Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230865

Analysis and Prediction of Bank Customer Loyalty Based on XGBoost Algorithm

Yuyan ZHANG[1], Ke CHEN[2], Ting CHEN[3]
College of Business, Hohai University, Nanjing, China

Abstract- At present, the homogenization of banking products and the vigorous development of internet finance have intensified the competition among banks. Customers are the core assets of banks, whose size and loyalty is crucial to any bank. Loyal customer's repeat purchases or recommending products to people around creates higher value for banks. Therefore, in order to improve customer loyalty, a method of identifying customer loyalty is urgently needed which prioritizes providing more personalized services for loyal customers. Based on bank's long-term customer resource data, this paper divides customer groups by means of feature selection and data processing, compares the experimental results of multiple machine learning models such as GBDT, and selects the optimal XGBoost model to predict customer's long-term loyalty to banks, in order to predict potential customer churn for banks, attempt to retain high-value customers as much as possible, and to increase potential revenue.

Keywords-Machine learning; Customer loyalty; XGBoost model; Bank data analysis

1. Introduction

Bank customer loyalty, a quantitative concept, refers to the degree of bank customer's loyalty. Bank customer loyalty refers to the degree to which customers develop feelings for a certain bank due to influences such as their own attributes, product quality, banking services, thereby having preference for and long-term and repeated purchasing behavior of bank's products or services.[1]

Researches on customer loyalty in the banking industry are earlier seen in western countries. Solimun and Adji A R (2018) proposed that customer loyalty is affected by the performance evaluation of financial institutions, and is prone to generate certain intermediate damage[2].Gomezcruz M E (2019) proposed that the distribution of bank outlets, the quality of financial products, the professionalism of banking employees, company culture, and customer affective commitment will all affect customer loyalty[3]. Teeroovengadum V (2020) built a model for specific situations of bank customer's satisfaction through a scale, and analyzed the interrelationship of various factors that affect bank customer loyalty[4].

Domestic researches on customer loyalty in the banking industry started relatively later. Yu Chengwei (2018) proposed that every 5% increase in loyal customers can lead

[1]Corresponding Author: Yuyan ZHANG, College of Business, Hohai University; e-mail: sparkspark@foxmail.com
[2]Ke CHEN, College of Business, Hohai University; e-mail: 3014020014@qq.com
[3]Ting CHEN, College of Business, Hohai University; e-mail: ct345i@163.com

to 20%-80% surge in profits by summarizing the data of existing research results. Therefore, Yu Chengwei believed it is necessary to improve the loyalty of bank customers[5]. Jin Wenzhao (2019) analyzed the model of bank customer loyalty based on the construction of existing ECSI and CCSI models[6]. Zheng Hanhui (2019) focused on the nature of customer loyalty and analyzed the views of consumers with different levels of loyalty on special products or services of banks[7]. Hu Zhiting (2020) believed that in the digital era, enterprises need to put more efforts on improving the quality of the service climate, through which to enhance customer's loyalty to enterprises in a positive manner. [8].

When analyzing customer loyalty, few machine learning algorithms are used in China, so is true with studies on comparing the prediction effects of different models. This paper draws on research ideas and methods of customer loyalty evaluation based on long-term customer resource data, and subdivides customer groups through feature selection and data processing. Besides, the paper also establishes prediction models and compares model effects based on eight algorithms such as XGBoost and GBDT, so as to meet the diverse needs of customers and increase banking revenues with its technical innovation and feasibility.

2. Research Methodology

2.1. Feature Selection for the Prediction Model

Since the resource data obtained from bank's long-term customers in reality is often incomplete, machine learning models are often unable to effectively identify and extract information from it. Data and features determine the upper limit of machine learning, while models and algorithms can only approach this upper limit. After data is collected, the main step of machine learning modeling is feature selection.

The process of feature selection is based on selecting the most consistent, relevant and non-redundant subset of features from the original feature vector. It not only reduces training time and model complexity, but also facilitates in preventing overfitting in the end.

The easiest and most direct way to measure customer loyalty is repeat purchases. However, with the deepening and development of theoretical researches, customer loyalty can no longer be simply taken as the equivalent to repeated purchasing behavior. Short-term customer loyalty analysis can analyze different customer's purchasing dependence on bank products through product's purchasing data in order to provide better sales services. However, the long-term customer loyalty analysis to be studied in this paper is to identify characteristics from customer resource data, pinpoint customer churn factors, and predict potential customer that might been lost.

