RESEARCH ON BANK CUSTOMER CHURN PREDICTION MODEL BASED ON ENSEMBLE LEARNING ALGORITHM

Shang Xinping, Wang Yi

Artificial Intelligence, Dongguan City University, Dongguan, Guangdong, China

ABSTRACT

With the rise of Internet finance, the competition of banking industry is becoming increasingly fierce. To gain more accurate and comprehensive insight into customer needs and improve customer loyalty, it is essential to establish a customer churn analysis model. This kind of model can help banks identify customers who are about to lose, facilitate business decisions, retain relevant users, and ensure that bank interests are not affected. Under this background, this paper establishes a customer churn prediction model using ensemble learning algorithm. Experimental data show that the model can effectively predict and analyse the loss of bank customers.

KEYWORDS

customer churn; data preprocessing; XGBoost

1. INTRODUCTION

With the continuous opening of the financial market and the increasing competition, the loss of bank customers has become an important factor affecting the operational efficiency of banks and has always been one of the hot issues that enterprises pay attention to. [1]. To effectively reduce customer loss, improve customer satisfaction and loyalty, strengthen customer classification management, attract potential customer groups, and improve service quality, banks need to use advanced forecasting models to identify potential lost customers and improve the core competitiveness of enterprises [2]. Churn, also known as churn or churn, is the phenomenon of a customer ceasing a business relationship with a company or service provider. In the banking industry, customer churn is a critical issue as it directly affects revenue, profitability and market share. As financial markets continue to open up and competition intensifies, banks are under intense pressure to retain existing customers and attract new ones.As a powerful machine learning technology, ensemble learning algorithm can improve the accuracy and stability of the whole model by combining the prediction results of multiple learners, so it has a broad application prospect in the customer churn prediction of banks.

Over the years, customer churn prediction has evolved from traditional statistical models to advanced machine learning algorithms. Earlier studies relied heavily on statistical methods such as logistic regression, survival analysis, and decision trees. These models provide a basic understanding of customer churn but are often oversimplified and have limited predictive power. With the advent of machine learning, researchers began experimenting with algorithms such as random forests, support vector machines (SVMS), and neural networks. These algorithms

David C. Wyld et al. (Eds): CAIML, ICAIT, ITCSE, CRYPIS, NLCA – 2024
pp. 117-130, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.141809

provide improved prediction accuracy but can still be prone to overfitting or instability. Ensemble learning, which combines predictions from multiple learners, has become a powerful way to predict customer churn. Algorithms such as gradient enhancer (GBM), Random Forest with feature selection, and Stacking ensembles have shown significant improvements in accuracy and stability. Among them, XGBoost (Extreme Gradient Boosting) is popular for its efficiency, scalability, and flexibility.

This paper takes the data of a bank as the research object. The dataset contains 14 variables and 10,000 samples. Firstly, the feature data is analysed and pre-processed, including data cleaning, feature construction and selection; Machine learning is then used to integrate XGBoost algorithms to predict and simulate customer churn. Through the establishment of a prediction model, the customer turnover rate can be effectively predicted, user activity can be improved, the effect of customer retention and care can be achieved, and the cost of retaining and caring can be reduced.

2. CONSTRUCTION OF BANK CUSTOMER CHURN FORECASTING MODEL

2.1. Data Exploration and Preprocessing

At this stage, the data for each indicator (feature) needs to be systematically cleaned and transformed to improve the performance of subsequent predictive models. This process involves several key steps:

Step 1: Check and deal with missing values

Start with a comprehensive look at missing values in the data. The presence of missing values weakens the predictive power of the model. For different types of data (such as numeric or subtyped), different interpolation strategies can be adopted, such as using the mean, median or mode to fill the numeric missing values, and for classified data, the most common category may be selected as the fill. If the proportion of missing values of a feature is too high (such as more than 50%), the feature may lose value because it contains too much unknown information and should be considered directly deleted.

Step 2: Identify and process duplicate values

There may be duplicate records in the data set due to entry errors or inadvertent data merging. These duplicates distort the true distribution of the data, which in turn affects the accuracy of the modelling results. Therefore, it is necessary to use appropriate functions or methods to identify and remove these duplicates to ensure the uniqueness and accuracy of the data set.

Step 3: Remove irrelevant or low-variance variables

Some features may not be directly related to the predicted target or exhibit very low variance (such as all values being the same), which is not only beneficial for model training, but may increase the computational burden. These irrelevant or low-information features should be identified and deleted by means of correlation analysis or variance detection to improve the efficiency and accuracy of the model.

