Proceedings of ICEMGD 2025 Symposium: Innovating in Management and Economic Development
DOI: 10.54254/2754-1169/2025.LH23930

Bank Customer Churn Prediction Based on Stacking Model

Ruixuan Li

Institute of Shenzhen Audencia Financial Technology, Shenzhen University, Shenzhen, China
liruixuan.li@audencia.com

Abstract: With the increasingly fierce competition in the financial industry, customer churn prediction has become a key research topic. Accurate prediction of which customers are more likely to churn can help banks take timely retention measures to reduce business losses. This paper adopts a data-driven approach and uses the public bank customer churn dataset to deeply analyze the distribution of data characteristics and deal with the problem of data imbalance, and proposes a customer churn prediction method based on stacked ensemble model. In this study, random forest, XGBoost, CatBoost and LightGBM were used as the basic model, and XGBoost was used as the meta-learner to establish a two-layer stacked ensemble framework. Compared with the traditional single model and simple ensemble methods, the experimental results show that the proposed method is significantly ahead in Accuracy, Recall, AUC, F1-score and other indicators, which verifies its advanced and precise capabilities in customer churn prediction.

Keywords: Churn prediction, Stacking model, XGBoost, Data analytics, Imbalanced data.

1. Introduction

In a highly competitive banking market, customer churn, where customers stop using banking services or switch to a competitor, directly leads to fewer customers and lower revenue. The cost of retaining an existing customer is much lower than acquiring a new one, so the ability to effectively predict and retain customers who are about to leave is crucial for the banking industry.

In the past, traditional models such as logistic regression and decision tree were mostly used for customer churn prediction [1]. These methods are simple to implement, but they are difficult to capture complex nonlinear relationships and have shortcomings in the processing of imbalanced data, so the prediction accuracy is limited. In order to improve the prediction performance, ensemble learning methods have been introduced into this field, such as random forest (Bagging) and XGBoost (Boosting), which improve the robustness and accuracy of the model to a certain extent by fusing the results of multiple base models [2]. However, a single ensemble model is still limited by the limitations of its base learner and cannot give full play to the complementary advantages of different algorithms.

To solve the above problems, Stacking provides a feasible solution [3]. In this method, a meta-learner is trained to integrate the prediction results of multiple base learners to make full use of the advantages of different algorithms to reduce the error. Practice shows that compared with single model, stacked model can achieve higher accuracy in some financial prediction tasks. Therefore, applying Stacking to customer churn prediction is expected to further improve the model performance.

DOI: 10.54254/2754-1169/2025.LH23930

In addition, the churn data of bank customers generally has class imbalance, that is, churn samples are far less than non-churn samples [4]. This will bias the model training towards the majority class and weaken its ability to identify the minority class (churn). Common methods for dealing with imbalance include undersampling, oversampling and synthetic samples. As a combination of oversampling and undersampling technology, SMOTETomek achieves data balance by generating minority samples and deleting overlapping samples, and significantly improves the recognition effect of the model for minority classes [5].

In addition, customer churn is affected by many factors, and the feature dimension is high and there is redundancy. Using all features directly may increase model complexity and introduce noise, which is not conducive to model generalization. Thus feature selection is needed to extract key factors. Random forests provide a built-in feature importance evaluation that can be used to select the features that contribute the most to churn prediction, eliminating redundant information, simplifying the model, and improving prediction performance [6].

In response to the above challenges, this paper proposes a bank customer churn prediction model that integrates multiple techniques. Firstly, a stacked ensemble architecture was constructed, with multiple differentiated classification algorithms as the base learner, and XGBoost was used as the meta-learner to fuse the prediction results of each base model, so as to improve the overall prediction accuracy. Secondly, SMOTETomek was introduced to resample the training data to balance the proportion of classes and enhance the detection ability of the model for minority class loss customers. Finally, random forest was used to analyze the importance of features and select important features, eliminate invalid information, and maintain the simplicity and robustness of the model.

In summary, the model constructed in this paper fuses multiple models by stacked ensemble learning, SMOTETomek balanced and imbalanced data, and random forest screening features, which effectively makes up for the shortcomings of traditional prediction methods. This method not only improves the prediction accuracy and recall rate, but also provides a new technical means for bank customer relationship management, which has important theoretical significance and application value.

