European Journal of Operational Research 269 (2018) 760–772

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees

Arno De Caigny[a], Kristof Coussement[a], Koen W. De Bock[b,*]

[a] Department of Marketing, IESEG School of Management, (LEM, UMR CNRS 9221), Université Catholique de Lille, 3 Rue de la Digue, F-59000 Lille, France
[b] Audencia Business School, 8 Route de la Jonelière, F-44312 Nantes, France

ARTICLE INFO

Article history:
Received 18 September 2017
Accepted 5 February 2018
Available online 12 February 2018

Keywords:
OR in marketing
Hybrid algorithm
Customer churn prediction
Logit leaf model
Predictive analytics

ABSTRACT

Decision trees and logistic regression are two very popular algorithms in customer churn prediction with strong predictive performance and good comprehensibility. Despite these strengths, decision trees tend to have problems to handle linear relations between variables and logistic regression has difficulties with interaction effects between variables. Therefore a new hybrid algorithm, the logit leaf model (LLM), is proposed to better classify data. The idea behind the LLM is that different models constructed on segments of the data rather than on the entire dataset lead to better predictive performance while maintaining the comprehensibility from the models constructed in the leaves. The LLM consists of two stages: a segmentation phase and a prediction phase. In the first stage customer segments are identified using decision rules and in the second stage a model is created for every leaf of this tree. This new hybrid approach is benchmarked against decision trees, logistic regression, random forests and logistic model trees with regards to the predictive performance and comprehensibility. The area under the receiver operating characteristics curve (AUC) and top decile lift (TDL) are used to measure the predictive performance for which LLM scores significantly better than its building blocks logistic regression and decision trees and performs at least as well as more advanced ensemble methods random forests and logistic model trees. Comprehensibility is addressed by a case study for which we observe some key benefits using the LLM compared to using decision trees or logistic regression.

1. Introduction

In an era of increasingly saturated markets that have intensified competition between companies, customer defection poses a real problem (Colgate, Stewart, & Kinsella, 1996). Therefore it has become clear to companies and managers that the historical customer information, which can be used to create models, in the existing customer base is one of the most important assets to combat customer churn (Ganesh, Arnold, & Reynolds, 2000). The search and identification of customers who show a high inclination to abandon the company or customer churn prediction is of crucial importance (Ganesh et al., 2000; Keaveney, 1995; Shaffer & Zhang, 2002) as part of a customer-oriented retention strategy that aims to reduce customer churn (Blattberg, Kim, & Neslin, 2010). Concretely, in customer churn prediction a scoring model allows the estimation of a future churn probability for every customer based on the historical knowledge of the customer. In practice these scores can be used to select targeted customers for a retention campaign.

Customer churn has been tackled from two different angles in previous research. On the one hand, researchers focus on improving customer churn prediction models in which more complex models are being developed and proposed in order to boost the predictive performance (Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012). On the other hand, researchers want to understand what drives customer churn and defined important drivers of customer churn such as customer satisfaction (Gustafsson, Johnson, & Roos, 2005; Hansen, Samuelson, & Sallis, 2013; Johnson, Nader, & Fornell, 1996). They consider customer churn prediction as a managerial problem that is driven by the customer's individual choice. Therefore action ability of customer churn prediction models is a key concern in which researchers can help managers to better understand the drivers of customer churn in order to make better informed decisions in combatting customer churn (Gustafsson et al., 2005; Verhoef, 2003). Hereby many authors point out the managerial value for customer segmentation (Athanassopoulos, 2000; Chan, 2008; Hansen et al., 2013; Seret, Verbraken, Versailles, & Baesens, 2012). By taking into account the main concerns of these two research angles, customer churn prediction models

* Corresponding author.
E-mail addresses: a.decaigny@ieseg.fr (A. De Caigny), k.coussement@ieseg.fr (K. Coussement), kdebock@audencia.com (K.W. De Bock).

https://doi.org/10.1016/j.ejor.2018.02.009
0377-2217/

Table 1
Overview of literature in churn prediction modeling after 2011.

Authors	Title & Journal	Year	What?	Techniques	Dataset – #cust. – #feat. -public(1) / private(2)	Metrics – sampling – feat. selection – validation
De Bock K.W. & Van den Poel D.	Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models – *Expert Systems with Applications*	2012	Application and evaluation of GAMensPlus on six real life datasets and case study	Generalized additive models (GAM & GAM as ensemble), bagging, random forests, random subspace method, logistic regression	Bank, DIY, supermarket, telecom and mail order garments – 3827 to 43,305 cust. – 15 to 529 feat. – (2)	Accuracy, AUC & TDL – undersampling – bagging – 5×2 cross validation, non-parametric Friedman test followed by Holm's procedure
Verbeke W., Dejaeger K., Martens D., Hur J. & Baesens B.	New insights into churn prediction in the telecommunications sector: A profit driven mining approach – *European Journal of Operational Research*	2012	A new measure to select the optimal model and fraction of customers to include and benchmarking experiment evaluating various classification techniques in telecom sector	Decision trees, logistic model tree, bagging, boosting, random forests, nearest neighbors, neural networks, rule induction techniques, logistic regression, naive Bayes, Bayesian networks, SVM	Telecom – 2180 to 338,874 cust. – 15 to 727 feat. – (1) & (2)	Maximum Profit, AUC &TDL – oversampling – fisher score – holdout, non-parametric Friedman test followed by the post-hoc Nemenyi test
Ballings M. & Van den Poel, D.	Customer event history for churn prediction: How long is long enough? – *Expert Systems with Applications*	2012	Time window optimization with respect to predictive performance in a newspaper company	Logistic regression, classification trees, bagging	Newspaper – 129,892 cust. – 1733 feat.- (2)	AUC – no sampling – Stepwise (logistic regression) – holdout, DeLong test
Chen Z.-Y., Fan Z.-P., Sun M.	A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data – *European Journal of Operational Research*	2012	This study presents a framework for customer churn prediction directly using longitudinal data	SVM based techniques, neural networks, decision tree, random forests, boosting, logistic regression, proportional hazard model	Food, Adventure and telecom – 633 to 8842 cust. – 20 to 36 feat. – (1)	PCC, sensitivity, specificity, Maximum profit, H, AUC, TDL – undersampling – adaptive feature selection – holdout
Coussement K. & De Bock K.W.	Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning – *Journal of Business Research*	2013	A comparison between single algorithms and their ensemble counterparts in the online gambling industry	Decision trees, random forests, Generalized additive models (GAM & GAM as ensemble)	Online gambling operator – 3729 cust. – 60 feat. – (1)	TDL, lift index (LI) – Bootstrap – no variable selection – 5×2 cross validation, nonparametric Wilcoxon-signed rank
Tang L., Thomas L., Fletcher M., Pan J. & Marshall A.	Assessing the impact of derived behavior information on customer attrition in the financial service industry – *European Journal of Operational Research*	2014	Use of derived behavior information to improve customer attrition models in the financial service industry	Probit-hazard model	Bank – 19,774 cust. – 22 feat. – (2)	AUC – no sampling – no feature selection – holdout, *t*-test
Moeyersoms J. & Martens D.	Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector – *Decision Support Systems*	2015	The use of high-cardinality attributes to predict churn in energy industry	Decision tree, logistic regression, SVM	Energy – > 1,000,000 cust. – 10 feat. – (2)	True positive rate, precision, TDL, AUC – 10 fold cross validation with holdout, Wilcoxon signed rank
Coussement K., Lessman S. & Verstraeten G.	A comparative analysis of data preparation algorithms for customer churn prediction: A case study in telecommunication – *Decision Support Systems*	2017	Study of the effect of data preparation techniques on predictive performance in telecommunication	Logistic regression, bagging, Bayesian network, Naive Bayes, decision tree, neural network, random forests, SVM, SGB	Telecom – 30,104 cust. – 956 feat.– (2)	AUC, TDL – undersampling – Hall correlation based feature selection – holdout, non-parametric test of De Long

should have good predictive performance and lead to actionable insights.

In customer churn prediction decision trees (DT) and logistic regression (LR) are very popular techniques to estimate a churn probability because they combine good predictive performance with good comprehensibility (Verbeke et al., 2012). While both techniques are useful and have their strengths, they have their flaws as well. DT handle interaction effects between variables very well but have difficulties to handle linear relations between variables. For LR the opposite is true: it handles linear relations between variables very well but it does not detect and accommodate interaction effects between variables. In this paper, the logit leaf model (LLM) is proposed as a new hybrid classification algorithm that uses a combination of decision trees and logistic regression and that is developed to reduce the weaknesses of DT and LR while maintaining their strengths. Conceptually the decision tree in the LLM splits the data into more homogenous subsets on which a logistic regression is fit for every subset. The added value of this method lies in the fact that it may improve the predictive performance of the LR and DT and that it offers a more actionable model for which segments are created and main drivers are detected on segment level.

