The causes of bank customer churn based on XGBoost and LightGBM models: the evidence from the Kaggle dataset

Jiayi Cai

Abstract:

As the digital economy continues to grow, the expansion of Internet finance introduces new challenges for the conventional banking industry. Banks must deal with multiple pressures, such as digital transformation, declining customer loyalty, and fintech competition. Analyzing the potential factors of bank customer churn from multiple perspectives and constructing models for predicting churn can help bank managers understand the causes of churn, identify problems, detect potential churn customers promptly, and develop efficient retention strategies based on customer characteristics and preferences. In this paper, we used a combination of visualization, data mining, and machine learning methods to analyze the factors used to predict bank customer churn from multiple perspectives, such as feature selection (Random Forest Feature Importance Ranking), feature extraction (PCA), visualization, etc. We also constructed two churn prediction models based on the gradient boosting tree algorithms, XGBoost and LightGBM, compared the evaluation measures before and after feature selection and before and after tuning parameters, and interpreted the model through SHAP methods. After the paper, the following conclusions were drawn: (1) Total Trans Amt, Total Trans Ct, and Total Revolving Bal are pivotal in analyzing and predicting customer churn; (2) the SHAP Summary Plot can react to the visual analysis of predictors of customer churn to a certain extent; (3) the effect of feature selection on the assessment of the results is sometimes insignificant; (4) tuning parameter settings can enhance model performance to a certain extent, but the optimal parameters may vary based on the preprocessing method employed. These conclusions will assist banks in comprehending customer churn factors more deeply, constructing a higher performance churn prediction model, and conducting a comprehensive result synthesis analysis.

Keywords: customer churn; predictors of churn; churn prediction models; machine learning

1. Introduction

The construction of the digital economy and the development of the fintech sector have significantly changed the financial industry’s business models and service methods, which has brought great challenges to the traditional banking industry. More and more FinTech and BigTech firms have arisen as suppliers of digital financial services [1], constructing FinTech platforms and offering customers a diverse range of services such as payment and remittance, lending, enterprise financial management, crowdfunding, enterprise technologies for financial institutions, trading and capital markets, insurance, personal financial management, wealth management, and digital banking [2]. This trend has accelerated disturbance in the financial sector and contributed to the democratization of financial services while strengthening the close relationship between financial service providers and customers and reducing switching costs for customers to change service providers [3]. Many consumers believe that traditional banking intermediaries are obsolete and that FinTech is the “disintermediating” retail money and finance [4]. Digital finance can affect competition within the banking sector through three intermediary mechanisms: deposit effect, loan effect, and cooperation effect. For example, the evolution of digital finance has affected commercial banks’ savings deposit and household loan operations [5]. Although some studies have shown that fintech credit and banking may be complementary [6] and that fintech credit also contributes to bank stability [7], banks may no longer maintain a competitive edge in the processing of soft credit information and relationship lending compared to information technology-advanced technology companies. Tech companies can utilize their information advantage to target the distribution of debt products and capture market share, thus affecting the profitability of the traditional banking industry [6].

Furthermore, within the blockchain domain, the potential of blockchain applications extends beyond financial institutions and payment systems. There is also a rising interest in blockchain technology [8], with millennials, particularly, showing a strong interest in fintech products and services [9]. Notably, the COVID-19 global pandemic has expedited the digital transformation of financial services, making people’s lives more virtual. Technology companies are also seizing this market opportunity and attempting to integrate their digital products into the daily lives of individuals [9]. A study has shown that during the COVID-19 global pandemic, companies like Amazon and Tescent that utilized fintech innovations demonstrated strong resilience and achieved significant revenue growth [2]. In addition, as a result of the worldwide economic recession and financial storm, the services provided by traditional banks can no longer meet their customers' needs [10]. At the same time, digital transformation and technological advances have enabled numerous technological companies to meet their customer's needs cost-effectively [9]. As a result, the switching motivation of customers will be enhanced, and they will be more proactive in seeking financial institutions that have profit guarantees or offer more cost-effective services [10]. Unprecedented disintermediation has intensified competition in the financial sector and threatened the business model of traditional banks [3]. One study suggests that 66.8% percent of existing bank customers already use or intend to use bank accounts from emerging fintech companies within the next three years [11]. In addition, the CustomerGauge 2018 NPS and CX Benchmarking Report [12] reveals a worrying reality for the banking industry: the industry performs poorly in terms of Average Net Promoter Score (NPS), which is at the lower end of the cross-industry scale, while also facing challenges of relatively low Average Retention Rate and Average Return on Retention.

However, customer loss can adversely affect banks' profitability, cost structure, brand reputation, and risk management. Customers are the most important source of profit for banks [13], and losing customers means that banks may lose their source of revenue, which reduces profitability. Meanwhile, customer retention helps improve financial performance and supports the bank's sustainability [14]. It is widely acknowledged that the expense associated with acquiring new customers surpasses the expense of retaining existing ones [15]. This is because of the necessity for banks to allocate more resources to attract new customers to substitute for the missing ones and spend more time establishing stable relationships with new customers. In addition, it has been shown that brand reputation is associated with customer inertia and customer retention [10]. Customer churn may damage a bank's brand reputation, further affecting customer retention and new customers' trust. Customer churn may also lead to a decline in the quality of the bank's loans, thus creating a higher credit risk for the bank. Therefore, in the information-intensive financial services sector, it is not enough to focus on customer acquisition; banks also need to analyze in-depth the factors that contribute to customer churn and identify trends in churn in a timely manner. Especially in a situation where customers can independently terminate their use of banking services at any time and freely choose services from different financial institutions, understanding customer churn tendencies in advance and conducting an analysis of customer churn will help banks allocate resources more efficiently to retain their customers, ensure sustainable profit [15], and achieve long-term sustainability.

The financial industry, as a highly relationship-related industry, is in dire need of improving customer brand experience (BE) as well as facilitating customer engagement (CE) [16] to increase customer inertia and loyalty and thus achieve long-term customer retention. However, to date, the measures implemented by conventional retail banks have not demonstrated adequate in motivating customers and maintaining long-term customer loyalty [17]. At this juncture, some efficient methods, such as visualization, data mining, and machine learning, play an important role. By adopting these methods, bank managers can understand the causes of customer churn and identify the problems in customer management and business areas that existed before the current situation to optimize and improve them. Moreover, compared to a comprehensive marketing approach, bank managers can use the information they already have to categorize customers, recognize individuals with favorable credit situations or future growth potential, and implement customer retention programs tailored to their characteristics and preferences [10]. This helps to meet the needs of different customer groups and enhance customer brand experience (BE), thereby increasing customer satisfaction and loyalty [16]. In addition, an effective churn prediction model can predict customer churn in a timely manner, which helps the bank's marketers take early steps to enhance customer engagement (CE) and provide a better customer experience [16]. Therefore, the purpose of this paper is to conduct a comprehensive analysis and exploration of the factors employed for predicting customer churn in banks from multiple perspectives through the application of visualization, data mining, and machine learning methods and to develop two churn prediction models utilizing the Gradient Boosting Tree algorithm, XGBoost, and LightGBM, and then make targeted recommendations after comparative evaluation. This will help banks to better understand the factors of customer churn, construct higher performance churn prediction models, and be able to perform a thorough analysis of the outcomes to optimize the use of resources and avoid significant losses due to customer churn. It will also help to improve the competitiveness of banks and enable them to achieve sustained growth in the face of fierce competition.

The innovation of this paper is the use of a comprehensive approach to probe and analyze from multiple perspectives. Next, the paper is organized as follows: Section 2 will present a literature review on methods for predicting customer churn. Section 3 will present the research methodology and analyze and preprocess the dataset. Section 4 will provide an in-depth analysis of the factors used for customer churn prediction, such as performing feature selection (random forest feature importance ranking), feature extraction (PCA), and visualization analysis. Section 5 will perform a comparative analysis of models for predicting customer churn. Finally, Section 6 will summarize the findings and make relevant recommendations.

2. Literature review of Customers Churn Prediction Methods

Customer churn prediction is an intricate task that requires processing large amounts of data and exploring multiple methods to find the best solution. Existing research on customer churn prediction methods focuses on statistical analysis methods, time series analysis, machine learning, cluster analysis, ensemble learning, and hybrid methods. The purpose of this chapter is to provide the reader with the chance to make a more informed choice of methods for real-world applications by reviewing the research and applications of different churn prediction methods.

Statistical modeling approaches were first used in the marketing and financial industry to solve churn analysis and prediction tasks [18]. For example, survival analysis that models the occurrence and timing of events. Analysis of Variance (ANOVA) is used to reveal customer behavior. T-tests and chi-square statistics were used to predict customer behavior and perceptions. Khodadadi et al. [19] presented ChOracle, an oracle that forecasts user churn by modeling user return times to a service utilizing a blend of Temporal Point Processes and Recurrent Neural Networks. And they showcase ChOracle's outstanding performance across various real datasets. On the other hand, Oskarsdóttir et al. [20] utilized a time-series approach to forecast customer churn in the telecommunications industry, utilizing a time-series that represents the dynamics of customer behavior. This approach of exploiting dynamic behavior makes sense instead of exploring previous static classification applications.