2. Method

Stacking is a high-level ensemble method that combines multiple models to improve prediction accuracy [7]. Unlike voting methods, stacked models learn how to best fuse the outputs of multiple base learners in the first layer by training a single meta-learner (the second-layer model) [7].

In this study, a two-layer stacked model is constructed: the first layer contains four models XGBoost, CatBoost, LightGBM and random forest as the base learners, and the second layer similarly selects XGBoost as the meta-learner. We use XGBoost as a metamodel because its tree model performs well in dealing with high-dimensional and nonlinear features, and it effectively reduces the risk of overfitting through mechanisms such as regularization and first stopping. In order to avoid information leakage, we set 5-fold cross validation (cv=5) in StackingClassifier, so that the output of each base learner will be used to train the meta-model when it predicts the "unknown" part of the data. In this way, the metamodel will only see features predicted by the base learner when this example was not used in its training, which ensures that no information is "passed through" during training. After cross-validation, all base learners are fit again on the full training set (StackingClassifier does this by default), and finally the XGBoost metamodel combines the predictions of all base models to make a decision. Thanks to this idea of multi-model fusion, stacked models can often achieve better generalization performance than single models in most scenarios. In this study, to highlight the overall gain of the stacking strategy, we adopt relatively stable default parameter Settings for each base model and metamodel, and focus on the accuracy improvement after fusion.

DOI: 10.54254/2754-1169/2025.LH23930

3. Data analysis and feature distribution research

3.1. Data analysis

In this study, the basic properties of this data set are carefully counted and analyzed. By visualizing variables such as customer gender, credit card type, education, customer status, and income category, as shown in Figure 1.

The bar chart of customer churn broken down by gender (Figure 1) shows that existing customers account for the majority of both female and male customers, but the total number of female customers is slightly higher than that of male customers. The results suggest that although gender itself is not the main factor affecting customer churn, it can still be used as an auxiliary information in customer segmentation to further explore the interaction effect of gender on churn risk combined with other variables (such as consumption behavior and credit usage), as shown in Figure 1.

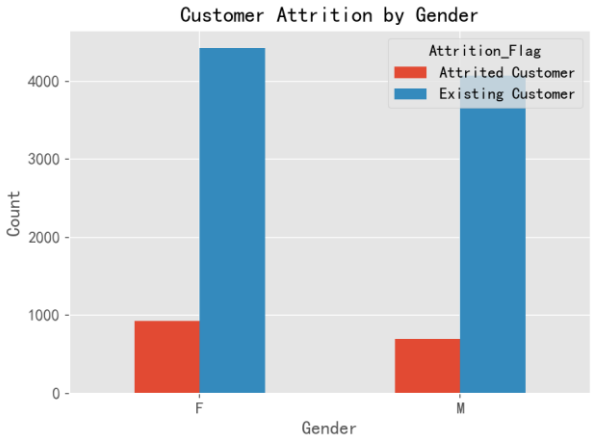

Figure 1: Customer attrition by gender

The distribution of credit card types reveals that the vast majority of customers use the basic Blue card, while the percentage of customers with advanced cards such as Silver, Gold, and Platinum is low. Such distribution characteristics show that low-end credit card users occupy a dominant position in the bank customer group. Therefore, when formulating customer retention strategies, it is necessary to focus on the differences in consumption and service experience between different types of cards, especially how to retain high-end card customers by improving service quality or increasing value-added services, as shown in Figure 2.

