Decision Support Systems 95 (2017) 27–36

Contents lists available at ScienceDirect

Decision Support Systems

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

A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry

Kristof Coussement [a,*], Stefan Lessmann [b], Geert Verstraeten [c]

[a] IESEG School of Management, Université Catholique de Lille (LEM, UMR CNRS 9221), Department of Marketing, 3 Rue de la Digue, F-59000 Lille, France
[b] Humboldt-University of Berlin, Unter den Linden 6, D-10099 Berlin, Germany
[c] Python Predictions, Avenue R. Van den Driessche 9, B-1150 Brussels, Belgium

ARTICLE INFO

Article history:
Received 12 April 2016
Received in revised form 24 November 2016
Accepted 27 November 2016
Available online 29 November 2016

Keywords:
Predictive analytics
Data preparation techniques
Churn prediction

ABSTRACT

Data preparation is a process that aims to convert independent (categorical and continuous) variables into a form appropriate for further analysis. We examine data-preparation alternatives to enhance the prediction performance for the commonly-used logit model. This study, conducted in a churn prediction modeling context, benchmarks an optimized logit model against eight state-of-the-art data mining techniques that use standard input data, including real-world cross-sectional data from a large European telecommunication provider. The results lead to following conclusions. (i) Analysts better acknowledge that the data-preparation technique they choose actually affects churn prediction performance; we find improvements of up to 14.5% in the area under the receiving operating characteristics curve and 34% in the top decile lift. (ii) The enhanced logistic regression also is competitive with more advanced single and ensemble data mining algorithms. This article concludes with some managerial implications and suggestions for further research, including evidence of the generalizability of the results for other business settings.

1. Introduction

Many companies suffer the substantial problem of customer defection, due to fierce competition resulting from saturated markets, dynamic market conditions, and continuous introductions of new competitive offerings. In response, many of them have switched from an offer-centric strategy, designed to sell as many offerings as possible, to a customer-oriented retention approach that explicitly seeks to reduce churn [2]. A key enabler of targeted retention programs is the capacity to perform computerized searches for and identifications of customers who exhibit a high propensity to end their relationship with the company, or customer churn prediction [24,56]. Concretely, customer churn prediction is the practice of assigning a churn probability to each customer in the company database, according to a predicted relationship between that customer's historical information and its future churning behavior. Practically, the probability to end the relationship with the company is then used to rank the customers from most to least likely to churn, and customers with the highest propensity to churn receive marketing retention campaigns. Two challenges impact the success of these campaigns. First, it is crucial to develop appropriate marketing tactics, to convince potential churners to stay. A field experiment designed to test the impact of three types of retention actions revealed, for example, that targeting at-risk customers with a customer satisfaction survey yield the best retention performance [8]. Second, companies could improve the returns on their investments in retention campaigns by distinguishing potential churners who are more susceptible to marketing actions (i.e., *persuadable customers*) from those who will leave anyway, whether they will receive a retention offer or not (i.e., *non-persuadable customers*). This effort is known as net effect or uplift modeling [36,58].

The smart selection of customers, using predictive modeling, thus is of crucial importance. Done well, it can result in substantial additional profits compared with random selections of customers for targeted retention campaigns [46,61]. The ample variations in predictive performance across various methods also have impacts on the bottom line. In one study, a company with 5 million customers that contacted 10% of them for a customer retention campaign attained additional profits in the hundreds of thousands of dollars when it chose the most accurate method [49].

Yet customer churn prediction remains a complex process, containing various decision points for analysts [43]. The established Cross-Industry Standard Process for Data Mining (CRISP-DM) breaks it down into six distinct stages: business understanding, data understanding,

* Corresponding author.
E-mail address: K.Coussement@ieseg.fr (K. Coussement).

http://dx.doi.org/10.1016/j.dss.2016.11.007
0167-9236/

data preprocessing, modeling, evaluation, and deployment. For this study, we focus on one of the most time-consuming, critical steps, data preprocessing[1] [11]. As is true for any predictive modeling setting (e.g., direct marketing [14,21], demand modeling [40]), the "garbage in, garbage out" rule applies to churn prediction. That is, the choices that researchers make during the data preprocessing step influence the ultimate success of the classification [49]. Occurring immediately before the predictive modeling step, preprocessing consists of two main tasks, as indicated in Fig. 1: data reduction and data preparation [53]. For example, imagine a churn data set that contains ten customers (*Customer ID* as a unique identifier), two independent variables (continuous variable *Age* and categorical variable *Income*, with low, medium, and high categories), and a dependent variable that indicates whether a customer has churned (*Churn?*) (Fig. 1, Panel a). Data reduction techniques aim to reduce the dimensionality of the data set by extracting the most relevant variables, with the greatest power for discriminating between churners and non-churners (*variable selection*) (Fig. 1, Panel b), or else selecting the most representative customers (*sampling*) (Fig. 1, Panel c). Data preparation methods[2] often refer to the transformation of the original variables into a form that supports a particular classification algorithm, i.e. variable transformation methods. For example, a logistic regression cannot directly handle categorical variables such as *Income*, so a variable transformation method converts the information in the original *Income* variable into a new form that fits a logistic regression model. In our example, we create two new dummy variables, *Income#Low* and *Income#Medium* (Fig. 1, Panel d).

Previous research also addresses customer churn prediction, as recently reviewed by Risselada, Verhoef & Bijmolt [55] and Verbeke et al. [61] and summarized in Table 1. In reviewing prior studies' uses of data reduction and preparation methods, whether they contrast multiple churn prediction algorithms, and whether they use logistic regression as benchmark prediction model, we derive several conclusions from Table 1.

Most churn studies report their data reduction procedures related to the independent variables. Only a few of them provide insights into their data preparation techniques; only Moeyersoms & Martens [46] benchmarks the impact of different data preparation treatment (DPT) strategies on customer churn predictive performance. However, their case study considers DPTs for only high-cardinality, categorical variables in a small-scale setup, containing ten variables. As a result of their dataset, these authors define only three variables as high-cardinal, and thus treatable by multiple data preparation methods to investigate their impact on the predictive churn performance. Therefore, we extend their work and provide a large-scale (in terms of variables) benchmarking case study to investigate the impact of multiple DPTs for both categorical and continuous variables on customer churn prediction performance.

Moreover, Table 1 shows that churn prediction literature tends to contrast various algorithms within the same paper, and the majority of these papers acknowledge logistic regression as a viable benchmark algorithm [61]. From this review, we therefore derive two key research questions:

- Does fine-tuning the DPT method, for categorical and continuous variables, exert a positive impact on logistic regression performance?
- Can logistic regression models with fine-tuned data compete with more advanced benchmark classifiers that use standard DPTs?

These questions are pertinent for two main reasons. First, from an academic perspective, the focus on developing and testing novel, advanced classification algorithms in churn prediction literature might be misplaced if a proper DPT suffices to erode their advantages over simple, standard methods, such as logistic regression in particular. Second, from a managerial point of view, DPT is a non-invasive approach to increase predictive accuracy. It offers the potential to improve deployed churn models and increase the effectiveness of customer retention management, without invoking high costs to implement novel prediction solutions (e.g., software packages, user training, consulting).

