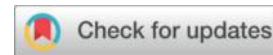


Research on credit card transaction risk prediction model based on multi-source data fusion and its application

Ziang Qi

Fuqua School of Business, Duke University, Durham, NC, United States

Email :ziang.qi@outlook.com



Abstract: In this paper, the credit card risk control and early warning model is constructed to predict and control the credit card risk by fusing multi-source data through data mining technology. After analyzing the theoretical application of XGBoost algorithm on credit card transaction risk, its model parameters are optimized by particle swarm algorithm to construct PSO-XGBoost credit card transaction risk prediction model. The credit card transaction risk prediction performance of the PSO-XGBoost model is verified and applied to the abnormal transaction risk assessment of Bank A. The AUC value, accuracy, F1 value, precision rate, and recall rate of the PSO-XGBoost model are the largest among all the algorithms, and the correct detection rate of the PSO-XGBoost model is significantly higher than that of other algorithms, and the error detection rate significantly lower than other algorithms, with the best risk prediction performance. Among the 10 Bank A credit card customers, customer 4 and customer 5 have very high risk in their credit card transactions, customer 2 and customer 3 have high risk, customer 1, 5, 7 and 8 have medium risk in their credit card transactions, and customer 9 and customer 10 have low risk.

Keywords: multi-source data fusion; data mining; PSO-XGBoost; risk prediction; credit card transactions

1. Introduction

In recent years, with the deepening of the financial system reform and the gradual liberalization of the financial system, the competition in the financial market has become more and more intense, and a variety of advanced service tools and service means in the western financial management have been introduced to mainland China [1-3]. As a new credit payment tool in the financial field, the use of credit card

transactions is a simple credit service as a method of non-cash payments and transactions [4-5]. Through banks or credit card companies in accordance with the credit and financial payment degree of the user's user card to the user-cardholder, the cardholder can use the credit card to pay for cash that does not need to be paid [6-8]. Credit cards in circulation in China started in this century, especially in recent years, the development is very fast, both in terms of the number of cards issued, the number of authorized merchants, the number of ATM/POS and the number of agent outlets have a very large increase, but the average card deposit balance, the average card transfer amount, the average card consumption is a declining trend, and the risk of card transactions has become more and more prominent day by day [9-12]. Therefore, the application of predictive modeling of credit card transaction risk has become a development trend that credit card business must pay attention to [13].

A predictive model is a model that can predict future events by analyzing historical data [14-15]. In credit card transactions, predictive modeling can be used to identify risks and predict the probability of fraudulent transactions [16]. Common credit card fraud prediction models include logistic regression models, neural network models, and multi-source data fusion models [17-18]. Each of these models has its advantages and shortcomings, and banks can choose the suitable model for prediction according to their actual needs [19-20]. However, no matter which model is used, it must rely on a large amount of historical data for training in order to get accurate prediction results [21].

The article uses data mining techniques in credit card risk prediction and management to design credit card risk control model and credit card risk warning model. Then, XGBoost algorithm is used as a benchmark model to explore the theoretical feasibility of XGBoost algorithm in credit card transaction risk prediction, optimize the parameters of XGBoost model by particle swarm optimization algorithm, and construct XGBoost credit card transaction risk prediction model based on particle swarm optimization (PSO-XGBoost). The PSO-XGBoost model is juxtaposed with other methods to verify the prediction performance of the PSO-XGBoost model on credit card transaction risk by means of precision rate, recall rate, F1 value, accuracy rate, AUC value, correct detection rate, and incorrect detection rate. Finally, the PSO-

XGBoost model is utilized to detect the transaction risk level of credit card customers of Bank A.

2. Credit card transaction risk management based on data mining

2.1 Application of data mining in credit card risk prediction and control

Credit card business is a high investment, high risk, high return business [22], the key is to be able to effectively control the risk and reduce the operating costs of the enterprise, so in this direction, the application of data mining is effective and is a major trend in the application of bank credit card business.

Fraudulent behavior often occurs in the credit card business, such as credit card fraud, malicious overdrafts and other situations, these commercial banks and business caused huge losses. If you want to predict such fraudulent behavior, even if the prediction accuracy is low, it can reduce the chances of fraud opportunities to occur, thus reducing the loss. The relationship between fraud identification is mainly through the relationship between normal behavior and fraud, to derive some of the characteristics of fraudulent behavior, such as business based on these characteristics, can alert policy makers.

Data mining techniques can automatically predict trends and behaviors [23]. Data mining techniques can automatically predictive information found in the bank's database, built risk prediction models to help the bank to detect potential incidents of fraudulent transactions. Fraud detection can be done by cluster analysis of data. Since fraudulent payments usually have some unique functional characteristics such as multiple payments in a short period of time, mostly occurring in the new business of processing credit cards, the analysis can be done by dividing the data into a group of possible fraud's and unlikely data for cluster analysis.

