Research on Customer Credit Scoring Model Based on Bank Credit Card

Maoguang Wang, Hang Yang

Research on Customer Credit Scoring Model Based on Bank Credit Card

Maoguang Wang [1] and Hang Yang [1]

[1] Central University of Finance and Economics, School of Information，China
{Mgwangtiger,Yanghangv}@163.com

Abstract. With the development of China's economy, especially the maturity of the market economy, credit is important to the society and individuals. At present, credit system is mainly divided into two parts. Enterprise credit system is an important part of social credit system. But at the same time, as the foundation of social credit system, the establishment of the personal credit system is of great significance to reduce the cost of collecting information and improve the efficiency of loan processing. At the bank level, this paper discretizes the credit card data of a bank, selects the features by calculating Weight of Evidence and Information Value, and information divergence, then uses Logistic Regression to predict. Finally, the results of the Logistic Regression are transformed into visualized credit scores to establish a credit scoring model. It is verified that this model has a good prediction effect.

Keywords: Personal Credit System, Information Value, Information Divergence, Logistic Regression.

1 Introduction

The construction of the financial system in the 21st century is inseparable from the support of the credit system. The problems and risks reflected in the credit are followed, which shows that there are still some shortcomings in the credit system of our country: firstly, there is a lack of relevant detailed laws and regulations; secondly, the customer information data sets used by various enterprises are different, the reliability and quality of the data set used by credit agencies need to be improved.

As bank credit is the foundation of credit system, individual customer is the main part of bank customer group. The research of personal customer credit rating is of great significance. The establishment of Individual Credit Investigation System helps to predict risks in advance for commercial early warning analysis of banks. The Credit System presents the customer's credit report in the form of score, and the result is concise.

2 Related work

The concept of credit scoring originated from the concept of overall division put forward by Fisher (1936) in the field of statistics. Durand (1941) realized that the idea of

"division" could be used in the field of economics to divide the "good" and "bad" of loans. With the emergence of credit card, Credit Scoring gradually appears and is used in banking and other fields. In short, the development of credit scoring system is mainly divided into two stages. In the initial stage of market economy, the traditional credit scoring system is also called expert scoring system. The core of credit scoring in this way is "5C" element. With the development of economy, the modern scoring method mainly uses the skills of mathematical statistics to quantify indicators. At present, there are many credit scoring systems based on data mining and big data algorithm. Among the modern scoring methods, the first one to be used is discriminant analysis. Durand (1941) first used discriminant analysis in scoring, and fair (1958) established a credit scoring system on this basis. Myers (1963) used discriminant analysis and regression analysis to establish a credit scoring system, and predicted the credit scoring.In addition to the discriminant method, the regression analysis method is widely used. Under this method, there are many branches. For example, Henley (1995) used linear regression for credit scoring, Wigington (1980) used logistic method. At the same time, this method was still the most commonly method in credit scoring, which could overcome the defects in linear regression. In addition, Nath Jackon (1992) also applied the method of mathematical programming, but it has been proved that the effect of the method of mathematical programming is equivalent to that of the method of linear programming[1].data mining was also used in credit scoring model, which was widely used in credit decision-making and fraud prevention. In the field of data mining, Decision Tree algorithm, Neural Network algorithm, and other methods such as Support Vector Machine(SVM) and Bayesian Network could be used[2].According to the characteristics of the data set in this paper, and the characteristics that logistic regression could effectively screen variables, this paper used logistic regression to analyze the selected eigenvalues[3]. In the study of credit scoring card, most of the results[4] in recent years were expressed by the method of binary classification. This paper uses the calculation method for reference[5], transforms the logistic results into the form of scoring, breaks the situation of binary classification, and makes the results more intuitive by grading.

This paper obtains the credit card data of a bank customer for six months. After analyzing the obtained data, this paper will use the method of Weight of Evidence, Information Value and information divergence to select the logarithmic features, and consider the prior rules properly in the process of feature selection, so as to try to find the best feature extraction method. Through the comparison of multiple groups, the optimal feature system will be selected, and the selected features are input into the established Logistic Regression model. At the same time, the credit scoring model is built to convert the results into the credit score, and the customers are classified according to the scores. The classification result has value to the bank, and it is also convenient for customers to view their own credit rating.

3 Construction of Feature Selection System

Data and features determine the upper limit of experimental results. Therefore, feature engineering is important for a model algorithm. This paper mainly uses feature selection in feature engineering to process data.

