Frontiers in Economics and Management Volume 4 Issue 12, 2023
ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0035

Design and Development of Personal Credit Score Card based on Logistic Regression Model

Jiayue Ma

School of Business, Xianda College of Economics & Humanities Shanghai International Studies University, Shanghai, China

*mjysknsyl@163.com

Abstract

Based on the analysis of China's economic trends, the scale of the credit market, business credit risks, and individual consumption concepts, a modeling approach using big data technology is proposed, taking into account the existing personal credit data protection modes and the status of the personal credit system in various commercial banks and financial institutions. However, the current statistical and artificial intelligence methods have improved the accuracy of the models, but their robustness and generalization ability are not ideal. Therefore, it is necessary to address the existing issues. Simply optimizing the model is not enough, and the selection and preprocessing of indicators are also crucial. Based on literature research on models, this paper concludes that the logistic regression model has advantages in terms of data requirements, regulatory requirements, explanatory requirements, and application breadth. Therefore, this model is chosen as the foundation for designing and developing a personal credit scoring card. Finally, after analyzing the advantages and disadvantages of the established credit scoring card model, future research directions and trends of big data technology in financial credit are discussed, and specific recommendations are proposed.

Keywords

Credit Score Card; Logistic Regression Model; Credit Risk; Credit Assessment; Machine Learning.

1. Introduction

In recent years, credit consumption has gradually become one of the main businesses for various financial institutions in China [10], and the scale of various personal credit consumer loans is rapidly expanding [1]. According to statistics, China's personal credit has been growing at an annual rate of 20%, and it is estimated that the scale of personal credit will reach 41.1 trillion yuan by 2019. At the same time, financial institutions are facing increasingly serious personal credit risk issues.

The rise of financial technology and internet finance in China has brought opportunities and challenges to the personal credit scoring business. The application of big data in the field of financial risk control has demonstrated the prospects of addressing credit risks. With the development of social informatization, the number of market-oriented credit reporting agencies is increasing, and the indicators for personal credit scoring are also emerging, relying on an increasing amount of data. Currently, China's personal credit scoring system mainly relies on traditional data such as personal financial activity information, which is controlled by commercial banks and the central bank's credit reporting center. Therefore, it is of great significance to utilize existing customer historical information for effective personal credit assessment.

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0035

Based on the above background and significance, this paper designs and develops a personal credit scoring card model using big data mining techniques, based on the limited customer historical credit information. The model aims to provide an effective credit assessment to help financial institutions determine whether to provide loans to customers.

Firstly, this paper discusses the research background, content, and methodology, and conducts a study on relevant literature to identify the increasing personal credit risks faced by third-party financial institutions in China. Therefore, it proposes the establishment of a personal credit scoring card model based on limited credit data, providing reference for third-party financial institutions to mitigate personal credit risks. Secondly, this paper provides a brief introduction to the Chinese personal credit scoring system and the theoretical models required for the study. It analyzes the correlation between attributes and the target variable using two indicators, WoE (Weight of Evidence) and IV (Information Value). Preprocessing techniques such as random forest, defined functions, and data binning are used to reduce the risk of model overfitting and improve model stability. The paper then builds a modeling framework based on logistic regression model using data mining techniques, evaluates the model training results using the AUC curve, and applies the trained model to credit scoring of the test dataset. The conclusion is drawn that the model has good evaluation performance, reflecting the customer's credit situation, and can assist third-party financial institutions in making loan decisions for new users and mitigating credit risks.

2. Research Status and Indicator Selection

2.1. Research Status

This paper points out that personal credit assessment methods can be divided into statistical methods, operations research methods, non-parametric methods, artificial intelligence methods, and composite scoring models. Previous studies have shown that artificial intelligence methods have relatively high model accuracy but low stability. To adapt to the rapid development of credit business in China, statistical methods with high stability have become the preferred choice for financial institutions. Therefore, this paper chooses to develop a personal credit scoring model using statistical methods.

In the field of statistical methods, discriminant analysis and logistic regression analysis are widely applied. Previous research has shown that logistic regression analysis performs better in terms of predictive power and stability [12], and can provide the probability of customer default [3], thereby mitigating default risk. Therefore, logistic regression analysis has become one of the preferred statistical methods for financial institutions, outperforming discriminant analysis in terms of classification effectiveness [22]. Based on these reasons, this paper selects logistic regression analysis as the basic model for modeling.

