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/© 2018 Elsevier B.V. All rights reserved.

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