Considering that current banks are close to maturity in product designs, the breadth and depth of bank's coverage are quite similar, the main financial products are strong in homogeneity, and that this paper focuses on dividing customer groups in order to provide corresponding products and services in accordance with customer's habits and preferences, product attributes are not considered in the model proposed in this paper.

As shown in the Table 1, this paper summarizes the long-term customer loyalty indicators, and selects 12 features including customer gender (Gender), customer age (Age), customer annual income (Tenure) and so on as the core features of the long-term customer loyalty prediction model.

Table 1. Core features of long-term customer loyalty prediction model

Dimension	Characteristic	Explanation
Basic Information	Gender	The gender of the customer
	Age	Age of the customer
	EstimatedSalary	Customer personal annual income
Account opening situation	CreditScore	Credit qualification, where higher values indicate higher credit
	Tenure	Account age, the length of time the customer has deposited with the bank, in years
	Balance	AUM, financial assets of the customer
Transaction behavior	NumOfProducts	Number of products purchased by the customer
	IsActiveMember	Customer activity status, 1 if the customer is active, 0 otherwise
	HasCrCard	Customer's credit card status, 1 if the customer has a credit card, 0 otherwise
Composite index	IsActiveStatus	Characteristics of new and regular customers' activity level
	IsActiveAssetStage	Characteristics of customer activity by deposit amount
	CrCardAssetStage	Characteristics of credit card holding status of different financial assets

It should be noted that the new and regular customer activity feature (IsActiveStatus) is a composite index based on the account age and customer active status. Among them, the account age interval [0, 3] is a new customer, and the account age range (3,6] refers to stable customers and the ones in (6, ∞) are regular customers. The activity characteristics of new and regular customers are encoded according to the rules.

The characteristics of customer activity of different financial assets (IsActiveAssetStage) are composite indicators based on the asset stage and customer active status. Among them, the financial asset range [0, 50000] is marked as low assets, and those in (50000, 90000] is marked as medium lower assets, (90000, 120000] is marked as medium and upper assets, and those exceeds 120000 is marked as high assets. The activity characteristics of new and regular customers are encoded according to the rules.

The credit card holding status characteristics (CrCardAssetStage) of different financial assets are composite indicators based on the asset stage and credit card holding status. The division of asset stages is consistent with the above-mentioned statement. The activity characteristics of new and regular customers are encoded according to the rules.

2.2. Principle of the Model

Extreme Gradient Boosting (XGBoost)[9], designed by Dr. Tianqi Chen from the University of Washington, is dedicated to allowing multiple decision trees to break through their own computing limits, so as to achieve the engineering goals of fast computing and high performance. This paper is the first attempt to apply the XGBoost algorithm to the prediction of long-term loyalty of bank customers.

The main workings of XGBoost are to randomly select certain sample variables as the base learners, and repeat the fitting residual to minimize the objective function. For a data set containing n pieces of m dimensions, the XGBoost model can be expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F(i = 1,2,...,n) \tag{1}$$

where $F = \{f(x) = w_{q(x)}\}(q : R^m \to \{1,2,...,T\}, w \in R^T)$ is the set of decision tree structures, q is the tree structure of samples mapped to leaf nodes, T is the number of leaf nodes, and w is the real fraction of leaf nodes. The objective function of the XGBoost model is divided into an error function term L and a model complexity function term Ω:

$$Obj = L + \Omega \tag{2}$$

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

$$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

Among them, γT is the regular term of L_1 and $\frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ is the regular term of L_2.

When optimizing the model with training data, the original model is required to remain unchanged and a new function is added to the model to minimize the objective function, namely:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{5}$$

where, $\hat{y}_i^{(t)}$ represents the predicted value of the model for the t time, $f_t(x_i)$ is the new function added for the t time. At this time, the objective function is:

$$Obj^{(t)} = \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_i(x_i)))^2 + \Omega \tag{6}$$

In the XGBoost algorithm, in order to quickly find the parameters that minimize the objective function, the second-order Taylor expansion of the objective function is carried out, and the approximate objective function is obtained:

$$Obj^{(t)} \approx \sum_{i=1}^{n} [(y_i - \hat{y}_i^{(t-1)})^2 + 2(y_i - \hat{y}_i^{(t-1)}) f_i(x_i) - h_i f_t^2(x_i)] + \Omega \tag{7}$$

