ISSN 2616-5902 Vol. 4, Issue 10: 38-42, DOI: 10.25236/AJBM.2022.041007

Research on the Construction of Personal Credit Score Model Based on WOE Analysis and Logistics Model

Shuqi Liang[1,*], Li Tan[2]

[1]School of Information Engineering, Zhujiang College, South China Agricultural University, Guangzhou, 510900, China
[2]School of Information Management, Xinjiang University of Finance and Economics, Urumchi, 830000, China
**Corresponding author: 1498421482@qq.com*

Abstract: *As the concept of credit consumption enters people's lives, personal credit loans are gradually becoming a consumer demand, which makes the problem of credit risk in the bank lending business even more serious. To better protect the interests of financial institutions, investors and consumers, and to make the financial world more balanced and secure, scoring and modelling personal credit can reduce the likelihood of credit risk by identifying and quantifying risks in advance, reducing losses and making reasonable and effective loan plans. This has important implications for financial institutions, investors and consumers alike, and plays a very important role in economic development. In this paper, we will study the model construction of individual credit scores by obtaining data from customers' basic attributes, repayment ability, credit transactions, property status, loan attributes, other factors and time windows, processing and analysing them, and using WOE analysis to determine whether the indicators are economically meaningful, and correlation analysis to check the relevance of variables and IV screening variables. The logistic regression model was then converted to a standard scorecard format through WOE transformation and the model was then tested to obtain the scoring criteria.*

Keywords: *Credit risk; credit score; WOE analysis method; logistic regression model*

1. Introduction

The rapid growth of the national economy, the continuous improvement of people's quality of life, and the rapid development of the Internet have also brought people a different way of life[1]. Credit loan prepayment consumption has become the mainstream of most people's living consumption: buying a house loan, buying a car loan, and even purchasing instalments and other consumption methods can increase the credit risk brought by credit consumption [2] .

Personal credit score is a method commonly used in the world to assess personal credit risk, and it is also an important means to manage credit risk. By judging the customer's repayment ability, profitability, development ability and other aspects, assess the customer's credit risk level, and then formulate corresponding plans and take measures according to the customer's situation, so as to ensure the security of credit.

Between Bensic [3] the logistic approach is considered to be the relatively most accurate of the traditional models and should be promoted in the credit scoring industry. Chuang and Huang [4] compared the neural network approach with other methods on the credit scoring neighbourhood and the study empirically showed that the neural network model has higher accuracy when the variables present complex non-linear relationships. The lack of personal credit scoring technology has become a bottleneck for the development of consumer credit business of commercial banks in my country. Compared with the mature and complete FICO scoring system in foreign countries, my country's online loan platform currently does not have a unified, standard and effective credit evaluation mechanism. In this paper, the collected data will be sorted and analyzed, and the variables will be screened after determining the variables and their correlations. Then, the logistic regression model [5-7] can be transformed into a standard scorecard format by using WOE transformation, and then the model can be tested and the scoring standard can significantly increase the total value of the loan platform loan by improving the accuracy of default prediction.

ISSN 2616-5902 Vol. 4, Issue 10: 38-42, DOI: 10.25236/AJBM.2022.041007

2. Data processing

2.1. Data acquisition

As the data belongs to personal consumer loans, relevant data is obtained from credit scores to build a credit scoring system. Data is obtained from the following areas: basic attributes, solvency, credit transactions, property status, time windows and other factors.

2.2. Data preprocessing

Before working with data, it is necessary to understand the missing values and outliers of the data. Python's Data Loading, Data Accuracy and Distribution gives an idea of the missing values, mean and median of the data. From the processed data, it can be seen that the monthly income and the number of dependents of the borrower are missing. There are 29,731 missing values for monthly income and 3,924 missing values for the number of dependents. As the missing values would make it impossible to apply some of the mathematical modelling and analysis methods, these missing values need to be dealt with in the first step of developing the credit risk rating model. The treatment includes removing the missing values and filling in the missing values. The missing rate for the lender's monthly income is relatively large, so filling in the missing values is done for correlation between the variables. The random forest method is used here to fill in the missing values.

The missing values for the number of family members of the lender are relatively small and the missing values are removed directly, which has little impact on the overall model. To deal with missing values, outliers also need to be dealt with. An outlier is a value that deviates significantly from the majority of the sampled data. For example, when the age of a single customer is zero, that value is usually considered an outlier. Finding outliers in the sample population is often done using outlier detection. Also, a good customer in the dataset is 0 and a default customer is 1. Consider the normal understanding that a customer who can perform the contract and pay interest properly is 1, so negate.