2.2 Credit card risk control model design

By establishing predictive or descriptive models, the work related to credit card

risk management can be accomplished more effectively. Credit card risk control models mainly include dynamic credit limit adjustment models (evaluation of customers' current credit rating), credit scoring models (evaluation of customers' initial credit rating) and fraud monitoring models (analysis of abnormal customer spending).

1) Credit scoring model

The credit scoring model is to characterize the taste distribution of the emergent credit customers with the basic information of existing customers, relevant historical transaction data [24]. At the same time, the credit scoring model predicts the basic information about applying for new credit card customers to fill the score and determine the initial level of new customers.

In the actual algorithm, if Y is a dichotomous response variable, i.e., $Y = 0$ or 1 , and X belongs to a P -dimension explanatory variable, it is derived from Logistic:

$$\pi(x) = P(Y = 1) = \frac{\exp\left[\beta_0 + \sum_{i=1}^p \beta_i X_i\right]}{1 + \exp\left[\beta_0 + \sum_{i=1}^p \beta_i X_i\right]} \quad (1)$$

Where $\pi(x)$ belongs to the probability of the occurrence of $Y = 1$ when defined in terms of $X = x$. According to the characteristics of Logistics regression, it is possible to divide the process of processing into the following three aspects:

(1) Firstly, the distinctive feature field in the data corresponding to credit card is determined by the data mining method as a factor, and if the factor $X_i = 1$ indicates the occurrence of the risk factor i .

(2) Response variables are defined in terms of $y=1$ card issuance (low-risk customers) and $y=0$ card rejection (high-risk customers).

(3) In learning, all factor coefficients (β_i) are determined to give the corresponding model. The result of this calculation can be used as a reference information for the bank to issue credit cards, and can also be used to define the criteria for granting credit limits to customers, e.g., for $\pi(x)$ for the customer to be granted a card, and then according to the actual value of $\pi(x) > 0.5$ customers $\pi(x)$ to give a

different credit rating, this categorization scheme is both secure and positive.

2) Credit limit dynamic adjustment model

Credit limit dynamic adjustment model is mainly to solve the credit limit of credit card holders to grade and repayment delays in collection and other issues. That is, when a customer repayment delay, it must need to be collected, such as collection, collection efforts should be how to control. How should a bank adjust credit limits when a customer often does not have enough credit or can only use a small portion of it.

A dynamic adjustment model with customer past behavioral data and legal credit to determine the probability of future customer credit quality. This probability will be on the creditworthiness of customers with good credit, and delays cannot be changed immediately, even if the collection just reminds the customer not to call immediately. It not only maintains a positive relationship with customers, but also reduces operating costs.

Because the customer's consumer behavior is constantly changing, so the credit rating should be changed accordingly. The analysis of basic information about the customer, transaction data and other meaningful data is analyzed to characterize the distribution of the current credit rating and to predict the trend of credit risk management in the next stage of the customer's development, the cycle is usually six months, one year, or when the trend changes from time to time to make adjustments, and, in accordance with the predicted trend to adjust the credit rating of the credit accordingly.

2.3 Credit card risk early warning model design

Credit risk early warning models are fraud detection models. Since credit card transactions knot no password, than a debit card, credit card fraud is more likely, credit card fraud losses account for a large part of the cost of business, so foreign banks or credit card companies commonly used fraud monitoring model to monitor each transaction to minimize fraud losses.

Fraud modeling is based on the use of probability of the probability distribution of the main detection methods: the principle of small probability event detection of

very small, determine the distribution function of the number of credit card overdrafts and repayments, the boundary value of the small probability event, usually will be the probability value of 0.05. Calculate the formula is as follows:

$$\int_{\lambda_2}^{\lambda_1} f(x)dx = 0.95 \quad (2)$$

The value of λ_1 , λ_2 can be calculated. The formula $F(x)$ is the distribution function of a random variable, which is generally capable of the number of overdrafts and repayments or the amount of overdrafts and repayments. The difficulty lies in using this method in order to determine the distribution of the random variable, in order to not very precise conditions, the exponential or normal distribution is desirable. This method is usually used for long-term monitoring of unusual trading behavior. When the value of the random variable of the monitored account is in the interval $[\lambda_1, \lambda_2]$, an early warning is given to the credit card account.

