International Journal of Innovation in Science and Mathematics
Volume 12, Issue 6, ISSN (Online): 2347–9051

Research on Customer Bank Product Purchase Intention Prediction Based on XGboost Algorithm

Kecheng Ye
School of Computer Scienceg, Hangzhou Dianzi University Information Engineering College, Hangzhou, China.
Corresponding author email id: filetransfer1@163.com

Date of publication (dd/mm/yyyy): 30/11/2024

***Abstract* – With the rapid development of "Internet Plus" and the deepening application of big data technology, various industries have accumulated a large amount of user behavior information. Based on better meeting the needs of users, user-centric behavior analysis has become the focus of various industries, and providing good user experience services is the key to enterprise development. Traditional commercial banks have a large business volume and a large number of users. Banks expect to analyze the historical behavior of users and predict the banking products that users will buy in the future, so that they can adjust their business according to the needs of users and adapt to the trend of the development of Internet finance. Based on the above background, this paper researches and analyzes the historical behavioral data of bank users, and proposes a prediction model of customers' intention to purchase bank products based on integrated learning. Firstly, the original dataset is preprocessed to deal with the problems of missing data and anomalies, while the feature coding is applied to transform the non-numeric features. For feature engineering, embedded feature selection is used to make selection of features. Subsequently, the construction of the customer intention prediction model based on the XGboost model model is carried out and the model evaluation is conducted, and the analysis finds that all the indicators of XGboost are higher, indicating that the model has a better performance, and it can be used to accurately predict the customer's intention of purchasing the bank products, so as to improve the efficiency of the bank's work and the level of its business, and to provide a reference for the commercial banks when providing products to customers.**

***Keywords* – Bank Customer Behavior, Embedded Feature Selection, Model Comparison, XGboost.**

I. INTRODUCTION

In recent years, China's economic output has been rising, the per capita income of the residents has been rising, but the economic growth rate shows a downward trend year by year, the downward pressure on the economy is increasing, the financial and economic performance is not good and the price level is rising, the disposable cash in the hands of the residents continues to depreciate, and a variety of consumer spending is numerous, the residents face the pressure of all aspects of life, and they need to have a method of preserving the value of the cash. Bank financial products are favored by the public for their value preservation and appreciation. The development of China's financial industry started relatively late compared to the western developed countries, most residents in the financial investment knowledge and concepts are relatively backward, so the bank needs to analyze the customer's purchase intention, and make accurate predictions, which is conducive to promoting the sustainable development of the banking business [1].

In order to predict customers' needs and purchase intentions, banks need to have a large amount of customers' purchase history data and analyze it accurately. On the basis of the accurate analysis of customers' historical purchasing behavior, banks can design and recommend customers' purchasing behavior according to the advantages of their own financial products and services. When the staff negotiates with customers, the analysis results must be accurately displayed on the staff's computer or mobile device, which can greatly improve the efficiency of the bank staff and reduce the unnecessary waste of human resources. Therefore, it is necessary to

use effective methods to analyze data, establish mathematical models, and make accurate recommendations. Currently, the rapid development of big data, cloud computing and data processing technology, the use of related technology to track and capture the needs of customer activities, can maximize the approximation of customer demand, through the continuous integration of various types of resources to match the customer's operational activities, the next step of the customer's purchasing behavior to make accurate predictions, and then recommend to the customer the appropriate bank-related products, which is crucial to the development of the bank.

II. Related Work

Back in 2006, Long, from the perspective of information technology, under the support of the theory of financial innovation and bank informatization, combined with his own work practice, analyzes the great role of information technology in the bank's innovation of service content and improvement of service quality, focuses on the study of how to make full use of information technology to promote the innovation of bank products, and puts forward specific measures to suggest [2].

In 2007, Zhao takes commercial bank products as a clue, on the basis of a clear definition of modern commercial bank financial products, in accordance with the People's Bank of China on China's commercial bank business data statistical standards, with China Construction Bank, Bank of Communications and Chongqing Municipal Commercial Bank as the representatives of the state-owned four major commercial banks, joint-stock commercial banks and city commercial banks, carefully and deeply analyzed the current stage of China's commercial banks in the product marketing situation [2]. Commercial banks in product marketing in the current stage of the situation, and further pointed out the existence of major problems and the reasons for the problems, and ultimately to explore the ideas and specific strategies to solve the marketing problems of China's commercial banks [3].

