RESEARCH ON INTEGRATED LEARNING ALGORITHM MODEL OF BANK CUSTOMER CHURN PREDICTION

Shang Xinping , Wang Yi

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

ABSTRACT

With the rapid growth of Internet finance, competition within the banking industry has intensified significantly. To better understand customer needs and enhance customer loyalty, it has become crucial to develop a customer churn prediction model. Such a model enables banks to identify customers at risk of leaving, support data-driven business decisions, and implement strategies to retain valuable clients, thereby safeguarding the bank's interests. In this context, this paper presents a customer churn prediction model based on an ensemble learning algorithm. Experimental results demonstrate that the model effectively predicts and analyzes potential customer churn, providing valuable insights for retention efforts.

KEYWORDS

customer churn; data preprocessing; XGBoost

1. INTRODUCTION

With the ongoing expansion of financial markets and heightened competition, customer churn has emerged as a critical factor influencing banks' operational efficiency and remains a top concern for businesses [1]. To effectively reduce customer attrition, enhance satisfaction and loyalty, refine customer segmentation, attract potential clients, and improve service quality, banks must leverage advanced prediction models to identify customers at risk of leaving and boost their competitive advantage [2]. Customer churn, also known as customer attrition, refers to the process where a customer ends their relationship with a company or service provider. In the banking sector, this issue is especially significant, as it directly impacts revenue, profitability, and market share. As market competition intensifies, banks are under growing pressure to retain existing customers and attract new ones.

Ensemble learning, a robust machine learning approach, offers a solution by improving the overall accuracy and stability of predictive models through the combination of multiple algorithms, making it ideal for customer churn prediction in banks.

Historically, customer churn prediction has progressed from traditional statistical methods to more advanced machine learning techniques. Early models, such as logistic regression, survival analysis, and decision trees, provided a basic understanding of churn but were often limited in their predictive capabilities. As machine learning evolved, more sophisticated algorithms like random forests, support vector machines (SVMs), and neural networks were introduced, offering better accuracy but often suffering from overfitting and instability issues. Ensemble learning, which integrates the predictions of multiple models, has emerged as a more reliable and accurate solution. Methods like Gradient Boosting Machines (GBM), Random Forest with feature

DOI: 10.5121/ijdms.2024.16501

selection, and Stacking have shown marked improvements in predictive performance and stability. Among these, XGBoost (Extreme Gradient Boosting) stands out for its efficiency, scalability, and flexibility.

This research focuses on a bank's dataset, consisting of 14 variables and 10,000 samples. The study begins by analysing and pre-processing the data, which includes tasks like data cleaning, feature engineering, and feature selection. XGBoost, an ensemble learning algorithm, is then utilized to predict and model customer churn. The resulting prediction model enables banks to accurately forecast customer attrition, increase user engagement, improve retention strategies, and reduce the costs associated with retaining customers.

2. CONSTRUCTION OF BANK CUSTOMER CHURN FORECASTING MODEL

2.1. Data Exploration and Preprocessing

At this stage, it is crucial to systematically clean and transform the data for each feature to enhance the performance of the predictive models. This process involves several important steps:

Step 1: Handling Missing Values

Begin by thoroughly examining the dataset for any missing values, as these can weaken the predictive power of the model. Different strategies should be employed depending on the type of data (numeric or categorical). For numerical data, methods such as filling in missing values with the mean, median, or mode can be applied. For categorical data, the most frequent category may be used as a replacement. However, if a feature has an excessive proportion of missing values (e.g., more than 50%), it may lose its significance due to the large amount of unknown information and should be considered for removal.

Step 2: Identifying and Removing Duplicate Values

Duplicate entries can occur due to data entry errors or accidental data merging, which can distort the true distribution and negatively affect the accuracy of the model. Identifying and removing these duplicates using appropriate techniques ensures the dataset’s integrity and reliability.

Step 3: Eliminating Irrelevant or Low-Variance Features

Certain features may have little relevance to the target variable or exhibit extremely low variance (e.g., nearly identical values across the dataset), adding unnecessary complexity to the model. These irrelevant or low-variance features should be identified through correlation analysis or variance testing and removed to enhance the model’s efficiency and accuracy while reducing computational overhead.

Step 4: Detecting and Handling Outliers

Outliers can distort statistical models and reduce predictive accuracy. Statistical techniques, such as the interquartile range (IQR), combined with visual tools like box plots, can be used to detect these outliers. Depending on the context, outliers can either be removed or transformed (e.g., through logarithmic transformation or binning) to minimize their impact on the model. In some cases, data standardization or normalization is necessary to address the differences in scale between features.

