Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/62/20231347

Credit Card Logistic Modeling Principles, Methods and Wind Control Strategy Building

Yiming He[1,a,*]

[1]*Department of Finance, Northeastern University at Qinghuangdao University, Taishan Road No. 143 Haigang District, Qinghuangdao City, Hebei Province, China*
a. yiminghe@ldy.edu.rs
**corresponding author*

Abstract: Now there are many methods of credit card modeling, of which Logistic regression is the most commonly used. Logistic regression has been modified since its introduction. The core of logical regression is theoretically supported by linear regression, with the introduction of nonlinear factors to deal with the secondary classification problem through the sigmoid function. This paper aims to explore the principles and processes of logic regression specifically, and to use logical returns for rating card modeling. It is therefore recommended that banks use credit cards for effective risk management. Logical regressions are more influenced by the data, so in the data processing link, we use sample partitions and distribute conversions to WOE values. At the same time, use examples to validate conclusions and to present the advantages and disadvantages of Logical Regression. This article also proposes the need for back-end monitoring of scorecard models and requires artificial judgment of whether the data type of the client reveals valid characteristics.

Keywords: logistic regression, rating card, credit score, risk management

1. Introduction

The credit card model was first introduced by the US credit rating giant FICO in the 1960s and is widely used in the fields of credit risk assessment and financial risk control. The rating card model gives credit scores of different grades, thereby determining the quality of the customer [1].

The credit card modeling methodology has been very well constructed. Orgler introduced linear regression analysis into a credit card model for personal consumer loans [2]. Winginton used the logistic regression method for the first time in traditional credit scorecards to represent the training sample requirements in two categories of good and bad people and presented a credit rating model [3]. Jolly Cloud used the logistic regression model to predict the probability of default of borrowers, using AUC as an evaluation indicator, and the model worked well [4]. Since then, the credit rating model has been stagnant for years, mostly because of the improvement of the rules of judgment, and Hand and others believe that development should be focused on developing more complex models [5]. Khandani and others used machine learning technology to predict nonlinear parameters of consumer credit risk models and found that machine learning methods were more interpretative than linear regression models [6]. Using big data analysis methods and data mining techniques, Huang Qing and Wang Lin designed and constructed a risk measurement model for retail customers of commercial banks' e-commerce platforms and established a risk rating model [7]. In the process of

DOI: 10.54254/2754-1169/62/20231347

introducing the decision-tree algorithm in machine learning into the character subbox, the model prediction effect is improved [8]. Yuan Jiang and others have proposed the concept of a multi-source data universality model, using the XGBoost algorithm to generate a sub-rating model, then convert the sub-grading model to a scorecard, and verify its validity by means of empirical analysis [9].

Wang Xian and Zhang Yuan believe that building a risk strategy requires the modeling of big data using machine learning techniques [10]. Xi Jinping and others believe that it is necessary to use artificial intelligence technology to analyze and build risk strategies in customer credit assessment, borrowing risk assessment, and fraud detection [11]. This article describes the logistic regression method.

2. Background technology

In general, the logical regression model is the most common type of credit rating model, used in the pre-loan approval phase to judge the lender's level of risk. The theoretical basis of the model is relatively well-defined and very interpretative, and it is the main method for the development of commercial bank credit scorecards. The logical regression factor variable is whether the borrower fails or not and is generally referred to as a "good sample" or "bad sample", expressed by 0 or 1. The bank then builds up a model based on its own information to calculate the probability of the borrower's default [12].

The first step in logistic regression is the processing of data, i.e., variable filtering, and the calculation of IV values. The full name of IV is information value, which means information value, which can be used to measure the predictive ability of self-variables. The calculation of the IV value requires the use of the WOE value. WOE's full name is Weight of Evidence. WOE is a form of coding for the original self-variable. For example, the original value of the answer to a judgment question is right and wrong, and the WOE report corresponds to different WOE values. After the conversion is completed, for and "wrong" values are replaced by the respective WOE values. The equation (1) to (3) for WOE and IV are:

$$WOE_i = \ln\left(\frac{y_i}{y_T}\right) \tag{1}$$

$$IV_i = \left(\frac{y_i}{y_T} - \frac{n_i}{n_t}\right) \times \left(\frac{\frac{y_i}{y_T}}{\frac{n_i}{n_t}}\right) \tag{2}$$