Step 4: Detect and process outliers

Outliers, or outliers, can significantly distort the parameter estimates and predictions of statistical models. Statistical methods such as interquartile intervals (IQR) can be used in conjunction with visual tools such as box plots to identify outliers. Once an outlier is identified, it can be selected to delete or perform appropriate transformations (such as logarithmic transformations, box splitting, etc.) according to the actual situation to reduce its impact on the model. In some cases, it is also necessary to standardize or normalize the data to eliminate the impact of dimensional differences between different features.

Step 5: Balance the data set

The problem of data imbalance (i.e. the number of samples in some categories is much larger than others) is a common challenge in classification tasks. In the attrition pie chart shown in Figure 1, you can see that the ratio of non-attrition users to attrition users is close to 4:1, and this imbalance may cause the model to favour predicting most classes (i.e., non-attrition users). To solve this problem, techniques such as oversampling (e.g. SMOTE) or under sampling can be used to adjust the proportion of samples of various categories in the dataset to achieve a balance.

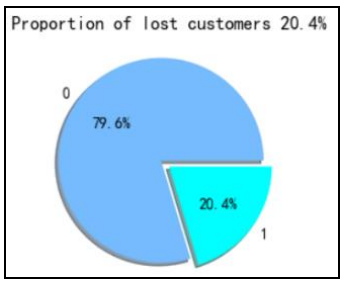

Figure 1. Pie Chart of Loss Rate

Further analysis of the relationship between the target variable and other variables:

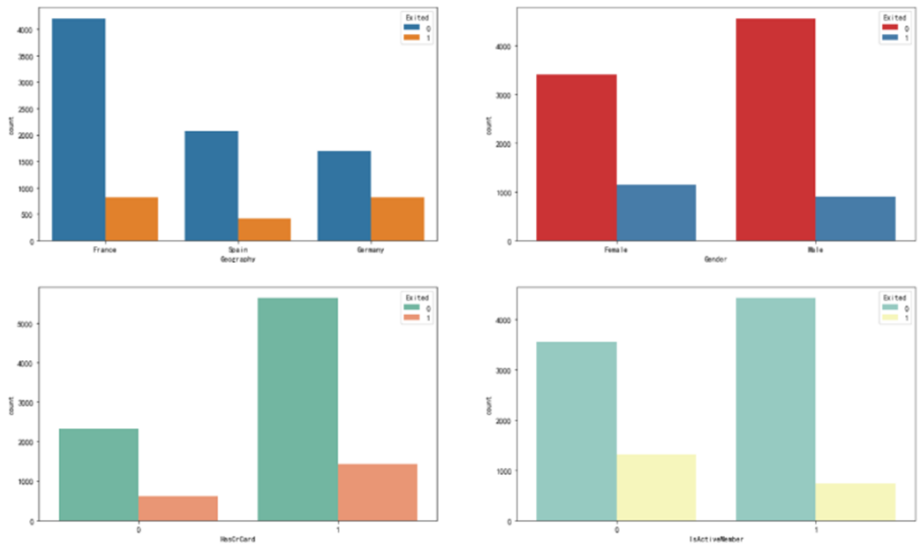

Figure 2. Diagram1 of the Relationship Between the Target Variable and Other Variables

The following questions can be seen in Figure 2:

1) Germany has the fewest customers and France the most, but the proportion of lost customers is reversed. This indicates that banks may not allocate enough customer service resources in areas with fewer customers.
2) The total number of male customers is higher than that of female customers, but the turnover ratio is lower than that of female customers, indicating that the bank's service strategy is not comprehensive enough.
3) Customers with credit cards churn more than customers without credit cards.
4) Inactive customers have a higher churn rate. However, the overall proportion of inactive customers is quite high, so banks should give relatively preferential policies to inactive customers and turn inactive customers into active customers to reduce the loss of customers.

