Proceedings of CONF-BPS 2026 Symposium: Sustainability Transitions and Regional Economic Restructuring in New Contexts
DOI: 10.54254/2754-1169/2026.NJ30790

Machine Learning-Based Bank Customer Churn Prediction

Boyu Shang

Department of Medical Informatics, Harbin Medical University (Daqing), Daqing, China
tina006633@gmail.com

Abstract. In the context of rapid fintech development and increasingly fierce industry competition, customer churn has become a key issue affecting the sustainable development of banks. Therefore, accurately predicting customer churn, analyzing its main influencing factors in depth, and taking effective measures to retain existing customers are crucial for banks to achieve stable growth in a highly competitive environment. This study investigated the issue of bank customer churn based on machine learning techniques by building a classification prediction model to identify churned and non-churned customers. The study first conducted an exploratory data analysis (EDA) of the bank customer data, including missing value handling, feature distribution analysis, and correlation analysis. Next, data preprocessing was performed. To mitigate the prevalent class imbalance in customer churn data, the SMOTE oversampling method was employed. By applying feature engineering, the dataset was transformed through encoding and normalization, and several churn-related derived variables were added. Afterwards, multiple machine learning models, such as Logistic Regression, Random Forest, and XGBoost, were trained, with parameter tuning conducted via cross-validation and grid search. The findings showed that among all tested models, XGBoost delivered the strongest predictive performance for customer churn, achieving an accuracy of 0.8595, an F1-score of 0.6298, and an AUC value of 0.8652. It effectively identifies potential churn customers and quantifies the impact of various features, providing decision support for banks to implement proactive and precise customer retention strategies.

Keywords: Customer churn, XGBoost, Random Forest, Logistic Regression

1. Introduction

In recent years, alongside the swift advancement of financial technology and the increasingly fierce competition in the banking industry, customer liquidity has increased significantly, and customer churn has become one of the main challenges affecting the sustainable development of banks. Since attracting new customers typically requires substantially higher costs than keeping current ones, accurately identifying potential churned customers has become crucial for banks to enhance their competitiveness and realize long-term customer value.

In the early studies on customer churn issues, traditional statistical methods were mostly employed for churn prediction, such as the work by Ye Jin et al., who conducted predictive analysis on the churn situation of telecommunications customers based on Bayesian networks [1].

DOI: 10.54254/2754-1169/2026.NJ30790

Subsequently, machine learning algorithms began to be widely used in the problem of predicting customer churn. Ma used Random Forest (RF), XGBoost, and LightGBM models to predict customer churn. Based on evaluation metrics such as the Kappa coefficient, LightGBM was ultimately selected as the final customer churn early warning model due to its superior prediction performance and operational efficiency [2]. Zhang used the SMOTE algorithm to solve the sample imbalance problem and built a user churn prediction model using Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM) and XGBoost. The final results showed that the XGBoost model outperformed the other models [3]. Balaji et al. compared several prediction models, including LR, DT, RF, and Gradient Boosting Classifiers (GBC). They employed extensive feature engineering and targeted feature selection to improve data representation and mitigate overfitting. The model's performance was comprehensively assessed using accuracy, precision, recall, and AUC metrics. The empirical findings indicate that gradient boosting and random forest models have excellent ability to capture complex patterns associated with customer churn [4]. Latheef et al. addressed the issue of predicting bank customer churn by proposing the use of the LSTM model for prediction. They also employed the SMOTE technique to preprocess the data to address the data imbalance problem. The evaluation indicated that the proposed churn prediction system achieved an accuracy of 88%, markedly outperforming the system without SMOTE implementation [5].

This paper, based on multiple machine learning models, investigated the problem of predicting bank customer churn, with the goal of systematically evaluating the predictive capabilities of various models and determining the principal factors affecting customer churn, thereby providing a basis for banks to formulate precise customer retention strategies.

2. Methodology

2.1. Data source

This research employed the publicly available 'Bank Customer Churn' dataset provided on Kaggle. This dataset contains a total of 10,000 sample records of customers and their 14 features [6].

