Prediction of Credit Card Loan Risk Based on Multilayer Perceptron Neural Network Model

Zejie Yu *

School of Automation Science and Engineering, South China University of Technology, 510000, Guangzhou, China

*Corresponding author Email: 202030462456@mail.scut.edu.cn

Abstract. Under the background of Internet finance and the popularization of intelligent terminal equipment, the scale of data and the diversity of information have developed exponentially rapidly, while the bad debt rate of personal credit is vague and much higher than that of commercial banks in the traditional technology period. Most of the traditional credit evaluation is driven by models, and the robustness of the risk control system is poor, which cannot meet the increasingly complex demand for default risk prediction. Since commercial banks have accumulated large-scale data assets, it has great significance to combine machine learning technology to help banks extract effective information from massive data and achieve risk assessment of borrowers, thereby reducing the risk of default in borrowing. This paper selects a credit card related data set, uses principal component analysis to extract feature vectors, and obtains a risk assessment index system for borrowers. Then, based on MLP neural network, a credit card loan risk prediction model is constructed. The study reads: The MLP model based on principal component analysis has high accuracy, fast running speed and strong stability. It is an ideal model for commercial banks to evaluate the credit risk of borrowers. This study can provide a new reference for commercial banks to solve the problem of credit risk prediction.

Keywords: artificial neural network, multilayer perception, machine learning, risk evaluation.

1. Introduction

Since the outbreak of COVID-19, the macroeconomic market has been depressed. Out of concern for the economy, the traditional financial institutions represented by commercial banks have become increasingly reluctant to lend, resulting in the simultaneous existence of credit grudging and insufficient demand for credit. At the same time, more and more financial institutions have begun to pay attention to personal credit risk. How to help banks solve the credit evaluation of borrowers has become the focus of academic discussion.

Yet the traditional credit evaluation is mostly driven by the model, and the robustness of the risk control system is poor, which can not meet the increasingly complex needs of default risk prediction. Meanwhile emerging technologies such as big data, cloud computing, artificial intelligence, and mobile Internet have developed rapidly. The financial industry has begun to integrate deeply with the new round of technological revolution of artificial intelligence, which has had a significant impact on business models, applications, processes, and products in the financial field. Therefore, using machine learning technology to continuously optimize the risk assessment model has become a major goal of the construction of the current credit risk assessment system.

2. Literature Review

Initially, the credit risk assessment mainly used qualitative analysis methods, such as 5C method and 5P method, that is, from Character, Capacity, Collateral, Capital and Condition or People, Purpose, Payment, Protection and Perspective to evaluate the borrower. Yet qualitative analysis relies too much on subjective experience and cannot quantify risk. With the evolution of mathematical statistics and the development of modern information technology, there are many researches on quantitative analysis and forecasting models of credit risk assessment, which are mainly divided into statistical measurement methods and machine learning methods.

Early scholars preferred to use statistical analysis method to establish credit rating model to evaluate personal credit risk in traditional commercial banks. In 1941, Durand [1] first used statistical methods to evaluate personal credit. He divided the overall data according to different characteristics, and then completed the credit risk research on loans. In the 1950s, statistical techniques such as regression analysis began to be used in credit scoring to quickly process a large number of credit applications. In 1956, Bill Fair and Earl Isaac invented the FICO credit scoring model based on logistics regression model, which has clear business logic and strong interpretability [2]. In 1980, Wiginton et al. [3] proved that the proportion of correct classification of logistic regression applied to personal credit risk assessment was significantly higher than that of linear discriminant model. In 2016, Zhang et al. [4] found that the personal credit scoring model established by Adaptive Lasso-Logistic regression method was superior to the traditional Logistic and Lasso-Logistic methods in variable selection, interpretation and prediction accuracy.

Yet the traditional statistical measurement methods have strict assumptions. The classical logistic regression model and discriminant analysis model are mainly aimed at linear problems, which cannot analyze the nonlinear relationship between variables, and have high requirements on data quality, sensitive to noise. But with the rapid development of Internet finance, the traditional statistical measurement model has great limitations in the application of actual financial risk control scenarios. Its shortcomings such as single information dimension and time lag are no longer suitable for model development and integration under the background of big data. Therefore, more and more scholars begin to pay attention to the use of machine learning and deep learning methods to build risk assessment models.

