Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 17 (2013) 911 – 918

Procedia
Computer Science

Information Technology and Quantitative Management (ITQM2013)

The Model and Empirical Research of Application Scoring Based on Data Mining Methods

Lai Hui[a], Shuai Li[a], Zhou Zongfang[a, *]

[a] *School of management and economics, UESTC, Chengdu, 610054 ,China*

Abstract

Personal credit scoring has played an extremely important role in the credit risk management of commercial banks. Specially, application scoring provides an important basis for the approval of customers' credit application for the first time. In this paper, firstly, the classification of the personal credit scoring is sorted out and the definition of application scoring is given; then T test method is used to do the indicators selection; further, the author establishes the static application scoring model based on the data mining methods of Logistic regression and MCLP. The results show that among the methods that used in the application scoring, the effect of MCLP is better and it's more suitable for commercial application and promotion.

Keywords: Application scoring; T test; Data mining; MCLP

1. Introduction

With the development of personal loans and credit card business, credit scoring technology has been widely applied into lending decisions, asset pricing and post-loan management of the commercial banks. Credit scoring techniques is capable to help the commercial banks reduce the artificial one-sidedness when making the loan approval decisions, thus lowering the loss. Therefore, the personal credit scoring has played a very important role in the credit risk management of commercial banks.

The research home and abroad on personal credit scoring methods mainly focuses on the discrimination analysis, regression analysis, mathematical programming and neural network methods. In the foreign literature, Durand (1941) firstly applied the discrimination analysis method into the credit risk assessment system [1]. Orgler (1970) used linear regression analysis for consumer loans credit risk assessment, and Logistic regression model was applied to study the effect of the credit score analysis by Wiginton (1980) [2-3]. Then Freed (1981) made the use of the linear programming method to do classification of personal credit risk [4]. Because of the

* Corresponding author.: Lai Hui. Tel.:1-398-176-4773; .
E-mail address: laihui198923@163.com.

doi:10.1016/j.procs.2013.05.116

powerful nonlinear mapping ability, neural network was proved to be effective to solve a lot of problems, so that it was also applicable for credit scoring as well [5]. While domestic scholar WangChunfeng successively used the discrimination analysis method, combination forecasting method and statistical methods combined with neural network technology to do empirical researches on credit scoring [6-7]. And HuangHuizhong (2010) proposed an improved LMBP algorithm in order to overcome the defects of BP neural network when it was applied in personal credit scoring model [8]. In recent years, with the development of computers, more and more data mining methods and techniques have been used in personal credit scoring. SVM and the improved SVM algorithms like ES-based Lq LS-SVM, LS-SFM, ES-ALqG-SVM in the credit assessment and analysis by Li Jianping et al [9-12]. Fang Hongquan established a data side based on 1333 actual loan records and built a two-stage credit risk assessment system by using the method of OLAM [13]. Shi extended the linear programming model based on data mining technology and proposed multiple criteria linear programming (MCLP) to do classification of real credit data [14]. On the base of MCLP, ES-MK-MCP was proposed for credit decision making by Li Jianping et al [15]. Hsieh, Lin and Sohn, respectively proposed dynamic credit scoring model based on data mining, they both first do cluster, and then built classifier[16-17].Lin combined logistic regression with neural network and proposed dual-model scoring system, which was similar to the mixed methods of static scoring[18]. Nie used logistic regression and decision trees to predict credit card churn [19]. As to personal credit scoring: firstly, the studies existed fail to subdivide the personal credit scoring further; secondly, data mining model methods research on personal credit scoring mainly focuses on the dynamic credit scoring, however, it needs to rely on historical data during a certain period of time. Additional, due to the confidentiality of personal customers in commercial banks, it makes personal credit scoring empirical studies based on data mining methods more difficult to carry out. Therefore, from another point of view, for the application scoring, the sub-types of the personal credit scoring, this paper uses data mining method to build a static application scoring model, based only on application background information as well as loan data of the personal customers when applying for loans.

This paper first classifies the kinds of personal credit scoring and gives the definition of applicant scoring; then based on real credit data of personal customers of a foreign commercial bank, application scoring model is established. As to the data analysis and indicators selection, the statistical t test method is used; and as to the model, MCLP model is used as well as the traditional Logistic regression model. The results of both models are compared and some conclusions are achieved.