Fourteen customer churn datasets from different industries are used on which the LLM is benchmarked against four conceptually related and well-known algorithms in customer churn prediction: DT, LR, random forests (RF) and logistic model trees (LMT) (Neslin, Gupta, Kamakura, Lu, & Mason, 2006; Verbeke et al., 2012). The predictive performance and comprehensibility are used as performance criteria.

The purpose of this study is twofold. Firstly, LLM is proposed as a new hybrid classification algorithm that enhances LR and DT. It helps analysts who are facing data with heterogeneity between customers and it is benchmarked against popular algorithms in customer churn prediction. Secondly, a visualization for LLM is proposed and how this helps to better understand a churn prediction model is discussed by means of a case study.

This paper is structured as follows. In the next section previous research in customer churn prediction and the trade-off between accuracy and comprehensibility is briefly discussed. Section 3 presents the LLM and clarifies how it is linked with the benchmark algorithms. The 4th section handles the experimental design. The results on predictive performance and a case study where the output of the LLM is compared with the output of LMT

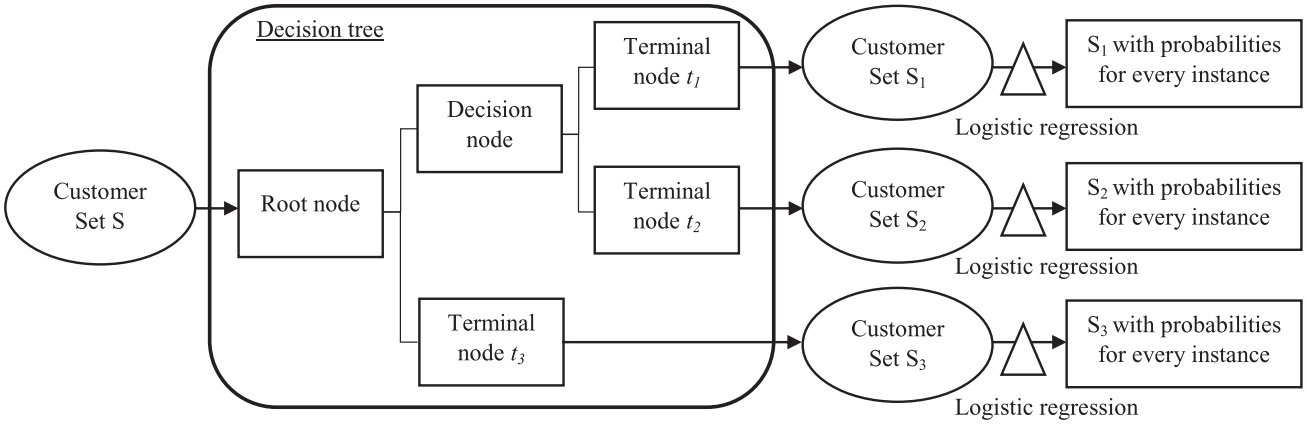

Fig. 1. Conceptual presentation of the logit leaf model.

Table 2
Overview trade-off between predictive performance and comprehensibility for considered algorithms.

Classifier\ criteria	Pred. perf.	Comprehensibility
Logistic regression (LR)	+	+ +
Random forests (RF)	+ +	–
Logistic model trees (LMT)	+ +	–
Decision tree (DT)	–	+ +
Logistic Leaf model (LLM)	+ +	+

(Landwehr, Hall, & Eibe, 2005), LR and DT are discussed in Section 5. The conclusions are presented in Sections 6 and 7 states limitations and areas for further research.

2. Predictive modeling in customer churn

The importance of customer churn prediction as a research discipline is discussed in this section. It introduces the predictive performance and comprehensibility as two main requirements for customer churn prediction models and explains why there is often a trade off between these requirements.

Customer churn predicition is an important research discipline as part of customer relationship management because it is a lot more profitable to retain and satisfy existing customers than to attract new customers for several reasons: (1) Successful companies have long term relationships with their customers which allow them to focus on their customer needs instead of looking for new and potentially not very profitable customers who are typically characterized by a higher attrition rate (Dawes & Swailes, 1999; Reinartz & Kumar, 2003); (2) Customers who leave the company can influence other customers within their social network to do the same (Nitzan & Libai, 2011); (3) long-term customers have beneficial effects on both profit and cost side. On the profit side, long term customers tend to buy more and they can refer people to the company by positive word of mouth (Ganesh et al., 2000). On the cost side, they are less costly to serve because a company possesses already more information about them and knows their customer needs which decrease the service cost (Ganesh et al., 2000). (4) Competitive marketing actions have less effect on long term customers (Colgate et al., 1996); (5) Losing customers increases the need and the cost to attract new customers and decreases the profits by the missed sales and lost opportunities for cross and up sale. These effects cause that that the cost to retain an existing customer is several times smaller than the cost of acquiring a new customer (Torkzadeh, Chang, & Hansen, 2006). Consequently customer churn prediction is necessary in a retention strategy.

Previous customer churn prediction studies investigated an extensive range of algorithms in different industries such as telecom (Verbeke et al., 2012) and financial services (Kumar & Ravi, 2008; Tang, Thomas, Fletcher, Pan, & Marshall, 2014). An overview of the literature until 2011 in customer churn prediction modeling is found in Verbeke, Martens, Mues, and Baesens (2011). Table 1 provides an overview of the most recent customer churn prediction literature after 2011.

Customer churn prediction models are often solely evaluated on their predictive performance, or their ability to discriminate between churners and non-churners. Over the last 20 years more advanced techniques have been proposed and evaluated. DT and LR are very popular techniques in customer churn. The LR became the standard in customer churn prediction because of good predictive performance and robust results (Coussement, Lessman, & Verstraeten, 2017; Neslin et al., 2006). Also RF has been used in various churn prediction studies where it has been proven to achieve excellent predictive performance (Coussement & Van den Poel, 2008; Kumar & Ravi, 2008). An extensive benchmark study can be found in Verbeke et al. (2012) that confirms the good predictive performance of LR and RF in customer churn prediction and show the extent of algorithms that have been used in the field. In the extensive set of algorithms that have been tested in customer churn prediction, LMT (Landwehr et al., 2005), which are discussed in more detail in Section 3.2, is the algorithm that shows the most conceptual similarities with LLM. LMT received so far only limited attention in customer churn prediction, however, it shows good predictive performance (Verbeke et al., 2012).

Other authors focused on the comprehensibility of customer churn prediction models (De Bock & Van den Poel, 2012; Martens, Vanthienen, Verbeke, & Baesens, 2011). Firstly, in customer retention management it is important to know why customers are leaving the company to better act on the drivers of churn. Therefore more comprehensible models are preferred over non-comprehensible alternatives. Secondly, if the model has to be justified towards the management a more comprehensible model helps to show that it is in line with current domain knowledge.

To measure the comprehensibility of a classification algorithm the size and the classification output type are important determinants (Martens et al., 2011). Classification output types define the format in which the results of the model are returned to the researcher. Commonly used classification output types are linear models, non-linear models, rule-based models and tree based models. The size depends on the output types; the size of linear and non-linear output types is measured by the number of terms while the size of rule-based and tree based output types is given respectively by the number of rules and the number of leaves

Model creation phase
Split D_{tot} in D_{tr} and D_{val}
Input: (training) data $D_{tr} = \{(Xi, Yi)\}_{i=1}^{N}$
1 : Calculate initial decision tree on D_{tr} spanning the total space S
2 : Define subspaces S_t based on a set of terminal nodes T for which $S = \bigcup_{t \in T} S_t, \forall\, t \neq t'$: $S_t \cap S_{t'} = \emptyset$
3: **For** $i = 1$ *to* T **do:**
4: Define starting logit
5: **Repeat**
6: add variable max decrease AIC
until stopping criteria
7: **End for;**
8: Combine results M_k in model M
Output: model M

Prediction phase
Input: (new) data $D_{val} = \{(Xi, Yi)\}_{i=1}^{N}$
1: Apply decision rules of model M on D_{val} spanning the total space S, resulting in subspace S_t based on a set of terminal nodes T for which $S = \bigcup_{t \in T} S_t$, $\forall\, t \neq t'$: $S_t \cap S_{t'} = \emptyset$
2: **For** $i = 1$ *to* T **do:**
3: Apply logistic regression specific for S_t
4: **For** $j = 1$ *to* n_i **do:**
5: Calculate predictions for all n_i instances in S_t
6: End For;
7: End For;
8: Combine predictions
Output: one prediction for every instance in S

Fig. 2. Pseudo-code logit leaf model.

