies.

To calculate the maximum profit, in the remainder of this paper the values of the parameters γ, *CLV*, δ, and c in Eq. (3) are set equal to respectively 0.30, 200€, 10€, and 1€. The average values of the retention rate and the CLV are estimated by telco operators with high accuracy: huge budgets are spent on customer retention campaigns, which are typically conducted on a monthly basis and target millions of customers. Consequently, telco operators need to analyze the costs and benefits of such campaigns into detail, requiring accurate estimates of the average CLV of retained customers and average retention rates. Of course, the values may well differ for different segments (prepaid vs. postpaid, private vs. professional, etc.), and for different providers. The average values of the retention rate and the CLV that are used in the study are based on values found in the scientific literature (Neslin et al., 2006; Burez and Van den Poel, 2007), and based on information provided by data mining specialists of several telco providers. However, these values do not necessarily apply to each setting and may need to be adjusted when applying the MP measure. Moreover, it needs to be stressed that the contribution of this paper lies in the development and the application of the maximum profit criterion, rather than in the reported values of the maximum profits resulting from these parameter values.

Mozer et al. (2000) is to our knowledge the first and thus far the only study in the literature on customer churn prediction that acknowledges the importance of selecting the optimal threshold churn probability for a customer to be included in a retention campaign to maximize the expected cost savings to the operator. Mozer et al. (2000) refer to this selection problem as the *optimal decision-making policy*, and analyze a number of contour plots for a range of values of the parameters that impact the generated profit, leading to an indication of the cost savings that can be achieved by a retention campaign. However, the idea of the *optimal decision-making policy* is not formalized, nor is it applied as a performance measure to assess the performance of customer churn prediction models.

Finally, it should be stressed that optimizing the fraction of customers to include in a retention campaign in order to maximize the profit generated by a customer retention campaign is an innovative, and highly valuable, key-insight for practitioners in the field of customer churn prediction. As will be shown in the next sections, optimizing the included fraction of customers, as well as selecting the optimal customer churn prediction model by using the maximum profit criterion, can save a company significant amounts of money.

4. Experimental design of the benchmarking study

The experimental design of the benchmarking study consists of a full factorial experimental setup, in order to assess the effects of three different factors on the performance of a churn prediction model. The first factor concerns the *classification technique*, and has 21 possible levels, i.e. one per technique that is evaluated. The main workings of the applied techniques will be briefly explained in Section (4.1). The second factor, *oversampling*, consists of two possible levels. Level zero means that the original data set is used, while level one indicates that oversampling is applied to improve the learning process, as will be explained in Section 4.2. Finally, the third factor represents *input selection*, and also has two levels. Level zero means no input selection, and level one means that a generic input selection scheme is applied which will be presented in Section 4.3. This results in a $21 \times 2 \times 2$

experimental setup. The aim of the benchmarking study is to contrast the different levels and combinations of levels of these three factors, in order to draw general conclusions about the effects of classification technique, oversampling, and input selection on the performance of a customer churn prediction model. A full factorial experimental design allows to statistically test the effect of each factor separately, as well as the effects of interactions between the factors.

4.1. Classification techniques

Table 1 provides an overview of the classification techniques that are included in the benchmarking experiments (Lessmann et al., 2008). Previous studies reporting on the performance of a technique in a churn prediction modeling setting are referred to in the last column of the table. The included techniques are selected based on previous applications in churn prediction and expectations of good predictive power. An extensive overview of classification techniques can be found in Tan et al. (2006) or Hastie et al. (2001).

Where appropriate, default values for the hyperparameters of the various techniques are used, based on previous empirical studies and evaluations reported in the literature. If unknown, a parameter optimization procedure is performed which calculates the performance of a model trained on 2/3 of the training data and evaluated on the remaining validation set, for a range of parameter values. The values resulting in the best performing model are selected, and the final model is trained on the full training set using the selected parameter values. E.g. this procedure is performed for neural networks in order to determine the optimal number of hidden neurons, and for SVMs and LSSVMs to tune the kernel and regularization parameters. The benchmarking experiments are performed using implementations of the classification techniques in Weka, Matlab, SAS, and R.

4.2. Oversampling

Typically, the class variable in a customer churn prediction setting is heavily skewed, i.e. the number of churners is much smaller than the number of non-churners. This may cause classification techniques to experience difficulties in learning which customers are about to churn, resulting in poor classification power. Learning from imbalanced data sets has received a growing amount of attention from the data mining community (Chawla, 2010). In order to improve learning, sampling schemes to balance the class distribution have been proposed, such as over- and undersampling.

Fig. 5 illustrates the principle of oversampling. Observations of the minority class in the training set are simply copied and added to the training set, thus changing its distribution. Oversampling does in fact not add any new information to a data set, but only makes parts of the available information more explicit. Note that the class distribution of the test set is not altered, because a trained classifier is always evaluated on a pseudo-realistic data sample, in order to provide a correct indication of the future performance. Alternatively, the class distribution of the training set can also be altered by removing observations from the majority class, which is called undersampling. However, undersampling reduces the available amount of information, and therefore oversampling is applied. Table 3 in Section 6.1 summarizes the number of observations in each data set included in the benchmarking study, and the class distribution of the original and oversampled data set. The degree of sampling affects the performance of the resulting classifier. Classification techniques typically perform best when the class distribution is approximately even, and therefore the data sets are oversampled to approximate an even class distribution.

Finally, a number of advanced sampling schemes have been proposed in the literature, such as SMOTE (Chawla, 2002) which constructs synthetic data instances of the minority class in order to balance the class distribution. However, an extensive study of the impact of sampling schemes on the performance of customer churn prediction models is beyond the scope of this study, and left as a topic for future research.

4.3. Input selection

The third factor that possibly impacts the performance of a churn prediction model is the variable selection procedure. In practice, often a limited number of highly predictive variables is preferred to be included, in order to improve the comprehensibility of classification models, even at the cost of a somewhat decreased discrimination power (Martens et al., 2007; Piramuthu, 2004). Therefore a procedure can be applied in order to select the most predictive attributes and to eliminate redundant attributes. In this study, a generic variable input selection procedure is applied which iteratively reduces the number of variables included in the model, i.e. a wrapper approach (Tan et al., 2006). Previous to applying the generic input selection procedure a number of redundant variables are already filtered from the data set using the Fisher score. This filter is applied since the computational requirements to apply the wrapper approach scales exponentially with the number of variables that is present in the data set. The Fisher score does not require discretisation of the variables, and is defined as follows:

$$\textit{Fisher score} = \frac{|\bar{x}_C - \bar{x}_{NC}|}{\sqrt{s_C^2 + s_{NC}^2}} \tag{5}$$

with $\bar{x}_C$ and $\bar{x}_{NC}$ the mean value, and s_C^2 and s_{NC}^2 the variance of a variable for respectively churners and non-churners. Typically, the 20 variables with the highest Fisher scores, indicating good predictive power, are selected. As will be shown in the results section, a subset of 20 variables suffices to achieve optimal performance.

Algorithm 1: Pseudo-code of input selection procedure

choose initial number of attributes k to start input selection procedure
split data in training data $\mathcal{D}_{tr}$, and test data $\mathcal{D}_{te}$ in a 2/3, 1/3 ratio
calculate Fisher score of attributes in $\mathcal{D}_{tr}$
select k attributes with highest Fisher scores and continue with this reduced data set $\mathcal{D}_{tr}^k$
for $i = k$ to 1 **do**
  for $j = 1$ to i **do**
    train model excluding attribute j from $\mathcal{D}_{tr}^i$
  calculate performance P_j^i of model j
  end for
  remove attribute A_m from $\mathcal{D}_{tr}^i$ with $P_m^i = \max_j\left(P_j^i\right)$ resulting in $\mathcal{D}_{tr}^{i-1}$
end for
plot $\left(i, P_m^i\right)$ with $i = 1,\ldots,k$
select cut-off value i with optimal trade-off between performance and number of variables

The input selection procedure starts from this reduced data set. In each step as many models are trained as there are variables left. Each of these models includes all variables except for one. The variable that is not included in the model with the best performance is removed from the data set, and a next iteration is started with one variable less in the data set. Hence the procedure starts with 20 models that are calculated, with each model including only

Table 1
Summary of techniques evaluated in the benchmarking study.

Classification technique		Previous studies in churn prediction
Decision tree approaches		
A decision tree is grown in a recursive way by partitioning the training records into successively purer subsets. A minimum number of observations needs to fall into each subset, otherwise the tree is pruned. The metric to measure the pureness or homogeneity of the groups differs for the different techniques. C4.5 uses an entropy measure, while CART uses the Gini criterion. ADT is a boosted decision tree (see Ensemble methods) which distinguishes between alternating splitter and prediction nodes. A prediction is computed as the sum over all prediction nodes an instance visits while traversing the tree.		
C4.5 Decision Tree (Quinlan, 1993)	(C4.5)	Mozer et al. (2000), Wei and Chiu (2002), Au et al. (2003), Hwang et al. (2004), Hung et al. (2006), Neslin et al. (2006) and Kumar and Ravi (2008)
Classification and regression tree (Breiman et al., 1984)	(CART)	
Alternating decision tree (Freund and Trigg, 1999)	(ADT)	
Ensemble methods		
Ensemble methods use multiple base-classifiers resulting in better predictive performance than any of the constituent models, which are built independently and participate in a voting procedure to obtain a final class prediction. Random forests incorporates CART as base learner, Logistic model tree utilizes logit, and both bagging and boosting use decision trees. Each base learner is derived from a limited number of attributes. These are selected at random within the RF procedure, whereby the user has to predefine the number. LMT considers only univariate regression models, i.e. uses one attribute per iteration, which is selected automatically. Bagging repeatedly samples with replacement from a data set according to a uniform probability distribution, and trains the base classifiers on the resulting data samples. Boosting adaptively changes the distribution of the training examples so that the base classifiers, will focus on examples that are hard to classify		
Random forests	(RF)	Buckinx and Van den Poel (2005), Lariviere and Van den Poel (2005), Burez and Van den Poel (2007), Burez and Van den Poel (2009), Coussement and Van den Poel (2008) and Kumar and Ravi (2008)
Logistic model tree (Landwehr et al., 2005)	(LMT)	
Bagging	(Bag)	Lemmens and Croux (2006)
Boosting	(Boost)	Lemmens and Croux (2006)
Nearest neighbors		
Nearest neighbor methods classify an instance based on the k-most similar or nearest instances. kNN methods measure the analogy or similarity between instances using the Euclidean distance. Following Baesens et al. (2003b), both k = 10 and k = 100 are included in the experiments		
k-Nearest Neighbors $k = 10$	(kNN10)	(Datta et al., 2000)
k-Nearest Neighbors $k = 100$	(kNN100)	
Neural networks		
Neural networks mathematically mimic the functioning of biological neural networks such as the human brain. They consist of a network of neurons, interconnected by functions and weights which need to be estimated by fitting the network to the training data. By applying the trained network on a customer's attributes, an approximation of its posterior probability of being a churner is obtained		
Multilayer perceptron (Bishop, 1996)	(NN)	Datta et al. (2000), Mozer et al. (2000), Au et al. (2003), Hwang et al. (2004), Buckinx and Van den Poel (2005), Hung et al. (2006), Neslin et al. (2006) and Kumar and Ravi (2008)
Radial basis function network	(RBFN)	Kumar and Ravi (2008)
Rule induction techniques		
Rule induction techniques result in a comprehensible set of if-then rules to predict the minority class, while the majority class is assigned by default. RIPPER is currently one of the dominant schemes for rule-learning, operating in two stages. First an initial rule set is induced, which is refined in a second optimization stage to filter contradictory rules. PART on the other hand infers rules by repeatedly generating partial decision trees, combining rule learning from decision trees with the separate-and-conquer rule-learning technique		
RIPPER (Cohen, 1995)	(RIPPER)	Verbeke et al. (2011)
PART (Frank and Witten, 1998)	(PART)	
Statistical classifiers		
Statistical classifiers model probabilistic relationships between the attribute set and the class variable. Posterior probabilities are estimated directly in logistic regression. Naive Bayes estimates the class-conditional probability by assuming that attributes are conditionally independent, given the class label, so that class-conditional probabilities can be estimated individually per attribute. Bayesian Networks allow a more flexible approach and extend Naive Bayes by explicitly specifying which pair of attributes is conditionally independent		
Logistic regression	(Logit)	Eiben et al. (1998), Mozer et al. (2000), Hwang et al. (2004), Buckinx and Van den Poel (2005), Lariviere and Van den Poel (2005), Lemmens and Croux (2006), Neslin et al. (2006), Burez and Van den Poel (2007), Burez and Van den Poel (2009), Coussement and Van den Poel (2008) and Kumar and Ravi (2008)
Naive Bayes	(NB)	Neslin et al. (2006)
Bayesian networks	(BN)	Neslin et al. (2006)
SVM based techniques		
SVM based classifiers construct a hyperplane or set of hyperplanes in a high-dimensional space to optimally discriminate between churners and non-churners, by maximizing the margin between two hyperplanes separating both classes. A kernel function enables more complex decision boundaries by means of an implicit, nonlinear transformation of attribute values. This kernel function is polynomial for the VP classifier, whereas SVM and LSSVM consider both a radial basis and a linear kernel function		
SVM with linear kernel (Vapnik, 1995)	(linSVM)	Coussement and Van den Poel (2008) and Kumar and Ravi (2008)