In 1986, Coffman [5] first applied the classification tree model to the field of personal credit risk assessment, which achieved satisfactory results. In terms of automatic selection of explanatory variables and accuracy, the decision tree model had obvious advantages. In 1990, Odom et al. [6] used neural networks in the first personal credit evaluation, and concluded that neural networks are more robust than discriminant analysis in reducing sample size. In 2010, Khandani et al. [7] used machine learning technology to construct a prediction of consumer credit default probability. The empirical results showed that after the financial crisis, machine learning algorithms can effectively reduce the cost of credit lines. In 2014, Oreski et al. [8] proposed a genetic algorithm hybrid neural network algorithm (HGA-NN) to identify the optimal feature subset, which improved the classification accuracy and scalability of credit risk assessment. In 2019, Setiawan et al. [9] proposed a binary particle swarm optimization algorithm based on support vector machine (BPSOSVM) to select features for data sets, and used extreme random tree (ERT) and random forest (RF) as classifiers to predict whether a loan will become a bad debt. Since the ensemble learning model (EM) can effectively improve the accuracy and stability of the model, many scholars suggest using ensemble learning to construct a default risk assessment model. Lean et al. [10] verified the effectiveness of the proposed multi-level neural network ensemble learning model in credit risk assessment through two public credit data sets. Sun [11] sampled a new model of Bagging integrated learning algorithm based on SMOTE algorithm and differential sampling rate when studying unbalanced enterprise credit evaluation. The results show that the new model is significantly better than the other five models and is effective for enterprise credit evaluation imbalance. In 2021, according to the characteristics of financial data, Wang [12] combined the SMOTE algorithm with the MK-FCM method to construct an integrated model for unbalanced credit risk prediction and improved the algorithm.

This paper proposes the use of MLP model to predict the borrowers' credit risk, extracts the feature vector based on principal component analysis, and integrates performance index on personal loan risk assessment into artificial neural networks, transparentizing impact path of the black box. The marginal contribution of this paper is to break through the singularity of the traditional statistical measurement model and the endogeneity of its explanatory variables, with the use of the ANN-MLP model to describe the impact of relevant indicators on the credit risk of borrowers, and the impact path is dynamic and nonlinear. This study is helpful for commercial banks to judge the credit risk of

borrowers, providing an effective reference for commercial banks' loan decisions, and verifies the applicability of artificial neural network in the study of credit risk assessment.

3. Method

3.1 Introduction of artificial neural network

Artificial Neural Network (ANN) describes the mathematical theory and network structure of artificial neurons by simulating the principle and process of nerve cells in biology. [13] Since ANN imitates the human brain nervous system to a certain extent, it also imitates some basic functional characteristics of the human brain, that is, the human brain finally makes feedback through a complex neural network system after receiving external information. In this way, the artificial neural network algorithm has the ability to deal with problem of complex information, uncertain environment, unclear relationship and imperfect background knowledge. In addition, the neural network algorithm also has adaptive technology. It does not need to make a priori assumption. It only needs to learn and train the sample data to find the functional relationship hidden inside the sample data. ANN can use the historical training model to make the correct response to the samples in the future training. It can accurately process and analyze the nonlinear system through continuous training and learning, and the prediction accuracy is high. [14]

3.2 Architecture of artificial neural network

The most basic unit of artificial neural network model (ANN) is the neuron model, which can simulate the biological neural network system to interact with real objects. The model is divided into three layers: input layer, hidden layer and output layer. In the neural network model, the neuron receives other weighted neuron signals and outputs the results by using a nonlinear activation function after comparison. The non-linear mapping ability of neural network model is outstanding, which can learn the rule of data through model training from historical sample characteristics. The neural network model is often used to deal with problems such as classification and regression, and shows good generalization ability. [15]

3.3 Multilayer perceptron (MLP)

The MLP neural network, a typical representative of neural networks, includes an input layer, one or more hidden layers, and an output layer. The layers are fully connected by neurons with nonlinear activation functions. Each connection is assigned different weights, so that the input of each layer of neurons is the weighted sum of the output values of the previous layer of neurons, namely

$$y^{l}_{i} = f(\sum_{j=1}^{n} y_j^{l-1} w_{ji}^{l-1} + b_i^{l}) \tag{1}$$

In the formula, y^{l}_{i} is the output of neurons i of the layer l; f is the activation function of the layer l; n is the number of neurons of layer $l-1$; y_j^{l-1} is the output of neuron j of the layer $l-1$; w_{ji}^{l-1} is the connection weight of the neuron j and the neuron i in the layer $l-1$; b_i^{l} is bias for neurons i of the layer l.

