Proceedings of the 3rd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/151/2024.19327

Probabilistic Statistical Models and Their Applications in Economic Issues

Bokun Dong[1,a,*]

[1]*Southwestern University of Finance and Economics, Chengdu, 611100, China*
a. 1979480683@qq.com
**corresponding author*

Abstract: With the rapid development of global economies, loans have become one of the key driving forces for economic growth. However, the outbreak of the subprime mortgage crisis served as a wake-up call for governments worldwide. To prevent malignant financial crises triggered by large-scale debt defaults, financial institutions in various countries should strengthen their regulatory mechanisms for financial loans. Based on this premise, this paper employs four probabilistic statistical models—Logistic Regression, Naive Bayes, Cox Proportional Hazards Model, and Chi-Square Test—to analyze the Credit Risk Dataset from the Kaggle platform. The study delves into the impact of various variables on loan defaults, the origins of credit risk, and its potential effects on the financial system. The results indicate that variables such as the loan-to-income ratio and loan interest rates significantly influence the risk of loan defaults. By comparing different models, this paper provides practical and effective evidence for financial risk regulation, enhancing the accuracy and reliability of credit risk assessments. It offers a new perspective for ensuring the long-term stable development of the economy.

Keywords: Probability Theory, Statistical Models, Economy, Credit Loans, Risk Control.

1. Introduction

Against the grand backdrop of economic globalization, central banks worldwide leverage interest rate adjustments to manage their nations' economic recessions and growth, rendering interest rates a crucial component of financial markets globally. Concurrently, with the rapid economic development of various countries, an increasing number of individuals tend to invest in and consume assets such as real estate and automobiles through loans. Consequently, personal loan consumption has evolved into a significant driving force for economic development. However, as loan services expand, credit issues have become increasingly severe, with large-scale loan defaults posing a significant risk of triggering financial crises. For instance, the 2007 U.S. subprime mortgage crisis was a direct result of massive mortgage defaults. According to the Global Financial Stability Report released by the International Monetary Fund (IMF) on May 9, 2008, the financial turbulence induced by the U.S. subprime mortgage crisis was projected to cause nearly a trillion dollars in global losses and was spreading from the subprime mortgage sector to areas such as prime mortgages, consumer credit, and corporate credit[1]. This highlights the critical need to assess the credit risk of loan customers, which has become a pressing challenge in the financial industry.

DOI: 10.54254/2754-1169/151/2024.19327

In 1941, David Durand was the first to propose the use of discriminant analysis for credit evaluation, developing a quantitative scoring method based on borrowers' personal information[2]. Classical risk assessment models can be broadly categorized into two types: one involves rating (scoring) methods, and the other focuses on measuring default risk indicators. However, as personal loan credit risk is influenced by multiple factors, it does not adhere to a simple linear relationship or remain constant. Since the actual circumstances of borrowers change continuously over time, traditional credit scoring methods are inadequate for handling multidimensional data and complex economic environments. Therefore, this paper will conduct analysis and discussion using probabilistic models in economic problems, employing Logistic Regression, Naive Bayes, Cox Proportional Hazards Model, and Chi-Square Test to analyze the dataset in depth. The aim is to comprehensively evaluate the default probability of loan customers and explore the relationship between personal loan credit and default.

This provides financial institutions with more reliable evidence for credit approval and risk management, enhances the stability and security of economic systems, and holds significant theoretical and practical value.

2. Probabilistic Statistical Models in Economic Issues

In the study of personal credit risk and economics, probabilistic statistical models provide essential inferential methods for exploring and understanding complex variable characteristics and economic phenomena. This paper will use four models—Logistic Regression, Naive Bayes, Cox Proportional Hazards Model, and Chi-Square Test—to deeply examine their mathematical principles and applications in economic problems.