In 2010, Chi tested the customer usability of product innovation, and found that banks can obtain the evaluation information of customers on their products in time and make effective adjustments to the products, so as to improve the competitiveness of the banks [4]. Zhao solved the problem of selecting strategies to build a sustainable product innovation capability in the environment based on the experience of product innovation of the international first-class banks, and established an evaluation index system of the organizational and management capability of product innovation suitable for China's banking industry [5]. Chen analyzes the risk of private banking business and its control, and proposes the use of fuzzy comprehensive evaluation to evaluate the operational risk of private banking business. At the same time, Bank of China Shenzhen Branch conducts a case study, including product innovation, risk control and other aspects, and puts forward suggestions for the development of its private banking business [6].

In 2013, Xiao based on the theory of banking product innovation, analyzed the current situation of product innovation in state-owned C commercial banks in detail, proposed the direction for the development of the bank's product innovation, and formed a basic management system framework and specific strategic recommendations [7]. Chen analyzed the factors affecting the sales of banks' financial products, understood the influence mechanism and intensity of these factors, and analyzed the results to guide banks to carry out their banking business more scientifically and rationally, and to promote the rapid and healthy development of the banking financial products market. Chen analyzes the factors affecting the sales of bank financial products

based on the key retail business indicators, customer demographic characteristics and customer financial asset characteristics, and understands the mechanism and intensity of the influence of these factors, and guides the bank to carry out the bank financial business more scientifically and rationally through the results of the analysis, so as to promote the rapid and healthy development of the market of bank financial products [8].

In 2019, Xiao used the public dataset of Banco Santander, Spain. Based on the structure of statistical analysis of data, and then feature extraction of data through feature engineering, the data features of this paper are analyzed and constructed from three levels of analysis, such as initial features, simple features and complex features. Further, the model is used to predict the purchasing tendency of bank customers, based on the historical behavior records of users' purchasing products in the previous 17 months, to predict the new purchasing of products by users in the same month, and to provide a decision-making basis for banks to recommend financial products to their customers [9]. Li predicts and analyzes the product marketing response through BP neural network, which guides the setup of marketing programs and improves the efficiency. The marketing data evaluated and predicted by the simulation model, compared with the actual data, the two sets of experimental errors are small, and the BP experimental results are feasible and save costs and reduce energy consumption. It shows that the new model prediction scheme and data processing method proposed by integrating artificial neural network prediction is feasible [10].

In 2020, Liu analyzed the macro-analysis environment of the political environment, economic environment, social environment and industry technology environment of domestic mobile banking based on PEST tools, and combined with the SWOT analysis of mobile banking products of Bank of Communications Gansu Branch to revise the market positioning. Thirdly, according to a series of marketing problems currently faced by Bank of Communications mobile banking products, through the analysis of the causes of the problems, find out the causes of the problems, and propose corresponding solutions to the problems currently faced[11].

In 2023, Zhou divided the development of intelligent investment banking products into two parts: product design and product marketing, and organized the core processes of the business from three aspects: service system, system development and core technology to form an overall intelligent investment banking product program. Finally, based on the business characteristics, JS Bank formulated the intelligent investment banking product development strategy, so as to enhance product competitiveness [12].

III. Empirical Analysis

A. *Data Source*

In this paper, the data of a commercial bank's customer data, the data recorded in the banking system of the customer's basic information, including: the customer's occupation (job), the customer's age (age), the customer's marital status (marriage), the customer whether there is a breach of contract (default), the customer whether the mortgage and other information (housing); at the same time, also saves the bank and the customer to contact the intensity of the data, including the number of times in this marketing campaign with the customer (campaign), the bank and the customer from the last marketing campaign to contact the length of time (duration), the bank and the customer from the last marketing campaign to contact the time (duration), the bank and the customer from the last marketing campaign to contact the time of day (duration). The intensity of the bank's contact with the customer, including the number of times the customer was contacted in the current

marketing campaign (campaign), the length of time the bank has been in contact with the customer since the last campaign (duration), and the time interval between the bank's contact with the customer and the customer's contact in the last campaign (p-days); in addition, there are the current market conditions: the rate of change in employment (emp_var_rate), Consumer price index (cons_price_index), interbank lending rate (lending_rate) and so on. The training dataset has 22,500 user records and the test dataset has 7,500 user records. Each row of data in the training data set records the information related to the purchase of each customer and the result of whether to buy bank products, while each row of data in the test data set records the personal information of each user and does not give the result of whether the customer buys bank products or not. Partial presentation of the raw data is shown in Table 1:

Table 1. Data source (First five lines).