2. Application scoring

2.1. Classification of the personal credit scoring

Personal credit scoring model is an advanced techniques that uses data mining and statistical analysis. Based on the analysis of the data including the customers' demographic characteristics, credit history, behavior records, trading records, we dig out the behavior patterns and credit characteristics embedded in the data [20]. Then a certain predictive model will be built after identifying the relationships between the historical information and future credit performance. And finally a fraction will be available to assess comprehensively of personal customers' credit performance in the future. Personal credit scoring is divided into three categories: credit bureau scoring, application scoring and behavior scoring. Among them, the credit bureau scoring is mainly achieved by external credit scoring companies; while application scoring and behavior scoring require commercial banks to develop their own scoring models. The following is a comparison between the classifications of personal credit scoring of commercial banks from the scoring time, data base, the range of applications.

Table.1. Classification of the Personal Credit Scoring

	Application scoring	Behavior scoring
Scoring time	A certain period after opening account	Customers over the next 6-12 months
Data basis	Background information and loan application data	Loan use, repayment history, account information and other data.
Application	First time apply for a loan approval	Post-loan management, customer relationship management

As can be seen from Table 1, we get the definition of the application scoring: the commercial banks use a certain credit scoring model to assess the probability of default based on the background information and loan application data, when personal customers apply for a loan approval for the first time after opening an account for a period. Therefore, the application scoring plays an important role in personal customers' application for loans for the first time.

2.2. Sample data for application scoring

In the application scoring model, the factors that affect the default probability of personal customers can be divided into five modules: basic conditions, solvency, life stability, credit history, guarantees. In order to study the impact of factors affecting the application scoring, this paper chooses real data of personal customers' application for loans from a foreign commercial bank (the same data in literature [8]).

2.3. Sample data preprocessing

In the sample data, various indicators in the index system have already been assigned. Before modeling based on the sample data, there is a need for cleaning the sample data, such as filling in missing values and dealing with exception values.

2.4. Indicator variable selection

The purpose of the indicator variables selection is to sort out the most predictive indicator variables, and delete indicator variables with weak predictive ability, thus reducing the number of candidate indicator variables, however, improving the effect of the model we need to build.

2.4.1. T test

T-test is to judge whether the means of the variables in default and non-default are equal, based on the prerequisites that the variables overall follow a normal distribution. It is intended to see whether the variables are predictive. Assuming H^0 is that there is no significant difference between default and non-default sample means, namely there is no significant difference between the difference mean with 0. We build statistics

$$t = \frac{\bar{D}}{S/\sqrt{n}} \qquad （1）$$

$\bar{D}$ is the mean difference, S is the sample variance. Taking the confidence level with $\alpha = 0.05$, we reserve the variables whose p -value is less than α.

2.4.2. Indicator selection

The original variables are X_1 Age in years, X_2 Personal status and sex, X_3 Number of people being liable to provide maintenance for, X_4 Job, X_5 Present employment since, X_6 Housing, X_7 Present residence since, X_8 Installment rate in percentage of disposable income, X_9 Property, X_{10} Status of existing checking account, X_{11}Other installment plans, X_{12} Credit amount, X_{13} Savings account/bonds, X_{14} Duration in month, X_{15} Credit history, X_{16} Number of existing credits at this bank, X_{17} Other debtors / guarantors. The independent samples of the 17 variables were tested with the T test method. Table 2 is the result of the variables whose p -value is more than $\alpha = 0.05$:

Table.2. Variables with p -value > $\alpha = 0.05$

Variables	X_2	X_3	X_4	X_6	X_7	X_8	X_{11}	X_{17}
P-value	0.7264	0.9018	0.5092	0.4358	0.9654	0.9250	0.2413	0.2851

The variables whose p -value is more than α will be deleted, therefore, the following 9 variables are reserved: X_1 Age in years, X_5 Present employment since, X_9 Property, X_{10} Status of existing checking account, X_{12} Credit amount, X_{13} Savings account/bonds, X_{14} Duration in month, X_{15} Credit history, X_{16} Number of existing credits at this bank.

3. Model methods

For personal credit scoring model algorithms, firstly, the training process neural network is very complex and it is difficult to explain; secondly, the decision tree has better interpretability, however, due to over-fitting or unreasonable pruning, its accuracy performs poorly on test data set; then linear regression, K-Nearest Neighbors accuracy is lower compared with other model algorithms, but Logistic regression has better performance. Among the data mining methods: due to the characteristics of simple and straightforward modeling process, the flexibility to modify the parameters, its applicable to multi-class classification problem, and higher classification accuracy, multi-objective linear programming model (MCLP) has been widely used , this paper will make use of Logistic regression and MCLP to build model respectively.

3.1. Logistic regression

Defining: $y = 1$ represents default and $y = 0$ is non-default. This method uses the sample data to build model, in order to predict the default probability of the borrower. In the logistic regression model, set:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_k X_k \qquad (2)$$

p is the probability of $y = 1$; X_k is the variable describing the characteristics of the borrower; $\frac{p}{1-p}$ is occurrence ratio.

