Credit Default Prediction Based on Multivariate Regression

Yingzi Sun[1], Lirui Yang[2], and Ruonan Zhao[3](✉)

[1] University of Arizona, Tucson, USA
ysun1@email.arizona.du
[2] Guangzhou Foreign Language School ISA Wenhua IB Programme, Guangzhou, China
ryangbos@163.com
[3] Pearl River College, Tianjin University of Finance and Economics, Tianjin, China
18404191@masu.edu.cn

Abstract. Credit default is a wide-spread credit derivative instrument. As it becomes more and more popular, an appropriate supervision system has to be established. In this paper, a multiple factor regression models are constructed in order to investigate the feasibility for credit default prediction based on R program. Since risks are unavoidable, some measures should be taken to predict them in order to help the banks that sell credit default swaps to minimize their risks. According to the analysis, a model is successfully created. These results shed light on guiding further exploration focusing on credit default prediction.

Keywords: Credit Default · risk · logistic regression

1 Introduction

Credit default swaps (CDSs) are financial derivative instruments because their financial value is derived from the value of an underlying financial asset, usually a bond. The ability to trade derivatives allows the various risks of an asset to be transferred to counter-parties willing to bear them without the underlying asset being involved in the trade–or even being held by either the buyer or the seller [1]. In other words, the seller of the CDS provides the buyer with some kind of insurance against some reference default. The buyer of the CDS pays a series of fees to the seller. In exchange, the buyer may expect to receive a payoff if the asset defaults. The origin of the CDS market dates back to the early 1990s, in the aftermath of the Exxon Valdez oil spill of March 1989, the United States bank, JP Morgan, bought protection against a possible Exxon default from the European Bank for Reconstruction and Development (EBRD) [2]. This contract reduced JP Morgan's exposure to Exxon and increased the return on EBRD reserves that could only be used to lend to high rated borrowers [2].

CDCs are getting more and more popular nowadays. The evidence is that this trend continues; participants in the market for credit protection have increased in number and

Y. Sun, L. Yang and R. Zhao—Contributed equally.

Y. Jiang et al. (Eds.): ICFIED 2023, AEBMR 237, pp. 16–23, 2023.
https://doi.org/10.2991/978-94-6463-142-5_3

level of activity. As the market for credit default swaps grows in size and importance, few major financial institutions, whether they are natural buyers of protection (usually banks), sellers of protection (typically hedge funds, insurers, and reinsurers), or both (dealers), are without a credit derivatives operation of some kind [3].

According to some recent research, credit default swaps can enhance the credit service capacity of small and medium enterprises (SMEs). On the other hand, they can help alleviate the credit risk pressure that banks may incur by expanding credit support to SMEs under industrial transformation and economic environment uncertainty. Thus, it plays a positive role in credit resources to stimulate the transformation and development of SMEs.

However, CDCs always come with numerous default risks. Default risk arises any time bank funds are extended, committed, invested, or otherwise exposed [5]. Therefore, it is necessary for banks to minimize the risks. Sound sanctioning process, appropriate administration, measurement and monitoring processes and adequate controls and mitigation systems for default risk are effectively contributing to reduce default risks [5]. In addition to these means, our commentary below will mainly focus on the way to predict the potential risk of credit default swap for a firm and predict whether it can lend money with low risk [6–9].

We have used several methods to determine the risk. Firstly, with R studio, the multivariate correlation analysis was done, and according to the correlation coefficient matrix heat map, the variables PAY and BILL_AMT are strongly positively correlated, and the variable PAY_AMT is weakly negatively correlated. Secondly, the logistic regression was done with R studio as well, the VIF of the strongly correlated factors was calculated, and the regression model is reliable with no multicollinearity. Thirdly, we have applied logistic regression.

The rest part of the paper is organized as follows. The Sect. 2 will introduce the model establishment and data origination, which shows the possibilities of each individual (differed by gender, education level, marital status, and age) to pay back the loan or not. In Sect. 3 of the paper, with the data from Kaggle Website, logistic regression is adopted to re-calculate the possibilities in Sect. 2. Eventually, a brief summary is given in Sect. 4.

