Analysis of Individual User's Financial Debt Default -- Based on the Construction of Logistic Regression Model

Tianhao Yu

Business, Central University of Finance and Economics, Beijing, 100081, China

2022311390@email.cufe.edu.cn

Abstract. As an important part of the economy and society, financial institutions bear the important task of providing financial support and financial services to the real economy. However, due to various reasons, customers of financial institutions may default, which will not only bring economic losses to financial institutions, but may also have a negative impact on the stability of the entire financial system. This article uses Logistic regression to establish a customer debt default warning model, and analyzes and evaluates customer data from financial institutions. It is found that the occurrence of customer defaults is mainly related to the customer's occupation (whether a student) and the customer's balance. Through the establishment of this model, it aims to discover potential financial default customers and help financial institutions reduce the risks caused by debt defaults, which has certain research value and practical significance.

Keywords: Financial institution, customer debt default, logistic regression, early warning model.

1. Introduction

Financial default refers to the behavior in financial transactions where the borrower is unable to repay the debt according to the agreed time limit and conditions, resulting in losses for the creditor [1]. In recent years, the fluctuation of the global economy and the increasing complexity of financial markets have made financial defaults more common and attracted widespread attention. As a branch of financial default, personal default has become an important social and economic issue. Personal defaults mainly refer to credit card defaults, consumer loan defaults, etc. This type of default usually occurs when there is a problem with a person's financial situation, such as loss of income due to unemployment, illness, etc., or over-consumption leading to debt. Once individual consumers default, they will face consequences such as credit damage, increased interest and fines, collections and even legal proceedings. With the rapid development of the financial market and the continuous improvement of the credit system, financial defaults by individual users are becoming more and more common. Financial default by individual users will not only bring economic losses to financial institutions, but may also affect users' credit records and quality of life. In addition, financial defaults by individual users may also trigger risk contagion effects and have a negative impact on the stability of the entire financial system. Therefore, it is of great practical significance and research value to explore the impact and causes of financial defaults by individual users, as well as how to prevent and control default risks.

Regression model based on the financial debt default situation of 10,000 users, including three independent variables: occupation (whether they are students), balance and income, to explore whether the occurrence of user financial default is related to these three variables. If If there is a correlation, the model can be used to mine whether existing users have financial defaults. It can also predict potential future financial default customers, helping financial institutions identify which users should issue debt warnings, thereby reducing the impact of default users on financial institutions. losses and maintain the stability of the financial system. In the past, relevant research on financial debt defaults mainly focused on corporate defaults [2]. Few studies have conducted in-depth studies on individual users' debt defaults. This article can fill the gap in research on individual users' debt defaults and assist financial institutions in making the best financial decisions. has important research value and practical significance.

2. Data description

2.1. Selection of samples

The sample data selected in this article comes from the data set collected on the kaggle website, with a total of 10,000 users. related information. The author will use Logistic regression to build a model based on these 10,000 pieces of data. The data connection is as follows: https://www.kaggle.com/datasets/creepycrap/finance-dataset.

2.2. Introduction to indicators

In the data of these 10,000 users, the main information included is the user's occupation (whether he is a student), the user's balance, the user's income and whether the user has defaulted. The indicators used in this model are detailed in the following table:

Table 1. Variables and descriptions

symbol	name	illustrate
X_1	User occupation	Is the user a student?
X_2	User balance	User's bank card balance
X_3	user income	User's total revenue
p	Probability	probability of default occurring

3. Method

3.1. Logistic regression model

The logistic regression model is a type of linear regression model. It expresses the probability of an event by mapping the predictor variables to (0, 1) by establishing a linear regression function [3]. It is used for prediction and analysis of binary classification problems. It is simple and easy to understand, easy to implement and other advantages, it is mainly divided into three basic types: binary logistic regression, polynomial logistic regression, and ordinal logistic regression. Currently, logistic regression is widely used in machine learning, medicine, economics, sociology and other fields. There are two possible outcomes for the categorical response of binary logistic regression, which is used in this study to predict users' debt defaults. When the predicted probability p is greater than 0.5, we believe that there is a default phenomenon on the debt, and the prediction result is treated as a dummy variable 1; when the predicted probability p is less than 0.5, there is no default phenomenon on the debt, and it is treated as a dummy variable 0. The general functional form of logistic regression is:

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

Among them, Z is the linear combination of the independent variables X, which can be expressed as $Z = C_0 + C_1X_1 + C_2X_2 + \cdots + C_nX_n$, C_0, C_1, $C_2 \cdots C_n$ are the regression coefficients [4]. The derivation process is: If it is assumed that the probability of an event occurring is p, then its opposite event is $1 - p$. Comparing the two is the chance ratio (odds). After calculating the natural logarithm of p, a regression equation with X as the independent variable is established:

$$\ln\frac{p}{1-p} = C_0 + C_1X_1 + C_2X_2 + \cdots + C_nX_n \tag{2}$$

Taking the same exponential function on both sides, we get:

$$p = \frac{1}{1+e^{-(C_0+C_1X_1+C_2X_2+\cdots+C_nX_n)}} \tag{3}$$

3.2. Application of logistic regression of personal debt default

In this study, the author hopes to use the Logistic regression model to help financial institutions predict whether users will default on their debts. This article selects the historical data of 10,000 users

of a financial institution, including the user's default status, educational background, account balance and income. When building the model, try to select three variables: academic qualifications, account balance, and income as independent variables of the debt default prediction model to explore whether there is a linear relationship between the occurrence of user debt default and these three independent variables. If there is a linear relationship, it means that there is a correlation between the user's debt default and academic qualifications, account balance, and income. Banks and other financial institutions can use this model to rate the user when lending money to the user to determine whether the potential user will default on the debt. Depending on the situation, decide whether to borrow and the amount to borrow.

4. Results

4.1. Data preprocessing

In this study, SPSS was mainly used to analyze and process the data. In the preparatory stage of building a model, we first need to process the independent variables. Since the independent variable of whether the user is a student is not a numeric variable, we create a dummy variable, using the number 1 to indicate that the user is a student, and the number 0 to indicate that the user is not a student, to facilitate subsequent modeling. Secondly, since the user balance and user income amount in the sample are relatively concentrated, and the dimensions are different from the dummy variables we just created, we need to eliminate the impact of the dimension and standardize the user amount and user income to convert the values into A number between 0-1 to achieve this purpose. Use the normalization formula [5]:

$$Z = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4}$$

X_{max} , X_{min} respectively represent the maximum value and minimum value in the two independent variable data of user balance and user income. The normalized result Z is successfully obtained through this formula. Finally, we checked the sample and found that among the 10,000 user data, only 333 had defaults, accounting for 3.33% of the total sample. The ratio of the number of default samples to the number of non-default samples was roughly 7: 200. The ratio is very unbalanced, which will lead to There is an emphasis on the majority category during prediction, making the majority category prediction more accurate, while the accuracy of minority class samples is lower. Therefore, the author downsampled a larger number of non- default sample data, that is, reducing the number of non-default samples used for model construction. In this study, the author selected a total of 666 samples, including 333 samples with defaults and 333 samples without defaults, thus improving the accuracy of model predictions. In order to make the research conclusions more general, the author iterated the sample selection five times, obtained five different sets of data, and calculated the average value. .