(Martens et al., 2011). There is a general consensus that smaller output size is more comprehensible but the opinion about output types is rather subjective. For example some authors prefer rule based or tree based models, like DT, over models with linear output types, like LR (Martens et al., 2011), while others consider linear output types more interpretable than rule based output types (Coussement, Harrigan, & Benoit, 2015). Non-linear output types are generally considered the least interpretable of the above mentioned output types but additional analyses such as rule extraction techniques can be used to make these models with non-linear output types more comprehensible (Baesens, Setiono, Mues, & Vanthienen, 2003; Martens, Baesens, Van Gestel, & Vanthienen, 2007; Verbeke et al., 2011). Martens et al. (2011) suggested that to in order to obtain comprehensible classification models one should (1) try to build rule-based models, (2) combine different output types to find the optimal trade-of between simple techniques with good comprehensibility and more advanced techniques with extended flexibility and generalization behavior, and last but not least (3) visualization is key.

By consequence customer churn prediction is a complex process which requires informed decisions from an analyst or researcher at various stages (Lima, Mues, & Baesens, 2011). It is important that these decisions are data-driven whenever it is possible (Lessmann & Voß, 2009). The main decision point for an analyst in the modeling step is which of the classification algorithms one should use for which he or she faces a trade-off between comprehensibility and predictive performance. This explains why DT and LR are popular techniques in customer churn prediction as they combine good predictive performance with good comprehensibility (Hwang, Jung, & Suh, 2004; Kumar & Ravi, 2008; Mozer, Wolniewicz, Grimes, Johnson, & Kaushansky, 2000; Wei & Chiu, 2002). A summary of this trade-off between comprehensibility and predictive performance for the considered algorithms is provided in Table 2. Due to its hybrid nature and reduced simplicity, the LLM has been awarded a single + for comprehensibility. Nevertheless, it is noteworthy that the hybrid LLM algorithm has the ability to detect segments, which has been proven valuable in customer churn research due to the heterogeneity in the customers' base (e.g. Glady, Baesens, & Croux, 2009), and to detect relevant customer churn drivers specific to those segments. Therefore it provides valuable and actionable insights to the marketing department that can aid to set up segment-specific targeted retention initiatives based on the significant customer churn drivers. Consequently the LLM may be preferred over the LR or DT in the setting of customer churn prediction.

3. Logit leaf model

3.1. Description of the LLM algorithm

The LLM is a two-step hybrid approach that constructs a decision tree in the first step to identify homogenous customer segments, and in a second step applies logistic regressions to each of these segments.

Table 3
Data set characteristics: dataset name, industry, number of observations, number of attributes, churn percentage and source.

Data set	Industry	# observations	# attributes	% churn (%)	Source
Ds1	Financial services	117,808	237	3.55	European financial services provider
Ds2	Retail	32,371	47	25.15	European retailer
Ds3	DIY	3827	16	28.14	European DIY supplier
Ds4	Newspaper	427,833	165	11.14	European newspaper company
Ds5	Telecom	71,047	87	29.00	Duke[a]
Ds6	Financial services	102,279	138	5.99	European financial services provider
Ds7	Telecom	47,761	43	3.69	European telecom operator
Ds8	Telecom	50,000	303	7.34	European telecom operator
Ds9	Financial services	631,627	232	2.53	European financial services provider
Ds10	Financial services	573,895	232	2.57	European financial services provider
Ds11	Financial services	398,087	232	4.50	European financial services provider
Ds12	Financial services	316,578	232	6.45	European financial services provider
Ds13	Financial services	602,575	232	3.16	European financial services provider
Ds14	Energy	20,000	33	10.00	European energy company

[a] Center for Customer Relationship Management Duke University, February2014. http://www.fuqua.duke.edu/centers/ccrm.

Decision trees constitute a class of predictive models that score well on efficiency and comprehensibility, mainly because of their simplicity. They use a process where the data is recursively split into smaller and purer subsets by repeatedly applying a greedy search through the space of possible decision trees branches and choosing optimal splits based upon a splitting criterion. This process starts in the root node, which is a node without parent nodes and iteratively determines optimal splitting criteria that divide the data over two child nodes. This process terminates, when no further splits are desirable or possible, with a set of nodes without child nodes called terminal nodes or leaves. As such, the entire customer set S that is spanned by all attributes of the data can be split by the tree structure that consists of a set of leaves or terminal nodes T, in a disjoint subdivision of S into customer subsets S_t where every subset is represented by a leaf t in the tree:

$$S = \bigcup_{t \in T} S_t; \forall\, t \neq t' : S_t \cap S_{t'} = \emptyset$$

The repeated splitting gradually leads to more complex models and consequently, introduces a risk of overfitting. This undesirable complexity can be contained through the introduction of a number of hyperparameters that govern the splitting process on the one hand, and through the application of post-hoc pruning on the other hand. Pruning is a complexity-reduction technique that removes parts of the tree that do not provide enough power to classify customers. The complexity of the tree is also governed by several parameters of the decision tree algorithm such as the minimal leaf size parameters which controls the minimum number of observations in nodes and this constitutes a stopping criterion for the splitting process. Detailed information on the parameter settings in this study and how these parameters are optimized, is discussed in Section 4.4.

Conventional DT assign predictions to an observation based upon the class tendency of the terminal node in which that observation falls. In the LLM algorithm, in a second step, a logistic regression with forward variable selection is fitted at every terminal node.

LR is a frequently used standalone predictive technique in marketing (Bucklin & Gupta, 1992). Also in the domain of customer churn prediction it has been proven a valuable technique for two reasons: (1) in a LR posterior probabilities are estimated directly which make it a lot more comprehensible than other, more complex "black box" methods; (2) logit modeling has been shown to provide good and robust results in benchmarking studies for churn prediction (Neslin et al., 2006; Verbeke et al., 2012) and can compete with more advanced techniques (Coussement et al., 2017) in customer churn prediction. The logistic regressions in the LLM deploys forward selection which makes that the algorithm (1) has a built-in feature selection mechanism and (2) selects the most important variables for each group separately.

Fig. 1 shows a conceptual representation of the LLM displaying the flow of the data. In this representation, the entire customer set S has been divided into three subsets S_1, S_2 and S_3 by the decision tree. A logistic regression is fitted for every subset separately resulting in probabilities for every instance in the subsets. Fig. 2 shows the pseudo code of the LLM algorithm.

3.2. Related methods

The core mechanism in the LLM that consists of an assignment of different data points to different constituent classifiers is found in a number of related methods. First, *classifier selection* involves the creation of an ensemble of classifiers whereby one or more members are selected in function of the scoring task at hand, based upon an estimate of member competence for that task (Kuncheva, 2014). As such, the most competent member classifier, or set of member of classifiers, is selected for the data to be scored. Some methods depend on a preliminary assessment of member competence for predefined subspaces in the data such as the clustering and selection approach (Kuncheva, 2000, 2002). Other methods that dynamically assess member model competence in function of the data to be scored are referred to as *dynamic* classifier selection algorithms (Didaci, Giacinto, Roli, & Marcialis, 2005; Giacinto & Roli, 2000). An important difference of these methods to the proposed LLM is that classifier selection methods train all ensemble members on the full training data set first and then perform a competence-based selection during the scoring phase while LLM first identifies segments and subsequently trains segment-specific models (logistic regressions). Hence, LLM delivers truly segment-specific models that can be interpreted as such, which is not the case in classifier selection. Moreover, dynamic classifier selection methods are known to be computationally expensive in the scoring phase (Kuncheva, 2014, p. 239).

Second, a noteworthy subcategory of classifier selection that builds upon the principles of mixture models is found in *mixture of experts* (ME) models (Jacobs, Jordan, Nowlan, & Hinton, 1991; Kuncheva, 2014). In a ME architecture, originally defined for ensembles of artificial neural networks, both the member models and a meta-model called the gating network are simultaneously trained so that members learn to predict well for a subspace of the data and the gating networks learns to identify the most competent member to score a particular input. A final component, the selector, is chosen as a heuristic rule for combining the member's predictions. ME models differ from other ensemble selection methods

Table 4
Classification methods and meta-parameters settings considered in this study.

Classifier	No. models per algorithm[1]	Meta-parameter	Candidate settings[2,3]	Code[4]
Decision tree[6]	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]	
Logistic regression	1	N.a.	N.a.	R
Random forests	30	No. of CART trees	100, 200, 500, 750, 1000	R
		No. of randomly sampled variables[5]	$\sqrt{v}$*[0.1, 0.25, 0.5, 1, 2, 4]	
Logistic model tree	1	Internal cross validation	N.a.	W
Logit leaf model [6]	36	Confidence threshold for pruning	0.01, 0.15,..., 0.30	S
		Min. leaf size	n*[0.01,0.025,0.05,0.1,0.25,0.5]	

[1] A set of candidate values is defined for every parameter and we create models for all possible value combinations. For example DT offers two meta-parameters with both 6 candidate values, so we create 6*6 = 36 classification models.
[2] Candidate settings are based on (Coussement et al., 2017).
[3] The variable n denote the number of observations and v denote the number of independent variables in the data set.
[4] Symbols represent the following sources: W = RWeka (Hall et al., 2009), R = R core system, S =Publicly available implementation in the new LLM package (De Caigny et al., 2018) accessible via https://CRAN.R-project.org/package=LLM.
[5] The number of randomly sampled variables is always rounded to the next integer.
[6] C4.5 decision tree is used.

in the sense that they are trained and the data is partitioned simultaneously so that local models each learn a part of the data as well as possible. This is a fundamental difference to the LLM where the partitioning of the data and the subsequent training of logistic regression models occurs sequentially. Another difference to the LLM is that the gating network in the ME makes soft partitions of the input space while in the LLM only hard partitions of the input space are created by the decision tree (Yuksel, Wilson, & Gader, 2012). Moreover, as both the gating network as the model members are complex, the interpretation of neural network-based ME models is very challenging. It is however important to note that variations upon the ME framework have been applied to leverage linear base learners with the aim of obtaining flexible, nonlinear mixture models for classification and regression. Noteworthy contributions to ME literature include mixtures of support vector machines (Fu, Robles-Kelly, & Zhou, 2010; Zhu, Chen, & Xing, 2011a,b), generalized linear models and Gaussian processes (Rasmussen & Ghahramani, 2002).

