BU.510.650

Dec 9, 2021

**BU.510.650 -** **Categorical Data Project**

**Loan Prediction**

1. **Introduction**

For most lending industries such as Lending Club and Sallie Mae, also known as P2P lenders, they use online trading platforms as lending channels, undisturbed by traditional financial intermediaries such as Banks. P2P platforms have become popular recently due to the face that it reduces funding costs, and it also brings higher profitability to both borrowers and lenders. Borrowers benefit from lower interest rates; lenders get a higher return than they get from the bank, and a better rate of return can be efficiently calculated. However, in small loans, assessing the credit of loan applicants is a common challenge, and loans are usually unsecured. Companies wanting to be profitable not only require a certain number of customer groups but a way to predict the customer payment status and reduce potential risk are important indicators of sustainable profitability. Therefore, the purpose of our project is to determine whether the applicant's loan can be approved, eventually to use a logistic regression model to maximize the correct classification of potential and non-defaulters.

**2.** **Data**

To automate this process, the company collects a set of information when lending to the lender, identifying the customer base that qualifies for the loan amount so that they can target those customers specifically. The loan dataset we use in this project is the publicly available on Kaggle. The data they collect includes Gender, Married, Education, Self-employed, Credit History, Loan\_Status, these are binary and categorical predictors. Dependents and Property\_Area are categorical predictors with three different levels, Dependents contain categories 1, 2, and 3+, Property\_Area includes Urban, Rural, Semiurban. We have four continuous variables, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term, Applicant\_Income. Finally, the response variable for this dataset is Loan\_Status, everything else will be our predictors.

**3.** **Data exploration**

The raw data contains a total of 615 observations and 13 variables. However, not all of the predictors are useful for our models, specifically loan ID, and thus we remove it first. We also have a lot of missing values in this dataset, they are a total of 149 missing values from different predictors. By looking at the dataset, I realized that there is no relationship between these missing values, so I assumed that missing value are randomly missing, and I have imputed the missing values in different ways. By exploring the data visually, I created a data visualization for each predictor, and made contingency tables to check the percentage of the predictor in relation to loan status. The imbalanced dataset is a typical issue in the social lending platform, in this dataset, we had 422 people who paid the loan, and 193 people who did not. Predicting loan risk from an imbalanced dataset is imprecise because the imbalanced data affects the model's ability to maximize the correct classification of potential and non-defaulters.

**4.** **Feature expansion**

The first component in the model is to extract features from raw data via data mining techniques. Our techniques included data cleaning, data transformation, correlation analysis, and deriving new attributes. Its main purpose is to improve the reliability of the data by cleaning the data and selecting the data feature subset with the maximum discriminating ability. The outliers and missing or empty values ​​are detected through data exploration, then, I took a different approach to deal with the missing values. First, I replaced the missing values using the mean of the non-missing observation for that predictor, similarly for categorical variables, I used the category that appears the most frequently. Finally, I converted the Dependents variable to a continuous variable in order, for the Property\_Area variable with multiple classifications, I divided each level into a new indicator variable such as Urban, Rural, and Semiurban. After completed all missing value, I replaced outliers with lower and upper cutoff values. For example, the upper limit is computed as 1.5 \* IRQ, where IRQ is equal to 3rd Quartile minus 1st Quartile. Once the data has been cleaned, I tried to do some log transformation since I found that CoapplicantIncome and ApplicantIncome are really right-skewed, but since these predictors have a lot of zero attribution, the log transformation would be undefined, therefore, I decided to keep the original data for further analysis. Finally, the correlation between each predictor is then calculated based on the loan state to better understand the data and its attributes, and there is no multicollinearity issue.

**5.** **Methodology**

Firstly, I split the cleaned data into training sets and test sets, then normalized continuous variables, and finally created a relatively stable data set for our model training. To the best of our knowledge, logistic regression is one of the good performances in binary classification. Therefore, we started with logistic regression by using all the variables. It was found that many variables were not significant, so it was considered to eliminate these insignificant variables. Stepwise regression method was used here to select variables. Through purpose selection, our final model included Gender, Credit\_History, Rural, Urban, LoanAmount, and CoapplicantIncome. From this model, the P values of all variables were less than 0.1, which passed the significance test and retained the relatively important variables. Although the variables of the model passed the significance test, it is also necessary to ensure that the whole model is significant, so the Hosmer-Lemeshow test is also performed on this model, which here gives a p-value of 0.59, indicating that this model does not have an obvious lack of fit. Next, we did model evaluation, For the training set, the accuracy is 0.84 and the AUC is 0.83. For the test set, the accuracy is 0.81 and the AUC is 0.65. It is a good model because the accuracy and AUC do not have big difference between the training and test sets. But the specificities for two sets are as low as 0.37, so I tried to find the optimal threshold point 0.77 that maximizes the specificity (TN rate) and sensitivity (TP rate). However, when I applied this threshold to the validation dataset, the result was not very good. The model was accurate in predicting customers not in default, but very poor in predicting customer default on the loan. I think the possible reason is that the model is very sensitive to unbalanced data

**6.** **Conclusion/ Limitations and future work**

My final model is logit () = -2.48 + 0.58x1 + 0.33x2 + 4.1x3 – 0.32x4 – 0.64x5 – 1.15x6 (X1 = {1= Man, 0=Otherwise}; X2 = CoappIicatincome; X3 = {1= Credit History,0=Otherwise} X4 = LoanAmount; X5 = {1= Urban, 0=Otherwise}; X6 = Rural). Credit history is the most important in determining if the loan will be approved, and the estimates odds of loan will be approved for those with credit history is e4.1 = 66.7 times the odds for those without credit history, holding all other variables constant. However, since we are using unbalance data to predict the outcomes of our data, this accuracy level in my model is still somewhat misleading and will tend to be optimistic. To avoid this, with a sufficient amount of data, we can also consider using sampling techniques like up or down- sampling to deal with imbalance dataset, or try some different algorithm such as decision tree.