Research on the Prediction Model of Borrowers' Severe Delinquency Behavior Based on Multi-Dimensional Credit Indicators

Zihong Chen [1*]

[1] School of Economics, Guangdong Peizheng College, Guangzhou, China

Abstract. In the context of the expansion of the consumer credit market, borrowers' credit behaviors have become diversified and complicated. This study is committed to integrating multi-dimensional credit indicators such as revolving credit utilization rate, debt-to-income ratio, and overdue records of different periods to build a high-precision prediction model, so as to evaluate the probability of borrowers' severe delinquency behavior in the next 2 years. The research uses 16,715 personal loan records from openML with dataset ID 46543 as data support, applies a linear regression model for analysis, and carries out the research by defining variables such as credit risk, debt repayment risk, and credit activity-to-age ratio. The results show that variables such as credit activity-to-age ratio, real estate debt dependence, and severe delinquency prediction have a significant linear relationship with the dependent variable, and there is no multicollinearity among the variables. Finally, the corresponding regression equation is obtained. This study provides a scientific basis for financial institutions' risk management, credit approval, and post-loan monitoring, and helps to reduce the non-performing loan ratio and optimize the allocation of credit resources.

1 Introduction

In the wave of digital transformation in the financial industry, the consumer credit market is experiencing rapid expansion, and borrowers' credit behaviors have also shown a trend of diversification and complication. The revolving credit utilization rate can reflect the intensity of borrowers' use of credit lines, the debt-to-income ratio can be used to measure borrowers' repayment ability, and overdue records of different periods can reveal their historical credit risks. However, most of the current related studies focus on single indicators or simple linear combinations, lacking systematic integration and in-depth mining of multi-dimensional credit data. At the same time, the rapid development of machine learning and big data technology has provided new methods and ideas for complex credit risk modeling, making it an urgent need to build a prediction model that combines traditional indicators and advanced algorithms to better adapt to the dynamically changing credit risk environment. Accurately assessing credit risk is crucial to preventing systemic financial risks. The traditional credit evaluation method relies on a single or a small number of indicators, which makes it difficult

* Corresponding author: faopa@ldy.edu.rs

to fully capture the complex characteristics of credit risk. Building a prediction model by integrating multi-dimensional credit data can more comprehensively depict the credit status of borrowers, which is of great significance for financial institutions to reduce the non-performing loan ratio, optimize the allocation of credit resources, and effectively improve their risk control capabilities and market competitiveness. Based on this, this study focuses on multi-dimensional credit indicators, aiming to integrate indicators such as revolving credit utilization rate, debt-to-income ratio, and overdue records of different periods, and use linear regression model to build a high-precision prediction model to evaluate the probability of borrowers' severe delinquency behavior in the next 2 years, to provide a scientific basis for financial institutions' risk management, credit approval and post-loan monitoring.

2 Data collection plan

2.1 Data and its statistical characteristics

This study uses quantitative data covering borrowers' basic information, credit behavior data, and risk labels. The data comes from 16,715 personal loan records with dataset ID 46543 in openML. Among them, the revolving credit utilization rate reflects the intensity of borrowers' use of credit lines through the ratio of balance to credit limit. Age, to a certain extent, reflects the financial stability cycle of borrowers [1]. The number of overdue days of 30-59 days, 60-89 days, and more than 90 days, respectively, reveals the historical credit risks of different periods [2]. The debt-to-income (or asset) ratio is used to measure the debt burden of borrowers. Monthly income reflects the repayment ability [3]. The number of outstanding credit lines and loans reflects the credit activity. The ratio of the number of real estate loans or credit lines to the total number of outstanding credit lines and loans (adding 1 to the denominator to avoid zero) is used to measure the degree of dependence on real estate debt. The number of people dependent on the borrower can help judge the family's economic pressure. The target variable (0 means no, 1 means yes) of whether severe delinquency behavior occurs in the next 2 years is used as the dependent variable.

2.2 Selection of variables and their significance

In the financial field, the assessment of borrowers' credit risk is crucial for financial institutions' risk management, credit approval, and post-loan monitoring. In order to accurately predict whether borrowers will have severe delinquency behavior in the next 2 years, this study selects variables from multiple dimensions, comprehensively considers key factors such as borrowers' credit status, repayment ability, and credit history, and recombines them into the following 5 new variables.

(1) Credit Risk Indicator (CRRI) is the credit risk indicator, which is used to reflect the current degree of borrowers' use of credit lines. This indicator holds that the higher the utilization rate, the greater the repayment pressure the borrower may face, which is of great significance for credit risk assessment. Its calculation method is (rev_util × 0.3) + (late_30_59 × 0.2) + (late_60_89 × 0.25) + (late_90 × 0.25) [4].