2.2. Data preprocessing and feature engineering

Before modeling, the data underwent systematic preprocessing and feature engineering to build a high-quality customer churn prediction model. The main steps include outlier handling, feature construction, data standardization, hierarchical data partitioning, and SMOTE oversampling. Specifically, the validity of data on elderly customers (over 80 years old) was verified and age segmentation features were created.

Categorical variables were labeled (gender) or one-hot coded (geographic location), and eight derivative features were added based on exploratory data analysis to enhance the ability to capture behavioral patterns. After removing irrelevant features, 17 core features were retained. Numerical variables were scaled to zero mean and unit variance, and the dataset was partitioned into training and test subsets in an 80:20 ratio using stratified sampling. For the approximately 20.4% of churned customers, SMOTE was applied for oversampling to alleviate class imbalance and provide a balanced data foundation for model training.

DOI: 10.54254/2754-1169/2026.NJ30790

2.3. Model building

2.3.1. Logistic Regression model

This study used Logistic Regression (LR) as the baseline model and constructed an L2-regularized LR model on data processed with feature engineering and SMOTE, employing the "liblinear" solver for parameter estimation. The model coefficients and their odds ratios were then analyzed to identify key influencing factors. The results showed that a single product, the number of products, and age were associated with a higher churn probability, while customer activity, the French region, and older age groups exhibited lower churn risk.

2.3.2. Random Forest model

Random Forest (RF) introduces randomness at the sample and feature levels by ensembled decision trees and outputting results through majority voting, thereby reducing overfitting and improving generalization ability. Model parameters were optimized using RandomizedSearchCV combined with 5-fold cross-validation, with the F1 score as the primary evaluation metric.

The hyperparameter search space included the number of estimators, tree depth limits, minimum samples required for splitting and for leaf nodes, the strategy for feature subsampling, and the use of bootstrap sampling. After tuning, the model achieved an F1 score of 0.8783 in cross-validation, demonstrating high stability and predictive performance. Finally, the model was trained using the optimal parameters, and an assessment of feature importance was performed, revealing that age, customer activity, and estimated salary were the key influencing factors.

2.3.3. XGBoost model

XGBoost, as a high-efficiency gradient boosting framework, continuously improves model prediction performance by iteratively building decision trees to correct residuals. Once the model is trained, it is subsequently applied to generate predictions on the test set. Simultaneously, feature importance analysis is performed to identify key influencing factors. The results show that customer activity, number of products, and holding multiple products are the most significant influencing factors.

3. Results

The three models were rigorously assessed under a unified test set for performance comparison. The results showed that each model had good customer churn identification ability, but there were significant differences in discriminant ability and classification performance (Table 1). In terms of overall ranking ability, the XGBoost model performed best, achieving an AUC-ROC of 0.8652, which reflects its superior capability in separating churners from non-churners. The RF model's AUC-ROC is 0.8522, indicating a stable overall performance; the AUC-ROC of the LR model is 0.8431, which is slightly lower than that of the tree model but still remains at a good level. In terms of binary classification performance, RF maintains a relative balance between precision and recall (Precision = 0.5865, Recall = 0.6413, F1-score = 0.6127), indicating that the model performed moderately well in identifying real churned customers and controlling false positives.

XGBoost exhibits high precision and moderate recall (Precision = 0.6790, Recall = 0.5872, F1-score = 0.6298), indicating that it is better able to identify high-risk customers with a higher confidence level, but misses some real churned customers. In contrast, LR exhibits a clear

DOI: 10.54254/2754-1169/2026.NJ30790

characteristic of "high recall, low precision" (Precision = 0.4715, Recall = 0.7322, F1-score = 0.5736), which can capture more than 70% of real churned customers, but has a high false positive rate among customers predicted as churned. Overall, XGBoost is superior in terms of overall discrimination ability and accuracy, RF performs well in terms of the balance between precision and recall, while LR has a clear advantage in recall.