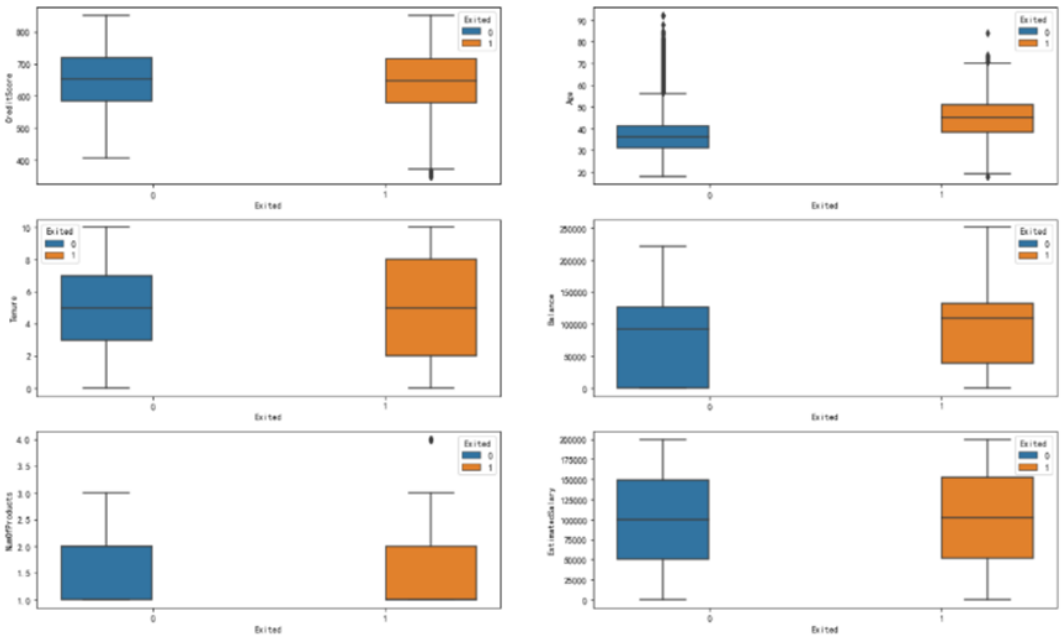

Figure 3. Diagram2 of the Relationship Between the Target Variable and Other Variables

In Figure 3, it can be seen that:

1) There is no significant difference in the distribution of credit scores between churn and non-churn customers.
2) Older customers churn more than younger ones, so banks need to adjust retention strategies for customers of different age groups.
3) In terms of tenure, clients at the extremes are more likely to churn.
4) The bank is losing customers with large bank balances, the bank may lack of loan funds, and the profit margin will be compressed.
5) Product and salary have no significant effect on the likelihood of churn.

Step 6: Data conversion and normalization

Features at different scales have different effects on machine learning algorithms. To ensure that all features contribute equally to model performance, features need to be normalized or normalized. This usually involves scaling the data to a small specific interval (such as [0,1]) or subjecting it to a standard normal distribution (mean 0, variance 1). This transformation helps to improve the convergence speed and prediction accuracy of the model.

After completing the cleaning process, the dataset should be checked again to ensure that all missing values, duplicates, and outliers have been handled appropriately. The final dataset should be structured in a way that is ready for model training, with balanced classes and relevant features selected.

2.2. Feature Construction and Selection

After data preprocessing, it is common practice to observe correlations between feature variables using the correlation coefficient matrix. By displaying the correlation coefficient matrix in the form of a heatmap, you can intuitively see the intensity of the correlation between each characteristic variable.

As can be seen from the heat map shown in Figure 4 below, the correlation between feature variables is weak, which means that these features can be considered relatively independent when building the model, so all features can be included in the model building process.

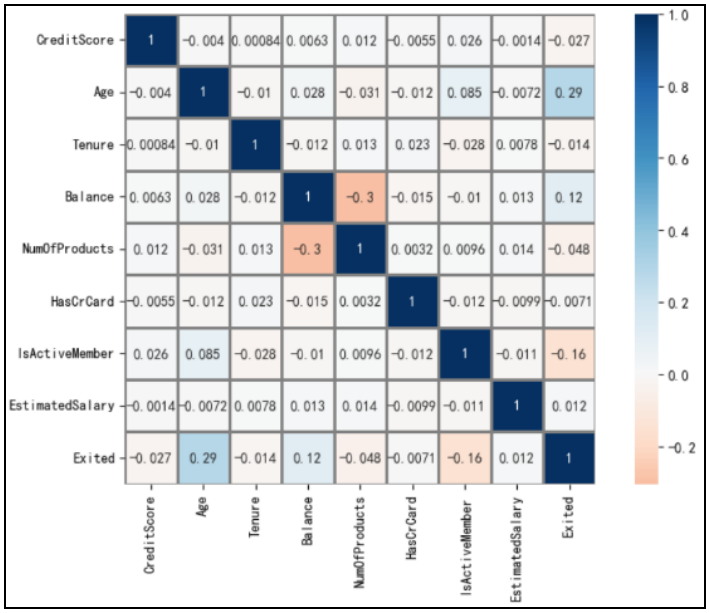

Figure 4. Relationships Among Features

According to Pearson correlation coefficient [3], further analysis of the degree of correlation between customer churn and each dimension is shown in Figure 5, from which we can see that age characteristics have the greatest impact on customer churn; The impact of different geographies is also different. The loss rate of users in Germany is significantly higher than that in other countries. In terms of gender, the loss rate of women is higher than that of men. The loss rate of active users is significantly lower than that of inactive users, which also indicates that active customers have higher loyalty than inactive customers.