3. Credit card transaction risk prediction model based on XGBoost

3.1 Theoretical feasibility analysis of the XGBoost model

The core idea of XGBoost is that when predicting the risk of credit card transactions [25], weaker performing models such as decision trees may not give good predictions, for which improvements are needed in two ways. Firstly, a set of decision trees is generated [26], and secondly, XGBoost is iteratively boosted against a set of decision trees to improve the prediction of credit card transaction risk based on the gradient direction in each round of iteration. In the final generation of the prediction of credit card transaction risk, XGBoost uses equation (3) to integrate the combination of credit card transaction decision trees generated during the iteration process:

$$y = \sum_{m=1}^M g_m(x) \quad (3)$$

Where $g_m(x) = \sum_{j=1}^J r_j I(x \in R_j)$, represents the prediction results of different

decision trees combined according to the weights r_j , I represents the output of indication to the prediction results of the decision trees, such as predicting that a credit card default will occur when Output 1, otherwise output 0. During the iteration process, XGBoost uses equation (4) to update the weights of each decision tree's prediction of credit card transaction risk:

$$h_k(x) = h_{k-1}(x) + \eta \left(\sum_{j=1}^{J_k} \gamma_{jk} I(x \in R_{jk}) \right) \quad (4)$$

where η is used to control the magnitude of the update in the iteration, which is used to control the accuracy of the fit to the prediction of credit card transaction risk.

3.2 XGBoost model for credit card transaction risk prediction

In this paper, in order to facilitate the analysis of the credit card transaction risk prediction problem, each credit card transaction is classified into normal transactions as well as abnormal transactions, therefore, the problem is a binary prediction problem, and equation (5) is used to represent the credit card transaction risk prediction problem in the XGBoost model:

$$L(y_i, \hat{y}_i) = 2 \sum_{i=1}^q \log(1 + \exp(-2y_i p_i)) \quad (5)$$

Where $\{(x_i, y_i)\}$ represents the sample combination of credit cards, $x_i = (x_{1i}, x_{2i}, \dots, x_{qi})$ represent the different indicators in the sample. Eq. (6) is used to represent the XGBoost for credit card transaction risk prediction problem, where K denotes the number of decision trees in XGBoost:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (6)$$

Where f_k represents the prediction function finally produced by the k th decision tree, the prediction objective of XGBoost for credit card transaction risk can be expressed as equation (7):

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (7)$$

Eq. (7) is mainly used to represent the risk prediction results produced by different decision trees when risk is generated for credit card transactions, and the weight of each decision tree's prediction on the final prediction results, and Eq. (8) is used to perform multiple rounds of iterations to iteratively update the objective prediction function of Eq. (7) in the direction of the gradient, in order to improve the prediction effect of the credit card risk:

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_i(x_i)) + \Omega(f_t) + \text{const} \quad (8)$$

where $\hat{y}_i^t$ represents the prediction of the risk state of sample i during the iterative computation of the t round. In order to improve the accuracy of iterative updating, the above objective function is expanded according to equation (9):

$$Obj^{(t)} = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (9)$$

Where $f_t(x) = \omega_{q(x)}$, $\omega \in R^t$, $q: R^d \rightarrow \{1, 2, 3, \dots, T\}$, ω represents the proportion of leaf nodes in the decision tree in XGBoost, and q represents the decision tree structure in XGBoost. In order to improve XGBoost's ability to predict credit card risk out-of-sample, the regular term is introduced: $\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2$, where T is the number of leaf nodes in the decision tree. XGBoost selects different metrics to be used to generate the branching structure when predicting the risk of credit card transactions, and the metrics need to be partitioned, which can be represented by $I_j = \{i \mid q(x_i) = j\}$, which is introduced and then transforms Eq. (9) into Eq. (10):

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^n \left[g_i \omega_{q(x_i)} + \frac{1}{2} h_i \omega_{q(x_i)}^2 \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \\ &= \sum_{i=1}^n \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \end{aligned} \quad (10)$$

Assuming, $G_j = \sum_{i \in I_j} g_i$, $H_j = \sum_{i \in I_j} h_i$, in order to solve for the optimal value of the objective function, the calculation is done by taking the derivative of ω_j by taking the

derivative of ω_j to obtain equation (11):

$$\omega_j^* = -\frac{G_i}{H_j + \lambda} \quad (11)$$

Based on the above calculation process, the objective function for predicting the risk of credit card transactions in XGBoost can be transformed into Equation (12):

$$Obj = -\frac{1}{2} \sum_{j=1}^T \frac{G_i}{H_j + \lambda} + \gamma T \quad (12)$$

In XGBoost's prediction of credit card transaction risk, equation (13) is used to calculate the information gain of different credit card metrics for risk prediction, and the node with the highest information gain is selected for branching through the information gain in constructing the structure of XGBoost's decision tree:

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \lambda \quad (13)$$

In the final XGBoost structure, the model can judge the influence weight of the indicator on the prediction of credit card transaction risk according to the size of the value of Gain of each indicator, which helps the credit card business personnel to analyze the business more intuitively according to the indicator situation.