However, as data availability increases and problem complexity increases, there is a growing tendency to adopt machine-learning methods that can handle large-scale data and complex features. In machine learning, numerous algorithms exist to facilitate various tasks.

Some of the traditional single machine learning algorithms include Logistic Regression (LR), Decision Trees, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (K-NN). However, all these methods have some limitations. For example, when using Logistic Regression, the unbalanced data impacts the accuracy and precision of estimated parameters, leading to the model's poor predictive ability [21]. It has also been shown that the performance of models such as Decision Trees [22] can be affected by imbalanced datasets. In addition, machine learning algorithms, such as deep neural networks [23, 24] and cluster analysis, are usually unsuitable for directly dealing with imbalanced datasets. However, an approach has been proposed by researchers such as Carl Yang [25], who developed a coherent and robust framework, called ClusChurn, for interpretable new user clustering and churn prediction. This framework is capable of delivering real-time data analytics and prediction results, thus benefiting multiple aspects.

However, some machine learning algorithms and methods for unbalanced data also exist, such as ensemble learning, algorithm optimization, hyperparameter tuning, and hybrid methods.

Geiler et al. [18] recommend ensemble approaches, such as AdaBoost, Gradient Boosting, or XGBoost, for predicting churn. In particular, XGBoost has demonstrated superiority in handling imbalanced datasets [26]. However, some studies suggest that XGBoost can be used in combination with other ensemble methods for achieving state-of-the-art performance [27]. In addition, Khoh et al. [28] have proposed an optimized weighted soft voting ensemble learning model, shown through empirical results to have higher prediction accuracy than machine learning and deep learning models, among others, for customer churn prediction systems. Algorithmic optimization methods were also utilized by Vani Haridasan et al. [29], who employed an arithmetic optimization algorithm (AOA) in conjunction with a stacked bidirectional long short-term memory (SBLSTM) model, demonstrating potential in performance compared to recent approaches. Thakkar et al. [30] proposed a novel class-dependent cost-sensitive boosting algorithm called AdaboostWithCost that reduces errors and misclassification costs, thereby reducing the cost of churn. Arshad et al. [31] on the other hand, proposed a hybrid model called "A Hybrid System for Customer Churn Prediction and Retention Analysis via Supervised Learning (HCPRs)" which used Synthetic Minority Over-Sampling Technique (SMOTE) and Particle Swarm Optimization (PSO) to solve the problem of imbalance class data and feature selection. Extensive experiments are conducted to assess the model's performance across Random Forest (RF), Linear Regression (LR), Naïve Bayes (NB), and XGBoost. The results indicate that the proposed model has a higher accuracy under the curve (AUC) of 98% when used with the XGBoost classifier compared to other methods.

Nowadays, propelled by the advancements in machine learning, the exploration and application of customer churn prediction methods have become more and more popular. The models used to predict customer churn are becoming increasingly sophisticated, but the predicted results are becoming more accurate and reliable. However, the generalization ability of many models has yet to be verified. While the choice of machine learning models is contingent on dataset characteristics, a limited number of studies have conducted experiments across diverse churn datasets from various domains to establish the models' validity and assess their performance [15]. Therefore, in the future, there is still a need to focus on this aspect of generalization ability, to construct more advanced hybrid models with better generalization ability and novel feature engineering methods, among others.

3. Research Methodology and Data Preprocessing

The crucial steps in Customer Churn Prediction (CCP) encompass data analysis, feature extraction, identification of pivotal features influencing retention, and the reasonable selection of a classification model. Understanding the data is paramount before applying machine learning algorithms involving data cleaning and feature selection. [32]. Therefore, this study plans to use visualization, data mining, and machine learning methods to preprocess the dataset and then explore and analyze the dataset from multiple perspectives. And visualize and analyze the dataset after extracting the key predictor features of customer churn. In addition, this study plans to construct two churn prediction models based on Gradient Boosting Tree algorithms, XGBoost and LightGBM, for comparative evaluation and targeted comments.

This section will be divided into four parts: sample selection, outlier handling, descriptive statistics, and unbalanced data processing using SMOTE.

3.1 Sample Selection

The study is based on a set of data from Kaggle called Credit Card Customers. This dataset contains 10127 credit card customers, each with 23 features. Out of the 23 features, there are seven types of float-type features, ten types of int-type features, and six types of object-type features. These features can be roughly customized into two categories: the basic characteristics of the customer and the characteristics and usage of the customer's credit card.

In the basic characteristics of the customer, in addition to the CLIENTNUM (Customer Number), Attrition Flag, Age, Gender, Education Level, Marital Status, and Income Category, there is a Dependent count (The number of dependents of the customer).

The characteristics and usage of the customer's credit card include: Card Category, Months on book (Length of time the customer has held the credit card), Total Relationship Count (Total number of products held by the customers), Months Inactive 12 mon (The number of months inactive in the last 12 months), Contacts Count 12 mon (The total number of interactions or contacts made between the customer and the bank within the past 12 months), Credit Limit (The maximum amount of credit that the bank extends to a customer), Total Revolving Bal (The unpaid amount that carries off on the next credit card's cycle), Avg Open To Buy (The average amount of credit available for a customer to use or spend), Total Trans Amt (The total amount of transactions made by a customer in the last 12 months), Total Trans Ct (The total count or number of transactions made by a customer in the last 12 months), Avg Utilization Ratio (The average percentage of available credit that a customer has utilized or borrowed), Total Amt Chng Q4 Q1 (The net change in the total transaction amount during a specific period, typically comparing the fourth quarter (Q4) to the first quarter (Q1) of the same year) and Total Ct Chng Q4 Q1 (The net change in the total transaction count during a specific period, typically comparing the fourth quarter (Q4) to the first quarter (Q1) of the same year) , etc.

3.2 Outlier Handling

The Education Level, Marital Status, and Income Category features contain unknown labels. First, replace them with nulls, as they do not provide any information. Then, they were calculated based on the K-Nearest Neighbors (K-NN) values to remove all nulls. However, the analysis revealed that all three variables had very little effect on the dependent variable and could be used individually or eliminated. Exclusion was chosen in this study. In addition, this study also chose to exclude CLIENTNUM and the last two complex features, as none of these features affect customer churn.

3.3 Data Transformation

This study transforms six types of object-type features into data to facilitate subsequent machine learning. The two features, Attrition Flag and Gender, are distinguished by 0 and 1, respectively. The remaining four features, Education Level, Marital Status, Card Category, and Income Category, are all ranked from 1. The transformations are presented in Table 1.

Table 1. Data Transformation

Attrition Flag	Existing Customer=1 Attrited Customer=0
Gender	M=1 F=0
Education Level	Uneducated=1 High School=2 College=3 Graduate=4 Post-Graduate=5 Doctorate=6
Marital Status	Single=1 Married=2 Divorced=3
Income Category	Less than $40K=1 $40K - $60K=2 $60K - $80K=3 $80K - $120K=4 $120K +=5
Card Category	Blue=1 Silver=2 Gold=3 Platinum=4

3.4 Descriptive Statistics

Descriptive statistics were applied to the dataset, and the outcomes are displayed in Table 2.

Table 2. Descriptive Statistics

	Count	Mean	Std	min	50%	max
Attrition Flag	7081.0	0.842819	0.363997	0.0	1.000	1.000
Age	7081.0	46.347691	8.041225	26.0	46.000	73.000
Gender	7081.0	0.523372	0.499489	0.0	1.000	1.000
Dependent count	7081.0	2.337805	1.291649	0.0	2.000	5.000
Education Level	7081.0	3.065810	1.404962	1.0	3.000	6.000
Marital Status	7081.0	1.664031	0.619564	1.0	2.000	3.000
Income Category	7081.0	2.343313	1.355904	1.0	2.000	5.000
Card Category	7081.0	1.082757	0.328819	1.0	1.000	4.000
Months on book	7081.0	35 981359	8.002609	13.0	36.000	56.000
Credit Limit	7081.0	8492.773831	9126.072520	1438.3	4287.000	34516.000
Total Relationship Count	7081.0	3.819376	1.544444	1.0	4.000	6.000

Months Inactive 12 mon	7081.0	2.342607	0.995104	0.0	2.000	6.000
Contacts Count 12 mon	7081.0	2454456	1.104917	0.0	2.000	6.000
Total Revolving Bal	7081.0	1167.501624	812.315606	0.0	1282.000	2517.000
Avg Open To Buy	7081.0	7325.272207	9131.217585	3.0	3250.000	34516.000
Total Amt Chng Q4 Q1	7081.0	0.760584	0.223139	0.0	0.735	3.397
Total Trans Amt	7081.0	4394.299816	3468.461606	510.0	3831.000	17995.000
Total Trans Ct	7081.0	64 503319	23 809330	10.0	67.000	134.000
Total Ct Chng Q4 Q1	7081.0	0.711508	0238693	0.0	0.700	3.714
Avg Utilization Ratio	7081.0	0.282313	0.278731	0.0	0.186	0.999

3.5 Unbalanced data processing using SMOTE

When it comes to classification problems, class imbalance often occurs, which significantly impacts the model [33]. Therefore, churn datasets with class imbalance problems (CIP) must be treated appropriately before being used as input to machine learning algorithms [15]. Typically, diverse sampling methods are employed to alter the class distribution [18]. This enhances the accuracy of models and imparts greater stability to them [33]. Sampling methods can be broadly categorized into three types: oversampling (e.g. SMOTE, ADASYN), undersampling (e.g. Tomek Links, ENN), and hybrid sampling. Although it has been shown that undersampling tends to overtake oversampling [18], Tékouabou et al. [33] do not recommend an undersampling approach as it can lead to the loss of information that could have otherwise contributed to the model, and instead propose a succinct and detailed process of machine learning model construction including combining SMOTE to balance the data and employing ensemble methods, complemented by cross-validation.