After removing the constant term, it's clear that the objective function is only related to the first and second derivatives of the error function. At this point, the objective

function is:

$$Obj^{(t)} \approx \sum_{i=1}^{n}[g_i w_{q(x_i)} + \frac{1}{2}h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2}\sum_{j=1}^{T} w_j^2$$
$$= \sum_{j=1}^{T}[(\sum_{i \in I_j} g_i)w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda)w_j^2] + \gamma T \quad (8)$$

2.3. Model Evaluation

2.3.1. Confusion Matrix

After the model is established, evaluation of the model's performance is required. The confusion matrix is a summary of various situations of the predictive results of the classification model, and reveals the real classification in the data set and the classification results predicted by the model in the form of a matrix. As shown in the Table 2, TP means that actual and predicted results are both positive, and FP means that the actual results are negative and the predictive results are positive; FN means the actual results are positive and the predicted results are negative; and TN means both the actual and predicted results are negative.

Table 2. Confusion matrix

	Predicted to be positive category	**Predicted to be negative category**
Actual is positive category	TP	FN
Actual is negative category	FP	TN

2.3.2. Evaluation Metrics of Model Training

Based on the confusion matrix, different model evaluation metrics are provided: accuracy, precision, recall, F1-Score.

a) Accuracy

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (9)$$

For the number of samples correctly predicted, the total number of samples is the prediction accuracy rate, and the larger the value, the better.

b) Precision

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

Accuracy refers to the proportion of samples predicted to be in the minority class that are actually in the minority class. The larger the value of accuracy, the higher the accuracy of the model in predicting minority class samples. Conversely, a small value of accuracy indicates that the model misclassifies many majority classes.

c) Recall

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{11}$$

The recall refers to the proportion of the number of samples that are correctly predicted to be among the true minority class samples. The higher the value of the recall, the more minority classes the model acquires, and vice versa, indicating that the ability of the model to acquire minority classes is relatively weaker. The recall is inversely proportional to the precision, and the balance between the two means that the minority class is captured based on the requirements of the majority class. Which one to prefer should be judged according to actual business needs: whether it is more costly to accidentally damage the majority class, or fail to capture the minority class.

d) F1

In order to take both precision and recall into account, the harmonic mean of the two can be used as a comprehensive indicator when considering the balance between the two, that is, F1-Score.

$$F1 = \frac{2}{\frac{1}{\mathrm{Pr}esicion} + \frac{1}{\mathrm{Re}call}} = \frac{2 * \mathrm{Pr}esicion * \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \tag{12}$$

When the harmonic mean of the two numbers is close to the small number of the two, a higher F1-Score should be selected to ensure a higher recall and precision of the model. The range of F1-Score is [0,1], and the smaller its difference value from 1, the better.

3. Empirical Research

3.1. Data Preprocessing

The data in this paper originates from a domestic bank's user data disclosed by the "BdRace" data mining competition platform (BdRace), with a total of 9,300 data source samples, each of which corresponds to a certain customer's statistical information. Features contained are common in the banking industry, such as gender, age, income, and so on, and are relatively referential to the prediction of long-term customer loyalty.

As shown in Table 3, the data set contains a total of 13 variables, including 7 numerical variables and 5 variable subtypes. The data is firstly pre-processed based on the following guidelines: (1) removing samples with significant data gaps; (2) removing outliers that are against objective reality.

Through further analysis of the data, it can be seen that problems such as uneven distribution of the "Exited" (customer churn) features and the large span of numerical distribution of each value prevail. Specifically, the number of customers who have not been lost is more than three times the number of customers who have been lost; the maximum value of customer credit qualification reaches 250,000, and the status of customer activity is 0 and 1, which means that the sample is unbalanced and the order of magnitude of each measurement index is different.

Table 3. Summary of data set features

Name of feature	Description	Type of variable
CustomerId	Customer ID	Numerical
CreditScore	Qualification for credit	Numerical
Gender	The gender of the customer	Categorical
Age	Age of the customer	Numerical
Tenure	Age of account	Numerical
Balance	financial assets of the customer	Numerical
NumOfProducts	Number of products purchased by the customer	Numerical
HasCrCard	Customer's credit card status	Categorical
IsActiveMember	Customer activity status	Categorical
EstimatedSalary	Customer personal annual income	Numerical
Exited	Customer churn	Categorical
Status	Customer status	Categorical
AssetStage	Asset stage	Categorical

For sample imbalance, this paper adopts SMOTE model oversampling to achieve sample balance: SMOTE method is a data preprocessing technology applied to imbalance proposed by Chawla et al[10]. Different from the mechanism of simple copying samples of random oversampling, SMOTE synthesizes new samples between two minority class samples through linear interpolation, thus effectively alleviating the overfitting caused by random oversampling.