3.3 Optimization of XGBoost model parameters based on particle swarm algorithm

From the analysis of the risk prediction process of XGBoost for credit card transactions, it can be seen that the important factors that have an impact on the final XGBoost objective function include H_j , λ , γ , T and other parameters, where H_j represents the depth of the decision tree in the XGBoost model, and λ Represents the magnitude of the improvement of the accuracy of the objective function by XGBoost in each iteration, with γ representing the weight of the regular term and T representing the number of decision trees in XGBoost. By controlling these parameters, the effect of XGBoost on credit card default will change, and how to choose the value of the parameters to achieve the optimal prediction ability of XGBoost on

credit card risk belongs to the problem of parameter optimization. The common algorithms for constant optimization can be chosen from particle swarm algorithm, genetic algorithm and so on. From the comparison of these two algorithms, the particle swarm algorithm is more conducive to solving to get the global optimum, rather than falling into the local optimum, and the calculation speed is faster.

For this reason, this paper is based on the particle swarm algorithm [27], to construct the parameter optimization of the XGBoost model for credit card risk prediction, the particle swarm algorithm randomly initializes each parameter such as H_j , λ , γ , and T , and generates a group of particle swarms, and then adopts a certain group of parameters in the particle swarms to be applied in the XGBoost model to model the samples, use the AUC value to estimate the prediction effect of the model, and sort the particles according to the AUC value, in which G_{best} is used to represent the parameter corresponding to the particle with the best prediction effect, and at the same time, P_{best} is used to represent the parameter corresponding to the same particle that achieves the best effect in the iterative process. Parameters. The process of solving the optimal parameters of XGBoost using the particle swarm algorithm is to update G_{best} as well as P_{best} through a series of iterations to gradually narrow the prediction error, and ultimately solve the optimal parameters. The specific optimization process is shown in Figure 1.

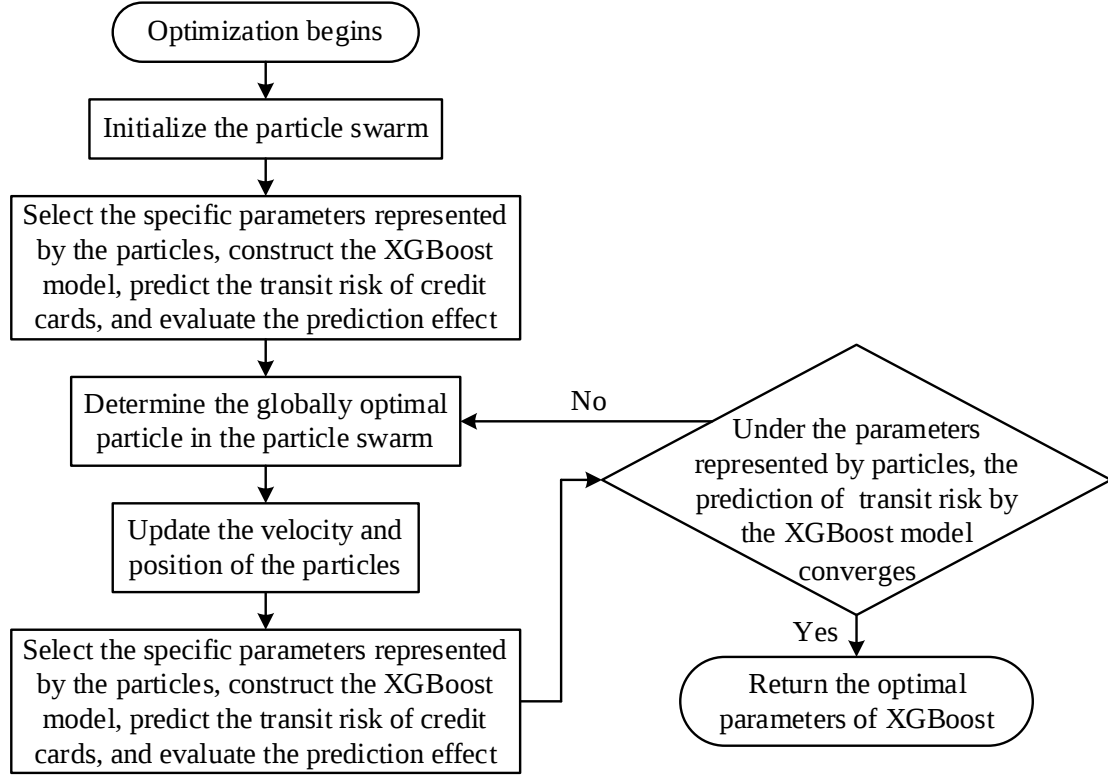


Figure 1 The process of particle group optimizing the XGBoost model

First, initialize each parameter, this paper chooses the particle swarm size of 1000, iterative calculation 3000 times, while randomly initializing the parameters of the XGBoost model such as H_j , λ , γ , T and so on. Second, the best measurement model for credit card default risk is selected based on the AUC predicted by the model. Third, calculate the AUC value predicted by the model under the corresponding parameters of each particle in the particle swarm, and record the AUC calculation results of each model. Fourth, the correlation value of the particle swarm is constantly updated according to the following formula, and in this paper, the size of c_1 and c_2 is chosen to be 0.5. Finally, when the result of the improvement in the effect of the model's AUC prediction is less than 0.01 or the number of iterations reaches 3,000, the optimization process is ended, and the optimal parameters are returned:

$$v_i = w \times v_i + c_1 \times rand \times (P_{best}(i) - x_i) + c_2 \times rand \times (G_{best} - x_i) \quad (14)$$

$$x_{i+1} = x_i + v_{i+1} \quad (15)$$

4. Risk prediction analysis

4.1 Predictive performance analysis

In this paper, we introduce a total of five metrics, precision rate, recall rate, F1 value, accuracy rate, and AUC value, to judge the classification effect. Precision rate measures the proportion of true positives out of the number of predicted positives. The recall rate measures the proportion of correctly predicted positives to all positives. The F1 value is the reconciled average of the precision rate and the recall rate, which measures the model's ability to combine the precision of the positive predictions and the ability to find positives. Accuracy is the probability of correct prediction. The AUC value measures the probability that the model predicts a true positive or true negative.

The experimental data were selected as 70% of the samples as training set and the rest as test set. In this paper, Logit, Gaussian Plain Bayes (GNB), Decision Tree (DT), Random Forest (RF), AdaBoost, ID3, and BP Neural Networks are used to conduct experiments along with this paper's XGBoost Credit Card Transaction Risk Prediction Model based on Particle Swarm Algorithm respectively, which is fitted on a balanced dataset (balanced with a sample ratio of normal to fraud class of 1:1.), and on the test set for classification, and the experimental results are shown in Table 1.

After analyzing the experimental results in Table 1, it can be found that the AUC value, accuracy, F1 value, precision, and recall of the XGBoost model based on particle swarm optimization in this paper are 0.9959, 0.9974, 0.9082, 0.8998, and 1.0000, respectively, which are the best results of all the prediction models, and the prediction performance of the PSO-XGBoost model is the best result among all the prediction methods and has a relative advantage in the test index.

Table 1 Experimental results

Method	AUC	Accuracy	F1	Precision	Recall
Logit	0.5989	0.9845	0.8755	0.8739	0.6106
Logit+Balance	0.5732	0.9729	0.8705	0.8475	0.6019
Logit+SMOTE	0.6464	0.9247	0.8375	0.8602	0.6061
Logit+ROS	0.6703	0.9716	0.8777	0.8186	0.6811
GNB	0.9355	0.9299	0.8314	0.8431	0.6855
GNB+Balance	0.9519	0.8654	0.8416	0.8563	0.6135

GNB+SMOTE	0.9182	0.8889	0.8758	0.8366	0.6787
GNB+ROS	0.8867	0.9398	0.8378	0.8828	0.6799
DT	0.5641	0.9146	0.7636	0.8865	0.7341
DT+Balance	0.6602	0.8931	0.8308	0.8784	0.7921
DT+SMOTE	0.6482	0.8433	0.8074	0.8798	0.7496
DT+ROS	0.6596	0.8705	0.8144	0.8584	0.6897
RF	0.9414	0.9316	0.8649	0.8171	0.8233
RF+Balance	0.8971	0.9529	0.8801	0.7901	0.8035
RF+SMOTE	0.9013	0.9347	0.8733	0.8034	0.7767
RF+ROS	0.8502	0.8856	0.8292	0.8644	0.8136
AdaBoost	0.8132	0.9478	0.8094	0.8199	0.7506
AdaBoost+Balance	0.8089	0.9616	0.7336	0.8359	0.7633
AdaBoost+SMOTE	0.8338	0.9039	0.8242	0.7979	0.6728
AdaBoost+ROS	0.9314	0.9231	0.7785	0.8104	0.6219
ID3	0.8175	0.8895	0.8106	0.8555	0.8034
ID3+Balance	0.7755	0.8865	0.8243	0.8123	0.7807
ID3+SMOTE	0.7295	0.9107	0.7846	0.8488	0.7593
ID3+ROS	0.7707	0.9395	0.8395	0.8142	0.7829
BP	0.7908	0.8995	0.7756	0.7954	0.6833
BP+Balance	0.8151	0.8759	0.8246	0.8083	0.6195
BP+SMOTE	0.8541	0.9074	0.7682	0.8303	0.5967
BP+ROS	0.8505	0.9253	0.8169	0.8196	0.6711
PSO-XGBoost	0.9959	0.9974	0.9082	0.8998	1.0000

Based on the above discussion, the corresponding training data and test data are selected, and the detection model is built using SPSS Clementine 11 (a data mining tool) for further detection experiments, and the detection results obtained are shown in Table 2. The performance of detection is mainly determined by two factors: ① Whether high-risk transactions are correctly detected as high-risk, which is expressed by the correct detection rate TP. ② How many normal transactions are detected as high-risk transactions, expressed by the false detection rate FN.