3.5.1 Churn rate analysis

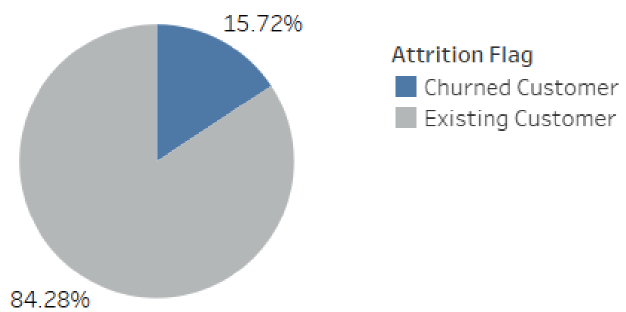

Figure 1

As shown in Figure 1, only 15.72% percent of the customers are churned, which indicates an imbalance between churned customers and existing customers. A single traditional machine learning algorithm model is unsuitable for predicting customer churn in this dataset, as it cannot precisely and comprehensively identify customers with a propensity for churn due to its limitations. Therefore, before constructing a churn prediction model, this study will balance the dataset by cross-validating combinations of SMOTE, supplemented with heat maps for visualization and analysis.

3.5.2 Correlation Heat map after SMOTE

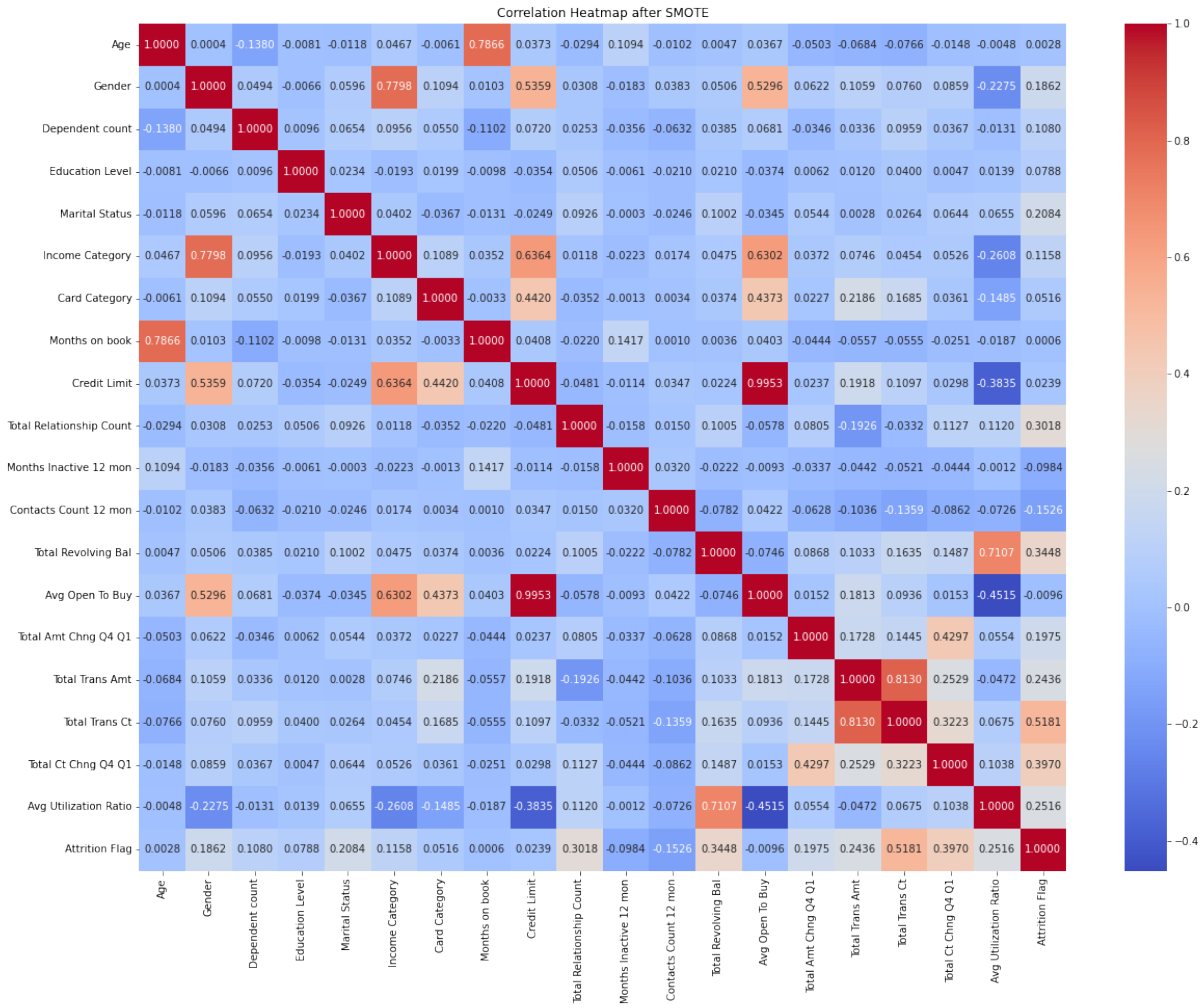

Figure 2

Figure 2 is a visual heat map after balancing the classes using SMOTE. In Figure 2, by using colors, we can easily determine which features are positively correlated with each other and which are negatively correlated. Red means it is positively correlated; conversely, blue means it is negatively correlated. In addition, we can quickly know the degree of correlation between individual features. The redder the color, the more positively correlated; conversely, the bluer the color, the more negatively correlated. Attrition flags are positively correlated with Total Trans Ct, Total Ct Chng Q4 Q1, Total Revolving Bal, and Total Relationship Count, and negatively correlated with Contacts Count 12 mon and Months Inactive 12 mon.

Analysis of factors predicting customer churn

In churn datasets, many features are usually involved, and dealing with such a large number of features leads to high computational costs and storage space [34]. Therefore, dimensionality reduction methods must be used to reduce overfitting, thereby enhancing the data's performance and the generalization of prediction models [18]. Feature selection and feature extraction, as two different feature engineering methods, are critical steps in the machine learning process as they enhance the machine learning model performance, reduce dimensionality, reduce computational cost, and improve the interpretability of models. Combining feature selection and feature extraction into one analysis helps better understand the data. In this study, we chose to apply the methods of feature selection and feature extraction separately to the balanced dataset and visualize and analyze them after extracting the key predictor features of churn.

4.1 Random Forest Feature Importance Ranking

The Random Forest Feature Importance Ranking is a feature selection method that utilizes the Random Forest model to assess the importance of data features. This technique allows us to understand the relative importance of each feature and thus determine which features are critical to model performance. By performing the Random Forest feature importance ranking multiple times, we can derive the seven most important factors used to predict customer churn, consistent with those shown in Figure 3: Total Trans Amt, Total Trans Ct, Total Revolving Bal, Total Ct Chng Q4 Q1, Avg Utilization Ratio, Total Relationship Count and Total Amt Chng Q4 Q1.