Third, similar due to the functioning of the LLM, a number of related hybrid methods exist whereby in a first training phase a model-based partition of the data space is obtained and in a subsequent modeling phase one or more classifiers are constructed in function of this data partitioning. In Li, Weng, Shao, & Guo (2016), decision trees are used to create a data partitioning that is afterwards incorporated in logistic regression through the inclusion of a set of dummy variables. This approach outperformed conventional logistic regression. A more advanced technique is the homogenous ensemble method LMT that deploys decision trees to assign different data points to a different base classifiers (Landwehr et al., 2005). The proposed LLM is closely related to this homogenous ensemble method LMT that is using logit as a base learner and has a built in variable selection method (Landwehr et al., 2005). LMT grows a tree starting with a logistic model at the root using the logitBoost algorithm (Friedman, Hastie, & Tibshirani, 2000) where an internal cross validation mechanism determines an appropriate number of iterations. When a node splits, the logitBoost algorithm proceeds on the data of the child nodes, which are de facto subsets of the total data, refining the model of the parent node. Here it considers only univariate regression models, i.e. only one attribute is used per iteration, which are selected automatically. Similar to other tree-structured classifiers a pruning process takes care of unnecessary complexity to improve generalizability of the model. At the terminal nodes posterior probabilities estimates are calculated by using logistic regression (Landwehr et al., 2005).

While LLM is conceptually related with LMT there are some important differences. Firstly, LLM fits logistic regressions with forward selection only at the terminal nodes while LMT fits a logistic regression at every node that is warm-started with the results of the parent node. Secondly, logistic regression in LMT uses the logitBoost algorithm as an ensemble method with univariate logistic regressions as base learners, while LLM fits logistic regressions with forward selection of variables. Thirdly, in LLM the tree building that is used for segmentation and logistic regression fitting to optimize the predictions in the leaves happen in separate steps, while in LMT it occurs at the same time and the tree is pruned later.

4. Experimental set-up

4.1. Data and experimental design

Experiments are conducted on fourteen churn datasets originating from different sources. Table 3 provides an overview of the different datasets.

The experimental design of this study is chosen to assess the performance of the LLM compared to four benchmark algorithms. Because it is necessary for most algorithms to optimize parameter settings, a variation of the popular 5×2 cross-validation was implemented that splits the data in training, selection and validation sets each containing one third of the data resulting in a 5×3 cross-validation experimental design (Burez & Van den Poel, 2009; Dietterich, 1998).

4.2. Data preprocessing

First, missing value imputation is applied. Missing values are treated differently based on the percentage of missing values in an attribute in analogy with Verbeke et al. (2012). Imputation procedures are used for attributes with more than 5% of the values missing. Depending on the variable, zero imputation, median imputation or modus imputation is used (Coussement et al., 2017). Dummy variables are created flagging variables where missing variables are imputed. For attributes with less than 5% of the values missing, the instances containing the missing value are removed from the data in order to limit the impact of imputation procedures (Verbeke et al., 2012). Categorical variables are transformed into binary variables using dummy encoding. This technique creates v–1 dummy variables, where v equals the number of distinct values of the categorical variable (Pyle, 1999). These newly created variables indicate the presence or absence of a particular characteristic.

Second, outlier detection and treatment is applied. Outliers are unusual values that are typically defined as being more than three standard deviations away from a variable's mean value (Anderson, Sweeney, Williams, Freeman, & Shoesmith, 2010). Winsorization

Table 5
Example of a confusion matrix for binary classification.

	Actual 1	0	
Predicted	1 True positive (TP)	False positive (FP)	Predicted positives (PP)
	0 False negative (FN)	True negative (TN)	Predicted negatives (PN)
	Actual positives (AP)	Actual negatives (AN)	

is applied to transform outliers into "acceptable" values that are within three standard deviations.

A last preprocessing step involves undersampling. Typically, the class variable in a churn prediction setting is heavily skewed, i.e. the number of churners is often much lower than the number of non-churners. As shown in Table 3, the datasets in this study suffer from skewed class distributions as well with churn class incidences ranging from 2.53 to 29.00%. To remedy this, undersampling is applied to the training data: the size of the majority class, in this case non-churning customers, is reduced to the size of the minority class, churners, through random sampling of non-churning customers (Burez & Van den Poel, 2009; Ling & Li, 1998).

4.3. Variable selection

Input selection procedures have several advantages. Firstly, a classifier yield better predictive performance when trained on a small set of well-chosen and highly predictive variables than on a model that has been trained on an extensive set containing much redundant or noisy data (Verbeke et al., 2012). The input selection procedure prevents the model to overfit on noisy data and increases the prediction model's stability through reducing the collinearity and thus improve predictive performance (James, Witten, Hastie, & Tibshirani, 2014). Secondly, variable selection results in more concise models since the number of variables are limited which improve the interpretability of the models (Lessmann & Voß, 2009). Thirdly, models with less variables are in general more efficient to industrialize (Maldonado, Flores, Verbraken, Baesens, & Weber, 2015). Lastly, a good feature selection technique can reduce the feature acquisition costs in the future (Maldonado, Pérez, & Bravo, 2017).

In a first step to reduce the number of variables to twenty the simple but effective fisher score is used (Verbeke et al., 2012). The fisher score is defined as follows in a churn prediction context:

$$Fisher\ score = \frac{\left|\bar{X}_c - \bar{X}_{nc}\right|}{\sqrt{S_c^2 + S_{nc}^2}},$$

with $\bar{X}_c$ and $\bar{X}_{nc}$ the mean value, and S_c^2 and S_{nc}^2 the variance of an independent variable for respectively churners and non-churners. Note that a secondary variable selection is inherent to some of the algorithms as explained in Section 3.

4.4. Parameter settings

All of our benchmark algorithms have already been used in previous studies (Coussement et al., 2017; Verbeke et al., 2012) and parameter settings are chosen exactly the same as in these previous studies. An overview of all parameter settings is presented in Table 4. The decision tree parameter ranges during optimization are identical for the standalone DT and the decision tree in the LLM, as presented in Table 4. However, the selection of the optimal parameters for both approaches depends on the selection set performance and therefore the values chosen can be different for both algorithms. For the DT, parameters are chosen to deliver optimal performance of that tree while in the LLM the parameters for the decision tree are selected based on the optimal performance of the entire LLM (thus after the logistic regressions are fit on the leaves of the decision tree). Therefore it is possible and even likely that there is a difference in the selected parameters for the standalone DT and the decision tree in the LLM on the same data. The number of segments for the LLM is not set by the researcher but depends on the two decision tree hyperparameters in this experiment. The pruning strategy for decision trees in the LLM and standalone DT is both entropy based.

4.5. Evaluation criteria

The predictive performance of the different classifiers is assessed by the area under the receiver operating characteristics curve (AUC) and top decile lift (TDL) (Coussement et al., 2017; Lemmens & Croux, 2006; Verbeke et al., 2012) . The AUC and TDL can be derived from the confusion matrix. Table 5 presents a confusion matrix for binary classification.