Step 5: Balancing the Dataset

Data imbalance, where certain classes significantly outnumber others, is a common challenge in classification tasks. In this case, the churn distribution may be imbalanced, with non-churn customers outnumbering churn customers by approximately 4:1 (as shown in Figure 1). This imbalance can cause the model to favor the majority class (non-churn customers). To address this, techniques like oversampling (e.g., SMOTE) or undersampling can be applied to adjust the class proportions, ensuring a more balanced dataset and improving model performance for minority classes.

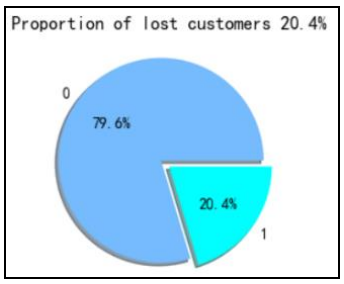

Figure 1. Pie Chart of Loss Rate

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

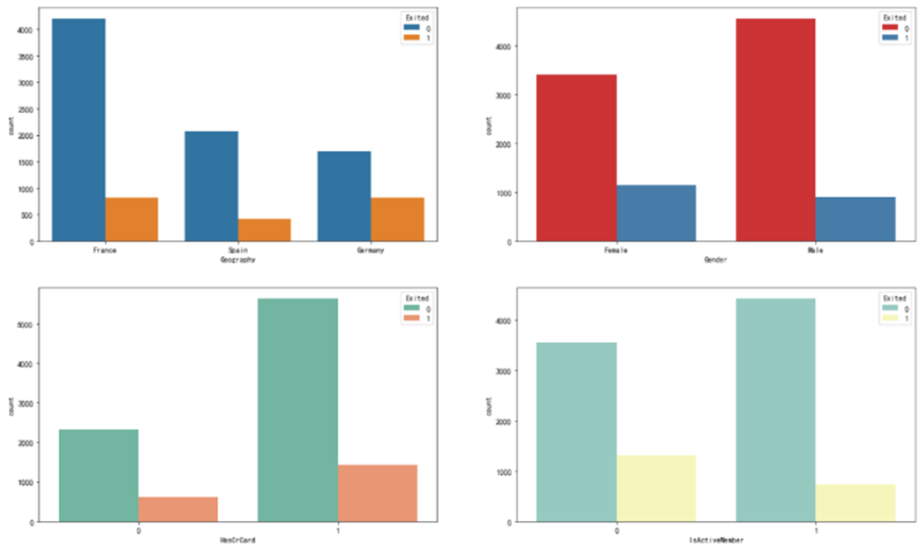

Figure 2. Diagram1 of the Relationship Between the Target Variable and Other Variables

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.

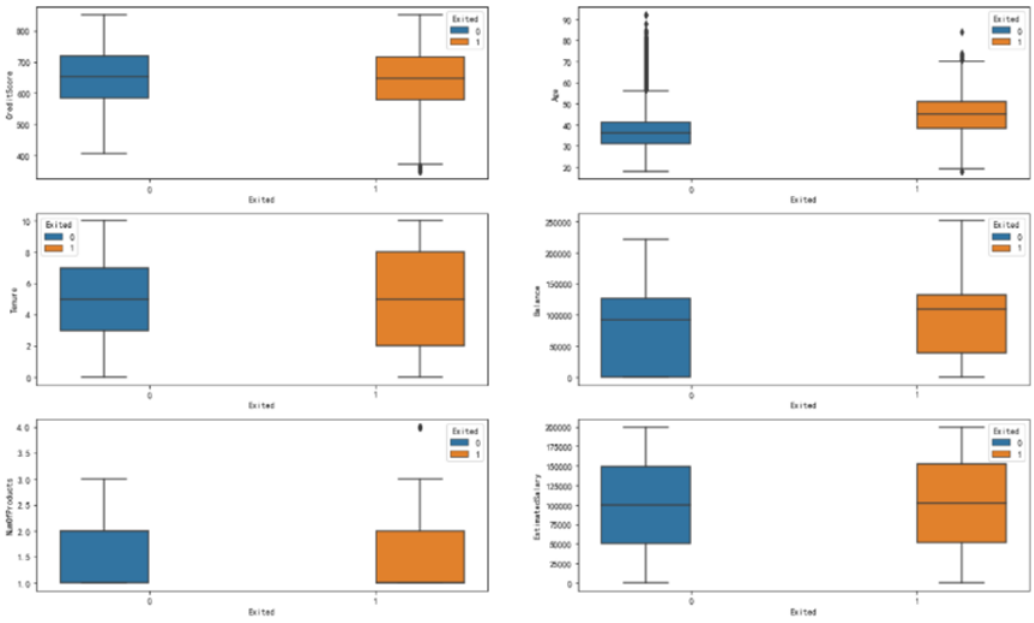

Figure 3. Diagram2 of the Relationship Between the Target Variable and Other Variables