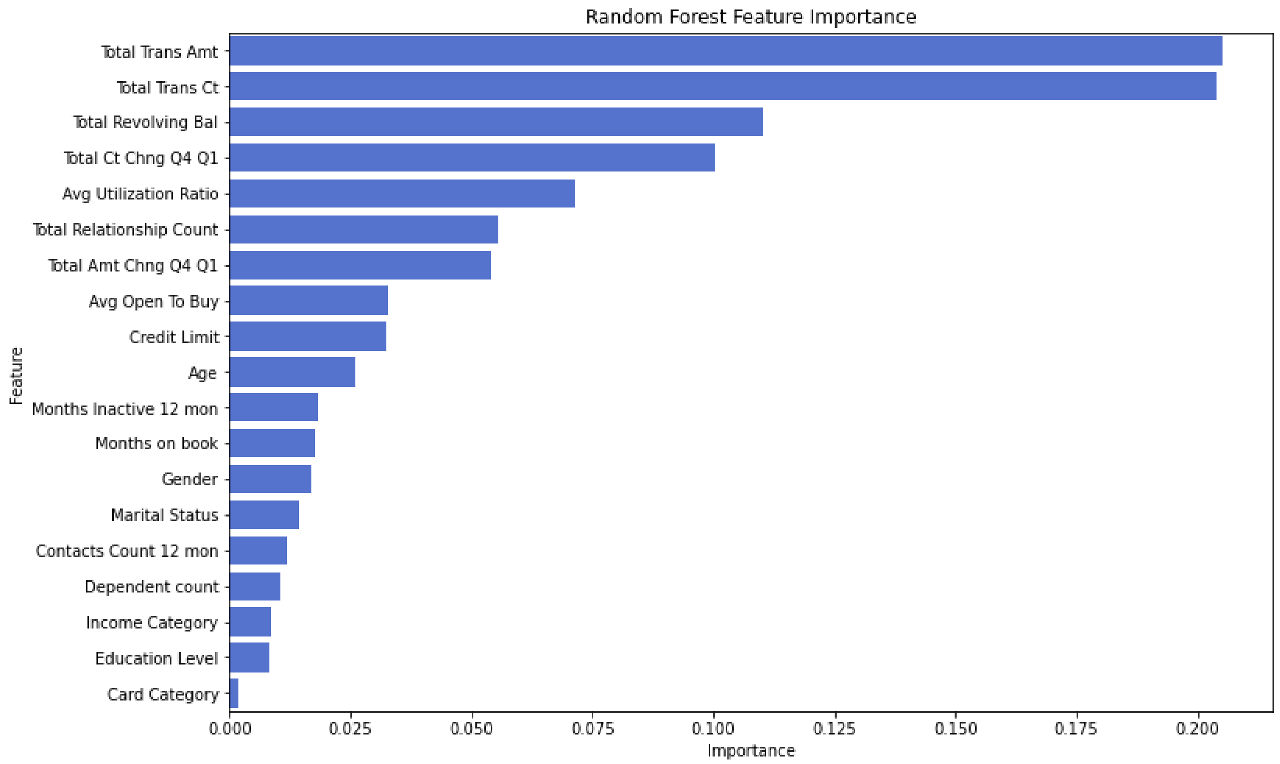

Figure 3

4.2 Visualization of principal component weights

Principal Component Analysis (PCA) is a feature extraction method and product of linear algebra mathematics [35]. PCA transforms the original data into a new set of axes by linear transformation, i.e., principal components are obtained by linear combination. Thus, PCA can transform multi-dimensional data into low-dimensional data, which assists in feature selection [32].

In the PCA technique, we can utilize eigenvalues to select key features. Higher feature values imply more significant features and higher principal component scores, which indicate that they contribute more to the data and can be used for prediction. Additionally, we can visualize the principal component weights, which helps analyze how much each raw feature contributes to the principal components, identify which raw features significantly impact the principal components, and understand the data more deeply.

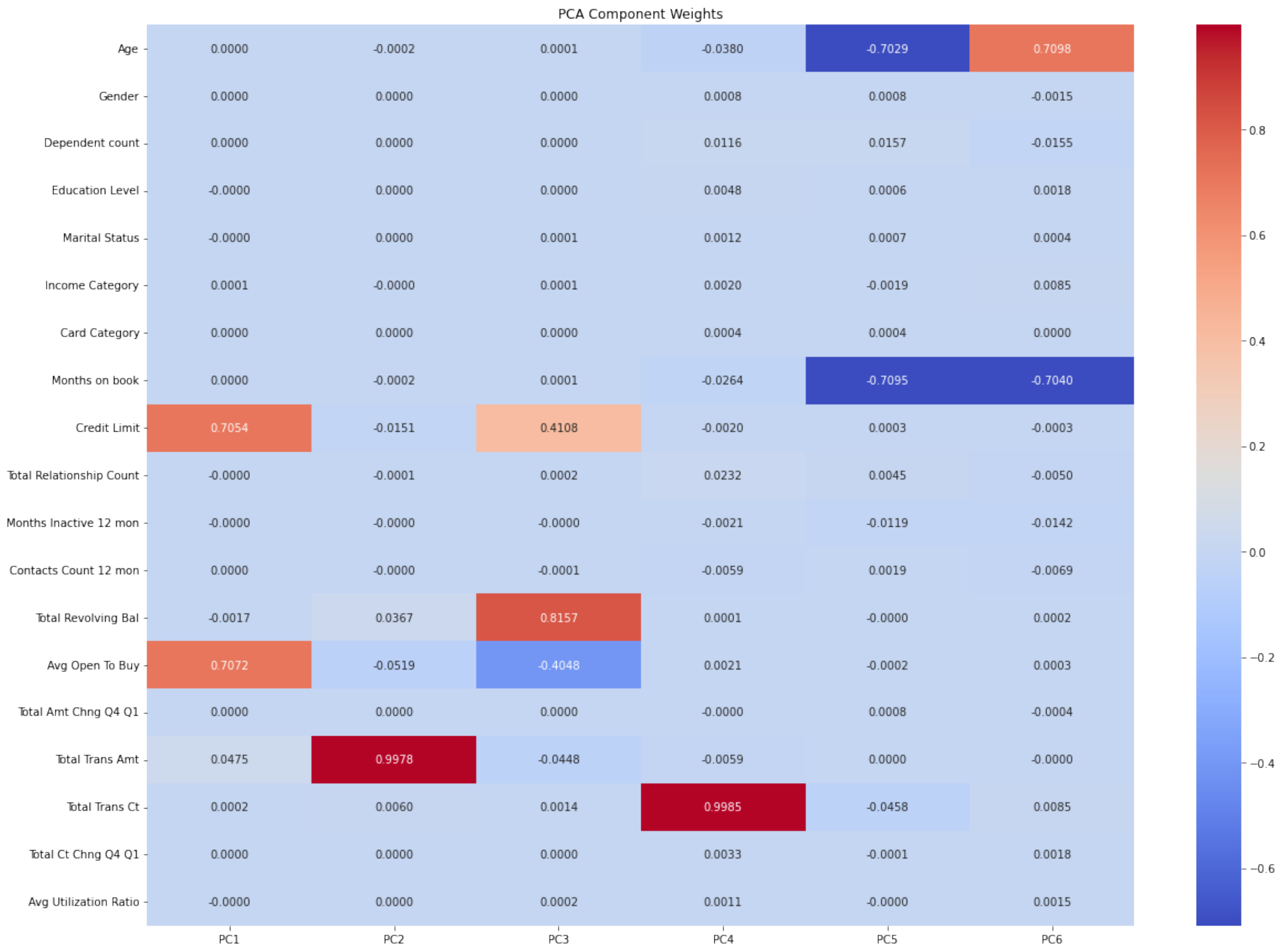

Figure 4

Figure 4 shows the visualization of principal component weights after extracting six principal components by principal component analysis. In Figure 4, we can easily see that the seven original features of Age, Months on the book, Credit Limit, Total Revolving Bal, Avg Open To Buy, Total Trans Amt, and Total Trans Ct have contributed more to the six extracted principal components, especially Total Trans Amt and Total Trans Ct, the degree of contribution to PC2 and PC4, respectively, has reached 0.99. In addition, the positive and negative correlations and the magnitude of the weights can be judged by the color and shade.

Feature selection and feature extraction, as two different feature engineering methods, have different purposes and methods for extracting features. As a result, the features they extract also tend to be different. Therefore, the seven most important predictors of customer churn in the balanced credit card customer dataset extracted by the feature selection (Random Forest Feature Importance Ranking) method are Total Trans Amt, Total Trans Ct, Total Revolving Bal, Total Ct Chng Q4 Q1, Avg Utilization Ratio, Total Relationship Count, and Total Amt Chng Q4 Q1, while the seven predictors of customer churn extracted using feature extraction (PCA) method are Age, Months on book, Credit Limit, Total Revolving Bal, Avg Open To Buy, Total Trans Amt, and Total Trans Ct. Most of the extracted features are different under both methods, but there are the same three features: Total Trans Amt, Total Trans Ct, and Total Revolving Bal. These three features are both the top three in the random forest feature importance ranking and the top three with the highest principal component weights. However, the choice as to whether to go for feature selection or feature extraction depends both on the specific task and requirements and on the subsequent machine learning algorithm used. If one wants to preserve the interpretability and meaning of the original features, feature selection may be more appropriate. Feature extraction may be more appropriate if one wishes to reduce data dimensionality and eliminate redundancy. Furthermore, the feature selection (K-Best) method demonstrated superiority over the feature extraction (PCA) method when paired with four different classification algorithms (XGBoost, Random Forest Classifier, Logistic Regression, and Support vector classifier) in a particular scenario [32].

4.3 Visualization Analysis

The three features, Total Trans Amt, Total Trans Ct, and Total Revolving Bal, are both the top three in the importance ranking of Random Forest features and the top three with the highest principal component weights. Therefore, they are important in the study of customer churn.