In the next section, we elaborate the alternative DPTs for categorical and continuous variables. After describing our focal data set, logistic regression model, and variable selection procedure, we detail our research design and evaluation metrics. We then present our empirical results and conclude.

2. Data preparation treatments

To provide a clearer view on DPTs and their necessity, we propose a conceptual framework that decomposes the data preparation process into two steps: value *transformation*, followed by value *representation*. The decisions made in each step lead to a particular DPT for the independent variables in any given data set. The first step converts independent variables into discrete variables that contain fewer unique data points. It minimizes the complexity of the input data and increases the scalability of the modeling process, while still preserving classification performance [22,44]. Furthermore, value transformation methods anticipate handling the missing values as a separate category and thus avoid applying additional missing value strategies. This tactic is interesting, because logistic regression has no natural means to deal with missing values [31].

The value representation step ensures that the format of the discrete variables produced in the first step is fit for churn prediction. That is, the representation methods ensure that the independent variable characteristics overcome the (statistical) constraints of the logistic regression model, including its inability to directly estimate the impact of multi-category discrete variables on the churning event and its lack of means to incorporate nonlinearity between the logit of the independent variables and the churn behavior [31]. In Table 2, we describe the most popular methods, according to our proposed DPT framework.

The value transformation methods differ in their ability to tackle outliers (*Outliers*) and to use the homogeneity in the churn behavior to convert independent variables into discrete variables that contain fewer unique data points (*Churn link*). Outliers distort logistic regression performance [35,52]. However, some value transformation methods offer an indirect solution, in that they reduce the outliers' negative impact on predictive performance by incorporating those outliers into a larger category, within the new discrete variable. Other methods convert independent variables into discrete variables during the transformation process whereby the categories in the new discrete variable are created based on the homogeneity of the churn behavior within these categories. To illustrate the use of churn behavior information within value transformation methods, consider the conversion of an independent variable *tariff plan* with n distinct levels into a transformed variable with $k < n$ levels. Such conversion can use churn behavior information in that it merges category levels that are associated with similar churn frequencies in the training set (see Section 3.1 below). This idea can be implemented by means of a univariate decision tree, which categorizes the original variable *tariff plan* so as to maximize the homogeneity of churn behavior (i.e., churner vs. non-churner).

The value representation method characteristics also differ along two dimensions. First, some methods do not increase the variable dimensionality, whereas others do (*Dimensionality*). Second, some value representation methods tend to rely on customer churn behavior information obtained from the dependent variable (*Churn link*). We next describe the most popular DPT methods in more detail, as they relate to various business classification fields, including customer churn prediction.

[1] We do refer to data preprocessing as the process that follows after the data collection, data extraction and feature engineering (or variable construction) phases.

[2] We do acknowledge that other data preparation methods like data cleaning, missing value handling and outlier detection strategies exist. In the remainder of the paper, the term data preparation method refers to variable transformation method.

Visualization of data reduction and preparation

Panel a: Churn dataset

Customer ID	Age	Income	Churn?
1	24	Low	Yes
2	50	High	No
3	32	Low	Yes
4	40	Medium	No
5	26	Low	Yes
6	65	High	No
7	71	Medium	No
8	43	High	No
9	45	High	No
10	59	Medium	No

Panel b: Data reduction - variable selection

Customer ID	Age	Income	Churn?
1	24	Low	Yes
2	50	High	No
3	32	Low	Yes
4	40	Medium	No
5	26	Low	Yes
6	65	High	No
7	71	Medium	No
8	43	High	No
9	45	High	No
10	59	Medium	No

Panel c: Data reduction - sampling

Customer ID	Age	Income	Churn?
1	24	Low	Yes
2	50	High	No
3	32	Low	Yes
4	40	Medium	No
5	26	Low	Yes
6	65	High	No
7	71	Medium	No
8	43	High	No
9	45	High	No
10	59	Medium	No

Panel d: Data preparation

Customer ID	Age	Income#Low	Income#Medium	Churn?
1	24	1	0	Yes
2	50	0	0	No
3	32	1	0	Yes
4	40	0	1	No
5	26	1	0	Yes
6	65	0	0	No
7	71	0	1	No
8	43	0	0	No
9	45	0	0	No
10	59	0	1	No

Fig. 1. Visualization of data reduction and preparation.

2.1. Value transformation step

To create a discrete variable with fewer unique data points, this first step uses either a remapping strategy for categorical variables or discretization methods for continuous variables.

2.1.1. Categorical variables: remapping

A remapping strategy reassigns the values of a categorical variable [50], to obtain an optimized, discrete variable in which the original categories remap onto a new categorical variable, according to their relationships with the dependent variable. Decision trees are well suited for this transformation task for categorical variables, because they use frequency comparisons to build the churn model and can handle many categories simultaneously. The decision tree–based remapping process starts with the entire customer data set, or the root node, then successively splits the data into smaller subsets or internal nodes on the basis of the values of the categorical variable. The purpose is to find homogenous groups of customers who belong to the same predefined target group—for this study, churners or non-churners. Customers who belong to the same terminal node or leaf then get grouped

Table 1
Data reduction and preparation methods in prior churn prediction literature.

Study	Data Reduction		Data Preparation	Contrasting Churn Algorithms	Logit Model as Benchmark
	Variable Selection	Sampling			
Eiben, Koudijs and Slisser [20]		X	X	X	X
Datta, Masand, Mani and Li [16]	X	X			
Mozer, Wolniewicz, Grimes, Johnson and Kaushansky [47]				X	X
Wei and Chiu [62]	X	X			
Au, Chan and Yao [1]	X	X	X	X	
Hwang, Jung and Suh [33]	X			X	X
Buckinx and Van den Poel [7]				X	X
Lariviere and Van den Poel [41]				X	X
Hung, Yen and Wang [32]	X	X		X	
Lemmens and Croux [42]	X	X		X	X
Neslin, Gupta, Kamakura, Lu and Mason [49]	X	X		X	X
Burez and Van den Poel [8]				X	X
Coussement and Van den Poel [13]		X		X	X
Kumar and Ravi [38]	X	X		X	X
Xie, Li, Ngai and Ying [64]	X	X		X	
Risselada, Verhoef and Bijmolt [55]	X	X		X	X
Verbeke, Martens, Mues and Baesens [61]	X	X	X	X	X
Moeyersoms and Martens [46]			X	X	X
This study	X	X	X	X	X

Table 2
Building blocks of the DPT framework.

Data Preparation Step	Variable Type	Description	Transformation		Representation	
			Outliers	Churn link	Dimensionality	Churn link
Transformation	categorical	No regrouping				
		Decision tree–based remapping	X	X		
	continuous	Decision tree–based discretization	X	X		
		Equal frequency discretization	X			
		Equal width discretization				
Representation		Dummy coding			↑	
		Incidence replacement			=	X
		Weight-of-evidence conversion			=	X

Note: X = present; ↑ = increase; = no increase.

together and receive the same category label within the new discrete categorical variable.