(*continued on next page*)

Table 1 (*continued*)

Classification technique	Previous studies in churn prediction	
SVM with radial basis function kernel	(rbfSVM)	Coussement and Van den Poel (2008) and Kumar and Ravi (2008)
LSSVM with linear kernel (Suykens and Vandewalle, 1999)	(linLSSVM)	
LSSVM with radial basis function kernel	(rbfLSSVM)	
Voted perceptron (Freund and Schapire, 1999)	(VP)	

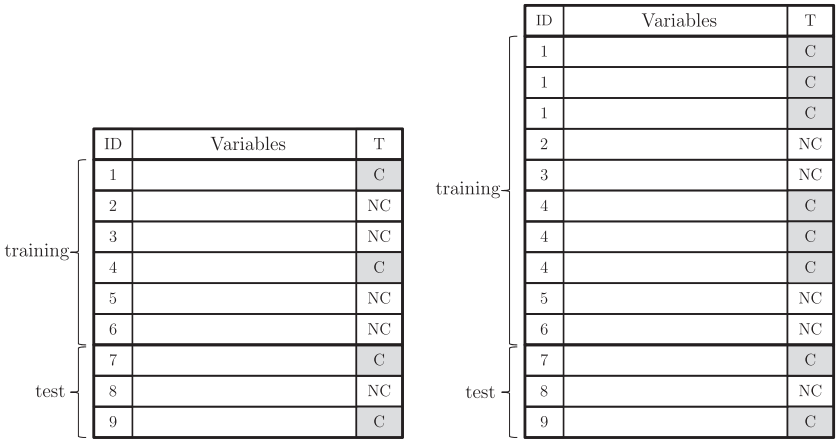

Fig. 5. Illustration of the principle of oversampling. A small data set with target variable T and nine observations (left panel) is split into a training set of six observations and a test set of three observations. Training instances classified as churners ($T = C$) are repeated twice in the oversampled data set (right panel).

nineteen variables. The variable excluded in the model with the best performance is then effectively removed from the data set, thus leaving a data set with only nineteen variables. This procedure is repeated, and eighteen models are estimated on the reduced data set. Again the variable excluded in the best performing model is removed from the data set. The procedure continues until no variables are left. A formal description of this procedure can be found in Algorithm 1. Fig. 6 illustrates the input selection procedure by plotting the performance of the sequentially best classifiers with a decreasing number of attributes. The number of attributes is shown on the X-axis, and the Y-axis represents the performance measured in terms of AUC, as will be explained in Section 5.2.

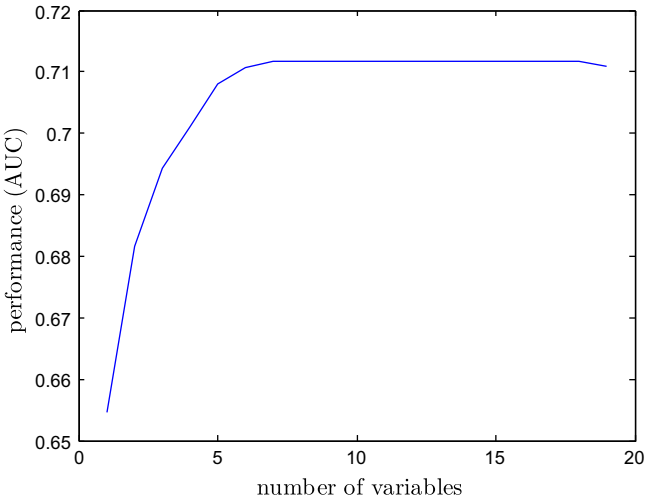

Fig. 6. Example of the evolution of the performance during the input selection procedure for a decreasing number of variables (technique ADT applied on data set KDD without oversampling, cfr. infra). The X-axis represents the number of variables included in the model, while the Y-axis represents the performance of the model measured in terms of AUC.

As can be seen in Fig. 6, removing a variable typically does not substantially impact the performance of a classifier when the number of variables remains large. The performance of the classifier drops however when the number of attributes left in the data set becomes too small. The model with the number of variables at the elbow point is generally considered to incorporate an optimal trade-off between minimizing the number of variables and maximizing the discriminatory power. The performance at the elbow point is the result of the input selection procedure that is reported in Section 6.

5. Experimental setup

A robust experimental setup and the use of appropriate test statistics and performance measures are crucial to draw valid conclusions. Section 5.1 describes the methodology that is followed in preprocessing the raw data sets, and Section 5.2 provides a non-exhaustive review of statistical (as opposed to profit centric) measures to assess the performance of classification models. This allows to correctly interpret the reported performance results in Section 6. Finally, Section 5.3 describes the statistical tests that will be applied to check the significance of differences in performance.

5.1. Data preprocessing

The general process model of the development of a customer churn prediction model described in Section 2 is followed to apply the selected classification techniques on the collection of data sets. A first important step in this process concerns the preprocessing of the raw data. Missing values are handled depending on the percentage of missing values of an attribute. If more than 5% of the values of an attribute are missing then imputation procedures

Table 2
The confusion matrix for binary classification.

		Actual	
		+	–
Predicted	+	True positive (TP)	False positive (FP)
	–	False negative (FN)	True negative (TN)

were applied. If less than 5% is missing, then the instances containing the missing value are removed from the data set in order to limit the impact of imputation procedures. Since missing values from different attributes often seem to occur for the same instances, i.e. usually for the same customers multiple data fields are missing, and instances are only removed if less than 5% of the values of an attribute are missing, the overall number of removed instances remained small. In case of categorical variables with many categories, coarse classification using hierarchical agglomerative clustering with the Euclidean distance is applied to reduce the number of categories to four (Tan et al., 2006). Finally all categorical variables are turned into binary variables using dummy encoding. No further preprocessing steps or transformations have been applied on the data.

5.2. Statistical performance measures

5.2.1. Percentage correctly classified

The *percentage correctly classified* (PCC) observations measures the proportion of correctly classified cases on a sample of data. Although straightforward, the PCC may not be the most appropriate performance criterion in a number of cases, because it tacitly assumes equal misclassification costs for false positive and false negative predictions. This assumption can be problematic, since for most real-life problems, one type of classification error may be much more expensive than the other. For instance in a customer churn prediction setting the costs associated with not detecting a churner will likely be greater than the costs of incorrectly classifying a non-churner as a churner (i.e. the costs associated with losing a customer tend to be greater than the costs of including a non-churner in a retention campaign). A second implicit assumption when using the PCC as evaluation criterion is that the class distribution (class priors) among examples is presumed constant over time, and relatively balanced (Provost et al., 1998). As mentioned in Section 4.2, the class distribution of a churn data set is typically skewed. Thus, using the PCC alone proves to be inadequate, since class distributions and misclassification costs are rarely uniform, and certainly not in the case of customer churn prediction. However, taking into account class distributions and misclassification costs proves to be quite hard, since in practice they can rarely be specified precisely, and are often subject to change (Fawcett and Provost, 1997).

5.2.2. Sensitivity, specificity, and the receiver operating characteristic curve

Class-wise decomposition of the classification of cases yields a confusion matrix as specified in Table 2. If TP, FP, FN, and TN represent the number of *true positives*, *false positives*, *false negatives*, and *true negatives*, then the *sensitivity* or *true positive rate* measures the proportion of positive examples which are predicted to be positive (TP/(TP + FN)) (e.g. the percentage of churners that is correctly classified), whereas the *specificity* or the *true negative rate* measures the proportion of negative examples which are predicted to be negative (TN/(TN + FP)) (e.g. the percentage of non-churners that are correctly classified).

Using the notation of Table 2, we may now formulate the overall accuracy as PCC = (TP + TN)/(TP + FP + TN + FN). Note that sensitivity, specificity, and PCC vary together as the threshold on a classifier's continuous output is varied between its extremes. The *receiver operating characteristic curve* (ROC) is a 2-dimensional graphical illustration of the sensitivity on the *Y*-axis vs. (1-specificity) on the *X*-axis for various values of the classification threshold. It basically illustrates the behavior of a classifier without regard to class distribution or misclassification cost, so it effectively decouples classification performance from these factors (Egan, 1975; Swets and Pickett, 1982). An example of a ROC curve is shown in Fig. 7.

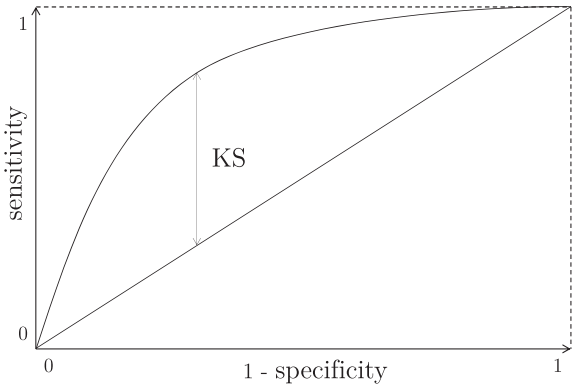

Fig. 7. Example of ROC curve with Kolmogorov–Smirnov statistic indicated.

5.2.3. Area under the receiver operating characteristic curve

In order to compare ROC curves of different classifiers, one often calculates the *area under the receiver operating characteristic curve* (AUROC or AUC). Assume a classifier produces a score $s = s(x)$, function of the attribute values x, with corresponding probability density function of these scores for class k instances $f_k(s)$ and cumulative distribution function $F_k(s)$, with only two classes $k = 0,1$. Then the AUC is defined as (Krzanowski and Hand, 2009):

$$AUC = \int_{-\infty}^{\infty} F_0(s) f_1(s)\, ds. \tag{6}$$

The AUC provides a simple figure-of-merit for the performance of the constructed classifier. An intuitive interpretation of the AUC is that it provides an estimate of the probability that a randomly chosen instance of class 1 is correctly rated or ranked higher than a randomly selected instance of class 0 (e.g. the probability that a churner is assigned a higher probability to churn than a non-churner). Note that since the area under the diagonal corresponding to a pure random classification model is equal to 0.5, a good classifier should yield an AUC much larger than 0.5.

5.2.4. Gini coefficient and Kolmogorov–Smirnov statistic

A measure that is closely related to the AUC is the *Gini coefficient* (Thomas et al., 2002), which is equal to twice the area between the ROC curve and the diagonal, i.e. $Gini = 2 * \text{AUC} - 1$. The Gini coefficient varies between 0 (i.e. the ROC curve lies on the diagonal and the model does not perform better than a random classification model) and 1 (i.e. maximum ROC curve and perfect classification).

Another performance measure related to the ROC curve is the *Kolmogorov–Smirnov* (KS) statistic. The KS statistic gives the maximum distance between the ROC curve and the diagonal at a specific cut-off value. Again, a value of the KS performance measure equal to one means a perfect classification, and KS equal to zero means no better classification than a random classifier. The KS measure is indicated in Fig. 7.

Table 3
Summary of data set characteristics: ID, source, region, number of observations, number of attributes, number of ordinal attributes, percentage churners in the original and oversampled data set, and references to previous studies using the data set.

ID	Source	Region	# Obs.	# Att.	# Ord.	%Churn original	% Churn sampled	Reference
O1	Operator	North America	47,761	53	42	3.69	50.01	Mozer et al. (2000)
O2	Operator	East Asia	11,317	21	12	1.56	47.44	Hur and Kim (2008)
O3	Operator	East Asia	2904	15	7	3.20	55,52	Hur and Kim (2008)
O4	Operator	East Asia	2969	48	30	4.41	45.99	Hur and Kim (2008)
O5	Operator	East Asia	2180	15	3	3.21	55.97	Hur and Kim (2008)
O6	Operator	Europe	338,874	727	679	1.80	50.00	
D1	Duke[a]	North America	93,893	197	135	1.78	49.75	Neslin et al. (2006) Lemmens and Croux (2006) Lima et al. (2009)
D2	Duke[a]	North America	38,924	77	36	1.99	49.81	
D3	Duke[a]	North America	7788	19	8	3.30	56.49	
UCI	UCI[c]	–	5000	23	15	14.14	50.28	Lima et al. (2009) Verbeke et al. (2011)
KDD	KDD Cup 2009[b]	Europe	46,933	242	173	6.98	50.56	

[a] www.fuqua.duke.edu/centers/ccrm/datasets/download.html.
[b] www.kddcup-orange.com.
[c] www.sgi.com/tech/mlc/db.

5.3. Statistical tests

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:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{7}$$

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 (i.e. $N > 10$ and $k > 5$), which is the case in this study when comparing different techniques ($N = 11$ and $k = 21$). When comparing the levels of the factors oversampling and input selection, k equals two and exact critical values need to be used to calculate the statistic.

If the null hypothesis is rejected by the Friedman test we proceed by performing the post hoc Nemenyi (Nemenyi, 1963) test 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:

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} \tag{8}$$

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.

The previous tests are applied to compare the results over multiple data sets, and to draw general conclusions about the impact of a factor on the performance of a model. To compare the performance, measured in AUC, of two classifiers on a single data set, the test of DeLong, DeLong, and Clarke-Pearson is used. After complex mathematical calculus, following a non-parametric approach whereby the covariance matrix is estimated using the theory on generalized U-statistics, DeLong et al. (1988) derive the following test statistic:

$$\left(\widehat{\Theta} - \Theta\right)c^T[cSc^T]^{-1}c\left(\widehat{\Theta} - \Theta\right)^T, \tag{9}$$

which has a chi-square distribution with degrees of freedom equal to the rank of cSc^T with $\widehat{\Theta}$ the vector of the AUC estimates, S the estimated covariance matrix, and c a vector of coefficients such that $c\Theta^T$ represents the desired contrast.