(2) Debt Repayment Risk Indicator (DCRI) is the debt repayment risk indicator, which measures the debt burden of borrowers by means of the debt-to-income (or asset) ratio (debt_ratio). Its calculation method is debt_ratio × [1 + (late_30_59 + late_60_89 + late_90) / 10] [5].

(3) Credit Activity-to-Age Ratio (CIAR) refers to the credit activity-to-age ratio, in which age can reflect the financial stability cycle of borrowers to a certain extent. This indicator is calculated by the ratio of open_credit to age, that is, open_credit / age [6].

(4) Real Estate Debt Dependence (RERD) represents real estate debt dependence, which is the ratio of the number of real estate loans or credit lines (real_estate) to the total number of outstanding credit lines and loans, and can reflect the degree of borrowers' dependence on real estate debt. To avoid the denominator being zero, 1 is added to the denominator. The specific calculation method is real_estate / (open_credit + 1) [7].

(5) Severe Delinquency Prediction Indicator (DLQ_PRED) is the severe delinquency prediction indicator. Its theoretical basis is that past overdue records are an important basis for predicting future defaults. The longer the overdue time and the greater the number of overdue items, the greater the possibility of future severe delinquency. At the same time, the higher the monthly income, the stronger the repayment ability, and the lower the possibility of severe delinquency. Its calculation method is (late_30_59 × 0.1) + (late_60_89 × 0.2) + (late_90 × 0.3) + (monthly_inc / 10000) × (- 0.4) [8].

In credit risk assessment, original variables such as revolving credit utilization rate and overdue records of a certain period can only reflect one aspect of borrowers' credit status and cannot fully capture complex risk characteristics [9]. The purpose of recombining variables is to build comprehensive indicators with stronger explanatory power by integrating information from multiple dimensions [10]. Combining revolving credit utilization rate, the number of overdue accounts of 30-59 days, 60-89 days, and more than 90 days into a credit risk indicator according to different weights can comprehensively reflect the intensity of borrowers' credit use and overdue situations of different degrees. The debt repayment risk indicator combines the debt-to-income ratio with the sum of overdue records of different periods, reflecting the interaction between debt burden and default history. The credit activity-to-age ratio, through the ratio of the number of outstanding credit lines and loans to the borrower's age, explores the relationship between credit activity and the borrower's financial stability cycle. The real estate debt dependence indicator measures the degree of borrowers' dependence on real estate debt by the ratio of the number of real estate loans or credit lines to the total number of outstanding credit lines and loans. The severe delinquency prediction indicator integrates overdue history and repayment ability, which enhances the predictive effect of historical default behavior and repayment ability on future risks. These variables can improve the accuracy of the model in predicting borrowers' severe delinquency behavior in the next 2 years through the logical correlation between variables, and provide a more systematic evaluation basis for financial institutions' risk management.

3 Model construction and analysis

3.1 Model construction

This study uses a linear regression model for analysis, taking multi-dimensional credit indicators such as CRRI, DCRI, CIAR, RERD, and DLQ_PRED as independent variables, and taking whether borrowers have severe delinquency behavior in the next 2 years (target variable DLQ_2yrs, 1 for occurrence and 0 for non-occurrence) as the dependent variable to build the model.

3.2 Analysis method

In the process of model analysis, the significance of variable coefficients (judged by whether the P-value is less than 0.05) is tested to evaluate the impact of each indicator on the dependent variable, and Variance Inflation Factor (VIF) analysis is used to judge the multicollinearity between variables (taking VIF value greater than 10 as the judgment threshold for severe multicollinearity), to ensure the reliability of the model and the independence of indicators.

4 Empirical analysis

4.1 Coefficient analysis

The regression coefficient results are shown in Table 1. This table presents the regression coefficients, standard errors, t-values, P-values, and 95% confidence intervals of each independent variable (credit risk indicator CRRI, debt repayment risk indicator DCRI, credit activity-to-age ratio CIAR, real estate debt dependence (RERD), severe delinquency prediction indicator DLQ_PRED) on the dependent variable (whether severe delinquency behavior occurs in the next 2 years Dlq_2yrs), which can be used to judge the significance and degree of impact of each variable. It can be seen from Table 1 that the P-value of the credit activity-to-age ratio (CIAR) is 0.000, which is less than 0.01, indicating that it is significant at the 1% level, with a coefficient of 0.18915, showing a positive impact. The P-value of real estate debt dependence (RERD) is 0.005, which is less than 0.01, significant at the 1% level, with a coefficient of-0.09449, showing a negative impact. The P-value of the severe delinquency prediction indicator (DLQ_PRED) is 0.000, less than 0.01, also significant at the 1% level, with a coefficient of 0.01293, showing a positive impact. However, the P-value of the credit risk indicator (CRRI) is 0.488, and the P-value of the debt repayment risk indicator (DCRI) is 0.820, both greater than 0.1, indicating that these two variables have no significant impact on the dependent variable.

