Capstone project: Lending Club Loan Status

Jiangong Li

**Project Outline:**

The historical loan data is issued by Lending Club. The dataset has over 100 features, such as 'addr\_state', 'earliest\_cr\_line', 'loan\_amnt', 'loan\_status', 'mort\_acc'. The responses of the model is the loan status which is categorized to “bad\_loans” and “good-loans”. The goal is to build a model to predict the chance of default for a loan based on the feature information provided by applicants.

This project was the course project (STAT 542: Statistic learning) that I have taken in Spring 2019. We were supposed to use three models to classify the loan status and to pick up the best model with justice.

**Data source:**

The data of the Lending Club are list here:

* <https://www.kaggle.com/wordsforthewise/lending-club>.

Training: Data from 2017 to 2018

Testing: Data in 2018Q

**Methods:**

1. *Exploratory data analysis:*

Some features have too many NA values, and some features have too many categorical levels. Some columns contain duplicated information. Some raw feature has the meaning but is naming inconsistent. Based on the results of exploratory data analysis, I decided 17 numerical variables and 7 adjusted categorical variables kept for model development.

1. *Data pre-processing and organization*

The imputation of the testing data is to replace the missing values of numeric feature to be the median, and replace the missing values of categorical feature to be the highest frequency. Then, the selected data matrix of was processed to sparse matrix (127 features) where all categorical variables were adjusted to binary.

1. *Model development and comparison*

We tried logistic regression with lasso, Naïve Bayes, and Xgboost. The results were showing in Table 1. We use log-loss to check the model performance. Naïve Bayes method requires less time but has worst performance. Xgboost method requires many parameters. Tuning parameters is time consuming. Logistic regression method with lasso (10-fold cross-validation) have good performance and require less parameter studies. We decide to use logistic regression. As shown in Figure 1, the lambda selected Logistic regression method with lasso is the minimum lambda. As shown in Figure 2, the AOC of our model is 0.72.

**Table 1. Time spend of model training and performance of model testing.**

|  |  |  |
| --- | --- | --- |
|  | Time spend | Log loss |
| Logistic regression | 55.5min | 0.4484 |
| NaiveBayes | 10min | 3.57 |
| Xgboost | 1min | 0.55 |

**A screenshot of a cell phone

Description automatically generated**

**Figure 1. lambda selection for Logistic regression method with lasso.**

**A screenshot of a cell phone

Description automatically generated**

**Figure 2. ROC curve for Logistic regression method with lasso (AUC=0.72).**