World Scientific Research Journal Volume 6 Issue 11, 2020
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0019

Research on Personal Credit Risk Assessment based on Machine Learning Algorithm

Keting Ye

College of Economics, Jinan University, Guangzhou 510000, China.

Abstract

The main purpose of this paper is to use machine learning algorithm to establish different classification models to evaluate and predict personal credit risk. In this paper, we take the data of give me some credit in the kaggle competition as an example. We take the seriousdlqin2yrs variable which is overdue for more than 90 days or worse as the target variable, and take other characteristic variables of the data as independent variables for modeling and analysis. In the stage of data preprocessing, this paper first uses the k-nearest neighbor method to fill in the missing values in the data, then processes the outliers in the data, and tests the multicollinearity among variables. In the process of model construction, logistic regression model is constructed by step-by-step screening method, decision tree model is constructed by cart algorithm, classification prediction model is constructed by SVM algorithm, and integrated model is constructed based on three algorithms. With the help of AUC value and ROC curve, by comparing the prediction effect of different models in training data set and test data set, it is found that the integrated learning model performs better, has higher classification effect and has stability.

Keywords

Credit evaluation; Logistic model; Decision tree model; SVM algorithm; Integrated learning.

1. INTRODUCTION

With the rapid development of China's economy, people's living standards have been generally improved, and the consumption capacity of residents is also gradually rising. Especially in recent years, the rapid development of Internet Finance provides a lot of convenience for residents' daily consumption, which makes the growth trend of residents' credit consumption more and more obvious. Personal housing mortgage loans, personal micro loans, credit card consumer loans and other credit products can be seen everywhere. Relevant data show that with the growth of individual consumption credit capacity in China, the loan scale of individual in many financial institutions, such as commercial banks, has been steadily increasing, gradually surpassing the loan scale of enterprises. However, with the development of individual credit economy, the risk of default of personal credit loan is increasing year by year, which brings huge capital risk to commercial banks and other financial institutions. Therefore, it is particularly important for financial institutions to effectively assess personal credit risk according to customer information when making loans, so as to reduce the economic losses caused by default.

In the era of big data, with the rapid development of Internet finance, personal credit consumption grows rapidly, and credit card users gradually increase, accompanied by a large number of transaction data. The traditional credit audit mainly relies on manual work, which decides whether to approve loans based on manual experience. When facing a large number of

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0019

micro credit applications, the cost of manual audit is high and the audit time is long. This traditional credit card risk assessment method has been unable to meet the needs of rapid increase, and the credit evaluation model can quickly and automatically complete batch processing, not only fast, but also accurate. Therefore, the establishment of a more perfect and more powerful personal credit risk assessment system has become the urgent need of many financial institutions.

In western developed countries, personal credit risk assessment originated in the United States. The methods of personal credit risk assessment can be divided into statistical analysis and nonparametric machine learning. In the research of statistical methods, personal credit evaluation is essentially a method to divide a whole into several different groups according to different characteristics. This idea of dividing the population into different groups was first proposed by Fisher(1936)in statistics. At the same time, the logistic regression model was established for the first time. Although the model established by statistical method has high accuracy and precision, it can not reflect the credit status of customers completely in many cases, especially the information with real analysis value is difficult to be extracted.

In recent years, nonparametric statistics and machine learning methods have also developed rapidly in credit evaluation. Makowski(1985)first applied the decision number model to the field of personal credit evaluation. Because of its good interpretation and ability to deal with missing data, it has attracted extensive attention. Support vector machine (SVM) is a supervised learning method, which has obvious advantages in high-dimensional data classification, pattern recognition and small sample data. Baesens and gestel (2003) first applied support vector machine (SVM) in the field of personal credit evaluation, and through empirical analysis found that the classification effect of SVM credit risk assessment model was better than that of neural network and linear regression.

Domestic scholars have done a lot of work on the research of personal credit risk assessment methods, and have achieved considerable results. Kuang Nan F, Guijun Z and Huiying Z(2014)applied lasso logistic model to personal credit score. The research results show that lasso logistic model has better effect than logistic model with full variable and logistic model with stepwise regression to screen variables. Bing W and Qi F (2006) established a personal credit risk assessment model by using support vector machine theory, and used the grid 5-fold crossover method to find the optimal parameters of different kernel functions, and compared the SVM model with linear discriminant analysis, logistic regression analysis, nearest neighbor, classification regression tree and neural network, and found that SVM has better prediction effect.