4.2. Model construction

The preprocessed data was imported into SPSS for analysis, and a binary logistic model was selected for construction. In order to make the research conclusions more general, the author iterated the sample selection five times, obtained five different sets of data, and calculated take the average. The specific results are as follows:

Table 2. The coefficient of the independent variable

Number of iterations	variable	coefficient	significance level	Prediction accuracy
1	X1	1.09	0.023	88.6
	X2	16.464	0	
	X3	-0.527	0.605	
	constant	-8.65	0	
2	X1	0.735	0.097	89.5
	X2	16.212	0	
	X3	0.638	0.516	
	constant	-8.742	0	
3	X1	0.668	0.125	89.2
	X2	16.287	0	
	X3	0.049	0.958	
	constant	-8.543	0	
4	X1	0.272	0.125	85.3
	X2	14.895	0	
	X3	1.026	0.958	
	constant	-8.164	0	
5	X1	0.458	0.284	87.2
	X2	14.074	0	
	X3	0.044	0.964	
	constant	-7.413	0	

It can be seen from the data in the table that the significance level of user balance and whether the user is a student is less than 0.05, which means that the occurrence of user default is strongly related to the user balance and whether the user is a student, while the significance level of user income is significantly greater than 0.05. This shows that there is no strong correlation between user default and user income, that is, the independent variable of user income can be eliminated in the model, and only the linear relationship between user default, user balance, and whether the user is a student is considered. Finally, the author averaged the five sets of data of the three independent variables of constant, user balance and user's occupation (whether a student), and obtained the following logistic regression model formula:

$$Y = \ln \frac{p}{1-p} = -8.302 + 0.645X_1 + 15.586X_2 \quad (5)$$

They X_1, X_2 respectively represent the user's occupation (whether he is a student) and the user's balance, and Y represents whether the user has committed a breach of contract. X_1 The value of is 0 or 1, X_2 and the value of is any constant greater than 0. When $0.5 < Y < 1$, from this model, we can infer that the user may potentially breach the contract; when $0 < Y < 0.5$, we believe that the user will basically not default. From this our model is constructed. [6]

4.3. Model testing

In order to test the accuracy and feasibility of the model, we need to test the model. In this article we mainly use the R OC curve to evaluate the model. The ROC curve is called the receiver operating characteristic curve, which reflects the comprehensive index of sensitivity and specificity continuous variables. [7] By setting different critical values, a series of sensitivity and specificity can be calculated, and then a curve is drawn with sensitivity as the ordinate and (1-specificity) as the abscissa. The larger the area under the curve, the higher the diagnostic accuracy. On the ROC curve, the point closest to the upper left corner of the graph is the critical value with higher sensitivity and specificity. [8] The following is the ROC curve of the debt default model and the interpretation of the results:

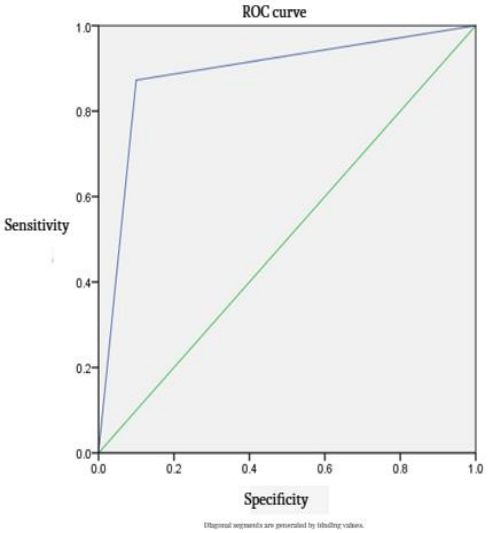

Figure 1. ROC curve

Table 3. Area under curve

Test result variable : default				
area	Standard errora	Asymptotic significanceb	Asymptotic 95% confidence interval	
			lower limit	upper limit
.895	.014	.000	.868	.922
The test result variable default has at least one binding value between the positive actual state group and the negative actual state group. Statistics may be biased.				
a. According to non-parametric assumptions				
b. Null hypothesis: True area = 0.5				

It can be seen from the test results that the area AUC under the ROC curve is 0.895. It is generally believed that an AUC value greater than 0.7 has a high diagnostic value, indicating that the construction of the model has high accuracy and credibility. The construction of the model compares success.