2.1. Logistic Regression Model

The Logistic Regression model is an improvement over the Linear Regression model and is a type of "generalized linear regression model." It is one of the most commonly used statistical models for classification problems[3]. Its core idea is to linearly weight the variables and map the results to a probability range [0,1] through a logistic function, thereby estimating the relationship between the independent and dependent variables. The specific expression of logistic regression is shown in Equation (1):

$$P(y = 1|x) = \frac{1}{1+e^{-\left(\beta_0+\sum_{i=1}^{k}\beta_i x_i\right)}} \quad (1)$$

Here, $P(Y = 1|X)$ represents the probability of the event occurring; β_0 is the intercept, $\beta_1, \ldots, \beta_p$ are regression coefficients, and $X_1, \ldots, X_p$ are feature variables. The S-shaped curve characteristic of the logistic function allows it to compress an unbounded linear combination into a probability value. Using maximum likelihood estimation, the model can maximize the probability of the observed data under the model. Due to the strong interpretability of the logistic regression model, it can provide the marginal effects of each independent variable and is therefore widely applied in personal credit evaluation and default risk assessment to probabilistically predict individual credit ratings and default probabilities.

2.2. Naive Bayes Model

The Naive Bayes Classifier (NBC) originates from classical mathematical theory, with a solid mathematical foundation and stable classification efficiency.[4]. Its core principle is based on Bayes' theorem and the assumption of conditional independence among features. It uses Bayes' theorem to

DOI: 10.54254/2754-1169/151/2024.19327

calculate the posterior probability of each class given a sample and selects the class with the highest posterior probability as the predicted output.The formula is shown in Equation (2):

$$P\left(y|x_1, x_2, \ldots, x_k\right) = \frac{P(y)\cdot P\left(x_1,x_2,\ldots,x_k|y\right)}{P\left(x_1,x_2,\ldots,x_k\right)} \tag{2}$$

Here, the prior probability usually comes from historical experience or records. For example, when assessing personal credit risk, the individual's historical default probability is often used for calculation. Since Bayes' theorem combines prior and posterior probabilities, it avoids subjective bias caused by relying solely on prior probabilities and also prevents overfitting that may result from using only sample information. Therefore, Naive Bayes has relatively high accuracy. Additionally, due to its fast computation speed and strong robustness, Naive Bayes is widely applied in scenarios involving high-dimensional data.

2.3. Cox Proportional Hazards Model

The Cox Proportional Hazards Model was proposed by British statistician D.R. Cox. It is used to analyze the effects of multiple factors on survival time, can handle censored survival data, and does not require estimation of the survival distribution. It is currently one of the most commonly used multifactor analysis methods in academia[5]. Since censored data exists in personal credit datasets and the distribution type of default time is unknown, we choose to use the Cox Proportional Hazards Model to analyze various factors influencing personal credit risk. This allows for the analysis of time-related events (such as default or repayment time) and their relationship with one or more risk factors of interest.

By specifying the parameter form of each covariate and the unrestricted baseline hazard function $h_0(t)$, the model focuses on proportional hazard modeling to study the relationship between various factors and survival time.Its general expression is shown in Equation (3):

$$h(t|x) = h_0(t) \cdot \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right) \tag{3}$$

Based on the flexibility of the Cox model and its low dependency on data distribution, it can be used in credit scenarios to handle censored data. By analyzing customer characteristics (such as credit scores or repayment history) and their relationship with default time, it provides financial institutions with dynamic risk warnings.

2.4. Chi-Square Test

The Chi-Square Test, as a type of hypothesis test, is a non-parametric test method based on the chi-square distribution[6] It examines whether there is an association between two categorical variables by comparing the differences between observed and expected values. If the difference is significant, it is considered that there is a statistical dependency between the two variables. The definition of the chi-square statistic is shown in Equation (4):

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

The basic logic of the Chi-Square Test is to construct a null hypothesis (usually assuming the variables are independent) and then calculate the chi-square statistic. By comparing it with the critical value in the chi-square distribution table, one determines whether to reject the null hypothesis. In personal credit analysis, the Chi-Square Test can be used to examine the associations between factors affecting loans. This helps risk managers understand the potential impacts of different factors on risk, enabling them to make more accurate decisions.

