Original Research Article

Assessment and early warning of personal credit default behavior: Analysis based on logistic regression model

Yusheng Kang, Xianghui Li*, Hanyan Zhang
School of Mathematics and Statistics, Henan University of Technology, Zhengzhou, Henan, 450001, China

Abstract: With the rapid development of digital finance, the assessment and early warning of personal credit default behavior have become critical components of financial institutions' risk control. Based on a real-world credit card transaction dataset containing over 65,000 records, this study employs the information gain method to select key predictive indicators and applies a logistic regression model to evaluate and provide early warnings for credit default behavior.The findings are as follows: First, the core factors influencing personal credit default behavior are primarily concentrated on borrowers' historical credit records and debt status, rather than on demographic characteristics. Second, the logistic regression early warning model constructed in this paper achieves a sound balance between predictive accuracy and risk identification capability, successfully identifying all defaulting clients. This result indicates that the model can effectively control the type of misclassification that carries the highest risk cost in financial risk management—Namely, misclassifying a "bad" client as a "good" one—Thereby demonstrating high practical value.Therefore, the personal credit risk early warning model based on logistic regression and the information gain method can accurately identify credit default behaviors among high-risk borrowers while maintaining interpretability. It is recommended that financial institutions, in their risk control practices, place particular emphasis on borrowers' historical credit blemishes, large outstanding debts, and records of bounced checks, supplemented by influencing factors such as age and household size.
Keywords: credit default; information gain; logistic regression; demographic characteristics

1. Introduction

Driven by mobile Internet, big data, and AI, society has entered the digital era. The integration of the Internet and finance has given rise to Internet finance, which enriches financing channels for enterprises and individuals, especially SMEs that face high barriers under traditional banks. Digital lending still faces risks such as information asymmetry and borrower default, harming investors and disrupting financial order. Scholars worldwide have conducted extensive research on big data-based personal credit assessment. In early years, credit evaluation mainly depended on experts' subjective judgment. Fisher [1] proposed the linear discriminant method and Durand [2] applied it to loan risk assessment. Statistical methods were later widely used in credit evaluation models. Orgler [3] first introduced linear regression into personal credit assessment, while Wiginton [4] pioneered the use of Logistic regression in this field. Many domestic researchers [5-8] combined Logistic regression with the information gain method to select evaluation indicators. In the big data era, machine learning algorithms have been gradually introduced into credit risk modeling. Zhang Likun [9] improved the BP neural network for credit assessment. Lin et.al. [10] proved that the random forest model has better stability than neural networks and support vector machines.

Credit default refers to failure to meet repayment obligations, divided into corporate and personal categories. This paper focuses on personal credit risk, reflected in repayment willingness (subjective) and capacity (objective)—Key factors in lending decisions. Information asymmetry and personal factors are major causes. China's credit system remains imperfect; some borrowers falsify or hide information on online platforms, putting investors at a disadvantage and increasing default risk. Some defaults also arise from intentional non-repayment or ignorance of credit rules.Thus, accurate personal credit risk assessment is essential. Using credit card data covering fraud and default, this paper builds a Logistic regression-based credit evaluation and early warning model for risk control.

2. Data preprocessing and indicator selection

2.1. Data preprocessing

This study adopts credit card transaction data from a financial institution, containing over 60,000 valid records. The dataset covers 26 indicators falling into four categories: personal basic information, financial status, credit records and household information. Customer category is defined as the dependent variable: Customers with overdue payments or bad debts are labeled as defaulting ("bad") customers, while others are non-defaulting ("good") customers. Indicators including blood type, zodiac sign, religious belief, customer number and application source are eliminated for having no practical correlation with credit default. The remaining 21 variables are quantified, as presented in **Table 2.1**.

Table 2.1. Credit card customer credit evaluation indicators and quantification grouping.