From the experimental test results in Table 2, it can be seen that the correct detection rate of XGBoost model optimized with particle swarm optimization of this paper (82.67%, 83.50%) is significantly higher than that of other algorithms, and meanwhile the error detection rate of this paper's model (0.42%, 0.29%) is significantly higher than that of other algorithms.

Table 2 Test results

Method	Real	Training set	Prediction set
--------	------	--------------	----------------

	value	Predict value		Detection rate	Predict value		Detection rate
		0	1	TP/FN%	0	1	TP/FN%
Logit	0	6325	375	FN=5.93	3548	252	FN=7.10
	1	103	197	TP=65.67	96	104	TP=52.00
GNB	0	6582	118	FN=1.79	3495	305	FN=8.73
	1	125	175	TP=58.33	102	98	TP=49.00
DT	0	6485	215	FN=3.32	3567	233	FN=6.53
	1	92	208	TP=69.33	82	118	TP=59.00
RF	0	6293	407	FN=6.47	3642	158	FN=4.34
	1	110	190	TP=63.33	80	120	TP=60.00
AdaBoost	0	6423	277	FN=4.31	3385	415	FN=12.26
	1	131	169	TP=56.33	87	113	TP=56.50
ID3	0	6114	586	FN=9.58	3459	341	FN=9.86
	1	106	194	TP=64.67	93	107	TP=53.50
BP	0	6428	272	FN=4.23	3682	118	FN=3.20
	1	96	204	TP=68.00	76	124	TP=62.00
PSO-XGBoost	0	6672	28	FN=0.42	3789	11	FN=0.29
	1	52	248	TP=82.67	33	167	TP=83.50

4.2 Risk assessment of unusual transactions

In this paper, the data related to credit card customers of sample bank A are first normalized and then used as the basis for the construction of principal component analysis. Factor analysis can usually be presented in terms of commonality analysis with a default threshold value of 1. The closer the commonality is to 1, the more common features are measured between the variable and other variables, and the less information is lost, indicating that the variable is more representative. The variance of the common factor is shown in Table 3, and the variance of commonality for all variables is above 0.75, indicating that the variables selected are representative of most of the information and have good explanatory power for each competency.

Table 3 Common factor variance

		Initial	Extraction
X1	Account time	1	0.931
X2	Time interval to last trading	1	0.844
X3	Business type	1	0.917
X4	Money account	1	0.877
X5	Days of trading of near 7 days	1	0.876
X6	The total amount of the trade in the near 7 days	1	0.934
X7	Average transaction on the last 7 days	1	0.901

X8	Days of trading of near 15 days	1	0.804
X9	The total amount of the trade in the near 15 days	1	0.796
X10	Average transaction on the last 15 days	1	0.791
X11	Days of trading of near 1 month	1	0.867
X12	The total amount of the trade in the near 1 month	1	0.828
X13	Average transaction on the last 1 month	1	0.919
X14	Days of trading of near 2 months	1	0.947
X15	The total amount of the trade in the near 2 months	1	0.957
X16	Average transaction on the last 2 months	1	0.893
X17	Days of trading of near 3 months	1	0.937
X18	The total amount of the trade in the near 3 months	1	0.896
X19	Average transaction on the last 3 months	1	0.789
X20	Days of trading of near 6 months	1	0.812
X21	The total amount of the trade in the near 6 months	1	0.948
X22	Average transaction on the last 6 months	1	0.797

Common factors are extracted immediately afterward. Generally speaking, the extraction of common factors is based on the cumulative contribution rate, which usually needs to be greater than 75% to better retain the original information. The analysis results are shown in Table 4, we extracted 6 principal components from 22 indicators, and their cumulative contribution rate is 88.237%, which means that the main information of 22 indicators has been included in them. For this reason, the paper uses the above 6 principal components to substitute the above 22 indicators.