2.3. Exploratory data analysis

Exploratory data analysis was carried out on the existing data prior to modelling, with as few a priori assumptions as possible. Histograms were used to show the distribution of customer age, customer income and monthly income, and the distribution was found to be approximately normal, as required for statistical analysis. However, the histogram presented a distribution of situations concentrated in one column and the distribution was not clear. Next, outliers were found using the boxplot, and then the outliers needed to be split up to eliminate them. There are 301 outliers in the monthly income, which is negligible in a data volume of 120,000.

2.4. Variable selection

This paper adopts the variable selection method of the credit scoring model, and uses the WOE analysis method to determine whether the indicator meets the economic significance by comparing the index binning and the default probability of the corresponding binning.

2.4.1. Binning processing

Optimal partitioning of continuous variables is chosen first, and isometric partitioning of continuous variables is then considered when the distribution of continuous variables does not meet the requirements for optimal partitioning. Optimal partitioning will be used for the recycling of unsecured amounts, age, debt ratio and monthly income in the dataset. For variables that cannot be optimally partitioned, continuous discretization is used.

2.4.2. Correlation analysis and IV screening

Next, this paper uses the cleaned data to look at the correlations between variables. Note that the correlation analysis here is only a preliminary check, and the VI (weight of evidence) of the model is further checked as the basis for variable screening.

Infomation Value(IV) of each variable is further calculated . IV metrics are generally used to determine the predictive power of independent variables. Its formula is:

ISSN 2616-5902 Vol. 4, Issue 10: 38-42, DOI: 10.25236/AJBM.2022.041007

$$IV = Sum\left(\left(Good\ \ Attribute - Bad\ \ Attribute\right) \times ln\left(\frac{Good\ \ Attribute}{Bad\ \ Attribute}\right)\right) \tag{1}$$

by the IV value are:< 0.02:unpredictive ;0.02 ~ 0.1:weak;0.1 ~ 0.3 :medium ;0.3 ~ 0.5 :strong;>0.5:suspicious.

3. Model building and analysis

3.1. Model analysis

WOE (Weight of Evidence): transformation can transform the Logistic regression model into a standard scorecard format [8] .

WOE analysis is to bin the indicators, calculate the WOE value of each gear, and observe the trend of the WOE value changing with the indicators. The mathematical definition of WOE is:

$$\mathrm{W}OE = ln\left(\frac{Good\ \ Attribute}{Bad\ \ Attribute}\right) \tag{2}$$

In the analysis, each indicator is sorted from smallest to largest and the WOE value of the corresponding position is calculated.The larger the positive index, the smaller the WOE value; the larger the negative index, the greater the WOE value. The greater the negative slope of the WOE value of the positive index and the greater the positive slope of the response index, it means that the index has a good ability to distinguish. The WOE value is close to a straight line, which means that the indicator's ability to judge is weak. If there is a positive correlation trend between the positive indicator and WOE, and the negative correlation trend between the negative indicator and WOE, it means that this indicator does not meet the economic significance and should be removed.

Introducing WOE transformation is not to improve the quality of the model, but some variables should not be included in the model, either because they cannot increase the value of the model, or because the error related to the correlation coefficient of its model is large, in fact, establish a standard credit scorecard It is also possible not to use WOE conversion.In this case, the logistic regression model needs to deal with a larger number of independent variables. Although this adds complexity to the modelling process, the resulting scorecards are all the same. Prior to modelling, the filtered variables were converted into WOE values for credit scoring purposes.

3.2. Model establishment and results

Since the weight of evidence transformation can transform the logistic regression model into a standard scorecard format, after replacing the variable data with WOE, directly call stats models to implement logistic regression. The output shows that all the variables in the logistic regression have passed the significance test and meet the requirements.

3.3. Model checking

The next test was validated using the test data set aside at the beginning of the modelling to assess the fit of the model through the ROC curve and the AUC. The AUC value is 0.85, indicating that the prediction effect of the model is good and the accuracy rate is high. an AUC value of 0.85 indicates a good prediction and high accuracy of the model.

3.4. Establish a scoring system

The ROC curve was used to verify the predictive power of the model and then the logic model was converted into a standard scorecard format.

3.4.1. Scoring criteria

The scoring formula is:

$$a = log(\frac{P\_good}{P\_bad}) \tag{3}$$

ISSN 2616-5902 Vol. 4, Issue 10: 38-42, DOI: 10.25236/AJBM.2022.041007

$$Score = offset + factor \times log(odds) \tag{4}$$

Before building a standard scorecard, choose a few scorecard parameters: base score, PDO (double the score to ratio) and good to bad ratio. Here, a score of 600 is used as the base score, the PDO is 20 (for every 20 points higher, the good to bad ratio is doubled) and the good to bad ratio is taken as 20.

The scoring standard is: personal total score = basic score + each part score.