Steps of MLP neural network for credit risk assessment are as follows:

Step 1: Samples are randomly divided into training set and test set.

Step 2: The sample features of the training set enter the network through the input layer, and the output is passed to the hidden layer. The input of the hidden layer is the weighted sum of the output of the previous layer, that is. The input value A is processed by the hidden layer activation function and passed to the output layer. The input of the output layer is , processed by the S-type function of the activation function of the output layer, and the final result is mapped in the (0,1) interval. At the same time, the network calculates the output value error by comparing the output value with the actual value, uses the algorithm to continuously update the network parameters, until the loss is minimized,

then stops training and gets the final model. Setting the model threshold to 0.5, if the final result is lower than the threshold, the enterprise is assessed as high credit risk. On the contrary, the enterprise is evaluated as low credit risk, and the training set is completed.

Step 3: Input the test set sample into the model and compare the model output value with the actual value to derive the model's evaluation capability.

4. Evaluation Construction of Algorithm Model

Through the evaluation index, this paper selects the appropriate method to help check the problems existing in each link of the analysis process, and corrects them in time. For accurately evaluating the prediction accuracy and model performance of the established MLP neural network, improving the model confidence, this paper provides a reliable reference index for optimizing the model and analysis results.

In this paper, two evaluation indexes of prediction effect, that is, sum of squared errors (SSE) and relative error (Er), are selected for evaluation. The basic calculation formulas are as follows:

$$SSE=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2 \tag{2}$$

$$E_r=\frac{\hat{Y}_i-Y_i}{Y_i} \tag{3}$$

In the formula, n is the number of samples; Y_i is measured value; $\hat{Y}_i$ is the predicted value.

SSE and Er are used to measure the deviation between the predicted value and the true value of the model. The closer the value is to 0, the better the model selection and fitting, and the more successful the data prediction.

The ROC curve (Receiver Operating Characteristic Curve). Formulas of TPR and FPR are as follows:

$$TPR=\mathrm{Re}call=\frac{TP}{TP+FN} \tag{4}$$

$$FPR=\frac{FP}{FP+TN} \tag{5}$$

The ROC is a curve with FPR as the horizontal axis and TPR as the vertical axis. Every point represents the FPR and TPR of the model under different thresholds.

5. Data and Indicator Construction

5.1 Data source and data preprocessing

This paper takes personal credit business as the research object, and selects information of 30,000 personal credit card borrowing and repayment data. The data is derived from the default of credit clients dataset provided by Kaggle, including the personal information, credit card consumption records, repayment records and other information from a bank of Hong Kong from April to September of a certain year. The data has 25 original features, one of which is the target attribute, that is, whether the next month is defaulted.

Data characteristics included in the sample include user personal information such as ID, GENDER, EDUCATION, MARRIAGE, and AGE, as well as credit amounts, repayments from April to September, billing amounts from April to September, and repayment amounts from April to September. Among them, GENDER, EDUCATION, MARRIAGE, and repayment from April to September are quantified as dummy variables. What we stipulates is as follows: for GENDER, 1 = male, 2 = female; for EDUCATION 1 = graduate, 2 = undergraduate, 3 = high school, 4 = other; for MARRIAGE 1 = married, 2 = single, 3 = other; for repayment from April to September -2 = two months in advance, -1 = one month in advance, 0 = timely payment, 1 = one month delay, ... 8 =

Repayment delay of 8 months. The ID is the users' number, which is meaningless for prediction and can be deleted directly. The remaining variables are continuous variables.