6. Empirical results

6.1. Data sets

Eleven data sets were obtained from wireless telecom operators around the world. Table 3 summarizes the main characteristics of these data sets, some of which have been used in previous studies referred to in the last column of the table. As can be seen from the table, the smallest data set contains 2180 observations, and the largest up to 338,874 observations. The applied techniques are evaluated using holdout evaluation on a single random split up of each data set into 2/3 training set and 1/3 test set, as commonly applied in large-scale benchmarking studies (e.g. Baesens et al., 2003b; Lessmann et al., 2008). The training set is used to learn a model, which is then evaluated on the test set to obtain an indication of the performance of the model. Holdout evaluation on a single split up provides a reliable indication of the performance if the data set is sufficiently large, and when multiple holdout evaluations with random splits into training and test set yield low variability. Preliminary tests indicated that the size of the data sets is sufficiently large and that a single split yields a reliable indication of the performance. Furthermore, multiple holdout splits (a.k.a. random subsampling) or cross-validation, which are typically applied when the size of the data set is small, would heavily increase the computational load of the experiments. Given the magnitude of the benchmarking study, this was an important factor as well in selecting the single split holdout evaluation methodology. For an introduction to evaluation methodologies as well as further references, one may refer to, e.g. Tan et al. (2006).

The data sets also differ substantially regarding the number of attributes, in a range from 15 up to 727. However, more attributes do not guarantee a better classification model. The final performance of a classifier mainly depends on the explanatory power of the attributes, and not on the number of attributes available to train a model. For instance, the number of times a customer called the helpdesk will most probably be a better predictor of churn behavior than the zip code. A large number of attributes heavily increases the computational requirements. Therefore the number of variables in data sets O6, D1, and KDD is reduced to a number of 50 using the Fisher score, in order to remove redundant

Table 4
Results of the benchmarking experiment evaluated using the MP per customer performance criterion.

Data set	O1	O2	O3	O4	O5	O6	D1	D2	D3	UCI	KDD	AR	AMP	O1	O2	O3	O4	O5	O6	D1	D2	D3	UCI	KDD	AR	AMP	
NN	0.10	0.05	0.34	1.38	0.00	0.03	0.00	0.00	0.22	5.36	0.26	**6.00**	0.70	0.07	0.00	0.33	1.40	0.14	0.00	0.00	0.00	0.09	3.97	0.33	**8.59**	0.57	
linSVM	0.00	0.00	0.06	1.41	0.05	0.00	0.00	0.00	0.00	3.05	0.09	14.14	0.42	0.00	0.00	0.45	1.31	0.05	0.01	0.00	0.00	0.06	3.72	0.09	**10.55**	0.52	
rbfSVM	0.03	0.00	0.28	1.37	0.11	0.00	0.00	0.00	0.00	4.92	0.09	**10.50**	0.62	0.00	0.00	0.00	1.16	0.05	0.00	0.00	0.01	0.00	5.16	0.09	**12.41**	0.59	
linLSSVM	0.04	0.00	0.48	1.42	0.15	0.00	0.00	0.00	0.16	3.80	0.21	**9.64**	0.57	0.04	0.00	0.53	1.32	0.14	0.01	0.00	0.00	0.11	3.78	0.23	**7.73**	0.56	
rbfLSSVM	0.08	0.00	0.46	1.49	0.14	0.00	0.00	0.00	0.11	4.96	0.07	**9.59**	0.66	0.04	0.00	0.00	1.05	0.14	0.00	0.00	0.00	0.00	4.81	0.09	**11.36**	0.56	
RIPPER	0.02	0.00	0.16	0.48	1.57	0.01	0.00	0.00	0.16	5.24	0.00	**10.86**	0.70	0.00	0.00	0.04	0.26	1.35	0.00	0.00	0.00	0.04	4.45	0.01	**12.55**	0.56	
PART	0.00	0.00	0.17	1.14	1.28	0.03	0.00	0.01	0.18	4.97	0.10	**7.64**	0.72	0.00	0.00	0.13	0.77	1.37	0.00	0.00	0.00	0.05	5.15	0.10	**11.18**	0.69	
C4.5	0.06	0.00	0.00	0.72	1.30	0.00	0.00	0.00	0.21	5.09	0.14	**10.59**	0.68	0.00	0.02	0.15	0.70	1.36	0.00	0.00	0.00	0.04	4.99	0.07	**11.23**	0.67	
CART	0.00	0.00	0.00	1.04	1.53	0.00	0.00	0.00	0.00	5.41	0.08	12.36	0.73	0.00	0.00	0.15	1.29	1.60	0.00	0.00	0.00	0.00	5.42	0.01	**11.09**	0.77	
ADT	0.12	0.01	0.08	0.98	1.57	0.00	0.00	0.00	0.21	5.21	0.30	**6.41**	0.77	0.12	0.01	0.07	1.39	1.61	0.00	0.00	0.00	0.08	5.21	0.30	**5.59**	0.80	
RF	0.06	0.00	0.36	1.03	1.30	0.00	0.00	0.00	0.18	5.76	0.14	**7.27**	0.80	0.08	0.01	0.22	1.38	1.32	0.00	0.00	0.00	0.22	5.80	0.14	**5.77**	0.83	
LMT	0.00	0.00	0.00	1.09	1.36	0.00	0.00	0.00	0.00	5.49	0.24	**11.18**	0.74	0.00	0.00	0.19	0.60	1.32	0.00	0.00	0.00	0.06	5.19	0.04	**10.50**	0.67	
Bag	0.09	0.02	0.41	1.13	1.65	0.04	0.00	0.00	0.19	5.63	0.29	**4.18**	0.86	0.02	0.00	0.32	1.22	1.54	0.00	0.00	0.02	0.12	5.36	0.19	**5.77**	0.80	
Boost	0.03	0.00	0.67	1.06	1.45	0.00	0.00	0.00	0.18	4.35	0.13	**9.64**	0.72	0.03	0.00	0.70	1.15	1.48	0.00	0.00	0.00	0.19	4.10	0.14	**8.18**	0.71	
RBFN	0.00	0.00	0.40	0.95	0.51	0.00	0.00	0.00	0.18	4.36	0.02	12.27	0.58	0.02	0.00	0.41	0.09	0.13	0.00	0.00	0.00	0.21	4.38	0.13	**9.45**	0.49	
VP	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.03	0.00	16.18	0.01	0.00	0.00	0.00	0.08	0.08	0.00	0.00	0.00	0.00	0.94	0.00	16.14	0.10	
Logit	0.08	0.00	0.40	1.38	0.14	0.03	0.00	0.00	0.19	4.00	0.24	**8.50**	0.59	0.08	0.00	0.48	1.25	0.15	0.01	0.00	0.00	0.11	3.97	0.22	**7.27**	0.57	
kNN10	0.00	0.00	0.11	1.07	0.05	0.00	0.00	0.00	0.04	4.02	0.05	12.59	0.49	0.00	0.00	0.17	0.97	0.30	0.00	0.00	0.00	0.02	3.77	0.00	**12.50**	0.48	
kNN100	0.00	0.00	0.23	0.98	0.06	0.01	0.00	0.00	0.10	4.28	0.04	13.00	0.52	0.00	0.00	0.16	1.32	0.05	0.00	0.00	0.00	0.06	3.99	0.00	**11.00**	0.51	
BN	0.04	0.00	0.59	1.40	1.22	0.05	0.00	0.00	0.35	4.08	0.25	**6.64**	0.73	0.00	0.00	0.00	0.05	0.47	0.00	0.00	0.00	0.11	3.71	0.25	**11.36**	0.42	
NB	0.00	0.00	0.42	1.07	0.11	0.00	0.00	0.00	0.21	4.57	0.11	**10.82**	0.59	0.00	0.00	0.45	1.02	0.14	0.00	0.00	0.00	0.21	4.57	0.11	**9.77**	0.59	
NN	0.05	0.01	0.49	1.22	1.45	0.01	0.00	0.03	0.09	3.66	0.21	**7.68**	0.66	0.08	0.00	0.25	1.11	1.53	0.02	0.00	0.00	0.15	5.86	0.28	**7.59**	0.84	
linSVM	0.02	0.00	0.60	1.51	0.08	0.00	0.00	0.01	0.06	3.65	0.00	**12.86**	0.54	0.00	0.00	0.39	1.30	0.00	0.00	0.00	0.00	0.04	3.81	0.11	14.91	0.51	
rbfSVM	0.05	0.00	0.69	0.94	0.98	0.00	0.00	0.00	0.00	5.06	0.09	**13.32**	0.71	0.00	0.00	0.00	1.33	1.39	0.00	0.00	0.00	0.05	5.61	0.11	**12.50**	0.77	
linLSSVM	0.01	0.00	0.46	1.26	0.08	0.00	0.00	0.01	0.02	3.97	0.20	**13.41**	0.55	0.03	0.00	0.50	1.37	0.11	0.00	0.00	0.02	0.02	3.97	0.20	**10.59**	0.56	
rbfLSSVM	0.05	0.00	0.45	1.41	1.05	0.00	0.00	0.01	0.22	5.45	0.16	**9.82**	0.80	0.06	0.01	0.00	1.15	1.39	0.00	0.07	0.00	0.20	5.37	0.16	**9.05**	0.76	
RIPPER	0.03	0.00	0.14	0.85	1.65	0.00	0.00	0.02	0.26	5.64	0.06	**9.55**	0.79	0.00	0.00	0.05	0.53	1.52	0.02	0.00	0.01	0.05	5.19	0.01	**11.45**	0.67	
PART	0.01	0.00	0.18	1.39	1.57	0.03	0.00	0.02	0.24	5.47	0.26	**6.86**	0.83	0.00	0.00	0.05	1.22	1.45	0.02	0.54	0.01	0.04	5.41	0.21	**8.86**	0.81	
C4.5	0.04	0.00	0.46	1.14	1.45	0.03	0.00	0.02	0.27	5.33	0.21	**6.36**	0.81	0.00	0.00	0.36	1.07	1.53	0.02	0.46	0.01	0.00	5.17	0.20	**10.23**	0.80	
CART	0.00	0.00	0.00	1.14	1.64	0.03	0.00	0.01	0.00	5.31	0.19	**11.18**	0.76	0.00	0.02	0.46	1.23	1.45	0.03	0.46	0.00	0.05	5.09	0.13	**8.77**	0.81	
ADT	0.05	0.00	0.64	1.05	1.46	0.02	0.00	0.01	0.12	5.21	0.31	**7.18**	0.81	0.02	0.00	0.64	1.10	1.46	0.02	0.00	0.02	0.10	5.21	0.29	**6.05**	0.81	
RF	0.01	0.01	0.38	1.14	1.45	0.03	0.66	0.02	0.16	5.73	0.13	**7.14**	0.88	0.01	0.00	0.53	0.95	1.35	0.02	0.67	0.02	0.15	5.69	0.13	**7.73**	0.86	
LMT	0.03	0.00	0.25	1.50	1.65	0.03	0.00	0.02	0.05	5.40	0.30	**6.77**	0.84	0.00	0.00	0.46	0.96	1.44	0.03	0.36	0.00	0.02	5.32	0.17	**11.14**	0.79	
Bag	0.07	0.01	0.45	1.13	1.65	0.02	0.14	0.00	0.08	5.63	0.34	**5.91**	0.86	0.02	0.01	0.30	0.91	1.45	0.03	0.67	0.00	0.15	5.46	0.13	**8.09**	0.83	
Boost	0.02	0.00	0.70	1.11	1.45	0.00	0.00	0.00	0.04	4.66	0.16	**12.77**	0.74	0.03	0.00	0.70	1.11	1.45	0.00	0.00	0.00	0.05	4.10	0.18	**10.27**	0.69	
RBFN	0.06	0.00	0.36	1.30	1.38	0.00	0.00	0.01	0.21	4.82	0.21	**10.00**	0.76	0.06	0.00	0.41	1.21	1.17	0.00	0.00	0.02	0.20	5.03	0.16	**10.14**	0.75	
VP	0.01	0.00	0.06	0.63	0.15	0.00	0.00	0.00	0.00	0.12	0.00	15.73	0.09	0.00	0.00	0.40	0.01	0.08	0.00	0.00	0.02	0.03	1.98	0.00	15.68	0.23	
Logit	0.07	0.00	0.49	1.39	0.03	0.01	0.00	0.02	0.05	4.00	0.18	**9.86**	0.57	0.09	0.00	0.49	1.26	0.03	0.01	0.00	0.02	0.03	4.04	0.17	**10.00**	0.56	
kNN10	0.02	0.00	0.33	1.11	1.27	0.01	0.00	0.00	0.08	5.07	0.24	**12.09**	0.74	0.00	0.00	0.54	1.27	1.33	0.01	0.60	0.00	0.04	4.98	0.24	**9.32**	0.82	
kNN100	0.02	0.00	0.33	1.11	1.27	0.00	0.00	0.00	0.08	5.29	0.19	**13.32**	0.75	0.02	0.00	0.40	1.02	1.45	0.01	0.03	0.00	0.13	5.36	0.23	**9.59**	0.79	
BN	0.05	0.00	0.57	1.27	1.53	0.00	0.00	0.02	0.26	4.28	0.36	**7.32**	0.76	0.01	0.00	0.57	1.09	1.47	0.01	0.54	0.01	0.04	4.17	0.33	**8.41**	0.75	
NB	0.03	0.00	0.44	1.27	0.81	0.00	0.00	0.02	0.23	4.56	0.19	**10.86**	0.68	0.03	0.00	0.44	1.27	0.81	0.00	0.00	0.02	0.23	4.58	0.19	**9.64**	0.69	

Without input selection

With input selection

Without oversampling

With oversampling

data before applying the classification techniques without input selection. As explained in Section 4.3, before applying the wrapper input selection procedure the number of variables in all data sets is reduced to a maximum of 20.