3.4.2. Partial scoring

Calculate the score of each variable:

Table 1: The score of each variable

Unsecured Lines	score	age	score	30-59Days PastDue NotWorse	score	90 Days Late	score	60-89Days PastDue NotWorse	score
<=0.0312]	24	<=33	-8	<=0.0	16	<=0.0	20	<=0.0	9
(0.0312,0.158]	22	(33,40]	-6	(0.0,1.0]	-27	(0.0,1.0]	-102	(0.0,1.0]	-60
(0.158,0.558]	5	(40,45]	-4	(1.0,3.0]	-52	(1.0,3.0]	-142	(1.0,3.0]	-88
>0.558	-20	(45,49]	-3	(3.0,5.0]	-70	(3.0,5.0]	-166	>3.0	-95
		(49,54]	-1	>5.0	-80	>5.0	-160		
		(54,59]	3						
		(59,64]	7						
		(64,71]	14						
		(71,103]	16						
Note: The base score is 795.									

3.4.3. Automatic scoring system

The score is calculated based on the variables, and the result is:

Table 2: Automatic scoring system calculation results

Good Customers and Bad Customers	Revolving Utilization of Unsrcured Lines	Age	30-59 Days PastDue NotWorse	90Days Late	60-89 Days PastDue NotWorse	score
1	-20	-4	-70	-142	-60	499
1	22	3	-27	-102	-60	631
1	5	-3	-27	-102	-60	608
1	24	14	-27	-102	-60	644
1	24	-6	-27	-102	-60	624
1	22	16	-27	-102	-60	644
1	-20	-3	-27	-102	-60	583
1	5	14	-27	-102	-60	625
1	22	-1	-27	-102	-60	627
1	5	-6	-27	-102	-60	605
1	-20	7	-27	-102	-60	593
1	22	3	-27	-102	-60	631
0	-20	-8	-27	-142	-60	538
1	22	16	-27	-102	-60	644
1	24	-8	-27	-102	-60	622
1	-20	-4	-70	-102	-60	539
1	24	14	-27	-102	-60	644
1	22	-4	-27	-102	-60	624
1	24	-3	-27	-102	-60	627
1	22	-6	-27	-102	-60	622
1	24	-1	-27	-102	-60	629
1	22	14	-27	-102	-60	642

ISSN 2616-5902 Vol. 4, Issue 10: 38-42, DOI: 10.25236/AJBM.2022.041007

4. Result

In an era of prosperous countries and happy people's lives, the rapid development of information has made borrowing a daily routine for most people. An accurate, efficient and stable personal credit scoring system is not only required by society and economy, but also closely related to everyone. On the basis of existing research, this paper analyzes and processes the data of existing customers and potential customers, and establishes an automatic credit scoring system. The innovations are reflected in:

Use univariate feature selection methods and methods based on machine learning models to select variables. After confirming variables, transform WOE of variables and estimate logistic regression. Model development, logistic regression coefficients and WOE, etc. to determine credit scores, and logistic model Convert to standard scoring to build an automatic credit scoring system. The error value is small and the accuracy rate is high, which meets the needs of the lending industry chain and reduces the risk of the financial industry.

In practical applications, the impact of emergencies on the personal credit score model should be further considered, and the corresponding risks should be analyzed and judged, and the model should be adjusted in time, that is, the personal loan score adjustment caused by unexpected factors should be analyzed [9].

References

[1] Liu Jing. Personal Credit Risk Rating Based on Cost-Sensitive Bayesian Classification [D]. Guangzhou: South China University of Technology. 2015

[2] Shi Qingyan, Jin Yunhui. A review of the main models and methods of personal credit scoring [J]. Statistical Research, 2003, 8(4): 36-39

[3] Mirta Bensic, Natasa Sarlija, Marijana Zekic-Susac. Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. [J]. Int. Syst. in Accounting, Finance and Management, 2005, 13(3):

[4] Chuang Chun-Ling Huang Szu-Teng. A hybrid neural network approach for credit scoring [J]. Expert Systems, 2011, 28(2):

[5] Yu Wenjian, Shen Yichang . Research on Personal Credit Score Based on Logistic Model [J]. Hainan Finance, 2007, 3(6): 83-85

[6] Wang Li. Research on Credit Score of Small and Medium Enterprises Based on Logistic Regression Model [D]. Hefei: Hefei University of Technology, 2008

[7] Su Cheng . Research on Credit Risk Assessment Based on Logistic Regression Model [J]. Review of Applied Economics, 2011, 6(13): 215-220

[8] Chen Chunzhao, Xie Rui, Zha Jingyi, Zhu Jiaming. Research on credit risk assessment and credit strategy of small, medium and micro enterprises based on big data [J]. Journal of Natural Science of Harbin Normal University, 2021, 3 7 (4); 2 9-30

[9] Shangzhou Xia. Value Optimization Method of Online Lending Platform Based on WOE - Logistic Credit Scorecard Default Prediction Model [D]. Wuhan: Central China Normal University. 2019. 17-18