Id	Age	Job	Marital	Education	...	nr_employed
1	51	admin.	Divorced	professional.course	...	5219.74
2	50	services	Married	high.school	...	4974.79
3	48	blue-collar	Divorced	basic.9y	...	5022.61
4	26	entrepreneur	Single	high.school	...	5222.87
5	45	admin.	Single	university.degree	...	4884.7

B. *Data Preprocessing*

In daily life, the data obtained is often not complete, there will be some missing values, and some features have a large difference between, and not stable enough to appear many outliers, at this time the data preprocessing is particularly important. Therefore, it is necessary to deal with this problem before modeling to ensure that the quality of the data can meet the task of data mining. In this paper, the work done in data cleaning has three parts: filling the missing values, exploring the outliers and transforming the data features.

The first is the treatment of missing values, there are many missing values in this dataset, in order to avoid the impact of missing values on the predictive performance of the model, it is necessary to fill the missing values. The approach taken in this paper is to exclude the feature if there are more missing values (more than 30% missing rate); if there are fewer missing values, the Lagrangian interpolation method is used to interpolate the missing data.

Then deal with outliers, through the data exploration found that some features contain some outliers, however, in the field of financial risk control, the outliers may represent some special cases, which can be regarded as a kind of risk, so this paper does not deal with the outliers.

Finally, the conversion of features, there are 11 categories of features in this dataset, which are 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'poutcome', ' subscribe', in order to better construct the measurement model, the category features need to be transformed, this paper uses label coding to encode the categorical variables. After all of them are completed,there are 21 features in the dataset.

C. *Feature Selection*

Feature selection is the process of finding and selecting the most useful features in a dataset and is a key step

in the machine learning process [13]. Not all of these features are useful for subsequent modeling, according to their role in modeling or not, the features can be divided into two categories of useful and useless, which can be called "relevant features" and "irrelevant features". The process of selecting relevant features from these features is called feature selection. Feature selection before modeling can not only reduce the dimensionality of the data, reduce the processing time of the useless features, which reduces the complexity of the computation time, but also improve the credibility of the model, which can fully understand the information implied in the features.

In this paper, we mainly use embedded feature selection for factor selection, embedded feature selection utilizes the characteristics of the machine learning algorithm itself to perform feature selection, which is very effective for high dimensional datasets. In embedded feature selection, feature selection and model training are carried out simultaneously, i.e., it integrates the feature selection and the learner training process, both of which are completed in the same optimization process, and feature selection is carried out automatically during the learner training process [14]. Some machine learning methods inherently have mechanisms for scoring features or can easily apply them to the feature selection task, which can make feature selection more efficient and accurate. Embedded feature selection is based on the principle of constraining the complexity of the model through regularization methods so that the model automatically selects important features. Commonly used methods include:

(1) Penalty term based feature selection method, that is, with regularization, such as L1,L2 paradigm, L1 regularization will make part of the weight of the features become 0, so as to achieve the effect of feature selection. L2 regularization will make the weight of all features become smaller, but not equal to 0. It is mainly used in algorithms such as linear regression, logistic regression and support vector machine (SVM);

(2) Tree model-based feature selection method, i.e., using decision tree ideas, including Decision Tree, Random Forest, Gradient Boosting and so on.

Considering the many features of the sample, the model can easily fall into overfitting. In order to alleviate the overfitting problem, then the regularization term needs to be introduced. The larger the regularization term is, the simpler the model is and the smaller the coefficients are. When the regularization term increases to a certain degree, all the feature coefficients will tend to 0. In this process, there will be a part of the features whose coefficients will become 0 first. Which also realizes the feature selection process.