There are no missing values in the original data. By observing the variables, it can be seen that the number of EDUCATION types is 7; the minimum value is 0 and the maximum value is 6; the number of MARRIAGE types is 4; the minimum value is 0. These data are meaningless and exceed the specified data meaning, so these data are equivalent to missing values and need to be interpolated.

The amount of data with values of 0,5,6 in EDUCATION is 345, and the amount of data with MARRIAGE value of 0 is 54, which is not large compared with the total of 30000. According to the actual meaning, these values can be attributed to the 'other ' class, that is, the variables with values of 0,5,6 in EDUCATION are replaced by 4 and the variable of 0 in MAGGIAGE is replaced by 3.

5.2 Construction of index system based on principal component analysis

Obviously, if the above indicators are used as MLP network input, it will inevitably increase the complexity of the network and computing time, reduce network performance, and weaken the generalization ability of neural networks. Therefore, it is necessary to comprehensively analyze these indicators to reduce the number of features on the premise of reducing information loss. In this paper, the principal component analysis method is used to extract the factors of the above characteristics and construct a few comprehensive principal component indexes. The starting point of the principal component analysis is to calculate a set of new features arranged from importance from a set of features. They are linear combinations of the original features and are not related to each other.

First, this paper test whether the index data is suitable for factor analysis, and then test the sample data by KMO measure and Bartlett sphere test. The test results show that the KMO statistic value is 0.804, greater than 0.5, indicating that the degree of overlap between the variables is not high. At the same time, the Bartlett sphere test values in the three-category model are significant at the 0.01 level, so it is suitable for factor analysis. Factor analysis-principal component method was used to extract factors in SPSS, and the principal component factor was determined by taking the eigenvalue greater than 1 as the extraction standard. According to the gravel diagram (Fig.1.), three principal component factors should be selected as common impact factors, and the results are shown in Table.1.

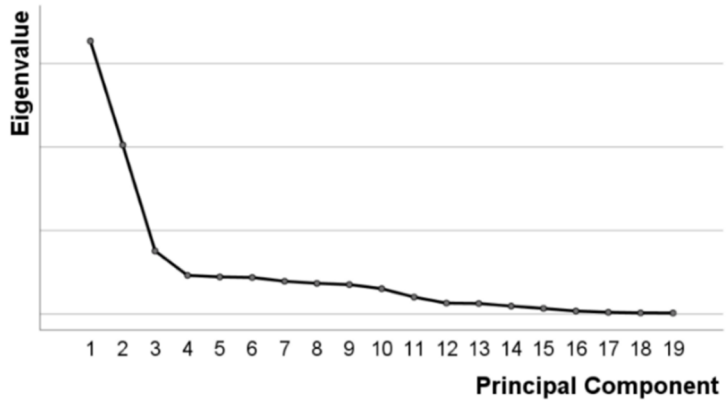

Figure 1. Eigenvalue spectrum of principal component analysis (self-drawn)

Table 1. The eigenvalue and contribution rate of the principal component of the sample data under the three classification mode (self-drawn)

Component	Initial eigenvalue			The sum of squared loads extracted			Sum of load squares after rotation		
	Grand total	Percentage of variance	Cumulative percentage	Grand total	Percentage of variance	Cumulative percentage	Grand total	Percentage of variance	Cumulative percentage
1	6.537	34.407	34.407	6.537	34.407	34.407	5.379	28.310	28.310
2	4.048	21.305	55.713	21.305	55.713	4.474	4.474	23.548	51.858
3	1.508	7.938	63.650	7.938	63.650	2.240	2.240	11.792	63.650

The three new features in Table.1. explain 63.650 % of the overall variance, that is, the information coverage rate of these three factors to the original data reaches 63.650 %, thus the results of the

extraction factors satisfactory. This paper uses the maximum variance orthogonal rotation method to rotate the factors, so that the load of each factor on each original variable is close to positive and negative 1 or 0, which contributes to a strong discrimination between the factors, benefiting of explaining the actual meaning of the factors. After orthogonal rotation, the load matrix of each factor is shown in Table.2. The vacant position in Table.2. indicates that the influence of this variable on the principal component is close to 0, which can be ignored for the convenience of analysis. This paper stipulates that variables $BILL\_AMT1$ to variables $BILL\_AMT6$ are respectively x_1 to x_6; PAY_0 is for x_7; PAY_2 to PAY_6 respectively x_8 to x_{12}; $LIMIT\_BAL$ is for x_{13}; PAY_AMT1 to PAY_AMT6 respectively x_{14} to x_{19}. From Table.2., the relationship between the selected principal components and the variables is as follows:

$$Z_1 = 0.942x_1 + 0.933x_2 + 0.926x_3 + 0.924x_4 + 0.911x_5 + 0.890x_6 \quad (6)$$

$$Z_2 = 0.706x_7 + 0.824x_8 + 0.865x_9 + 0.887x_{10} + 0.875x_{11} + 0.826x_{12} - 0.424x_{13} \quad (7)$$

$$Z_3 = 0.615x_{14} + 0.637x_{15} + 0.609x_{16} + 0.530x_{17} + 0.470x_{18} + 0.445x_{19} \quad (8)$$

Table 2. Component load matrix after rotation (self-drawn)

	Component		
	1	2	3
BILL_AMT2	0.942		
BILL_AMT1	0.933		
BILL_AMT4	0.926		
BILL_AMT3	0.924		
BILL_AMT5	0.911		
BILL_AMT6	0.890		
PAY_4		0.887	
PAY_5		0.875	
PAY_3		0.865	
PAY_6		0.826	
PAY_2		0.824	
PAY_0		0.706	
LIMIT_BAL		-0.424	
PAY_AMT2			0.637
PAY_AMT1			0.615
PAY_AMT3			0.609
PAY_AMT4			0.530
PAY_AMT6			0.470
PAY_AMT5			0.445

6. Construction of Ann-Mlp Model and Experimental Results

6.1 Sample partition

The neural network algorithm can gradually establish and improve the development path between the input variable and the output result through a large amount of historical data, that is, the neural network. In this neural network, the establishment of each nerve and the weight of the nerve are obtained through a large number of historical data training. The more data, the closer the neural network is to reality. After the neural network is established, the output can be predicted by different input variable values. Selecting historical data to build models, historical data is generally divided into two parts: training set and validation set. Many researchers use the first 70 % of data as a training set and the last 30 % as a validation set directly in the data order. If the data can be proved to be independent of each other, this is no problem, but in the process of data collection, the collected data is often not completely independent. Therefore, this paper uses a random number generator to

generate random seeds, and then uses the Bernoulli formula to randomly extract about 70 % of the data as the training set, the rest as the verification set, so as to avoid the data of the same attribute being classified into a data set, making the established model more effective.

A total of 30,000 samples were collected from the data set, of which 21023 samples were extracted as training sets, accounting for 70.1 % of the total samples, and the remaining 8977 samples were used as validation sets, accounting for 29.9 % of the total samples.