The AUC is used to evaluate the predictive performance of a binary classification system such as customer churn prediction with a simple one-figure score (Hanley & McNeil, 1982) and can be computed using following formula:

$$AUC = \int_0^1 \frac{TP}{AP} d\frac{FP}{AN}.$$

Its usage is justified as (1) it is an intuitive ranking based measure of posterior churn probabilities that can be easily understood. An intuitive interpretation of the AUC in a churn prediction context is that it provides an estimate of the probability that a randomly chosen churner is correctly rated or ranked higher than a

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

	Decision tree (DT)	Logistic model tree (LMT)	Logistic regression (LR)	Random forests (RF)	Logit Leaf model (LLM)
Ds1	0.734 (0.013)	0.747 (0.012)	0.691 (0.004)	0.782 (0.008)	0.739 (0.011)
Ds2	0.792 (0.008)	0.815 (0.004)	0.816 (0.004)	0.816 (0.003)	0.816 (0.004)
Ds3	0.712 (0.018)	0.753 (0.011)	0.753 (0.012)	0.748 (0.010)	0.754 (0.012)
Ds4	0.816 (0.005)	0.845 (0.001)	0.840 (0.001)	0.839 (0.001)	0.846 (0.001)
Ds5	0.620 (0.006)	0.624 (0.005)	0.595 (0.003)	0.639 (0.004)	0.626 (0.004)
Ds6	0.804 (0.007)	0.825 (0.005)	0.815 (0.003)	0.848 (0.004)	0.827 (0.004)
Ds7	0.678 (0.018)	0.729 (0.010)	0.731 (0.009)	0.735 (0.012)	0.726 (0.011)
Ds8	0.601 (0.008)	0.618 (0.006)	0.618 (0.005)	0.614 (0.006)	0.620 (0.005)
Ds9	0.861 (0.007)	0.885 (0.002)	0.879 (0.002)	0.877 (0.002)	0.887 (0.002)
Ds10	0.634 (0.015)	0.651 (0.010)	0.651 (0.009)	0.649 (0.009)	0.652 (0.009)
Ds11	0.768 (0.006)	0.792 (0.003)	0.783 (0.002)	0.796 (0.002)	0.794 (0.002)
Ds12	0.766 (0.015)	0.793 (0.003)	0.781 (0.003)	0.793 (0.003)	0.794 (0.003)
Ds13	0.843 (0.007)	0.867 (0.003)	0.861 (0.002)	0.848 (0.003)	0.867 (0.002)
Ds14	0.835 (0.005)	0.864 (0.002)	0.861 (0.002)	0.868 (0.002)	0.867 (0.002)

Table 7
Results of the benchmarking experiment using the TDL (10%) performance criterion.

	Decision tree (DT)	Logistic model tree (LMT)	Logistic regression (LR)	Random forests (RF)	Logit Leaf model (LLM)
Ds1	1.430 (1.125)	2.961 (0.411)	2.327 (0.097)	4.004 (0.235)	2.438 (0.668)
Ds2	1.109 (0.529)	2.777 (0.056)	2.762 (0.044)	2.740 (0.064)	2.776 (0.053)
Ds3	0.762 (0.373)	2.091 (0.144)	2.061 (0.137)	2.115 (0.115)	2.137 (0.153)
Ds4	0.741 (0.295)	4.496 (0.028)	4.415 (0.031)	4.478 (0.034)	4.498 (0.027)
Ds5	1.084 (0.288)	1.555 (0.044)	1.434 (0.020)	1.657 (0.045)	1.594 (0.042)
Ds6	3.901 (0.599)	4.460 (0.218)	4.123 (0.098)	4.967 (0.110)	4.506 (0.181)
Ds7	1.712 (0.614)	3.422 (0.163)	3.450 (0.176)	3.363 (0.200)	3.339 (0.282)
Ds8	0.768 (0.616)	1.805 (0.127)	1.819 (0.104)	1.679 (0.154)	1.745 (0.181)
Ds9	2.619 (1.746)	5.444 (0.212)	5.165 (0.066)	5.168 (0.000)	5.474 (0.256)
Ds10	0.615 (0.616)	2.093 (0.107)	2.083 (0.099)	2.008 (0.127)	2.096 (0.090)
Ds11	2.482 (0.906)	3.769 (0.110)	3.462 (0.045)	4.000 (0.064)	3.815 (0.066)
Ds12	2.391 (1.030)	3.721 (0.069)	3.559 (0.059)	3.651 (0.052)	3.781 (0.063)
Ds13	4.929 (0.174)	5.246 (0.150)	5.101 (0.050)	3.207 (1.662)	5.342 (0.217)
Ds14	2.625 (0.089)	5.294 (0.083)	5.143 (0.146)	5.189 (0.064)	5.358 (0.073)

Table 8
Average classifier ranks[1] across data sets for different performance measures.

Control:	Logit leaf model (LLM)	AUC 1.786	TDL 1.786
Benchmarks:	Decision tree (DT)	4.857* (0.000)	4.929* (0.000)
	Logistic model tree (LMT)	2.571 (0.377)	2.214 (0.473)
	Logistic regression (LR)	3.286* (0.036)	3.357* (0.026)
	Random forests (RF)	2.500 (0.377)	2.714 (0.241)

Between brackets is the adjusted p-value for Holm post-hoc test.
[1] Lower ranks are better.
* Indicates significance on 95% level.

Table 9
Overview of selected variables in cell2cell dataset.

Variable	Definition
Callwait	Mean number of call waiting calls
changem	% change in minutes of use
changem dummy	Dummy if %change in minutes of use is imputed
creditde	Low credit rating – de
custcare	Mean number of customer care calls
directas	Mean number of director assisted calls
eqpdays	Number of days of the current equipment
incalls	Mean number of inbound voice calls
models	# models issued
mou	Mean monthly minutes of use
opeakvce	Mean number of in and out off-peak voice call
outcalls	Mean number of outbound voice calls
phones	# handsets issued
recchrge	Mean total recurring charge
retcalls	Number of calls previously made to the retention team
revenue	Mean monthly revenue
setprcm	Missing data on handset price
webcap	Handset is web capable

randomly selected non-churner; (2) AUC is a cut-off independent measure that accounts for the overall performance of a classification technique since it considers all the possible cut-off values.

Besides the AUC, TDL is also considered as a metric for the predictive performance of a classifier. Lift reveals for a specific cut off value how much better (or worse) a classifier predicts compared to random selection. Lift can be defined using the confusion matrix for a specific cut off value as:

$$Lift = \frac{TP/(TP+FP)}{AP/(AP+AN)},$$

Top decile lift compares the proportion of churners in the entire dataset with the proportion of churners in the top decile containing customers with the highest predicted score according to the churn probabilities given by the classification model. A TDL score of 1 indicates that the density of churners in the top decile is the same as in the entire dataset. Scores higher than 1 indicate a higher density of churners in the top decile and vice versa. TDL is very valuable from a managerial perspective because it focusses on customers that are most at risk of leaving the company and therefore indirectly impacts the profitability of a customer retention campaign(Neslin et al., 2006).

To compare the predictive performance of LLM with the benchmarks a testing framework (Demsar, 2006) is applied based on the non-parametric Friedman test (Friedman, 1940). The Friedman statistic is defined as:

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

With R_j the average rank of an algorithm $j = 1,2,\ldots,k$ over N datasets. As a rule of thumb N and k should be at least 10 and 5, respectively, which is fulfilled in this study so the assumption holds that the Friedman test is distributed according to χ_F^2 with k-1 degrees of freedom under the null hypothesis that states that the results of all algorithms do not differ and thus the ranks R_j should be equal. If the null-hypothesis is rejected, LLM is pairwise compared with the benchmark algorithms using the Holm post-hoc test (García, Fernández, Luengo, & Herrera, 2010).

As discussed in Section 2, defining and measuring comprehensibility can be quite subjective. Two main drivers, the output type and the size of the output, have been defined as key elements that impact the comprehensibility (Martens et al., 2011). A case study is presented where three different output types, linear output from LR, tree based output from DT and hybrid output from LLM and LMT, are discussed. The output size of DT and LR are measured respectively by the number of leaves and the number of variables included (Martens et al., 2011). RF are not included in the case study since this ensemble method is considered not interpretable without additional analyses such as partial dependence plots and variable importance scores.

5. Results

In this section, the results are presented. In the first part the results with regard to the predictive performance of LLM and the four benchmarks can be found and in the second part a case study offers a visualization of the output of DT, LR, LLM and LMT on a public dataset.

Table 10
Visualization decision tree on cell2cell[2] dataset.

Segm	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Rule7	Rule8	Pred.[1]
1	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* ≤ −0.30	*set prcm* ≤ 0	*creditde* ≤ 0	*changem* ≤ −1.14	*mou* ≤ 1.31		1
2	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* ≤ −0.30	*set prcm* ≤ 0	*creditde* ≤ 0	*changem* ≤ −1.14	*mou* > 1.31		0
3	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* ≤ −0.30	*set prcm* ≤ 0	*creditde* ≤ 0	*changem* > −1.14	*incalls* ≤ −0.57	*phones* ≤ 0.23	0
4	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* ≤ −0.30	*set prcm* ≤ 0	*creditde* ≤ 0	*changem* > −1.14	*incalls* ≤ −0.57	*phones* > 0.23	1
5	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* ≤ −0.30	*set prcm* ≤ 0	*creditde* ≤ 0	*changem* > −1.14	*incalls* > −0.57		0
6	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* ≤ −0.30	*set prcm* ≤ 0	*creditde* > 0				0
7	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* ≤ −0.30	*set prcm* > 0					0
8	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* > −0.30	*models* ≤ 0.57	*creditde* ≤ 0				1
9	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* > −0.30	*models* ≤ 0.57	*creditde* > 0	*callwait* ≤ −0.51	*eqpdays* ≤ 0.18		1
10	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* > −0.30	*models* ≤ 0.57	*creditde* > 0	*callwait* ≤ −0.51	*eqpdays* > 0.18		0
11	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* > −0.30	*models* ≤ 0.57	*creditde* > 0	*callwait* > −0.51			0
12	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* > −0.30	*models* > 0.57	*recchrge* ≤ 0.49				1
13	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* > −0.30	*models* > 0.57	*recchrge* > 0.49				0
14	*changem dummy* ≤ 0	*retcall* > 0							1
15	*changem dummy* > 0								1

[1] Class 1 indicating churn and 0 indicating non-churn.
[2] Center for Customer Relationship Management Duke University, February2014. http://www.fuqua.duke.edu/centers/ccrm

Table 11
Visualization logistic regression on cell2cell[1]dataset.