Variable Name	Meaning	Quantification Grouping
Defective Account	Defective credit record	Yes=1, No=0
Loan Balance	Balance over 8 million yuan	Yes=1, No=0
Bounced Check	Bounced check record	Yes=1, No=0
Refusal Record	Credit refusal record	Yes=1, No=0
Compulsory Card Suspension	Forced card suspension record	Yes=1, No=0
Number of Cards	Quantity of credit cards	1=1, 2=2, 3=3, 4=4, >4=5
Usage Frequency	Card usage frequency	Daily=1, Often=2, Occasionally=3, Rarely=4, Never=5
Household Registration	Residential region	North=1, Central=2, South=3, East=4
Urbanization Degree	Regional urbanization level	Metropolitan=1, City=2, Town=3
Gender	Gender	Female=1, Male=2
Age	Age groups	15-19=1, 20-24=2, 25-29=3, 30-34=4, 35-39=5, 40-44=6, 45-49=7, 50-54=8, 55-59=9
Marital Status	Marital status	Unmarried=1, Married=2, Other=3
Education Level	Educational background	Primary & below=1, Junior High=2, Senior/Vocational=3, College=4, Bachelor & above=5
Occupation	Occupation types	Coded from 1 to 22 for different groups
Personal Monthly Income	Personal monthly income	Coded from 1 to 8 by income ranges
Personal Monthly Expenses	Personal monthly expenses	Coded from 1 to 5 by expense ranges
Housing	Housing condition	Rental=1, Dormitory=2, Self-owned=3, Spouse-owned=4, Parents-owned=5, Other=6
Family Monthly Income	Household monthly income	Coded from 1 to 6 by income ranges
Monthly Card Swipe Amount	Monthly card spending	Coded from 1 to 8 by amount ranges
Household Population	Number of co-residents	1=1, 2=2, 3=3, 4=4, 5=5, 6=6, 7=7, 8=8, ≥9=9
Family Economic Status	Household economic level	Upper=1, Upper-middle=2, Middle=3, Lower-middle=4, Lower=5
Customer Category	Customer type	Good customer=0, Bad customer=1

2.2. Indicator selection using information gain method

The original dataset includes 65,535 samples: 59,683 good customers and 5,852 bad customers. We split the data into a 7:3 training set and test set via R language. To ensure model effectiveness, 10,000 non-default samples and 2,500 default samples are selected from the training set for modeling.Information gain measures an indicator's classification capability. A higher information gain value means the variable exerts a stronger influence on customer classification. The core formula is defined as: $Gain(A)=I-E(A)$ where I refers to total information entropy and E(A) stands for conditional entropy. Taking gender as an example for calculation, the information gain of gender is 0.0014083.

We calculate the information gain of all 21 indicators with R language, and the results are shown in **Table 2.2**. In existing research, variables with an information gain above 0.01 are regarded as significant influencing factors.Finally, nine significant indicators are screened out: Defective Account, Loan Balance, Bounced Check, Refusal Record, Compulsory Card Suspension, Age, Occupation, Monthly Card Swipe

Amount and Household Population. These variables are adopted for subsequent empirical modeling.

Table 2.2. Information gain results for various indicators.

Indicator Name	Information Gain	Indicator Name	Information Gain
Defective Account	0.3173537	Education Level	0.0041873
Loan Balance	0.5892362	Occupation	0.0271123
Bounced Check	0.6151745	Personal Monthly Income	0.0049599
Refusal Record	0.4969194	Personal Monthly Expenses	0.0099303
Compulsory Card Suspension	0.70061681	Housing	0.0039261
Number of Cards	0.0024568	Family Monthly Income	0.0081524
Usage Frequency	0.0028261	Monthly Card Swipe Amount	0.02081524
Household Registration	0.0024836	Household Population	0.0175528
Urbanization Degree	0.002323	Family Economic Status	0.0039968
Gender	0.0014083	Age	0.0208188
Marital Status	0.0009216	—	—

3. Establishment of personal credit default assessment model

3.1. Binary logistic regression model

This study adopts the binary Logistic regression model for binary classification of customer default status, where the dependent variable is coded as 1 for default and 0 for normal repayment. Since the default probability p ranges from 0 to 1, a Logit transformation is applied. The core formula is as follows

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Let

$$\mathrm{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

Then:

$$p = \frac{e^{\alpha+\beta_1 X_1+\cdots+\beta_k X_k}}{1+e^{\alpha+\beta_1 X_1+\cdots+\beta_k X_k}}$$

3.2. Construction of the logistic regression model

Multicollinearity among independent variables will undermine model accuracy. We conducted a multicollinearity test on the nine selected indicators via R language with the cor and kappa functions. Several variables show high correlation coefficients above 0.9, yet the kappa value is 157 (below 1000), which means only slight multicollinearity exists. It is acceptable for prediction−oriented models, so subsequent regression analysis can proceed. The results show that the correlation coefficients between variables X2(Loan Balance), X3 (Bounced Check), and X4 (Refusal Record) are higher than 0.9, indicating the presence of multicollinearity. However, the kappa value is 157, which is less than 1000, suggesting that severe multicollinearity does not exist. When the model is used only for prediction, slight multicollinearity is tolerable. Therefore, further construction of the Logistic regression model for research is feasible.