3.1 Information Value

Information Value is a predictive ability to measure features, and the calculation of Information Value is based on the Weight of Evidence. Table 1 below lists the specific calculation method of WOE value[6].

Table 1. Weight of Evidence.

Formula	Meaning
$WOE_I = \ln\left(\frac{P_{y_i}}{P_{n_i}}\right) = \ln\left(\frac{\frac{y_i}{y_t}}{\frac{n_i}{n_T}}\right) = \ln(\frac{\frac{y_i}{n_i}}{\frac{y_t}{n_T}})$	P_{y_i}: Proportion of y = 1 samples in the current group to all y = 1 samples P_{n_i}: Proportion of samples with y = 0 in all samples with y = 0 y_t: Number of samples in group y = 1 y_T: Number of y = 1 in all samples n_i: Number of y = 0 in the group n_T: Number of y = 0 in all samples

From the formula in Table 1, it can be seen that the larger the woe value is, the better the prediction effect of this feature will be. However, it can also be seen that for each variable of each sample, the woe value contains plus and minus. If the woe value is used to measure the prediction ability of the whole feature, there may be a situation of positive and negative offsetting, which greatly reduces the overall prediction ability.

In order to make up for the deficiency of woe, it has been proposed that the calculation formula of IV based on woe is as follows[6].

$$IV_i = (py_i - pn_i) * WOE_i = (py_i - pn_i) * ln\left(\frac{P_{y_i}}{P_{n_i}}\right) = \left(\frac{y_i}{y_T} - \frac{n_i}{n_T}\right) * ln(\frac{y_i/y_T}{n_i/n_T}) \quad (1)$$

$$IV = \sum_i^n IV_i \quad (2)$$

Kindly According to the formula, the larger the IV is, the stronger the prediction ability of the feature is. But at the same time, in order to avoid the occurrence of extreme IV value, we need to make a reasonable discretization of the data before calculating IV.

3.2 Information Divergence

Information divergence is used to measure the contribution of a feature to the whole. It is often used for feature selection. The basis of information divergence is entropy. Entropy can be subdivided into information entropy and conditional entropy, and the calculation formula is shown in Table 2[7].

Table 2. Calculation formula of entropy.

Information Entropy	$H(S) = -\sum_{i=1}^{C} p_i log_2(p_i)$
conditional entropy	$\mathrm{H(C\|T)} = \mathrm{P(t)H(C\|t)} + \mathrm{P(\bar{t})H(C\|\bar{t})}$

The calculation of information divergence is based on conditional entropy and information entropy. The specific formula is as follows:

$$IG(T) = H(C) - H(C|T) \tag{3}$$

By writing Python program, the entropy of the whole dataset and the information divergence of each eigenvalue can be obtained.

3.3 Data Preprocessing

This paper selects the bank credit card data of a bank in Taiwan from April to September, 2005. There are 25 fields in the data set, including 23 features, as shown in Table 3 below.

Table 3. Data feature description.

Number	Feature	Concrete meaning
x_1	LIMIT_BAL	Overdraft amount
x_2	SEX	SEX
x_3	EDUCATION	EDUCATION
x_4	MARRIAGE	MARRIAGE
x_5	AGE	AGE
x_6	PAY_0	Repayment of customers in September
x_7	PAY_2	Customer repayment in August
x_8	PAY_3	Customer repayment in July
x_9	PAY_4	Customer repayment in June
x_{10}	PAY_5	Customer repayment in May
x_{11}	PAY_6	Customer repayment in April
x_{12}	BILL_AMT1	September bill amount
x_{13}	BILL_AMT2	August bill amount

x_{14}	BILL_AMT3	July bill amount
x_{15}	BILL_AMT4	June bill amount
x_{16}	BILL_AMT5	May bill amount
x_{17}	BILL_AMT6	April bill amount
x_{18}	PAY_AMT1	Repayment amount in September
x_{19}	PAY_AMT2	Repayment amount in August
x_{20}	PAY_AMT3	Repayment amount in July
x_{21}	PAY_AMT4	Repayment amount in June
x_{22}	PAY_AMT5	Repayment amount in May
x_{23}	PAY_AMT6	Repayment amount in April

Among them, sex (= 1: male, = 2: female); Education (= 1: postgraduate, = 2: University, = 3: high school, = 4: other, = 5 unknown, = 6: unknown); marriage (= 1: married, = 2: single, = 3 other); pay_0(= - 1: normal payment, = 1: delayed payment of one month, = 2: delayed payment of two months, = 8: delayed payment of eight months, = 9: delayed payment of nine months, and above).