2.1.2. Continuous variables: discretization

Discretization, or binning, is one of the most popular DPT techniques to convert continuous variables into a discretized form [22]. Three well-known discretization methods for continuous variables are equal frequency, equal width, and decision tree–based discretization. First, equal frequency discretization converts continuous variables into categorical ones. The goal is to create a certain number of bins b, such that each bin contains the same number of customers. Practically, the values of variable x get sorted in ascending order, and the size of the bins is calculated as

$$\frac{\text{total number of customers in dataset}}{b} \tag{1}$$

Second, equal width discretization is a simple method: It sorts the values of a continuous variable, then uses the range of these values to determine b equally ranged bins, where b is a parameter chosen by the analyst. For each continuous variable x, the following procedure applies: Given x_{min} as the minimum of variable x and x_{max} as its maximum, bin width Ω can be calculated as

$$\frac{x_{max} - x_{min}}{b} \tag{2}$$

Therefore, this method creates b bins with boundaries at $x_{min+i} \times \Omega$, where $i = 1, 2, \ldots, b - 1$.

Third, a decision tree can discretize a continuous variable. A decision tree first is built for variable x, and then the original values of the continuous variable x get regrouped, according to the leaf in the decision tree to which the customer belongs. Each leaf in the decision tree receives a unique value, which allows relabeling the original values of the variable x, after accounting for its relationship with the dependent variable.

2.2. Value representation step

Through value transformation, all categorical and continuous independent variables become discrete variables. The value representation step then depicts these discrete variables in a form fit for model building. We detail three well-known representation methods: dummy coding, weight-of-evidence conversion, and incidence replacement methods.

2.2.1. Dummy coding

A popular method in churn prediction studies represents the discrete variables for subsequent processing using dummy coding [61]. This technique creates $v - 1$ dummy variables, where v equals the number of distinct values of the discrete variable [53]. In a quantitative analysis, the dummy variable provides a numeric stand-in for a qualitative fact or category of the discrete variable. A dummy variable is binary, equal to 0 or 1, and indicates the absence or presence of a particular qualitative characteristic. The main disadvantage is that too many dummy variables can hinder the generalizability and interpretability of a prediction model [29].

2.2.2. Weight-of-evidence conversion and incidence replacement

To re-summarize discrete information into a continuous variable, taking the information of the dependent variable into account, a weight-of-evidence (WOE) conversion and its simplified version, incidence replacement, replace category labels with churn information about the dependent variable. Both methods are common ways to represent discrete independent variables in various business contexts [57, 59]. Whereas an incidence replacement method replaces the category labels with category-specific proportions of churners or churn incidence, the WOE technique represents a more advanced version that represents the strength of a category to separate churners from non-churners, according to the following formula for the new value of a category of a discrete variable:

$$\ln\left(\frac{\text{Proportion of churners in category}}{\text{Proportion of non}-\text{churners in category}}\right) \tag{3}$$

Consider the variable *city* as example. Both the WOE and incidence replacement methods calculate, for each level of the *city* variable, the proportion of customers who churned, which indirectly reveals the proportion of customers who did not churn. The incidence replacement method replaces the city name label in the variable *city* with the proportion of churners living in that city; the WOE method applies Eq. (3) to come up with the new value for the city name. Managerially, the value of WOE is 0 if the odds of $\left(\frac{\text{Proportion of churners in category}}{\text{Proportion of non}-\text{churners in category}}\right)$ equal 1. If the *Proportion of non-churners in category* is larger (smaller) than the *Proportion of churners in category*, the WOE value is a negative (positive) number. The larger the discrepancy between the nominator and the denominator, the better the discrimination power, and thus the larger the absolute WOE value is.

The main advantage of both methods is that each discrete variable gets replaced by only one output variable that contains summarized information. They thus avoid creating many additional variables, which is common for dummy coding. Moreover, these approaches take the relationship between the independent variable and the dependent variable into consideration. Finally, the value range of all independent variables is comparable following these procedures. Other than the slightly easier implementation offered by the incidence replacement method compared with the WOE method, the main difference between the two techniques is the way they use the churn information. Incidence replacement uses exactly the proportion of churners in a particular variable category as the value to replace that category label after the representation step (i.e., linear conversion); the WOE technique introduces

non-linearity in the representation step by converting the proportion of churners in a particular variable category according to Eq. (3). Furthermore, the WOE technique uses a logarithmic transformation, which pushes the distribution of the WOE-transformed variable toward a less skewed distribution, such that it might benefit certain classification methods.

3. Research procedure

3.1. Data

The residential database of a large European mobile telecommunication provider offers an ideal test environment, because many telecommunication companies experience fierce competition and high churn rates, and managing churn rates effectively is a key priority [42]. Customers are in contractual relationships with the firm, so the churn variable indicates whether they end their mobile subscription at the end of the contract. The total data set contains 30,104 customers with a churn incidence of 4.52%. Because we had access to the internal data structure, we included 956 churn drivers: 156 categorical and 800 continuous variables, including customer behaviors (e.g., minutes of outgoing calls), customer–company interaction variables (e.g., number of contacts with the support center), subscription-related variables (e.g., type of tariff plan), and customer demographics (e.g., gender). We cannot offer any more detailed information, because of our confidentiality agreement with the company.

To avoid problems of overfitting, we split this original data set into a training set (50% or 15,052 customers), a selection set (20% or 6021 customers), and a validation set (30% or 9031 customers). The training set serves to build the classification models; the selection set reveals the parameters for the optimal data preparation techniques and classifiers; the validation set mimics the real-life performance of the classifiers using the fine-tuned parameters. Both our research questions rely on the same data partitioning scheme.

3.2. Scoring model and variable selection

A logistic regression model serves as the baseline throughout this article. Such models are frequently used binary choice models for marketing, and they reflect random utility theory, which states that when decision makers face a set of choices, they choose the one associated with the maximum utility. For a given training set with N labeled training examples $\{(x_i,y_i)\}$ for $i = 1, 2, \ldots, N$ with input data $x_i \in R^n$, as well as corresponding binary target labels $y_i \in \{0,1\}$, a logistic regression estimates the probability $P(y = 1|\mathbf{x})$ by

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(w_0 + \mathbf{w}\ \mathbf{x}))}, \quad (4)$$

where $\mathbf{x} \in R^n$ equals an n-dimensional input vector, $\mathbf{w}$ is the parameter vector, and w_0 refers to the intercept [31]. Both w_0 and $\mathbf{w}$ are estimated using the maximum likelihood procedure.

Customer churn data sets usually contain many customer characteristics, and some DPT operations increase the number of independent variables, such as by introducing dummy indicators in the representation stage. However, a benchmark study has shown that not all customer metrics contribute to churn prediction performance [61]; usage variables seem to be the most solicited type of variable and the best predictors of churn, whereas socio-demographic, financial, and marketing variables contribute about equally in distinguishing churners from non-churners. These findings align with the recognition that customer churn data sets often contain multiple variables that are redundant or irrelevant, so variable selection is required. The advantages of a variable selection method are manifold [61]. First, too many redundant, irrelevant variables can hinder prediction performance. The result of variable selection is the improvement of the prediction model's stability, through reduced overfitting [34]. Second, variable selection builds neater models, which increases their interpretability for users. Third, when prediction models include fewer variables, operational efficiency improves during model industrialization. We use a correlation-based feature selection heuristic [25], which relies on an easy, rapid, correlation-based computation of variable subsets prior to the model-building stage. It scores the variable subsets by trading off the average correlation with the dependent variable against the average intercorrelation within a subset. If two variables are perfectly correlated, only one enters the final subset. Variable subsets with high average correlation with the dependent variable and low intercorrelations earn higher scores and are preferable. This heuristic is especially well-suited for churn prediction because it (1) can process data sets with both continuous and categorical variables, (2) avoids a computationally expensive estimation of the actual prediction model to assess variable candidates during the selection stage, and (3) can remove redundant variables, whereas approaches that score variables individually would be unable to detect these redundancies.