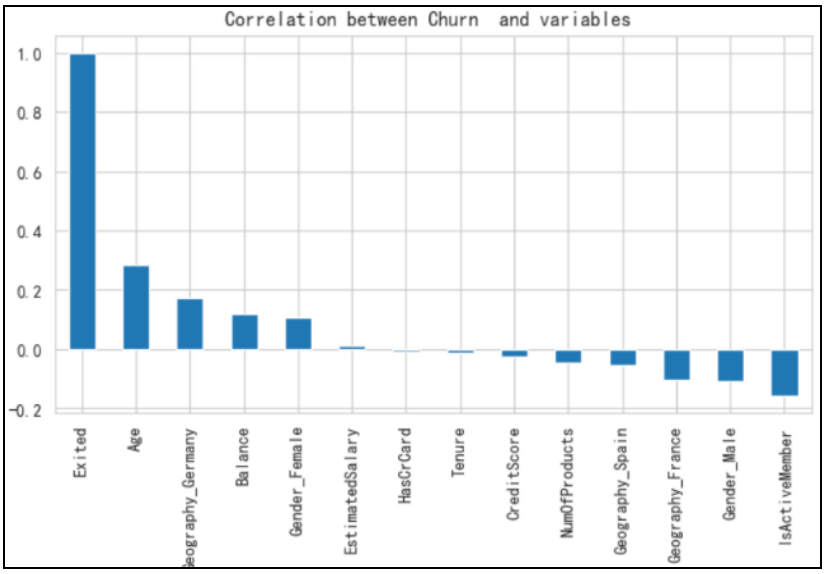

Figure 5. Relationship Between User Churn and Various Dimensions

However, a feature construction and selection process are still needed to optimize model performance.

Feature construction is to generate new features from the original data in a certain way, which can better express the intrinsic characteristics of the data and improve the predictive performance or interpretation of the model. The process of feature construction requires a deep understanding of the business context and data analysis objectives, combined with domain knowledge to create manually. Based on the new variables computed from the original features, multiple classes of the categorical variables are combined into fewer classes to reduce the dimension and complexity of the features, and the interactions between two or more features are considered to generate new interaction features that may contain additional information to help improve the performance of the model.

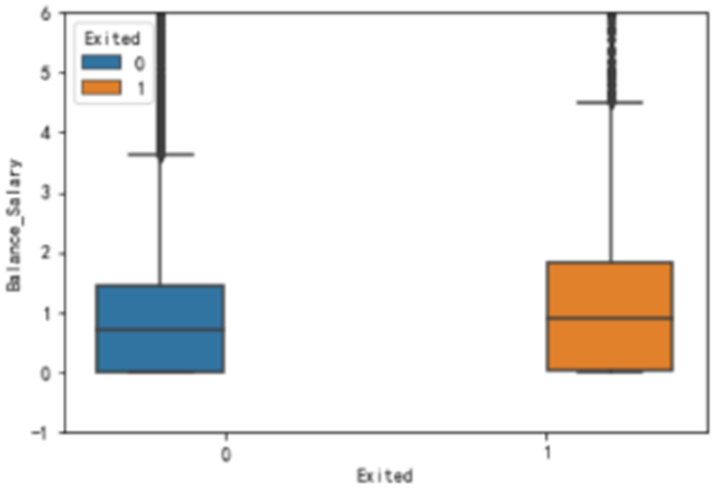

Figure 6. The Influence of Balance to Wage Ratio on Attrition Rate

As Figure 6 shows, while estimated wages do not have much effect on attrition, the balance to wage ratio does influence attrition rates. Customers with higher balance-wage ratios have a greater churn rate, which may discourage bank lending.

Feature selection refers to selecting the feature subset from the original feature set that is most useful for predicting the target before building the model. Feature selection can reduce the complexity of the model, improve the generalization ability of the model, and reduce the risk of overfitting.