Table 4 Total variance interpretation (%)

	Initial eigenvalue			Sum of squared loadings			Sum of squared rotated loadings		
	Total	Variance percentage	Cumulative percentage	Total	Variance percentage	Cumulative percentage	Total	Variance percentage	Cumulative percentage
1	6.137	27.896	27.896	6.137	27.896	27.896	5.845	26.568	0.266
2	4.103	18.65	465.546	4.103	18.650	465.546	3.597	16.350	0.430
3	3.171	14.414	60.960	3.171	14.414	60.960	3.288	14.946	0.579
4	2.650	12.046	73.006	2.650	12.046	73.006	2.974	13.518	0.714
5	1.880	8.545	81.551	1.880	8.545	81.551	2.115	9.614	0.810

6	1.47 1	6.686	88.237	1.47 1	6.686	88.237	1.59 3	7.241	88.237
7	0.76 4	3.473	91.710						
8	0.20 8	0.945	92.655						
9	0.19 9	0.905	93.560						
10	0.18 8	0.855	94.415						
11	0.18 2	0.827	95.242						
12	0.17 3	0.786	96.028						
13	0.16 0	0.727	96.755						
14	0.14 0	0.636	97.391						
15	0.13 6	0.618	98.009						
16	0.12 7	0.577	98.586						
17	0.11 8	0.536	99.122						
18	0.08 4	0.382	99.504						
19	0.05 6	0.255	99.759						
20	0.02 6	0.118	99.877						
21	0.01 5	0.068	99.945						
22	0.01 2	0.055	100.000						

The ratio of the variance contribution of each common factor to the total variance contribution of each factor was used as the weights, and the individual common factor scores were weighted and summed to obtain the composite factor F score. The composite factor F score reflects the magnitude of the potential risk of bank credit card customer transactions and meets the normalization criterion, i.e., the higher the F score, the smaller the risk of credit card transactions, and the lower the F score, the greater the

risk of credit card transactions. The results of the principal component weights are shown in Table 5.

Table 5 The main component weight result

	Variance interpretation rate (%)	Cumulative variance interpretation rate (%)	Weight (%)
F1	27.896	27.896	31.615
F2	18.65	465.546	21.136
F3	14.414	60.960	16.336
F4	12.046	73.006	13.652
F5	8.545	81.551	9.684
F6	6.686	88.237	7.577

This paper selects the data samples after clustering above, based on the analysis of the comparative effect of prediction models, selects the PSO-XGBoost model with the optimal effect, on the basis of principal component analysis, the PSO-XGBoost credit card transaction risk prediction model in this paper is applied by example, and part of the data of Bank A in February 2025, the normalized indicator data is inputted into the model, and the prediction results of abnormal credit card transaction risk are obtained for this bank. Credit card abnormal transaction risk prediction results, due to more data, here intercept the system randomly output the first 10 prediction results as an example shown in Table 6. Among them. Risk level 1, 2, 3 and 4 represent very high risk, high risk, medium risk and low risk respectively. From Table 6, it can be seen that the risk of credit card transactions of customer 4 and customer 5 of bank A is very high risk, customer 2 and customer 3 have high risk, customer 1, 5, 7 and 8 credit card transactions have medium risk, and customer 9 and customer 10 have low risk.

Table 6 A bank abnormal transaction risk prediction example

Customer serial	Corresponding output	Risk level
1	2	3
2	1	2
3	1	2
4	0	1
5	2	3
6	0	1
7	2	3
8	2	3
9	3	4

10	3	4
----	---	---

5. Conclusion

In the article, firstly, the risk management of credit card transactions is carried out through data mining technology, and secondly, the XGBoost algorithm is introduced to explore the feasibility of its application to credit card transaction risk prediction, and the model parameters are optimized by particle swarm optimization algorithm.

The AUC value, accuracy, F1 value, precision and recall of PSO-XGBoost model in this paper are 0.9959, 0.9974, 0.9082, 0.8998 and 1.0000 respectively, which are obviously better than other algorithms. The PSO-XGBoost model has the highest correct detection rate and the lowest error detection rate and the best prediction performance on the training and prediction sets of 82.67% and 83.50% and 0.42% and 0.29% respectively. In the prediction of transaction risk for the credit card customers of Bank A, those belonging to very high risk are customer 4 and customer 5, there is a high risk for customer 2 and customer 3, there is a medium risk for customers 1, 5, 7, and 8, and customer 9 and customer 10 are low risk.

References

- [1] Babar, M., & Habib, A. (2021). Product market competition in accounting, finance, and corporate governance: A review of the literature. *International Review of Financial Analysis*, 73, 101607.
- [2] Corbae, D., & Levine, R. (2018, September). Competition, stability, and efficiency in financial markets. In *Paper for the Jackson Hole Economic Symposium, Working Paper*, Haas School of Business, University of California, Berkeley (Vol. 14).
- [3] Kuznetsova Natalia, P., Pisarenko Zhanna, V., & Chernova Galina, V. (2016). Financial market institutions competitiveness and financial convergence. *New Challenges of Economic and Business Development–2016*, 442.
- [4] CARD, C. (2016, December). Credit card payment. In *Workshop in Prague* (Vol. 60, p. 100).
- [5] Singh, S., Rylander, D. H., & Mims, T. C. (2018). Understanding credit card payment behavior among college students. *Journal of Financial Services Marketing*, 23,

38-49.