The goal of this paper is to establish a credit risk assessment model by using machine learning algorithm. Through empirical research, this paper compares the applicability and robustness of different models in a large number of customer credit data, calculates the default probability of customers quickly and effectively, and looks for a more suitable credit risk assessment model, so as to provide certain reference value for financial institutions to build credit analysis and evaluation system.

2. CORRELATION THEORY

2.1. Logistic Regression

Logistic model [6] can be applied to regression problems, and also can be used to solve classification problems. In the classification problem, the model can calculate the probability of belonging to each category according to a set of independent variables. Logistic regression model is the most widely used multivariate quantitative analysis method for binary dependent variable (i.e. $y = 1$ or $y = 0$).

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0019

Binary logistic regression is used to describe the data that the dependent variable y obeys the 0-1 distribution. The relationship between dependent variabl and independent variable is nonlinear, and its conditional probability distribution is as follows:

$$P = p(Y=1|X,W) = \frac{e^{W^T X+b}}{1+e^{W^T X+b}} = \frac{1}{1+e^{-(W^T X+b)}} \tag{1}$$

$$p(Y=0|X,W) = 1-P = \frac{1}{1+e^{W^T X+b}} \tag{2}$$

Where P represents the probability of Y occurrence with the value of 1, X represents the input vector, that is, the independent variable, $X \in R^n$ and $X=(x_1,x_2,\cdots,x_n)$, W represents the weight vector, $W \in R^n$ and $W=(w_1,w_2,\cdots,w_n)$, b represents the transposes the parameter.

2.2. Support Vector Machine (SVM)

Support vector machine (SVM) [7] is an intelligent machine learning algorithm, which is mainly based on statistical learning theory. SVM makes full use of the learning mechanism, according to the principle of structural risk minimization, guarantees the minimization of empirical risk and confidence range, and achieves high statistical prediction. It is often used in classification, recognition, regression and other fields.

The initial application of support vector machine is to study linear separable problems. Taking linear separable SVM as an example, this paper introduces the idea and principle of support vector machine.

Suppose there are a certain number of N training samples $\{(x_i,y_i),i=1,2,\cdots,l\}$, which are composed of two categories. Suppose there is an optimal hyperplane,

$$wx+b=0 \tag{3}$$

If the samples of the same class fall on the same side of the optimal hyperplane, then the samples are considered to be separable. We may define the interval between the sample point and the optimal hyperplane as

$$\varepsilon_i = y_i(wx_i+b) = |wx_i+b| \tag{4}$$

The goal of support vector machine classification is to find an optimal hyperplane in the hyperplane satisfying the conditions, so $\frac{2}{\|w\|}$ as to reach the maximum, which is equivalent to $\frac{\|w\|^2}{2}$. The mathematical model is expressed as follows:

$$\begin{cases} \min \ \frac{1}{2}\|w\|^2 \\ s.t. \ \ y_i(wx_i+b) \geq 1, \ \ i=1,2.\cdots n \end{cases} \tag{5}$$

The Lagrange multiplier method is used to solve the model

$$\Phi(w,b,a_i) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i[y_i(wx_i+b)-1] \tag{6}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0019

Where is a_i Lagrange coefficient, and $a_i > 0, i = 1, 2, \cdots, n$.

2.3. Model Evaluation

The evaluation indexes of the classification model include accuracy, type I error (1-specificity), type II error (1-sensitivity) and AUC (area under curve) value [10]. The prediction result of most machine learning models for classification problem is probability, and the calculation of accuracy rate, the first kind of error and the second kind of error needs to convert the probability into category, which requires setting a threshold value, which will affect the calculation of accuracy and other indicators. This problem can be avoided by using AUC. Therefore, ROC (receiver operating characteristic) curve and AUC value are selected as the evaluation criteria of the model.

After the binary prediction is obtained, a confusion matrix can be constructed to evaluate the prediction effect of the binary classifier.

Table 1. Confusion matrix

	Prediction positive class P	Forecast negative class N
Actual positive class T	TP	FN
Actual negative class F	FP	TN

Among them, TP means that the prediction sample is actually positive, and the prediction result is also positive; FN means that the prediction sample is actually positive and the prediction result is negative; FP is that the prediction sample is actually negative and the prediction result is positive; TN is that the prediction sample is actually negative and the prediction result is also negative.