It can be seen from the above table that there are many features in this data set, so it is important to select the most meaningful feature from many features. Considering that both IV and information divergence are statistics to evaluate the importance of features, this paper uses IV and information divergence to screen features respectively, and compares the results, in order to select the better feature selection method for this data set.

First, preprocessing the data, dealing with the missing and abnormal values. The following is the basic description of this data set(see Fig.1).

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
count	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000	30000
mean	167484.3	1.603733	1.853133	1.551867	35.4855	-0.0167	-0.13377	-0.1662	-0.22067	-0.2662	-0.2911	51223.331	49179.075	47013.155	43262.949	40311.401	38871.76	5663.581	5921.1635	5225.6815	4826.0769	4799.3876	5215.503
std	129747.7	0.489129	0.790349	0.52197	9.217904	1.123802	1.197186	1.196868	1.169139	1.133187	1.149988	73635.861	71173.769	69349.387	64332.856	60797.156	59654.108	16563.28	23040.87	17606.961	15666.16	15278.306	17777.47
min	10000	1	0	0	21	-2	-2	-2	-2	-2	-2	-165580	-69777	-157264	-170000	-81334	-339603	0	0	0	0	0	0
25%	50000	1	1	1	28	-1	-1	-1	-1	-1	-1	3558.75	2984.75	2666.25	2326.75	1763	1256	1000	833	390	296	252.5	117.75
50%	140000	2	2	2	34	0	0	0	0	0	0	22381.5	21200	20088.5	19052	18104.5	17071	2100	2009	1800	1500	1500	1500
75%	240000	2	2	2	41	0	0	0	0	0	0	67091	64006.25	60164.75	54506	50190.5	49198.25	5006	5000	4505	4013.25	4031.5	4000
max	1000000	2	6	3	79	8	8	8	8	8	8	964511	983931	1664089	891586	927171	961664	873552	1684259	896040	621000	426529	528666

Fig. 1. Basic data description

In order to achieve the better fitting effect, 30000 pieces of data are divided into training set and test set according to the proportion of 7:3.

At the same time, in order to calculate the IV of the feature, this paper combines the method of Optimal Binning and equal depth segmentation to discretize each feature of the sample, according to the AUC calculated by different segmentation methods as the measurement standard.

3.4 Feature Selection System Based on IV

After the data binning, the IV values of each features has been calculated(see Fig.2).

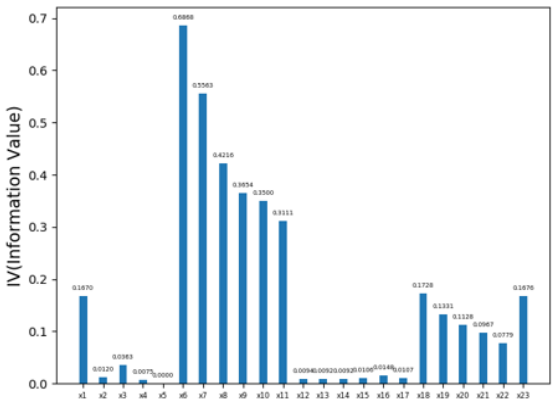

Fig. 2. Characteristic IV.

In general, the prediction ability of IV is measured according to table 4[8].

Table 4. Table captions should be placed above the tables.

IV	Predictive power
<0.02	Unpredictability
0.02-0.1	Weak prediction ability
0.1-0.3	General prediction ability
0.3-0.5	Strong prediction ability
>0.5	Suspicious

According to the above table, the variables with no prediction ability and weak prediction ability can be eliminated in this paper:$x_2, x_3, x_5, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{21}, x_{22}$

Remove variables with doubtful prediction ability:x_6, x_7.A feature selection system (A_1) is obtained. A_1 contains features:$x_1, x_8, x_9, x_{10}, x_{11}, x_{18}, x_{19}, x_{20}, x_{23}$.

Considering the influence of prior rules on data sets, it is decided to further consider the contribution of x_3, x_4, x_5 to the model on the basis of A_1 feature system, and obtain the feature system (C_1). C_1 contains features:$x_1, x_3, x_4, x_5, x_8, x_9, x_{10}, x_{11}, x_{18}, x_{19}, x_{20}, x_{23}$.