3.3. Experimental setting

Table 3 summarizes the DPTs we consider in the remainder of this study. To avoid overfitting, the DPT methods and their parameters are optimized using the training and/or selection set. The validation set serves to mimic real-life prediction performance and thus to benchmark the optimized DPT methods. Specifically, the Chi-square Automatic Interaction Detector (CHAID)[3] is used for the decision-tree based DPT techniques. They find the most ideal remapping (discretization) strategy for a categorical (continuous) variable by building a univariate decision tree for that variable on the training set. The CHAID decision trees use the 5% significance criterion as implemented in the original paper by [37]. To avoid overfitting, the decision trees get complemented by an additional pruning step, applied to the selection set, that searches for the smallest subtree that does not cause a significant drop in prediction performance [19]. To roll out the decision-tree based DPT method to the validation set (by the time of rolling out the model the dependent variable is not available), the values of the categorical (continuous) variable in the validation set are pushed through the optimal decision tree, which remaps (discretizes) the original categories (values) into new categories. Furthermore, to guarantee methodological consistency, we identify the boundary values of the bins for equal frequency and equal width discretization on the training set, then use the resulting settings for the selection and validation sets. Both use a bin size b equal to 10 [3,4,12,65]. Testing other bin sizes b with $b \in \{5,15\}$ indicated minor changes in classification performance but did not change the results or conclusions. Finally, incidence replacement and WOE conversion replace the category labels in the selection and validation set by the churn incidences of the corresponding categories in the training set.

Regarding our first research question, or the impact of DPTs for logistic regression modeling on churn prediction performance, we tested their impact using a $2 \times 3 \times 3$ between-subjects design, such that we created 18 versions of the initial data set by combining a transformation strategy for the categorical and continuous variables with a representation technique. Specifically, there are two levels of transformation methods for the categorical variables (TransCat: nog vs. dt_org_cat), three levels of transformation methods for the continuous variables (TransCont: dt_org_cont vs. efd vs. ewd), and three levels of representation methods (Rep: dum vs. inc vs. WOE). The number of variables after variable selection across DPTs ranges from 49 to 92. For the first research question, we randomly created 100 bootstrap samples for each of the 18 versions of the training set, then built a customer churn prediction model for each of the 1800 bootstrap samples that we applied to

[3] A CHAID decision tree acknowledges that when building decision trees, the significance of the chi-square test determines the next best split at each step [37]. It builds decision trees where more than two branches can be attached to a single node.

Table 3
Abbreviations for DPT.

Data Preparation Stage	Variable Type	Description	Abbreviations	
Transformation	Categorical	No regrouping	TransCat	nog
		Decision tree–based remapping		dt_org_cat
	Continuous	Decision tree–based discretization	TransCont	dt_org_cont
		Equal frequency discretization		efd
		Equal width discretization		ewd
Representation		Dummy coding	Rep	dum
		Incidence replacement		inc
		Weight-of-evidence conversion		WOE

the selection set. The performance measures for the latter data sets fill the $2 \times 3 \times 3$ design. Each of the 18 cells in the design thus contains 100 observations that represent the churn prediction performance measures for the selection set.

Our second research question pertains to how well a logistic regression model, built on the data set using an optimal DPT (LOGIT-DPT), performs relative to the advanced single and ensemble benchmark classifiers that rely on standard preparations. We therefore derived another version of the data set with a standard preparation in line with Moeyersoms & Martens [46], in which we encoded the categorical variables using binary dummy variables and normalized the continuous variables. The same data partition of the dataset in training (50%), selection (20%) and validation (30%) set is used as for the first research question. The performance measures from the validation set provide the benchmark for the LOGIT-DPT relative to the benchmark classifiers. Accordingly, we compared the performance of LOGIT-DPT against eight diverse classifiers, as shown in Table 4.

Several considerations based on our literature review motivated our choice of benchmark classifiers for the cross-sectional churn data. In particular, ensemble models that pool the predictions of multiple base classifiers predict customer attrition with high accuracy [49,61], so we compared our LOGIT-DPT against ensemble models based on classification trees such as bagging, boosting and random forests [13,42]. Furthermore, neural networks have shown to perform well [7], while kernel methods also are popular in machine learning and other disciplines; previous research suggests they are effective for churn modeling too [10]. Therefore, we included kernel methods in the form of support vector machines [13,15]. For completeness, we compared our LOGIT-DPT with some well-known (statistical) classifiers, such as naïve Bayes, decision trees, and Bayesian networks [61], because they nicely balance their performance on accuracy, comprehensibility and operational efficiency. Interested readers may find a more comprehensive description of the benchmark classifiers in Hastie et al. [29], for example.

Most benchmark classifiers include some meta-parameters that must be optimized, such as the number of hidden nodes in a neural network, the kernel function parameters in a support vector machine, or the number of base models in an ensemble classifier. Meta-parameters support the fine-tuning of an algorithm to a specific modeling task. The predictive performance of a classifier often depends on an appropriate choice of meta-parameters. To develop strong benchmarks, we defined a set of candidate values for each meta-parameter and empirically determined the most predictive setting with the selection set. We obtained candidate settings from prior literature and previous classifier comparisons in particular [61]. If a benchmark classifier exhibits multiple meta-parameters, we define candidate values for each parameter and test all possible value combinations. For example, random forests exhibit two meta-parameters: the number of decision trees in the ensemble and the number of independent variables included to create an individual tree [6]. Considering five and six candidate settings, respectively, for these meta-parameters, we created $5 \times 6 = 30$ random forests classifiers and assessed their performance using the selection data set. We employed the random forest model with the highest accuracy on the selection data set as a benchmark for the LOGIT-DPT on the validation data set. In total, we built 560 algorithm versions on the selection set to optimize the eight benchmark algorithms.

3.4. Evaluation metrics

To assess the performance of all classification methods, we used the area under the receiver operating characteristics curve (AUC)[4] and the top decile lift (TDL). The AUC is used for following reasons in this research study. First, it is frequently applied in customer churn prediction studies [61]. Second, compared with other evaluation metrics, such as the percentage correctly classified, the AUC also offers a more robust evaluation metric that accounts for the overall performance of a classification technique by considering all possible cut-off points on the receiver operating characteristics curve. Finally, this ranking-based measure of posterior churn probabilities is intuitively clear and offers clear statistical interpretations: The AUC is the estimated probability that a randomly chosen churner has a higher posterior churn probability than a randomly selected non-churner. Thus if a churn model indicated an AUC of 0.60, this means that if one randomly picks an actual churner and a non-churner from the dataset, then 60% of the times the churner will have a higher churn probability output by the classifier than the non-churner. A random model has an AUC of 0.50.