In view of different magnitudes of metrics, the Z-Score method is utilized in this paper to standardize the training data, and testing data is standardized by means of the mean and variance of the training data.

3.2. Testing and Selection Machine Learning Classification Algorithm

Table 4. Comparison of the effectiveness of eight machine learning algorithms

Machine learning algorithms	Training set accuracy	Validation set accuracy
XGBoost	0.99	0.912
GBDT	0.951	0.901
LightGBM	0.926	0.898
CatBoost	0.897	0.87
adaboost	0.832	0.827
Random Forest	0.834	0.82
Decision Tree	0.813	0.796
ExtraTrees	0.801	0.788

This paper selects eight machine learning algorithms that are more in line with the research objectives, namely XGBoost, GBDT, LightGBM, CatBoost, adaboost, Random Forest, Decision Tree and ExtraTrees, and divides the training set and test set according to an 8:2 ratio to test the training effect. The comparison results of XGBoost with other models are shown in the Table 4. It can be found that XGBoost has achieved the best results in the training set and puts up the best performance in the test set among all test algorithms. Therefore, the XGBoost algorithm is selected to train the prediction model in this paper.

3.3. Evaluation of Prediction Models Based on XGBoost

This paper builds a long-term loyalty prediction model for bank customers based on 12 indicators to predict customer churn. The XGBoost machine learning algorithm is used to train classification models. The trained models have yielded constructive results. The characteristics and evaluation of specific models are as follows.

3.3.1. The Ranking of Feature Importance

The importance of the 12 features is ranked, and the results are shown in the Table 5.

Table 5. Feature importance

Dimension	Characteristic	Importance
Basic Information	Gender	2.6%
	Age	13.2%
	EstimatedSalary	2.9%
Account opening situation	CreditScore	2.1%
	Tenure	4.8%
	Balance	4.2%
Transaction behavior	NumOfProducts	23.9%
	IsActiveMember	21.8%
	HasCrCard	2.5%
Composite index	IsActiveStatus	5.4%
	IsActiveAssetStage	10.7%
	CrCardAssetStage	5.9%

It can be seen that the importance of the number of products purchased by customers (NumOfProducts) reaches 23.9%, ranking first place in importance, indicating that the number of products purchased by customers can reflect customer loyalty, and can also tell customer's satisfaction with and trust in the bank. In addition, customer activity status (IsActiveMember) and customer age (Age) are also important factors, which will prompt banks to accurately implement service strategies for different customer groups.

3.3.2. Confusion Matrix Analysis

In this paper, 9300 pieces of long-term customer resource data of a bank are selected as the original data set. The data were preprocessed according to the following principles: (1) samples with serious data gaps were removed; (2) eliminating outliers that are against objective reality.

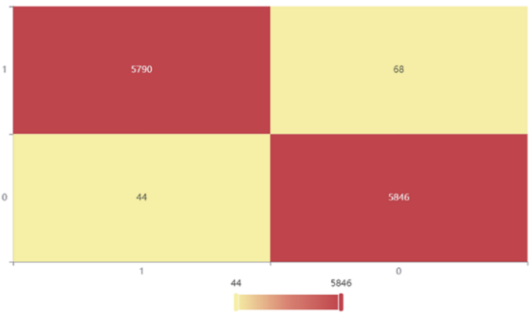

Figure 1. Confusion matrix heatmap on the training set

Figure 1 is the confusion matrix diagram of XGBoost on the training set. The results elaborate the judgment of customer churn. There are 11636 samples that are correctly classified by the model (quadrants 2 and 4), and 112 samples that are misclassified (quadrants 1 and 3).

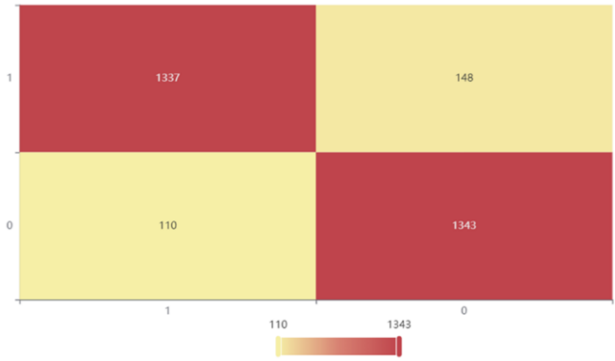

Figure 2. Confusion matrix heatmap on the test set

Figure 2 is the confusion matrix diagram of XGBoost on the test set. The results show that there are 2680 samples (quadrants 2 and 4) correctly classified by the model, and 258 samples (quadrants 1 and 3) that are misclassified.