Logistic regression, SVM, and decision trees can be used as base learners for regularized feature selection, and only algorithms that can obtain feature coefficients or can obtain feature importance can be used as base learners for embedded selection. For SVM and logistic regression, the parameter C controls the degree of sparsity, and the smaller C is, the fewer features are selected. In this paper, we use logistic regression and SVM as the base learner for selecting features, and use L2-paradigm regularization for factor screening to retain the factors left by both models.

Table 2. Number of factors after feature selection.

Base Classifier	Number of Retained Factors
Logical regression	7
SVM	8
Logistic regression & SVM	7

After data cleaning and feature selection, a total of seven most valuable features were retained from the original features for building the subsequent prediction model.

D. *Model Construction*

The bank product purchase intention studied in this paper is a dichotomous problem, i.e., two outcomes, purchase and no purchase. Therefore, five commonly used evaluation metrics for classification algorithms, namely Accuracy, Precision, Recall, Fl-score and AUC, are used to evaluate the classification effect of the model.

In the previous data preprocessing process, it was found that the dataset was unbalanced (the ratio of positive examples to negative examples was about 4:1), therefore, it was necessary to carry out the balancing process before model training, and this paper adopts the SMOTE algorithm to make balancing process for the dataset, which avoids the problem of affecting the results of model training due to the imbalance of the data. The oversampled data are divided according to the ratio of 4:1, which are used for model training and prediction respectively.

Firstly, weak classifiers like decision tree and logistic regression are utilized for model construction, and it is found that the classification effect is poor. Therefore, we consider integrated algorithms, respectively build decision tree-based bagging and random forest in integrated learning bagging algorithm, and AdaBoost, XGboost and Gradient Boosting Classifier in boosting algorithm, respectively build the models under the default parameters, and compare the models according to the evaluation indexes. Effects, as shown in the Table 3, to comprehensively analyze and select the appropriate model as the prediction model.

Table 3. Effect comparison of model.

Type	Classifier	Accuracy	Precision	Recall Rate	F1-Score	AUC
Weak Classifier	SVM (linear)	0.6884	0.6561	0.7817	0.7134	0.7134
	SVM (SGD)	0.7039	0.6858	0.7439	0.7137	0.7137
	Native Bayesian	0.6876	0.6811	0.6966	0.6887	0.6887
	LogisticRegression	0.7117	0.6991	0.7355	0.7168	0.7168
	DecisionTree	0.8269	0.8646	0.7720	0.8157	0.8157
	KNN	0.8473	0.7856	0.9523	0.8609	0.8609
Bagging	Bagging	0.9163	0.9306	0.8983	0.9142	0.9142
	Random Forest	0.8852	0.8735	0.8987	0.8859	0.8859
Boosting	AdaBoost	0.9213	0.9246	0.9161	0.9203	0.9203
	XGboost	0.9333	0.9518	0.9118	0.9314	0.9314
	Gradient Boosting Classifier	0.8924	0.8767	0.9114	0.8937	0.8937

Since this paper studies the purchase of banking products by customers, and the purpose is to be able to predict the users who have the intention to purchase banking products, the indicator of accuracy rate, although it can represent the overall correct rate and has some reference value, is not an appropriate indicator for what is studied in this paper, so it is not considered for the time being for the time being. For the rest of the indicators,

the recall rate and precision rate is the relationship between the two, the recall rate can be understood as the proportion of actual purchase customers can be predicted by the model, the precision rate refers to the proportion of customers predicted by the model as purchase customers actually make a purchase, often we pay more attention to the intention to buy customers can be identified, if the model to identify the intention to buy the user is very accurate, then it can be a better recommendation of the bank's products in order to obtain a better return, so in the precision rate, recall rate and precision rate, we can not consider this indicator for the time being. Therefore, among the three metrics of precision rate, recall rate and F1 score, the precision rate is considered relatively more important and is the metric to focus on. And AUC value tends not to be affected when the proportion of positive and negative samples changes, while other indicators will be affected more, so AUC value should also be focused on.

