An application of logistic regression on risky loan status prediction

Haonan Xu, Su Huang, Zihan Li

March 26, 2023

Abstract

Loan defaults (unsuccessful fully paying the loan in a prescript time) may cause financial losses for financial institutions, therefore it is important to evaluate the customers with their loan history and social characteristics before lending. In this paper, we proposed a logistic regression method which to predict the loan status and figure out the influential factors that result in risky loan status. In the modeling process, we focused on the selection of factors which was highly correlated or contain no/partial information originally. Moreover, we want to suggest a general idea of risky loan prediction for the lending platforms by explaining the predicted result.

1 Introduction

The probability of Default (PD) is an important assessment of risk connected with loans to organizations and individuals in credit risk modeling[1]. A defaulted customer is defined as one who is not able to fully pay the loan within the prescribed time (This is a standard description of default, for some countries/regions, financial institutions follow different laws[2]). Because numerous loan defaults caused huge financial losses and may lead to bankruptcy, it is important to evaluate customers before a valid loan to avoid losses to a greater extent.

Several efforts have been made to estimate loan credit risk with regression[3][4][5][6], neural networks[7], etc. This paper focuses on a logistic regression method that provides 1) a better fitness in predicting the binary result of the default status, and 2) high computational efficiency. With the generated model, this paper also finds out the significant factors that result in risky loan status. Thus the hypothesis is H_0: factors have no effect on loan default; H_1: at least one factor has an impact on loan default.

2 Data Description

The data is collected from an online lending platform LendingClub[8], which recorded customers' borrowing history and social characteristics between 2007-2011, with 42,535 observations and 115 variables. In this paper, 15 factors are selected after eliminating missing variables, unique identifications, and repetitions in order to build a model with high reliability and universal reference value. For the selected data, there are four data types: continuous quantitative, discrete quantitative, binary, and non-binary categorical variables.

1. Response Variable
   For the response variable, we define a bad customer(Defaulted) as who is not able to fully pay the loan within the prescribed time. The `Default` is a binary variable with 1 representing the customer who failed to fully pay the loan in the lending history.

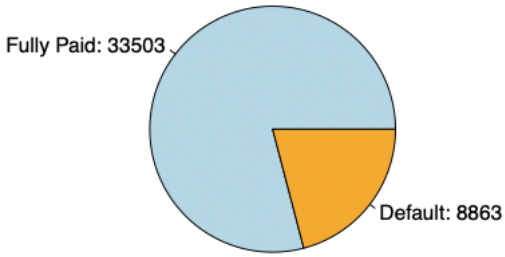

Figure 1: Distribution of Response Variable

2. Predictor
   Two notable changes are made to the original data: `emp_title` is transformed to `unemployed`, a binary categorical variable with value 1 representing the customer has no current employment; `grade` is converted from categorical A-G to a discrete numerical variable of 1-7, 1 represents the best and 7 repre-

sents the worst. These transformations are made for 1) better fitness to the regression model, 2) ignoring non-significant information, and 3) the convenience of computation.

After modification, the variables present as below:

1. annual_inc: Annual income provided by borrowers.
2. emp_length: length of employment in years.
3. unemployed:
4. fico_range_high: Highest FICO credit score of borrowers.
5. funded_amnt: The total amount committed to that loan at that point in time.
6. funded_amnt_inv: The total amount committed by investors for that loan at that point in time.
7. grade: LC assigned loan grade. We convert the scale from A-G to 1-7, where 1 indicates a grade A customer.
8. home_ownership: The homeownership status provided by the borrower during registration. We only keep the value with RENT, OWN, and MORTGAGE while excluding OTHER and NONE.
9. installment: The monthly payment owed by the borrower if the loan originates.
10. int_rate: Interest Rate on the loan. We convert the data from string to double.
11. loan_amnt: The listed amount of the loan applied for by the borrower.
12. default: Whether borrowers fully paid their loan. We set borrowers who fully paid the loan as 0 and all others as 1. We choose default as the response variable.
13. open_acc: The number of open credit lines in the borrower's credit file.
14. term: The number of payments on the loan. Values can be either 36 or 60 months.
15. verification_status: Indicates if income was verified by LendingClub, not verified, or if the income source was verified.

3 Exploratory Data Analysis

Based on the given dataset, a logistic regression method has been proposed since our response variable `default` has binary outcomes. Before constructing the model, we look through the correlative relationship between each factor (Figure 2). With the result, one significant challenge of constructing a specific model is some of the variables may highly correlated, or contain partial/no information on the default status. To address this issue, the correlation of each variable needs to be determined, then the factors of partial/no information should be deleted as well. After the final model has been founded, some work is necessary to evaluate and explain the model.