We set customer type Y as the dependent variable (0 = good customer, 1 = bad customer) and adopted stepwise regression to screen insignificant variables. After removing the insignificant variableX5 (Compulsory Card Suspension) and X4 (Refusal Record), seven indicators remain significant at the 0.1 significance level. The final regression results are displayed in **Table 3.1**.

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_{11} + \beta_6 X_{14} + \beta_7 X_{19} + \beta_8 X_{20}$$

The regression results are shown in **Table 3.1** below:

Table 3.1. Model regression results.

Variable	B	S.E.	Z value	Sig.
Constant	-2.76139	0.65515	-4.215	2.5e-05
X1	4.51436	0.33382	13.523	<2e-16
X2	6.20341	0.47283	13.120	<2e-16

Continued Table 3.1

Variable	B	S.E.	Z value	Sig.
X3	7.40345	0.41141	17.995	<2e-16
X11	-0.15668	0.06953	-2.253	0.0242
X14	0.05813	0.03246	1.791	0.0733
X19	-0.20484	0.11620	-1.763	0.0779
X20	-1.15979	0.15985	-7.255	4.0e-13

The final Logistic regression equation is:

$$\ln\left(\frac{p}{1-p}\right) = -2.76139 + 4.51436X_1 + 6.20341X_2 + 7.40345X_3 - 0.15668X_{11} + 0.05813X_{14} - 0.20484X_{19} - 1.15979X_{20}$$

Regression results indicate that defective account records, excessive loan balances, and bounced check records significantly raise default risks. In contrast, older age and larger household size are associated with a lower probability of default.

3.3. Model evaluation

We further verified the model with 5,000 test samples (see **Table 3.2**). The model achieves a prediction accuracy of 99.80% for good customers and 97.97% for bad customers. Notably, it identifies all defaulting customers accurately, which is critical for credit risk control. Overall, the model presents reliable predictive capability.

Table 3.2. Model prediction results.

Prediction / Actrual	Good Customer (0)	Bad Customer (1)	Total
Good Customer (0)	4556	9	4565
Bad Customer (1)	0	435	435
Total	4556	444	5000

4. Conclusion

Based on the above research findings, the following conclusions can be drawn: First, personal credit default behavior is primarily driven by historical credit records and debt status. Variables such as "defective account status," large outstanding loan balances, and "record of bounced checks" significantly increase the risk of default. In contrast, age and household size are negatively correlated with default risk, indicating that behavioral characteristics have greater discriminatory power for credit risk than demographic attributes. Second, the logistic regression model demonstrates satisfactory early-warning performance. On the test set, the model achieved an identification accuracy of nearly 99.8% for non-defaulting clients and 97.97% for defaulting clients, successfully identifying all defaulters. This effectively controls the type of misclassification that carries the highest risk cost in financial risk management.

In summary, the logistic regression model adopted in this paper offers both interpretability and predictive accuracy. It is recommended that financial institutions, in their risk control practices, place particular emphasis on borrowers' historical credit blemishes, large outstanding debts, and bounced check records, while also considering age and household size as auxiliary influencing factors.

Acknowledgements

This research was supported by the Young Key Teachers Program of Henan University of Technology.

About the author

*Coresponding author: Xianghui Li, Email: xianghuili_1107@163.com.

References

[1] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.

[2] Durand, H. F., & Soons, S. G. (1941). Repudiation of family trust settlements. Columbia Law Review, 41(2).

[3] Orgler, Y. E. (1970). Practical financial management models: Reply. Management Science, 16(12).

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

[5] Xu, S. F., & Wang, Y. C. (2003). Logistic model in personal credit evaluation. Journal of Tianjin Institute of Light Industry, (S1), 46-49. (in Chinese)

[6] Chen, X. L. (2020). Research on credit default risk assessment model of P2P borrowers based on Logistic regression model [Master's thesis]. Shanghai International Studies University. (in Chinese)

[7] Wang, M. J. (2015). Credit risk assessment of P2P lending platform borrowers based on Logistic regression model [Master's thesis]. Beijing Foreign Studies University. (in Chinese)

[8] Yu, G. L. (2017). Research on credit risk assessment of P2P online lending platform borrowers based on Logistic regression model [Master's thesis]. Lanzhou University. (in Chinese)

[9] Zhang, L. K. (2016). Research on personal credit risk assessment based on P2P online lending platform [Master's thesis]. Harbin University of Science and Technology. (in Chinese)

[10] Lin, C. D., & Peng, G. L. (2007). Application of random forest in determining the index system for corporate credit evaluation. Journal of Xiamen University (Natural Science Edition), 46(2), 199-203. (in Chinese)