Table 3 also indicates the class distribution, which is for all data sets heavily skewed. The percentage of churners typically lies within a range of 1–10% of the entire customer base, depending on the length of the period in which churn is measured.

The definition of churn also slightly differs over the data sets, depending on the operator providing the data set. Most of the data however is collected over a period of three to six months, with a churn flag indicating whether a customer churned in the month after the month following the period when the data was collected. The one month lag of the data on the churn flag gives the marketing department time to setup campaigns aimed at retaining the customers that are predicted to churn.

Four types of attributes can be identified in the data sets.

- *Socio-demographic data*: contains personal information about customers such as age, gender, or zip code.
- *Call behavior statistics* or *usage attributes*: are mostly summary statistics such as number of minutes called per month, or variables capturing the evolution or trends in call behavior such as increase or decrease in number of minutes called. Usage attributes are purely related to the actual consumption.
- *Financial information*: contains billing and subscription information. Examples are average monthly revenue, revenue from international calls, type of plan chosen, and credit class of a customer.
- *Marketing related variables*: contain information about interactions between the operator and the customer, such as promotions that are offered to a customer, calls from the retention team to a customer, but also calls from the customer to the helpdesk. Marketing related variables are of particular interest to learn about how customers can be retained (Bolton et al., 2006).

The results of the input selection procedure will allow to assess in Section 6.3 the type of information that is generally most important to predict customer churn.[2]

6.2. Results and discussion

In the first part of this section the results of the benchmarking experiment are evaluated to assess the impact of input selection, oversampling, and classification technique on the performance of a customer churn prediction model. The performance is measured and evaluated using both statistical performance measures, i.e. top decile lift and AUC, and the maximum profit criterion. The use of these performance measures is analyzed and discussed in the second part of this section.

Tables 4–6 report the results of the benchmarking study in terms of respectively maximum profit, top decile lift, and AUC, both with and without oversampling and input selection applied. In all three tables the Bonferroni–Dunn test is used to evaluate the average ranks (AR) resulting from the respective performance measures. The best performance is in bold and underlined, and results that are not significantly different from the top performance at the 95% confidence level are tabulated in bold. Statistically

[2] The variables in data set KDD are anonymized and cannot be interpreted, and UCI is a synthetic data set which is only included in the study to allow comparison with previous and future studies. Therefore the input selection results of these data sets will not be included in the analysis in Section 6.3.

Table 5
Results of the benchmarking experiment evaluated using the top decile lift performance criterion.

Data set	O1	O2	O3	O4	O5	O6	D1	D2	D3	UCI	KDD	AR	AL	O1	O2	O3	O4	O5	O6	D1	D2	D3	UCI	KDD	AR	AL	
NN	3.52	1.93	4.28	7.86	1.59	3.77	1.50	1.09	3.59	6.10	2.67	**6.77**	3.45	3.71	1.75	5.13	7.86	3.98	4.02	1.37	0.73	3.81	3.20	2.82	**8.14**	3.49	
linSVM	2.01	0.70	1.43	7.38	2.39	1.39	1.06	0.18	1.91	3.12	0.98	16.82	2.05	1.55	1.93	5.70	7.86	1.19	4.92	1.50	0.73	4.04	3.68	0.92	**10.27**	3.09	
rbfSVM	2.68	1.58	3.99	7.86	3.18	2.05	1.06	0.91	3.48	4.68	1.52	**11.68**	3.00	2.46	1.93	2.00	7.38	2.78	0.74	1.37	1.82	1.12	5.93	1.29	**11.45**	2.62	
linLSSVM	2.98	1.58	6.27	7.62	2.39	2.79	1.23	0.55	3.59	3.55	2.47	**10.18**	3.18	3.19	1.23	5.70	8.10	1.99	4.92	1.50	0.73	4.04	3.42	2.48	**8.05**	3.39	
rbfLSSVM	3.29	1.75	5.99	7.86	4.38	3.12	1.30	0.55	3.70	5.50	2.04	**8.64**	3.59	3.14	1.75	3.14	7.14	2.78	2.54	1.67	1.09	1.12	5.67	2.21	**10.18**	2.93	
RIPPER	1.04	0.88	3.14	3.49	8.35	0.90	1.33	0.55	1.79	6.36	0.86	13.91	2.61	1.85	1.23	3.42	2.86	7.96	1.48	1.33	0.91	2.02	4.20	1.41	14.55	2.61	
PART	2.58	1.23	4.28	6.74	7.16	4.67	1.60	0.91	3.81	6.02	2.25	**8.36**	3.75	1.77	1.75	3.42	4.29	7.16	1.39	1.26	1.09	2.47	5.28	1.95	13.05	2.89	
C4.5	2.61	0.88	2.00	4.88	7.16	0.74	1.33	0.55	2.02	6.17	1.48	13.82	2.71	1.62	1.93	3.71	4.76	7.16	1.97	1.02	0.91	2.47	5.32	1.92	12.91	2.98	
CART	0.84	0.88	2.00	6.05	8.35	0.74	1.33	0.55	1.12	6.23	1.35	14.27	2.68	2.11	1.40	4.28	6.90	8.35	1.64	1.43	0.73	2.80	5.84	1.87	**11.23**	3.40	
ADT	3.39	1.58	4.28	6.05	8.35	5.33	1.50	1.09	4.26	5.97	2.73	**6.14**	4.05	3.39	1.93	4.28	7.38	8.75	5.17	1.57	1.09	3.81	5.97	2.73	**4.50**	4.19	
RF	2.82	2.80	5.42	6.51	7.56	3.61	0.99	0.55	3.48	6.71	2.12	**8.82**	3.87	2.93	2.10	4.56	7.62	8.75	4.26	1.13	1.09	3.59	6.71	2.32	**6.45**	4.10	
LMT	0.84	0.88	2.00	6.05	7.16	0.74	1.33	0.55	1.12	6.54	2.65	13.23	2.71	1.94	1.93	4.56	5.48	8.75	3.85	1.26	0.55	2.02	6.10	1.99	**10.64**	3.49	
Bag	3.41	2.45	5.42	6.05	8.75	5.90	1.54	1.82	3.25	6.75	2.74	**4.05**	4.37	2.98	2.10	4.28	7.14	8.75	4.18	1.43	0.91	3.03	6.67	2.41	**6.41**	3.99	
Boost	2.87	2.28	5.99	6.74	9.15	2.87	1.67	1.27	3.93	4.42	2.43	**5.73**	3.96	3.05	2.10	6.27	5.95	9.15	2.87	1.67	1.27	3.93	3.55	2.50	**5.32**	3.85	
RBFN	2.60	1.58	5.13	6.74	5.17	3.53	1.09	1.27	4.15	4.59	1.60	**9.68**	3.40	2.31	1.93	5.42	2.62	3.18	3.20	1.06	1.45	4.38	4.46	2.46	**8.95**	2.95	
VP	0.84	0.88	2.00	1.16	0.80	0.74	1.33	0.55	1.12	0.95	1.06	17.77	1.04	2.04	0.88	1.71	1.43	1.19	2.79	1.33	0.73	1.12	1.43	1.16	17.27	1.44	
Logit	3.66	1.05	5.42	7.91	1.99	5.08	1.57	0.91	3.93	3.72	2.59	**7.09**	3.44	3.68	1.05	5.99	7.62	1.99	5.66	1.43	1.09	3.93	3.77	2.55	**7.09**	3.52	
kNN10	1.84	2.45	3.14	6.90	4.38	3.20	1.13	0.73	3.14	4.37	1.89	**11.86**	3.02	1.89	2.28	3.71	6.67	3.98	3.20	1.26	0.55	2.80	4.33	1.68	**12.36**	2.94	
kNN100	2.09	2.63	4.85	6.74	4.38	4.35	1.37	0.73	3.93	4.33	1.93	**9.36**	3.39	1.91	1.93	3.14	7.86	4.38	3.77	1.33	1.09	3.81	4.33	1.65	**10.18**	3.20	
BN	2.85	1.75	5.13	8.14	7.16	5.98	1.54	1.45	4.49	4.50	2.45	**4.95**	4.13	2.31	1.75	2.00	1.43	6.36	0.66	1.30	0.91	2.80	4.33	2.43	13.23	2.39	
NB	2.95	1.75	5.70	7.21	3.58	2.46	1.64	1.27	4.26	4.55	2.35	**6.86**	3.43	2.95	1.58	5.70	6.90	3.18	2.46	1.60	1.27	4.15	4.55	2.34	**7.77**	3.34	
NN	3.39	2.80	6.27	7.67	8.75	4.84	1.65	1.88	3.70	2.94	2.42	**6.82**	4.21	3.59	2.28	4.28	7.21	8.75	4.92	1.44	1.41	3.81	6.28	2.71	**7.82**	4.24	
linSVM	2.38	1.23	4.56	7.67	2.78	3.69	1.45	1.72	2.92	3.72	1.46	14.45	3.05	1.67	1.75	5.70	7.67	2.78	4.43	1.66	1.25	3.81	3.42	2.33	**13.27**	3.32	
rbfSVM	2.68	1.40	6.56	6.05	8.35	2.79	1.50	2.03	4.15	4.94	1.77	**11.09**	3.84	1.87	1.58	2.28	7.67	8.75	4.51	1.66	1.41	3.59	6.32	2.23	**12.05**	3.81	
linLSSVM	3.12	1.93	5.99	7.91	1.99	4.43	1.13	1.56	3.81	3.03	2.39	**10.64**	3.39	3.27	2.28	6.27	7.91	2.78	4.43	1.39	1.88	3.48	3.03	2.36	**10.36**	3.55	
rbfLSSVM	2.53	2.10	5.99	7.91	8.75	3.61	1.47	1.56	4.26	5.93	2.29	**9.41**	4.22	3.15	2.98	2.85	6.98	8.75	4.51	5.78	1.41	3.81	6.23	2.21	**9.59**	4.42	
RIPPER	1.01	0.88	3.14	4.88	8.75	0.74	1.04	1.41	2.24	6.54	1.00	16.45	2.88	1.75	1.93	2.85	4.88	8.75	4.35	7.22	2.03	3.03	4.46	1.78	13.82	3.91	
PART	2.98	1.58	4.28	8.37	8.75	4.35	1.65	2.19	4.04	6.58	2.75	**7.09**	4.32	2.24	2.28	2.85	6.51	8.75	4.76	7.59	1.88	2.13	6.15	2.53	**10.05**	4.33	
C4.5	2.31	0.88	5.13	6.74	8.75	3.20	1.04	1.41	2.36	6.49	1.80	14.45	3.65	2.24	1.23	4.85	6.74	8.75	4.02	7.49	2.34	2.13	5.37	2.50	**11.23**	4.33	
CART	0.84	0.88	2.00	6.28	9.15	3.20	1.04	1.56	1.12	6.58	2.59	13.82	3.20	2.09	1.93	5.42	6.51	8.75	4.84	7.44	2.19	3.14	5.11	2.00	**10.73**	4.49	
ADT	3.04	2.45	5.70	6.05	9.15	4.84	1.64	2.19	4.15	5.97	2.81	**5.82**	4.36	3.25	2.28	5.70	6.74	9.15	4.26	1.69	2.50	4.04	5.97	2.75	**6.77**	4.40	
RF	2.75	3.51	5.13	6.98	8.75	4.92	7.58	2.50	3.48	6.80	1.94	**6.59**	4.94	2.78	3.68	6.56	6.98	8.75	4.84	7.63	1.56	4.26	6.71	1.94	**5.77**	5.06	
LMT	2.80	0.88	3.42	7.91	9.15	4.92	1.04	1.72	3.70	6.02	2.82	**8.91**	4.03	1.99	2.80	6.27	6.74	8.75	4.92	7.19	2.81	2.92	6.02	2.29	**7.82**	4.79	
Bag	3.22	3.16	4.85	6.51	8.75	4.92	4.56	1.72	3.70	6.71	2.82	**6.18**	4.63	2.85	2.10	4.85	6.74	8.75	4.84	7.41	1.88	2.92	6.80	1.88	**9.77**	4.64	
Boost	2.66	2.10	6.27	6.98	9.15	2.87	1.54	1.56	3.48	4.03	2.54	**9.95**	3.93	3.07	2.63	6.27	6.98	9.15	2.95	1.58	1.09	3.25	3.55	2.59	**9.86**	3.92	
RBFN	3.34	2.63	5.42	7.91	8.75	4.43	1.55	2.34	4.38	5.37	2.52	**5.68**	4.42	3.20	2.45	5.99	6.98	8.35	4.43	1.54	1.88	4.60	5.11	2.50	**8.68**	4.28	
VP	0.89	0.88	2.28	4.19	1.19	0.82	1.04	1.25	1.12	1.00	1.04	19.23	1.43	2.09	0.88	5.42	0.93	3.18	2.87	1.33	1.56	2.80	1.73	1.10	17.73	2.17	
Logit	3.39	1.40	5.99	7.91	3.18	4.67	1.40	1.88	3.48	3.42	2.35	**9.86**	3.55	3.59	2.45	6.27	7.21	3.18	4.67	1.38	1.72	3.59	3.29	2.27	**10.36**	3.60	
kNN10	2.19	1.75	5.13	7.44	8.75	4.76	2.20	1.88	3.37	5.54	2.24	**10.14**	4.11	2.18	3.16	5.13	7.67	8.75	4.84	7.09	1.88	3.81	5.93	2.30	**7.77**	4.79	
kNN100	3.00	1.75	5.13	7.44	8.75	4.43	2.20	1.88	3.37	4.59	2.44	**9.45**	4.09	2.73	2.10	5.42	6.51	8.75	4.76	3.57	1.25	4.15	5.19	2.56	**10.09**	4.27	
BN	3.15	1.75	5.42	7.67	9.15	4.51	1.47	1.88	4.38	4.55	2.93	**6.77**	4.26	2.50	2.63	5.70	7.21	8.75	4.59	7.19	1.88	3.93	4.89	2.88	**7.23**	4.74	
NB	3.17	2.28	6.27	7.91	6.36	4.43	1.48	2.03	4.49	4.11	2.39	**7.18**	4.08	3.17	2.10	6.27	7.91	6.76	2.87	1.48	2.03	4.49	4.11	2.39	**9.23**	3.96	

Without input selection

With input selection

Without oversampling

With oversampling

significant differences in performance at the 99% level are emphasized in italics, and significantly different results at the 95% level but not at the 99% level are reported in normal script. In Table 6, as explained in Section 5.3, the results of the test of DeLong, DeLong, and Clarke-Pearson to compare the performance in AUC on each data set separately are reported following the same notational convention.