Feature construction enriches the data set by generating new features, while feature selection simplifies the model by eliminating redundant or unimportant features. These two steps complement each other to improve the predictive power and interpretability of the model.

2.3. Model Construction and Evaluation

Ensemble learning algorithms are a class of powerful machine learning frameworks that make final predictive decisions by combining the predictions of multiple base learners (usually decision trees, neural networks, etc.). The core idea of this approach is "brainstorming," the idea that a combination of multiple models often has better generalization and greater accuracy than a single model. In the area of customer churn prediction, integrated learning algorithms, especially XGBoost, are favored for their excellent performance and flexibility. This article uses the integrated learning algorithm XGBoost to model customer churn prediction, as shown in Figure 7 below.

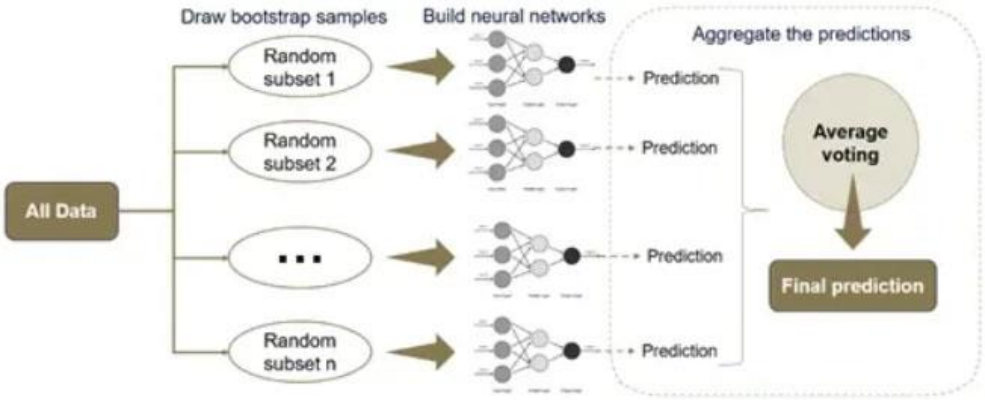

Figure 7. Integrated Learning Neural Networks

XGBoost (Extreme Gradient Boosting) is an efficient and flexible gradient boosting library for tasks such as classification, regression and sorting [4]. It is based on the gradient lifting framework and optimizes the performance of the entire model by iteratively adding weak learners (usually decision trees). Compared to traditional gradient lift algorithms, XGBoost is improved and optimized in several ways, including:

- Second order Taylor expansion of the loss function: XGBoost takes into account not only the first derivative of the loss function (i.e., the gradient), but also the second derivative (i.e., the Hessian matrix) at each iteration, which allows it to approximate the optimal solution more precisely, resulting in faster convergence and improved model accuracy.
- Regularization terms: In order to control the complexity of the model and prevent overfitting, XGBoost adds regularization terms to the objective function, including the

number of leaf nodes in the tree, L1 and L2 norms of leaf node weights, etc. These regularization terms help to improve the stability and generalization ability of the model.

- Parallel and distributed computing: XGBoost supports column sampling and parallel computing for efficient processing of large data sets. At the same time, it also supports distributed computing, making it possible to train models on large-scale distributed systems.

In customer churn forecasting, XGBoost improves model accuracy and stability by:

- Feature Importance Assessment: XGBoost automatically assesses the importance of features to help identify which factors have the greatest impact on customer churn. This helps the business team to better understand customer behavior and thus develop more targeted marketing strategies.
- Automatic processing of missing values: XGBoost automatically learns and processes missing values in training data without manual preprocessing. This simplifies data cleaning and preprocessing and reduces the impact of human error on model performance.
- Preventing overfitting: By adding regularization terms and using techniques such as early stopping, XGBoost can effectively prevent models from overfitting. This ensures that the model maintains good performance on both the training and test sets.
- Efficient model training: XGBoost uses a variety of optimization strategies to speed up the model training process, including caching mechanisms, feature preordering, and parallel computing. This enables XGBoost to complete model training on large data sets in a relatively short time.
- Flexible model tuning: XGBoost provides rich parameter setting options that allow users to flexibly adjust the model based on specific tasks and data set characteristics. By adjusting these parameters, the user can further optimize the performance of the model.

In summary, XGBoost, as an advanced integrated learning algorithm, shows excellent performance and stability in customer churn prediction. With its efficient model training, accurate prediction results, and flexible parameter adjustment capabilities, XGBoost has become an important tool for businesses to predict customer churn.