3.2. MCLP

MCLP uses the multi-objective linear programming approach to seek for the optimization combination of the scoring weight. Following is the pattern of the method:

$$Min(d_\alpha^{\ +} + d_\alpha^{\ -})^p + (d_\beta^{\ +} + d_\beta^{\ -})^p \tag{3}$$

Subject to

$$\alpha^* + \sum_i \alpha_i = d_\alpha^{\ -} - d_\alpha^{\ +}$$

$$\beta^* - \sum_i \beta_i = d_\beta^{\ -} - d_\beta^{\ +}$$

$$A_i X = b + \alpha_i - \beta_i, A_i \in G$$

$$A_i X = b - \alpha_i + \beta_i, A_i \in B$$

where A_i, α*, and β* are given, X and b are unrestricted, and $\alpha_i, \beta_i, d_\alpha^{\ -}, d_\beta^{\ +}$, $d_\alpha^{\ +}, d_\beta^{\ -} \geq 0$. α_i is the distance of that misclassified sample points deviate from the demarcation plane; while β_i is the distance of correctly classified points to the adjusted demarcation plane. α^* and β^* are the target values of α_i and β_i. $d_\alpha^{\ -}, d_\beta^{\ +}$, $d_\alpha^{\ +}, d_\beta^{\ -}$ are used to measure the extent that the actual value deviates from the target value. By determining the parameters that minimize the sum of the two distances that deviate from the target, the model may obtain an optimal sorting plane. In the prediction of the probability of default of personal customers, the customers are divided into default and non- default categories. Attributes A_i represents the values of the indicator variables of each personal customers, and the vector X represents the weight of each indicator variable, and then through maximum likelihood method, customers' default probability p can be estimated.

4. Analysis of results

4.1. Application scoring model based on Logistic regression

60% of the data are randomly selected as training samples of the model, while the remaining 40% as the testing samples. By means of WEKA software, do Logistic regression to the remaining 9 variables. The following model can be established based on the results of the regression:

$$\lg(\frac{p}{1-p}) = -1.4686+0.0122X_1+0.2364X_5-0.209X_9+0.6192X_{10}+0.1507X_{13}-0.0305X_{14}+0.4023X_{15}-0.3372X_{16} \tag{4}$$

The running results of WEKA software can be obtained in Table 3:

Table.3. Indicators of the classification results of the application scoring model based on Logistic regression

	Actual value	0	1	Prediction accuracy	Sensitivity	Specificity	ROC value
Predictive value	0	257	36	77．25%	0.591	0.824	0.773
	1	55	52				

4.2. Application scoring model based on MCLP

Set p=1, α*=0.001 and β* =10000000，b=1[21]，then the running results of WEKA software can be obtained in Table 4:

Table.4. Indicators of the classification results of the application scoring model based on MCLP

	Actual value	0	1	Prediction accuracy	Sensitivity	Specificity	ROC value
Predictive value	0	261	32	0.805	0.656	0.85	0.813
	1	46	61				

4.3. Conclusions of the Application scoring model

In application scoring model, T test method is done respectively to the sample data of indicator variables。And finally 9 variables are remained for the model, the possible explanations are: firstly, the variables, like X_2 Personal status and sex, X_3 Number of people being liable to provide maintenance for, X_7 Present residence since, do little impact to the result of the application scoring model; secondly, the correlation between part of the variables is too strong, such as X_4 Job and X_5 Present employment since, X_9 Housing and X_6 Property, thus one can be directly instead of them.

As can be seen from Table 3 and Table 4, based on the same sample data, under the condition that the number of variables is big while the correlation of the variables is strong, Application scoring model based on MCLP performs better than traditional Logistic model, no matter in the overall prediction accuracy rate or ROC value.

5. Conclusions

From the empirical analysis of this paper, the following conclusions can be achieved:

Through T test, the variables remained in the model have a good interpretation of realistic meanings. Through the comparison of the application scoring model based on MCLP and Logistic regression, the MCLP-based application scoring model has a better overall performance than the traditional logistic regression model. Therefore, the MCLP-based application scoring model is more suitable for commercial application and promotion. However, this paper still has some deficiencies: first, the model is built based on a small sample data, the assuming of T test that the sample data approximately follows a normal distribution is an idealized situation, so the model’s stability and reliability needs to be tested further. Therefore, the next step in the research work mainly focuses on the collection of a large sample data on one hand; and on the other hand, this paper just do the research work on static personal application scoring, the next work is to establish dynamic model based on dynamic data from personal accounts of domestic commercial banks, namely personal behavior scoring is the future direction.