Table 1. Model performance comparison

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.7785	0.4715	0.7322	0.5736	0.8431
Random Forest	0.8350	0.5865	0.6413	0.6127	0.8522
XGBoost	0.8595	0.6790	0.5872	0.6298	0.8652

Based on bank customer churn data, this study constructs and compares three machine learning models: LR, RF, and XGBoost. The results show that the XGBoost model achieves the best overall performance (F1 = 0.6298, AUC = 0.8652).

4. Discussion

This study conducts a systematic comparison of the performance of LR, RF, and XGBoost in customer churn prediction, revealing the significant differences in their underlying mechanisms when solving the same business problem. We will delve into the reasons behind the model performance, explore the optimization potential, and reflect on the limitations and future directions of this study.

4.1. Model advantages and disadvantages analysis

LR model exhibits high recall and low precision, mainly due to their linear decision boundary and limited nonlinear fitting ability [7]. To improve coverage of churned customers, the model tends to classify more samples as churned, leading to an increase in the false positive rate, thus exhibiting high recall and low precision. However, its significant advantage lies in its good interpretability. The positive and negative impacts of features on churn risk can be intuitively identified through the direction of the coefficients, which is of great reference value for formulating preliminary intervention strategies for a wide range of customers [8].

RF possesses high variance and low bias, enabling it to effectively learn complex patterns in data. However, random forests are, to some extent, black-box models, and their predictions are difficult to interpret linearly like those of LR. While the overall contribution of a variable can be assessed by calculating feature importance, the specific direction and extent of its influence cannot be determined [9]. Furthermore, its "compromise" performance in precision and recall may indicate that its appeal will decrease in business scenarios where extreme optimization of a certain metric is required, such as pursuing the highest precision to minimize retention costs.

In this study, XGBoost achieved the best overall discriminative ability (AUC=0.8652) and the highest accuracy (0.6790). Its superior performance is mainly attributed to the residual iteration correction brought about by the gradient boosting mechanism, the efficient optimization process based on the second derivative, and the effective control of model complexity by the regularization term [10]. Higher accuracy enables the model to identify high-risk customers more reliably, which helps to achieve more cost-effective marketing resource allocation. However, this performance improvement comes at the cost of increased model complexity, higher training costs, and decreased

DOI: 10.54254/2754-1169/2026.NJ30790

interpretability. Even with the help of methods such as SHAP, its interpretability is still significantly higher than that of the logistic regression model [11].

4.2. Model optimization

4.2.1. Logistic Regression model

To improve the performance of LR, we can start from two aspects: feature engineering and class imbalance handling. Given LR's dependence on the linear assumption, we can enhance its ability to characterize nonlinear patterns by introducing multinomial features, interaction terms, or domain-based derived variables [12]. At the same time, cost-sensitive learning can be used for imbalanced data. By adjusting the class weights and increasing the penalty for misclassification of the minority class, we can make the model more effective in identifying churned customers. Prior studies have likewise demonstrated that these strategies are crucial for enhancing model performance under imbalanced data conditions [13].

4.2.2. Random Forest model

Strategies such as Bayesian optimization are employed to fine-tune key parameters, including those affecting model complexity, decision boundary smoothness, and the number of trees, in order to fully explore the model's performance potential [14]. Furthermore, it is crucial to find and set the optimal classification decision threshold that meets specific business needs. Specifically, increasing the threshold helps improve precision, while decreasing the threshold can effectively improve recall, thereby achieving a precise alignment between model performance and business objectives. Adjusting the classification threshold is widely considered one of the core algorithmic strategies for solving the class imbalance problem [15].

4.2.3. XGBoost model

Given the large number of hyperparameters in XGBoost, refined hyperparameter search is key to unlocking its performance. In addition to general parameters, the coordinated adjustment of the learning rate and max_depth should be a primary focus. Practice shows that a smaller learning rate combined with more trees usually yields better performance, but this requires a trade-off in computational cost. In this process, automated hyperparameter tuning tools such as Bayesian optimization can more efficiently explore high-dimensional hyperparameter spaces and find globally optimal or near-optimal solutions [16].

