

Comparative Analysis of Several Models for Churning Customer Prediction

Zhaoyuan Tan*

School of Finance, Tianjin University of Finance and Economics, Tianjin 300222, China

Abstract. Customer churn prediction is critical for financial institutions to retain clients and optimize resource allocation. It is less expensive to keep current clients than to find new ones. There lots of research in this field, but their performance is often limited by data imbalance issues. This study compares three machine learning models: Random Forest, XGBoost Classifier, and Light Gradient Boosting Machine Classifier for predicting credit card customer churn using a dataset from Kaggle. The research addresses data imbalance issues through oversampling techniques (SMOTE, SMOTEENN, Borderline SMOTE) and evaluates model performance using accuracy and F1 score. Results show that the LGBM Classifier with Borderline SMOTE achieves the highest accuracy (97.43%) and F1 score (0.9259), outperforming other methods. This approach effectively balances precision and recall, improving minority class prediction. These findings provide actionable insights for financial institutions to implement proactive retention strategies. There are still limitations and future work to do. More different datasets, updated models for small datasets, and more feature engineering methods should be taken into consideration.

1 Introduction

Customer churn prediction is a key challenge in the financial industry. With the research of Verhoef, retaining existing customers is more cost-effective than searching for new ones [1]. Due to intense market competition and changes in customer preferences, credit card services are especially faced with a high churn rate. Previous studies have shown that machine learning models can effectively identify potential churn customers. For instance, Praveen Lalwani uses Adaboost to gain the highest accuracy and the AUC rate of 81.71%, but their performance is often limited by data imbalance issues [2, 3]. An insufficient number of samples in the minority class (churn customers) may cause models to be biased towards the majority class, weakening their practicality.

To address this issue, this study evaluates the performance of three machine learning models (Random Forest, XGBoost, and Light Gradient Boosting Machine, or LGBM) based on a real-world credit card customer dataset from Kaggle. In addition, it compares the optimization effects of three oversampling techniques (SMOTE, SMOTEENN, and Borderline SMOTE) on data balancing and minority class prediction. These findings can help

* Corresponding author: beholder323@stu.tjufe.edu.cn

banks and companies keep their customers by predicting who might leave early and then taking corresponding actions to keep these customers.

2 Researching method

2.1 Data description

The dataset used in the paper is from Kaggle, titled “Credit Card Customers” [4]. This content has a large amount of data from 2021 and the formal period. The classification goal is to predict whether the customers will leave their credit card service (variable y).

It can be seen that there are a total of 10126 records and 20 fields mentioning the age, salary, marital status, credit card limit, credit card category, and some more fields in the dataset. 9 of the fields are numeric fields, and the other 11 of are categorical fields.

To identify the null value status of each variable in the dataset, a commonly used null value checking method in data processing software was employed. Specifically, the built-in functions of the software were utilized to check each element of each variable in the dataset one by one. There is no null data in the dataset, the data is clean and well-prepared for the next step.

Use the method of label encoding for categorical features, which is meant to change the raw data into data that machine learning models can recognize directly.

Determine whether numerical features have any relationships or correlations with one another, and to compute the correlation between them, use a heat map to show the relationship and get some conclusions.