The evaluation index of the model can be calculated by the following formulas:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{7}$$

$$TPR = \frac{TP}{TP+FN} \tag{8}$$

$$FPR = \frac{FP}{FP+TN} \tag{9}$$

ROC curve is composed of a series of TPR and FPR. FPR is used as abscissa, TPR as ordinate.

AUC (area under curve) is the area under the ROC curve enclosed by the coordinate axis, and its value range is 0.5-1. In general, the larger the AUC, the better the classifier.

3. DATA SOURCES AND PROCESSING

3.1. Data Sources

Due to the privacy of personal credit data, it is not easy to obtain. In this paper, we will take the given me some credit data set published in the kaggle competition as an example, from the aspects of data preprocessing, modeling analysis and model comparison, to judge and identify the credit risk of credit customers based on the characteristic variables of credit customers. The data set has 150000 samples and 11 variables. In this paper, seriousdlqin2yrs variables representing overdue delinquency of more than 90 days or worse are taken as explanatory

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0019

variables: one is 0 (non defaulting customers), the other is 1 (defaulting customers), and the other 10 characteristic variables are explanatory variables. These features include age, debt ratio, monthly income, etc. Table 2 shows the meanings of variables and their fields in the dataset:

Table 2. Variable list

Variable	Field	Type
Y	SeriousDlqin2yrs	Y/N
X1	RevolvingUtilizationOfUnsecuredLines	percentage
X2	age	integer
X3	NumberOfTime30-59DaysPastDueNotWorse	integer
X4	DebtRatio	percentage
X5	MonthlyIncome	real
X6	NumberOfOpenCreditLinesAndLoans	integer
X7	NumberOfTimes90DaysLate	integer
Variable	Field	Type
X8	NumberRealEstateLoansOrLines	integer
X9	NumberOfTime60-89DaysPastDueNotWorse	integer
X10	NumberOfDependents	integer

3.2. Data Preprocessing

When carrying out data analysis, it is necessary to understand the overall situation of data. The lack of data, data anomalies and other reasons may lead to the original data can not meet the requirements of data mining modeling, which is not conducive to mining effective information. Therefore, data preprocessing is usually needed before data mining.

(1) Missing value processing. In general, due to the fear of personal privacy disclosure, inconvenient disclosure or data information record omission or error, there may be missing values in many important attributes of personal credit original data, such as age, income, etc. If we simply ignore the missing values, we may lose the potential valid information of the original data. In addition, a large number of data missing brings difficulties to modeling and analysis. Therefore, it is necessary to test and process the missing values of the original data set. First of all, observe the missing data, as shown in Figure 1, where red indicates data missing. It is not difficult to find that the variable X5 representing monthly income and the variable X10 of the number of family members have obvious missing values.

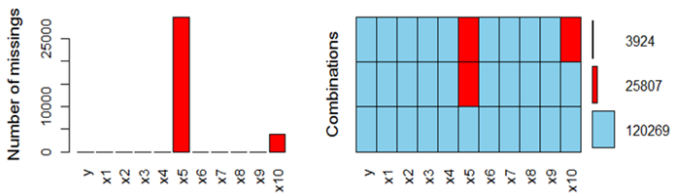

Figure 1. Distribution of data variables

There were 3924 missing values in the variable X10(2.6%) and 29731 missing values (19.8%) in variable X5. The missing proportion of variable X5 is high, but its variable meaning is monthly income, which plays an important role in modeling, and the lack of it has a greater impact on

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0019

the prediction results of the model. In order to overcome the shortage of direct deletion of data, this paper uses the data filling method to fill in the missing data. Considering the high proportion of missing data, this paper will use KNN algorithm to fill the missing data. For the missing observations that need to be filled, KNN algorithm mainly finds k nearest observations based on Euclidean distance, and then uses the data of these k nearest neighbors to get the interpolation value by distance inverse weighting method, and finally uses the value to replace the missing value.

(2) Exception handling. In the detection and processing of outliers, this paper adopts the method of combining qualitative and quantitative methods. For the numerical variables such as overdue times, non secured amount of revolving loan and loan quantity, the abnormal values of X3, X7 and variables representing overdue times are 96 and 98, so they are eliminated.

When analyzing the abnormal value of age variable, it is found that the abnormal value of the variable is 0. The age of the credit card customer is zero, which is obviously not in line with the actual situation. Therefore, the observation value of the sample with the age of 0 is eliminated. As for other characteristic variables such as income, although the data are quite different, they may exist in real life, so this paper does not make special treatment.