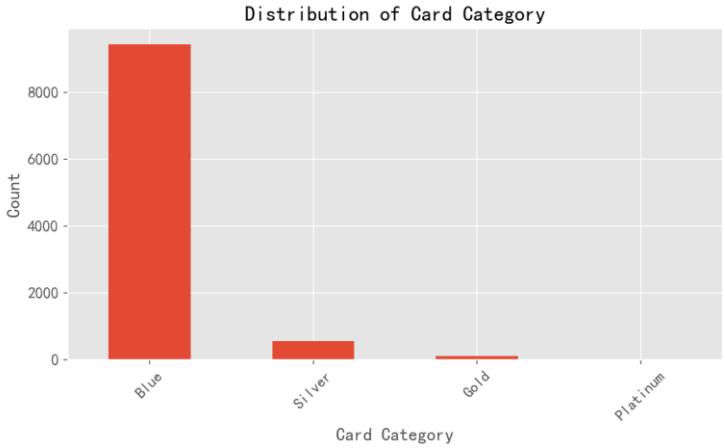

Figure 2: Distribution of card category

DOI: 10.54254/2754-1169/2025.LH23930

Breaking down the chart by gender shows that Blue cards are the dominant card type for both men and women, but there are slight differences in the percentage of mid - and high-end cards. This information provides a basis for banks to carry out precision marketing: for example, customized offers or exclusive services can be launched for middle and high-end card users, so as to improve customer viscosity, as shown in Figure 3.

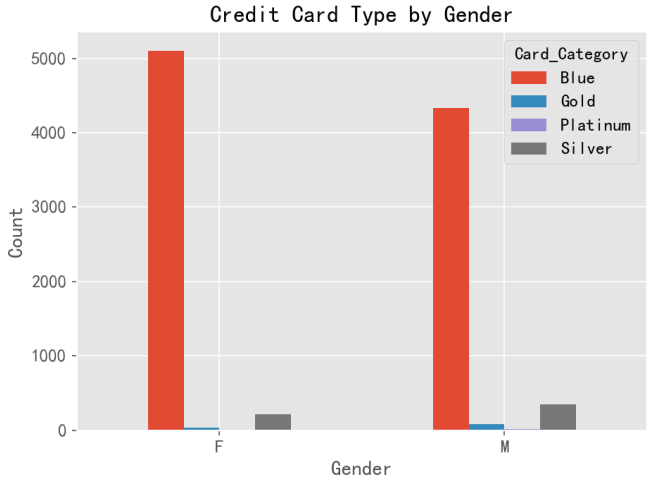

Figure 3: Credit card type by gender

By analyzing the chart of average credit utilization rate divided by education level, it is found that the distribution of customers with different education levels in credit limit use is relatively balanced, and the overall difference is not significant, indicating that the influence of education factor on credit use behavior is limited. Therefore, in the process of customer segmentation, the educational indicators can be used as auxiliary variables, and should not be used as the main division basis, as shown in Figure 4.

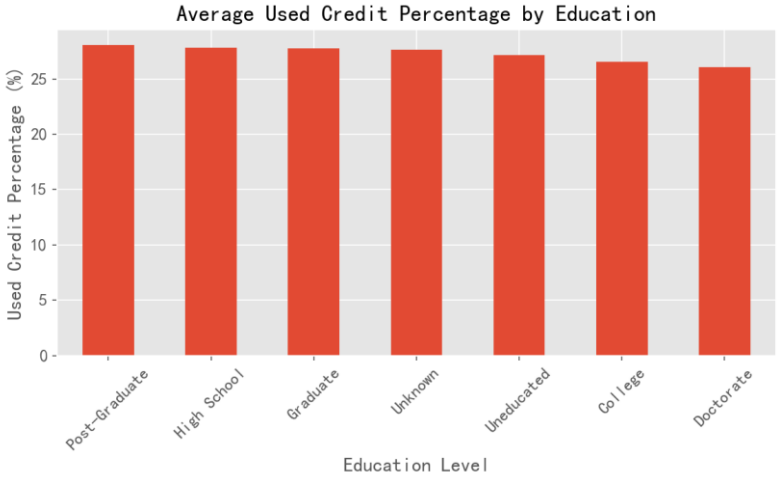

Figure 4: Average used credit percentage by education

In the chart of average credit utilization by customer status, the average credit utilization of existing customers is significantly higher than that of lost customers, which indicates that customers with higher dependence on credit cards tend to maintain long-term cooperation and may enjoy higher service stickiness. However, customers with low credit utilization have higher churn risk. This

DOI: 10.54254/2754-1169/2025.LH23930

provides an important identification indicator for banks in customer risk early warning and fine management, as shown in Figure 5.