From the data in the table, it can be seen that in weak classifiers, KNN works better with the highest return rate and AUC value, indicating that in weak classifiers, KNN has the best prediction effect. In Bagging, Bagging based on decision tree is better than Random Forest, and in Boosting, XGboost is better than AdaBoost and Gradient Boosting Classifier. Taken as a whole, the best performer among all the models in terms of recall and AUC is XGboost. XGboost. Therefore, in this paper, XGboost is finally used for the construction of bank customers' intention to buy products prediction model.

XGboost is a gradient boosting algorithm, residual decision tree, whose basic idea is to gradually add to the model one tree at a time, and to make the overall effect (with a decrease in the objective function) improve with each addition of a CRAT decision tree [15]. Multiple decision trees (multiple single weak classifiers) are used to form a combined classifier, and each time a new decision tree is constructed, the weights of the new decision tree are obtained by optimizing the loss function based on the residuals of the decision trees that have already been built, using a second-order Taylor expansion. It is fast to train because of the automatic use of CPU's multithreading for parallel computation, and also accuracy improvement in algorithmic accuracy is carried out.

The accuracy of the model obtained based on the default parameters is 93.33%. On this basis the grid search of sklearn is utilized for parameter adjustment. The specific process is as follows:

Step 1: Determine the number of estimators for learning rate and tree_based parameter tuning

max_depth is usually chosen between 3 and 10, with a starting value of 5.

min_child_weight picks a relatively small value, starting at 1.

Initial gamma=0.

subsample, colsample_bytree is generally taken as 0.5-0.9, with a starting value of 0.8.

scale_pos_weight=0.

First find the optimal number of decision trees needed based on the default value of learning rate of 0.1, in increments of 100, measured from 100 to 900. get the optimal n_estimators as 800.

Step 2: Tuning of max_depth and min_child_weight parameters

The two parameters are tuned first, as they have a big impact on the final result. The parameters are first coarsely tuned from a large range and then fine tuned from a small range. After outputting the results, the optim-

-al max_depth of 7 and min_child_weight of 1 are obtained.

Step 3: Tuning of gamma parameters

Search for gamma from [0,0.1,0.2,0.3,0.4] and get the optimal gamma as 0.1.

Step 4: Adjust the subsample and colsample_bytree parameters

The subsample and colsample_bytree parameters are incremented from 0.1 to 0.6 to 0.9. The ideal values of the subsample and colsample_bytree parameters are 0.8, 0.8. We then take values around this value in steps of 0.05. The ideal values of the subsample and colsample_bytree parameters are still 0.8, 0.8. _bytree parameters still have ideal values of 0.8, 0.8. then use them as the final ideal values.

Step 5: Regularization parameter tuning

The next step is to reduce the overfitting by regularization, here the tuning reg_alpha parameter is used. Here the ideal value is taken as 0.1.

Step 6: Reduced Learning Efficiency

Shrinking the learning rate by a factor of ten to 0.001 and increasing the number of trees to 5000 increased the score.

The final tuning results are shown in the Table 4:

Table 4. Table of tuning results.

Parameter Abbreviation	Parameter Meaning	Initial Value Parameter	Adjustment Results
max_depth	Maximum depth	None	9
min_child_weight	Decision tree	1	1
gamma	Number of	0	0.2
subsample	Minimum sample weight of leaf nodes	1	0.8
colsample_bytree	Random sample	1	0.9
reg_alpha	Spanning Tree Column Sampling	0	0.1
learning_rate	L1 of weights	0.1	0.001
n_estimators	Regularization term	60	5000

Finally this paper uses accuracy, precision, recall and F1-score to evaluate the model of XGboost model with tuned parameters and the evaluation results are shown in the Table 5.

Table 5. Table of tuning results.

Classification Report	Precision	Recall	F1-Score
Not Default	0.93	0.96	0.94
Default	0.95	0.93	0.94
accuracy			0.94
macro avg	0.94	0.94	0.94
weighted avg	0.94	0.94	0.94

The accuracy, precision, recall, F1-score and AUC values reach 94%, which indicates that the model performs better and can be used to identify bank customers with purchase intention. Thus, it can be assumed that the integrated learning algorithm improves the predictive ability of the model. In addition integrated learning has a strong advantage in comparison to the stability of the model, from which it can be inferred that the XGboost model has a huge advantage in identifying bank customers with purchase intention.