2 Methodology

2.1 Data

The dataset for this study was obtained from Kaggle Website [10]. This dataset contains 30000 rows and 24 columns. In the whole processing process, we were using RStudio application doing all the data analysis. The dataset can predict whether the customers can make a Default Credit Card payment for the next month or not, so it's our target variable (dependent). Rest all 23 variables considered independent variables, which have the information of the various details about the customers such as Balance, gender, age, marriage, and the details of their previous transactions. Here, the independent variables have a combination of both categorical and continuous variables. During the analysis, the categorical response variable was converted to factors, then proceeded with the further analysis.

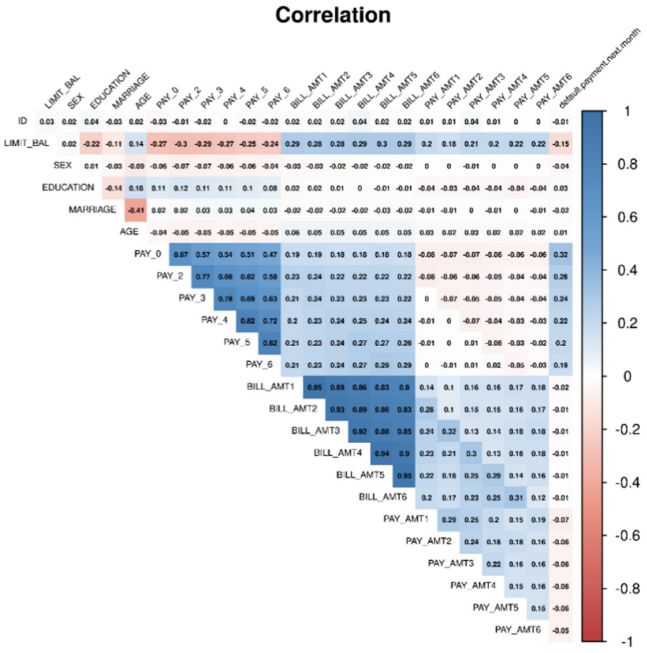

Fig. 1. Correlation coefficients.

2.2 Correlation Analysis

As shown in Fig. 1, the strength of the relationship between two variables, and its expressed mathematically by the correlation coefficient. It ranges from −1 to 1, where −1 indicates the negative correlation where both the variables are going in an opposite direction, and +1 indicates a positive correlation between the variables where both the variables are going in the same direction. Lastly, 0 indicates there's no correlation between the two variables. It can be mathematically described as:

$$r = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{\sqrt{([x\Sigma x\hat{}2 - (\Sigma x)\hat{}2][n\sum y\hat{}2 - (\sum y)\hat{}2)}} \tag{1}$$

where r = Pearson's coefficient, n = number of observations, $\sum$xy = summation of products x and y, $\sum$x = Summation of x, $\sum$y = Summation of y, $\sum$x2 = sum of squared x, $\sum$y2 = sum of squared y. The correlation matrix is given in Fig. 1, it can be observed that there exists a weak correlation of AGE, BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6 with our target variable default payment next month in feature selection, all these variables, which have a weak correlation with the target variables, have been removed for modeling to reduce the dimension of the overall data as there were 24 variables in the original model. Then, Logistic regression was built for all the remaining independent variables.

Table 1. Summary of the Logistic Regression

	z value	Pr(>|z|)
LIMIT_BAL	−4.436	9.16
SEX	−3.109	0.002
EDUCATION	−3.847	0.001
MARRIAGE	−4.826	1.4
PAY_0	27.136	0.001
PAY_2	3.301	0.001
PAY_3	2.72	0.006
PAY_4	0.831	0.406
PAY_5	1.448	0.148
PAY_6	0.228	0.82
BILL_AMT1	−4.829	1.37
PAY_AMT1	−4.034	5.49
PAY_AMT2	−3.989	6.63
PAY_AMT3	−1.469	0.142
PAY_AMT4	−2.338	0.019
PAY_AMT5	−1.949	0.051
PAY_AMT6	−2.551	0.011

2.3 Logistic Regression Models

A logistic regression equation is used to obtain the relationship between the dependent variables with one or more independent variables. Here our dependent variable has to be categorical, especially binary. In this study, we have chosen Logistic regression as our response variable is binary. At the same time, we also calculated the Variance Inflation Factor (VIF) and R-Squared to reveal the multicollinearity of the model. The summary of the Logistic regression model is shown in Table 1. From the summary of the model, overall, the model took six iterations to build this model, and the AIC value is 19711, which is high. If we observe the p-value, few variables have ($p < 0.05$), which indicates we can reject the null hypothesis and say the predictors are associated with changes in the response variable.