As the IV of the suspicious variable is close to 0.5, based on the feature selection system A_1, considering the influence of $x_6$and x_7 features on the model, the feature system (B_1) is obtained. B_1 contains features:$x_1, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{18}, x_{19}, x_{20}, x_{23}$.

At the same time, considering the characteristics of IV and the prior rule, the feature system (D_1) is obtained. D_1 contains features:$x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{18}, x_{19}, x_{20}, x_{23}$.

3.5 Feature Selection System Based on Information Entropy

After preprocessing the original data, the data set is input into the written Python program, and the entropy of the whole data set is 0.762353. It can be seen that the data set

of this paper is orderly and carries a lot of valuable information. Then calling the prepared function, calculating the conditional entropy of each feature, We can get the information divergence of each feature after sorting, shown in table 5.

Table 5. Information divergence of 23 features.

Feature	Information divergence
x_2	0.001
x_4	0.001
x_3	0.004
x_5	0.004
x_1	0.024
x_{11}	0.038
x_{10}	0.044
x_9	0.047
x_8	0.054
x_7	0.071
x_6	0.110
x_{22}	0.172
x_{21}	0.174
x_{23}	0.175
x_{20}	0.189
x_{19}	0.193
x_{18}	0.203
x_{17}	0.532
x_{16}	0.543
x_{15}	0.556
x_{14}	0.567
x_{13}	0.575
x_{12}	0.584

It can be seen from the table that the information divergence of the bill amount in September of x_{12} (BILL_AMT1) is the largest, which is the optimal feature. According to the number of features in A_1, B_1, C_1 and D_1 feature systems, four sets of feature systems are selected according to the information divergence. Four groups of characteristic systems are respectively recorded as A_2, B_2, C_2, D_2. A_2 selects features of the same size as A_1, and removes 14 features with small information gain. The selected feature types account for 39% of the total features. Therefore, A_2 includes features $x_{20}, x_{19}, x_{18}, x_{17}, x_{16}, x_{15}, x_{14}, x_{13}, x_{12}$. B_2 selects the number of features of the same size as B_1, removes 11 feature variables, and the selected feature types account for 52% of the total number of features. Therefore, B_2 includes$x_{21}, x_{23}, x_{20}, x_{19}, x_{18}, x_{17}, x_{16}, x_{15}, x_{14},$

x_{13}, x_{12}. C_2 selects the number of features of the same size as C_1, and removes 12 feature variables. The selected feature types account for 47% of the total number of features. C_2 includes features $x_{23}, x_{20}, x_{19}, x_{18}, x_{17}, x_{16}, x_{15}, x_{14}, x_{13}, x_{12}$.D2 selects feature numbers of the same size as D_1, and removes 9 feature variables. The selected feature types account for 60% of the total features. D_2 includes features $x_6, x_{22}, x_{21}, x_{23}, x_{20}, x_{19}, x_{18}, x_{17}, x_{16}, x_{15}, x_{14}, x_{13}, x_{12}$.

3.6 Comparison of Feature Selection System

In the first group, the WOE/IV method is compared with information divergence in feature selection. Information divergence emphasizes the feature with the greatest contribution. IV can more intuitively observe the importance of each feature. In order to verify that the information divergence and IV are more suitable for the field studied in this paper, four groups of specific comparisons are made in this paper, A_1 & A_2, B_1 & B_2, C_1 & C_2, D_1 & D_2. Among them, A_1, B_1, C_1 and D_1 are the feature systems obtained through IV, and A_2, B_2, C_2 and D_2 are the feature systems obtained based on entropy(see Fig.3). for the specific comparison. It can be seen from the figure that, in the case of the same type of sample features selected, the features constructed by IV are generally better than the model based on the features selected by information divergence.

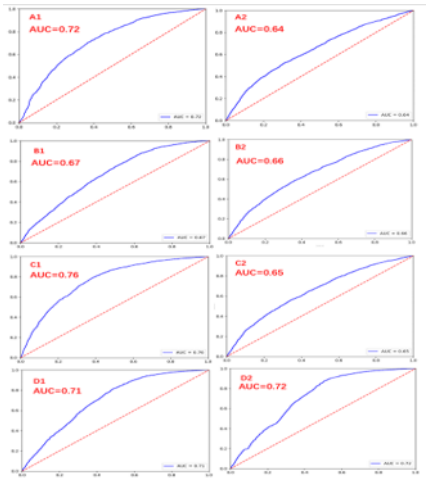

Fig. 3. Information Divergence VS. IV Results

In the second group, the AUC of A_1, B_1, C_1 and D_1 are shown(see Fig.4).It can be seen from the figure that all the models built based on C_1 have better effect. Therefore, this paper selects C_1 as the feature system of this paper, inputs Logistic Regression model and finally constructs the scoring model.