6.2.1. Input selection

The results of the experiments with and without input selection can be found respectively in the lower and upper panels of Tables 4–6. Applying the Friedman test to compare the results for each measure yields a p-value around zero, both when including results with and/or without oversampling. This indicates that classifiers yield a significantly better predictive performance when applying input selection. At first sight this result might seem somewhat counterintuitive. However, it makes sense that it is easier to learn from a smaller data set with few, yet highly predictive, variables, than from an extensive set containing much redundant or noisy data. This result indicates that it is crucial to apply an input selection procedure in order to attain good predictive power. Moreover, a model containing less variables is advantageous as it will be more stable, since collinearity is reduced. More importantly from a practical point of view, a concise model is also easier to interpret, since the number of variables included in the model is minimized. Fig. 8(a) plots the results of the input selection procedure on data set O1 for logistic regression, which is exemplary for most techniques and data sets. The Y-axis represents the performance measured in AUC, and the number of variables included in the model is plotted on the X-axis. From this figure, it can be seen that adding a variable improves the performance of the logit model dramatically when only few variables are included in the model (i.e. on the left side of the figure the curve has a steep, positive slope).

However, the positive effect of adding extra variables flattens at eight variables, then reaches a maximum at nine variables, and when including more than twelve variables the performance decreases. The optimal trade-off between the number of attributes that is included in the model and the predictive performance as required by a business expert lies at eight variables. Selecting less variables yields poor predictive power, while including more variables makes the model harder to interpret and adds only little predictive power. Fig. 8(a) can also be interpreted starting from the right side, with many variables included in the data set. At first removing variables improves the performance since mostly redundant attributes will be removed. When too much information is filtered from the data however, the performance drops.

In Fig. 8(b) a boxplot summarizes the number of variables of the different data sets that is used by the eight best performing techniques according to the MP criterion (cfr. infra). On each box, the central mark indicates the median number of variables, the edges of the box represent the 25th and 75th percentiles, the whiskers extend to the most extreme numbers that are not considered to be outliers, and outliers finally are plotted by crosses. ADT, NB, and C4.5 appear to be very efficient algorithms, which are able to produce powerful models with only a very small number of attributes. The number of variables needed by RF, NN, PART, and LMT on the other hand seems to be heavily dependent on the data (i.e. the boxes and whiskers are spread over a wide range). On average, these eight techniques only need around 6 or 7 variables to yield optimal performance.

This means that a surprisingly small number of variables suffices to build an effective and powerful customer churn prediction

Table 6
Results of the benchmarking experiment evaluated using the AUC performance criterion.

Data set	O1	O2	O3	O4	O5	O6	D1	D2	D3	UCI	KDD	AR	AA	O1	O2	O3	O4	O5	O6	D1	D2	D3	UCI	KDD	AR	AA
NN	73.0	53.6	72.4	90.0	55.2	84.0	54.2	53.8	62.2	**89.3**	67.8	**8.1**	**68.7**	**75.5**	**61.8**	76.0	**86.5**	67.3	81.5	55.1	**54.5**	68.3	83.8	**71.2**	**6.4**	**71.0**
linSVM	59.8	44.8	56.6	87.7	61.0	75.8	52.5	43.3	63.1	79.9	50.5	15.2	61.4	55.2	51.6	**87.0**	**88.3**	62.1	86.4	**56.7**	**53.6**	75.8	83.4	48.4	**10.2**	**68.0**
rbfSVM	67.5	**58.3**	82.6	88.1	63.9	75.9	53.1	48.7	64.3	88.8	58.3	10.3	68.1	61.6	**57.0**	50.0	**89.3**	73.0	50.4	54.0	**54.0**	59.6	**90.6**	57.7	**10.5**	**63.4**
linLSSVM	69.0	**57.8**	**89.5**	90.3	64.5	77.0	54.1	49.8	70.5	83.7	66.6	**8.5**	**70.3**	69.0	53.4	**86.8**	**89.8**	62.2	**86.9**	**57.5**	**51.4**	75.6	83.6	66.6	**7.2**	**71.2**
rbfLSSVM	70.3	**63.5**	**89.4**	88.4	74.1	60.2	53.0	49.8	63.0	**89.9**	64.4	**8.5**	**69.7**	69.2	**58.7**	78.3	**89.7**	67.2	62.2	55.4	**53.7**	63.5	**89.7**	65.2	**7.7**	**68.4**
RIPPER	50.6	50.0	55.4	63.0	**91.8**	50.8	50.0	50.0	55.4	86.9	50.0	14.5	59.4	54.4	51.3	57.0	61.4	**89.1**	53.6	50.5	50.4	54.4	78.7	59.7	16.3	60.0
PART	58.2	53.4	63.3	74.0	79.5	78.0	55.2	46.9	68.2	78.4	58.7	13.2	64.9	54.0	54.2	59.7	69.3	85.8	53.3	49.9	**52.5**	57.3	85.6	55.4	15.0	61.6
C4.5	56.4	50.0	50.0	60.6	79.6	50.0	50.0	50.0	55.6	82.5	57.1	15.7	58.3	53.1	55.2	59.4	70.6	85.7	56.6	49.2	50.0	57.0	82.7	55.2	15.8	61.3
CART	50.0	50.0	50.0	76.3	**91.5**	50.0	50.0	50.0	50.0	86.6	60.2	14.7	60.4	56.2	52.4	64.3	82.3	**91.8**	54.0	50.9	49.6	58.8	86.1	55.3	14.0	63.8
ADT	72.0	**64.3**	82.3	83.9	**94.5**	**89.2**	**59.8**	**59.5**	72.5	88.5	**70.9**	**3.6**	**76.1**	72.0	**65.3**	**82.2**	**87.6**	**93.9**	**88.7**	**59.1**	**59.2**	70.8	**88.5**	**70.9**	**2.6**	**76.2**
RF	66.0	**59.6**	79.4	82.1	88.4	63.9	49.9	47.7	63.4	**90.4**	62.2	10.7	68.5	66.1	56.3	72.0	**88.3**	**92.3**	67.3	52.1	**54.7**	66.8	**89.5**	63.0	**8.2**	**69.8**
LMT	50.0	50.0	50.0	83.2	**87.7**	50.1	50.0	50.0	50.0	**90.0**	68.0	12.9	61.7	59.2	**56.1**	74.0	77.9	**93.5**	74.2	49.2	45.0	54.1	**89.2**	60.0	12.1	66.6
Bag	**74.4**	**64.4**	**90.3**	81.2	**96.4**	**90.2**	**59.1**	**61.8**	71.0	**91.8**	**71.3**	**2.2**	**77.5**	67.1	56.5	73.9	**86.3**	**93.3**	66.4	53.6	46.3	64.1	**90.3**	65.3	**9.1**	**69.4**
Boost	68.3	**64.2**	**86.6**	82.8	**95.0**	81.5	**59.9**	**59.6**	69.8	85.0	68.4	**4.9**	**74.6**	69.5	**63.3**	**88.0**	**87.3**	**94.7**	81.5	**59.9**	**57.4**	69.9	83.7	68.8	**3.8**	**74.9**
RBFN	65.5	52.9	77.8	83.7	75.0	80.3	55.5	**53.1**	76.6	84.6	61.2	**9.3**	**69.7**	68.6	54.3	**82.7**	73.2	72.2	82.2	53.3	**50.7**	**78.3**	86.0	64.2	**9.2**	**69.6**
VP	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.4	51.2	18.0	50.1	65.9	54.5	48.6	52.2	56.2	80.8	52.6	**58.1**	50.0	62.3	53.7	14.9	57.7
Logit	**74.7**	53.9	86.7	**91.0**	62.8	87.3	**57.2**	53.3	75.6	84.6	68.2	**5.8**	**72.3**	**75.3**	54.7	**87.4**	**86.2**	62.1	**88.4**	**57.2**	**53.6**	75.6	84.4	68.1	**6.7**	**72.1**
kNN10	58.9	**59.7**	66.1	81.7	74.7	62.5	51.8	48.0	64.3	83.6	59.7	12.7	64.6	58.8	**59.7**	66.7	81.2	76.0	62.5	51.9	47.8	64.0	83.5	59.4	12.9	64.7
kNN100	65.6	**63.6**	75.3	**87.6**	67.6	77.3	**56.5**	44.5	76.6	80.8	64.1	**9.9**	**69.0**	64.2	**61.7**	73.7	**87.8**	72.1	76.9	53.8	46.5	**75.5**	82.4	63.6	**10.3**	**68.9**
BN	69.9	**54.4**	**88.8**	**89.8**	92.7	**89.6**	**58.3**	**54.4**	**80.8**	86.5	64.9	**4.5**	**75.5**	69.9	54.8	68.2	61.0	87.2	75.2	53.8	**52.6**	70.9	86.5	65.7	**9.1**	**67.8**
NB	69.9	55.9	84.1	86.4	75.2	79.3	55.6	**53.9**	**80.9**	83.4	67.1	**6.9**	**72.0**	66.3	54.3	**84.1**	**87.0**	74.9	79.2	55.6	**54.0**	**79.8**	80.8	65.7	**8.0**	**71.1**
Without oversampling														**With oversampling**												
NN	**72.8**	**64.5**	**90.3**	**91.0**	**50.4**	87.3	**59.4**	**60.6**	76.8	82.5	66.5	**7.6**	**72.9**	**72.8**	**64.8**	**88.7**	**88.0**	**94.8**	**87.6**	**59.1**	**60.6**	**77.0**	**91.7**	71.2	**4.4**	**77.8**
linSVM	65.3	**59.5**	84.8	**92.0**	46.2	84.1	**57.2**	**59.3**	71.9	82.8	63.1	13.3	69.7	67.0	**61.4**	**88.4**	**92.5**	67.9	85.8	**59.5**	**57.7**	76.6	83.4	65.9	**10.9**	**73.3**
rbfSVM	67.3	**62.6**	**89.1**	83.9	**45.8**	83.1	**59.8**	**59.1**	67.8	89.5	62.8	12.1	70.1	60.8	**64.4**	82.8	**90.5**	95.6	85.7	**55.0**	**58.7**	76.3	**90.8**	68.2	**9.7**	**75.3**
linLSSVM	69.1	**63.3**	89.3	**92.5**	44.0	85.1	54.2	**63.8**	76.8	83.2	65.8	**10.3**	**71.6**	71.1	**62.8**	**88.8**	**93.0**	65.4	85.5	**59.1**	**59.6**	77.0	83.2	65.6	**9.7**	**73.8**
rbfLSSVM	66.1	**64.5**	**90.3**	**93.3**	55.5	84.4	54.9	**63.6**	76.3	**90.9**	68.1	**7.6**	**73.4**	70.6	**70.1**	**87.0**	**88.6**	94.6	86.6	**54.8**	**63.1**	71.2	**90.8**	65.5	**8.6**	**76.6**
RIPPER	50.8	50.0	55.3	71.0	**93.8**	50.0	50.0	50.8	58.1	87.5	50.8	17.3	60.7	64.1	**63.4**	76.2	80.1	**93.8**	85.5	51.3	**55.1**	74.0	85.4	66.8	15.0	72.3
PART	70.9	**63.9**	82.4	**89.0**	**95.4**	86.6	**60.0**	**59.4**	**77.9**	88.7	70.1	**7.5**	**76.8**	64.3	60.6	76.2	88.1	**93.5**	**87.4**	52.0	**56.8**	71.5	87.5	68.5	12.6	73.3
C4.5	57.0	50.0	70.6	80.7	**94.8**	82.3	50.0	50.8	58.9	88.2	63.4	15.8	67.9	62.5	59.3	65.6	80.5	**93.9**	86.0	52.0	**55.6**	62.6	84.3	68.5	15.9	70.1
CART	50.0	50.0	50.0	78.0	94.9	82.3	50.0	**54.2**	50.0	88.4	67.2	15.9	65.0	55.8	60.0	71.7	88.1	**94.4**	87.2	51.3	54.9	60.7	85.3	65.8	15.4	70.5
ADT	71.9	**65.4**	**87.7**	86.9	**97.2**	87.3	**60.4**	**59.2**	**79.3**	88.5	**71.1**	**5.3**	**77.7**	70.6	**65.6**	**87.7**	88.3	**97.1**	86.3	**61.0**	**56.8**	**77.3**	88.5	**71.2**	**6.4**	**77.3**
RF	64.6	**63.3**	78.6	**89.6**	**94.9**	**87.7**	54.7	**58.7**	69.1	90.6	65.6	**10.5**	**74.3**	65.6	**63.7**	79.6	87.3	**94.6**	**87.0**	**54.8**	**58.8**	71.2	**91.4**	65.7	**11.2**	**74.5**
LMT	69.9	50.2	52.4	**90.8**	**95.5**	87.3	50.0	**54.7**	73.7	**90.7**	**71.8**	**9.4**	**71.5**	63.0	**65.6**	**88.7**	**87.8**	**95.2**	87.3	**57.1**	**62.8**	73.7	**90.3**	66.6	**7.5**	**76.2**
Bag	**73.6**	**69.7**	**89.6**	87.5	**97.0**	87.3	**60.5**	**60.4**	**78.5**	**91.8**	**71.2**	**2.7**	**78.8**	67.5	61.0	82.2	**89.6**	**94.1**	87.2	**55.8**	**61.5**	67.3	**90.6**	65.4	**10.8**	**74.7**
Boost	67.8	**64.4**	**88.0**	84.9	**95.1**	81.1	**59.9**	**54.7**	77.8	85.9	68.8	**10.5**	**75.3**	70.0	**64.8**	**88.0**	84.9	**95.1**	81.9	**60.0**	**55.7**	76.2	83.7	69.5	**10.0**	**75.4**
RBFN	**71.8**	**64.6**	**87.1**	**90.6**	93.2	86.1	**58.4**	**59.0**	**79.2**	88.9	69.4	**7.8**	**77.1**	71.7	59.8	**88.4**	**92.4**	91.2	85.7	**58.9**	**57.8**	**81.8**	89.0	69.8	**8.0**	**77.0**
VP	50.3	50.0	51.4	66.3	52.0	50.4	50.0	50.0	50.0	50.8	53.0	18.9	52.2	66.6	55.2	**86.9**	61.5	67.9	80.9	**57.5**	**57.1**	67.8	68.3	53.4	16.6	65.7
Logit	**72.8**	62.7	**88.7**	**93.1**	68.9	**86.9**	**59.2**	**60.4**	77.6	83.7	66.5	**8.3**	**74.6**	**73.1**	**62.4**	**89.5**	**92.2**	68.3	**86.7**	**59.6**	**60.3**	77.0	83.8	66.3	**7.7**	**74.5**
kNN0	62.5	**65.9**	**86.5**	88.8	93.8	**87.1**	**57.6**	**63.3**	**78.2**	89.0	70.3	**7.5**	**76.6**	62.4	**65.8**	**87.6**	**90.2**	**94.4**	**87.2**	**55.3**	**58.7**	**78.0**	89.1	67.9	**8.4**	**76.1**
kNN100	57.8	**65.9**	**86.5**	88.8	93.8	86.1	**57.6**	**63.3**	**78.2**	89.0	71.1	**8.0**	**76.2**	68.1	**65.1**	**87.1**	**89.0**	**94.5**	86.9	**56.3**	**61.5**	**78.4**	**90.7**	70.0	**6.5**	**77.1**
BN	71.6	58.4	**89.4**	90.3	**97.0**	86.3	**60.0**	**55.5**	**81.7**	84.0	**71.6**	**6.7**	**76.9**	68.1	**63.4**	**89.6**	**89.1**	**96.3**	86.1	**57.4**	**60.6**	**79.7**	83.2	**71.4**	**6.6**	**76.8**
NB	**71.8**	**64.7**	**88.6**	**92.1**	90.8	86.1	**58.7**	**59.5**	**81.8**	86.9	68.8	**7.1**	**77.3**	71.8	**64.1**	**88.6**	**92.0**	90.9	81.1	**58.7**	**59.5**	**81.8**	86.9	68.8	**7.9**	**76.7**