In addition, due to the relatively late development of credit consumer loan business in China, the personal credit scoring system is not yet perfect. Therefore, the research direction of Chinese personal credit scoring indicators mainly focuses on selecting and improving international cutting-edge indicators (represented by FICO scores) [11]. China's personal credit scoring system mainly relies on data provided by commercial banks and credit reporting agencies [11], but this leads to the problem of incomplete credit data for third-party financial technology institutions, which affects the accuracy of credit assessment. Currently, there is a lack of simple models to assess personal credit using existing data.

2.2. Selection of Personal Credit Scoring Indicators

Due to the late start of credit scoring research in China, a mature and unified credit scoring system has not yet been developed. Currently, Chinese commercial banks and financial institutions still rely on the relatively mature scoring system in the United States to develop

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0035

their own scoring systems. Moreover, the scoring indicators vary among different financial institutions, which is determined by their diverse business needs. For these reasons, this article refers to the indicators in the American FICO scoring system to collect data. Overview of Personal Credit Score Card Model

3. Data Overview of Personal Credit Score Card Model

This article selects the data from the Give Me Some Credit project in the Kaggle data mining competition in the United States. The data can be downloaded from the website www.kaggle.com. The sample dataset used in this article consists of 150,000 records, including 11 variables (attributes), as shown in the following table:

Table 1. All Variables and Variable Descriptions

Variable name	Variable description
SeriousDlqin2yrs(independent variable)	Over 90 days overdue.
RevolvingUtilizationOfUnsecuredLines	The total balance of credit cards and personal credit limits divided by the total credit limit, excluding real estate and non-installment debt such as car loans.
age	Borrower's age.
NumberOfTime30-59DaysPastDueNotWorse	The number of overdue instances where the borrower is in arrears for 30 to 59 days after the due date.
DebtRatio	Monthly debt, alimony, and living expenses divided by total monthly income.
MonthlyIncome	MonthlyIncome
NumberOfOpenCreditLinesAndLoans	The number of loans, such as installment payments, car loans, mortgage loans, and credit loans (such as credit cards).
NumberOfTimes90DaysLafe	The number of instances where the borrower is overdue for 90 days or more.
NumberRealEstateLoansOrLines	The number of mortgage loans and real estate loans, including home equity lines of credit.
NumberOfTime60-89DaysPastDueNotWorse	The number of overdue instances where the borrower is in arrears for 60 to 89 days after the due date.
NumberOfDependents	The number of dependents in the household excluding oneself (spouse, children, etc.).

Data Source: Compiled based on the dataset downloaded from the Kaggle website.

3.1. Data Preparation

This article's data is raw data, so it needs to be cleaned. Data cleaning generally involves handling duplicate values, missing values, and outliers.

By examining the overview of the raw data, it is found that the variable "MonthlyIncome" has some missing data. Typically, there are three ways to handle missing data: directly deleting the missing values, manually filling in specified missing values, or filling in missing data with the mean, mode, or predicted values. In this article, the random forest algorithm is first used to predict the missing data, and then the predicted values are filled in the missing values.

Outliers refer to data points with extremely large or small values for a particular feature. A box plot is used to divide the data for a specific feature into quartiles (0.25, 0.5, 0.75) based on the minimum and maximum thresholds, and the upper and lower edges of the box plot are determined. Data points outside the upper and lower edges of the box plot are considered outliers and need to be handled.

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0035

Observing outliers sometimes requires considering the practical significance of the feature. For example, in the dataset used in this article, there is a feature called "age" (customer age). Based on common knowledge of human lifespan, customers who are eligible for loans are typically between 18 and 100 years old. Therefore, data points below 18 and above 100 need to be deleted in this dataset.

3.2. Data Feature Processing

Data binning is the process of dividing the values of a collected feature into multiple bins and replacing the data within each bin with a unified value. In this article, binning is performed by specifying the number of bins and the interval. After observing the data, the first, second, fourth, and fifth columns are binned into 10 segments, while the third, sixth, seventh, eighth, ninth, and tenth columns are binned using intervals.