The algorithm model is used to learn 80% of samples as training sets, and 20% of samples as test sets to verify the learning ability of the model.

To comprehensively evaluate the performance of the model, this paper uses several indexes such as accuracy rate, recall rate and F1 score of the test set data [5]. These metrics are key measures of model performance, helping to understand and evaluate the model's performance in different aspects, so as to select the most appropriate model or adjust model parameters to optimize performance.

- Precision: The proportion of samples predicted by the model to be positive that are positive. A positive class is predicted to be a positive class (TP) and a negative class is predicted to be a positive class (FP)，i.e.

$$\text{precision} = \frac{TP}{TP + FP}$$

The precision reflects the reliability of the model prediction as positive. High accuracy means that the majority of the samples predicted to be positive are indeed positive, but it can also cause the model to be too conservative and miss some samples that are actually positive.

- Recall: The proportion of all positive samples that are correctly predicted by the model to be positive, i.e.

$$Recall = \frac{TP}{TP + FN}$$

The recall rate reflects the ability of the model to find all positive samples. A high recall rate means that the model can find most samples that are actually positive classes, but it can also cause the model to incorrectly predict some negative class samples as positive classes.

- F1 Score: This is the harmonic average of accuracy and recall for a comprehensive evaluation of model performance, i.e.

$$\text{F measure} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

The F1 score is a single metric that considers both the accuracy and comprehensiveness of the model. In scenarios where both precision and recall need to be a concern, F1 scores are a good choice.

The test results are evaluated based on precision, recall, F1 score, and accuracy, as shown in the following Figure 8, all of which are above 0.85, indicating good performance.

	precision	recall	f1-score
0	0.89	0.97	0.93
1	0.83	0.52	0.64
accuracy			0.88
macro avg	0.86	0.75	0.78
weighted avg	0.87	0.88	0.87

Figure 8. Test Results

It can also observe and judge the accuracy of the learner simply and intuitively by viewing the ROC curve [6] to understand the generalization performance of the learner. The ROC curve is drawn with True Positive Rate (TPR) as the vertical coordinate and False Positive Rate (FPR) as the horizontal coordinate under different threshold Settings. TPR represents the proportion of all samples that are positive examples correctly predicted by the model. FPR represents the proportion of all samples that are negative examples that are incorrectly predicted to be positive by the model. The closer the ROC curve is to the upper left corner (i.e. high TPR and low FPR), the better the classification performance of the model.

AUC is the area under the ROC curve, which is a performance evaluation index to measure the learner's quality. The AUC value is R and its value ranges from 0 to 1. The closer the AUC value is to 1, the better the prediction performance of the model is. When the AUC value is 0.5, the model performance is no different from random guessing. If the AUC value is less than 0.5, it indicates that the direction predicted by the model is completely opposite to the reality. The AUC value provides a quantitative standard to evaluate the predictive power of the model, making the performance comparison between different models more objective. It comprehensively considers the model's performance under all classification thresholds, so it can fully reflect the model's predictive power. Like ROC curves, AUC values also have good performance for categorically unbalanced datasets.

In summary, ROC curves and AUC values are important tools for evaluating the predictive power of classification models, not only providing an intuitive graphical display, but also quantitative

ways to help researchers and developers comprehensively and objectively evaluate model performance and make more rational decisions. The ROC curve shown in Figure 9 and AUC value of 0.913 are the test results of this model, which can strongly prove that the established model has good prediction effect and is suitable for related prediction tasks.

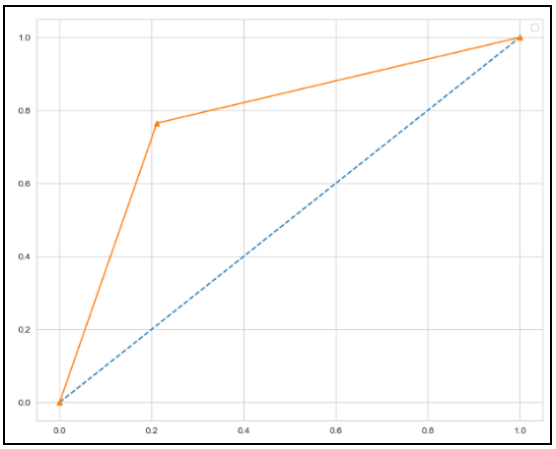

Figure 9. ROC Curve

Based on the performance feedback of the test sets, it is necessary to constantly adjust the model structure and hyperparameters, as well as try different optimization methods. Through continuous evaluation and optimization, the performance of the model can be gradually improved to better adapt to the actual application scenario [7].