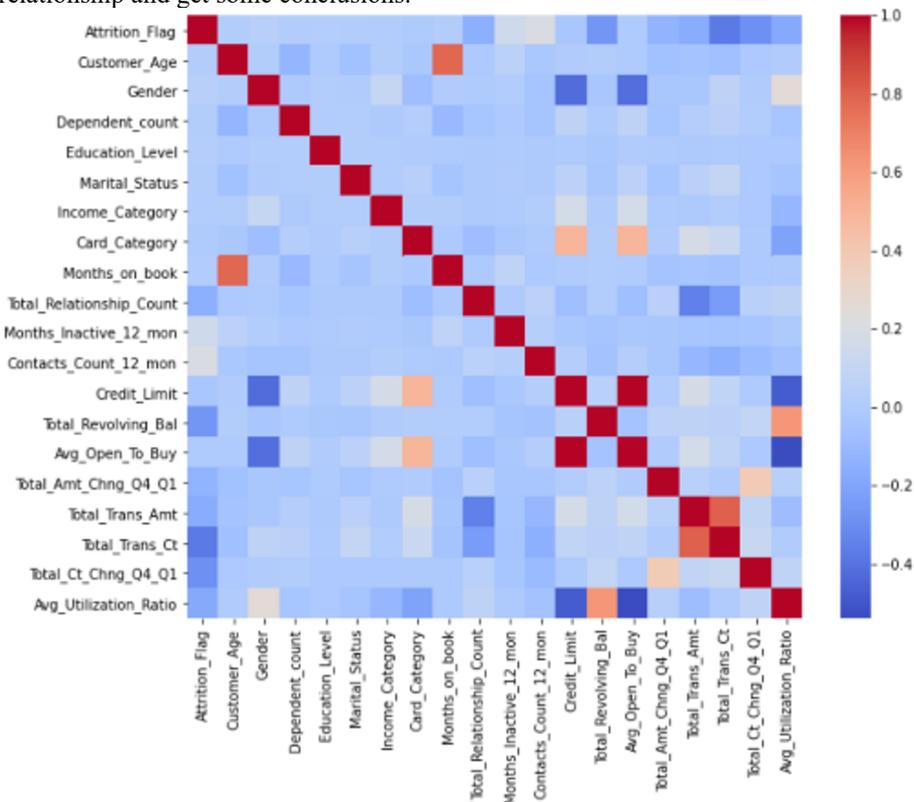


Fig. 1. The heat map: feature-by-feature visualization (Photo/Picture credit: Original).

As Fig. 1 demonstrates, the status of a customer, either existing or lost, is closely related to the number of months the customer was inactive and the number of contacts customers made in the past year. Also, the change of the transaction count provides a negative impact and may lead to the abandonment of the credit card.

2.2 Modeling

2.3.1 Train test split

This section is to separate the raw dataset into a test set and a training set, which helps to test whether the model is predictive. Then, 20 percent of the data is put into the test set.

2.3.2 Initial model attempts

Build three models for prediction, including Random Forest, XGB Classifier, and LGBM Classifier.

Random Forest operates by constructing an ensemble of decision trees trained on bootstrapped subsets of the input data and feature space. Each tree generates an independent classification, and the majority vote determines the final projection. This architectural design inherently mitigates overfitting by introducing randomness during the training phase, thus improving generalization capabilities.

XGBoost Classifier represents an advanced implementation of gradient-boosted decision trees. It sequentially builds models to minimize prediction errors by leveraging gradient descent optimization. A key innovation lies in its regularization techniques, which prevent overfitting while maintaining computational efficiency. XGBoost also incorporates parallel processing and columnar data storage, enabling scalable performance on large datasets.

The LightGBM Classifier adopts a leaf-wise growth strategy, which focuses more on reducing losses compared to the traditional model. By discretizing features using a histogram-based algorithm, it can significantly improve computational efficiency. Meanwhile, it uses the feature bundling technique to group features that are rarely non-zero simultaneously.

Find out the performance of the mentioned model with imbalanced data. Since the data is imbalanced, there are more records to train the majority class, which may cause the models to be biased towards the majority class. Compare the total accuracy and F1 score of the minority class as the standards of comparison.

2.3.3 Data balancing with different methods

The performance of different models is compared to determine the best model with the highest accuracy and F1 score. Based on the best model, multiple methods can be used to balance the dataset, including AMOTE, SMOTTEN, and Borderline SMOTE.

Firstly, try to oversample the minority class through SMOTE. It works by generating synthetic samples of the minority class in the feature space, thereby increasing the number of minority class samples in the training class and improving the performance of machine learning models on the minority class.

Secondly, use the method of SMOTEENN to balance the dataset, which is based on the oversampling by SMOTE and is improved by adding the undersampling method by ENN to clean the training set. It is a method that can optimize the dataset distribution and make the class boundaries clearer.

Thirdly, adopt the method of Borderline SMOTE, which, often combined with random undersampling, mitigates class imbalance. It focuses on generating synthetic samples for the minority class samples that are near the decision boundary and mainly enriches the data around the critical areas where misclassification is more likely to occur. Compare the total accuracy and F1 score of the minority class as the standards of comparison.

3 Result

3.1 Best model

Table 1. Comparison of Evaluation Metrics for RF, XGB, and LGBM Classification Models

	Accuracy	Precision	Recall	f1-score
Random Forest	96.49%	92%	83%	0.8761
XBGClassifier	96.79%	91%	87%	0.8930
LGBMClassifier	97.29%	91%	90%	0.9082

Table 1 illustrates a comparison of the evaluation metrics for three classification models: Random Forest (RF), XGBoost Classifier (XGB), and Light Gradient Boosting Machine Classifier (LGBM). In terms of accuracy, all three models demonstrate high levels, indicating their good capabilities in the classification task.

3.2 Best method for data balancing

Table 2. Comparison of Performance Metrics for SMOTE, SMOTEENN, and Borderline SMOTE Oversampling Methods

	Accuracy	Precision	Recall	f1-score
SMOTE	97.29%	91%	90%	0.9082
SMOTEENN	95.80%	84%	92%	0.8801
Borderline SMOTE	97.43%	91%	95%	0.9259

According to Table 2, it seems that the model does better both on the accuracy and the F1 score than the unbalanced data, achieving 97.29% and 0.9082, which means the model is more predictive after oversampling.

It seems that the result of SMOTEENN is not appropriate in the model or not good for the dataset, as the accuracy and the F1 score suffered a decline of only 95.80% and 0.8801, respectively; the accuracy is even lower than data that have not been balanced.