$$IV = \sum_i^n IV_i \tag{3}$$

y_i is the number of bad samples in this group, n_i is the amount of good ones in that group, y_T is the total number of goods in the sample, and n_t is the sum of all bad ones. Multiple linearity tests will also be required. Multiple linearity means that the model estimates are distorted or difficult to estimate accurately due to the presence of precise or highly related relationships between interpretative variables in a linear regression model and therefore need to be excluded during data processing. We measure multi-linearity by using the multi-differential expansion factor, assuming that there are multiple self-variables, and one of them returns to the X linearity with the remaining X variable, and we get R_i^2, and then the multi-linearization factor (equation 4) for x_i is

$$VIF_i = \frac{1}{1-R_i^2} \tag{4}$$

The second step is to use the existing data matching model. The model in equation (5) is:

DOI: 10.54254/2754-1169/62/20231347

$$ln(\frac{PD}{1-PD}) = \beta_0 + \beta_1 WOE_1 + \cdots + \beta_m WOE_m \quad (5)$$

where PD is the default rate.

Step three gets a score value. The logical regression model used above predicts probability, i.e., default probability estimates, but in the end we need to get a very simple, intuitive score, so we rely on the rating calibration to standardize the predicted probability value to a fractional value and get a score value table (i.e., how many points each of the variables takes). Score calibration is the use of linear conversion to obtain specific scores. We define the equation (6):

$$Odds = \frac{p}{1-p} \quad (6)$$

where p is the probability of a good sample, and we call odds the good-bad ratio, that is, the ratio of good and bad samples. Score values can be obtained by linear mapping, defining equation (7):

$$Score = A + B \times ln\,(odds) \quad (7)$$

PDO refers to the score at which the odds become twice as high as they originally needed to be. Then, according to the definition of PDO equation (8):

$$\begin{cases} Score = A + B \times ln\,(odds) \\ Score + PDO = A + Bln(2odds) \end{cases} \quad (8)$$

Then the equation (9):

$$\begin{cases} B = \frac{PDO}{ln2} \\ A = Score - Bln(odds) \end{cases} \quad (9)$$

The output of A and B leads to the original formula, which gives a score card value with different probabilities.

The fourth step is the evaluation of the quality of the model. While we can't use the scorecard model directly, we also need to verify its quality. It's usually divided into technical and non-technical perspectives. The technical perspective is the AUC indicator and the observation of KS values. Non-technical perspectives are security, interpretability, and complexity.

1. Technical point of view

(1) Calculation of the AUC (Area Under ROC Curve) value

As shown in Figure 1,the calculation of the AUC value requires drawing a ROC curve. The ROC curve is a coordinate chart consisting of the false positive rate (FPR) and the true positive rate of the vertical coordinate (TPR), while the area underneath, which is surrounded by the coordinate axis (ROC), is called the AUC value. The larger the AUC value, the better the performance of the classifier (Figure 1).

DOI: 10.54254/2754-1169/62/20231347

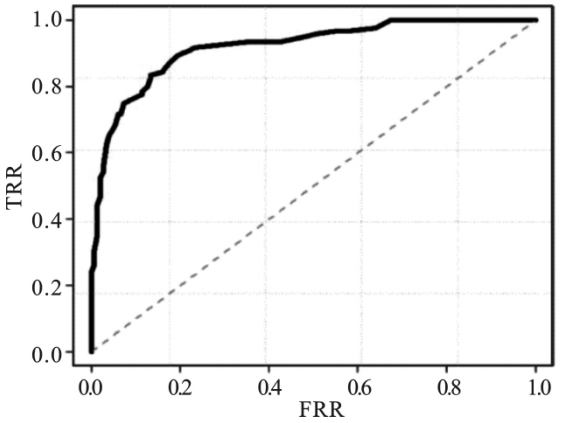

Figure 1: ROC curve map [12]

(2) Calculation of the KS (Kolmogorov-Smirnov) value

As shown in Figure 2, the KS curve is similar to the ROC curve. With different probabilities, draw the TPR and FPR curves; the distance between the two curves is the KS value, as shown in Table 1. The KS curve is used to measure the differential capacity of a model, allowing a small number of errors to be made to identify the bad sample as much as possible.