The following are common methods and strategies for hyperparameter tuning and model optimization:

1）Hyperparameter Tuning Methods:

- Grid Search: Exhaustively tries all combinations of hyperparameters but can be computationally expensive.
- Random Search: Randomly samples hyperparameters, often more efficient than Grid Search for large parameter spaces.
- Bayesian Optimization: Uses historical performance to predict the next best hyperparameters, often more efficient than Grid or Random Search.
- Hyperband: Based on Successive Halving, reduces computational costs by discarding poor-performing hyperparameters early.

2）Optimization Methods:

- Learning Rate Scheduling: Adjusts the learning rate dynamically to improve convergence and performance. Methods include Step Decay, Cosine Annealing, and Warm Restarts.
- Weight Regularization: Adds L2 or L1 regularization to prevent overfitting.
- Batch Normalization: Normalizes inputs at each layer to improve training speed and generalization.
- Gradient Clipping: Limits gradient values to prevent gradient explosion.

3）Early Stopping: Prevents overfitting by stopping training early when validation performance stops improving.

4）Loss Function Selection: Choose appropriate loss functions based on the task, like Cross-Entropy Loss for classification or Mean Squared Error for regression.

5）Optimizer Selection: Adaptive optimizers like Adam, RMSprop, and Adagrad dynamically adjust learning rates, suitable for various training scenarios.

6）Model Architecture Fine-tuning:

- Activation Function: Experiment with different activation functions (e.g., ReLU, Leaky ReLU, ELU) to find the best for training speed and accuracy.

By applying these methods, model performance can be significantly improved, better fitting the needs of real-world applications.

3. RETENTION STRATEGY

When the prediction accuracy of the model is 88%, customers will still be lost, and the recall rate is 0.52. These 52% customers are lost customers, and corresponding retention strategies for these lost customers need to be formulated as follows:

1) Early identification of potential lost customers

Using churn prediction models, banks can identify churn prone customers in advance. Through the model's analysis of customer behaviour, transaction data, etc., banks can understand which customers are likely to leave in the future. Based on this, banks can prioritize actions to communicate with these customers and offer personalized services and incentives to retain them.

Implementation strategy:

- For customers at high risk of churn, banks can provide customized offers, loan interest rate adjustments, or personalized financial advice to enhance customer satisfaction and loyalty.
- Initiate proactive service plans to regularly interact with high-risk customers, understand their needs and problems, and address them in a timely manner to avoid churn.

2) Customer segmentation and targeted retention plan

Losing customers is not a homogeneous group, and different types of customers may have different reasons for losing customers. The model can help banks segment customers by analysing the characteristics of different customer groups (such as age, income, trading habits, etc.). Based on this, banks can implement more targeted retention strategies.

Implementation strategy:

- For young customer groups, banks can introduce more attractive digital products or services, such as mobile payment, smart investment, etc., to meet their needs for convenience.
- For HIGH-NET-WORTH customers, banks can provide more advanced financial services and personalized wealth management solutions to enhance customer loyalty.

3) Improve customer satisfaction and experience

Through the model prediction, the key factors affecting customer satisfaction and churn can be found. Banks can optimize service processes, product design and customer interaction based on these factors. For example, if a certain type of service (such as slow loan approval) is found to be associated with customer churn, the bank can begin to improve the relevant processes to fundamentally reduce customer dissatisfaction.

Implementation strategy:

- Optimize the customer service experience of the bank, provide a faster and more efficient service response mechanism, and ensure that customer problems are resolved in a timely manner.
- Use customer feedback data to predict customers' future needs through models to provide personalized product recommendations or customized financial services.

4) Timely intervention

Continuously track customer behaviour data, timely detect customer churn risk signals, and respond quickly. For example, if a customer has recently reduced account activity or withdrawn funds, the system can send a prompt warning, prompting the bank to intervene.

Implementation strategy:

- Implement automated customer intervention mechanisms, such as automatic notification emails when customer behaviour is abnormal and provide customized service options.
- Personalized communication through account managers to retain customers when they are about to leave.

5) Optimize marketing and cross-selling

Churn prediction models not only help banks retain customers, but also identify potential cross-selling opportunities. For example, customers with high churn risk may not be interested in certain financial products, but through model analysis, banks can find these customers' potential demand for other financial products (such as credit cards, insurance, etc.), so as to target cross-selling.