5. Conclusion

In this study, we conducted an in-depth analysis of customer data from financial institutions, aiming to uncover potential financial defaulting customers. Through the logistic regression model, we found that whether a user defaults is significantly related to the user's balance and whether the user is a student. Among them, the user's balance proved to be an important predictor. The results of the logistic regression model show that the lower the user's balance, the greater the likelihood of default. This may be because when users have lower balances, they may face greater financial pressure, thereby increasing the risk of default. This finding provides an important risk reminder for financial institutions that they should pay close attention to those customers with lower balances to prevent possible defaults. In addition , we found that whether a user is a student is also related to breach of contract. In our sample, student users have a higher probability of default. This may be because student users typically do not have a stable source of income and are more susceptible to economic fluctuations. [9] At the same time, student users may be more inclined to take on higher debt loads, which also increases their risk of default. This finding suggests that financial institutions should be more cautious when providing financial services to student users and consider taking appropriate risk management measures. In this study, we did not find a strong correlation between user income and whether users default . This may be because users with higher incomes generally have stronger repayment capabilities, thus reducing the risk of default. However, it is important to note that high income does not mean the complete elimination of default risk. Financial institutions

still need to focus on customers with higher incomes but also higher debt levels to prevent possible defaults.

In summary, by analyzing customer data of financial institutions, we successfully unearthed factors that are significantly related to financial defaults. These factors include the user's balance, whether the user is a student, and the user's income. These findings provide financial institutions with valuable risk tips and management recommendations, helping to prevent potential financial defaults.

However, this study still has certain limitations. For example, there are few independent variables selected in this study, and the impact of factors such as the macroeconomic environment, policies and regulations on financial defaults is not taken into account [10]. Future research can further expand on this basis and incorporate more relevant variables into the model to improve the accuracy and reliability of predictions. In addition, this study mainly focuses on quantitative analysis, but does not provide an in-depth exploration of the intrinsic motivations and influencing factors of breach of contract. Future research can use qualitative research methods, such as interviews, focus groups, etc., to reveal the deep-seated causes and influencing factors of financial defaults.

Overall, this study reveals factors that are significantly related to financial defaults by analyzing and mining customer data from financial institutions. These findings help financial institutions better manage and prevent financial defaults, thereby contributing to the stability and development of financial markets.

References

[1] A. Rey, G. Land, The Role of Financial Reporting Quality in Accessing to Financial Debt: Evidence from Italy, International Journal of Accounting and Financial Reporting (3), 2018.

[2] G. Jing, Z. Xinmin, Aggressive Corporate Strategy, Credit Cycle and Debt Default. Foreign Economics & Management, 07, 38-53, (2021).

[3] X. Zhengde, L. Zhenxuan, X Yipeng, Research on Customer Default Risk of Internet Finance Based on Ordered Logistic Model. Systems Engineering (08), 29-38, (2017).

[4] P. Zeqing, Logistic regression early warning model for enterprise debt default risk, Shanghai Economic Research, 08, 73-83, (2017).

[5] N. Gaurav, H. Anil, T. Bharat, K. Chetan, Novel MIA-LSTM Deep Learning Hybrid Model with Data Preprocessing for Forecasting of PM2.5. Algorithms (1), (2023).

[6] Mi Z., Xing L., Mengyu H., Songshan H., Jiayang Z., A Downsample Strategy for AGV Life-Long Localization.. (eds.) Proceedings of 2020 the 7th International Conference on Automation and Logistics (ICAL 2020) (pp.42-46). ACM, (2020).

[7] Z. Guoping, On the analytical properties of category encodings in logistic regression. Communications in Statistics - Theory and Methods (6), (2023).

[8] Sitnik H., Stepanova Y., Biochemical Markers in the Tear Fluid of Patients with Keratoconus: Results of a ROC Analysis. Ophthalmology Eastern Europe (3), (2023).

[9] W. Yunyun, C. Songcan, Review of Classifier Evaluation and Design Based on AUC. Pattern Recognition and Artificial Intelligence (01), 64-71, 2011.

[10] D. Yuanzhu, Generation Mechanism and Risk Prevention of Corporate Debt Default (PhD dissertation, Jilin University), (2022).