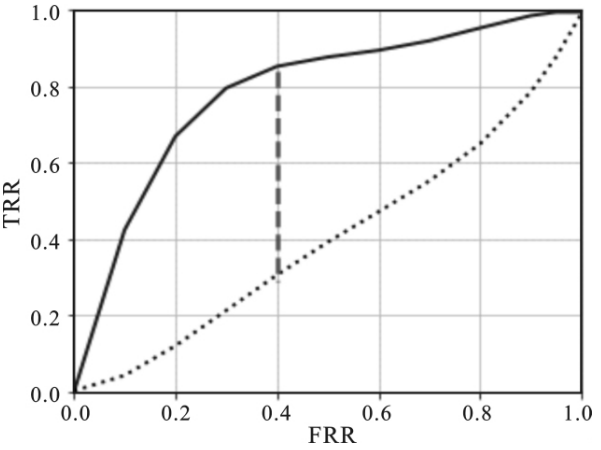

Figure 2: KS curve map [12]

Table 1: KS values [12]

KS Values	Results of evaluation
less than 0.2	Model Unidentifiable
0.2-0.4	Models are acceptable
0.41-0.5	The model has better differential capabilities
0.51-0.6	The model has good differentiation.
0.61-0.75	The model has very good differentiation
more than 0.75	Model abnormalities, possible problems

2. Non-technical perspectives

(1) Security modeling's ability to resist external attacks

(2) Explanation Modelling is easy to understand and variable-specific business implications

(3) Complexity Models are not easily too complex to reduce development and maintenance costs.

DOI: 10.54254/2754-1169/62/20231347

3. Substantive analysis

This article has developed a logical regression model using the data provided by a professor in the United Kingdom (the data is not disclosed), with the specific equation (10) as follows:

$$Y = 0.0210782x_1 + 0.2951075x_2 + 0.0540624x_3 + 0.2176936x_4 - 1.189183 \quad (10)$$

As shown in Figure 3, the AUC value is 0.7818, which indicates a good level of adaptation.

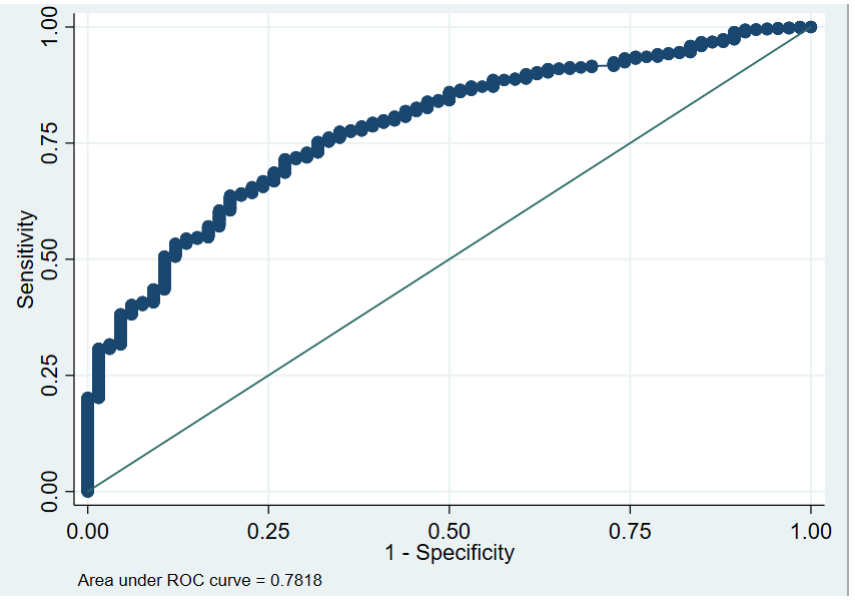

Figure 3: AUC values for matching models

4. Wind control strategy construction

Now the banks rely on the rating card for risk management. First, from the principle of logical regression, the choice of dataset is especially important. Wind control personnel should pay attention to the lender's daily behavior data and focus on extracting effective characteristics that are relevant to the economic business. The data dimension includes the basic information of the loan applicant, bank card information, whether there has been a default, etc. Secondly, the banks need to combine their own risk preferences and judge the business development characteristics of the enterprise, flexible use of credit card tools for credit approval, not relying on scores too much. Especially for borrowers with pre-failure section or a lower rating value, it is necessary to make a mixed judgement through expert judgment and combination of machine testing [13].

Thirdly, the rating card also needs backend monitoring, i.e., observing bad rates, to determine whether there are significant differences in risk performance within the range, which can be used by banks to optimize the credit card model to improve the model's predictability.