After binning the sample dataset, it is necessary to select and exclude irrelevant or redundant features. This article uses Weight of Evidence (WoE) and Information Value (IV) to assess the impact of a feature on the target variable and determine the selection of features.

The formulas for calculating WoE and IV are as follows:

$$\text{score} = \sum_{i=1}^{n} \left(\left(WoE_i * \beta_i + \frac{a}{n}\right) * factor + \frac{offset}{n}\right)$$

$$\text{WoEi} = \ln\left(\frac{Bad_i}{Bad_T} / \frac{Good_i}{Good_T}\right) = \ln\left(\frac{Bad_i}{Bad_T}\right) - \ln\left(\frac{Good_i}{Good_T}\right)$$

$$\text{IVi} = \left(\frac{Bad_i}{Bad_T} - \frac{Good_i}{Good_T}\right) * \text{WoEi}$$

$$\text{IV} = \sum_{i=1}^{n} IV_i$$

In the expression, Badi represents the number of bad samples in a single interval after feature data binning, BadT represents the total number of bad samples for that feature; Goodi represents the number of good samples in a single interval after feature data binning, GoodT represents the total number of good samples for that feature.

Based on the above expressions, this article calculates the IV values for 10 features. According to the strength of the correlation between IV and the target variable, the impact of these features is "moderate" or above. Therefore, when training and fitting the model, only the following 5 features are selected: Revolving Utilization Of Unsecured Lines, age, Number Of Times 30-59Days Past Due Not Worse,Number Of Times 90Days Late,Number Of Times 60-89Days Past Due Not Worse.

4. The Establishment of Logistic Regression Model

4.1. Transformation of Raw Data

Data Feature Processing Before establishing the logistic regression model, this article transforms the original data into corresponding WoE values to improve the effectiveness of model training and facilitate the subsequent calculation of credit scores.

4.2. Transformation of Raw Data

This article utilizes the logistic regression model for data fitting and programming in Python. The logistic regression model is used to predict the probability of a binary outcome, such as the probability of a normal customer or a default customer. The predicted results have only two categories, namely 'normal' or 'default'. For ease of computer programming, the label 'normal' is represented as 1 and 'default' as 0. The dependent variable of the logistic regression model is the natural logarithm of the odds ratio between the probability of a normal customer and the probability of a default customer.

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0035

The fitted model is as follows:

0.543796*RevolvingUtilizationOfUnsecuredLines+0.384398*age+0.990557*NumberOfTime30-59DaysPastDueNotWorse+1.713207*NumberOfTimes90DaysLate+1.306427*NumberOfTime60-89DaysPastDueNotWorse + 10.466645.

4.3. Evaluation of Model

Model evaluation refers to the assessment of the model's performance. There are various evaluation criteria for classifiers, such as confusion matrix, KS curve, ROC statistic, and AUC curve. In this article, the AUC curve is used as the evaluation criterion for the established model.

Geometrically, AUC is defined as the area under the ROC curve enclosed by the coordinate axes. The ROC curve is plotted based on a series of different binary classification methods, with true positive rate as the vertical axis and false positive rate as the horizontal axis, corresponding to true negative rate and false negative rate. The AUC curve plot for the fitted model in this article is shown below:

The AUC value of the model established in this article, as calculated by the computer, is 0.84. According to the aforementioned AUC evaluation criteria, the training effect of the algorithm is considered good.

Before establishing the model in this article, the raw data was divided into training set data and test set data (with a ratio of 70% for the training set data and 30% for the test set data). In the previous chapters of this article, the model was successfully built using the training set data. Now, the test set data is being examined. The distribution of credit scores for normal/default customers in the test set is calculated as follows:

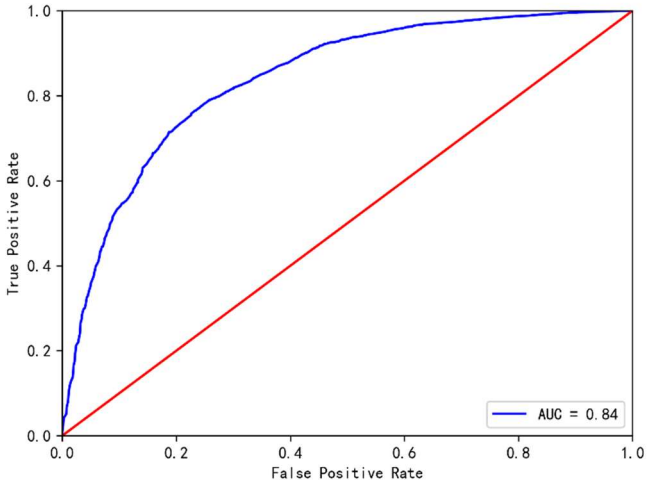

Figure 1. Area Under Curve of The Model

The AUC value of the model established in this article, as calculated by the computer, is 0.84. According to the aforementioned AUC evaluation criteria, the training effect of the algorithm is considered good.

4.4. Test of Model

Before establishing the model in this article, the raw data was divided into training set data and test set data (with a ratio of 70% for the training set data and 30% for the test set data). In the previous chapters of this article, the model was successfully built using the training set data. Now, the test set data is being examined. The distribution of credit scores for normal/default customers in the test set is calculated as follows:

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0035

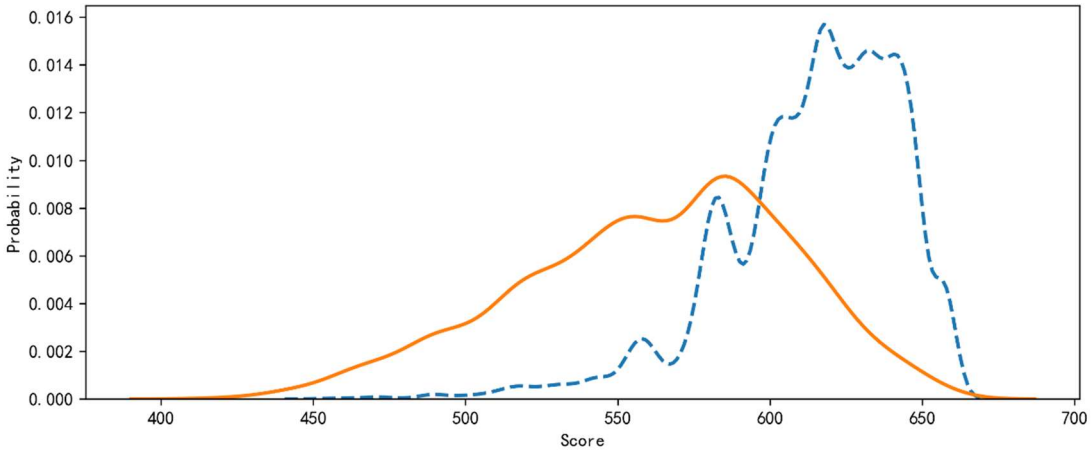

Figure 2. Description and Distribution of Credit Scores for Normal/Default Customers in the Test Set

In Figure 2, the blue dashed line represents the score distribution of normal customers, while the orange solid line represents the score distribution of default customers. The x-axis represents customer scores, and the y-axis represents the probability of normal/default customers. Based on Figure 2, it is evident that the average credit score of normal customers is higher. The distribution graph of scores for normal/default customers in the test set shows the distribution of credit scores for both types of customers. The normal customers are predominantly distributed in the higher score range, indicating that the model can reflect the true credit situation of customers.

5. Credit Scoring Card Development-Resolving Personal Risk

5.1. Credit Score Card based on Model Results

In the process of customer credit scoring, a personal credit scorecard is essentially a table composed of credit scores corresponding to different values of descriptive variables for the borrower. The credit scorecard can calculate credit scores for different borrowers. In this article, the basic score of the 600th percentile of the personal credit scorecard is taken, and the score corresponding to each feature variable can be calculated using the following formula: Weight of Evidence (WoE) multiplied by the fitting coefficient of the feature variable, added to the fitting intercept, multiplied by the scaling factor, and finally added with the offset. The mathematical formula is as follows:

$$\left(WoE_i * \beta_i + \frac{a}{n}\right) * factor + \frac{offset}{n}$$

Among them, WoE_i represents the WoE value of each variable; β_i represents the fitting coefficient; "a/n" represents the intercept; "scaling factor" represents the scaling factor; "offset/n" represents the offset.