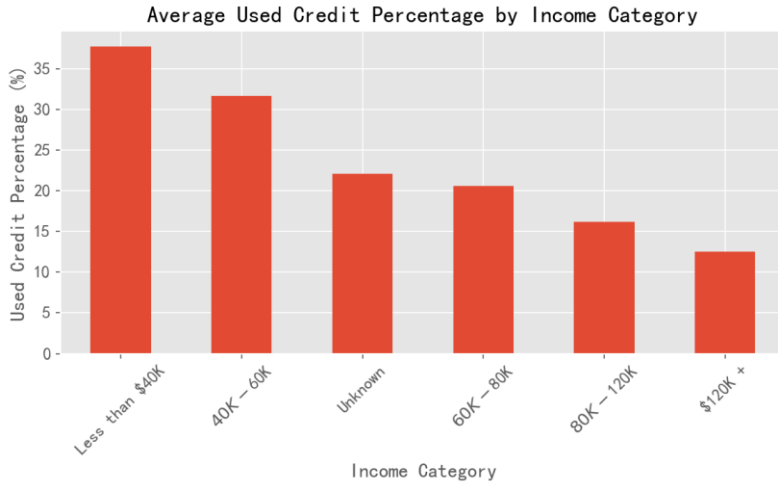

Figure 5: Average used credit percentage by income category

3.2. Feature selection

In data preprocessing, the main customer information features in the original data set are retained, and the feature selection is carried out by combining domain knowledge and quantitative analysis. In this study, highly correlated features were first eliminated to avoid multicollinearity. For example, the feature "average available amount" (Avg_Open_To_Buy) exhibits a Pearson correlation coefficient of 0.996 with "credit limit" (Credit_Limit).AVg_open_to_buy actually duplicates the credit limit information, so only Credit_Limit is retained. Avg_Open_To_Buy is removed. For the remaining features, the predictive ability is comprehensively evaluated by calculating the correlation with the target, chi-square test and random forest importance. The results show that, Customer demographics (such as age Customer_Age, number of Dependent_count), account history (Months_on_book, number of relationships with the bank Total_Relationship_Count), and customer activity (Mo number of inactive months in the past year nths_Inactive_12_mon, Contacts_Count_12_mon), consumption behavior (credit limit Credit_Limit, revolving balance Total_Revolving_Bal, total transaction amount Total_Trans_Amt, total number of transactions Total_Trans_Ct), behavioral changes (quarterly transaction amount change rate Total_Amt_Chng_Q4_Q1, quarterly transaction number change rate Total_Ct_Chng_Q4_Q1), and account utilization (Avg_Utilization_Ratio) are high for churn The explanatory power. Therefore, the above 14 features were selected to construct the prediction model in this study. These features cover basic customer attributes, account usage and behavior change trends, which can help to comprehensively characterize the possibility of customer churn. In relative terms, categorical variables such as gender, marital status, education level, income category, etc., did not show significant discriminative power in the preliminary analysis and thus were given low weights in the model training. However, a small amount of class information is still retained in the ensemble model to ensure that the model does not miss potential signals. In general, the filtered feature set not only ensures the information richness, but also avoids redundancy and noise, which is beneficial to the training effect and generalization ability of the model. as shown in Figure 6.