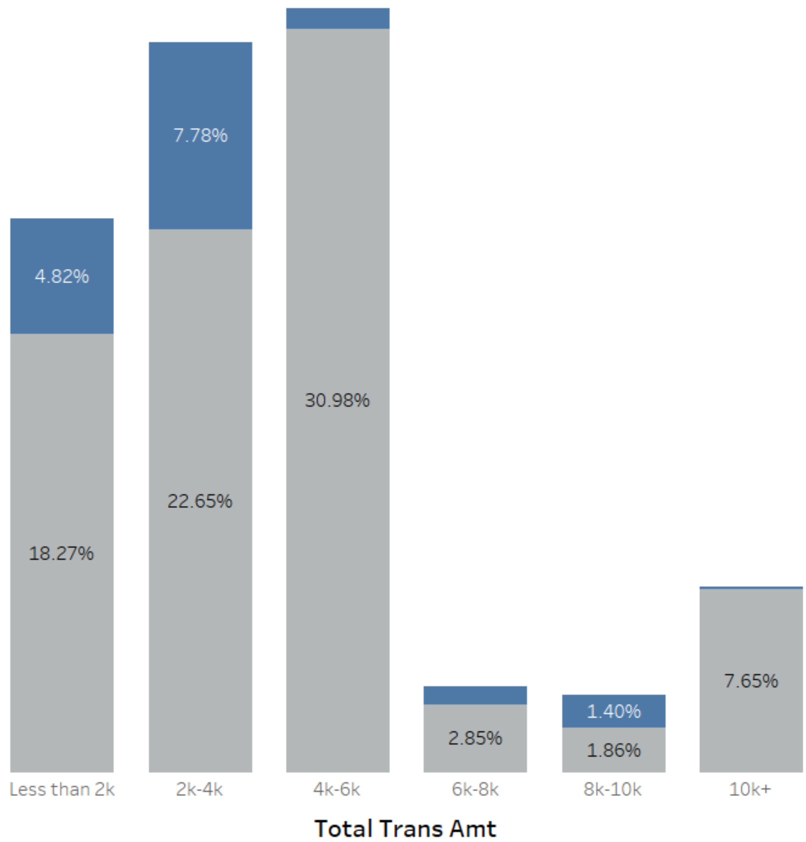

Figure 5

Figure 5 shows the cumulative number of transactions carried out by a customer within the past 12 months. In the dataset, Total Trans Amt has a minimum of 510 and a maximum 17,995. Because of the large span, it is divided into six categories: less than $2k, $2k-4k, $4k-6k, $6k-8k, $8k-10k, and $10k and above. Figure 5 shows that $2k-4k churned customers are the highest, followed by less than 2k. Therefore, customers with fewer total transactions may be more prone to churn. However, churned customers of $8k-10k account for a higher percentage of existing customers of $8k-10k, and hence, customers of $8k-10k also need to be noticed.

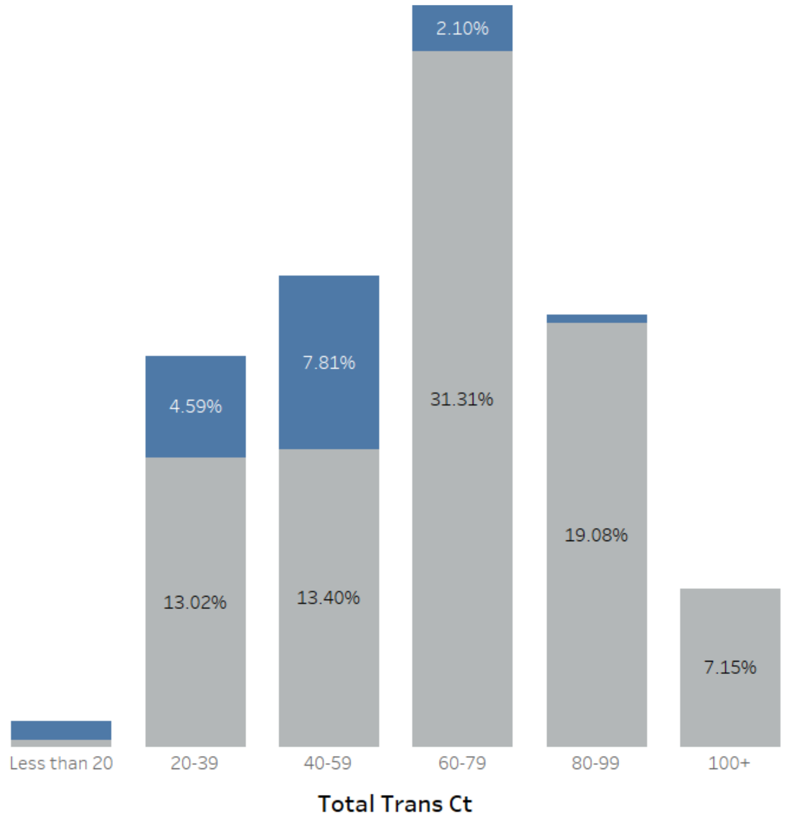

Figure 6

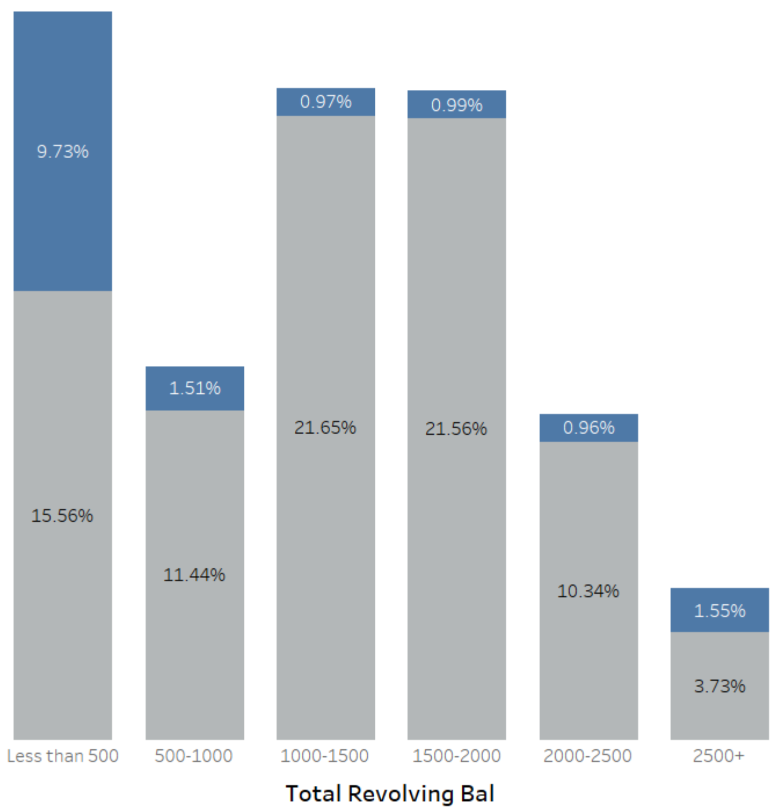

Figure 7

Figure 6 shows the total count of transactions conducted by a customer in the last 12 months. The dataset's minimum number of total transactions is 10, and the maximum is 134. For ease of analysis, it is also divided into six categories,: less than 20, 20-39, 40-59, 60-79, 80-99 and 100 and above. Clearly, the most churned customers are those with 40-59 transactions and customers with fewer transactions are also more likely to churn. When the number of transactions reaches 80 or higher, the churn rate among customers is very low.

Figure 7 shows the outstanding amount that will carry over to the next credit card cycle. The minimum amount outstanding is 132, and the maximum is 2,517, and again can be categorized into six categories: less than $500, $500-1,000, $1,000-1,500, $1,500-2,000, $2,000-2,500, and $2,500 and above. Customers with a total revolving amount of less than $500 are the ones who churn the most.

According to Figure 5, Figure 6, and Figure 7, customers with low total transaction amounts, low numbers of transactions, and low unpaid amounts are relatively more prone to churn. This phenomenon may be related to Customer Engagement (CE), as customers with lower characteristics may be less dependent on banking services, which means they may also be less loyal to the bank [16]. In addition, this also sidesteps the fact that customers with strong ties to financial institutions, owning a substantial number of goods and services and borrowing extensively from banks, are less inclined to close their accounts [11]. Therefore, banks should regularly identify and observe general service-delivery trends to detect customers at risk of churning. In addition, banks should also enhance the ability to predict customer churn and provide preventative structural measures in weak areas [36].

5. Comparative analysis and evaluation of customer churn prediction models

5.1 XGBoost and LightGBM

Both XGBoost and LightGBM are gradient-boosting-based decision tree integration algorithms commonly employed to address classification and regression issues [31]. However, they differ in their splitting strategy. Specifically, LightGBM distinguishes itself from XGBoost by implementing one-sided sampling to filter out data splits. It means that it does not need to sort out the feature values, improvings training speed and efficiency [35]. Despite their different approaches, they both employ a gradient-boosting framework, which can promote weak and strong learners and provide robust performance and accurate predictions [31].