4.3. Future research directions

Future work may explore advanced ensemble strategies, such as stacking, to integrate the complementary strengths of LR, RF, and XGBoost. As a heterogeneous ensemble framework, stacking has been shown to systematically combine diverse base learners, often achieving predictive performance superior to any single model [17].

5. Conclusion

This study constructs a machine learning-based framework for predicting bank customer churn and systematically evaluates the performance of three models: logistic regression, random forest, and

DOI: 10.54254/2754-1169/2026.NJ30790

XGBoost. Among them, XGBoost demonstrated the best overall prediction performance, achieving the highest accuracy and AUC-ROC value. Feature importance analysis shows that customer activity level, number of products, holding multiple products, and age have a significant impact on churn. Among them, low activity level, fewer products, and older age are associated with higher churn risk, which is consistent with banking business practices. These findings can provide a basis for banks to develop more targeted customer retention strategies.

Although the XGBoost model performed best in this study, it still has some limitations. On the one hand, there is still room for improvement in the recall rate, which may result in some high-risk customers not being identified promptly. On the other hand, it relies on structured static features, making it difficult to capture dynamic changes in customer behavior. Furthermore, the study is based on a single public dataset, and the model's generalization ability still needs further verification.

Future research can be deepened in the following directions: First, we introduce richer temporal and behavioral features and explore LSTM and Transformer to improve the ability to characterize dynamic patterns. Second, further optimize the category imbalance handling mechanism. Third, interpretability methods such as SHAP and LIME are incorporated to enhance model transparency. Fourth, consider using heterogeneous ensemble learning strategies such as Stacking to effectively fuse complementary information from models such as LR, RF, and XGBoost, thereby achieving better prediction performance than a single model.

References

[1] Ye, J., Cheng, Z., & Lin, S. (2005). Telecommunications customer churn prediction analysis based on Bayesian networks. Computer Engineering and Applications, (14), 212–214.
[2] Ma, Y. (2021). Research on a bank customer churn early-warning model based on LightGBM (Master's thesis). Huazhong University of Science and Technology.
[3] Zhang, X. (2021). Research on commercial bank user churn prediction model based on XGBoost (Master's thesis). Dongbei University of Finance and Economics.
[4] Balaji, G., Gowtham, N. A., Tarun, S., Prajapati, G., & Manikandan, N. (2024). Customer churn prediction using machine learning algorithms. In 2024 International Conference on Emerging Research in Computational Science (ICERCS) (pp. 1–6).
[5] Latheef, J., & Vineetha, S. (2021). LSTM model to predict customer churn in the banking sector with SMOTE data preprocessing. In 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS) (pp. 86–90).
[6] Kollipara, R. (2022). Bank customer churn [Data set]. Kaggle. https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn
[7] Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations Newsletter, 6(1), 20–29.
[8] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer.
[9] Molnar, C. (2022). Interpretable machine learning: A guide for making black boxes explainable. https://christophm.github.io/interpretable-ml-book/
[10] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.
[11] Richard, G. T., & Giri, S. (2019). Digital and physical fabrication as multimodal learning: Understanding youth computational thinking when making integrated systems through bidirectionally responsive design. ACM Transactions on Computing Education, 19(3), Article 17.
[12] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
[13] Tanha, J., Abdi, Y., Samadi, N., et al. (2020). Boosting methods for multi-class imbalanced data classification: An experimental review. Journal of Big Data, 7, 70.
[14] Xue, Y., Jing, X., Ye, Q., Du, K., & Li, B. (2025). Remote sensing monitoring of wheat stripe rust using constrained random forest and Bayesian optimization. Remote Sensing Technology and Application, 40(1), 69.

DOI: 10.54254/2754-1169/2026.NJ30790

[15] Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. Progress in Artificial Intelligence, 5, 221–232.
[16] Musthofa, A. R., Noersasongko, E., & Purwanto. (2024). Predicting national rice production using the XGBoost method with hyperparameter optimization. In 2024 International Seminar on Application for Technology of Information and Communication (iSemantic) (pp. 207–211).
[17] Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115, 105151.