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 Transformation and Normalization

Features with varying scales can impact machine learning algorithms differently. To ensure that all features contribute equally to the model's performance, it is essential to apply scaling through normalization or standardization. Normalization typically scales the data to a specific range (such as [0,1]), while standardization transforms the data to follow a normal distribution (mean of 0 and variance of 1). These transformations help improve the model's convergence speed and predictive accuracy.

Once the data cleaning process is complete, the dataset should be reviewed again to confirm that all missing values, duplicates, and outliers have been appropriately handled. The final dataset should be well-prepared for model training, with balanced classes and carefully selected, relevant features.

2.2. Feature Construction and Selection

After completing data preprocessing, it is standard practice to examine the correlations between feature variables using a correlation coefficient matrix. Displaying this matrix as a heatmap provides a clear, visual representation of the strength of the correlations between different features.

As shown in the heatmap in Figure 4, the correlations between the feature variables are relatively weak. This indicates that the features are largely independent of each other, making them suitable for inclusion in the model-building process without concerns about multicollinearity.

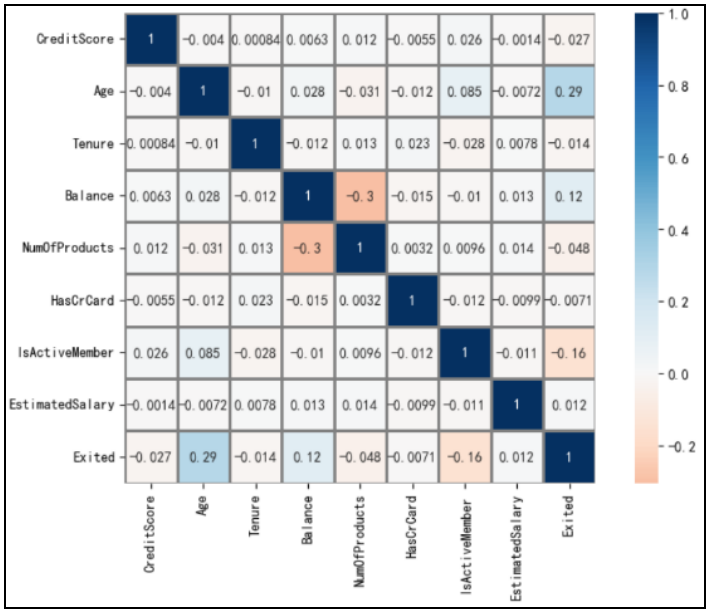

Figure 4. Relationships Among Features

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.