It seems that the result of Borderline SMOTE achieved a higher accuracy and F1 score than all attempts in former sections. Impressively, the model attains an accuracy of 97.4% and an F1 score of 0.9259, demonstrating outstanding performance and high-level effectiveness.

4 Discussion

4.1 Result analysis

The results demonstrate that the LGBMClassifier outperformed the Random Forest and XGBoost in imbalanced datasets. This superiority can be attributed to LGBM's leaf-wise growth strategy and histogram-based discretization, which enhance computational efficiency while maintaining high predictive accuracy. The model's ability to handle large-scale

datasets with high-dimensional features likely contributed to its robust performance in capturing complex patterns within customer behavior data.

Notably, the use of Borderline SMOTE significantly improved the model's ability to classify minority class samples, achieving the highest F1 score (0.9259) and accuracy (97.43%). This method generates synthetic samples near decision boundaries and is likely to enhance the model's attention to regions where there exists a high risk of misclassification and avoid adding samples that are easy to be confused blindly. In contrast, SMOTEENN's mixed approach resulted in reduced accuracy and F1 scores, possibly due to the removal of some important noise during undersampling, which may have eliminated informative minority class samples.

4.2 Limitation of the result

Several limitations should be acknowledged. First, the dataset is from a single source (Kaggle's "Credit Card Customers") and only includes approximately 10,000 samples. It might result in overfitting, whereas the model is able to remember all details of the training set, including the noise and null data, and perform well in this test set but may have poor performance in another dataset according to the research by Chiyuan Zhang [5].

To solve the problem, forefront models should be taken into consideration. For instance, Noah Hollmann put out research about a model named TabPFN that works better in a small or middle dataset [6]. Moreover, more different datasets from different fields can be taken into consideration to relieve the situation of overfitting from the research of Shai Ben-David [7].

Second, Pedro Domingos has mentioned the importance of feature engineering [8]. However, this study did not explore much about it, such as feature extraction and feature selection, which might enhance model performance, and only includes label encoding and correlation analysis.

Third, with the consideration put forward by Ninareh Mehrabi, it is crucial to ensure that these decisions do not reflect discriminatory behavior toward certain groups or populations [9]. Though there is a rank of factors that affect the decision, it is not transparent because the machine learning system does not give out a clear reason for which field causes the result to a single person.

To settle the issue, David Alvarez Melis put out a method to create a complex model with simple linear models [10]. It possesses the good interpretability properties of linear models locally while not limiting the model's performance.

5 Conclusion

The study shows that different machine learning models and data balancing methods can effectively predict customer churn for credit card services. First, to read the data and apply some preparatory measures such as data cleaning, label encoding, and correlation to the dataset, which helps the model to better input the data.

Then, three models are tested they are Random Forest, XGBClassifier, and LGBMClassifier, using the original imbalanced dataset. The LGBMClassifier is better at predicting customers who would leave (minority class) because it balances precision and recall well.

Finally, SMOTE, SMOTEENN, and Borderline SMOTE were used to improve the accuracy of the LGBM classifier. This method gave the best results; the LGBMClassifier achieved 97.4% accuracy and an F1 score of 0.93, which is much higher than other methods. In summary, the LGBMClassifier with Borderline SMOTE worked best. These findings can help banks and companies keep their customers by predicting who might leave early. There

are also some future works to complete. First, to add multi-source data and time-series analysis for further exploration to churning customers in different sight. Also, use forefront models such as TabPFN for small datasets to improve the performance of the model. Second, to add more about feature engineering, which helps to dig more about the inner value of the data and improve the model.

Last, to apply the proposed LGBM-Borderline SMOTE framework to high-churn industries, validating its universal applicability and exploring more existing deficiencies.

References

1. P. C. Verhoef, The role of relational information processes in customer retention. *J. Mark.* **67**(1), 31–44 (2003)
2. P. Lalwani, M. K. Mishra, J. S. Chadha, P. Sethi, Customer churn prediction system: a machine learning approach. *Comput.* **104**(2), 271–294 (2022)
3. S. Goyal, Credit card customers: predict churning customers. Kaggle. Available: <https://www.kaggle.com/datasets/arjunbhasin2013/credit-card-customers>
4. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30** (2017)
5. C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization. *Proc. Int. Conf. Learn. Represent. (ICLR)* (2017)
6. N. Hollmann, S. Müller, L. Purucker, Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326 (2025)
7. S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. Wortman Vaughan, A theory of learning from different domains. *Mach. Learn.* **79**(1-2), 151–175 (2010)
8. P. Domingos, A few useful things to know about machine learning. *Commun. ACM* **55**(10), 78–87 (2012)
9. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **54**(6), Article No. 115, 1–35 (2021)
10. D. Alvarez Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks. *Adv. Neural Inf. Process. Syst.* **31** (2018)