Without input selection

With input selection

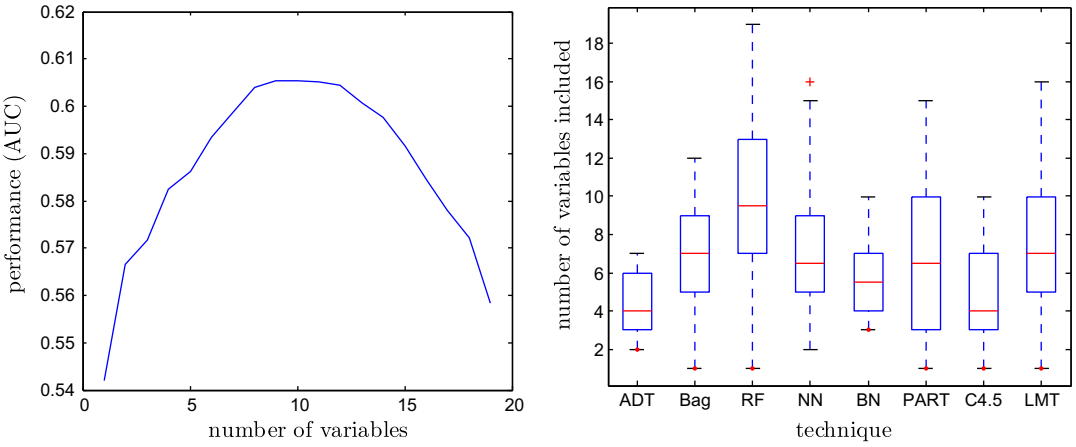

Fig. 8. Performance evolution of the input selection process for logistic regression applied on data set O1, and boxplot of the number of variables used by the eight best performing techniques using the MP criterion.

model. Hence from an economical point of view, to build or improve the predictive power of such a model, it is operationally more efficient to focus on collecting and improving the quality of a small set of well chosen attributes, than to collect as much data as possible. Data quality is an important issue in any data mining and KDD context, and generally speaking the return on investment to improve data quality depends on the structure of the processes within an organization. For instance, data entry processes should be carefully designed to avoid outliers and/or missing values. Section 6.3 will further analyze the results of the input selection experiments to indicate which type of information is required to predict customer churn.

6.2.2. Oversampling

For each of the performance criteria, when applying the Friedman test to compare the results with and without oversampling on an aggregate level, no significant impact on performance is detected. Hence, no clear conclusion can be drawn about the impact of oversampling on the performance of a customer churn prediction model.

The test of DeLong, DeLong, and Clarke-Pearson is applied to compare the performance of each classification technique with and without oversampling, for the results with input selection measured in AUC, which can be found in the right and the left lower panels of Table 6. The resulting *p*-values are reported in Table 7,

Table 7
The resulting *p*-values of the DeLong, DeLong, and Clarke-Pearson test applied to compare the performances of the classification techniques with and without oversampling, with input selection, on each data set separately. Performances that are not significantly different at the 95% confidence level are tabulated in bold face. Significant differences at the 99% level are emphasized in italics, and differences at the 95% level but not at the 99% level are reported in normal script. If the performance without oversampling is significantly better than the result with oversampling, the *p*-value is underlined.

Data set	O1	O2	O3	O4	O5	O6	D1	D2	D3	UCI	KDD
NN	**0.955**	**0.952**	**0.407**	**0.274**	**0.300**	0.012	**0.139**	**0.704**	**0.945**	0.000	0.000
linSVM	0.003	**0.928**	0.007	**0.694**	**0.718**	0.036	**0.133**	**0.643**	0.004	**0.616**	0.000
rbfSVM	0.000	**0.705**	**0.152**	0.013	**0.444**	0.000	0.026	**0.882**	0.022	**0.051**	0.000
linLSSVM	0.000	**0.869**	**0.643**	**0.339**	**0.670**	**0.183**	0.000	**0.365**	**0.853**	**0.549**	**0.730**
rbfLSSVM	0.000	**0.116**	**0.130**	**0.053**	0.000	0.000	0.000	**0.924**	**0.068**	**0.396**	0.002
RIPPER	0.000	0.000	0.000	0.019	**0.979**	0.000	0.000	**0.090**	0.000	**0.071**	0.000
PART	0.000	**0.367**	0.043	**0.661**	**0.283**	**0.071**	0.000	**0.580**	0.011	**0.293**	0.015
C4.5	0.000	0.013	**0.308**	**0.942**	**0.626**	0.000	0.000	0.021	**0.284**	0.002	0.000
CART	0.000	0.011	0.000	0.002	0.000	0.000	0.000	**0.849**	0.000	0.006	0.000
ADT	**0.137**	**0.937**	**1.000**	**0.196**	**0.444**	**0.051**	**0.919**	**0.373**	**0.220**	**1.000**	**0.617**
RF	**0.426**	**0.901**	**0.844**	**0.305**	**0.897**	0.023	**0.664**	**0.991**	**0.422**	**0.276**	**0.296**
LMT	0.000	0.000	0.000	**0.388**	**0.886**	0.000	0.000	**0.140**	**0.983**	**0.762**	0.000
Bag	0.000	**0.058**	0.011	**0.386**	**0.268**	**0.598**	0.000	**0.783**	0.000	**0.193**	0.000
Boost	0.019	**0.495**	**1.000**	**0.987**	**1.000**	**0.066**	**0.056**	**0.613**	**0.184**	0.013	**0.116**
RBFN	**0.839**	**0.169**	**0.338**	**0.214**	**0.156**	**0.295**	**0.052**	**0.773**	**0.090**	**0.851**	**0.374**
VP	0.000	0.044	0.000	**0.201**	0.001	0.000	0.000	**0.110**	0.000	0.000	**0.337**
Logit	**0.133**	**0.909**	**0.269**	**0.206**	**0.277**	**0.319**	**0.684**	**0.531**	**0.488**	**0.741**	**0.670**
kNN10	**0.957**	**0.985**	**0.701**	**0.552**	**0.821**	**0.319**	0.000	**0.232**	**0.891**	**0.760**	**0.499**
kNN100	**0.329**	**0.835**	**0.643**	**0.966**	0.003	**0.139**	0.000	**0.593**	**0.796**	0.020	0.000
BN	0.000	**0.234**	**0.850**	**0.714**	**0.591**	**0.371**	0.000	**0.179**	**0.223**	**0.551**	**0.489**
NB	**0.359**	**0.351**	**0.794**	**0.260**	**0.398**	0.000	**0.939**	**0.217**	**0.971**	**0.303**	**0.146**

indicating the probability that a difference in performance is due to chance. The *p*-values smaller than 0.05 and 0.01 indicate that a difference in performance is significant with a confidence level of respectively, 95% and 99%, and are tabulated in normal and italic script. Furthermore, when the effect of oversampling on the performance of a classification technique is found to be significant, but negative, then the reported *p*-value is underlined. Non-significant differences finally are tabulated in bold. The image of Table 7 is rather diffuse, since there seem to be as many positive as negative significant effects on performance, and in many cases the results are not significantly different.

As illustrated by Table 7, the effect of oversampling strongly depends on the data set and the technique that is applied. For instance, oversampling improves the performance of Ripper on 8 out of the 11 data sets, while the results of ADT and RBFN are never found to be significantly impacted. These last techniques are apparently able to learn properly even with a very skewed class distribution. Furthermore, differences in performance on data sets *O*4 and *D*2 are almost never found significant, and in case of data set *D*1, oversampling yielded a positive effect in 6 cases, a negative effect in 7 cases, and no significant effect in 8 cases, which illustrates the apparent randomness in the effect of oversampling on the predictive power. Therefore it is recommended to adopt an empirical approach when building a customer churn prediction model, and to consistently test whether oversampling provides better classification results or not.

6.2.3. Classification techniques

In the previous paragraphs input selection is found to be crucial in order to obtain good predictive power. Therefore only the aggregate results with input selection (i.e. the results in the lower two panels of Tables 4–6) are included to compare the classification techniques using the Friedman and post hoc Nemenyi tests. The Friedman test results in a *p*-value close to zero for each of the three performance measures, and both with and/or without oversampling, indicating significant differences in performance to exist among the applied techniques. We thus proceed by performing post hoc Nemenyi to compare all classifiers, as explained in Section 5.1. The results are plotted in Fig. 9(a)–(c), respectively for the maximum profit criterion, top decile lift, and AUC. The horizontal axes in these figures represent the average ranking of a technique on all data sets. The more a technique is situated to the left, the better its ranking. Techniques are represented by a line, the left end of this line depicts the actual average ranking, while the line itself represents the critical distance for a difference between two classifiers to be significant at the 99% confidence level. The dotted, dashed, and full vertical lines in the figure indicate the critical differences with the best performing technique at respectively the 90%, 95%, and 99% confidence level. A technique is significantly outperformed by the best technique if it is situated at the right side of the vertical line.