Table 1. Regression coefficient table

Dlq_2yrs	Coefficient	Std.err	t	P>|t|	[95& conf. interval]	
CRRI	-0.00004	0.00006	-0.69	0.488	-0.00017	0.00008
DCRI	-0.00000	0.00002	0.23	0.820	-0.00004	0.00005
CIAR	0.18915	0.03418	5.53	0.000	0.12216	0.25615
RERD	-0.09449	0.03361	-2.81	0.005	-0.16038	-0.02861
DLQ_PRED	0.01293	0.00091	14.26	0.000	0.01115	0.01470
_cons	0.04724	0.00794	59.46	0.000	0.45679	0.48793

4.2 Multicollinearity test

The results of the variance inflation factor (VIF) test are shown in Table 2. This table shows the VIF values and 1/VIF values of each variable, which are used to judge whether there is multicollinearity between variables. It can be seen from Table 2 that the VIF values of all variables are between 1.00 and 1.03, and the average VIF value is 1.01, which is much less than 10. These indicate that there is no severe multicollinearity between the variables, the variables are relatively independent and the results of the regression model are highly reliable.

Table 2. Variance inflation factor table.

Variable	VIF	1/VIF
DLQ_PRED	1.03	0.9
CIAR	1.02	0.98
RERD	1.01	0.99
CRRI	1.01	0.99
DCCI	1.00	0.99
Mean VIF	1.01	

5 Conclusion

In the regression model for exploring the severe delinquency behavior of borrowers in the next 2 years, the credit activity-to-age ratio (CIAR), real estate debt dependence (RERD), and severe delinquency prediction indicator (DLQ_PRED) are all significant at the $P < 0.01$ level, indicating that the three have substantial explanatory power for the equation. CIAR (coefficient 0.189) has a significant positive impact, indicating that the higher the adaptability between credit activity and age (such as the middle-aged group with active credit), the lower the probability of severe delinquency of borrowers in the future, which verifies the hypothesis that "the matching degree between credit behavior and life cycle affects the default risk". RERD (coefficient-0.094) has a significant negative impact, indicating that the higher the proportion of real estate debt, the greater the possibility of borrowers falling into default due to factors such as asset depreciation and superimposed repayment pressure. DLQ_PRED (coefficient 0.013) is highly significant with a positive coefficient. Its design integrating overdue history and repayment ability accurately captures the interaction of "default inertia + repayment resources" and is the core variable for predicting severe delinquency in the equation. In contrast, the credit risk indicator (CRRI) and debt repayment risk indicator (DCRI) are not significant at the $P > 0.1$ level, indicating that their explanatory power for the equation is weak. This may be due to the low correlation between borrowers' credit use and overdue behavior in the sample, single debt structure and other factors. It is necessary to optimize the indicator design (such as introducing debt type segmentation, income volatility and other variables) to improve the explanatory power.

In the table, the VIF values of DLQ_PRED (severe delinquency prediction indicator), CIAR (credit activity-to-age ratio), RERD (real estate debt dependence), CRRI (credit risk indicator), and DCRI (debt repayment risk indicator) are all between 1.00 and 1.03, 1/VIF is close to 1, and the average VIF is 1.01. This indicates that there is almost no multicollinearity among all independent variables, the variable combination is logically independent, and can truly reflect the causal relationship of "independent variables to future severe delinquency behavior". The regression results are reliable and the model stability is high, which verifies the effectiveness of the construction of credit risk composite variables.

References

1. S. Agarwal, et al., The age of reason: Financial decisions over the lifecycle (2007)
2. N. Teixeira, T. Costa, I. Lisboa, Managing customer credit to reduce the company's risk and overdue credits (2024)

3. P.A. McCoy, S.M. Wachter, Why the ability-to-repay rule is vital to financial stability, Geo. LJ **108**, 649 (2025)
4. J. Wang, L. Yu, Study on risk assessment model of commercial banks based on credit risk degree, J. Manag. Eng. **4**, 85-90 (2021)
5. X. Ye, Study on loss given default (LGD) in bank credit risk management, Shanghai Jiao Tong University (2022)
6. W. Wu, P. Rong, Q. Xu, Health and household asset selection, Econ. Res. J. **46**, 43-54 (2021)
7. Z. Huang, Interactive mechanism and empirical analysis of credit constraints, real estate prices and macroeconomy, Zhejiang University (2019)
8. Y. Sun, Study on optimal credit strategies of banks for small, medium and micro enterprises under information asymmetry - Default rate measurement model based on Logistic regression, Financ. Dev. Res. **6**, 78-84 (2021)
9. N. Chen, B. Ribeiro, A. Chen, Financial credit risk assessment: a recent review, Artif. Intell. Rev. **45**, 1-23 (2023)
10. G. Heinze, C. Wallisch, D. Dunkler, Variable selection-a review and recommendations for the practicing statistician, Biom. J. **60**, 431-449 (2018)