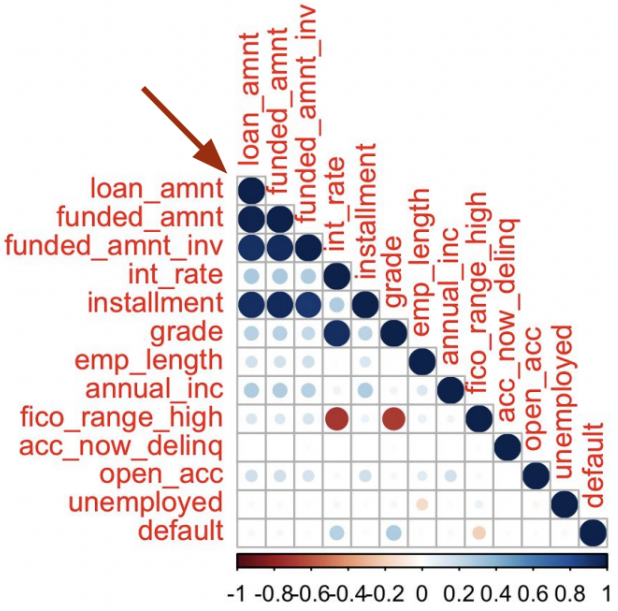

Figure 2: Correlation Plot of 15 Variables

4 Methods

1. VIF

   In order to find those highly correlated factors, a VIF test should be conducted before generating the final model. 5 variables with large VIFs are observed after constructing a full model logistic regression, which indicates strong multicollinearity between them. The largest VIF is 85.99 of `funded_amnt`, meaning most of its variation can be explained by other variables. We use VIF to evaluate correlations between each variable, the reduction of factors could lead us to construct an accurate model with higher computational efficiency. After dropping all possible combinations of these variables, we decided to keep `loan_amnt` only, which provides the model with the lowest VIF results. The new model reduces multicollinearity significantly, having the largest VIF of 3.04 for `grade`.

2. AIC

   Stepwise(Backward)-AIC is the first regression model applied to the rest of the variables. The

reason for using Stepwise AIC/BIC is that they are based on a rigorous statistical criterion and can handle a large number of predictor variables. In this case, the AIC removes the variable emp_length. The final AIC value of the model is 32010 and Residual Deviance is 31990. The only variable with a p-value over the level of $\alpha = 0.05$ is home_ownershipOWN with 0.372 containing 0 in its confidence interval. Besides that, all other variables are statistically significant.

3. BIC
   A Stepwise(Backward)-BIC method has also been applied. Comparing the result from AIC, BIC removes the variable emp_length, home_ownership, and open_acc. The final BIC value of the model is 32020 and Residual Deviance with 32000. It provides the model that all variables are statistically significant at the level of $\alpha = 0.05$.

4. Final Model
   We choose the logistic regression model under the BIC criterion because more penalization is given by BIC. The final logistic regression model given by the backward step function using the BIC criterion is:

$$\begin{aligned}\log(\frac{p}{1-p}) = 0.1488 &- 0.000008 \times \texttt{loan\_amnt} \\ &+ 0.3259 \times \texttt{term(60 months)} \\ &+ 0.4130 \times \texttt{grade} \\ &- 0.000002 \times \texttt{annual\_inc} \\ &- 0.3365 \times \texttt{vs(Source Verified)} \\ &- 0.1719 \times \texttt{vs(Verified)} \\ &- 0.0036 \times \texttt{fico\_range\_high} \\ &+ 0.0104 \times \texttt{open\_acc} \\ &+ 0.44 \times \texttt{unemployed(1)}\end{aligned}$$

5 Results

1. Prediction

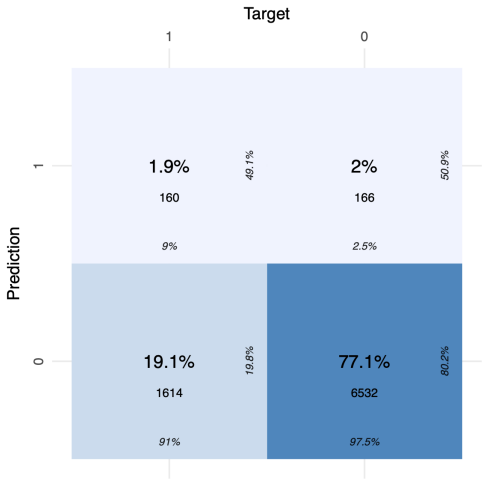

Figure 3: Confusion Matrix for Prediction

Figure 3 shows the prediction of the testing data using logistic regression. The overall accuracy is 77.1% (True Negative) + 1.9% (True Positive) = 79%, which is acceptable; however, when we compare the prediction result, 96.2% candidates in the testing group are labeled as fully paid. Only 3.8% are labeled as default, which is far from the actual default ratio of 21%.