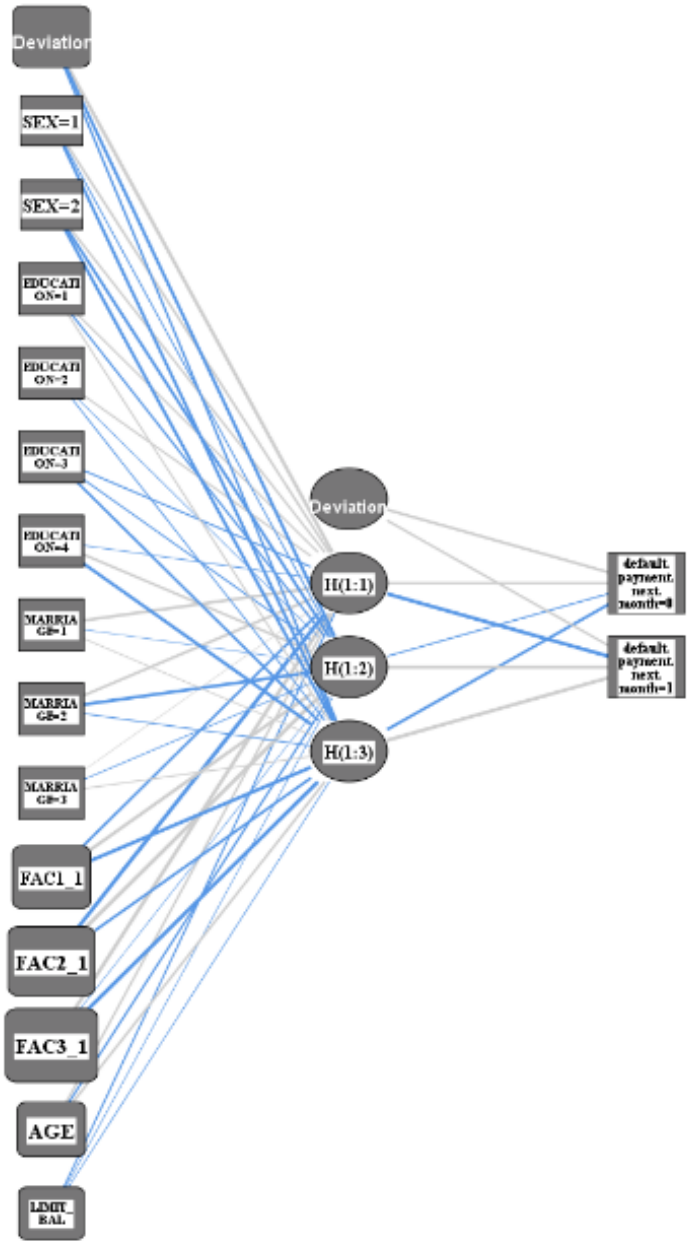

Figure 2. MLP neural network model (self-drawn)

6.2 Design of network structure, activation function setting and parameters

In general, the prediction accuracy of the neural network model is positively correlated with the number of layers of the network structure, but in the actual operation process, the prediction accuracy of the model is often affected by the sample size. Therefore, the optimal number of hidden layers is not the more the better, and the optimal number of hidden layers should be selected according to the actual situation and the needs of modeling.

Activation function, as an indispensable part of neural network model, controls the activation of input to output, performs function transformation of input and output, and transforms the input of possible infinite domain into the output of specified finite range. Without the activation function, the output of the neural network model is only a linear expression of the input. Common activation functions are Sigmoid function (a class of S-curve function), hyperbolic tangent function and ReLU function. This paper sets the hyperbolic tangent function as the hidden layer activation function and the Softmax function as the output layer activation function.

Since there is no uniform standard for the optimal number of hidden layers and the number of neurons in the neural network. The optimal network structure is mainly obtained by experience and continuous debugging. By comparing a variety of different MLP neural networks, this paper selects the model with the highest accuracy, that is, the number of hidden layers is 1 and the number of hidden layer neurons is 3, as the credit risk assessment model of credit card users. The MLP model network structure diagram is shown in Fig.2.

6.3 Model results and tests

After machine learning, the model training set is evaluated to be 80.9 % accurate, and the test set model is evaluated to be 80.1 % accurate, indicating that this machine learning training has a good effect and can basically achieve the prediction demand.

The importance of variables can be seen by observing Figure 3.

The ROC curve is shown in Figure 4.

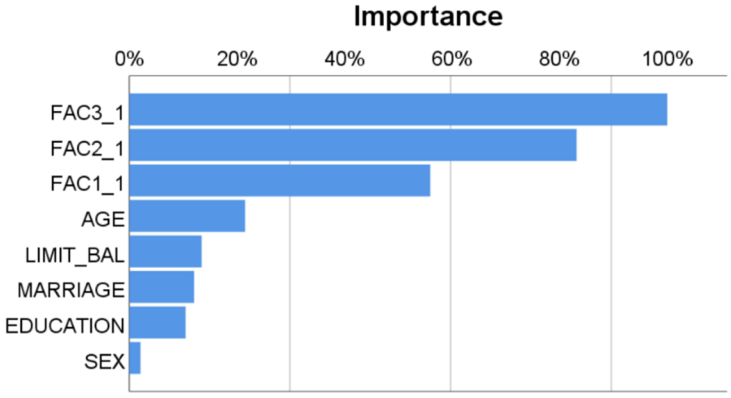

Figure 3. Normalization importance of variables (self-drawn)

