2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8334

A combination of multi-period training data and ensemble methods to improve churn classification of housing loan customers

Seppälä, Tomi; Thuy, Le
Department of Information and Service Management, Aalto University, School of Business, Finland

Abstract

Customer retention has been the focus of customer relationship management in the financial sector during the past decade. The first and important step in customer retention is to classify the customers into possible churners, those likely to switch to another service provider, and non-churners. The second step is to take action to retain the most probable churners.

The main challenge in churn classification is the rarity of churn events. In order to overcome this, two aspects are found to improve the churn classification model: the training data and the algorithm. The recently proposed multi-period training data approach is found to outperform the single period training data thanks to the more effective use of longitudinal data. Regarding the churn classification algorithms, the most advanced and widely employed is the ensemble method, which combines multiple models to produce a more powerful one. Two popularly used ensemble techniques, random forest and gradient boosting, are found to outperform logistic regression and decision tree in classifying churners from non-churners.

The study uses data of housing loan customers from a Nordic bank. The key finding is that models combining the multi-period training data approach with ensemble methods performs the best.

Keywords: *churn prediction, ensemble methods, random forest, gradient boosting, multiple period training data, housing loan churn*

1. Introduction

Customer retention has been the focus of customer relationship management research in the financial sector during the past decade (Zoric, 2016). Retaining existing customers is argued to be more economical over the long run for companies than acquiring new ones (Gur Ali & Ariturk, 2014). Van den Poel & Lariviere (2004), in their attempt to translate the benefits of retaining customers over a period of 25 years into monetary terms, concludes that an additional percentage point in customer retention rate contributes to an increase in revenue of approximately 7% (Van den Poel & Lariviere, 2004).The first step in customer retention is to classify the customers into binary groups of possible churners, indicating to customers that are likely to switch to another service provider, and non-churners, referring to those that are probably staying with the current provider. The second step in customer retention is to take action to retain the most probable churners to either minimize costs or maximize benefits. As a result, churn classification is an important first step in customer retention.

However, the main challenge in churn classification is the extreme rarity of churn events (Gur Ali & Ariturk, 2014). For example, the churn rate in the banking industry is usually less than 1%. In order to overcome this rarity issue, a great deal of research has been found to improve the two main aspects of a churn classification model: the training data and the algorithm (Ballings & Van den Poel, 2012). Regarding the training data, the recently proposed multi-period training data approach is found to outperform the single period training data thanks to the more effective use of longitudinal data of churn behavior (Gur Ali & Ariturk, 2014). Regarding the churn classification algorithms, the most advanced and widely employed is the ensemble method, which combines multiple models to produce a more powerful one (Yaya, et al., 2009). Two popularly used ensemble techniques are random forest and gradient boosting (Breiman, 2001), both of which are found to outperform logistic regression and decision tree methods in classifying churners and non-churners.

2. Research questions

To the best of the authors' knowledge, the proposed multi-period training data has not been applied with the ensemble methods in a churn classification model. As a result, in this study we examine whether the multi-period training data approach, when employed together with ensemble methods in a churn classification model, produces better churn prediction than with logistic regression and decision tree approach.

The research problem is detailed into the following research questions:

1. In models that employ logistic regression and decision trees, does the multi-period training data approach improve churn classification performance compared to the single period training data approach?

2. In models that employ single period training data, do random forest and gradient boosting improve churn classification performance compared to logistic regression and decision trees?

3. Do models that employ both the multi-period training data approach and ensemble methods perform better in churn classification than those in the first question?

4. What are the best churn predictors in the housing loan context?

3. Methods and Data

In order to answer the research questions, four methods are employed in this study as churn classification algorithms: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. All of the four methods are employed with both multi-period training data and single period training data to create competing models. Specifically, logistic regression and decision trees are used to run the baseline models with single period training data. The other models are then compared with the baseline models in order to answer the research question.

This study uses empirical data of housing loan customers from a Nordic bank. The data was collected and analysed for time period between August 1, 2015 and March 31, 2016, and was divided into observation and performance periods. The predictors employed in this study include information from all the four main groups as recommended in the literature: demography, customer behavior, characteristics of customer relationship and macro-environmental factors. The following variable selection methods are employed in the SAS Enterprise Miner before running the logistic regression models: a decision tree with the CHAID method, step wise selection, and the variable selection node using the R square procedure. The churn models are evaluated based on three criteria: misclassification rate, Receiver Operating Characteristics (ROC) index and top decile lift.

4. Results and discussion

This study validates that both multi-period training data and ensemble methods actually improve the churn classification performance compared to their counterparts in the housing loan context. More importantly, when employed together, the models with the combination

of the proposed multi-period training data approach and ensemble methods such as random forest and gradient boosting have the best performance among all the created models based on the misclassification rate, ROC index and top decile lift. Type II error refers to misclassifying churners as non-churners and it is more severe than misclassifying non-churners as churners (Type I error) since potential churners will be highly likely to churn without receiving any retention action. Using multi-period training data, the best models are produced with the random forest with a reduction of more than 10% in type II error rate compared with the worst performing models that employ single period training data and logistic regression without variable selection. Such improvement in the misclassified churn events can considerably prevent the bank from a considerable loss of those customers without taking any retention action. The improvement is mainly thanks to the more effective use of churn events that are usually scarce in real life data. Specifically, in contrast to the single period training data approach that captures churns only at a specific period of time and discards the churn events that have happened prior to that period, the multi-period training data approach allows the employment of historical churn events, providing the models with more churn events and mitigating the rarity issue in churn prediction. Therefore the imbalance between the classes is not as severe a problem when using the multiperiod training data approach. Consequently, the authors highly recommend other studies in churn classification to employ the multi-period training data approach together with ensemble methods to achieve the best possible classification models.

Regarding the last research question, this study shows that the most important churn predictors belong to the demographic group, in which the number of family members has the most significant effect on churning. It makes sense that a change in the number of family members, will considerably impact the decision related to a housing loan.

References

Breiman, L.(2001). Random Forests. *Machine Learning*, 45(1), 5-32.

Ballings, M. & Van den Poel, D. (2012). Customer Event History for Churn Prediction - How Long Is Long Enough?. *Expert Systems with Applications*, 39 (18), 13517-13522.

Gur Ali, Ö. & Ariturk, U. (2014). Dynamic Churn Prediction Framework with More Effective Use of Rare Event Data: The Case of Private Banking. *Expert Systems with Applications*, 41(17). 7889-7903.

Van den Poel, D. & Lariviere, B. (2004). Customer Attrition Analysis for Financial Services Using Proportional Hazard Models. *European Journal of Operational Research*, 157 (1), 196-217.

Yaya, X., Xiu, L., E.W.T., N. & Weiyun, Y. (2009). Customer Churn Prediction Using Improved Balanced Random Forests. *Expert Systems with Applications*, 36(3),5445–5449.

Zoric, A. B. (2016). Predicting Customer Churn in Banking Industry Using Neural Networks. *Interdisciplinary Description of Complex Systems*, 14(2). 116-124.