The choice of TDL also is not arbitrary, for following reasons. First, it is a popular evaluation metric in the churn prediction domain [42]. Second, the TDL evaluates the performance of classification methods in terms of their managerial value. It focuses on customers that the prediction algorithm indicates are most at risk of leaving the company. Practically, these customers are first sorted, from those predicted to be most likely to churn down to those predicted to be least likely to churn, according to the churn probabilities derived from the prediction model. The proportion of churners in the top 10% then gets compared with the proportion of churners in the total data set, and the increase in churn density is the TDL. For example, a TDL of 2 means that the density of churners in the top 10% is twice the density of churners in the total data set. A higher TDL indicates a better prediction algorithm, and a TDL >1 identifies a model that outperforms a random selection of customers. Third, as Neslin et al. [49] shows, the TDL is indirectly linked to the profitability of a retention campaign.

4. Results

4.1. Impacts of data preparation treatments on logistic regression performance

To answer the first research question, we conducted two three-way ($2 \times 3 \times 3$) analysis of variance (ANOVA) tests with Tukey post hoc tests on the selection set results, using AUC and TDL as dependent variables. Table 5 details the descriptive statistics (mean and standard deviation)

[4] Considering recent debates about the appropriateness of the AUC, we also compared the classifier performance in terms of the H-measure (e.g. [28]). However, the H measure results mimicked the tendencies we present in the main body of the paper, and are available upon request.

Table 4
Benchmark classification algorithms and meta-parameter settings.

Classifier		Models	Meta-parameter	Candidate settings[a]	Impl.[b]
Bagged CART [5]	Bag	9	No. of bootstrap samples	10, 20, …, 50, 100, 250, 500, 1000	M
Bayesian network [30]	B-Net	4	Approach to determine dependency network	K2, hill-climbing, tabu search, tree-augmented network	W
J4.8 decision tree [54]	DT	36	Confidence threshold for pruning	0.01, 0.15, …, 0.30	W
			Min. leaf size	n^*[0.01 0.025 0.05 0.1 0.25 0.5]	
Multilayer perceptron neural network [63]	NN	171	No. of hidden nodes	2, 3, …, 20	N
			Regularization penalty	$10^{(-4,\ -3.5,\ \ldots,\ 0)}$	
Naive Bayes [30]	NB	1	n.a.		W
Random forest [6]	RF	30	No. of CART trees	100, 250, 500, 750, 1000	M
			No. of randomly sampled variables	$\sqrt{m^*}[.1, .25, .5, 1, 2, 4]$	
Radial basis kernel support vector machine [15]	SVM	300	Regularization penalty	$2^{(-12,\ -13,\ \ldots,\ 12)}$	C
			Width of Rbf kernel	$2^{(-12,\ -13,\ \ldots,\ -1)}$	
Stochastic gradient boosting [23]	SGB	9	No. of boosting iterations	10, 20, …, 50, 100, 250, 500, 1000	W

[a] The variables n and m denote the number of observations and independent variables in a data set, respectively.
[b] Impl. = implementation of the classifier. Symbols represent the following sources: C = Chang & Lin [9], M = MATLAB core system, N = Nabney [48], W = Hall et al. [26].

of both performance measures for all three-way interaction combinations. The best three-way interaction is underlined and in italics; interactions that are not significantly different from the best combination are in italics. A detailed overview of the ANOVA results (Appendix A) and summary of the post hoc test results (Appendix B) are available in the Appendix.

For both AUC and TDL, the models are significant ($F_{Model\text{-}AUC}(18,1782) = 836.030$, $p < 0.01$; $F_{Model\text{-}TDL}(18,1782) = 23.024.9$, $p < 0.01$); the DPT technique influences customer churn prediction performance. We find significant variations in prediction performance across different DPT techniques, ranging from 0.570 to 0.667 in the AUC and from 1.570 to 2.381 in the TDL. Furthermore, the prediction results suggest the following optimal DPT scheme: decision-tree–based remapping for categorical variables, equal frequency binning for continuous variables, and WOE conversion as the representation method. Hereafter, we refer to this optimal DPT combination as the *dt_org_cat-efd-woe* treatment and the optimized logistic regression as LOGIT-DPT.

4.2. Competitiveness of logistic regression modeling using optimal data preparation treatment

Regarding the question of whether the LOGIT-DPT is competitive with benchmark classifiers that use standard data preparation, we report the performance measures on the validation set in Table 6. It also contains the statistical test results that benchmark our LOGIT-DPT against the advanced benchmark algorithms. To compare AUCs, we apply a commonly used test by Delong, Delong & Clarke-Pearson [17] that is designed to detect whether rank-based evaluation measures of two alternative classifiers differ significantly. However, this test is inapplicable for comparing classifiers according to the TDL, so we must use a different statistical test for these comparisons. We use the Breslow-Day test that relies on a cross-tabulation of predicted and actual responses of the two classifiers. To create cross-tabulations for the TDL comparison, we label all customers that one classifier predicts to be member of the top decile with a value of 1, and we apply a value of 0 to all other customers. Thus we obtain a discrete categorization of all customers, which is the input format required for the Breslow-Day test.

According to Table 6, LOGIT-DPT performs significantly better than more advanced single algorithms, such as neural networks, or support vector machines. Moreover, LOGIT-DPT exhibits very good performance compared with well-established ensemble methods. It performs significantly better than the bagging approache, and its performance is competitive with boosting or random forests. Specifically, the churn predictions of LOGIT-DPT display higher accuracy than those of random forests and boosting, in both the AUC and the TDL. However, the observed results do not provide sufficient evidence to reject the null hypothesis of equal performance.

Table 5
Descriptive statistics of AUC and TDL for all three-way interaction combinations.

TransCat	TransCont	Rep	Evaluation metric	
			AUC	TDL
nog	dt_org_cont	dum	0.637 (0.009)	2.035 (0.150)
	efd		0.632 (0.008)	2.002 (0.142)
	ewd		0.570 (0.001)	1.570 (0.001)
	dt_org_cont	inc	0.657 (0.007)	2.218 (0.143)
	efd		0.657 (0.008)	2.294 (0.150)
	ewd		0.610 (0.009)	1.748 (0.130)
	dt_org_cont	WOE	0.651 (0.008)	2.212 (0.155)
	efd		0.660 (0.006)	2.291 (0.165)
	ewd		0.626 (0.006)	1.979 (0.135)
dt_org_cat	dt_org_cont	dum	0.646 (0.008)	2.096 (0.159)
	efd		0.653 (0.006)	2.161 (0.148)
	ewd		0.617 (0.001)	2.008 (0.001)
	dt_org_cont	inc	0.657 (0.007)	2.295 (0.154)
	efd		0.662 (0.007)	*2.370 (0.149)*
	ewd		0.612 (0.008)	1.812 (0.117)
	dt_org_cont	WOE	0.657 (0.007)	2.249 (0.142)
	efd		*0.667 (0.008)*	*2.381 (0.182)*
	ewd		0.636 (0.006)	2.058 (0.125)

5. Conclusions

The general conclusions of our article suggest, first, that DPT strategies affect overall churn prediction performance, leading to improvements of up to 14.5% measured by AUC and 34% in TDL. Although few churn studies consider DPTs, our results align with the general idea that analysts should pay careful attention to data preparation [27]. Through its impact on the estimated churn probability, the choice of a DPT could improve the retention component for customer lifetime value [18].