The specific differences between the rankings in Fig. 9(b) and (c) will be discussed in detail in the next section. In general, we can conclude from Fig. 9(a)–(c) that a large number of techniques do not perform significantly different. Although the reported results in Tables 4–6 are widely varying, the majority of techniques yield classification performances which are on average quite competitive to each other. The conclusions of this section are comparable to previous benchmarking studies in credit scoring (Baesens et al., 2003b) and software defect prediction (Lessmann et al., 2008), which also reported a flat maximum effect and a limited difference in predictive power between a broad range of classification techniques. Hence, the impact of the classification technique on the resulting performance is less important than generally assumed. Therefore, depending on the setting other aspects beside discriminatory power have to be taken into account when selecting a classification technique, such as for instance comprehensibility or operational efficiency (Martens et al., 2011). In many business settings, a comprehensible model will often be preferred over a better performing black box model, since an interpretable model allows the marketing department to learn about customer churn drivers, and provides actionable information to set up retention initiatives, as will be discussed in Section 6.3. Furthermore, comprehensibility allows to check whether the classifier functions intuitively correct and in line with business knowledge. The most interpretable types of models are rule sets and decision trees, but also logistic regression and Bayesian techniques result in comprehensible classification models. Neural networks or SVMs on the other hand result in complex,

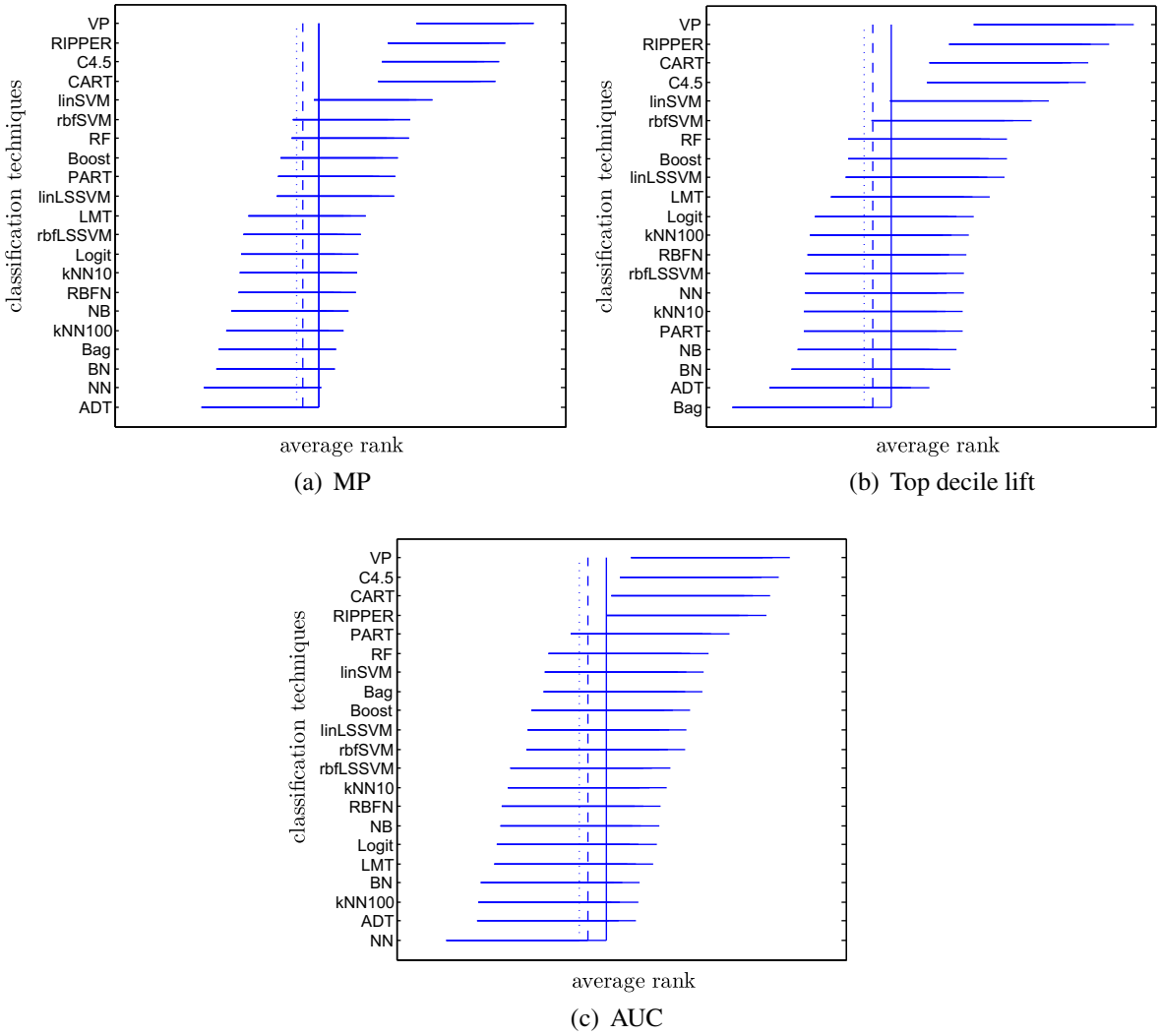

Fig. 9. Ranking of classification techniques and results of the post hoc Nemenyi test for the experiments with input selection using the three performance measures. The dotted vertical line indicates the 90% significance level, the dashed line the 95% level, and the full line the 99% level.

non-linear models, which are very hard to interpret and therefore called black box models. However, a combination of comprehensibility and good predictive power can also be achieved indirectly, by adopting a hybrid approach such as for instance rule-extraction from neural networks (Baesens et al., 2003a). The operational efficiency concerns the ease of implementation, execution, and maintenance of a model, which are represented by steps three and four in the process model discussed in Section 2. Rule sets and linear models are very fast in execution, and easy to implement and maintain. Nearest neighbor methods on the other hand do not result in a final model that can be implemented and executed straightforwardly, and have to calculate the k nearest neighbors each time a customer needs to be classified. Ensemble methods on the other hand involve a multitude of models that need to be executed, implemented, and maintained, and therefore typically score bad on this aspect.

Finally, Tables 4–6 provide a benchmark to churn prediction modeling experts in the industry to compare the performance of their customer churn prediction models.

6.2.4. Statistical performance measures vs. the maximum profit criterion

Fig. 10 compares the rankings of the classification techniques in terms of maximum profit, top decile lift, and AUC. For each performance measure, the techniques that are not significantly different at the 95% confidence level according to the Bonferroni–Dunn test are grouped within grey boxes. As can be seen from the figure, the rankings vary substantially over the different performance measures, but show as well some resemblances. At the top, ADT is best using MP and AUC, and second best using top decile lift, which indicates that this technique has an overall good performance. RF on the other hand, having the best top decile lift and third in terms of MP, is only classified fifteenth using AUC. RF apparently performs well regarding the customers it assigns the highest propensities to attrite, resulting in the best top decile lift. However, when taking into account the entire ranking of the customers the performance of RF declines sharply, as indicated by the AUC measure.

At the bottom of the rankings a number of techniques seems to perform bad for all three measures, such as for instance VP, rbfSVM, and linSVM. The bad performance of the latter two techniques is rather surprising, given the competitive performance results reported in the literature in other domains. However, this can be due to the fact that these classifiers had to be trained using smaller samples, possibly leading to poor discriminatory power. On the other hand, a remarkable difference in ranking exists regarding the rule induction techniques and decision tree approaches. Evaluating C4.5, RIPPER, CART, and PART using the maximum profit criterion yields average (RIPPER and CART) to good performance (C4.5 and PART). However, in terms of top decile lift

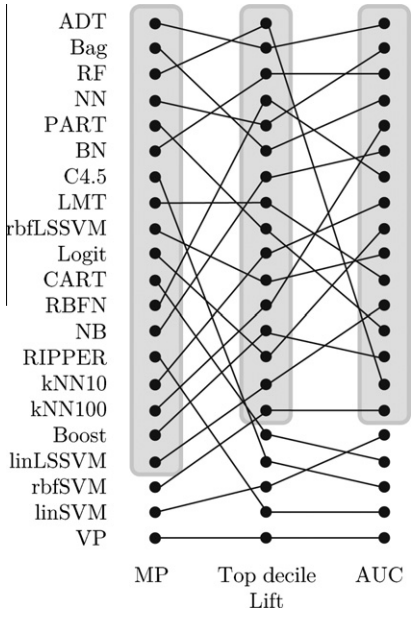

Fig. 10. Comparison of the rankings of classification techniques resulting from the benchmarking experiment using the maximum profit criterion, top decile lift, and AUC. The techniques that are not significantly different at the 95% confidence level according to a post hoc Nemenyi test, are grouped in the grey boxes for each performance measure.

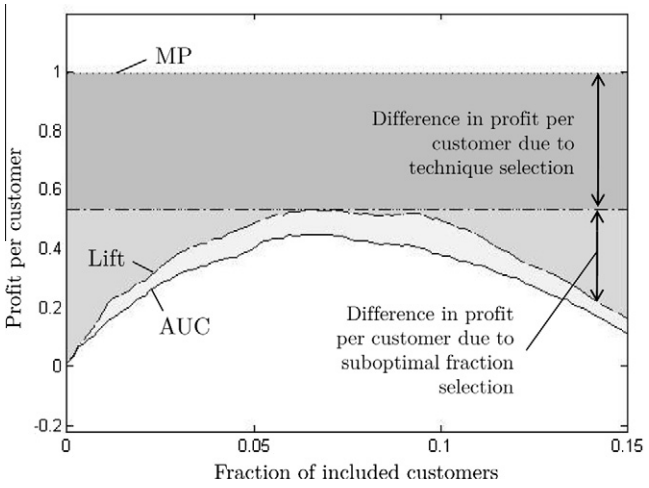

Fig. 11. Average profit per customer using maximum profit (dotted line), lift (dashed line), and AUC (full line).

or AUC, all except for PART are found to be significantly outperformed by the best performing technique. This can be explained by the fact that rule sets and decision trees do not provide a continuous output, and therefore their ROC curve has only as many points as there are rules or leaves, resulting in a discontinuous, piecewise monotone ROC curve and a fairly low AUC. Furthermore, these models only classify a fraction of the customers to be churners approximately equal to the fraction of churners in the data set, i.e. the base churn rate β_0. The base churn rate is in most data sets below 10% as indicated by Table 3, and therefore these techniques yield poor top decile lift. The optimal fraction of customers to include in a retention campaign however usually lies more near to the base churn rate. For instance, the average optimal fraction for technique PART on the data sets with input selection, both with and without oversampling, is equal to 3.38%, for C4.5 3.11%, for CART 3.45%, and for RIPPER 2.73%. The average over all techniques lies at 4%. Therefore the maximum profit criterion allows a more fair comparison regarding rule sets and decision trees.

Fig. 11 shows the average profit per customer over the eleven data sets, to illustrate the impact on the resulting profits of selecting a customer churn prediction model using each of the three performance measures. Both for top decile lift and AUC the resulting profit depends on the fraction of customers that is included in the retention campaign, whereas the MP criterion automatically determines the optimal fraction of customers to include and results in the maximum profit. Therefore, the average profit per customer generated by using the MP criterion is shown as a constant function in Fig. 11, represented by a dotted line and equal to 0.9978. The grey-most area between this function and the dash-dotted horizontal line below, indicating the maximum average profit using top decile lift, represents the difference in profit per customer resulting from suboptimal model selection using top decile lift. This difference equals 0.9978€ − 0.5321€ = 0.4677€ per customer. In this setting, for an operator with a customer base of one million customers, the difference in profit due to suboptimal customer churn prediction model selection yields half a million euros per retention campaign. On top of this, an additional difference in profit per customer will exist if a suboptimal fraction of customers is included in the retention campaign, indicated by the middle-grey area between the lower horizontal line and the top decile lift profit curve. For instance, if a model is selected using top decile lift, and the top decile of customers is effectively included in the retention campaign, then the difference in profit per customer amounts to 0.4677€ + 0.0352€ = 0.5321€.

6.3. Customer churn drivers and managerial insights

As a result of the input selection procedure, we are able to identify which type of data is most important to predict churn. The variables selected by the best performing techniques during the input selection procedure are analyzed and binned into the four categories. This results in the pie charts shown in Fig. 12, which represent the percentage of the selected variables belonging to each of the four types. Pie charts are shown for each data set separately, and on an aggregate level. Not all data sets include as many attributes of each type, resulting in some variance between the charts. However, as can be seen from the figure, most charts are similar to the aggregate chart. Therefore we can conclude that usage variables are the most solicited type of variable, and seem to be the best predictors of future churn. The other three types of data are used almost equally, each category representing roughly twenty percent of the selected variables. Hence, none of the four categories can be excluded from the data, and complete data sets containing information on each type will tend to yield better predictive performance.