DOI: 10.54254/2754-1169/151/2024.19327

3. Empirical Analysis

3.1. Data Selection

This study utilizes the Credit Risk Dataset from the Kaggle platform. The dataset contains 32,581 records, including information such as personal demographics, loan amounts, and loan history. The specific variables and their meanings are detailed in Table 1:

Table 1: Variables and Their Meanings

Variable Name	Explanation
person_age	Individual's age
person_income	Individual's annual income
person_home_ownership	Homeownership status (e.g., owned, rented, etc.)
person_emp_length	Employment length (in years)
loan_intent	Purpose of the loan (e.g., education, medical, etc.)
loan_grade	Loan grade, typically based on the borrower's credit status
loan_amnt	Loan amount
loan_int_rate	Loan interest rate
loan_status	Loan status, 0: normal repayment, 1: default
loan_percent_income	Percentage of loan amount relative to individual income
cb_person_default_on_file	Indicator of default history in credit bureau records (yes/no)
cb_person_cred_hist_length	Credit history length (in years)

3.2. Result Analysis

3.2.1. Logistic Regression and Naive Bayes

For the methods of Logistic Regression and Naive Bayes, this paper employs accuracy, confusion matrix, and ROC curve with AUC value for analysis. Figures 1-4 respectively illustrate the confusion matrices, ROC curves, and AUC values of the two methods.

The confusion matrix evaluates the classification ability of the models, showing their predictive performance across different categories. From the confusion matrix, the model's accuracy can be calculated to evaluate the proportion of correctly predicted samples out of the total samples. The calculation is shown in Equation (5):

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{5}$$

The ROC curve shows the changes in the true positive rate and the false positive rate under different thresholds. AUC, the area under the ROC curve, ranges from [0,1] and is used to measure the overall discriminatory power of the model. A higher AUC value indicates better model performance.

From the figures, it can be observed that the accuracy of the Logistic Regression model is 0.80, with an AUC value of 0.76. The Naive Bayes model achieves an accuracy of 0.81 and an AUC value of 0.77. Overall, the Naive Bayes model performs better than the Logistic Regression model.

DOI: 10.54254/2754-1169/151/2024.19327

The confusion matrix shows that Logistic Regression has a high correct recognition rate for non-defaulting customers (TN = 7491) but weaker recognition ability for defaulting customers (TP = 339), resulting in a higher miss rate. Therefore, Logistic Regression is better at predicting non-defaulting customers.

In contrast, Naive Bayes improves predictions for positive samples, while Naive Bayes enhances the identification of defaulting customers, it introduces more misclassifications in non-defaulting customer predictions.

In summary, Naive Bayes outperforms Logistic Regression overall. Logistic Regression is more suitable for predicting non-defaulting customers and controlling false positive rates, whereas Naive Bayes is better suited for scenarios focusing on high-risk customers.