The confusion matrix clearly says there were 6847 no default payments as listed in Table 2. The model has also predicted no default payment; 476 said yes for a default payment, and the model has also expected yes for the default payment. Hence, there were $6847 + 476$ correct predictions and $1472 + 205$ wrong predictions. The overall accuracy of this model was 0.81, which is closer to 1, so it can be said that the obtained model is good. Also, McFadden's R-squared value for the model was 0.117, which is also closer to 0.2, so it's a good model.

Table 2. Confusion matrix

log.prediction.rd	0	1
0	6847	1472
1	205	476

3 Results and Discussion

First, we need to perform correlation analysis to determine whether there is a correlation between the numerous independent variables and the dependent variable known in the study. When there is no correlation between the variables and the dependent variable after correlation analysis, regression analysis is not needed; if there is some correlation, regression analysis is performed to further verify the exact relationship between them.

Additionally, correlation analysis is done to test the degree of cointegration between the independent variables; if the correlation between the independent variables is high, it may indicate the existence of a cointegrating relationship.

According to the correlation analysis, we can clearly see that several independent variables, especially PAY0, PAY2 and so on, show a high positive correlation with the dependent variable. It shows that when a borrower pays back the loan on time in the past is positively correlated with whether he returns it as scheduled after this borrowing, and a high value of past repayment is associated with a high probability of returning it as scheduled after this borrowing. Accordingly, one can determine that it affects his probability of not repaying the loan next month when a borrower has borrowed in the past and failed to repay on time. Marker consideration ought to be given to this situation whether it is delayed by one month, two months or three months.

In the variable BILL_AMT, it is found that most of the data show a positive correlation with the dependent variable, i.e., when the borrower has a higher historical deposit balance, the higher the probability of repaying the loan on time. The correlation analysis of bill can be applied based on the borrower's historical deposit balance. It means that when the borrower's balance is higher, his ability to repay the loan is higher and the relative risk we take is lower.

Considering the PAY_AMT factor, it is weakly positively correlated with the dependent variable, which means that when the borrower's past spending power is strong, the likelihood that he will repay the loan on time is high. Therefore, when the borrower we assess has a strong past spending power, the borrower tends to have a stronger ability to repay the loan. However, since the correlation between this variable and the dependent variable is extremely weak, we cannot consider the effect of this factor too much.

It is also found that the parameters ID, LIMIT_BAL, SEX, and AGE were less correlated with the dependent variables. Therefore, for the evaluation of borrowers, we should favor the independent variables PAY, EDUCATION, MARRIAGE, and BILL. Therefore, regression analysis is used to solve it. In logistic modeling (seen from Table 1), we need to analyze the presence of multicollinearity. If there is multicollinearity in the model, it will lead to non-existence of parameter estimates, unreasonable significance of parameter estimates, loss of significance test of variables, possible exclusion of important

Table 3. Model Coefficients

LIMIT_BAL	SEX	EDUCATION	MARRIAGE	PAY_0	PAY_2
1.492	1.005	1.096	1.05	1.507	2.647
PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	PAY_AMT1
3.218	3.81	4.147	2.915	1.368	1.153
PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	
1.122	1.127	1.11	1.091	1.082	

independent variables from the model, and failure of the predictive function of the regression model. Therefore, after obtaining the value of R square, we need to calculate the Variance inflation factor to eliminate the multicollinearity.