3.3.3. Model Training Metric Evaluation

Table 6. Model training evaluation index values

	Accuracy	Recall	Precision	F1
Training set	0.99	0.99	0.99	0.99
Test set	0.912	0.912	0.912	0.912

It can be seen from Table 6 that the prediction model based on XGBoost has achieved good results on the test set with high accuracy, recall, precision, and F1-Score, and is suitable for predicting long-term loyalty of bank customers.

4. Conclusions

At present, many banks fail to conduct in-depth research and analysis of customer's actual needs and expectations. In order to improve customer loyalty to banks and bank's marketing volume, commercial banks are urgently needed to transfer their business concepts from a "product and sale-oriented" business model to a "customer-centric" one, thus solving problems brought by the diversity of individual customer needs and low loyalty and other issues, and thereby achieving win-win. Based on the research on long-term loyalty of bank customers, this paper constructs a customer loyalty evaluation system, divides customer groups, establishes a unified and effective long-term loyalty prediction model of bank customers based on XGBoost algorithm, and realizes optimization and innovation in this field.

Although advanced ensemble learning algorithms are applied to the research and ideas for building a long-term loyalty prediction model for bank customers are offered

in the paper, limitations remain salient: (1) subjectivity exists in constructing features in this paper whose new features can be constructed with automatic methods such as machine learning later on; (2) the user information collected in this paper is limited as the characteristics that affect bank customer churn are not limited to the factors listed in this paper. In the follow-up research, more user information must be deeply mined, so as to identify user characteristics in more details and provide the predictive effect of the model.

References

[1] XingRun mei. (2022). Based on the theory of service profit chain of commercial bank customer loyalty promotion strategy research in colleges and universities (master's degree thesis, nanjing university of posts and telecommunications). https://kns.cnki.net/kcms2/article/abstract?v=5faiAHckh4xgPgsMUlPmCKMjf0jDvKj4LCUOWiDpUgFGyCcHAugHfxGTxmn1P9H_3RTuWrOsieMZQ-lPrhWvwbc0ZGUD4XV9mGFBa3T8s48CPeBlzk5uNw==&uniplatform=NZKPT&language=CHS

[2] Solimun, Adji A R. (2018) The mediation effect of customer satisfaction in the relationship between service quality, service orientation, and marketing mix strategy to customer loyalty. Journal of Management Development,37(1):76-87.

[3] Gomezcruz M E.(2019)Electronic reference services: a quality and satisfaction evaluation. Reference Services Review, 47(2): 118-133.

[4] Teeroovengadum V. (2020) Service quality dimensions as predictors of customer satisfaction and loyalty in the banking industry industry: moderating effects of gender. European Business Review,4(2): 95-122

[5] Yu Chengwei. (2018). The hankou bank customer loyalty promotion countermeasures (a master's degree thesis, guangxi normal university). https://kns.cnki.net/kcms2/article/abstract?v=5faiAHckh4z_QQSKARgA0-S4fvDohPXcD_xHDohQ3sRw4XPNrd5A2A2PHSjpguYgJyy6t_Q5vccME406jGPdLH5miWBpibg65XE2cvo7t2vLDs4rZT3sRg==&uniplatform=NZKPT&language=C HS

[6] Jin Wenzhao. (2019) The influence of commercial banks' service quality on customer loyalty. Shopping Mall Modernization, 21(6):89-91.

[7] Zheng Han hui. (2019). China citic bank credit card center customer loyalty promotion countermeasures (a master's degree thesis, lanzhou university). https://kns.cnki.net/kcms2/article/abstract?v=5faiAHckh4xtUn13YdLEDYJ-c9m4qQbOn2sRs_J1aFGZ9MydXyd95XQqD-HqPZPWRbMTlIFMb3MvV-6IBTOBc0r-CaAKzEcXQq3gUawCjpkzIi07kvoH2A==&uniplatform=NZKPT&languag e=CHS

[8] Hu Zhiting. (2020) Review and future prospect of research on service atmosphere and customer satisfaction in the age of digital intelligence. Modern Business,(33):28-30.

[9] C CHEN T, HE T, BENESTY M.(2015)XGBoost：Extreme Gradient Boosting. R Package Version,0. 4-2,1(4):1-4.

[10] CHAWLA N V, BOWYER K W, HALL L O, et al. (2002)SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16(1): 321-357.