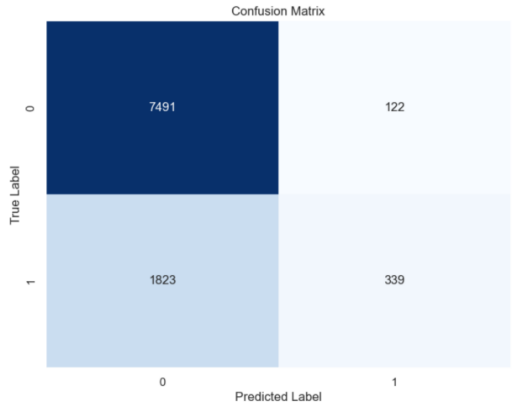

Figure 1: Logistic Regression Confusion Matrix

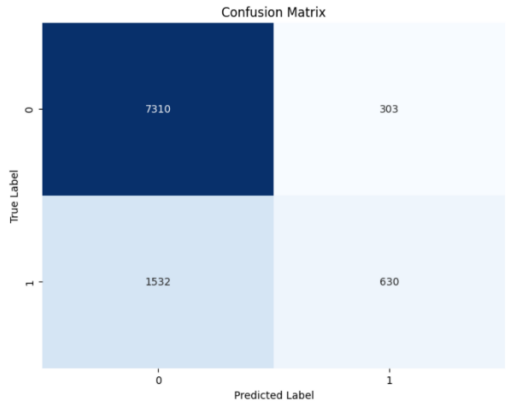

Figure 2: Naive Bayes Confusion Matrix

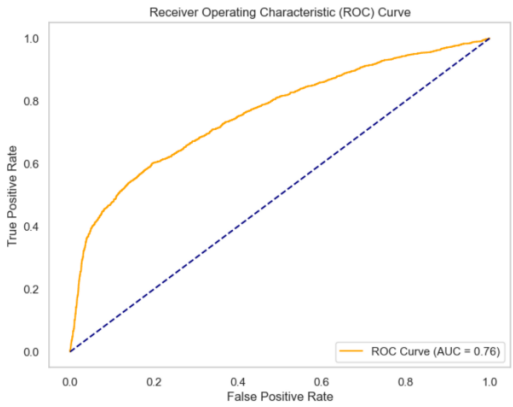

Figure 3: Logistic Regression ROC Curve

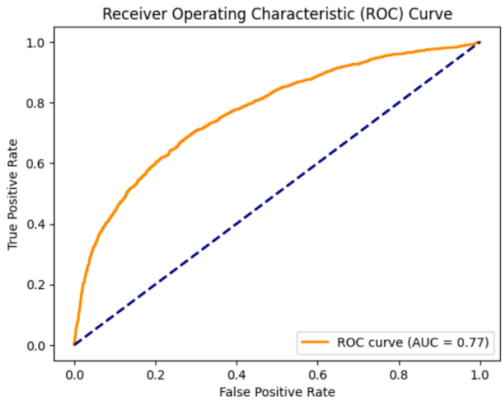

Figure 4: Naive Bayes ROC Curve

3.2.2. Cox Proportional Hazards Model

The results of the Cox Proportional Hazards Model illustrate the influence of multiple covariates on default time. Tables 3 through 6 present evaluations and interpretations of the Cox model results from different metrics. The explanations and meanings of the evaluation metrics for the Cox Proportional Hazards Model are summarized in Table 2:

DOI: 10.54254/2754-1169/151/2024.19327

Table 2: Explanation and Meaning of Cox Proportional Hazards Model Metrics

Metric	Meaning	Interpretation of Values
coef	Direction and magnitude of covariate's impact on default risk	Negative: Protective effect; Positive: Risk increases
exp(coef)	Relative impact of the variable on default risk	exp(coef)>1: Default risk increases; exp(coef)<1: Default risk decreases
se(coef)	Measures the uncertainty of the estimate	Larger values indicate greater uncertainty
p-value	Tests the significance of covariates' impact on default risk	p-value<0.05: Significant impact on risk

Table 3: Summary of Cox Model Coefficients

covariate	coef	exp(coef)	se(coef)	p-value
person_age	-0.252299	0.777013	0.004317	0
person_income	-0.000087	0.999993	1e-06	3.776e-18
loan_amnt	-0.000018	0.999982	4e-06	2.916e-06
loan_int_rate	0.19401	1.214109	0.004002	0
loan_percent_income	4.789	120.201	0.180706	9.055e-155

Table 4: Likelihood Ratio Test Results

Test Statistic	p-value	Degrees of Freedom
11694.82	<0.005	5

Table 5: Schoenfeld Residual Test Results

covariate	test_statistic	p-value
loan_amnt	42.13	<0.005
loan_int_rate	1.76	0.18
loan_percent_income	43.7	<0.005
person_age	282.46	<0.005
person_income	72.85	<0.005

By combining Table 2 and Table 3, it can be observed that negative coefficient variables (e.g., person_age and loan_amnt) indicate that an increase in these variables is associated with a reduction in default risk. Specifically, as an individual's age increases or the loan amount rises, the risk of default decreases. Conversely, positive coefficient variables (e.g., loan_int_rate and loan_percent_income) are positively correlated with an increase in default risk. Notably, the risk ratio (exp(coef)) for loan_percent_income is 120.201, suggesting that an increase in the loan-to-income ratio significantly raises default risk. Based on the p-values and the absolute values of the coefficients, the loan-to-income ratio is undoubtedly the most critical predictive variable. Its impact on risk far outweighs other variables, underscoring the vital role of the relationship between loan amount and income in assessing personal credit risk. Additionally, loan interest rates and age also have significant effects on default risk, with higher interest rates increasing default risk and advancing age reducing it.

From Table 4's results, the model's goodness of fit and accuracy can be effectively assessed. When the p-value is less than 0.005, it indicates that the overall model fit is significant. This means that the selected covariates in this study have statistical significance in explaining default time and can

DOI: 10.54254/2754-1169/151/2024.19327

effectively account for variations in default time. Additionally, the Harrell's C-index value is 0.8708, falling within the range [0.5,1] and closer to 1, which reflects the model's high predictive accuracy and discriminative ability. The model can accurately predict the level of default risk and effectively distinguish individuals at different risk levels.