5. Analysis of problems

This paper describes only the forward-end process of credit card modeling, i.e., the approval of customer loan applications, but requires the study of life-cycle credit card models, including automatic post-credit renewal, late default smart recovery, etc. Banks can establish credit scorecards and creditors' behavioral scorecards.

At the same time, the whole process of model monitoring, that is, observing the changes in the key indicators of the model, to determine whether the current model performance is still effective, model

DOI: 10.54254/2754-1169/62/20231347

surveillance to provide the basis for decision-making for the model optimization, can be found through monitoring. If the trigger conditions for model optimization are met, it is necessary to implement the model optimization and timely update the scorecard model.

6. Future prospects

It is worth noting that logical regression is explainable but relies on the reliability of the dataset and is affected by abnormal values. It should be noted that the dataset used for the logistic regression to establish a scorecard model is a "small sample". The "small sample" is relative to the "large sample". The "large sample" does not have statistical characteristics in real life. For example, a crime survey shows that the crime rate calculated by taking 2,000 samples in one region is often higher than 200,000, which may be due to the fact that there are still many good people in the world. And the logical regression process is cumbersome, and the data processing is demanding, so machine language modeling, such as XGBoost and LightGBM, is proposed. The emergence of machine language greatly improves the efficiency of modeling and reduces complicated data processing processes. The latter part of this article focuses on learning machine language programming and combining it with logical regression to solve the "black box problem".

7. Conclusion

This paper demonstrates the validity of the logistic regression establishment of credit cards by specifically describing the process of building credit cards using the Logistic regressive model and by example verification. But there are still disadvantages: the modeling process is cumbersome, requires data pre-processing and partitioning operations, and data sensitivity is high, i.e., when an anomaly exists, the model cannot be identified and eliminated, but instead results in errors in the score. And a lot of information is inevitably lost after a series of data processing. Therefore, when using logical regression modeling, banks need to focus on monitoring the data source and extracting valid characteristics from a large amount of data.

References

[1] *Qian WeiWay, Risk Control: Credit Scoring Card Model, 2021.1.27, 2023.8.29, https://www.biaodianfu.com/credit-score.html*

[2] *Orgler Y E. A Credit Scuring Model for Commercial Loans [J]. Journal of Money Credit & Banking, 1970, 2(4)*

[3] *Li Yang, Keliang Wang, Jianmin Wang, Comparative Methodology of Major Models of Credit Rating, Economic Management, 2008.3.20*

[4] *Liyun Zhu, Commercial Banking Credit Risk Analysis Based on Logistic Model, Brand Study, 2019(19)*

[5] *SrinivasanV, Yong H K. The Bierman-hausman Credit Granting Model: A Note [J]. Management Science, 1987, 33(10)*

[6] *Khandani A E, Kim A J, Lo A W Consumer Credit-risk Models via Machine-learning Algorithms [J]. Journal of Banking & Finance, 2010, 34(11)*

[7] *Changjun Huang, Lin Wang, Big Data Age Commercial Bank Risk Score Model Design Framework for Retail E-Commerce Customers, Investment Analysis, 20144.4.10*

[8] *Yujuan Xi, Applications of Characteristics Division Algorithms Based on Decision Tree in Credit Scoring Models for Business Banks, Zhengzhou University, 2020.9.1*

[9] *Zhigang Huang, Zhihui Liu, Jianlin Zhu. Construction and Application of Universal Model Stack Box for Multi-Source Data Credit Rating [J].Quantitative Economic and Technological Economic Studies,2019,36(4):155-168*

[10] *Qian Wang, Jun Zhang, Financial Technology Empowers Business Health Risk Digital Wind Control Study, Financial Science Age,2023,31(07)*

[11] *Kunlong Xu, Jingchang Ye, Zhicheng Sun, Bing Zheng, Fa Si, Wenxiang Yu, Chao Tong, Intelligent Wind Control Decision-making System Based on Digital Wholesale Finance, Information Technology and Standardization,2023(08)*

[12] *Zhihui Liu, Zhigang Huang, Heliang Xie, Big Data Wind Control Is Effective? - based on comparison analysis of statistical scorecards and machine learning models, Statistics and Information Forum, 2019.9*

DOI: 10.54254/2754-1169/62/20231347

[13] Deng Datsun, Zhao Yulong, Commercial Bank of China Small and Micro Enterprises Application for Credit Card Construction and Verification Research, Investment Research, 2017.5.10