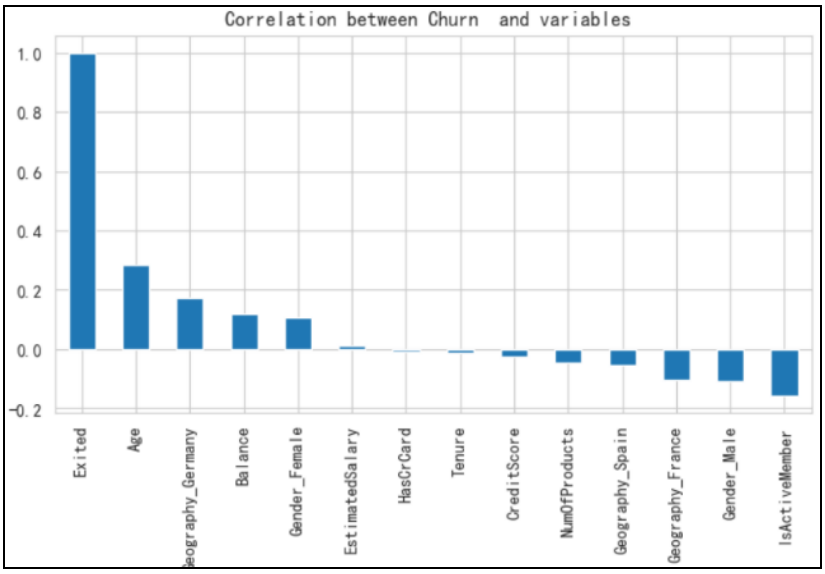

Figure 5. Relationship Between User Churn and Various Dimensions

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

Feature construction involves generating new features from the original data to better capture its intrinsic characteristics, thereby enhancing the model's predictive performance or interpretability. This process requires a deep understanding of the business context, data analysis objectives, and domain expertise to manually create relevant features. By deriving new variables from the original ones, categorical variables with multiple classes can be combined into fewer categories, reducing the dimensionality and complexity of the dataset. Additionally, interactions between two or more features can be considered to create interaction features that contain extra information, potentially improving the model's performance.

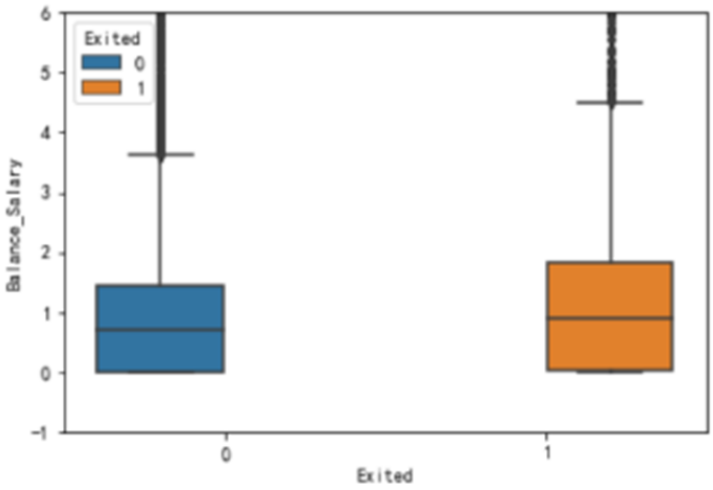

Figure 6. The Influence of Balance to Wage Ratio on Attrition Rate

As shown in Figure 6, while estimated wages have minimal impact on customer churn, the balance-to-wage ratio significantly influences churn rates. Customers with higher balance-to-wage ratios exhibit a greater likelihood of churn, which could deter banks from lending to such individuals.

Feature selection involves choosing the most relevant subset of features from the original dataset that are most effective for predicting the target variable. This process helps reduce model complexity, enhance the model's generalization ability, and lower the risk of overfitting.

While feature construction enriches the dataset by creating new features, feature selection simplifies the model by removing redundant or unimportant ones. These two processes work together to improve both the predictive accuracy and interpretability of the model.

2.3. Model Construction and Evaluation

Ensemble learning algorithms are a powerful class of machine learning frameworks that make final predictions by aggregating the outputs of multiple base learners, such as decision trees or neural networks. The core principle behind this approach is akin to "brainstorming"—the idea that combining the strengths of multiple models typically results in better generalization and higher accuracy than relying on a single model. In the realm of customer churn prediction, ensemble learning algorithms, particularly XGBoost, are highly regarded for their superior performance and flexibility. This study employs the XGBoost ensemble learning algorithm to model customer churn prediction, as illustrated in Figure 7 below.

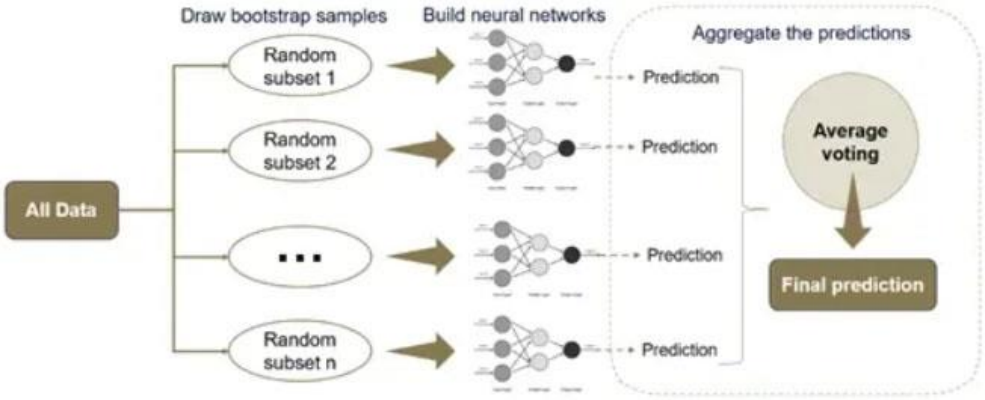

Figure 7. Integrated Learning Neural Networks

XGBoost (Extreme Gradient Boosting) is a highly efficient and flexible gradient boosting library designed for various tasks, including classification, regression, and ranking [4]. Built on the gradient boosting framework, it enhances model performance by iteratively adding weak learners, typically decision trees. XGBoost offers several improvements over traditional gradient boosting algorithms, including:

- Second-Order Taylor Expansion of the Loss Function: Unlike conventional methods that consider only the first derivative of the loss function (the gradient), XGBoost also incorporates the second derivative (the Hessian matrix) at each iteration. This enables a more precise approximation of the optimal solution, resulting in faster convergence and improved model accuracy.

- Regularization Terms: To manage model complexity and prevent overfitting, XGBoost includes regularization terms in the objective function. These terms account for the number of leaf nodes in the tree and the L1 and L2 norms of the leaf node weights, contributing to greater model stability and generalization.

- Parallel and Distributed Computing: XGBoost supports column sampling and parallel processing, making it well-suited for handling large datasets efficiently. It also facilitates distributed computing, allowing model training on large-scale systems.

In the context of customer churn prediction, XGBoost enhances model accuracy and stability through several mechanisms:

- Feature Importance Assessment: XGBoost automatically evaluates feature importance, helping to identify the factors most significantly impacting customer churn. This insight enables business teams to better understand customer behaviour and develop targeted marketing strategies.

- Automatic Handling of Missing Values: XGBoost can learn and manage missing values in the training data without the need for manual preprocessing. This streamlines the data cleaning process and minimizes the potential for human error in model performance.

- Overfitting Prevention: By incorporating regularization terms and employing techniques such as early stopping, XGBoost effectively mitigates the risk of overfitting. This ensures that the model performs well on both training and test datasets.

- Efficient Model Training: XGBoost employs various optimization strategies, including caching mechanisms, feature preordering, and parallel computing, to accelerate the model training process. This allows for quicker completion of model training on large datasets.

- Flexible Model Tuning: With a wide range of parameter settings, XGBoost allows users to adjust the model flexibly according to specific tasks and dataset characteristics. Fine-tuning these parameters can further enhance model performance.

In conclusion, XGBoost is an advanced ensemble learning algorithm that demonstrates exceptional performance and stability in predicting customer churn. Its efficient model training, accurate predictive capabilities, and flexible parameter adjustment options make it an invaluable tool for businesses aiming to forecast customer attrition.

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

The accuracy of the learner can be easily and intuitively assessed by examining the ROC curve [6], which provides insights into the model's generalization performance. The ROC curve is plotted with the True Positive Rate (TPR) on the vertical axis and the False Positive Rate (FPR) on the horizontal axis for various threshold settings. TPR indicates the proportion of actual positive samples correctly predicted by the model, while FPR represents the proportion of actual negative samples incorrectly classified as positive. A ROC curve that approaches the upper left corner of the plot (indicating high TPR and low FPR) signifies better classification performance.

The Area Under the Curve (AUC) quantifies the overall performance of the ROC curve and serves as a key metric for evaluating the learner's quality. The AUC value ranges from 0 to 1, with a value closer to 1 indicating superior predictive performance. An AUC of 0.5 suggests that the model's performance is equivalent to random guessing, while an AUC below 0.5 implies that the model's predictions are completely contrary to reality. The AUC provides a standardized metric for assessing model predictive power, allowing for objective comparisons between different models. It considers the model's performance across all classification thresholds, offering a comprehensive view of its predictive capabilities. Both ROC curves and AUC values perform well with imbalanced datasets.

In conclusion, ROC curves and AUC values are essential tools for evaluating the predictive power of classification models. They offer an intuitive graphical representation as well as a quantitative assessment, enabling researchers and developers to evaluate model performance comprehensively and objectively, thereby facilitating more informed decision-making. As illustrated in Figure 9, the model achieved an AUC value of 0.913, demonstrating strong predictive effectiveness and making it well-suited for related prediction tasks.

