Research on Consumer Credit Rating Model

Xinyu Wu*, Jielin Shang

PLA Dalian Naval Academy, Dalian, China

*E-mail:772774131@qq.com

Abstract. With the rapid growth of online consumer credit products in commercial activities, banks and financial institutions widely employ credit scoring models to assess customers, implementing varying credit limits and policies for different customer tiers in order to mitigate the risk associated with individual consumer credit. This paper presents a credit scoring model based on logistic regression methodology, assigning distinct scores to individuals based on their unique information, facilitating investors in making corresponding decisions in commercial endeavors.

Keywords: Scoring Model, Logistic Regression, Credit Risk

1 Introduction

In the contemporary landscape of business operations, consumer scenarios have become increasingly diverse and technologically driven. As of June 2023, the total outstanding balance of personal general consumer loans in China (excluding individual mortgages) stands at 18.74 trillion yuan, reflecting a year-on-year growth of 12.4%. [1]Within this context, the rapid proliferation of online consumer credit products has introduced a new challenge—how to effectively manage the risks associated with these credit products. This predicament is a shared concern among all financial institutions. Consequently, the most crucial facet of credit development is the effective mitigation of risks. Both banks and financial institutions aspire to devise a method that can not only control the default rate of individual consumer credit but also ensure substantial returns. Hence, the emergence of credit scoring models: a methodology that assigns varying scores to individuals based on their distinct information, enabling the assessment of customer credit risk and, thereby, augmenting profitability.

2 Credit Scoring Concept and Models

2.1 Credit Scoring Concept

Credit scoring is a process that utilizes an applicant's credit history information to generate varying levels of credit scores through the application of a credit scoring model. These credit scores serve multiple purposes: they can predict whether a cus-

A. Bhunia et al. (eds.), *Proceedings of the 2023 International Conference on Finance, Trade and Business Management (FTBM 2023)*, Advances in Economics, Business and Management Research 264,
https://doi.org/10.2991/978-94-6463-298-9_17

tomer is likely to default, categorize customers into good and bad segments based on these predictions, and also influence decisions regarding customer credit limits and interest rates.

Credit scoring can be categorized into various types based on the source and characteristics of data, making it adaptable to diverse scenarios. The evaluation of credit scoring models is a pivotal step in management decision-making and serves as a reflection of a company's competitiveness and profitability within its industry. In 2007, Anderson suggested that credit scoring should be approached from two fundamental aspects: "credit" and "scoring." In essence, "credit" signifies the concept of "consumption first, repayment in the future," while "scoring" involves the utilization of mathematical tools to sequentially rank and classify applicants, ensuring consistency and objectivity.

A credit scoring card effectively segregates applicants into two distinct categories: those deemed as good, indicating their capacity for timely repayment, and those categorized as bad, signifying their inability to meet their financial obligations. Additionally, it can predict the likelihood of default for new credit applicants. This prediction relies on an analysis of an applicant's historical loan characteristics and performance metrics, such as the number of loan inquiries made within the past six months, the time elapsed since the issuance of their first credit card, the total outstanding amount, and the cumulative number of months in arrears within the past six months, as recorded in the People's Bank of China credit report. By analyzing these known data, distinct characteristics that differentiate defaulting customers from non-defaulting customers are identified. Subsequently, a model is constructed to predict the probability of default for new loan applicants, providing a basis for making lending decisions.

2.2 Methods of Credit Scoring Modeling

Credit scoring models can be broadly categorized into two main groups: traditional statistical methods and artificial intelligence (AI) methods. Traditional statistical methods include techniques like Logistic Regression and Probit Regression, while AI methods encompass Decision Trees, Backpropagation (BP), Support Vector Machines (SVM), and others.[2]

In recent years, AI methods have gained more prominence in credit scoring applications. However, it's important to note that these AI methods are not yet fully mature. For instance, BP is considered a black-box operation with limited interpretability, and training SVM models can be time-consuming and challenging to explain. Although Logistic Regression may not match the predictive capabilities of AI methods, it offers strong interpretability and shorter model training times. Given the relative immaturity of AI methods and the advantages of Logistic Regression, this paper adopts the Logistic Regression approach for its credit scoring card. Logistic Regression plays a central role in the development of credit scoring cards, and the following section primarily focuses on its use within the context of credit scoring.