[6] Guttman-Kenney, B., Firth, C., & Gathergood, J. (2023). Buy now, pay later (BNPL)... on your credit card. *Journal of Behavioral and Experimental Finance*, 37, 100788.

[7] Stavins, J. (2020). Credit card debt and consumer payment choice: what can we learn from credit bureau data?. *Journal of Financial Services Research*, 58(1), 59-90.

[8] Liu, Y., & Dewitte, S. (2021). A replication study of the credit card effect on spending behavior and an extension to mobile payments. *Journal of Retailing and Consumer Services*, 60, 102472.

[9] Gan, C. E., Cohen, D. A., Hu, B., Tran, M. C., Dong, W., & Wang, A. (2016). The relationship between credit card attributes and the demographic characteristics of card users in China. *International Journal of Bank Marketing*, 34(7), 966-984.

[10] Lin, L., Revindo, M. D., Gan, C., & Cohen, D. A. (2019). Determinants of credit card spending and debt of Chinese consumers. *International Journal of Bank Marketing*, 37(2), 545-564.

[11] Yuan, Y., Rong, Z., Xu, N., & Lu, Y. (2021). Credit cards and small business dynamics: Evidence from China. *Pacific-Basin Finance Journal*, 67, 101570.

[12] Zielke, S., & Komor, M. (2020). Loyalty cards, credit options and economic market development. *International Journal of Retail & Distribution Management*, 48(6), 591-607.

[13] Xu, N., Rong, Z., & Yu, L. (2024). Credit cards and commercial insurance participation: Evidence from urban households in China. *Accounting & Finance*, 64(1), 1159-1182.

[14] Collins, G. S., & Moons, K. G. (2019). Reporting of artificial intelligence prediction models. *The Lancet*, 393(10181), 1577-1579.

[15] Afriyie, J. K., Tawiah, K., Pels, W. A., Addai-Henne, S., Dwamena, H. A., Owiredue, E. O., ... & Eshun, J. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. *Decision Analytics Journal*, 6, 100163.

[16] Gao, J., Zhou, Z., Ai, J., Xia, B., & Coggeshall, S. (2019). Predicting credit card transaction fraud using machine learning algorithms. *Journal of Intelligent Learning*

Systems and Applications, 11(3), 33-63.

[17] Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., ... & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. *Ieee Access*, 8, 201173-201198.

[18] Leong, O. J., & Jayabalan, M. (2019). A comparative study on credit card default risk predictive model. *Journal of Computational and Theoretical Nanoscience*, 16(8), 3591-3595.

[19] Gao, J., Sun, W., & Sui, X. (2021). Research on Default Prediction for Credit Card Users Based on XGBoost-LSTM Model. *Discrete Dynamics in Nature and Society*, 2021(1), 5080472.

[20] Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. *Procedia computer science*, 132, 385-395.

[21] Wang, C., & Han, D. (2019). Credit card fraud forecasting model based on clustering analysis and integrated support vector machine. *Cluster Computing*, 22(Suppl 6), 13861-13866.

[22] Xuelin Dong, Taolong Li & Yadi Yao. (2023). Research on Risk Management of Credit Card Business of Tangxia Branch of Guangzhou Rural Commercial Bank. *Financial Engineering and Risk Management*, 6(5).

[23] Kwanele Phinzi, László Bertalan, Gashaw Gismu Chakilu & Szilárd Szabó. (2025). Improving the spatial prediction of topsoil properties in a typical grazing area using multi-season PlanetScope spectral covariates and data mining techniques. *Earth Science Informatics*, 18(2), 222-222.

[24] Nikita Kozodoi, Stefan Lessmann, Morteza Alamgir, Luis Moreira Matias & Konstantinos Papakonstantinou. (2025). Fighting sampling bias: A framework for training and evaluating credit scoring models. *European Journal of Operational Research*, 324(2), 616-628.

[25] Lei Xu, Shaomu Wen, Hongfa Huang, Yongfan Tang, Yunfu Wang & Chunfeng Pan. (2025). Corrosion failure prediction in natural gas pipelines using an interpretable XGBoost model: Insights and applications. *Energy*, 325, 136157-136157.

[26] Mostafa Alizade Harakiyan, Amin Khodaei, Ali Yousefi, Hamed Zamani & Asghar

Mesbahi. (2025). Decision tree-based machine learning algorithm for prediction of acute radiation esophagitis. *Biochemistry and biophysics reports*,42,101991.

[27] Tian Zhang,Lianbo Ma,Shi Cheng,Yikai Liu,Nan Li & Hongjiang Wang. (2025). Automatic prompt design via particle swarm optimization driven LLM for efficient medical information extraction. *Swarm and Evolutionary Computation*,95,101922-101922.