Second, we offer evidence that a simplistic logistic regression model, built on a well-prepared data set, is just as viable as more advanced

Table 6
Statistical comparisons of LOGIT-DPT against benchmark algorithms.

Classifier	AUC	z-score	TDL	Chi^2 value (df = 1)
LOGIT-DPT	0.633		2.194	
Bag	0.610	1.682*	1.745	2.935*
B-Net	0.622	1.081	1.981	0.633
DT	0.500	9.504***	0.825	33.197***
NN	0.560	4.340***	1.274	13.536***
NB	0.598	2.301**	1.958	0.785
RF	0.631	0.222	2.123	0.069
SVM	0.535	7.449***	0.896	29.329***
SGB	0.610	1.847*	2.170	0.008

* $p < 0.10$.
** $p < 0.05$.
*** $p < 0.01$.

Table 7
Data set characteristics: dataset name, number of observations, percent incidence, number of categorical variables, number of continuous variables, number of independent variables, and number of independent variables after variable selection.

Data set	# observations	% incidence	# categorical variables	# continuous variables	# independent variables	# independent variables after variable selection
Credit Scoring	25,243	6.08%	56	86	142	19–47
Response Modeling	9822	5.97%	62	23	85	4–39

single and ensemble algorithms. This finding provides some new perspective on the ongoing searches for the best classification performance that hunt for new, more advanced algorithms, such as random forests [6], or support vector machines [60]. Applying a logistic regression model to correctly prepared data is generally less cumbersome than applying these advanced single algorithms, such as neural networks [7,47], or support vector machines. They require additional parameter tuning and often fail to provide direct churn probabilities, so they demand additional, post-processing steps such as calibration [51]. Furthermore, ensemble algorithms greatly increase model development time, because analysts must decide and optimize the base classifier to use in the ensemble, the number of bootstrap samples in the ensemble, and the aggregation method for the base classifiers' outputs [39]. Computing an ensemble prediction also requires individual predictions from all base models, so the time required to generate predictions from an ensemble model exceeds that of a logit model by several magnitudes. This difference can constitute a crucial advantage of the latter, especially if an application requires churn predictions to be computed in real time. Finally, from a user perspective, the choice of a particular prediction algorithm depends not only on prediction performance but also on critical criteria such as comprehensibility and justifiability [45]. Intuitively, more complex algorithms lack convincing insights about the comprehensibility and justifiability dimensions. The added complexity makes the churn prediction process more costly, compared with relying on a classic logistic regression model.

Although this study thus adds value to extant literature, it is not without some limitations that in turn can serve as bases for interesting future research paths. First, we use a case study approach, with one churn data set, to explore the impact of multiple DPTs on prediction performance and thereby shed new light on an underinvestigated topic. However, further research is needed to enhance the generalizability of our findings by validating them with data sets from other application domains and business contexts. Some initial experiments with the two alternative data sets detailed in Table 7—related to credit scoring and response modeling—suggest the validity of our findings beyond churn prediction.

For these follow-up experiments, we used the research procedures from Section 3 and applied the same data partitioning, scoring model, variable selection, experimental settings, and evaluation metrics to answer our two research questions. In Table 8 we reveal the impact of DPT choice on the selection set prediction performance for the credit scoring and response modeling data sets. Again, the best DPT technique is underlined and in italics, while the DPTs that do not differ significantly from this best combination are in italics only. The best DPT technique per evaluation metric serves to build the LOGIT-DPT. The results in Table 8 confirm the significant impact of the DPT strategy on predictive performance for both data sets, in line with our previous conclusion on research question 1: Significant improvements in predictive performance result from correct decisions during the DPT phase.

With regard to the second research question, in Table 9 we also compare the performance of the LOGIT-DPT on the validation set against the set of benchmark algorithms for both application settings. These results confirm that a simplistic logistic regression model built on well-prepared data is competitive with more advanced single and ensemble algorithms. We thus have an initial indication of the validity and generalizability of the conclusions.

Second, we focus on the most popular value transformation and representation methods for inclusion in our research study. Further research could include other, less common DPT strategies and investigate their impacts on churn prediction.

Third, we evaluated our prediction models according to the AUC and TDL, which both have their merits, as we noted previously. But they also lack a *direct* profit evaluation, as might be generated by prediction models. Although recent work by Verbeke et al. [61] provides a new evaluation metric, namely, the maximum profit (MP) criterion, we would have to gather additional information to calculate customer lifetime value, as would be required to implement the MP criterion in this research study. Therefore, further research is needed to verify the differences in evaluation metric uses across DPT strategies.

Table 8
Descriptive statistics of AUC and TDL for the credit scoring and response modeling data sets.

TransCat	TransCont	Rep	Credit scoring		Response modeling	
			Evaluation Metric		Evaluation Metric	
			AUC	TDL	AUC	TDL
nog	dt_org_cont	dum	0.835 (0.003)	4.062 (0.106)	0.766 (0.007)	3.194 (0.216)
	efd		0.831 (0.003)	4.097 (0.108)	0.763 (0.009)	3.264 (0.226)
	ewd		0.832 (0.003)	4.054 (0.120)	0.760 (0.006)	3.177 (0.165)
	dt_org_cont	inc	0.841 (0.002)	*4.228 (0.087)*	0.756 (0.002)	3.408 (0.052)
	efd		0.840 (0.002)	4.150 (0.086)	0.756 (0.001)	3.409 (0.056)
	ewd		0.839 (0.002)	4.085 (0.081)	0.763 (0.003)	3.502 (0.126)
	dt_org_cont	WOE	0.841 (0.002)	*4.257 (0.100)*	0.756 (0.002)	3.386 (0.047)
	efd		*0.842 (0.002)*	*4.236 (0.080)*	0.757 (0.001)	3.384 (0.051)
	ewd		0.833 (0.002)	4.049 (0.140)	0.764 (0.003)	3.473 (0.138)
dt_org_cat	dt_org_cont	Dum	0.835 (0.003)	4.045 (0.126)	0.765 (0.008)	3.218 (0.210)
	efd		0.831 (0.003)	4.086 (0.115)	0.764 (0.007)	3.282 (0.229)
	ewd		0.831 (0.003)	4.039 (0.111)	0.760 (0.006)	3.162 (0.131)
	dt_org_cont	inc	0.840 (0.002)	4.184 (0.099)	*0.781 (0.004)*	*3.671 (0.118)*
	efd		0.836 (0.003)	4.052 (0.095)	0.777 (0.007)	*3.710 (0.158)*
	ewd		0.835 (0.003)	4.067 (0.081)	0.778 (0.004)	3.633 (0.178)
	dt_org_cont	WOE	*0.842 (0.001)*	*4.235 (0.077)*	*0.779 (0.006)*	*3.651 (0.155)*
	efd		0.837 (0.002)	4.124 (0.091)	0.778 (0.006)	*3.689 (0.158)*
	ewd		0.832 (0.002)	4.064 (0.096)	0.775 (0.003)	3.605 (0.169)

Table 9
Statistical comparisons of LOGIT-DPT against benchmark algorithms for the credit scoring and response modeling data sets.