In solving the Class Imbalance Problem (CIP), Boo and Choi [37] compared different ensemble techniques, including Random Forest (RF), Extra-Trees, and XGBoost, as well as various oversampling and undersampling methods. The study findings indicated that the XGBoost model best predicted after SMOTE oversampling. In addition, Mirabdolbaghi and Amiri [35] pointed out that feature reductions can effectively improve speed. Among the feature reduction algorithms, Xgboost performs well. However, in terms of imbalanced metrics and accuracy, LightGBM is more outstanding, especially in evaluating imbalanced related metrics; the Bayesian-based LightGBM algorithm performs excellently. In addition, they also emphasized that boosting algorithms as robust classifiers requires tuning of several hyperparameters, such as learning rate and depth, as optimization of the parameters can enhance the model's performance and accuracy. Therefore, this study used two algorithms, XGBoost and LightGBM, to construct the customer churn prediction model, respectively. After 5-fold cross-validation, the evaluation measures before and after feature selection and before and after adjusting parameters were compared and analyzed.

5.2 Evaluation measures

The evaluation measures are crucial in determining the most effective model for churn prediction methods [35]. These evaluation measures help evaluate the model performance and directly affect the business decisions and outcomes. When evaluating a prediction model [18], different metrics can be relied upon, such as Accuracy, ROC curves and AUC values, F1 scores, and Mathews Correlation Coefficient (MCC). This study used Accuracy, AUC, and f1-Score as evaluation measures.

Accuracy is the proportion of correctly predicted samples out of the total number of samples [38], which can help us understand the correct classification ratio of the model in the overall sample. Typically, accuracy takes precedence as a primary criterion for evaluating the performance of churn prediction models [35]. In general, higher accuracy means that the model categorizes the data more accurately, so high accuracy is usually the goal of the model. However, in the case of class imbalance, the accuracy may be affected by the uneven distribution of categories.

The F1 score, denoting the harmonic mean of precision and recall [39], is a comprehensive evaluation metric, especially well-suited for dealing with unbalanced datasets. A superior F1 score generally suggests that the model maintains a more effective equilibrium between recall and precision [35]. A perfect model has an F1 score of 1 [39].

The ROC curve is a visual representation illustrating the balance between the True Positive Rate and False Positive Rate for a binary classification model at different thresholds. On the other hand, the AUC, or Area Under the Curve, is a comprehensive metric employed to evaluate a model's classification ability [35]. Often, comparing ROC curves of different models can be challenging; therefore, opting for AUC is preferable [40]. A greater AUC value indicates superior model performance [39].

5.3 Feature Selection Methods Evaluation

In section 4.1, this study identifies the top 7 predictors of customer churn using feature selection (Random Forest Feature Importance Ranking) and feature extraction (PCA) methods, respectively. Therefore, this part chose to combine these factors for analysis. At this point, there are eleven features in total, which are Total Trans Amt, Total Trans Ct, Total Revolving Bal, Total Ct Chng Q4 Q1, Avg Utilization Ratio, Total Relationship Count, Total Amt Chng Q4 Q1, Age, Months on book, Credit Limit and Avg Open To Buy.

The dataset used in this section is the SMOTE-balanced dataset. In addition, the customer churn prediction model's target variable is Attrition Flag.

First, the eleven features selected in the dataset for predicting customer churn are brought into the parameter-free XGBoost model for splitting. Then, it is cross-validated with five folds. Finally, the evaluation measures before and after feature selection are derived. Same for the LightGBM model. The results of the evaluation are presented in Table 3.

Table 3. Evaluation results before and after feature selection

	Accuracy	AUC F1-Score	
Parameter-free XGBoost before feature selection	0.897033	0.961250	0.851721
Parameter-free XGBoost after feature selection	0.894938	0.947880	0.849457
Parameter-free LightGBM before feature selection	0.893515	0.917931 0.847805	
Parameter-free LightGBM after feature selection	0.893598	0.892431 0.847765	

Table 3 reveals that the XGBoost model, following SMOTE oversampling, exhibits a slight performance advantage over the LightGBM model in all evaluation measures. As mentioned earlier, the XGBoost model after SMOTE oversampling performs well in prediction, especially in the AUC evaluation metrics [35]. This also indicates that the SMOTE oversampled XGBoost model is superior in classification ability, i.e., it can effectively differentiate between positive and negative categories relative to the SMOTE oversampled LightGBM model. However, the effect of feature selection on the evaluation results is not significant for either the XGBoost model or the LightGBM model. The following reasons may cause this: 1. The dimensionality of the dataset is already relatively low, and feature selection is no longer necessary. Feature selection may be more suitable for datasets with fewer samples and more features than the present dataset; 2. Combining feature selection and feature extraction adds complexity and uncertainty. If one of the methods already provides sufficient performance for the model, then further methods may introduce unnecessary complexity. Therefore, careful trade-offs and separate experiments are needed; 3. Evaluation metrics are less sensitive to feature selection, e.g., accuracy is usually less affected by feature selection; 4. Distributional characteristics of the data may affect the effectiveness of feature selection.

5.4 Hyperparameter Tuning Implementation

Fine-tuning the parameters not only enhances the model's overall performance but also mitigates the risk of overfitting, thereby improving the model's generalization capability. Li et al. [41] have proposed parameter settings for XGBoost and LightGBM models for the original dataset. However, this study proposes an alternative parameter setting for XGBoost and LightGBM models due to different data preprocessing methods. Given that the combined feature selection and feature extraction exert minimal influence on the evaluation outcomes, the features used in this section are all the features in the dataset after SMOTE balanced except the target variable (Attrition Flag). Table 4 displays specific parameter configurations.

Table 4. Parameterization of XGBoost and LightGBM models

The parameters set by Li [41] et al.	XGBoost	n_estimators	300
		max_depth	4
		learning_rate	0.2
		booster	“gbtree”
	LightGBM	num_boost_round	162
		learning rate	0.1
		max_depth	7
		num_leaves	65
		feature_fraction	0.8
		bagging_fraction	0.6
Parameters improved under this method	XGBoost	learning_rate	0.4
		max_depth	5
		n_estimators	500
	LightGBM	learning_rate	0.4
		max_depth	5
		n_estimators	300

After adjusting the parameters, the evaluation outcomes are presented in Table 5.

Table 5. Evaluation results before and after adjustment of parameters

	Accuracy	AUC F1-Score	
Parameter-free XGBoost before feature selection	0.897033	0.961250	0.851721
XGBoost with the parameters set by Li [41] et al.	0.896530	0.954835 0.852231	
XGBoost with improved parameters	0.902060	0.972499 0.863830	
Parameter-free LightGBM before feature selection	0.893515	0.917931 0.847805	
LightGBM under the parameters set by Li [41] et al.	0.896949	0.944565 0.852412	
LightGBM with improved parameters	0.902060	0.969959 0.860289	

The scores of XGBoost and LightGBM are roughly similar, but there is a slightly larger difference in AUC. Compared to the parameter settings proposed by Li et al. [41], the parameters proposed in this study were slightly better in terms of evaluation results under the data preprocessing method of this study. After adjusting the parameters, the scores of both the XGBoost and LightGBM models improved, although the overall change was small. The small change may be due to the following reasons: 1. the models may already have relatively good performance, and thus by adjusting the parameters, it may not be possible to significantly enhance the models' performance; 2. the number of parameters adjusted and the range of parameter choices are somewhat limited, which may cause the models to miss out on better parameter combinations due to the failure to adequately cover a wide range of the parameter search space; 3. the data itself may limit the models.

5.5 Interpretability of the model

Given the importance of churn prediction, a deep understanding of the model is crucial [35]. The SHAP method is part of the feature interpretation, quantifying the extent to which each feature contributes to the predictions made by the model and provides a more comprehensive analysis of feature importance. This aids in comprehending the model's predictions and their impact on the final predictions.

There are many ways to apply SHAP, such as SHAP Feature Importance, SHAP Summary Plot, SHAP Interaction Values, KernelSHAP, and TreeSHAP. Given that the visualization methods are more intuitive and the XGBoost model with improved parameters in the previous paper is slightly better regarding evaluation results. Therefore, this study will be based on the improved parameterized XGBoost algorithm, using the SHAP Feature Importance and SHAP Summary Plot methods to visualize and analyze the influence degree of features as a whole.

5.5.1 SHAP Feature Importance

The SHAP Feature Importance is determined by calculating the mean of the absolute values representing the extent to which a feature influences the target variable, which indicates that feature's importance. The Feature importance shows the extent to which each feature contributes to the predictive power, and the larger the SHAP value, the more influential the feature is. While SHAP Feature Importance, Feature Selection, and Feature Extraction all involve the analysis and processing of features, Feature Selection and Feature Extraction are typically used for data preprocessing to reduce the dimensionality of the dataset, and SHAP value is used to understand the model predictions. As a result, a different ordering of importance is derived. The results of the SHAP Feature Importance are shown in Figure 8.