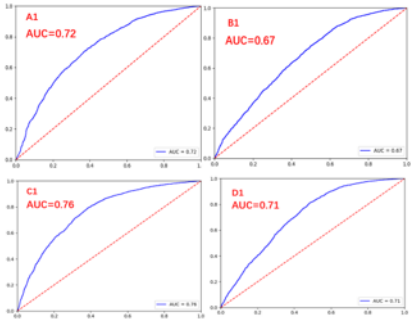

Fig. 4. AUC Based on IV Selection

In the third group, compare the four feature systems selected based on information divergence. The AUC calculated based on A_2, B_2, C_2 and D_2 is shown(see Fig.5). It can be seen from the pictures that the model effect of D_2 is the best, that is, when the selected feature types account for 60% of the total features, the model will have a better effect.

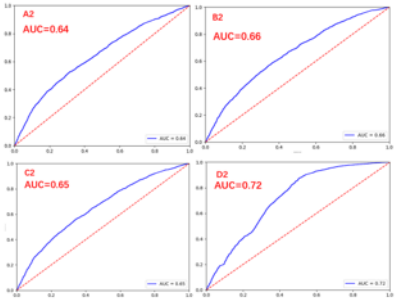

Fig. 5. AUC Based on Information Divergence

Through the above comparison, several rules with reference value can be obtained:

1. Selecting multiple indicators, the model based on IV is better than that based on information divergence.
2. In general, when the calculated IV is slightly greater than 0.5, it is better to consider the impact of this feature on the overall data.
3. When selecting features based on information divergence, the information divergence can be sorted from large to small, and the effect of selecting 60% of the total feature types is better.

4 Construction of Feature Selection System

Logistic Regression mode can explain the dependent variables, and is often used to solve the prediction problem of data subject to Normal distribution. Moreover, Logistic

Regression model overcomes the defects of linear Regression model, and has strong applicability in credit rating, which is suitable for this model. After inputting the C_1 feature system into the Logistic Regression model, and through the AUC obtained after inputting the feature into the model in the third section, it can be found that the prediction ability of this model is better. The logistic model[9] can be represented in table 6.

Table 6. Logistic Regression model.

Probability of event occurrence under the condition of characteristic x	$P(y=1\|x) = \frac{1}{1+e^{-g(x)}}$	x=$(x_1, x_2, \ldots, x_n)$ g(x)=$w_0 + w_1x_1+\ldots+w_nx_n$
Probability of event not occurring under the condition of characteristic x	$P(y=0\|x) = \frac{1}{1+e^{g(x)}}$	
Event ratio odds	odds=$\frac{P}{1-P}$ log($\frac{P}{1-P}$)=g(X)= $w_0 + w_1x_1+\ldots+w_nx_n$	

The logarithm form of odds has been obtained in the above table. In this paper, the logarithm form of probability occurrence ratio is expressed as the linear combination of feature variables, and the woe of each feature is multiplied by the regression coefficient of the variable plus the regression intercept, the scale factor is multiplied by the migration amount, and the corresponding score of each feature is obtained according to formula(4)[5], Among them, odds is the ratio of good and bad customers. Based on historical experience, this paper takes the ratio of good and bad customers as 20, and in order to make the calculated score positive, this paper stipulates that the basic score is 200 at this time, and when odds doubles, the score increases by 20, so that the calculation results of factor and offset can be obtained:

$$(woe_i * \beta_i + \frac{w_0}{n}) * \text{factor} + \frac{offset}{n} \tag{4}$$

$$factors = 20/\log(2) \tag{5}$$

$$offset = 200 - \log(20) * factors \tag{6}$$

The credit score of all characteristics of a customer is obtained by sorting out:

$$score = log(odds) * factor + offset = (\sum_{i=1}^{n} woe_i * \beta_i + w_0) * factor + offset = \sum_{i=1}^{n} woe_i * \beta_i * factor + w_0 * factor + offect \tag{7}$$

$$Basescore = w_0 * factor + offect \tag{8}$$

After calculation, factor = 28.85, offset = 113.56, base core = 154. The larger the score is, the higher the customer's credit rating is. According to the credit scoring model, the total score of customers in this paper is within the range of [0200]. After the study of customer scores, it is decided to divide customers into class I [0,40), class II [40,80),

class III [80,120), class IV [120,160) and class V [160,200] customers by using equidistant segmentation, and the customer's trustworthiness gradually decreases. The histogram of overall customer classification is shown(see Fig.6). It can be seen from the classification that category I has the most customers and category V has the least customers, indicating that most customers have high credit value.