IV. CONCLUSION

With the advent of the big data era, more and more user data are collected and integrated by the major industries, these data are valuable resources, big data technology has made some achievements in some fields. At present, commercial banks have a lot of customer resources, and have been focusing on the use of big data technology in the bank's business activities. The research object of this paper is the purchase behavior of bank customers, using the user's past purchase behavior data to analyze the possible purchase behavior of future users, so as to provide technical support for the development of bank products and business. In this paper, we preprocess the historical behavior data of bank customers and use embedded feature selection for factor screening, and by comparing multiple machine learning models, we finally establish the XGBoost model in integrated learning for the prediction of bank customers' intention to purchase products, and use the corresponding indexes to evaluate the effect of the model, and get the following conclusions:

(1) In the exploratory analysis and feature engineering of the data, it is understood that the dataset used in this paper is an unbalanced dataset (the ratio of users who buy products to those who don't is about 1:4), and subsequently a total of 7 features are retained after the embedded feature selection, which is filtered out with a high quality of the features.

(2) When comparing the modeling effect of weak classifier and integrated model, it can be found that the predictive ability of weak classifier is poor and unstable when the data volume is large, while the integrated model performs better, and the model is found to be in the optimal state after integrating various indexes, which determines that the integrated learning model is more superior in predicting the customers' intention to purchase products.

(3) Since buyers generally account for a relatively small percentage of the actual bank product purchases, i.e., there is a positive and negative sample imbalance in the dataset, at which time the accuracy, precision, and F1 score are affected, while the recall and AUC values are almost unaffected, the XGboost model is chosen as the final prediction model for the bank customer's intention to purchase a product, which is used for predicting whether or not the customer will purchase the bank products.

1. In summary, the XGboost-based prediction model of bank customers' intention to purchase products constructed in this paper can provide a reference for commercial banks when offering banking products to their customers.

REFERENCES

[1] Shen Gang. Reflections on AI-assisted banking product marketing [J]. Agricultural Bank of China, 2024, (04): 8-12. DOI: 10.16678.
[2] Long Fan. Research on product innovation of commercial banks based on information technology [D]. Chongqing University, 2006.
[3] Zhao X. Marketing Research on Commercial Bank Products [D]. Southwest University, 2007.
[4] CHI Guotai, ZHAO Zhihong, LI Zhanjiang. A customer usability testing model for bank product innovation and its empirical study [J]. Investment Research, 2010, (04): 10-13.
[5] ZHAO Zhihong. Research on the evaluation model of bank product innovation ability [D]. Dalian University of Technology, 2011.
[6] Chen Qi. Research on the Development of Private Banking in China [D]. Central South University, 2010.

[7] Xiao Jin. Research on product innovation management strategy of state-owned C commercial banks [D]. Nanchang University, 2013.
[8] Chen Liangkai. An empirical study on the factors influencing the sales of financial products in commercial banks [D]. Southwest Jiaotong University, 2014.
[9] Xiao Li. Research on the prediction model of bank customers' intention to purchase products [D]. Zhongnan University of Economics and Law, 2019.
[10] Li Tingting. Research on bank product marketing prediction model based on BP neural network method [J]. Modern Marketing (Lower Decade), 2019, (10): 92-93.
[11] Zuo Chengkun. Research on the prediction model of bank customers' willingness to purchase products [D]. Anhui University, 2023. DOI: 10.26917.
[12] Zhou Yan. Research on product development of bank intelligent investment advisor [D]. Southeast University,2023.
[13] Hong Zhou, Yang Gang, Yang Jinsong, etc. Feature selection of label library for joint filtering and embedded samples [J]. Electronic Design Engineering, 2024, 32(22): 146-150.DOI: 10.14022.
[14] Wu Xiaojun, Zhou Wenxin, Dong Yongxin. An improved embedded feature selection algorithm and its application [J]. Journal of Tongji University (Natural Science Edition), 2022, 50(02): 153-159.
[15] Zhao Peng, Wang Wenjian, Wu Di, etc. Blockchain abnormal transaction detection based on XGBOOST and random forest [J / OL]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 1-8[2024-11-22].

AUTHOR'S PROFILE

Kecheng Ye, School of Computer Scienceg, Hangzhou Dianzi University Information Engineering College, Hangzhou, China.