Logistic regression is a type of generalized linear regression that divides variables into dependent (outcome) and independent (predictor) variables. The dependent variable is typically a binary categorical variable, often representing two classes: 0 for

normal customers and 1 for defaulting customers. The independent variables describe customer information. The essence of credit scoring lies in utilizing historical customer data to extract characteristics that differentiate good customers from bad ones and subsequently categorize new customers[3].

Let's begin by assuming there are a total of n observations for the dependent variable y, which contains two values: 0 for normal customers and 1 for defaulting customers. Furthermore, there are r independent variables, $x_1,...,x_r$. For the i-th observation, the values of y and x are represented, as illustrated in Table 1.

Table 1. Logistic regression data symbol

Observed value	Independent variable			Result
	x_1	...	x_r	y
1	x_{11}	...	x_{1r}	y_1
2	x_{21}	...	x_{2r}	y_2
...	...	...	...	...
n	x_{n1}	...	x_{nr}	y_n

In a logistic regression model, we calculate the probability of event ($y=1$) as follows:

$$\Pr\{y=1\}=\frac{\exp(\beta_0+\beta_1x_1+...+\beta_rx_r)}{1+\exp(\beta_0+\beta_1x_1+...+\beta_rx_r)} \tag{1}$$

constant $\beta_0,\beta_1,...,\beta_r$ represents the parameters of the model. constant β_0 represents intercept term.

We can express predictor variables x as a vector based on the parameters β, as shown below:

$$X^T=[1 \quad x_1 \quad ... \quad x_r] \tag{2}$$

$$\beta^T=[\beta_0 \quad \beta_1 \quad ... \quad \beta_r] \tag{3}$$

Therefore, we can simplify the Logistic regression in Equation 1 to the following form:

$$\Pr\{y=1\}=\frac{\exp(\beta^Tx)}{1+\exp(\beta^Tx)} \tag{4}$$

Another alternative form of Equation 1 is:

$$\Pr\{y=1\}=\frac{1}{1+\exp(-\beta^T x)} \tag{5}$$

To simplify the symbols in the above equations and express the probability of event ($y=1$) as p , we can rewrite Equation 5 as follows:

$$p=\frac{1}{1+e^{-z}} \tag{6}$$

$$z=\beta_0+\beta_1 x_1+\ldots+\beta_r x_r \tag{7}$$

We can transform Equation 6 to:

$$\ln\left(\frac{p}{1-p}\right)=z \tag{8}$$

Here, p represents the probability of event (customers default), and $1-p$ represents the probability of event (normal customers). The ratio $p/(1-p)$ represents the odds of two events.

Maximum likelihood estimation is the most classical method for building a logistic regression model. The posterior probability for a single sample is given by (Equation 9):

$$p(y|x,\beta)=(h_\beta(x))^r(1-h_\beta(x))^{1-r} \quad (\ y=1/\ y=0\ ) \tag{9}$$

The likelihood function is:

$$\begin{aligned} L(\beta|x,y)&=\prod_{i=1}^{m}\mathrm{p}(\mathrm{y}^{(\mathrm{i})}|\mathrm{x}^{(\mathrm{i})};\beta)\\ &=\prod_{i=1}^{m}(h_\beta(x))^{y^{(i)}}(1-h_\beta(x))^{1-y^{(i)}} \end{aligned} \tag{10}$$

The log-likelihood function is:

$$\begin{aligned} l(\beta)&=\log(L(\beta|x,y))\\ &=\sum_{i=1}^{m}y^{(i)}\log(h(x^{(i)}))+(1-y^{(i)})\log(1-h(x^{(i)}) \end{aligned} \tag{11}$$

Then, we iterate until convergence.