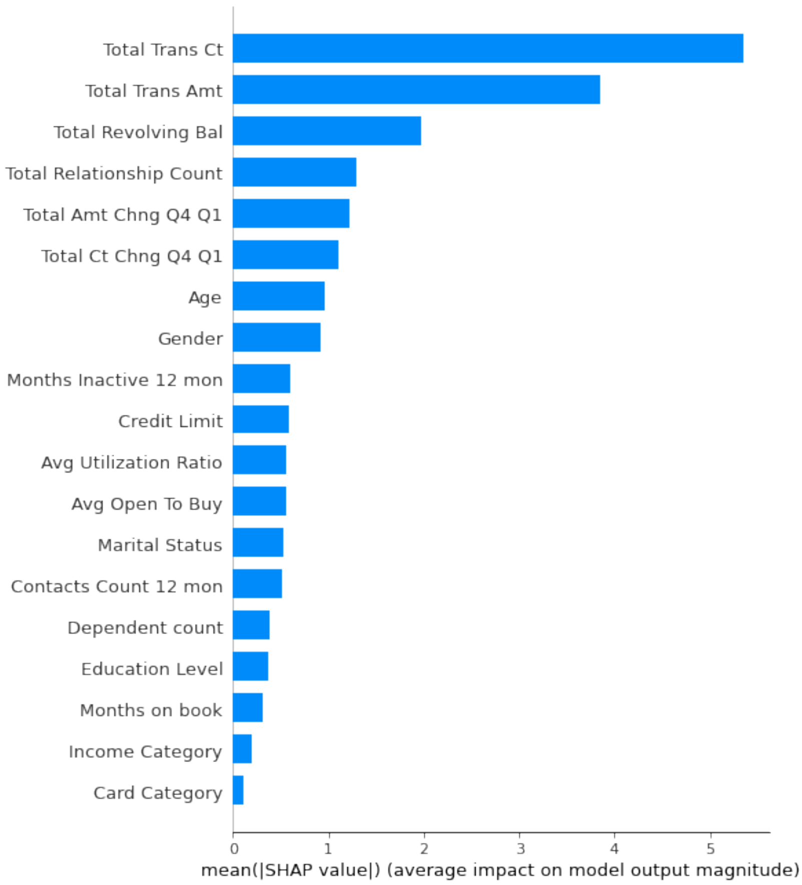

Figure 8

As can be seen from Figure 8, Total Trans Ct is the most important feature, changing the predicted churn probability by more than five percentage points on average (more than five on the x-axis). In addition, Total Trans Amt and Total Revolving Bal ranked second and third in importance, respectively. This is consistent with the first three features extracted earlier through feature selection and feature extraction.

5.5.2 SHAP Summary Plot

The SHAP Summary Plot combines the significance of features with their impact, providing a comprehensive representation of both positive and negative relationships between predictors and target variables. This helps to understand the overall pattern and detect predictive outliers in time. In the SHAP Summary Plot, the horizontal coordinate is the SHAP value, the magnitude of which indicates the extent to which the feature contributes to predicting customer churn. A positive SHAP value indicates that a feature positively contributes to increasing the probability of the model's predicted outcome; conversely, a negative SHAP value suggests a negative influence on the predicted outcome's probability. In addition, each row in the plot corresponds to a feature, and each point represents a sample. The color gradient reflects the feature values, with redder shades indicating higher values and bluer shades indicating lower values. The overlapping points along the y-axis provide insights into each feature's distribution of SHAP values. The SHAP Summary Plot is typically organized by the magnitude of SHAP values, representing their respective importance.

Figure 9

With Figure 9, it is intuitive to see that the dots on the far right of the Total Trans Ct row are essentially red, indicating that the larger the feature, the more positively it affects customer retention. Similarly, the total revolving balance is positively correlated with the target variable. However, the larger the Total Trans Amt is, the more negatively it affects customer retention, as the leftmost point on the Total Trans Amt row is essentially red. And while this is, as mentioned earlier, customers with fewer Total Trans Amt may be more prone to churn, customers in the $8k-$10k range need to be looked at as well, since $8k-$10k churned customers make up a higher percentage of existing customers in the $8k-$10k range. It is important to note that all of the impacts in this section only describe the model's behavior and do not equate to a causal relationship that necessarily exists in the real world.

The SHAP methodology and related visualization tools better enhance the analysis and interpretation of machine learning models' predicted results, which helps focus on the customer and understand the root causes of churn, leading to improved models, increased efficiency, and better management decisions [35].

6. Conclusions and Recommendations

This paper aims to investigate the factors and methodologies for predicting the churn of bank customers. Using the "Credit Card Customer" dataset on Kaggle and combining visualization, data mining, and machine learning methods, this paper comprehensively explores and analyzes the key factors that predict customer churn in banks from multiple perspectives. In addition, two churn prediction models utilizing the Gradient Boosting Tree algorithm, XGBoost, and LightGBM, are also constructed.

This study found several conclusions as follows. First, Total Trans Amt, Total Trans Ct, and Total Revolving Bal are critical in analyzing and predicting customer churn. They are not only the top three most important common features extracted by the Feature Selection (Random Forest Feature Importance Ranking) method and the Feature Extraction (PCA) method, but their importance was also demonstrated by the SHAP Feature Importance method. Second, the visual analysis of the factors predicting customer churn can be responded to to some extent by the SHAP Summary Plot. Third, the effect of feature selection on the evaluation results is sometimes insignificant, such as in this dataset after the data preprocessing methods of this study. Fourth, the tuning parameter will enhance the model's performance to some degree, yet the optimal parameter settings will be different depending on the preprocessing method.

In light of the findings above, this paper makes the following recommendations: First, banks need to regularly monitor trends in customer transaction behavior because customers with low total transaction amounts, low number of transactions, and low outstanding amounts are relatively more likely to churn. This implies that banks need to provide timely interactions and enhance customer engagement (CE) rather than focusing solely on the functional attributes of the service [16]. In addition, banks also need to enhance the brand experience (BE), such as offering customized products or services to enhance the customer's consumer experience. This also helps to maintain customer loyalty. Second, banks can enhance the combined visualization methods and machine learning use. Visualization tools provide bank managers with intuitive insights, such as a better visualization of customer usage, needs, and preferences, understanding market data, and monitoring transaction trends, which can help to better understand the current state of affairs and thus formulate relevant policies and strategies. Machine learning, on the other hand, helps bank managers analyze large amounts of financial data to identify potential risks, such as fraud and credit risk. In addition, machine learning helps banks to analyze customer behavior and strengthen their ability to predict customer churn, as well as to predict future market trends and investment risks, etc., to provide early warning and take targeted preventive measures at potential weak points [36]. The combined use of visualization and machine learning methods will empower bank managers to acquire a holistic and profound understanding of various aspects such as business dynamics, customer needs, and potential risks, thus enabling them to make better use of data to make decisions, improve business efficiency, and maintain a competitive edge in the highly competitive financial market. Third, the data preprocessing method has an important impact on feature selection and model performance. Therefore, banks should carefully select and carefully optimize data preprocessing methods to ensure that the model can maximize the use of available data and thus improve forecast accuracy. Fourth, adjusting model parameters can improve performance but require different settings depending on the data preprocessing method. Banks should invest time and resources in optimizing model parameters to ensure the best model performance is obtained. In summary, the above recommendations can help banks better understand the factors that contribute to customer churn, improve their products and services, enhance their customer relationship management, and develop more effective policies and strategies to increase customer inertia, which can result in cost savings, reduced churn, and improved performance. In addition, banks need to pay close attention to customer utilization and apply appropriate measures to actual business decisions.

This paper combines various methods, such as visualization, data mining, and machine learning, to explore and analyze the factors that predict bank customer churn from multiple perspectives. In addition, two churn prediction models, XGBoost and LightGBM, were constructed, compared, and analyzed regarding feature selection and model tuning. This paper also utilizes SHAP methods to explain the model results in depth. This research provides useful ideas for future research on related methods. As machine learning continues to evolve, models for predicting customer churn have become more complex, but their predicted results have become more accurate and reliable. However, there is still little research on integrated approaches combining visualization and machine learning and multi-perspective analysis and validation. Therefore, future research should emphasize the integrated use of methods for a more comprehensive and effective analysis. In addition, the generalization ability is still an important challenge, and future research can explore the construction of hybrid models applicable to domains with better generalization ability or adopt novel feature engineering methods. Also, future research could consider broadening the parameter search and trying different model architectures with trade-offs and experiments based on specific problems and data characteristics. In addition, there is a need to regularly monitor and refine the model's performance, which is essential to enhance the performance and applicability of the model.

Nonetheless, there are limitations to this study. First, it combines two different feature engineering methods, feature selection, and feature extraction, which increases the complexity and uncertainty and creates challenges for the subsequent application of machine learning methods. Second, although model tuning was performed, only a few parameters were adjusted within a certain range. Thus, the impact of tuning on the evaluation results still needs to be verified in more detail. Finally, the analysis in this study exclusively relies on the dataset of "credit card customers," necessitating further verification to assess the model's generalization capabilities.

In summary, to enhance the credibility of this study, future validation on several different datasets can be considered, and more evaluation metrics and tuning strategies can be explored to find the optimal model configuration. In addition, customer segmentation methods that can be studied from multiple perspectives, such as using the k-means algorithm for finer segmentation of customers, can also be explored to achieve more accurate customer churn prediction and management. Constructing more stable and efficient customer churn models can be explored for unbalanced datasets, and optimization methods are not limited to parameter tuning. Most importantly, however, a combination of approaches can be considered for a holistic analysis.