Intercept	eqpdays	retcall	creditde	changem	Changem dummy	recchrge	custare	models	setprcm	opeakvce	revenue	outcalls	webcap
0.14	0.19	0.96	-0.34	-0.10	1.28	−0.07	−0.06	−0.08	−0.14	0.05	−0.08	0.05	−0.09

[1] Center for Customer Relationship Management Duke University, February2014. http://www.fuqua.duke.edu/centers/ccrm

Table 12
Visualization logit leaf model on cell2cell[1] dataset.

	1st step: decision tree					2nd step: logistic Regression									
Seg	Rule 1	Rule 2	Rule 3	Rule 4	# obs.[2]	Shared variables							Segment specific variables		
						Incpt	retcalls	changem	Changem dummy		creditde	setprcm	eqpdays	directas	
1	*eqpdays* ≤ −0.31				5590	0.46	1.12	−0.07	0.96		−0.28	−0.63	0.23	−0.52	
						Incpt	retcalls	changem	Changem dummy		creditde	setprcm			
2	*eqpdays* > −0.31	*eqpdays* ≤ −0.06			1904	0.19	0.98	−0.24	1.59		−0.50	0.51			
						Incpt	retcalls			recchrge	creditde	custcare			
3	*eqpdays* > −0.06	*webcap* ≤ 0			1210	0.23	0.17			−0.19	−0.35	−0.20			
						Incpt	retcalls	changem	Changem dummy				revenue	mou	Outcalls
4	*eqpdays* > −0.06	*webcap* > 0	*callwait* ≤ −0.51	*changem* ≤ −0.05	1649	0.11	0.18	−0.29	1.66				0.42	−0.38	−0.20
						Incpt	retcalls	Changem		recchrge			models		
5	*eqpdays* > −0.06	*webcap* > 0	*callwait* ≤ −0.51	*changem* > −0.05	1192	−0.30	0.19	0.29		−0.14			−0.20		
						Incpt	retcalls	changem	Changem dummy	recchrge	creditde	custcare	opeakvce	callwait	
6	*eqpdays* > −0.06	*webcap* > 0	*callwait* > −0.51		2193	0.21	0.59	−0.10	1.37	−0.11	−0.36	−0.09	0.16	−0.21	

[1] Center for Customer Relationship Management Duke University, February2014. http://www.fuqua.duke.edu/centers/ccrm
[2] Number of training instances that fall within the segment.

5.1. Predictive performance

Tables 6 and 7 present the average cross-validated results over the different datasets in terms of AUC and TDL. The best performing classifier in each dataset is underlined. Table 8 shows the average ranks for every algorithm. These ranks are calculated based on the average results of every algorithm in terms of AUC and TDL per dataset in which lower ranks indicate better performance. These average ranks are the basis of a statistical analysis of model performance. The Friedman statistic is distributed according to chi square with four degrees of freedom and equals 30.47. The null-hypotheses that all classifiers are equal for both AUC ($p < 0.000$) and TDL ($p < 0.000$) are rejected, subsequently pairwise comparisons of the benchmark algorithms with LLM are performed using the Holm post-hoc test (García et al., 2010). The results of this analysis are summarized in Table 8 where the adjusted p-values for the Holm post-hoc test are given between brackets. It is clear from Table 8 that the hybrid LLM algorithm has the lowest average rank over all 14 datasets on both predictive performance metrics. It performs significantly better than its building blocks DT and LR. The performance of LLM is at least as good as the conceptually related LTM and the RF ensemble methods in this benchmark.

5.2. Results comprehensibility

In this section, the comprehensibility of the LLM is showcased by means of a case study focusing on customer churn prediction in a telecom setting. The output of the DT, LR, LLM and LMT on the cell2cell dataset are presented. This public dataset is well documented and has been used in previous customer churn studies

Table 13
Visualization logistic model tree on cell2cell[1] dataset.

Seg	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	# Obs[2]	Icpt	mou	recchrge	changem	custcare	revenue	outcalls	incalls	peakvce	opeakvce	callwait	phones	Models	eqpdays	creditde	webcap	retcalls	setprcm	retcall	Chagem_d
1	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* ≤ −0.30	*set prcm* ≤ 0	*creditde* ≤ 0	*changem* ≤ −1.14	398	0.62	−0.16	0.09	−0.13	−0.08	−0.11	−0.04	0.07	0.10	0.08	−0.07	−0.03	−0.11	0.16	−0.24	−0.27	.03	−0.33	.35	.62
2	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* ≤ −0.30	*set prcm* ≤ 0	*creditde* ≤ 0	*changem* > −1.14	2571	0.13	0	−0.02	0.03	0.01	−0.05	−0.03	−0.03	0.02	0.06	0.02	0.02	−0.02	0.15	−0.24	−0.15	.03	−0.33	0.35	0.62
3	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* ≤ −0.30	*set prcm* ≤ 0	*creditde* > 0		601	−0.16	0	0.05	−0.04	−0.03	0.02	−0	−0.03	0.04	−0.09	0.03	0.10	−0.09	0.12	−0.24	0.17	0.03	−0.33	0.35	0.62
4	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* ≤ −0.30	*set prcm* > 0			1778	0.24	0.14	−0.13	0.01	−0.02	0.03	0.07	−0.07	−0.02	−0.08	0.02	−0.13	0.08	−0.18	0.05	−0.48	0.03	−0.33	0.35	0.62
5	*changem dummy* ≤ 0	*retcall* ≤ 0	*eqpdays* > −0.30				7761	0.26	−0.05	−0.03	−0.07	−0.04	0.03	0.04	−0.06	0	0.05	−0.03	0	−0.1	−0.06	−0.19	−0.08	0.03	−0.06	0.35	0.62
6	*changem dummy* ≤ 0	*retcall* > 0					517	−0.10	−0.03	0.07	−0.13	−0.01	0.07	−0.03	0.03	−0.16	0.18	−0.12	0.02	−0.08	0.02	−0.1	0.09	0.04	−0.05	0.35	0.62
7	*changem dummy* > 0						112	−0.68	0.10	−0.03	−0.05	0.21	0	0.02	−0.03	−0.12	0.18	−0.04	0.34	−0.22	0.09	0.66	0.54	0.12	0.66	−0.40	0.62

[1] Center for Customer Relationship Management Duke University, February2014. http://www.fuqua.duke.edu/centers/ccrm
[2] Number of training instances that fall within the segment.

Table 14
Average number of terminal nodes in decision tree and logistic leaf model.

	Logit leaf model	Decision tree	Logistic model tree
Ds1	13.00 (5.48)	16.13 (4.24)	19.20 (8.64)
Ds2	2.50 (0.63)	14.30 (3.39)	2.67 (1.88)
Ds3	2.03 (0.18)	5.20 (2.17)	2.13 (0.35)
Ds4	5.40 (2.27)	11.13 (2.45)	7.07 (3.97)
Ds5	11.73 (4.95)	20.53 (4.58)	15.73 (4.51)
Ds6	4.93 (2.61)	23.03 (4.22)	5.33 (2.80)
Ds7	2.30 (0.88)	18.57 (8.92)	1.00 (0.00)
Ds8	2.33 (0.88)	5.07 (2.19)	2.87 (1.17)
Ds9	7.03 (2.32)	9.57 (2.35)	9.20 (4.43)
Ds10	1.97 (0.18)	8.93 (4.53)	1.40 (1.04)
Ds11	4.40 (1.50)	16.37 (4.23)	7.80 (4.70)
Ds12	8.97 (3.89)	15.73 (4.61)	6.67 (5.95)
Ds13	6.83 (2.74)	12.13 (2.06)	4.80 (2.11)
Ds14	2.43 (1.04)	13.07 (2.24)	3.73 (2.08)
Avg. Rank[1]	1.29	2.95*	1.79*

[1] Lower ranks indicate on average less terminal nodes Friedman statistic = 19.86, $p = 0.00$.
* Significant different ($p \leq 0.05$) from LLM using Holm post-hoc procedure.

Table 15
Average number of variables included in logistic regression and logistic leaf model.