	Credit scoring				Response modeling			
Classifier	AUC	z-score	TDL	Chi^2 value (df = 1)	AUC	z-score	TDL	Chi^2 value (df = 1)
LOGIT-DPT	0.829		4.030		0.715		3.039	
Bag	0.789	5.467***	3.505	5.125**	0.716	0.091	3.039	1.050
B-Net	0.813	3.852***	3.743	3.604*	0.728	1.275	3.101	0.190
DT	0.744	9.021***	2.814	21.108***	0.507	7.021***	1.303	29.948***
MLP	0.826	0.952	4.196	0.032	0.727	0.875	2.977	0.021
NN	0.813	3.119***	3.886	3.267*	0.710	0.260	2.853	3.160*
RF	0.824	1.469	3.982	0.519	0.742	2.599***	3.163	0.084
SVM	0.716	10.219***	2.718	11.364***	0.670	2.430***	2.109	1.752
SGB	0.815	3.013***	4.053	1.831	0.728	0.901	2.853	0.768

* $p < 0.10$.
** $p < 0.05$.
*** $p < 0.01$.

Appendix A. ANOVA results for AUC and TDL

Performance metric	df	Sum of squares	F-statistic	Probability > F
AUC				
TransCat	1	0.064	1301.44	<0.01
TransCont	2	0.696	7104.96	<0.01
Rep	2	0.176	1802.40	<0.01
TransCat * Rep	2	0.046	469.57	<0.01
TransCont * Rep	4	0.041	207.68	<0.01
TransCat * TransCont	2	0.016	164.10	<0.01
TransCat * TransCont * Rep	4	0.022	114.64	<0.01
Model	18	736.81	836,030	<0.01
Error	1782	0.09		
TDL				
TransCat	1	6.486	335.44	<0.01
TransCont	2	51.582	1333.90	<0.01
Rep	2	14.577	376.96	<0.01
TransCat * Rep	2	2.226	57.57	<0.01
TransCont * Rep	4	5.559	71.88	<0.01
TransCat * TransCont	2	1.414	36.57	<0.01
TransCat * TransCon * Rep	4	2.492	32.23	<0.01
Model	18	8013.320	23,024.90	<0.01
Error	1782	34.450		

Appendix B. Post hoc test results

			AUC				TDL			
	i	*j*	M_i	M_j	t(1782)	*p*	M_i	M_j	t(1782)	*p*
TransCat	nog	dt_org_cat	0.633	0.645	36.08	<0.01	2.039	2.159	18.31	<0.01
TransCont	dt_org_cont	efd	0.651	0.655	10.54	<0.01	2.184	2.250	8.16	<0.01
		ewd		0.612	97.56	<0.01		1.862	40.09	<0.01
	Efd	ewd	0.655	0.612	108.10	<0.01	2.25	1.862	48.25	<0.01
Rep	Dum	inc	0.626	0.643	41.10	<0.01	1.979	2.122	17.98	<0.01
		woe		0.650	58.46	<0.01		2.195	26.96	<0.01
	inc	woe	0.643	0.650	17.36	<0.01	2.122	2.195	8.99	<0.01

References

[1] W.H. Au, C.C. Chan, X. Yao, A novel evolutionary data mining algorithm with applications to churn prediction, IEEE Trans. Evol. Comput. 7 (6) (2003) 532–544.
[2] R.C. Blattberg, B.-D. Kim, S.A. Neslin, Database marketing: Analyzing and Managing Customers, Springer, New York, 2010.
[3] M. Boullé, Khiops: a statistical discretization method of continuous attributes, Mach. Learn. 55 (1) (2004) 53–69.
[4] M. Boullé, MODL: a Bayes optimal discretization method for continuous attributes, Mach. Learn. 65 (1) (2006) 131–165.
[5] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[6] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[7] W. Buckinx, D. Van Den Poel, Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting, Eur. J. Oper. Res. 164 (1) (2005) 252–268.
[8] J. Burez, D. Van den Poel, CRM at a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services, Expert Syst. Appl. 32 (2) (2007) 277–288.
[9] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27:1–27:27.
[10] Z.-Y.Y. Chen, Z.-P.P. Fan, M. Sun, A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data, Eur. J. Oper. Res. 223 (2) (2012) 461–472.

