Bank Churn Prediction Using Random Forest and Logistic Regression

Shangxuan Du

School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Shenzhen, Guangdong Province, 518172, China

118010052@link.cuhk.edu.cn

Abstract. In the banking industry, customer churn is a growing problem. Solving this problem effectively and choosing the appropriate forecasting model is important. To avoid such problems and select an appropriate model, in this paper, random forest and logistic regression models are used to predict bank churn based on a specific data set. Different situations are set to evaluate the models. During prediction, the parameters of the model and variables are changed slightly for comparison. The accuracy, recall, precision, and stability of models are compared. The accuracy of random forest is about 86%, nearly 3 points higher than logistic regression. After removing the least correlative factor, the accuracy of the random forest remained almost unchanged, while logistic regression had a 4-point decline. Fluctuation brought by removing the least correlative variable is smaller in the random forest which means better stability. Though this study has shown a random forest's better performance, removing the least correlative factor leads to a decline in both models. This is contradicted by the author's hypothesis. Hence, further study with enough features will be a good way to compare more about these two models in bank churn prediction.

Keywords: Random forest model (RFM), logistic regression model (LRM), bank churn, machine learning (ML)

1 INTRODUCTION

Customers terminate their accounts with this banking institution and transfer their assets to another financial entity, commonly called customer churn. Due to the sharp rise in the number of services in various industries, there is no shortage of places for banking customers to invest their money. Consequently, one of the main concerns for the majority of banks is now customer churn and engagement [1]. The impact of customer churn is intuitive and dramatic. Acquiring a new customer means spending more money on sales, marketing, advertising, and payroll. It costs a lot of money and most new customers do not generate as much profit as those that leave [2]. Therefore, to increase the company's competitiveness, it is strategically important to avoid business issues brought on by customer churn. To reduce the damage of customer churn, the company needs to predict their client behavior with a convincing accuracy, build

A. Bhunia et al. (eds.), *Proceedings of the 2024 2nd International Conference on Finance, Trade and Business Management (FTBM 2024)*, Advances in Economics, Business and Management Research 304,
https://doi.org/10.2991/978-94-6463-546-1_2

relationships between losing customers, and keep management over those aspects. Machine learning (ML) models are an effective way to predict and mitigate customer churn. The process of data analysis known as ML automates the creation of analytical models. Algorithms from ML learn from data iteratively. With the help of ML, systems can uncover hidden patterns without explicitly being told where to look [3]. From the analysis of ML, ML has the advantages that easily identifying trends, handling multi-variety data, and wide applications [4]. So it is efficient and suitable for predicting the loss of bank customers.

There are lots of algorithms in ML that can finish the task of prediction. So far, some researchers have already figured out that an improved balanced random forest is found to improve prediction accuracy significantly compared with other algorithms [5]. Michael has pointed out that an analysis method named logistic regression supports the analysis of binary outcomes with 2 mutually exclusive levels [6]. In other words, logistic regression is also adapted to the bank churn prediction. Bank churn prediction is a binary classification problem that is quite normal in people's lives. Although this study is limited to a specific bank data set, the study and analysis could apply to many other service industries even to various binary classification problems. It helps people anticipate outcomes and adjust their behavior accordingly.

To date, many methods have been used to predict bank churn problems, but only a few reports have explored how different algorithms perform on the same data set. The comparison of different methods can help people choose appropriate methods to predict more accurate results in problems. Thus, this paper is aimed at comparing the performance of two different methods on the same data set. The data set is taken from a website called Kaggle. It contains around ten thousand information on bank customers who either left the bank or continue to be a customer.

This paper looks at the correlation between the factors at the beginning. Then Random Forest Model and Logistic Regression Model are applied to the bank data set to predict the bank churn. Compare the performance of the two models in predicting the bank churn.

2 METHOD

Data used in this study is taken from a website called Kaggle, a platform offering plenty of data sets for learners to practice and compete. Each row represents a set of information about a customer. Each column represents a feature of the customer, like gender, age, geography, and so on. More than ten thousand rows and fourteen features are given by the data set.

2.1 Data Preprocessing

Before predicting, this study processes the data in three steps: cleaning, transition, and scaling. While gathering and analyzing data is crucial, data quality is still a common and challenging issue in practically all large organizations. The inaccuracy and inconsistency of data will greatly affect the result of model analysis [7]. The data set is cleaned first. Because the database this study used is quantitative enough, missing data and duplicate data are dropped. Besides, based on logic, a feature that does not affect predicted results is discarded. For example, feature index and name are dropped in this study. After cleaning, part of the data is adjusted by the encoder. The database is composed of numerical values and categorical variables. So categorical variables like gender and geography are transformed into numerical values. Usually, the model will understand that the variable with 10000 is much more important than the variable with value 0.5, which is not necessarily the case, so the values should be scaled to the same scale so that the proportion between them stays the same but the huge gap is reduced around a fixed value. The author scales the value of each feature to the same level where one is the max value and zero is the minimal value. So far, the preprocessing of the data set has been completed.