	Logit Leaf Model	Logistic Regression	Logistic Model Tree
Ds1	3.63 (0.66)	10.07 (1.14)	17.08 (1.84)
Ds2	9.89 (1.30)	14.13 (1.48)	16.81 (1.70)
Ds3	3.50 (0.69)	4.47 (1.70)	7.10 (3.85)
Ds4	11.20 (1.81)	17.68 (1.42)	19.35 (1.27)
Ds5	5.66 (1.27)	10.60 (1.69)	15.86 (2.92)
Ds6	9.72 (2.26)	15.13 (1.17)	19.77 (0.29)
Ds7	8.98 (1.50)	11.47 (1.28)	13.13 (3.38)
Ds8	3.71 (0.92)	5.00 (0.74)	4.33 (1.38)
Ds9	8.18 (1.47)	16.53 (1.11)	19.93 (0.12)
Ds10	7.03 (1.14)	8.47 (1.78)	12.88 (6.21)
Ds11	11.09 (1.36)	16.40 (1.04)	19.02 (1.48)
Ds12	8.00 (2.51)	16.20 (1.19)	19.53 (0.46)
Ds13	8.55 (1.57)	14.67 (1.27)	19.73 (0.31)
Ds14	12.55 (1.45)	14.73 (0.78)	19.32 (1.02)
Avg. Rank[1]	1.00	2.07*	2.93*

[1] Lower ranks indicate on average less variables retained by the logistic regressions Friedman statistic = 26.14, $p = 0.00$.
* Significant different ($p \leq 0.05$) from LLM using Holm post-hoc procedure.

Table 16
Results predictive performance case study.

	LLM	LR	DT	LMT	RF
AUC	*0.630*	0.594	0.609	0.630	0.644
TDL	*1.561*	1.444	0.916	1.629	1.744

(Verbeke et al., 2012) . The dataset is also included in the benchmark part of this study as *ds5*. Therefore the data is preprocessed as described in Section 4.2 and variables are selected according to the methods described in Section 4.3. Table 9 gives an overview of the selected variables through applying the fisher selection. LLM combines different output types with an important role for the rule based output type which is in line with the recommendations of Martens et al. (2011) to obtain comprehensible classifiers. Visualizations of DT, LR and LLM are presented in Tables 10, 11 and 12, respectively.

Table 10 shows the output of the DT. It contains 15 terminal nodes with the number of decision rules ranging from 1 to 8. An analyst can simply follow the rules to find the prediction for the resulting class. As there are relatively many rules and terminal nodes in the DT, the LR can be considered more interpretable in this case.

Table 17
Results of the benchmarking experiment using the runtime (in s) performance criterion.

	Decision tree (DT)	Logistic model tree (LMT)	Logistic regression (LR)	Random forests (RF)	Logit Leaf model (LLM)
Ds1	0.08 (0.01)	1.54 (0.79)	1.93 (0.38)	4.14 (2.34)	1.92 (0.45)
Ds2	0.16 (0.01)	10.00 (3.85)	4.56 (0.58)	18.00 (3.92)	3.68 (0.34)
Ds3	0.04 (0.06)	0.19 (0.03)	0.16 (0.07)	0.87 (0.47)	0.30 (0.06)
Ds4	0.88 (0.08)	88.12 (42.70)	35.24 (2.53)	125.21 (34.45)	23.43 (2.17)
Ds5	0.47 (0.06)	9.89 (3.20)	10.16 (2.51)	49.31 (14.44)	7.07 (1.05)
Ds6	0.12 (0.01)	4.41 (1.40)	3.84 (0.49)	11.68 (2.95)	3.39 (0.28)
Ds7	0.06 (0.01)	0.89 (0.32)	0.97 (0.23)	2.54 (0.90)	0.99 (0.11)
Ds8	0.06 (0.01)	0.64 (0.21)	0.93 (0.11)	1.19 (0.89)	0.80 (0.15)
Ds9	0.32 (0.04)	19.36 (6.77)	12.27 (0.79)	29.34 (6.71)	8.23 (0.73)
Ds10	0.05 (0.01)	0.80 (0.34)	1.11 (0.35)	2.96 (1.02)	0.82 (0.14)
Ds11	0.37 (0.06)	19.05 (4.89)	10.79 (0.70)	47.03 (10.43)	8.86 (0.96)
Ds12	0.47 (0.05)	22.47 (5.92)	14.03 (1.26)	55.01 (9.38)	9.43 (2.10)
Ds13	0.41 (0.03)	21.48 (6.34)	12.03 (1.34)	22.99 (12.12)	9.21 (0.83)
Ds14	0.28 (0.05)	16.99 (5.36)	9.30 (0.90)	30.53 (6.04)	7.86 (1.19)

The LR includes 13 of the initial 20 selected variables and Table 11 provides a visualization of the LR. The advantage of this technique is that one can easily detect variables that have a positive or negative influence on customer churn. For example customers that have made a call to the retention team will have, ceteris paribus, a higher probability to churn, while high mean total recurring charge will lower the probability to churn for a given customer.

The output of the LLM is visualized in Table 12. The LLM divides the customer base in six segments based on 4 variables. The division into segments helps to increase the action ability of the model. An analyst can use the model output to detect exactly the drivers for each and every segment. Afterwards the analyst can take appropriate actions for every segment and tackle the drivers of the segment instead of the drivers for the entire data set. Notice that the number of terminal nodes is considerably lower than the 15 of the full DT, leaving enough entropy to be explained by the segment specific logistic regressions. The number of terminal nodes of DT, LMT and LLM for the other datasets are presented in Table 14 where the same observation can be made.

The same effects as in the full LR for the variables discussed above are detected in the shared variables but they are more, less or not important according to the segment that is analyzed. Some additional insights directly pop out. The variable percentage change of minutes of use has in general a negative impact on the churn probability but in the 5th segment, which is a segment with higher values for this variable, it, ceteris paribus, increases the churn probability. The segment specific variables also contain insights for analysts. For example the mean number of director assisted calls has a negative impact on the churn probability for customers in the first segment for which the number of days of current equipment is low. This insight was not captured in the standalone LR in which that variable is not included. These additional insights have a beneficial effect on the predictive performance as well since LLM beats both DT and LR in this dataset as the results present in Table 16.

Finally, the LMT is visualized in Table 13 using an identical representation as the LLM. We would like to stress two important points regarding the comprehensibility of the LMT. First, a larger number of terminal nodes emerges in the LMT in comparison to the LLM. This comparison is presented in Table 14. Therefore the output size is significantly larger and the LMT tends to become less comprehensible. Second, the logitboost algorithm in LMT retains significantly more variables in the regressions in the leaves as summarized in Table 15. In the LLM, logistic regressions with forward selection limit the number of variables and detect the churn drivers within each segment more clearly in this experiment.

6. Conclusions

In this section, the main conclusions are presented and the implications of the results are discussed. Typically, the choice for a classification technique is a trade-off between predictive performance and comprehensibility. Therefore the LLM is proposed as a viable choice that scores high on both requirements. As proven in this benchmark study, the LLM (1) provides more accurate models than using its building blocks, LR and DT, as standalone classification techniques. It performs at least evenly well as two homogenous ensemble methods RF and LMT, which are among the best performing classification techniques; (2) The LLM delivers a comprehensible method with benefits regarding the action ability of the model, which is its main advantage compared to the LMT and RF; (3) LLM can enrich both DT and LR . On the one hand, the leaves in a decision tree get a lot more depth by adding the coefficients of a logistic regression. On the other hand, several logistic regressions are fitted which can take into account specific group characteristics that remained otherwise unknown if only a single LR was trained. This method provides a conceptually easy, but efficient and accurate model. Table 17 provides the average model training time for the different classifiers in which the LLM scores well.

7. Limitations and future research

This section introduces several potentially interesting topics for future research with regard to the LLM.

As a topic of further research, there are a lot of op opportunities in model variations. On the hand one can use other models that are used in churn literature in the leaves like SVM based techniques (Lessmann & Voß, 2009; Verbeke et al., 2012), RF (Breiman, 2001) or naive Bayes (Neslin et al., 2006). Using different models in the leaves such as RF, potentially improve the predictive performance further but the comprehensibility and efficiency might suffer. On the other hand, variations of the decision tree can be researched and in this respect we refer to the rule induction techniques like PART (Frank & Witten, 1998) and RIPPER (Cohen, 1995) because they result in a similar structure as decision trees. Also rules extracted from black box models can be used as a first step in leaf modeling. Related to these model variations are model improvements. While churn data sets are rather large by nature, we do note that smaller datasets can pose a challenge for the LLM. The parameter settings for the minimum leaf size are important as they determine the number of clients in the leaves. Although logistic regressions are quite robust, the number of customers in the leaves should be high enough for a well-fitted logistic regression model. It might be possible to further improve the model by im-

posing more complicated rules for the number of leaves and the leaf sizes.