DOI: 10.54254/2754-1169/2025.LH23930

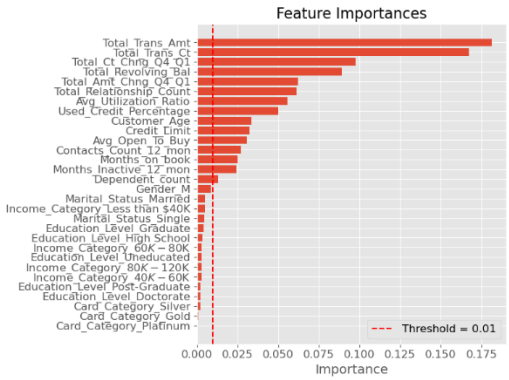

Figure 6: Feature importances

3.3. Handling imbalanced data

Customer churn data usually has class imbalance problem. In this dataset, the positive class (churn customers) is only about 16.1%, and the remaining 83.9% are non-churn customers. A model trained directly without processing may tend to output the majority class (no churn), thus achieving ostensibly high accuracy but neglecting churn identification. To solve this problem, this study used SMOTETomek method to resample and balance the training data in the model training stage [8]. SMOTETomek is a hybrid sampling method researchgate.net which combines SMOTE (Synthetic Minority Over-sampling Technique) with Tomek link under-sampling. Specifically, this study first uses SMOTE to synthesize new samples according to the existing minority samples, so as to increase the number of samples of lost customers (minority class). Then, the Tomek Links method is applied to detect the overlapping boundary sample pairs between the majority class and the minority class, and the majority class samples are removed to purify the decision boundary. Through the processing of SMOTETomek, the ratio of lost and non-lost samples in the training set is close to 1:1, and the imbalance phenomenon is significantly alleviated. This method helps the model to learn the minority class pattern more fully and improve the identification ability of churn customers. According to literature reports, using SMOTETomek for imbalanced data can effectively improve the performance of classification models [8]. In this experiment, only the above sampling processing is applied to the training set, while the original distribution of the validation/test set is maintained to check the generalization ability of the model to real imbalanced data in the evaluation phase. Compared with simple under-sampling or over-sampling, SMOTETomek makes full use of the minority class information and reduces the noise samples, which is especially beneficial to improve the robustness and recall rate of the model, as shown in Figure 7, Figure 8, and Figure 9.

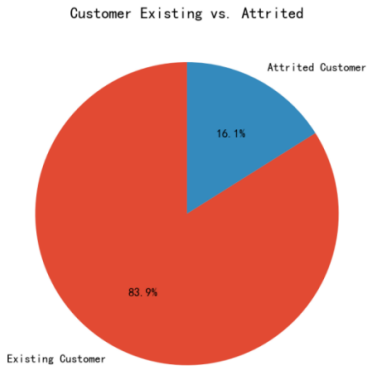

Figure 7: Customer existing vs attrited

DOI: 10.54254/2754-1169/2025.LH23930

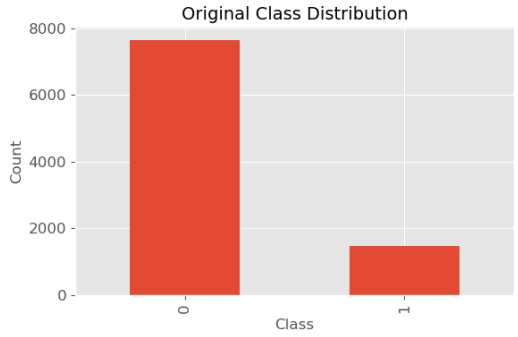

Figure 8: Original class distribution

Figure 9: Balanced class distribution

3.4. Correlation analysis

The plot reflects the Pearson correlation coefficient between the continuous variables by color intensity, and the values range from −1-1−1 to 111. The closer a color is to a warm color (yellow or red), the stronger the positive correlation; The closer a color is to a cool color (blue or green), the more negative or weak the correlation. It can be observed in the heat map that some variables related to transaction behavior (such as total transaction amount and total transaction number) show high positive correlation, indicating that customers with higher transaction amount tend to have more transactions in this data set. At the same time, features such as credit limit and available limit are also highly redundant, which need to be carefully handled or dimensioned in feature engineering. Overall, this plot provides an intuitive basis for understanding the underlying relationships and degree of redundancy among the variables, which will help in the subsequent screening of the most valuable features for churn prediction. as shown in Figure 10.

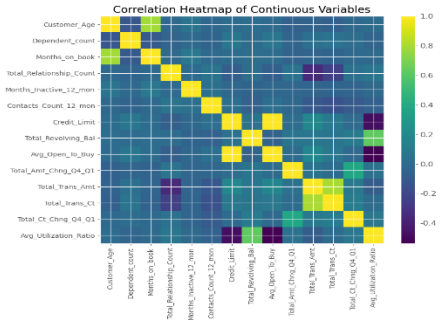

Figure 10: Correlation heatmap of continuous variables

DOI: 10.54254/2754-1169/2025.LH23930

4. Results

After the above process, we compared the stacked model with logistic regression, random forest, XGBoost, and voting ensemble models on the processed dataset, and evaluated their performance on an independent test set. The main evaluation metrics include Accuracy, area under the receiver operating characteristic curve (AUC), F1-score, etc. The accuracy reflects the overall prediction accuracy, AUC measures the ability of the model to distinguish between positive and negative classes, and F1-score considers both precision and recall, which is more suitable for evaluating the grasp of the positive class (lost customers) under imbalanced data. Table 1 summarizes the performance metrics of each model:

Table 1: Performance metrics of each model

Model	precision	recall	F1-score
Logistic regression	0.8857	0.8470	0.8592
Random forest	0.9555	0.9556	0.9555
XGBoost	0.9690	0.9694	0.9690
Voting ensemble model	0.9690	0.9694	0.9691
Stack model	0.9732	0.9733	0.9732

As can be seen in Figure 11, different models achieve different degrees of effect on the test set. The ROC curve of Logistic Regression tended to the bottom left corner, which was significantly backward in AUC (0.9092) compared with other models. This is confirmed by the classification metric (0.8592 F1-score). This partly illustrates the limitations of linear models in characterizing complex churn patterns. Compared with Random Forest, XGBoost, CatBoost, LightGBM and other tree models show better discrimination ability in the ROC space, and the AUC of XGBoost has reached 0.9916. The AUC of random forest is 0.9846, and the voting ensemble model (" soft voting "strategy, including XGB, CatBoost, LightGBM and random forest) combines the advantages of each model while achieving an AUC of 0.9915, which is very close to the performance of XGBoost. Its corresponding F1-score (0.9691) is also slightly higher than that of single XGBoost (0.9690), indicating that simple vote fusion can bring a small performance improvement. Stacking ensemble further combines the features of the multi-base model and the information of the output layer, and integrates the prediction results of each base model through the meta-learner, achieving the best performance of this experiment: the accuracy is about 0.9753, the AUC is about 0.9929, and the F1-score is 0.9732, which is better than other models in all indicators. This shows that the stacked ensemble effectively combines the strengths of each base model through the meta-learner, and significantly improves the prediction accuracy and recall ability of churn customers. as shown in Figure 11.

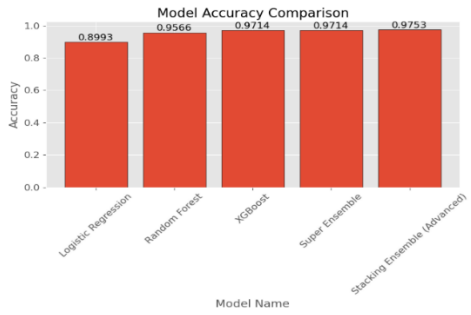

Figure 11: Model accuracy comparison

DOI: 10.54254/2754-1169/2025.LH23930

Bar graph comparing F1-score of different models. The bar chart visually shows the F1-score of each model on the test set (the numerical values are labeled at the top of the bar). As you can see, logistic regression has the lowest F1-score, random Forest and XGBoost are medium, voting ensemble improves, and stacked model has the highest column, indicating that it performs best in balancing precision and recall. as shown in Figure 12.