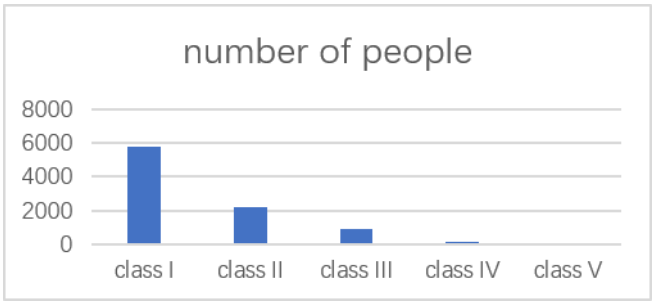

Fig. 6. Customer classification

In order to better observe the classification results, this paper makes statistics on the proportion of good and bad customers in the five categories of customers, as shown in Table 7.

Table 7. Proportion of good and bad customers in 5 types of customers.

Customer level	Proportion of good customers	Proportion of bad customers
Class I	88.81%	11.19%
Class II	68.50%	31.50%
Class III	40.02%	59.98%
Class IV	29.25%	70.75%
Class V	0%	100%

From the results of the credit scoring model in this paper, it can be seen that the two categories of customers with high credit rating, category I and II, are mostly composed of good customers, category III and IV, are mostly composed of bad customers, and category V completely untrusted customers are totally composed of bad customers. The experimental results are consistent with the actual laws, which further shows that the model in this paper has reference value.

5 Conclusion

In this paper, the bank credit card data, WOE/IV and information divergence are used for horizontal comparison. In vertical comparison, the method of feature selection considering prior rules and not considering prior rules is used. The C_1 group features of IV and prior rules are selected. 11 features are extracted from 23 features as the scoring

basis, which reduces the complexity of data processing. In the process of feature selection, after many comparative tests, this paper obtains three prior rules. Then, using logistic regression model, input the calculated woe into the model, and the AUC is 0.76. The regression coefficient of each feature is obtained, so as to build a credit rating model and get customer credit rating. Users are classified according to user rating. After comparing with the actual data, it is found that the classification results in this paper are consistent with the actual. The customer credit rating model based on the bank credit card is constructed in this paper. It makes the bank refine the customer classification through the form of generating the rating, which reflects the intuitiveness of the result for the user classification and has certain warning and reference value for the bank's credit business.

References

1. Qingyan Shi., Yunhui Jin.: A Summary of the Main Models and Methods of Personal Credit scoring[J]. Statistical Research, 2003(08):36-39.
2. Shan He., Zhendong Liu., Xiaolin Ma.: A Comparative Review of Credit Scoring Models -- Based on the Comparison Between Traditional Methods and Data Mining[J]. CREDITREFERENCE, 2019, 37(02):63-67.
3. Fengming Ma.:Study on the Application of Logistic Regression in Personal Credit Rating in China[D].Shanghai:Shanghai University of Finance and Economics,2008.
4. Yuwen, Deng.: Study on Credit Card Risk of Commercial Banks in Prefectural-Level Banks Based on Logistic Model[J]. Economic Management Journal:Chinese and English version, 2018(1):69-80.
5. Yuhua Li.:The establishment of credit scoring card model，SCIENCE & TECHNOLOGY INFORMATION,2010(13):48-49.
6. Berger.F.Gleisner.:Electronic marketplaces and intermediation:An empirical investigation of an online lending marketplace[D].University of Frankfurt,2007.
7. Jancheng Liu., Xinhua Jiang., Jinpei Wu.: The Realization of a Knowledge Reasoning Rule Induction System[J]. Systems Engineering, 2003, 21(3): 108-110.
8. Liang Shan.,Xiaolin Mao.:Modeling and application of consumer credit scoring in the Internet Finance Era.PUBLISHING HOUSE OF ELECTRONICS INDUSTRY.
9. LIANG Qi.: Distress Prediction: Application of the PCA in Logistic Regression[J]. Journal of Industrial Engineering/Engineering Management, 2005, 19(1).