Implementation strategy:

- Provide tailor-made product or service packages for departing customers to motivate them to re-engage with banking.
- Use insights from predictive models combined with historical customer data to launch attractive promotions or offers.

6) Increase customer Lifecycle Value (CLV)

By anticipating churn, banks can better manage customer lifecycle value and maximize long-term revenue per customer. According to the retention measures suggested by the model, banks can extend the life cycle of customers and increase the total revenue of customers.

Implementation strategy:

- Use predictive models to regularly assess customer lifecycle value and implement special retention strategies for high-value customers.

- Increase customer engagement and engagement through enhanced customer loyalty programs, thereby increasing overall customer value.

By combining the results of the attrition prediction model with practical application scenarios, banks can better understand the drivers behind customer attrition and develop more accurate and efficient customer retention strategies. This not only reduces customer churn, but also increases customer satisfaction and loyalty, ultimately boosting the bank's business performance.

4. CONCLUSIONS

This article conducted descriptive statistical analysis and feature importance analysis on customer data of a certain bank and constructed a customer churn prediction model based on ensemble learning XGBoost algorithm. Through this model, we can analyse the churn situation of specific customers, identify the reasons for customer churn as soon as possible, and enable bank staff to carry out customer retention work as soon as possible, achieve precise marketing, and improve bank efficiency.

However, the data set used in this study comes from a specific bank, whose customer group may have a specific geographical, economic and cultural background, resulting in limited representation of the data set. This sample bias may affect the model's ability to generalize across other banks or different customer groups. In addition, the study used historical static data and did not include time series information about customer behaviour. Therefore, failure to capture the dynamic changing characteristics of customer behaviour may affect the accuracy of churn prediction. If other banks or new data sources provide more important features, the performance of the current model may be limited. Future research could improve the model's predictive power by integrating more external data sources. For example, combining social media data, mobile payment data, market data, etc., to enrich the customer profile and further refine the analysis of the reasons for customer churn. Time series data on customer behaviour can also be introduced to dynamically capture changes in customer behaviour patterns, which can improve the accuracy of churn forecasts and help banks to intervene in a timely manner. In customer churn forecasting, churn customers usually fall into a few categories. Future studies can address the class imbalance by adding a small number of class samples using data enhancement techniques (e.g. SMOTE, ADASYN, etc.) or by using a weighted loss function. Hierarchical sampling can also be used to ensure that the proportion of different customers in the training set is more balanced, so as to improve the accuracy of the model's churn prediction for different types of customers.

Although the XGBoost model was used in this study, the introduction of deep learning methods could be explored in the future, particularly suitable for processing time series data and complex customer behavior data. You can also try to introduce graph neural networks that use customer relationship networks to predict churn risk. It is possible to explore how to enhance the interpretability of the model so that the banking staff can more clearly understand the decision-making process of the model. To provide guidance for subsequent research and further promote the application and development of customer churn prediction model.

ACKNOWLEDGEMENTS

Here, I would like to express my heartfelt thanks to all the members who participated in this study. In the academic discussions, your insights and suggestions have inspired me and broadened my horizons. He gave me great help and support in data collection, experimental design and paper revision. Without your cooperation and support, this study would not be able to proceed smoothly. I would like to express my sincere thanks to all the people who have given me help and support.

REFERENCES

[1] Shi Danlei, Du Baojun. Prediction of bank customer churn based on BP neural network [J]. Science and Technology Innovation, 2021 (27): 104-106.
[2] Zhao J. Research on key technologies of bank customer analysis management based on data mining [D]. Zhejiang University,2005.
[3] Shi Yixuan Research on bank customer churn prediction based on data mining [D]. Inner Mongolia University, 2022.
[4] Fu Lei Bank customer churn early warning and model interpretability analysis [D]. Huazhong Agricultural University, 2022.
[5] Zhang Wen, Zhang Lili. Prediction and analysis of bank customer churn based on GA-SVM [J]. Computer and Digital Engineering, 2010,38 (04): 55-58.
[6] Chen Chenli. The bank customer churn model based on data mining technology research [D]. North China institute of aerospace industry, 2023. The DOI: 10.27836 /, dc nki. GBHHT. 2023.000085.
[7] Xie Bin Bank N customer churn analysis and marketing strategy research based on big data mining [D]. Zhejiang University of Technology, 2020.

AUTHOR

Shang Xinping, master, research direction "Artificial intelligence and machine learning", working in the School of Artificial Intelligence, Dongguan City University, full-time teacher. Currently studying at St. Paul University Philippines, Doctor in Information technology, has published several high-quality research papers.