MAT-8416
Dr. B
Hai Du
May 6, 2020

MAT 8414 - 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.

1. 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 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.

2. 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 then compared these models together. 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 used multiple imputations to estimate values that reflect the uncertainty around the true value, by inspecting the distribution of the original and imputed datasets. The density of the imputed datasets has a similar distribution to the original as expected. Second, 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

To the best of our knowledge, logistic regression is one of the good performances in binary classification. First, I tried modeling with logistic regression with the purposeful selection base on the multiple imputation dataset. My final model predictors included Married, Credit_History, LoanAmount, CreditHistory*LoanAmount. However, there are some possible limitations for multiple imputation. Recent studies have found that rounding off imputations for dummy variables can sometimes lead to bias in parameter estimation. Sometimes, our model will generate inconsistent results if I run the multiple imputation more than once. That is because multiple imputations have the property of being unbiased. Since I am not able to select the best dataset to analyze the potential issue. I also decided to run automatic backward selection multiple times for different complete datasets and selected the variable that appears most frequently. I found that the smallest AIC will always Include Married, Credit_History and Property_Area, then CoapplicantIncome, LoanAmount, and Education also appear frequently.

In order to avoid the inconsistent issue, I used the second method to replace the missing values and outliers, then created a relatively stable dataset. Similarly, through purpose selection, our final model included Married, Education, Credit_History, Rural, Urban, LoanAmount, and CoapplicantIncome. This model included every predictor in the first model. For this model, the AUC score is 0.81 which was slightly higher than that of model one, and its accuracy is 0.82. Our binary classification model predicted the outcome with 82% accuracy which is considered as good. Sensitivity is 0.98, which refers to the part of a loan that the model predicts is positive and not in default, and specificity 0.45 measures negative samples that actually default on the loan. Since we are given an imbalanced dataset, accuracy tends to emphasize majority class, which may mislead the true performance of the model. Therefore, in order to balance both sensitivity and specificity, I used G_Mean = $\sqrt{(sensitivity * specificity)}$ to avoid bias towards the majority class. In my final model, the G_mean is about 0.67, which is still not considered as good as I think. Therefore, I conducted the Hosmer-Lemeshow goodness of fit test for our logistic regression model, which gives a p-value of 0.06623, indicating that our model does not have an obvious lack of fit.

Finally, in order to optimize prediction, I tried lasso regression with these final predictors, all of the predictors are kept in the model, this proves to a certain extent that variables have significant contribution to our final model.

6. Conclusion/ Limitations and future work

My final model is logit ($\hat{\pi}(x)$) = $-2.47 + 0.51x_1 + 0.5x_2 + 4.2x_3 – 0.95x_4 – 0.81x_5 - 0.005x_6 + 0.00016x_7$ (X1 = {1= Married, 0=Otherwise}; X2 = {1= NotGraduate, 0=Otherwise}; X3 = {1= Credit History,0=Otherwise} X4 = {1= Rural, 0=Otherwise}; X5 = {1= Urban, 0=Otherwise}; X6 = Loan Amount; x7 =CoappIicatincome). 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 $e^{4.2}$ = 66.7 times the odds for those without credit history, holding all other variables constant. However, since we are using our 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 partition the data into a training set and a validation set and also consider using sampling techniques like up or down- sampling to deal with imbalance dataset. Second, there is a limitation for purposeful selection, because purposeful selection is not considered to force all dummy variables into the model. Therefore, I may consider using other selections to formulate the final model.