Furthermore, the Schoenfeld residual test is used to verify the proportional hazards assumption of the Cox model. From Table 5, it can be seen that, except for loan_int_rate, the p-values of all other variables are less than 0.005. Therefore, the experimental variables in this study satisfy the proportional hazards assumption, and the model performs well in explaining default risk. Particularly, the analysis of the impact of the loan-to-income ratio on default risk exhibits high significance and predictive capability. This further demonstrates the effectiveness and reliability of the model in personal credit risk assessment.

3.2.3. Chi-Square Test

Table 6 presents the results of the Chi-Square Test. The Chi-Square statistic in this context measures the degree of deviation between observed values and expected values. A larger Chi-Square statistic indicates a stronger association between variables. From the table, it is evident that loan_grade (loan rating) has a Chi-Square statistic of 5609.18, significantly higher than that of other variables. In contrast, loan_intent (loan purpose) has a Chi-Square statistic of 520.51, which is much smaller than that of loan_grade, suggesting a weaker association between loan purpose and loan status. Although loan purpose may have some impact on loan status, its influence is relatively minor compared to loan rating. Thus, the Chi-Square Test clarifies the strength of associations between various variables and loan status, providing a critical basis for further analysis of personal loan credit risk.

Next, the p-value is used to determine whether the relationships between variables are statistically significant. A p-value less than 0.05 is considered significant. In this analysis, all variables have p-values smaller than 0.05, indicating that the relationships between all variables and loan status are statistically significant.

Additionally, Cramér's V is a measure of effect size, ranging from [0, 1], representing the strength of association between variables. From the table, it can be seen that loan_grade has a Cramér's V value of 0.415, the only value greater than 0.3, indicating that loan rating is one of the most critical factors influencing loan status. Lower ratings are typically associated with higher default risk. Other variables exhibit moderate correlations with loan status.

Table 6: Chi-Square Test Results

Variable	Chi-Square Statistic	p-value	Cramér's V
person_home_ownership	1907.980698188821	0	0.24199410325695123
loan_intent	520.5115614374077	2.980681669776041e-110	0.12639589959746061
loan_grade	5609.184186567319	0	0.414923130194783
cb_person_default_on_file	1044.4395947711112	3.934660154785392e-229	0.17904387098348779

3.2.4. Correlation Between Feature Variables and Loan Status

Combining the four methods and experimental results, this study uses point-biserial correlation coefficients and absolute correlation to measure the relationship between features and loan status, as illustrated in Figures 5 and 6.

The point-biserial correlation coefficient is used to assess the linear relationship between loan status and other features (e.g., income, loan ratio). A higher positive correlation coefficient indicates a greater likelihood of loan default, while a higher negative correlation coefficient suggests a reduced

DOI: 10.54254/2754-1169/151/2024.19327

likelihood of default. From Figure 5, among the 12 related variables, loan_percent_income and loan_int_rate show significant positive correlations with default risk. This indicates that higher loan-to-income ratios and interest rates are associated with a greater likelihood of default. Conversely, features such as loan_grade_A show negative correlations with default risk, suggesting that higher credit ratings help reduce the likelihood of default.

In addition, absolute correlation converts all correlation coefficients to positive values, disregarding their direction, to measure the strength of the association between features and the target variable. This allows for a more straightforward comparison of the impact strength of different features on loan status. From Figure 6, loan_percent_income, loan_int_rate, and loan_grade_D have the highest absolute correlations with loan status, indicating that they are critical factors influencing default risk.