2.2 Correlation

To better conduct data prediction analysis and comparison, the correlations among all features are found and plotted. This helps the author arrange different situations to test the models that this study used.

2.3 Prediction

Random forest model (RFM) and logistic regression model (LRM) are used to predict the data and do the comparison. There will be two sections in predicting. In the first section, the evaluation and accuracy of the two models are extracted and compared. This is an attempt to find which model suits the bank churn prediction well. In the second section, the models are going to be tested in different situations to do the comparison. The situations in this study are changing the parameters of models and removing the feature that has the lowest correlation.

The author gives the hypothesis that removing the least important variable will improve the accuracy of the model. Optimizing the parameters in the model will improve the accuracy of the model.

3 RESULTS

This study applies RFM and LRM to a set of churn data in a bank. In the beginning, the two models are used with default parameters to find which model would perform better in bank churn prediction. With the train set and the test set, the prediction is finished by returning the result 0&1, where 0 means the customer will choose to terminate his account and 1 means the customer will continue his account in this bank. Classification reports also known as evaluation of models are shown in Table 1&2.

Table 1. Evaluation of random forest model in prediction.

	Precision	Recall	F1-score	Support
0	0.88	0.94	0.91	45554

1	0.72	0.53	0.61	12141
Accuracy			0.86	57695
Macro avg	0.80	0.74	0.76	57695
Weighted avg	0.85	0.86	0.85	57695

Table 2. Evaluation of logistic regression model in prediction.

	Precision	Recall	F1-score	Support
0	0.84	0.96	0.90	45554
1	0.67	0.34	0.45	12141
Accuracy			0.83	57695
Macro avg	0.76	0.65	0.67	57695
Weighted avg	0.81	0.83	0.80	57695

From the table 2 shown above, with default parameters, the random forest model is superior to logistic regression in terms of precision, recall, and accuracy. Without any optimization, RFM is appropriate to bank churn prediction better. To further find the application of the two models in this problem, more steps are put into effect. The correlation results are in Fig. 1 and Table 3.

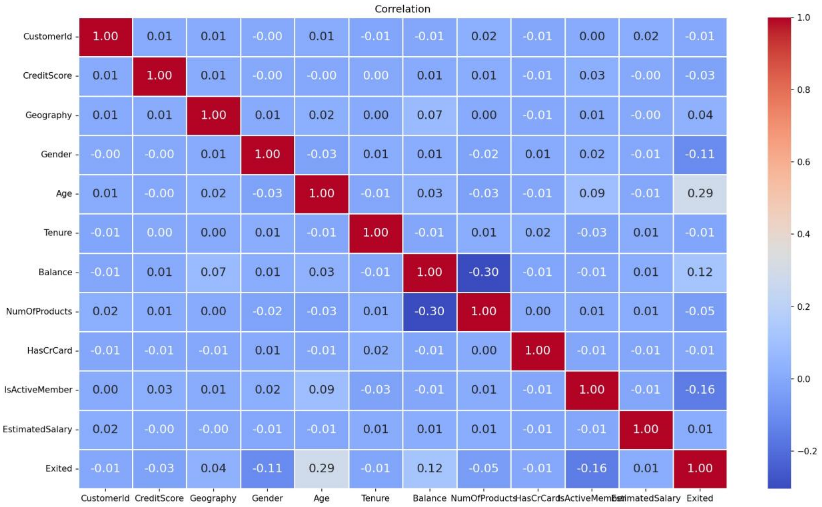

Fig. 1. Correlation among all factors by Heatmap (Original).

Table 3. Details of correlation.

	Exited
IsActiveMember	-0.1600
Gender	-0.1065
NumOfProduct	-0.0478
Creditscore	-0.0270
Tensure	-0.0141
HasCrCard	-0.0071
CustomerId	-0.0063
EstimatedSalary	0.0120

Continue Table 3.

Geography	0.0359
Balance	0.1186
Age	0.2854
Exited	1.0000

Associated with exited results, age has the highest correlation with prediction results while the customer has the lowest correlation. Thus, different situations are set to test the performance of two models in prediction. Situation 1(S1) is set as the control group with default parameter: estimators n=100 for random forest and C=100 for logistic regression. Situation 2(S2) is set with different estimators n=500 and C=300. Situation 3(S3) is set with estimators n=100 or C=100 and the data set is adjusted by removing the least correlative factor ‘CustomerId’. The accuracy is shown in Table 4 and Table 5.