As can be inferred from the above formula, the score of the scoring card can be calculated using the following formula:

$$\text{score} = \sum_{i=1}^{n} \left(\left(WoE_i * \beta_i + \frac{a}{n}\right) * factor + \frac{offset}{n}\right)$$

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0035

The calculation method for the proportion factor (factor) and offset/ n in the scorecard formula is as follows: Assuming a normal/default client ratio of 50/1 corresponds to a score of 600, adding 20 to the base score doubles the normal/default client ratio. Therefore, the formula is as follows:

$$600 = \log(50)*factor + offset$$
$$620 = \log(100)*factor + offset$$
$$factor = 20/\log(2)$$
$$offset = 600-factor*\log(50)$$

5.2. The Pros and Cons of Personal of Personal Credit Score Cards

The personal credit scorecard model has strong interpretability and simple calculation. The logistic regression model has better robustness. Second, the modeling uses real information data. Although the data in this article is from the Kaggle website, it is all real collected data, ensuring the accuracy of the model. Third, the personal credit scorecard is convenient to apply. This article aims to provide credit scoring tools for third-party financial institutions, so that they can efficiently obtain the credit score of customers in need of loans.

The design of the credit scorecard is relatively simple and not universally applicable. The personal credit scorecard developed in this article is only applicable to third-party financial institutions and is not suitable for major commercial banks. Second, the original data used for modeling is not balanced data. Third, the original data used to establish the credit scorecard model is based on American data. This article did not use Chinese credit data for modeling because credit data involves customer privacy and China has strict regulations on personal privacy data. Instead, this article used credit data from the Kaggle website.

6. Conclusion

The quantitative analysis of personal credit scoring, as a product of financial technology, has developed rapidly. Therefore, how to effectively manage credit risk is an urgent issue for financial institutions to address. This article first selects the logistic regression analysis method by analyzing the development and research status of credit scoring. Then, data preprocessing is conducted, followed by model establishment and evaluation. Finally, the evaluation and validation results show that this scorecard model has good accuracy and stability. Therefore, third-party financial institutions can use this credit scorecard as a reference to make decisions on whether to grant loans to customers, thereby mitigating personal credit risk.

Lastly, the recent development of artificial intelligence has provided more possibilities for financial technology. In credit model research, the results show that artificial intelligence has higher accuracy than statistical methods, but there are still issues such as poor interpretability. Therefore, in future research, it is worth exploring the combination of models to see if better methods can be improved. Moreover, credit scoring does not have to be limited to mitigating credit risk but can also be applied to activities related to loans, such as evaluating account profitability, assisting in financial fraud detection, and determining subsequent credit limits for accounts. The application of credit scoring in the financial industry can improve the quality and efficiency of credit services, leading to significant profit increase in the industry.

References

[1] Bai, J. (2012). Application of Logistic Regression-based Neural Network Model in Personal Credit Assessment. Inner Mongolia: Inner Mongolia University, 2012:1-34.

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0035

[2] Jiang, M., Xu, P., Ren, X., & Che, K. (2015). Development of Personal Credit Scoring Models And Analysis of Optimization Algorithms. Journal of Harbin Institute of Technology, 47(5), 40-45.

[3] Xu, P. (2019). Optimization Research on Personal Credit Scoring System in Commercial Banks. Harbin: Harbin Institute of Technology, 2017:1-158.

[4] Yu, X., Li, Z., & Duan, M. (2019). Comparative Study on Personal Credit Scoring Systems and Their Contemporary Value. Journal of Jiangxi Normal University (Philosophy and Social Sciences Edition), 52(4), 138-144.

[5] Yan, Y., & Jiang, H. (2010). Comparative Study on The Application of Credit Scoring Models. Statistics and Information Forum, (5), 30-35.

[6] Srinivasan V, Kim Y H. (1987) Credit Granting: A Comparative Analysis pf Classification Procedures. Journal of Finance, 42(3): 665-683.