Lending Club Loan Status Analysis

Analyst: Eugene Wen

Date: Feb. 28, 2018

Overview

- Background and Business Problems
- Data Description
- Data Wrangling and Manipulation
- Model Choice
- Results Summary
- Conclusion

Lending Club

- Lending Club (NYSE: LC) is the world's largest peer-to-peer lending platform. The reported revenue in 2016 is US$ 501 million.
- Lending Club enables borrowers to apply for loans between US$1,000 and US$40,000.
- Investors can search the loan listings on Lending Club platform to select loans they intend to invest.
- The investment profit gained by investors is the loan interests.

Business Problem

- The Risk Analysis Team would like to build a predictive model for loan default.
- Any late payment more than 30 days would be considered as risky → likely to default in the future.
- By constantly monitoring the current loan status and variables changing over time, Lending Club would have a better control on the loan default risk.

Data Description

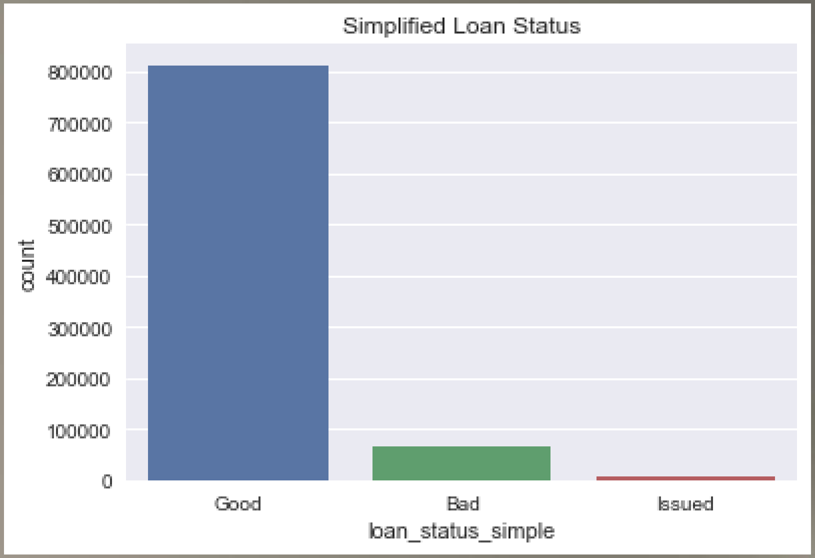
- The dataset was obtained from Kaggle, which is a combined dataset of 2007 – 2015 loans issued downloaded from Lending Club website.
- The dataset includes the current loan status (Current, Late, Fully Paid, etc.) and latest payment information as well as borrowers' past credit history.
- Totally there are more than 800,000 rows and 74 columns.

Data Wrangling

- The data integrity was first examined and some discrepancy was found.
  - Removed variables without descriptions.
- All variables having < 5% information were dropped.
- Irrelevant and (near) zero variance variables were dropped from dataset.
  - E.g. id, member_id and policy_code.
- Some variables have incorrect data types and were converted.
  - E.g. date strings were converted to datetime type.

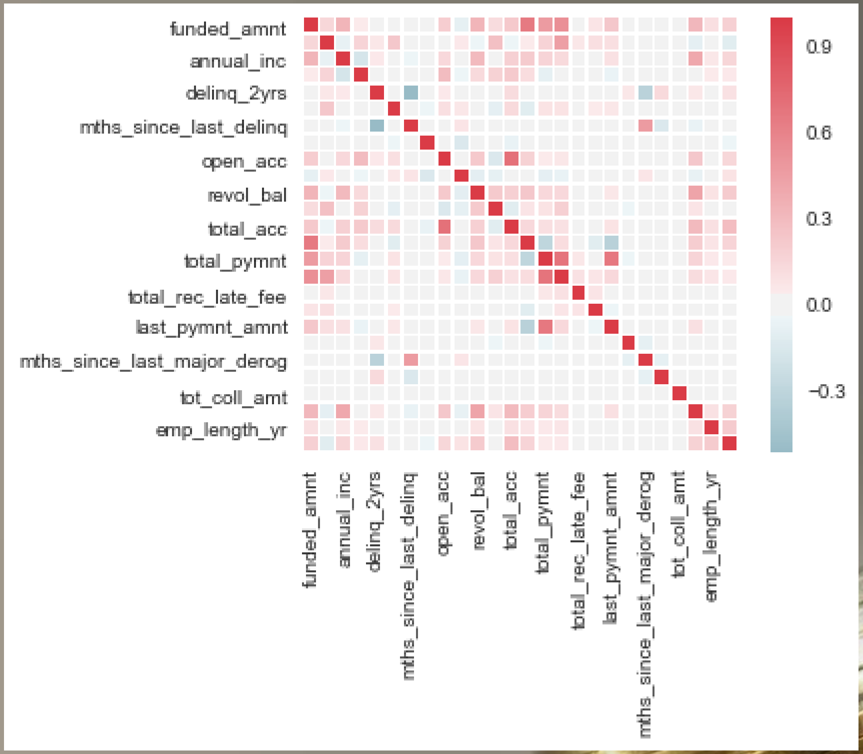
Recreate Target Variable

- The loan status has quite detailed categories.
- A new variable named "loan_status_simple" was created including only three categories:
  - issued (same as before)
  - good (no late payment)
  - bad (has late payment or default)



Processing Numerical Variables

- Two variables dti and total_rev_hi_lim show maximum values are 9999.0 and 9999999.0. They were capped using the maximum non-filler values.
- The missing values in each numerical variable was filled using column mean.
- Any correlated pairs was processed by dropping one of them to avoid high correlation in predictors.