A second interesting area for further research is the application of the leaf modeling technique in other settings where the same trade-off between predictive performance and comprehensibility is present such as credit scoring where businesses want to predict good applicants versus bad applicants (Baesens et al., 2003; Thomas, Edelman, & Crook, 2002; Lessman, Baesens, Seow, & Thomas, 2015) or fraud prediction (Fawcett & Provost, 1997). Also in medical diagnosis LLM can be used in applications like e.g. dementia prediction (Pazzani, Mani, & Shankle, 2001).

Thirdly, AUC and TDL are used to assess the predictive performance of the model. These methods have their advantages but they lack a direct profit criterion. The recently introduced maxim profit (MP) criterion (Verbeke et al., 2012) provides a new evaluation metric that incorporates profit. This evaluation metric requires additional information to calculate the customer lifetime value which was not possible to gather for all datasets in this benchmark study. Therefore further research is needed to verify the performance of the LLM using this maximum profit (MP) criterion.

Fourthly, efficiency is briefly touched in this study, but a more elaborated framework to objectively compare the efficiency of different classification techniques may help analysts to make better decisions regarding what techniques to use. Especially in cases where real time predictions are needed and models have to be trained on the fly, model training time is an important metric to take into consideration.

Lastly, other authors have experimented with segmented modeling using unsupervised techniques (Hung, Yen, & Wang, 2006; Seret et al., 2012). It will be interesting to compare these techniques with the LLM and compare the segments that are created in LLM and in the unsupervised technique.

References

Anderson, D. R., Sweeney, D. J., Williams, T. A., Freeman, J., & Shoesmith, E. (2010). *Statistics for business and economics*. Cengage: Andover *(2nd ed.)*.

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

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

Ballings, M., & Van den Poel, D. (2012). Customer event history for churn prediction: How long is long enough. *Expert Systems with Applications, 39*, 13517–13522.

Blattberg, R. C., Kim, B. D., & Neslin, S. A. (2010). *Database marketing: Analyzing and managing customers*. New York, NY: Springer.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Bucklin, R. E., & Gupta, S. (1992). Brand choice, purchase incidence, and segmentation: An integrated modeling approach. *Journal of Marketing Research, 29*(2), 201–215.

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

Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert Systems with Applications, 34*, 2754–2762.

Chen, Z.-Y., Fan, Z.-P., & Sun, M. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of Operational Research, 223*(2), 461–472.

Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the twelfth 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., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research, 66*(9), 1629–1636.

Coussement, K., Harrigan, P., & Benoit, D. F. (2015). Improving direct mail targeting through customer response modeling. *Expert Systems with Applications, 42*(22), 8403–8412.

Coussement, K., Lessman, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication. *Decision Support Systems, 95*, 27–36.

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.

Dawes, J., & Swailes, S. (1999). Retention sans frontieres: Issues for financial service retailers. *International Journal of Bank Marketing, 17*(1), 36–43.

De Bock, K. W., & Van den Poel, D. (2012). Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. *Expert Systems with Applications, 39*, 6816–6826.

De Caigny, A., Coussement, K., & De Bock, K.W., 2018. LLM: Logit Leaf Model Classifier for Binary Classification, R Package version 1.0.0.

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

Didaci, L., Giacinto, G., Roli, F., & Marcialis, G. L. (2005). A study on the performances of dynamic classifier selection based on local accuracy estimation. *Pattern Recognition, 38*(11), 2188–2191.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923.

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

Frank, E., & Witten, I. (1998). Generating accurate rule sets without global optimization. In *Proceedings of the fifteenth international conference on machine learning* (pp. 144–151).

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

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistic, 38*(2), 337–374.

Fu, Z., Robles-Kelly, A., & Zhou, J. (2010). Mixing linear SVMs for nonlinear classification. *IEEE Transactions on Neural Networks, 21*(12), 1963–1975.

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

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced non parametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences, 180*, 2044–2064.

Giacinto, G., & Roli, F. (2000). Dynamic classifier selection. In *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science: vol 1857*. Berlin, Heidelberg: Springer.

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

Gustafsson, A., Johnson, M. D., & Roos, I. (2005). The effects of customer satisfaction, relationship commitment dimension, and triggers on customer retention. *Journal of Marketing, 69*(4), 210–218.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations Newsletter, 11*(1), 10–18.

Hansen, H., Samuelson, B. M., & Sallis, J. E. (2013). The moderating effects of need for cognition on drivers of customer loyalty. *European Journal of Marketing, 47*(8), 1157–1176.

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

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

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.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3*(1), 79–87.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.

Johnson, M. D., Nader, G., & Fornell, C. (1996). Expectations, perceived performance and customer satisfaction for a complex service: The case of bank loans. *Journal of Economic Psychology, 17*(2), 163–182.

Keaveney, S. (1995). Customer switching behavior in service industries: An exploratory study. *Journal of Marketing, 59*, 71–82.

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.

Kuncheva, L. I. (2000). Clustering-and-selection model for classifier combination.. In *Proceedings of the fourth international conference on knowledge-based intelligent engineering systems and allied technologies: 1* (pp. 185–188). IEEE.

Kuncheva, L. I. (2002). Switching between selection and fusion in combining classifiers: An experiment. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 32*(2), 146–156.

Kuncheva, L. I. (2014). Classifier selection. *Combining pattern classifiers* (pp. 230–246). John Wiley & Sons, Inc.

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

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

Lessman, S., Baesens, B., Seow, H., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*, 124–136.

Lessman, S., & Voß, S. (2009). A reference model for customer-centric data mining with support vector machines. *European Journal of Operational Research, 199*, 520–530.

Li, J., Weng, J., Shao, C., & Guo, H. (2016). Cluster-based logistic regression model for holiday travel mode choice. *Procedia Engineering, 137*, 729–737.
Lima, E., Mues, C., & Baesens, B. (2011). Monitoring and back testing churn models. *Expert Systems with Application, 38*, 975–982.
Ling, C., & Li, C. (1998). Data mining for direct marketing problems and solutions. In *Proceedings of the fourth international conference on knowledge discovery and data mining (KDD-98)*. New York, NY: AAAI Press.
Maldonado, S., Flores, A., Verbraken, T., Baesens, B., & Weber, R. (2015). Profit-based feature selection using support vector machines – General framework and an application for customer retention. *Applied Soft Computing, 35*, 740–748.
Maldonado, S., Pérez, J., & Bravo, C. (2017). Cost-based feature selection for support vector machines: An application in credit scoring. *European Journal of Operational Research, 261*(2), 656–665.
Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research, 183*, 1466–1476.
Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems, 51*(4), 782–793.
Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems, 72*, 72–81.
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.
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.
Nitzan, I., & Libai, B. (2011). Social effects on customer retention. *Journal of Marketing, 75*, 24–38.
Pazzani, M., Mani, S., & Shankle, W. (2001). Acceptance by medical experts of rules generated by machine learning. *Methods of Information in Medicine, 40*, 380–385.
Pyle, D. (1999). *Data Preparation for Data Mining*. San Francisco, CA: Morgan Kaufmann Publishers.
Rasmussen, C. E., & Ghahramani, Z. (2002). Infinite mixtures of Gaussian process experts. *Advances of Neural Information Processing Systems, 14*, 577–584.
Reinartz, W. J., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing, 67*, 1.
Shaffer, G., & Zhang, Z. J. (2002). Competitive one-to-one promotions. Management Science, , *48*(9), 1143–1160.
Seret, A., Verbraken, T., Versailles, S., & Baesens, B. (2012). A new SOM-based method for profile generation: Theory and application in direct marketing. *European Journal of Operational Research, 220*, 199–209.
Tang, L., Thomas, L., Fletcher, M., Pan, J., & Marshall, A. (2014). Assessing the impact of derived behavior information on customer attrition in the financial service industry. *European Journal of Operational Research, 296*, 624–633.
Thomas, L., Edelman, D., & Crook, J. (2002). *Credit scoring and its applications*. Philadelphia, PA: SIAM.
Torkzadeh, G., Chang, J. C.-J., & Hansen, G. W. (2006). Identifying issues in customer relationship management at Merck-Medco. *Decision Support Systems, 42*, 2.
Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research, 218*(1), 211–229.
Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Application, 38*, 2354–2364.
Verhoef, P. C. (2003). Understanding the Effects of customer relationship management efforts on customer retention and customer share development. *Journal of Marketing, 67*(4), 30–45.
Wei, C., & Chiu, I. (2002). Turning telecommunications call details to churn prediction:A data mining approach. *Expert Systems with Applications, 23*, 103–112.
Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems, 23*(8), 1177–1193.
Zhu, J., Chen, N., & Xing, E. P. (2011a). Infinite latent SVM for classification and multi-task learning. *Advances in Neural Information Processing Systems*, 1620–1628.
Zhu, J., Chen, N., & Xing, E. P. (2011b). Infinite SVM: A Dirichlet process mixture of large-margin kernel machines. In L. Getoor, & T. Scheffer (Eds.), *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 617–624). Bellevue, Washington, USA.