The strong tendency might indicate an underfitting model, as our regression model might be too simple for the analysis. We will address this issue later in the discussion part.

2. Fitted vs Actual Response

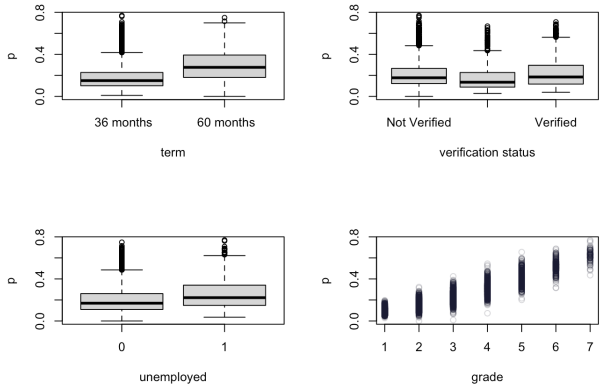

Figure 4: PD for Discrete and Categorical Variables

The plot Fitted vs Actual response shows the relationship between the variable in the final model and the fitted value. From Figure 4, we can see that variable term and unemployed have a clear relationship between fitted values when the category

converts from one to another. The fitted value increases when the variable grade level gets worse. However, for the variable verification status, the distribution is not that obvious from this box plot when the category changes.

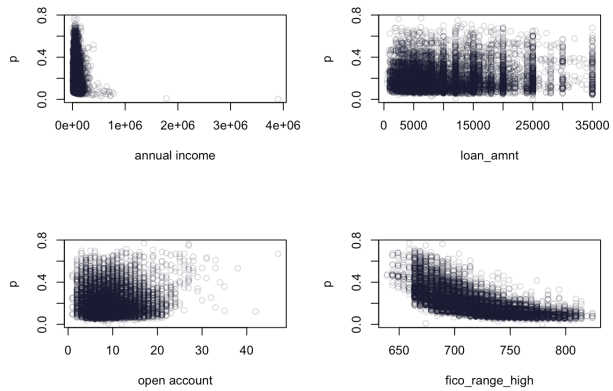

Figure 5: PD for Continuous Variables

From Figure 5, we can see when annual income and fico score increase, the fitted value decreases. For the variable loan_amnt and open account, there are no significant changes in fitted value when they change.

3. Cook's Distance

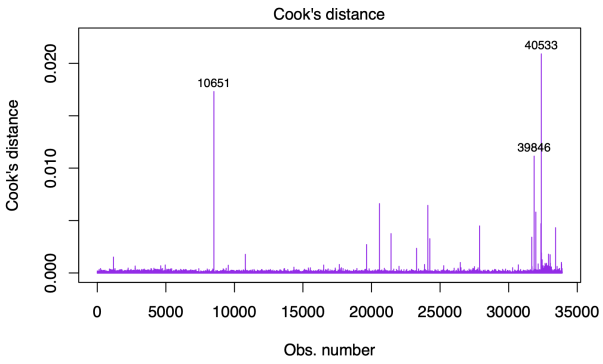

Figure 6: Cook's Distance for All Observations

Figure 6 shows the three most influential observations with the largest Cook's Distance. We examined these three defaulted examples, and one of them shows a good FICO score of 714 with a verified income of \$125,0000. It is an unusual case.

The other two individuals have decent incomes and credit scores, but their source of income is unverified, which might explain the default on loans.

4. Odds Ratio
Using the coefficients of logistic regression under the BIC criterion, we want to see how the odds ratio changes with respect to the change in each variable.

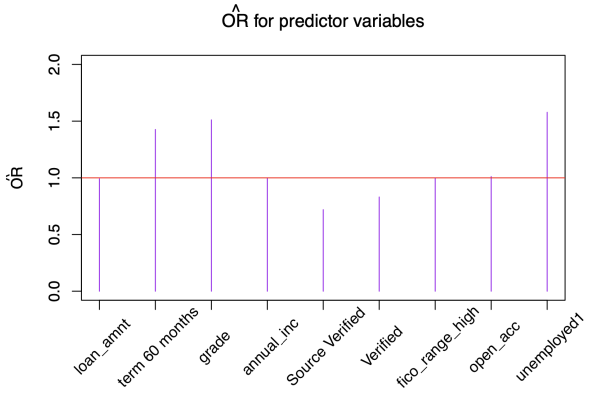

Figure 7: $\widehat{OR}$ for BIC Selected Predictors

Figure 7 shows the change in odds ratio when each variable increases or shifts by one unit. There are several variables that significantly affect the odds of default. The variable `unemployed1` has the largest $\widehat{OR}$ of 1.5778, meaning when an individual's employment status moves from employed to unemployed, the odds of defaulting on a loan is about 1.58 times higher. The $\widehat{OR}$ are similarly large for `term` and `grade`, which meets our expectations. In real-world applications, longer-term and worse grades are very likely to increase the odds of default.