References

[1] Cuadros-Solas, P. J., Cubillas, E., & Salvador, C. 2023. Does alternative digital lending affect bank performance? Cross-country and bank-level evidence. International Review of Financial Analysis, 90

[2] Murinde, V., Rizopoulos, E., & Zachariadis, M. 2022. The impact of the FinTech revolution on the future of banking: Opportunities and risks. International Review of Financial Analysis, 81.

[3] Rühl, A., & Zurdo, R. P. 2020. Does technology contribute to financial democratization? The collaborative economy and fintechs as catalysts for change. Revesco-Revista De Estudios Cooperativos(133).

[4] Langley, P., & Leyshon, A. 2021. The Platform Political Economy of FinTech: Reintermediation, Consolidation and Capitalisation. New Political Economy, 26(3): 376-388.

[5] Gao, C. Y., & Wang, Q. 2023. Does digital finance aggravate bank competition? Evidence from China. Research in International Business and Finance, 66.

[6] Kowalewski, O., & Pisany, P. 2022. Banks' consumer lending reaction to fintech and big tech credit emergence in the context of soft versus hard credit information processing. International Review of Financial Analysis, 81.

[7] Liem, N. T., Son, T. H., Tin, H. H., & Canh, N. T. 2022. Fintech credit, credit information sharing, and bank stability: some international evidence. Cogent Business & Management, 9(1).

[8] Varma, P., Nijjer, S., Sood, K., Grima, S., & Rupeika-Apoga, R. 2022. Thematic Analysis of Financial Technology (Fintech) Influence on the Banking Industry. Risks, 10(10).

[9] Elsaid, H. M. 2021. A review of literature directions regarding the impact of fintech firms on the banking industry. Qualitative Research in Financial Markets.

[10] Wang, Y. H., Shih, K. H., & Huang, Y. C. 2012. Measurement of Switching Cost on the Customer Retention in the Banking Industry. Journal of Testing and Evaluation, 40(6): 923-930.

[11] Lemos, R. A. D., Silva, T. C., & Tabak, B. M. 2022. Propension to customer churn in a financial institution: a machine learning approach. Neural Computing & Applications, 34(14): 11751-11768.

[12] <2018 customergauge nps & cx benchmarks report.pdf>.

[13] Domingos, E., Ojeme, B., & Daramola, O. 2021. Experimental Analysis of Hyperparameters for Deep Learning-Based Churn Prediction in the Banking Sector. Computation, 9(3).

[14] Park, W., & Ahn, H. 2022. Not All Churn Customers Are the Same: Investigating the Effect of Customer Churn Heterogeneity on Customer Value in the Financial Sector. Sustainability, 14(19).

[15] De, S. M., & Prabu, P. 2022. Predicting customer churn: A systematic literature review. Journal of Discrete Mathematical Sciences & Cryptography, 25(7): 1965-1985.

[16] Rasool, A., Shah, F. A., & Tanveer, M. 2021. Relational dynamics between customer engagement, brand experience, and customer loyalty: An empirical investigation (vol 20, pg 273, 2021). Journal of Internet Commerce, 20(4): 508-508.

[17] Hübner, F., Herberger, T. A., & Charifzadeh, M. 2023. Determinants of customer recovery in retail banking-lessons from a German banking case study. Journal of Financial Services Marketing.

[18] Geiler, L., Affeldt, S., & Nadif, M. 2022. A survey on machine learning methods for churn prediction. International Journal of Data Science and Analytics, 14(3): 217-242.

[19] Khodadadi, A., Hosseini, S., Pajouheshgar, E., Mansouri, F., & Rabiee, H. R. 2022. ChOracle: A Unified Statistical Framework for Churn Prediction. Ieee Transactions on Knowledge and Data Engineering, 34(4): 1656-1666.

[20] Oskarsdóttir, M., Van Calster, T., Baesensa, B., Lemahieu, W., & Vanthienen, J. 2018. Time series for early churn detection: Using similarity based classification for dynamic networks. Expert Systems with Applications, 106: 55-65.

[21] Salas-Eljatib, C., Fuentes-Ramirez, A., Gregoire, T. G., Altamirano, A., & Yaitul, V. 2018. A study on the effects of unbalanced data when fitting logistic regression models in ecology. Ecological Indicators, 85: 502-508.

[22] Branco, P., Torgo, L., & Ribeiro, R. P. 2016. A Survey of Predictive Modeling on Im balanced Domains. Acm Computing Surveys, 49(2).

[23] Wang, S. J., Liu, W., Wu, J., Cao, L. B., Meng, Q. X., Kennedy, P. J., & Ieee. 2016. Training Deep Neural Networks on Imbalanced Data Sets. Paper presented at the International Joint Conference on Neural Networks (IJCNN), Vancouver, CANADA.

[24] Zhou, F. N., Yang, S., Fujita, H., Chen, D. M., & Wen, C. L. 2020. Deep learning fault diagnosis method based on global optimization GAN for unbalanced data. Knowledge-Based Systems, 187.

[25] Yang, C., Shi, X. L., Luo, J., Han, J. W., & Acm. 2018. I Know You‘ll Be Back: Interpretable New User Clustering and Churn Prediction on a Mobile Social Application. Paper presented at the 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), London, ENGLAND.

[26] Zhao, Z., Peng, H., Lan, C. W., Zheng, Y., Fang, L., & Li, J. Y. 2018. Imbalance learning for the prediction of N⁶-Methylation sites in mRNAs. Bmc Genomics, 19.

[27] Luo, R. S., Dian, S. Y., Wang, C., Cheng, P., Tang, Z. D., Yu, Y. M., Wang, S. X., & Iop. 2018. Bagging of Xgboost Classifiers with Random Under-sampling and Tomek Link for Noisy Label-imbalanced Data. Paper presented at the 3rd International Conference on Automation, Control and Robotics Engineering (CACRE), Chengdu, PEOPLES R CHINA.

[28] Khoh, W. H., Pang, Y. H., Ooi, S. Y., Wang, L. Y. K., & Poh, Q. W. 2023. Predictive Churn Modeling for Sustainable Business in the Telecommunication Industry: Optimized Weighted Ensemble Machine Learning. Sustainability, 15(11).

[29] Haridasan, V., Muthukumaran, K., & Hariharanath, K. 2023. Arithmetic Optimization with Deep Learning Enabled Churn Prediction Model for Telecommunication Industries. Intelligent Automation and Soft Computing, 35(3): 3531-3544.

[30] Thakkar, H. K., Desai, A., Ghosh, S., Singh, P., & Sharma, G. 2022. Clairvoyant: AdaBoost with Cost-Enabled Cost-Sensitive Classifier for Customer Churn Prediction. Computational Intelligence and Neuroscience, 2022.

[31] Arshad, S., Iqbal, K., Naz, S., Yasmin, S., & Rehman, Z. 2022. A Hybrid System for Customer Churn Prediction and Retention Analysis via Supervised Learning. Cmc-Computers Materials & Continua, 72(3): 4283-4301.

[32] Anjana, K. V., & Urolagin, S. 2021. Churn Prediction in Telecom Industry Using Machine Learning Algorithms with K-Best and Principal Component Analysis, Singapore.

[33] Tékouabou, S. C. K., Gherghina, S. C., Toulni, H., Mata, P. N., & Martins, J. M. 2022. Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods. Mathematics, 10(14).

[34] Liu, Z. H., Lai, Z. H., Ou, W. H., Zhang, K. B., & Zheng, R. J. 2020. Structured optimal graph based sparse feature extraction for semi-supervised learning. Signal Processing, 170.

[35] Mirabdolbaghi, S. M. S., & Amiri, B. 2022. Model Optimization Analysis of Customer Churn Prediction Using Machine Learning Algorithms with Focus on Feature Reductions. Discrete Dynamics in Nature and Society, 2022.

[36] Lappeman, J., Franco, M., Warner, V., & Sierra-Rubia, L. 2022. What social media sentiment tells us about why customers churn. Journal of Consumer Marketing, 39(5): 385-403.

[37] Boo, Y., & Choi, Y. 2022. Comparison of mortality prediction models for road traffic accidents: an ensemble technique for imbalanced data. Bmc Public Health, 22(1).

[38] Wu, S. L., Yau, W. C., Ong, T. S., & Chong, S. C. 2021. Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. Ieee Access, 9: 62118-62136.

[39] Sana, J. K., Abedin, M. Z., Rahman, M. S., & Rahman, M. S. 2022. A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection. Plos One, 17(12).

[40] Xiao, J., Xiao, Y., Huang, A. Q., Liu, D. H., & Wang, S. Y. 2015. Feature-selection-based dynamic transfer ensemble model for customer churn prediction. Knowledge and Information Systems, 43(1): 29-51.

[41] Li, J. F., Bai, X., Xu, Q., & Yang, D. X. 2023. Identification of Customer Churn Considering Difficult Case Mining. Systems, 11(7).