Seen from Table 3. There is no Variance inflation factor value greater than 10 for the relevant variables. Besides, the values of SEX, Pay, and Marriage are very close to 1, indicating that the multicollinearity in this model is very light. Therefore, with VIF close to 1, we consider that the fluctuations of the regression coefficients are very small and do not bring a lot of uncertainty to our parameters. Therefore, one can assume that for the evaluation of borrowers, it is able to carry out linear regression model predictions based on the regression coefficients of other parameters, except for the weakly correlated parameters.

Then, we proceed to Optional, which is tested according to the parametric test, and set null hypotheses and alternative hypotheses. One can predict the future repayment of borrowers based on all the monthly repayment data surveyed. With this in mind, null hypothesis is set up, i.e., assuming that borrowers do not repay after borrowing. By looking at the p-values of variables with high correlation, we found that most of the values are greater than 0.05. Thereby, one should reject null hypothesis at 5% significance level. It is 95% confident that this borrower can repay after borrowing.

There are some limitations and shortcomings of the paper. First, the data factors we can collect do not cover all aspects, even though the current data already includes some factors that have very low correlation with credit. To be specific, we do not know exactly whether a relative or friend of the borrower can help the borrower to repay the debt, which also has an impact on our judgment. The second point is our assessment of the uncertainty of an entrepreneur applying for a loan, since it is unable to acquire enough about many areas to support the ability to accurately determine whether an entrepreneur will be able to get his earnings within the expected time frame. In addition, it also affects whether we borrow or not, and we are not sure how much risk we will take. The third point is that our forecasts are made in a stable economic market environment and will change in the event of a sudden economic crisis-like event.

4 Conclusion

In summary, this paper investigates the feasibility to predict credit fault based on multiple linear regression. According to the analysis, it is found that the model we used well

fitted with the training data. Based on reviewing the final Squared test, the effectiveness of the model is verified. In the future, it is hoped that there could be more research about different places and wider ranges of time scheduled while collecting the data. For investor lending, the data from multiple institutions can be obtained and utilized to evaluate, similar to the industry outlook, and partner qualifications. In addition, in the face of economic crisis and other situations we need to make backup measures to help the lending platform through difficulties. This kind of research could be used in order to better develop the bank's safety problems and to better allocate the money value to other people who need the money the best. According to the analysis in this paper, it could help the bank predict whether the person is capable of paying back the bill in the next month, if the person is at risk of cannot afford the next month's payment, the bank could take action first in order to protect the bank's losses. Overall, these results offer a guideline for building multi-factorial regression models for credit default forecasting and further exploration on this topic.

References

1. K. Cherny, and B. R. Craig. "Credit default swaps and their market function." Economic Commentary July 2009 (2009).
2. Online information "Credit default swaps." Trade and Development Report, annual 2010, p. 33+. Gale General OneFile, Avaliable at: link.gale.com/apps/doc/A239089452/GPS?u=chgzfls&sid=bookmark-GPS&xid=14d9dba8.
3. L. Marshall, "Credit default swaps and the issue of restructuring: documentation, market practice, and risk management. (Credit Default Swaps)." The RMA Journal, vol. 85, no. 8, May 2003, p. 32.
4. R. M. Stulz, "Credit Default Swaps and the Credit Crisis." Journal of Economic Perspectives, vol. 24 (1), 2010, pp. 73-92.
5. A. Lipton, A. Sepp. "Credit value adjustment for credit default swaps via the structural default model." The Journal of Credit Risk, vol. 5(2), 2009, pp. 127-150.
6. J. C. Hull, A. D. White, "Valuing credit default swaps II: Modeling default correlations." The Journal of derivatives, vol. 8(3), 2001, pp. 12-21.
7. H. Wen, "Analysis of the causes of credit bond defaults and countermeasures from the perspective of investor protection." China General Accountant, vol. 12, 2021, pp. 115-117.
8. P. Jorion, G. Zhang, Good and bad credit contagion: Evidence from credit default swaps. Journal of Financial Economics, vol. 84(3), 2007, pp. 860-883.
9. S. Dieckmann, T. Plank, "Default risk of advanced economies: An empirical analysis of credit default swaps during the financial crisis." Review of Finance, vol. 16(4), 2012, pp. 903-934.
10. Dataset from Kaggle, available at: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset?resource=download