The verification status has a negative impact on the odds of default. A customer having either a source verified or verified income will have fewer odds of default on a loan because when the variable of source verified and verified is equal to zero, a part of that customer group has their income unverified, which is a signal to default.

Other variables have a $\widehat{OR}$ very close to 1, having minimal effects on the odds of default.

6 Discussion and Conclusion

1. Conclusion
Back to the hypothesis we have, factors are all statistically significant in the final model which means we can reject H_0 and conclude there is data supporting that at least one factor has an effect on default. The variable unemployed has the strongest effect on default.

2. Limitation

   There are a few limitations in our application of the regression analysis. First of all, there are 115 variables, and we only selected 15 of them, which is an oversimplification of the data structure. We might ignore some statistically significant variables that could help us better predict the odds of default.

   Moreover, during our analysis, we excluded all interactions between the predictor variables. However, it is possible that there is some relationship between the person's credit history and social characteristics. Excluding the interaction might enlarge the residual and reduce the prediction accuracy.

   Finally, we chose to ignore the outliers. There are some variables, such as `annual_inc` or `open_acc`, having outliers that are far from the masses of the distribution. These values have minimal reference value, therefore excluding them could benefit the prediction accuracy.

3. Future Study

   First of all, we want to address the issues discussed in the previous section. We will add more predictor variables to the regression, include the interaction among variables, and examine the outliers. Among the variables that are worth considering, one specific factor is the issued date and loan type. The data is collected between 2007 to 2011 and it includes short-term mortgage loans. The 2008 financial recession brought a huge amount of mortgage default. Therefore, it is necessary to take systematic risk into account.

   Moreover, our regression model shows a strong tendency of labelling all candidates as fully paid. We might consider changing the probability threshold, which may give us a higher accuracy.

   In the future, we may consider more advanced models such as deep dense convolutional networks[11], SVM[12].

7 Acknowledgement

We would like to express our sincere appreciation to Prof. Cardoso and Pete Zuckerman for giving us great help on such an intriguing project and providing us with the opportunity of applying the statistical methods in a practice field.

References

[1] Costa E Silva E, Lopes IC, Correia A, Faria S. A logistic regression model for consumer default risk. J Appl Stat. 2020 May 5;47(13-15):2879-2894. doi: 10.1080/02664763.2020.1759030. PMID: 35707418; PMCID: PMC9041570.

[2] Beck R., Jakubik P., and Piloiu A., Key determinants of non-performing loans: New evidence from a global sample, Open Econ. Rev. 26 (2015), pp. 525–550. doi: 10.1007/s11079-015-9358-8

[3] Karan M., Ulucan A., and Kaya M., Credit risk estimation using payment history data: A comparative study of Turkish retail stores, Cent. Eur. J. Oper. Res. 21 (2013), pp. 479–494. doi: 10.1007/s10100-012-0242-y

[4] Kuangnan Fang H.H., Variable selection for credit risk model using data mining technique, J. Comput. 6 (2011), pp. 1868–1874.

[5] Mestiri S. and Hamdi M., Credit risk prediction: A comparative study between logistic regression and logistic regression with random effects, Int. J. Manage. Sci. Eng. Manage. 7 (2012), pp. 200–204.

[6] Westgaard S. and van der Wijst N., Default probabilities in a corporate bank portfolio: A logistic model approach, Eur. J. Oper. Res. 135 (2001), pp. 338–349. doi: 10.1016/S0377-2217(01)00045-5

[7] Crook J.N., Edelman D.B., and Thomas L.C., Recent developments in consumer credit risk assessment, Eur. J. Oper. Res. 183 (2007), pp. 1447–1465. doi: 10.1016/j.ejor.2006.09.100

[8] `https://data.world/jaypeedevlin/lending-club-loan-data-2007-11`

[9] RStudio 2022.07.2+576 "Spotted Wakerobin" Release (e7373ef832b49b2a9b88162cfe7eac5f22c40b34, 2022-09-06) for macOS.

[10] R version 4.1.2

[11] Kim, J-Y, Cho, S-B. Predicting repayment of borrows in peer-to-peer social lending with deep dense convolutional network. Expert Systems. 2019; 36:e12403. https://doi.org/10.1111/exsy.12403

[12] Zhou, X., Wang, H., Xu, C. et al. Application of kNN and SVM to predict the prognosis of advanced schistosomiasis. Parasitol Res 121, 2457–2460 (2022). https://doi.org/10.1007/s00436-022-07583-8

8 Appendix

For this paper/project, Haonan Xu, Su Huang, and Zihan Li work on result analysis, methods and modeling, introduction and data separately. The project is now open with all the source code and can be found at `https://github.com/doghanl/loan_default_prediction-LRM-.git`