Combined the conclusions of this paper with the actual situation of our country, in terms of the application scoring, the following two suggestions are proposed: firstly, when considering the indicators that affect the result of application scoring, we should select the indicators with the characteristics of comprehensive， independent and predictive, based on statistical analysis; secondly, commercial banks, credit bureaus and industry bodies should continue to collect and improve the customer application data, only in that way, thus making it possible for commercial banks to develop their own internal scoring method.

Acknowledgements

This research has been supported by National Natural Science Foundation of China (No.71271043), the special research Foundation of Ph.D program of China（No.20110185110021）and SiChuan Province Science and technology support project (No.2012SZ0001).

References

[1] Durand,D. Risk elements in consumer instalment financing. National Bureau of Economy Research, New York;1941, p.189-201.

[2] Orgler, Y. E.A Credit Scoring Model for Commercial Loans, Journal of Money, Credit, and Banking; 1970,2, Iss.4, p. 435-445.

[3] Wiginton, J. C. A note on the comparison of logit and discriminate models of consumer credit behavior. Journal of Financial and Quantitative Analysis; 1980,15,p. 757-770.

[4] Freed, N.& Glover, F. A linear programming approach to the discriminant problem. Decision Sciences ;1981,(12),p.68-74.

[5] Cheng B. & Titterington. D. M.. Neural Networks: A Review from a Statistical Perspective. Statistical Science; 1994, 9,p.:2-30.

[6] Wang Chunfeng, Wan Haihui, ZhangWei. Application of Combining Forecasts in Credit Risk Assessment in Banks. Journal of Industrial Engineering and Engineering Management;1999, 13(1),p.5-8

[7] Wang Chunfeng, Wan Haihui, ZhangWei. Credit Risk Assessment in Commercial Bank Using Neuralworks. Systems Engineering-Theory&Practice; 1999, 13(1),p.5-8

[8] Huang Huizhong, Zhou Zongfang, Yu Jike. LMBP neural network model and its applications in personal credit scoring, Management scientists; 2010,p. 1-3.

[9] Li, J., J. Liu, W. Xu and Y. Shi. Support Vector Machines Approach to Credit Assessment, in P. M. A. Sloot et al, eds., ICCS 2004, LNCS 2658, Springer, Berlin; 2004,p. 892-899.

[10] Liwei Wei, Zhenyu Chen, Jianping Li. Evolution Strategies Based Adaptive Lp LS-SVM. Information Sciences; 2011, 181 (14),p. 3000–3016.

[11] Jianping Li, Zhenyu Chen , Liwei Wei, Weixuan Xu and Gang Kou. Feather Selection via Least Squares Support Feature Machine. International Journal of Information Technology & Decision Making; 2007, 6(4),p. 671-686.

[12] Jianping Li, Gang Li, Dongxia Sun, Cheng-Few Lee. Evolution Strategies Based Adaptive Lq Penalty Support Vector Machine with Gauss Kernel for Credit Risk Analysis. Applied Soft Computing; 2012, 12(8),p. 2675-2682.

[13] Fang Hongquan, Zengyong. An Application of On-line Analytical Mining in Database to CreditRisk Assessment in Commercial Banks. China Soft Science; 2004,(10),p.126-130.

[14] Y. Shi, Y. Peng, W. X. Xu and X. W. Tang. Data mining via multiple criteria linear programming: Applications in credit card portfolio management. International Journal of Information Technology & decision making; 2002, 1(1),p. 131-151.

[15] Jianping Li, Liwei Wei, Gang Li and Weixuan Xu. An Evolution-Strategy Based Multiple Kernels Multi-criteria Programming Approach: The Case of Credit Decision Making. Decision Support Systems; 2011, 51(2),p.292-298.

[16] N. C. Hsieh. An integrated data mining and behavioral scoring model for analyzing bank customers. Expert Systems with Applications; 2004, 27(4),p.623-633.

[17] M. K. Lim, S. Y. Sohn. Cluster-based dynamic scoring model. Expert Systems with Applications; 2007, 32(2), p.427-431.

[18] Y. C. Lin. Improvement on behavior scores by dual-model scoring system. International Journal of Information Technology & Decision Making; 2002, 1(1), p. 153-164.

[19] G. Nie, W. Rowe, L. Zhang. et al. Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications; 2011, 38(12), p. 15273-15285.

[20] WangYin, Nie Guangli, Y. Shi. An Research on Customers Default Rate of Commercial Banks in China Based on Credit Scoring Models, Management Reviews; 2012, 24(2), p. 111-119

[21] Y. Shi,Zhong Huayi, Zhang Jianjun, Liu Yuji. Multiple Criteria Linear Programming Decision System: Theory and Applications, Higher Education Press; p.290-300.