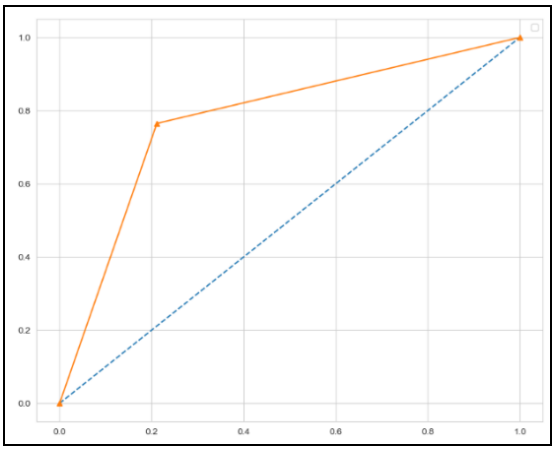

Figure 9. ROC Curve

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

Even with a prediction accuracy of 88%, banks may still lose customers, as evidenced by a recall rate of 0.52. This means that 52% of the lost customers need targeted retention strategies, which can be implemented as follows:

1) Early Identification of Potential Lost Customers
   Using churn prediction models, banks can proactively identify customers at risk of leaving by analyzing their behavior and transaction data. This allows banks to prioritize communication and offer personalized services.

   - Offer tailored incentives, such as customized loan rates or financial advice, to at-risk customers.
   - Establish proactive service plans to regularly engage high-risk customers and address their needs.

2) Customer Segmentation and Targeted Retention Plans

   Different customer segments have varying reasons for churn. By analyzing customer characteristics, banks can implement more targeted retention strategies.

   - For younger customers, introduce digital services like mobile payments to meet their convenience needs.
   - For high-net-worth individuals, offer advanced financial services and personalized wealth management.

3) Improve Customer Satisfaction and Experience

   The model can identify key factors affecting customer satisfaction and churn. Banks should optimize services and processes based on these insights.

   - Enhance customer service efficiency and response times.
   - Use feedback to predict future needs and provide personalized product recommendations.

4) Timely Intervention

Continuously monitor customer behavior to detect churn risks and respond quickly.

- Set up automated alerts for abnormal customer behavior, prompting timely interventions.
- Utilize account managers to communicate with customers showing signs of leaving.

5) Optimize Marketing and Cross-Selling

Churn prediction models can help identify opportunities for cross-selling other financial products.

- Create customized product packages for at-risk customers to encourage re-engagement.
- Launch attractive promotions based on insights from predictive models and customer data.

6) Increase Customer Lifetime Value (CLV)

By anticipating churn, banks can enhance customer lifecycle management and maximize long-term revenue.

- Regularly assess customer CLV and implement retention strategies for high-value clients.
- Boost engagement through loyalty programs to enhance overall customer value.

By integrating churn prediction results with practical strategies, banks can better understand the factors contributing to customer attrition and develop effective retention plans. This approach not only reduces churn but also enhances customer satisfaction and loyalty, ultimately improving the bank's performance.

4. CONCLUSIONS

This study analysed customer data from a specific bank using descriptive statistics and feature importance analysis and built a customer churn prediction model based on the XGBoost ensemble learning algorithm. The model helps analyse churn patterns, identify potential reasons for churn, and enables bank staff to take timely action for customer retention, leading to more precise marketing and improved efficiency [8].

However, the dataset is limited to one bank, with specific geographical, economic, and cultural factors, which may affect the model's generalizability to other banks or customer groups. Additionally, the use of static historical data without time-series information on customer behaviour may limit the model's ability to capture dynamic behaviour patterns. To improve prediction accuracy, future research could incorporate more external data sources (e.g., social media, mobile payment, and market data) and use time-series data to capture behavioural changes. Addressing class imbalance through techniques like SMOTE or weighted loss functions could also enhance model performance.

Future work could explore deep learning approaches, which are better suited for time-series and complex behavioural data or use graph neural networks to incorporate customer relationships into churn prediction. Improving model interpretability would also help bank staff better understand and apply the model's insights, supporting the ongoing development of churn prediction models in the banking sector.

ACKNOWLEDGEMENTS

I would like to extend my heartfelt gratitude to everyone who contributed to this study. Your valuable insights and suggestions during our academic discussions have greatly inspired me and expanded my perspective. Your assistance in data collection, experimental design, and paper revision has been immensely helpful. This research would not have been possible without your collaboration and support. I am deeply thankful to all who have offered their guidance and assistance throughout this process.

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.
[8] Shang Xinping, Wang Yi. Research on Bank Customer Churn Prediction Model based on Ensemble Learning Algorithm. 13th International Conference on Artificial Intelligence and Machine Learning (CAIML 2024), Toronto, Canada, 2024.

AUTHORS

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.