Table 4. Accuracy for random forests in different situations.

	S1	S2	S3
Accuracy	0.8581	0.8587	0.8558

Table 5. Accuracy for logistic regression in different situations.

	S1	S2	S3
Accuracy	0.8264	0.8264	0.7873

The result shows that increasing estimators of random forest would improve its performance which is consistent with the hypothesis. Accuracy for logistic regression remains the same. However, removing the least correlative factor hurts predicting. This is contrary to the hypothesis. At the same time, the adjustment of data affects the logistic regression to a greater extent.

4 DISCUSSION

The comparison in result shows readers that in the bank churn prediction problem, RFM performed better than LRM. A model’s accuracy may be simply calculated and is a great overall average of how well it can forecast. Furthermore, recall is the capacity to locate all pertinent instances, whereas precision is the percentage of the data points that the model deems pertinent. This is what Kaitlin Kirasich used to tell us [8]. Accuracy, recall, and precision are good criteria to judge the model. The random forest model is superior to the logistic regression model in every respect during this prediction. This is consistent with the author’s hypothesis. The reason for that can be concluded by previous studies. Bank churn prediction is a binary classification problem. With the help of previous studies of other scholars, it is recommended that the original version of random forest with default parameters has a better performance than logistic regression as a binary classification tool [9].

Limitation also exists in this study. When the author tried to further test the performance of models by removing the least correlative factor, the result contradicted with hypothesis. After removing the least correlative factor, both models have a decline in their accuracy. It is known from other scholars that, the effectiveness of ML and prediction in classifying client behavior has greatly increased, but feature sizes, class imbalance, and diverse data processing still severely restrict their capabilities [10]. Logically, dropping the most irrelevant factor is like removing or pulling out a decayed tooth. This could be for the following reasons. An insufficient quantity of features made each feature non-negligible. Fourteen features seem to be not enough for this procedure. If people want to see a rise by removing the least correlative feature, a future study with quantity features is necessary. Despite this limitation, it also shows that random forests have better stability when facing fluctuation of data.

Hence, the random forest model could be a nice choice when people encounter bank churn prediction even binary classification problems.

5 CONCLUSION

This study uses LRM and RFM to forecast customer attrition in the banking sector and compares their effectiveness. Compared with logistic regression, random forest has higher accuracy, recall, and precision in bank churn prediction. Through fluctuation tests, the random forest also shows strong stability. In summary, random forest is more effective than logistic regression in bank churn prediction. Bank churn prediction is a standard binary classification problem, so this study can also be employed in other similar classification problems. This study also helps companies to select an appropriate model to settle their matters on customer churn.

However, this study has an obvious limitation. From the result shown above, removing the least relative feature leads to an accuracy decline in both models. However, the author gives the hypothesis that removing the least correlative feature could improve the model. Though this test showed better stability for random forests, its influence contradicted the hypothesis. After analysis, the contradiction could be for two possible reasons. One is insufficient quantity of features makes each feature indispensable to the prediction. Another one is that the importance of features needs more evaluation rather than only correlation. Future studies can explore this by adding more features to the data set and building an evaluation system to evaluate the features to decide which feature should be removed.

REFERENCES

1. Rahman, M., Kumar, V: Machine learning based customer churn prediction in banking. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1196-1201 (2020).
2. Zhaohan S, Bingxiang L.: Decision tree method for customer churn crisis analysis. Journal of Management Science and Engineering, (02), 20-25 (2005).

3. Saran Kumar, A., Chandrakala, D.: A survey on customer churn prediction using machine learning techniques. International Journal of Computer Applications, 975, 8887 (2016).
4. Khanzode, K. C. A., Sarode, R. D.: Advantages and disadvantages of artificial intelligence and machine learning: A literature review. International Journal of Library & Information Science (IJLIS), 9(1), 3 (2020).
5. Xie, Y., Li, X., Ngai, E. W. T., Ying, W.: Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), 5445-5449 (2009).
6. LaValley, M. P.: Logistic regression. Circulation, 117(18), 2395-2399 (2008).
7. Hellerstein, J. M.: Quantitative data cleaning for large databases. United Nations Economic Commission for Europe (UNECE), 25, 1-42 (2008).
8. Kirasich, K., Smith, T., Sadler, B.: Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Science Review, 1(3), 9 (2018).
9. Couronné, R., Probst, P., Boulesteix, A. L.: Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics, 19, 1-14 (2018).
10. Tékouabou, S. C., Gherghina, Ș. C., Toulni, H., Mata, P. N., Martins, J. M.: Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods. Mathematics, 10(14), 2379 (2022).