[11] K.J. Cios, W. Pedrycz, R.W. Swiniarski, L. Kurgan, Data Mining: A Knowledge Discovery Approach, Springer, US, 2007.
[12] K. Coussement, P. Harrigan, D.F. Benoit, Improving direct mail targeting through customer response modeling, Expert Syst. Appl. 42 (22) (2015) 8403–8412.
[13] K. Coussement, D. Van den Poel, Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques, Expert Syst. Appl. 34 (1) (2008) 313–327.
[14] S.F. Crone, S. Lessmann, R. Stahlbock, The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing, Eur. J. Oper. Res. 173 (3) (2006) 781–800.
[15] D.P. Cui, D. Curry, Prediction in marketing using the support vector machine, Mark. Sci. 24 (4) (2005) 595–615.
[16] P. Datta, B. Masand, D.R. Mani, B. Li, Automated cellular modeling and prediction on a large scale, Artif. Intell. Rev. 14 (6) (2000) 485–502.
[17] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametetric approach, Biometrics 44 (3) (1988) 837–845.
[18] B. Donkers, P.C. Verhoef, M.G. de Jong, Modeling CLV: a test of competing models in the insurance industry, Quant. Mark. Econ. 5 (2) (2007) 163–190.
[19] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, 2001.
[20] A.E. Eiben, A.E. Koudijs, F. Slisser, Genetic modelling of customer retention, Lect. Notes Comput. Sci 1391 (1998) 178–186.
[21] A. Even, G. Shankaranarayanan, P.D. Berger, Evaluating a model for cost-effective data quality management in a real-world CRM setting, Decis. Support. Syst. 50 (1) (2010) 152–163.
[22] U.M. Fayyad, K.B. Irani, On the handling of continuous-valued attributes in decision tree generation, Mach. Learn. 8 (1) (1992) 87–102.
[23] J.H. Friedman, Stochastic gradient boosting, Computational Statistics and Data Analysis 38 (4) (2002) 367–378.
[24] J. Ganesh, M.J. Arnold, K.E. Reynolds, Understanding the customer base of service providers: an examination of the differences between switchers and stayers, J. Mark. 64 (3) (2000) 65–87.
[25] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA 2000, pp. 359–366 (inproceedings).
[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. Newsl. 11 (1) (2009) 10–18.
[27] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001.
[28] D.J. Hand, Measuring classifier performance: a coherent alternative to the area under the ROC curve, Mach. Learn. 77 (1) (2009) 103–123.
[29] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, New York, 2001.
[30] D. Heckerman, Bayesian networks for data mining, Data Min. Knowl. Disc. 1 (1) (1997) 79–119.
[31] D.W.J. Hosmer, S. Lemeshow, Applied Logistic Regression, second ed. John Wiley, New York, 2000.
[32] S.-Y. Hung, D.C. Yen, H.-Y. Wang, Applying data mining to telecom churn management, Expert Syst. Appl. 31 (3) (2006) 515–524.
[33] H. Hwang, T. Jung, E. Suh, An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry, Expert Syst. Appl. 26 (2) (2004) 181–188.
[34] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, Springer Publishing Company, Incorporated, 2014.
[35] D.E. Jennings, Outliers and residual distributions in logistic regression, J. Am. Stat. Assoc. 81 (396) (1986) 987–990.
[36] K. Kane, S.Y.V. Lo, J. Zheng, Mining for the truly responsive customers and prospects using true-lift modeling: comparison of new and existing methods, J. Mark. Anal. 2 (4) (2014) 218–238.
[37] G.V. Kass, An exploratory technique for investigating large quantities of categorical data, Appl. Stat. 29 (2) (1980) 119.
[38] D.A. Kumar, V. Ravi, Predicting credit card customer churn in banks using data mining, Int. J. Data Anal. Tech. Strateg. 1 (1) (2008) 4–28.
[39] L.I. Kuncheva, Combining pattern classifiers: methods and algorithms, Combining Pattern Classifiers: Methods and Algorithms, John Wiley & Sons Inc., 2004
[40] T.P. Kunz, S.F. Crone, J. Meissner, The effect of data preprocessing on a retail price optimization system, Decis. Support. Syst. 84 (2016) 16–27.
[41] B. Lariviere, D. Van den Poel, Predicting customer retention and profitability by using random forests and regression forests techniques, Expert Syst. Appl. 29 (2005) 472–484.
[42] A. Lemmens, C. Croux, Bagging and boosting classification trees to predict churn, J. Mark. Res. 43 (2) (2006) 276–286.
[43] E. Lima, C. Mues, B. Baesens, Monitoring and backtesting churn models, Expert Syst. Appl. 38 (1) (2011) 975–982.
[44] H. Liu, F. Hussain, C.L.I.M. Tan, M. Dash, Discretization an enabling technique, Data Min. Knowl. Disc. 6 (2002) 393–423.
[45] D. Martens, J. Vanthienen, W. Verbeke, B. Baesens, Performance of classification models from a user perspective, Decis. Support. Syst. 51 (4) (2011) 782–793.
[46] J. Moeyersoms, D. Martens, Including high-cardinality attributes in predictive models: a case study in churn prediction in the energy sector, Decis. Support. Syst. 72 (2015) 72–81.
[47] M.C. Mozer, R. Wolniewicz, D.B. Grimes, E. Johnson, H. Kaushansky, Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry, IEEE Trans. Neural Netw. 11 (3) (2000) 690–696.
[48] I. Nabney, NETLAB: Algorithms for Pattern Recognition, Springer, 2002.
[49] S.A. Neslin, S. Gupta, W. Kamakura, J. Lu, C.H. Mason, Defection detection: measuring and understanding the predictive accuracy of customer churn models, J. Mark. Res. XLIII (May) (2006) 204–211.
[50] A. Niculescu-Mizil, C. Perlich, G. Swirszcz, V. Sindhwani, Y. Liu, P. Melville, ... M. Singh, Winning the KDD cup orange challenge with ensemble selection, The 2009 Knowledge Discovery in Data Competition (KDD Cup 2009) Challenges in Machine Learning, vol. 3, 2009, p. 21.
[51] J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in Large Margin Classifiers 1999, pp. 61–74.
[52] D. Pregibon, Logistic Regression Diagnostics, Ann. Stat. 9 (4) (1981) 705–724.
[53] D. Pyle, S. Editor, D.D. Cerra, Data Preparation for Data Mining. Applied Artificial Intelligence, vol. 17, Morgan Kaufmann Publishers, San Francisco, CA, 1999.
[54] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.
[55] H. Risselada, P.C. Verhoef, T.H.A. Bijmolt, Staying power of churn prediction models, J. Interact. Mark. 24 (2010) 198–208.
[56] G. Shaffer, Z.J. Zhang, Competitive one-to-one promotions, Manag. Sci. 48 (9) (2002) 1143–1160.
[57] N. Siddiqi, Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring, Wiley, 2005.
[58] M. Sołtys, S. Jaroszewicz, P. Rzepakowski, Ensemble methods for uplift modeling, Data Min. Knowl. Disc. 29 (6) (2015) 1531–1559.
[59] L.C. Thomas, D.B. Edelman, J.N. Crook, Credit Scoring and Its Applications, Society for Industrial and Applied Mathematics, 2002.
[60] V. Vapnik, The Ns, Expert Syst. Appl. 23 (2) (1999) 103–112.
[61] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens, New insights into churn prediction in the telecommunication sector: A profit driven data mining approach, Eur. J. Oper. Res. 218 (1) (2012) 211–229.
[62] C.P. Wei, I.T. Chiu, Turning telecommunications call details to churn prediction: A data mining approach, Expert Systems with Applications 23 (2) (2002) 103–112.
[63] P.M. West, P.L. Brockett, L.L. Golden, A comparative analysis of neural networks and statistical methods for predicting consumer choice, Mark. Sci. 16 (4) (1997) 370–391.
[64] Y. Xie, X. Li, E.W.T. Ngai, W. Ying, Customer churn prediction using improved balanced random forests, Expert Syst. Appl. 36 (3 PART 1) (2009) 5445–5449.
[65] Y. Yang, G.I. Webb, Discretization for naive-Bayes learning: managing discretization bias and variance, Mach. Learn. 74 (1) (2009) 39–74.

Dr. Kristof Coussement, Kristof is full professor of marketing analytics, director of the expertise center for marketing analytics and academic director of the M.Sc. in Big Data Analytics for Business at IÉSEG School of Management – Catholic University of Lille (Lille, France) (EQUIS, AACSB, AMBA). His main research interests are all aspects in database marketing and social media & online community intelligence using data- and text-mining techniques. He has published his work in international peer-reviewed journals from the marketing, information systems and operations research field, while he co-authored multiple quantitative marketing books.

Dr. Stefan Lessmann, Stefan received a M.Sc. and Ph.D. in Business Administration from the University of Hamburg (Germany) in 2001 and 2007, respectively. Since 2006, he has served as a visiting research fellow at the Centre for Risk Research at the University of Southampton (United Kingdom). In 2014, he joined the Humboldt-University of Berlin, where he holds a professorship in information systems. His research concentrates on managerial decision support, big data, and business analytics; he is especially interested in predictive modeling to solve planning problems in marketing, consumer finance, and operations management. He has published several papers in leading scholarly outlets and conducted consultancy projects in various domains.

Dr. Geert Verstraeten, Geert obtained his Ph.D. in business administration from Ghent University (Belgium) and is now managing partner at Python Predictions, a Belgian niche player with expertise in the domain of predictive analytics. He currently has over 10 years of "hands-on experience" in different industries such as retail, mail-order, telecom, banking, utilities, subscription services, postal services and fundraising. More specifically, his interests lie in delivering highly performing yet interpretable predictions of future (customer) events. Since 2012, he is program chair of Predictive Analytics World (PAW) London. PAW is a unique vendor-neutral business conference, bringing together experts, practitioners, managers in the domain.