Processing Categorical Variables

- Each categorical variable's frequency table was examined. All risk indicators and date variables were dropped from analysis.
- Missing values in each categorical variable were filled using "MISSING“, which was dropped when doing dummy coding.
- A few variables were engineered:
  - **emp_length**, by extracting the year value, recode "< 1 year" as 0 and "10+ year" as 10, then fill the n/a as the mean.
  - **home_ownership**, by consolidating ANY, NONE and OTHER as OTHER to reduce the category number.
  - **earliest_cr_line**, by combining this variable and issue_d to calculate length of credit history in years (cr_hist_yr), calculated as cr_hist_yr = issue_d.year() - earliest_cr_line.year().
- The missing values occurred in the two newly engineered numerical variables were filled using column means.

Training / Test Set Preparation

- The target variable is not balanced!
- All numerical variables were standardized, and all categorical variables were dummy coded.
- Then separated the dataset by loan status into three datasets.
  - The dataset with status = Issued was removed from the modeling.
  - The dataset with status = Good was undersampled to 10% of original size which yielded 81149 rows.
  - The undersampled dataset was appended to the dataset with status = Bad to form the train_test_set.
  - The train_test_set was then splitted into X_train, y_train, X_test and y_test for further modeling. The split proportion used was 30%.

Feature Selection

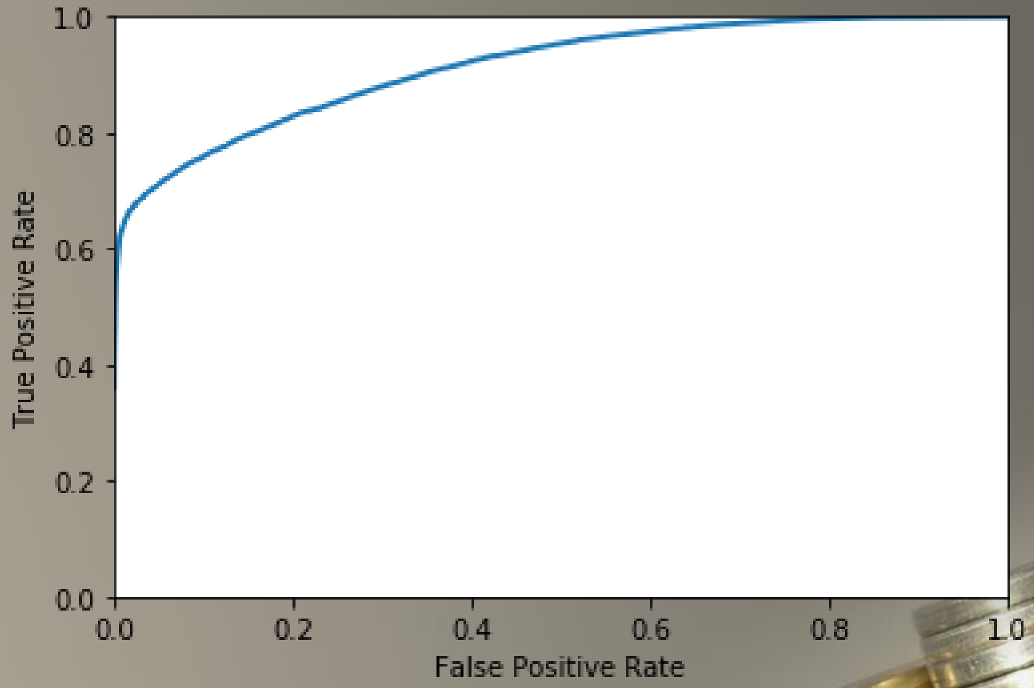
- Extra Trees Classifier was used to build the first model with all features.
- Based on feature importance and some experiments of cutoff values → cutoff = 0.05 for optimal subset.
  - No impact on the predicting accuracy
- The selected "optimal feature list" was then used for further model tuning using Grid Search and performance evaluation.

Model Tuning and Selection

- Model candidates (from experience and literature)
  - Logistic regression
  - Random forest classifier
  - Support vector machine classifier
  - Naïve Bayesian classifier
- 10-fold cross-validation mean accuracy as the performance metric.
- Best model: Random Forest (with parameter tuning)
  - Reported 10-fold CV accuracy: 84.1%

Prediction Result

- Test set prediction
  - Accuracy: 84.2%
  - Precision: 94.0%
  - Recall: 69.6%
  - AUC: 0.91



Conclusion

- Through this loan default risk analysis, we were able to use only a few indicative predictors to train a random forest classifier with optimized hyperparameters.
- The prediction results on test set reaches overall accuracy of 84%.
- For bad loan prediction, we could reach to 94% in precision.