(3) Correlation test. The correlation of variables of personal credit original data is analyzed, as shown in the figure 2. The color represents the correlation, and the deeper the color, the higher the correlation. The correlation between the explained variable y and the characteristic variable, as well as each characteristic variable, is not high. Further through the correlation coefficient matrix, it is not difficult to find that the correlation coefficients between variables are consistent less than 0.45. The characteristic variable that is most closely related to the target variable y is X7 (overdue times in 90 days), and the correlation coefficient is 0.314. This also shows that the occurrence of customer's default behavior is closely related to the previous overdue times. The number of individual previous defaults is more, and the probability of the next loan overdue is greater. The correlation coefficient between target variable y overdue repayment and characteristic variable X2 (age) is -0.113, showing a negative correlation, which indicates that the older the customer is, the less likely the loan is to be overdue. There is also a negative correlation between monthly income and target variables, which is also in line with the actual situation. The higher the customer's income, the stronger the repayment ability, and the lower the possibility of default. In addition, because the data itself is not a lot of characteristic variables, the model processing is not complex, so this paper will not filter the features.

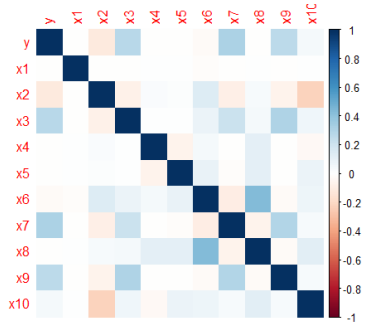

Figure 2. Thermal diagram of correlation coefficient

(4) Standardized treatment. Because there are dimensional differences among different numerical variables, which may have a great impact on the establishment of the model, it is necessary to standardize the data to eliminate the dimensional differences of variables, and the

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0019

ten data are comparable. In this paper, the normalization method is used to transform the original data into the interval [0,1]. The conversion function is as follows:

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \qquad i = 1, 2, \cdots, n \tag{10}$$

Where Xmax is the maximum value of sample data and Xmin is the minimum value of sample data.

(5) Data set segmentation. In order to train and test the model, the original data set should be divided into training data set and test data set. The purpose of the test data set is to prevent the model from over fitting the training data and to compare the prediction effect of the model. Through the comparison of test set prediction results, the optimal model is selected. In addition, the sample data set is unbalanced. 93.4% of all observations were made with the target variable series Dlqin2yrs equal to 0, while only 6.6% of the observations with Seriousdlqin2yrs equal to 1. Therefore, I use the equal proportion sampling method to divide the data set into 70% training set and 30% test set. The proportion of training set and test set is basically the same as the original.

4. THE ESTABLISHMENT OF PERSONAL CREDIT MODEL

4.1. Establishment of Logistic Model

According to the data preprocessing in the previous chapter, the training data set and test data set can be obtained. Firstly, for the training data set, the logistic regression model of all variables is carried out by GLM function. From the results of the significance test, the regression of the total variables shows that the fitting effect of the model is not very good, among which the p value of the variables X1 and X6 fails to pass the test, so the model needs to be improved and optimized.

In view of the problems, this paper considers to use the step-by-step screening method to select variables. The results are shown in table 3:

Table 3. Significance test of variables

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.822e+00	5.363e-02	-33.980	<2e-16 ***
x2	-2.855e-02	1.059e-03	-26.945	<2e-16 ***
x3	5.571e-01	1.307e-02	42.617	<2e-16 ***
x4	-2.920e-05	1.371e-05	-2.129	0.033238 *
x5	-1.126e-05	3.041e-06	-3.704	0.000212 ***
x7	8.532e-01	2.025e-02	42.127	<2e-16 ***
x8	7.576e-02	1.197e-02	6.329	2.47e-10 ***
x9	7.424e-01	2.765e-02	26.852	<2e-16 ***
x10	4.985e-02	1.173e-02	4.250	2.14e-05 ***

From the test results, compared with the full variable model, after excluding the variables X1 and X6, all the coefficients of the logistic model passed the test. At the significance level of 5%, it can be considered that the model is effective.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0019

4.2. Establishment of Decision Tree Model

In view of the problem that there are many variables and large amount of data in customer's historical credit data, this paper uses cart training algorithm to establish decision tree model, and draws the decision tree tree diagram, as shown in Figure 3,