3 Model scoring

Let the estimated probability of default be represented as p , then the estimated probability of non-default is $1\text{-}p$.[4]These two events are mutually exclusive, meaning that their probabilities sum to 1. Therefore, we have:

$$Odds = \frac{p}{1-p} \tag{12}$$

The calculation formula for the probability of default (pp) is as follows

$$p = \frac{Odds}{1+Odds} \tag{13}$$

The scoring card sets the score scale by defining the score as a linear expression of the logarithm of the odds ratio. The result is as follows:

$$Score = A - B\log(Odds) \tag{14}$$

The calculation of the odds ratio in logistic regression is as follows:

$$\log(Odds) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{15}$$

constant $\beta_0, \beta_1, ..., \beta_r$ represents the parameters of the model. constant β_0 represents intercept term. By using the scores specified based on the following two assumptions, we can determine the constants A and B. These two assumptions are:

1.A specific expected score is set for a specific odds ratio.

2.The score for an odds ratio doubling (PDO) is specified[5].

First, assume that when the odds ratio is equal to θ , the score is P , and when the odds ratio is equal to 2θ , the score is $P + PDO$. By substituting them into (14), you can obtain the following two equations:

$$P = A - B\log(\theta) \tag{16}$$

$$P + PDO = A - B\log(2\theta) \tag{17}$$

Solving these two simultaneous linear equations yields the constant values for A and B.

$$B = \frac{PDO}{\log(2)} \tag{18}$$

$$A = P + B\log(\theta) \tag{19}$$

Solving the system of two linear equations. We can find the constant values for A and B. The results are:

$$B = \frac{PDO}{\log(2)} \tag{20}$$

$$A = P + B\log(\theta) \tag{21}$$

With the typical assumption that the scoring card results in a score of 600 when the odds ratio is 1:60 (default: non-default), and with PDO set to 20, and given B=28.85 and A=481.86, you can substitute these values into the score calculation formula:

$$Score = 481.86 - 28.85\log(Odds) \tag{22}$$

Usually, A is referred to as the offset, and B is referred to as the scale[6].

By combining Equation (14) and Equation (15), you get the formula for calculating the scoring card score:

$$Score = A - B\left\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right\} \tag{23}$$

4 Conclusions: Risk Management Strategies

Following the computation of the credit score through the scoring card as outlined above, the credit scoring model ultimately categorizes customers into good and bad borrowers based on their overall scores. The overall score is contingent on the individual attribute scores of the variables incorporated into the model, as discussed in the preceding section. With consideration of the distribution of customer overall scores, corresponding score thresholds can be established. For instance, if it is observed that customers with scores below 450 exhibit a default rate exceeding 90%, risk management strategies can be devised as follows: Customers with scores below 450 are designated as high-risk borrowers, prompting measures such as loan rejection, increased interest rates, or the introduction of financial collateral to mitigate associated risks.

These risk management strategies aim to harness customer credit scores for the purpose of managing and mitigating credit risk, thereby ensuring the financial institution maintains a robust performance in its lending and credit operations.

References

1. Durand D. Risk Elements in Consumer Instalment Financing, Technical Edition[M]// Risk Elements in Consumer Instalment Financing. National Bureau of Economic Research, Inc, 1941.
2. Koh H C, Tan W C, Goh C P. A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques[J]. International Journal of Business and Information, 2004,1(1):96-118.

3. Wiginton J C. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior[J]. Journal of Financial and Q uantitative Analysis, 1980,15(3):757-771.
4. Arminger, Enache, Bonne. Analyzing Credit Risk DATA: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feedforward Networks[J]. Social Science Electronic Publishing, 1997,12(2):293-310.
5. Tsaih R, Liu Y J, Liu W, et al. Credit Scoring System for Small Business Loans[J]. Decision Support Ststem, 2004,38(1):91-99.
6. Bensic, Sarlija, Zekic-Susac. Modelling Small-business Credit Scoring by Using Logistic Regression, Neural Networks and Decision Trees[J]. Intelligent Systems in Accounting, Finance and Management, 2005,3(3):133-150.