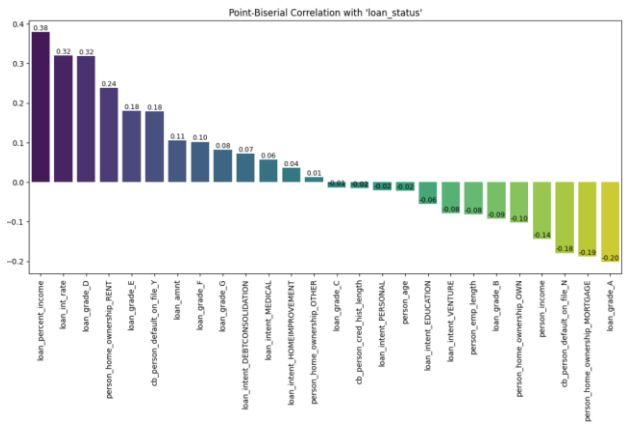

Figure 5: Point-Biserial Correlation Between Features and Loan Status

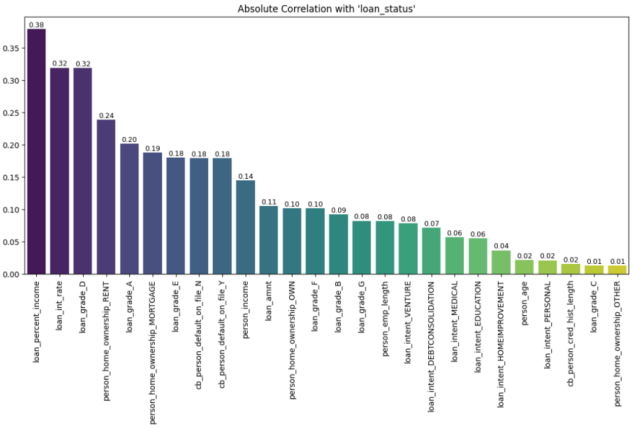

Figure 6: Absolute Correlation Between Features and Loan Status

4. Conclusion

This study employs methods such as Logistic Regression, Naive Bayes, Cox Proportional Hazards Model, and Chi-Square Test to conduct a multi-perspective quantitative analysis and research on personal loan default risks. The results show that different models have their respective strengths and weaknesses in risk assessment and can be combined for in-depth analysis from multiple perspectives. The Logistic Regression model, with its strong interpretability, has significant advantages in the field of risk control, especially in predicting non-defaulting customers. The Naive Bayes model significantly improves recall rates and excels in identifying high-risk customers, making it suitable for high-risk customer scenarios. Furthermore, the results of the Cox Proportional Hazards Model reveal time-related credit risk factors, particularly showing high significance and predictive power when analyzing the impact of loan amounts and income proportions on default risks. The Chi-Square Test is used to verify the statistical association between loan characteristics and default status, clearly identifying the significant impact of key variables on defaults.

In summary, credit risk ratings should prioritize loan_percent_income (loan amount as a percentage of income), loan_int_rate (loan interest rate), loan_grade_D, and historical default records (cb_person_default_on_file) as key predictive variables. For borrowers with low ratings and a history of defaults, stricter loan approval processes should be implemented. Additionally, measures such as increasing interest rates or shortening repayment periods can be applied to high-risk borrowers to implement stronger risk management.

Through the integrated application of different statistical models, this study provides a more comprehensive credit risk assessment approach, enhancing the accuracy of identifying and predicting

DOI: 10.54254/2754-1169/151/2024.19327

loan default risks. It addresses the deficiencies in existing theories, provides empirical evidence for financial institutions in credit approval and risk management, and contributes to the long-term stable development of the economy, holding significant practical guiding value.

References

[1] *Wang Chenghao, Li Jianqiang, Dong Jichang. Insight into Real Estate Credit Risk Prevention from the U.S. Subprime Mortgage Crisis[J]. Systems Engineering Theory & Practice, 2010, 30(3): 437-446.*

[2] *Wang Rui. Prediction Model for Repayment Probability of Overdue Customers in Microfinance[D]. Tianjin: Tianjin University of Commerce, 2018.*

[3] *Zhang Libin, Wu Zongwen. Comparison between Credit Scoring Card Based on XGBoost Machine Learning Model and Logistic Regression Model[J]. Journal of South-Central University for Nationalities (Natural Science Edition), 2023, 42(6): 846-852. DOI:10.20056/j.cnki.ZNMDZK.20230616.*

[4] *An Yingbo, Zhang Yujing, Zhang Jiannan. Research on Credit Risk Early Warning of Rural Banks Based on Naive Bayes[J]. Wireless Internet Technology, 2015(22): 109-111. DOI:10.3969/j.issn.1672-6944.2015.22.048.*

[5] *Liu Junpeng. Research on Personal Credit Risk Measurement Based on Actuarial Methods[D]. Henan: Zhengzhou University, 2018.*

[6] *Zhang Wen, Chen Jianping, Liu Cunhe, et al. Application of Chi-Square Test in Statistical Homogeneity Zoning of Fractured Rock Mass[J]. Chinese Journal of Geotechnical Engineering, 2011, 33(9): 1440-1446.*