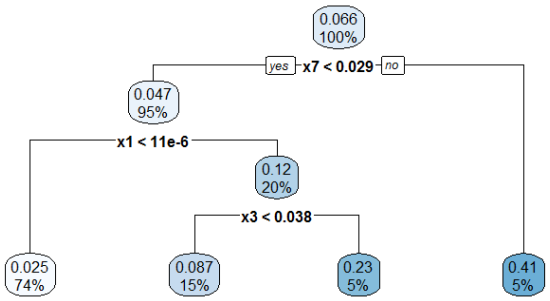

Figure 3. Decision tree graph

According to the decision tree model, the characteristic variables x1, X3, X7 have important reference value for customer credit risk identification. If the customer's historical loan is overdue for more than 90 days, it can be judged that the customer has a high probability of default again and is a potential default customer. If the customer's historical overdue times are small and have a low credit limit, it can be judged that the probability of default is small.

4.3. Establishment of SVM Model

The establishment of SVM model customer credit card data belongs to high dimension, often large amount of data, often linear inseparable. In order to solve this problem, this paper uses sigmoid kernel function to establish SVM classification model.

According to the above three models, the training set and test data set are predicted and evaluated respectively, and the ROC curve is drawn as shown in Figure 4:

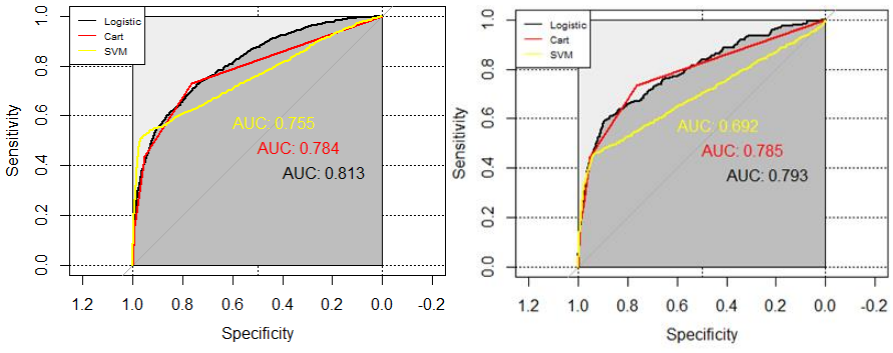

Figure 4. ROC curve of training set and test set

According to the ROC curve of a single model in the training set and test set, it is not difficult to find that the logistic model performs better, has higher classification accuracy, and has better stability. Due to the imbalance of data, the ratio of defaulting customers is low. Decision tree model and support vector machine model can deal with this kind of data in general, and the algorithm has certain limitations.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0019

4.4. Establishment of Integrated Model

The first mock exam is more sensitive and unstable than the other models. Based on the idea of ensemble learning, this paper uses stacking ensemble learning strategy to construct a credit risk identification portfolio model. According to the analysis in Chapter 3.3, compared with decision tree model and SVM model, logistic model has higher classification accuracy. In this paper, logistic regression, decision tree and support vector machine model are used as base classifiers, and logistic regression with better performance is used as sub classifier to construct an integrated model of credit risk identification.

To be specific, we first construct different base classifier models to obtain the initial overdue prediction probability. Then, the prediction probabilities of different base classifiers are taken as input variables, and the secondary classifiers are used for secondary training, and the final prediction results are obtained.

According to the ensemble learning strategy, the combination model is constructed, and the model is tested by training set and test set. The ROC curve is shown in Figure 5,

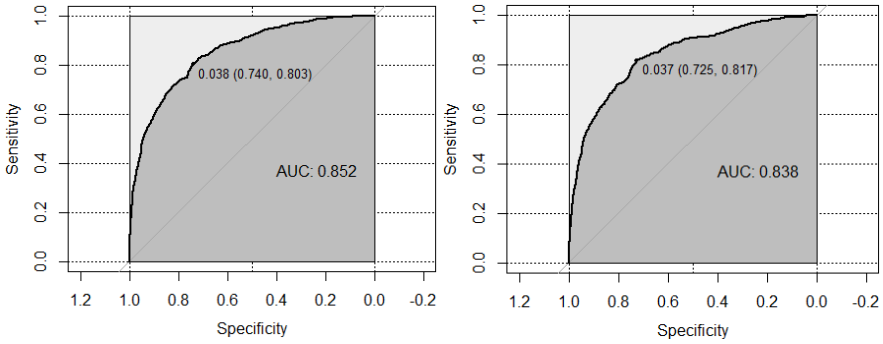

Figure 5. ROC curve of training set and test set of integrated model

According to the empirical results of different models, the classification and recognition effects of logistic model, decision tree model, SVM model and integrated model in training set and test set are arranged as follows:

Table 4. AIC value of model classification effect

Model	Train data	Test data
Logistic	0.813	0.793
Cart	0.784	0.785
SVM	0.755	0.692
Integrated model	0.852	0.838

According to the ROC curve and AUC value, the classification prediction effect of integrated learning model and single logistic regression model is good, which can ensure high AUC value in training street and test set. The function of test set is to prevent the model from over fitting and to test the robustness of the model. From the result, both models have certain generalization ability.

From the AIC value alone, the ensemble learning model performs better and has higher AIC value, which shows that the combined model has higher classification accuracy. Based on the first mock exam, the model is constructed by using the idea of ensemble learning. The combination model can effectively improve the classification accuracy and stability.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0019

5. CONCLUSION

Aiming at the problem of personal credit risk assessment, this paper uses logistic regression, decision tree model, SVM algorithm to construct classification model. From the data preprocessing, model building, empirical research, model comparison and other aspects are elaborated and analyzed. In this paper, first of all, we test whether the data is missing, and find that some variables have a high proportion of missing, and then we use k-nearest neighbor method to fill in the data; then, in the processing of outliers, we use the box chart to detect the abnormal values, and delete some abnormal values. At the same time, this paper analyzes the correlation between variables to detect whether there is multicollinearity. In the process of constructing the model, loistic model, decision tree model, SVM model and integrated learning model are established successively. In the empirical research stage, the model is trained by training set, and then the robustness of the model is tested and analyzed by using test set. Through the first mock exam of two models, we find that the integrated model has better classification effect than single model, and it can effectively improve classification accuracy and can be applied in the field of personal credit evaluation effectively.

REFERENCES

[1] Fisher R A.The use of multiple measurement in taxonomic problems[J].Annuals of eugenics,1936(7):179-188.

[2] Makowski P. Credit scoring brancher out [J]. Credit world, 1985, 75: 30-37.

[3] Baesens B,Van Gestel T, Viaene S. Benchmarking state-of-the-art classification algorithms for credit scoring[J].Journal of the operational research society,2003(54): 627-635.

[4] Kuang Nan Fang, Guijun Zhang, Huiying Zhang. An early warning method of personal credit risk based on lasso logistic model [J]. Research on quantitative economy, technology and economy, 2014,31 (02): 125-136.

[5] Wenbing Xiao, qi Fei. Research on the personal credit evaluation model and optimal parameter selection based on support vector machine [J]. System engineering theory and practice, 2006 (10): 73-79.

[6] Wenqi Sun. Research on risk control of personal credit loans of commercial banks based on logistic model [D]. Zhejiang University, 2018:11-12.

[7] Hang Li. Statistical learning methods [M]. Tsinghua University Press. 2012:55-75.

[8] Cheng Lian. Research on personal credit evaluation method of Internet Finance Based on support vector machine [D]. Zhejiang University of Finance and economics, 2017.

[9] An Chen. Research on credit card risk assessment based on machine learning [D]. Jiangxi University of Finance and economics, 2018.

[10] Huixuan Gao.Applied multivariate statistical analysis [M]. Beijing: Peking University Press, 2004 (12).

[11] Tengfei Zhang. Research on the application of data mining in P2P personal credit risk model [D]. Chongqing University, 2017.

[12] Dechuang You. Research on personal credit risk assessment based on hybrid ensemble algorithm [D]. Guangdong University of technology, 2018.

[13] Xunhan Li, Chao Li. Research on credit risk evaluation of P2P online loan borrowers based on SVM [J]. Journal of Hefei University of Technology (SOCIAL SCIENCE EDITION), 2018,32 (02): 28-34.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0019

[14] Debao Dai, Liping Ni, Ming Xue. Application of bank personal credit evaluation based on K-means and SVM [J]. Journal of Jiangsu University of science and Technology (NATURAL SCIENCE EDITION), 2017,31 (06): 836-842.

[15] Xiaoqun He. Multivariate statistical analysis [M]. Beijing: China Renmin University Press. 2004.