The attributes present in the data sets used in this study are rather diverse, nevertheless a number of interesting findings can be reported regarding specific variables that are relevant to predict churn. As mentioned, marketing variables are of specific interest to the marketing department since they provide actionable information. Attributes related to the hand sets provided by the operator to the customer, such as the price and age of the current equipment, generally seem to be very relevant in order to predict churn. Also the number of contacts between operator and customer is typically a good predictor. Concerning socio-demographic variables, the age of a customer turns out to have good predictive power, but zip code or similar information on the other hand not at all, as might be expected. Examples of often selected financial variables are mean total monthly recurring charge, the type of plan, and the credit class of a customer. Usage statistics that are often solicited are the mean number of outbound voice calls (to a specific competitor), or even simply the total number of call detail

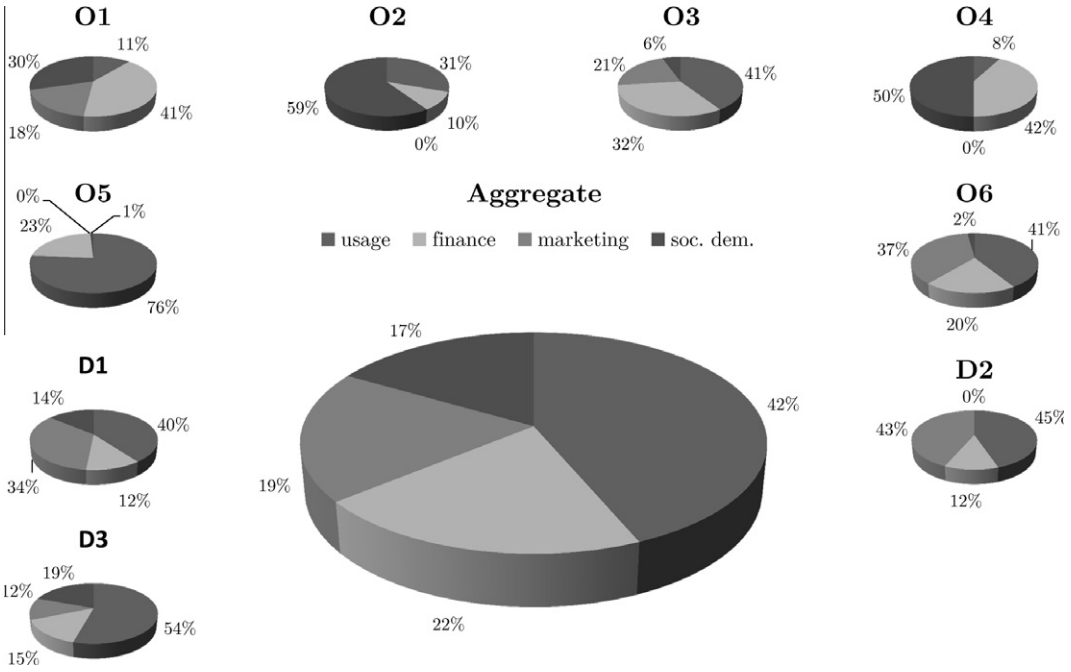

Fig. 12. Pie charts of the type of variables selected by the best performing techniques.

records. Trend variables such as recent to early usage ratio, this month to last months usage ratio, or simply an indicator of a positive or negative evolution seem to be frequently selected as well.

A typical issue when selecting variables is whether an attribute is a predictor for future churn, or a symptom of occurring churn. For instance, as a result of the modeling process a usage attribute indicating a strong drop in total minutes called might show to be strongly correlated with churn. However, this drop is likely to occur post-factum, when the event of churn has already taken place but is not logged in the data yet. Also a sudden peak in usage just before churn can typically be observed in churn data sets, and might therefore be considered a good predictor. This peak however usually indicates that the customer already decided to change from operator, and is consuming all the remaining minutes he already paid for. Therefore, such attributes cannot be used to predict customer churn since they do not allow to give an early warning, preferably even before the customer is actively considering to attrite. To successfully retain customers an early detection is of crucial importance in order to allow the marketing department to act on the results of the model, before the customer has made a final decision. This problem can partially be solved by lagging the data sufficiently to the churn event indicator. The lead of the churn flag on the attributes in the data sets included in this study is at least one month. A lead of more than three months on the other hand is expected to result in weak predictive power. Finally, a model always has to be checked and interpreted by a business expert to validate the selection of predictive variables, which again illustrates the need for a comprehensible model.

7. Conclusions and future research

Customer churn prediction models are typically evaluated using statistically based performance measures, such as for instance top decile lift or AUC. However, as shown in Sections 3 and 6 of this paper, this can lead to suboptimal model selection and a loss in profits. Therefore, in the first part of this paper a novel, profit centric performance measure is developed. Optimizing the fraction of included customers with the highest predicted probabilities to attrite yields the maximum profit that can be generated by a retention campaign. Since reducing the cost of churn is the main objective of customer churn prediction models, this paper advocates that the maximum profit should be used to evaluate customer churn prediction models.

In the second part of the paper a large benchmarking experiment is conducted, including twenty-one state-of-the-art predictive algorithms which are applied on eleven data sets from telecom operators worldwide, in order to analyze the impact of classification technique, oversampling, and input selection on the performance of a customer churn prediction model. The results of the experiments are tested rigorously using the appropriate test statistics, and evaluated using both the novel profit centric based measure and statistical performance measures, leading to the following conclusions.

Applying the maximum profit criterion and including the optimal fraction of customers in a retention campaign leads to substantially different outcomes. Furthermore, the results of the experiments provide strong indications that the use of the maximum profit criterion can have a profound impact on the generated profits by a retention campaign.

Secondly, the effect of oversampling on the performance of a customer churn prediction model strongly depends on the data set and the classification technique that is applied, and can be positive or negative. Therefore, we recommend to adopt an empirical approach, and as such to consistently test whether oversampling is beneficial.

Third, the choice of classification technique significantly impacts the predictive power of the resulting model. Alternating Decision Trees yielded the best overall performance in the experiments, although a large number of other techniques were not significantly outperformed. Hence, other properties of modeling techniques besides the predictive power have to be taken into account when choosing a classification technique, such as comprehensibility and operational efficiency. Rule induction techniques, decision tree approaches, and classical statistical techniques such as logistic regression and Naive Bayes or Bayesian Networks score well on all three aspects, and result in a powerful, yet comprehensible model that is easy to implement and operate. Therefore these

techniques are recommended to be applied for customer churn prediction modeling. Comprehensibility or interpretability is an important aspect of a classifier which allows the marketing department to extract valuable information from a model, in order to design effective retention campaigns and strategies. The comprehensibility of a model however also depends on the number of variables included in a model. Clearly a model including ten variables is easier to interpret than a model containing fifty variables or more.

This leads to a fourth conclusion, i.e. input selection is crucial to achieve good performance, and six to eight variables generally suffice to predict churn with high accuracy. Consequently, from an economical point of view it is more efficient to invest in data quality, than in gathering an extensive range of attributes capturing all the available information on a customer. Furthermore, the input selection procedure has shown that usage attributes are the most predictive kind of data. However, also socio-demographic data, financial information, and marketing related attributes are indispensable sources of information to predict customer churn. Moreover, marketing related attributes such as the hand set that is provided to a customer by the operator, are important sources of actionable information to design effective retention campaigns. Finally, this paper also provides benchmarks to the industry to compare the performance of their customer churn prediction models.

As a topic for future research, a fifth type of information remains to be explored on its ability to predict churn, i.e. social network information. Call detail record data is usually present in abundance, and can be analyzed to extract a large graph, representing the social network between the customers of an operator. Initial results of a pilot study indicate that social network effects play an important role in customer churn (Nanavati et al., 2008). A model that incorporates these effects as an extra source of information to predict churn therefore promises to yield improved performance. Finally, as mentioned in Section 3, uplift modeling and the mutual dependency between the probability of a customer to be retained, the customer lifetime value, and the predicted probability to churn are indicated as prime topics for future research.

Acknowledgements

We extend our gratitude to the Flemish Research Council for financial support (FWO Odysseus Grant B.0915.09), and the National Bank of Belgium (NBB/11/004).

References

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

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

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J., 2003b. Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society 54 (6), 627–635.

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

Bishop, C., 1996. Neural networks for pattern recognition. Oxford University Press, Oxford, UK.

Bolton, R., Lemon, K., Bramlett, M., 2006. The effect of service experiences over time on a supplier's retention of business customers. Management Science 52, 1811–1823.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Chapman & Hall, New York.

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

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.

Chawla, N., 2002. Smote: Synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 231.

Chawla, N., 2010. Data mining for imbalanced datasets: An overview. In: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer, US, pp. 875–886.

Cohen, W.W., 1995. Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning, pp. 115–123.

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.

Coussement, K., Van den Poel, D., 2008. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications 34, 313–327.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A., Joshi, A., 2008. Social ties and their relevance to churn in mobile telecom networks. In: Proceedings of the 11th international conference on Extending Database Technology: Advances in database technology, EDBT'08, pp. 697–711.

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

DeLong, E., DeLong, D., Clarke-Pearson, D., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44, 837–845.

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

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

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

Eiben, A., Koudijs, A., Slisser, F., 1998. Genetic modeling of customer retention. Lecture Notes in Computer Science 1391, 178–186.

Fawcett, T., Provost, F., 1997. Adaptive fraud detection. Data Mining and Knowledge Discovery 1-3, 291–316.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), 1996. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

Frank, E., Witten, I., 1998. Generating accurate rule sets without global optimization. In: Proceedings of the 15th International Conference on Machine Learning, pp. 144–151.

Freund, Y., Schapire, R., 1999. Large margin classification using the perceptron algorithm. Machine Learning 37 (3), 277–296.

Freund, Y., Trigg, L., 1999. The alternating decision tree learning algorithm. In: Proceedings of the 16th International Conference on Machine Learning, pp. 124–133.

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

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.

Glady, N., Baesens, B., Croux, C., 2009. A modified pareto/NBD approach for predicting customer lifetime value. Expert Systems with Applications 36 (2), 2062–2071.

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., Sriram, S., 2006. Modeling customer lifetime value. Journal of Service Research 9 (2), 139–155.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Springer.

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.

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

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

Kumar, D., Ravi, V., 2008. Predicting credit card customer churn in banks using data mining. International Journal of Data Analysis Techniques and Strategies 1 (1), 4–28.

Landwehr, N., Hall, M., Eibe, F., 2005. Logistic model trees. Machine Learning 59 (1), 161–205.

Lariviere, B., Van den Poel, D., 2005. Predicting customer retention and profitability by using random forest and regression forest techniques. Expert Systems with Applications 29 (2), 472–484.

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

Lessmann, S., Baesens, B., Mues, C., Pietsch, S., 2008. Benchmarking classification models for software defect prediction: A proposed framework and novel findings. IEEE Transactions on Software Engineering 34 (4), 485–496.

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.

Martens, D., De Backer, M., Haesen, R., Snoeck, M., Vanthienen, J., Baesens, B., 2007. Classification with ant colony optimization. IEEE Transactions on Evolutionary Computation 11 (5), 651–665.
Martens, D., Vanthienen, J., Verbeke, W., Baesens, B., 2011. Performance of classification models from a user perspective. Decision Support Systems 51, 782–793.
Mizerski, R., 1982. An attribution explanation of the disproportionate influence of unfavourable information. Journal of Consumer Research 9, 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.
Nanavati, A., Singh, R., Chakraborty, D., Dasgupta, K., Mukherjea, S., Das, G., Gurumurthy, S., Joshi, A., 2008. Analyzing the structure and evolution of massive telecom graphs. IEEE Transactions on Knowledge and Data Engineering 20 (5), 703–718.
Nemenyi, P., 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.
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.
Piatetsky-Shapiro, G., Masand, B., 1999. Estimating campaign benefits and modeling lift., 185–193.
Piramuthu, S., 2004. Evaluating feature selection methods for learning in data mining applications. European Journal of Operational Research 156 (2), 483–494.
Provost, F., 2005. From data mining to data science: Evaluation for predictive modeling. Course Notes for B20.3336.30: Data Mining for Business Intelligence, Stern School of Business, New York University, New York, NY, USA.
Provost, F., Jensen, D., 1999. Evaluating machine learning, knowledge discovery, and data mining. In: Tutorial presented at the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99) and at the Sixteenth National Conference on Artificial Intelligence (AAAI-99).
Provost, F., Fawcett, T., Kohavi, R., 1998. The case against accuracy estimation for comparing induction algorithms. In: Proceedings of the Fifteenth International Conference on Machine Learning. Morgan Kaufmann, pp. 445–453.
Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
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.
Rust, R., Zahorik, A., 1993. Customer satisfaction, customer retention, and market share. Journal of Retailing 69 (2), 193–215.
Stum, D., Thiry, A., 1991. Building customer loyalty. Training and Development Journal 45 (4), 34–36.
Suykens, J., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Processing Letters 9 (3), 293–300.
Swets, J., Pickett, R., 1982. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic Press, New York.
Tan, P., Steinbach, M., Kumar, V., 2006. Introduction to Data Mining. Addison Wesley, Boston, MA.
Thomas, L., Edelman, D., Crook, J. (Eds.), 2002. Credit Scoring and its Applications. SIAM.
Van den Poel, D., Lariviere, B., 2004. Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research 157 (1), 196–217.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York, Inc.,, New York, NY, USA.
Verbeke, W., Martens, D., Mues, C., Baesens, B., 2011. Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications 38, 2354–2364.
Wei, C., Chiu, I., 2002. Turning telecommunications call details to churn prediction: A data mining approach. Expert Systems with Applications 23, 103–112.
Zeithaml, V., Berry, L., Parasuraman, A., 1996. The behavioural consequences of service quality. Journal of Marketing 60 (2), 31–46.