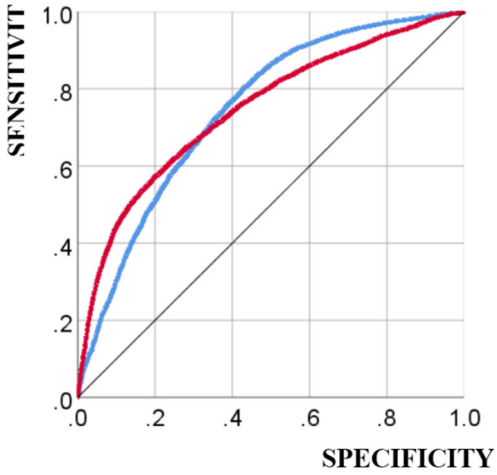

Figure 4. ROC curve (self-drawn)

7. Conclusion

Firstly, this paper summarizes and combs the research conclusions of loan credit risk at home and abroad. Based on the principal component analysis method, the evaluation index is reduced and three

principal component common factors are extracted. On this basis, combined with the neural network technology based on multi-layer perceptron (MLP), the credit card loan risk prediction model under the three-classification model is constructed. The accuracy of the prediction results based on the model reaches 80.1 %, which proves the validity and rationality of the model. The specific experimental conclusions are as follows:

(1) MLP neural network performs well in approximation ability, classification ability and learning speed, so it is effective to use MLP neural network to predict credit card credit problems.

(2) Multilayer perceptron has a large algorithm optimization space. In the later stage, algorithms can be added to make it more optimized, such as generalized multilayer perceptron, particle algorithm of BP neural network, convolutional neural network and so on.

This study is helpful for commercial banks to predict the credit risk of borrowers, providing an effective reference for commercial banks' loan decision-making, and it verifies the applicability of artificial neural network in the study of credit risk assessment. However, there are still some problems in this paper, such as sample data has room for expansion, variables may not be fully considered and so on, which can improve the relevant research in the future.

References

[1] Durand D. Risk Elements in Consumer Installment Financing[M]. NewYork: National Bureau of Economic Research.1941:60-129.

[2] Jiang Lin. Review of American FICO Scoring System [J]. Business Research. 2006(20):81-84.

[3] John C.W. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior[J]. Journal of Financial and Quantitative Analysis, 1980, 15(3): 757-770.

[4] Zhang Tingting, Jing Yingchuan. Adaptive Lasso-Logistic regression analysis of personal credit score [J]. Practice and understanding of mathematics.2016,46(18):92-99.

[5] J.Y. Coffman. The proper role of tree analysis in forecasting the risk behavior of borrowers[J]. Management Decision Systems.1986, (3):47-59.

[6] Odom D, Sharda R. A Neural Network Model for Bankruptcy Prediction[C]. Neural Networks.IEEE,1990:688-700.

[7] Khandani A. E., Kim A. J., Lo A. W. Consumer credit risk models via machine learning algorithms[J]. Journal of Banking &Finance, 2010, 34(11):2767-2787.

[8] Stjepan Oreski, Goran Oreski. Genetic algorithm-based heuristic for feature selection in credit risk assessment[J]. Expert Systems WithApplications,2014,41(4).

[9] Netty Setiawan, Suharjito, Diana. A Comparison of Prediction Methods for Credit Default on Peer to Peer Lending using Maehine Learning[J]. Proeedia Computer Science,2019,157.

[10] Yu L., Wang S.Y., Lai K. K. Credit risk assessment with a multistage neural network ensemble learning approach[J]. Expert Systems With Applications, 2007, 34(2):1434-1444.

[11] Jie, Sun, et al. Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates[J]. Information Sciences An International Journal, 2018.

[12] Wang L. Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization[J]. 2021.

[13] Xia Yulu. (2019). A Review of the Development of Artificial Neural Networks. Computer Knowledge and Technology (20),227-229.

[14] Li Jinfeng. (2020). Application and Development Trend of Artificial Neural Network in Petrochemical Enterprises. Petrochemical Design (02),53-58.

[15] Zhao Chongwen. (2020). Summarization of Artificial Neural Network. Shanxi Electronic Technology (03),94-96.