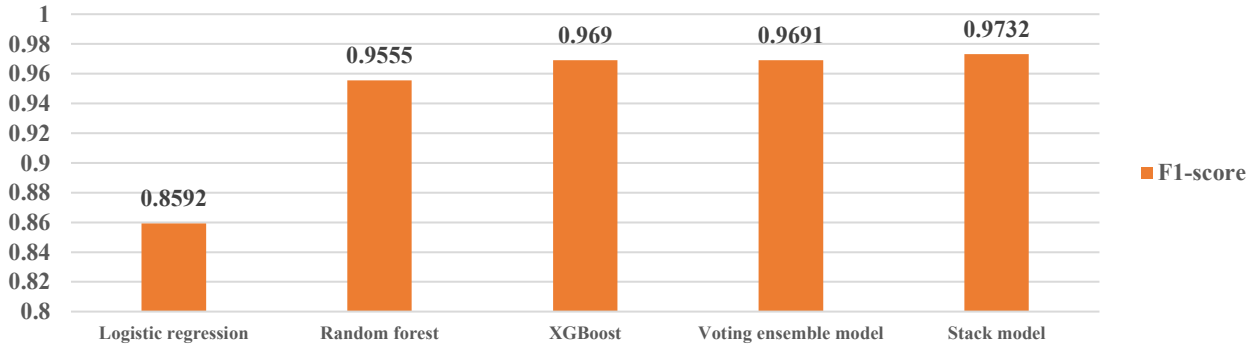

Figure 12: F1-score

Comparison of ROC curves for different models: The figure 12 plots ROC curves for Logistic regression (LR), Random forest (RF), XGBoost, voting ensemble, and stacked models, along with a diagonal reference for random guessing. The AUC values of each model are annotated below the curve. From the ROC curve, it can be seen that the curve of logistic regression (red) is close to the diagonal, and the True Positive Rate (TPR) climbs slowly in the low false positive rate region, with an AUC of about 0.9092, and the effect is limited. The curves for random forest and XGBoost (blue, purple) are clearly raised to the upper left, and TPR is higher for almost the entire range of false positive rates, especially when the false positive rate is below 0.2, both TPR exceeds 0.8, indicating that the model can identify most churn customers with low false positive rates. They all have an AUC of over 0.98, which is an excellent performance. The ROC curve of the voting ensemble model (pink) almost coincides with XGBoost, but extends slightly closer to the top left under some thresholds, showing comparable or slightly better classification performance. The stacked model has the most prominent ROC curve (blue), which can maintain higher TPR at all thresholds, and the overall curve is almost above the other models, with the largest area under the diagonal (AUC $\approx$ 0.9929). This further proves that the stacking model has achieved better classification results than other models under each threshold, and can more comprehensively separate lost and non-churn customers. Combining the histogram and ROC curve analysis, it was determined that the stacked model performed best in this study. as shown in Figure 13.

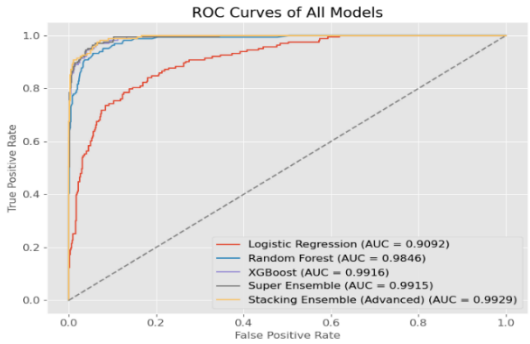

Figure 13: ROC curves of all models

DOI: 10.54254/2754-1169/2025.LH23930

5. Conclusion

In summary, this paper conducts customer churn prediction research based on BankChurners dataset. A multi-model ensemble scheme including logistic regression, random forest and XGBoost is adopted, and the imbalanced data is processed by SMOTETomek, which achieves high prediction performance on the test set. In particular, the two-layer stacked model shows the best results, which is significantly better than other comparison models in accuracy, AUC, F1 and other indicators. Experiments show that stacked ensemble can effectively fuse the advantages of different models and improve the recognition accuracy and recall rate of churn customers. This has important value for real business, helping financial institutions detect potential churn customers earlier and more accurately and take intervention measures. Of course, the performance of the model depends on the quality of the data and feature selection, and it can be further improved by introducing richer features and more advanced models in the future. In general, customer churn prediction based on stacking model provides an effective idea for solving such classification problems, and is expected to be popularized and applied in the field of customer relationship management.

References

[1] *Xu, X., & Xia, Y. (2021). Research on customer churn prediction model based on probability calibration. Statistics and Applications, 10(4), 634–641.*

[2] *Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Applied Sciences, 2(7), 1308.*

[3] *Odegua, R. (2019, March). An empirical study of ensemble techniques (bagging, boosting and stacking). In Proc. conf.: deep learn. indabaXAt.*

[4] *Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., ... & Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. Ieee Access, 4, 7940-7957.*

[5] *Wang, Z. H. E., Wu, C., Zheng, K., Niu, X., & Wang, X. (2019). SMOTETomek-based resampling for personality recognition. IEEE access, 7, 129678-129689.*

[6] *Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic.*

[7] *Aliyev, A., & Hasanova, R. (2024). Improving churn prediction in the banking sector using stacked generalization. Communications in Applied Information Technology, 12(1), 1–12.*

[8] *Li, S., & Shen, Z. (2024). Explainable customer churn prediction model based on deep learning. In Proceedings of the 3rd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2024) (pp. 282–287). Association for Computing Machinery.*