Summary - Project McNulty

Loan default prediction using Lending Club data

Jiamin Han

Project Design

In this project, I aimed to train a classification model to predict bad loans on a major peer-to-peer (P2P) lending platform, Lending Club. In a typical P2P lending, borrowers submit their loan applications to Lending Club, then individual lenders can directly browse and select loan applications that they want to fund. Eventually, borrowers pay interests and principals back to lenders. P2P lending is supposed to simplify the personal loan business by connecting investors and borrowers directly, thus bring down the cost of personal loans as compared to borrowing from a traditional financial institution. However, investors always run the risk of investing in a bad loan. In fact, the default rate of P2P loans is much higher than that of traditional loans. Therefore, it is important for the P2P lending industry to improve risk management by providing investors with comprehensive risk assessment in decision making. To address this issue, I will develop a predictive model to identify bad loans by using information available on loan applications. In that way, investors can make more objective and data-driven assessment of loan applications to minimize risk.

Methods

I started with exploratory data analysis (EDA), in which I examined data structure, type, and distributions. Based on correlation analyses, I filtered out features that did not have a strong impact on the outcome and features that could not be available before a loan is issued. A total of 28 features were included for modeling purpose. Then I cleaned the selected features by converting their data types into appropriate format, and removed extreme values in annual income that were higher than mean plus 3 standard deviations of income ($550,000). In addition, I recoded and created a number of features to be used in modeling. The recoded features include length of employment, home ownership, and loan purpose. The newly created features include average FICO score, length of credit history, current employment status, and region of residence in the US. The outcome variable, final loan status, was coded as 1 if the loan was charged off, payed off late, or default. For model training, I fitted 7 classification models on my training data (60% of sample), followed by cross validation and grid search to select the best model parameters. Then I validated my models on the test data (40% of sample), and selected the best-performing classification model by comparing their AUC and F1 scores. To examine the influence of imbalanced distribution of outcome on model performance, a secondary analysis was performed in a under-sampled dataset with equal number of good and bad loans.

Results

The performance metrics of the 7 models are shown in Figure 1. Gradient boosting (GBT), support vector machine (SVM) and random forest (RF) yielded the highest AUC scores, which were between 0.72 to 0.73. By zooming in the ROC curves, I found the three models performed similarly to each other, although GBT generated a higher F1 score compared to SVM and RF. However, the F1 scores were still low because of the imbalanced data. In the secondary analysis, I created a subset dataset of approximately 550,000 loans with balanced distribution of good and bad loans. The models generated similar AUC scores in the balanced dataset, but much higher F1 scores (Figure 2). Compared to the results of the full dataset, the imbalance in outcome distribution had a big impact on prediction accuracy as measured by F1 score.

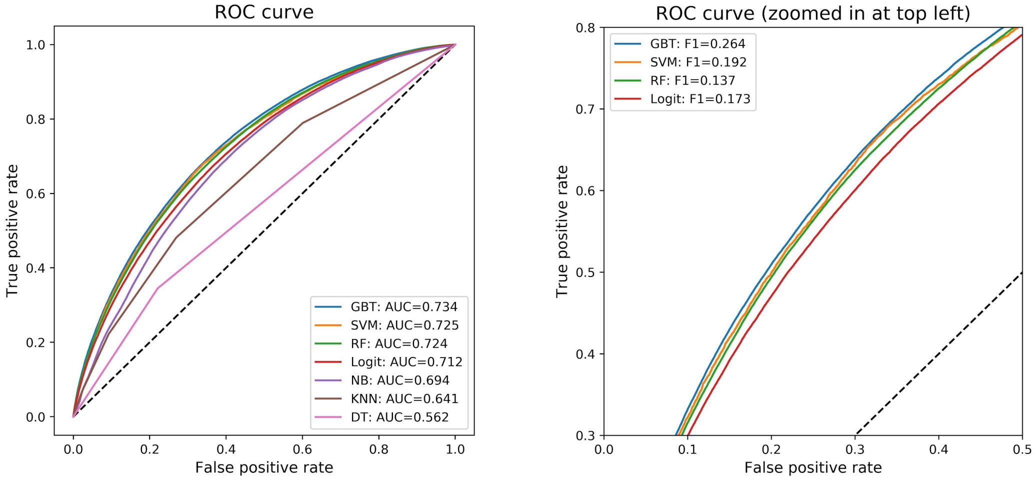

Figure 1. Performance of 7 classification models based on the full dataset.

By examining the performance of individual features in the GBT model, the top 10 features with the highest impact on loan outcomes are listed in Table 1. The high-impact features were mostly related to applicant’s credit health and income.

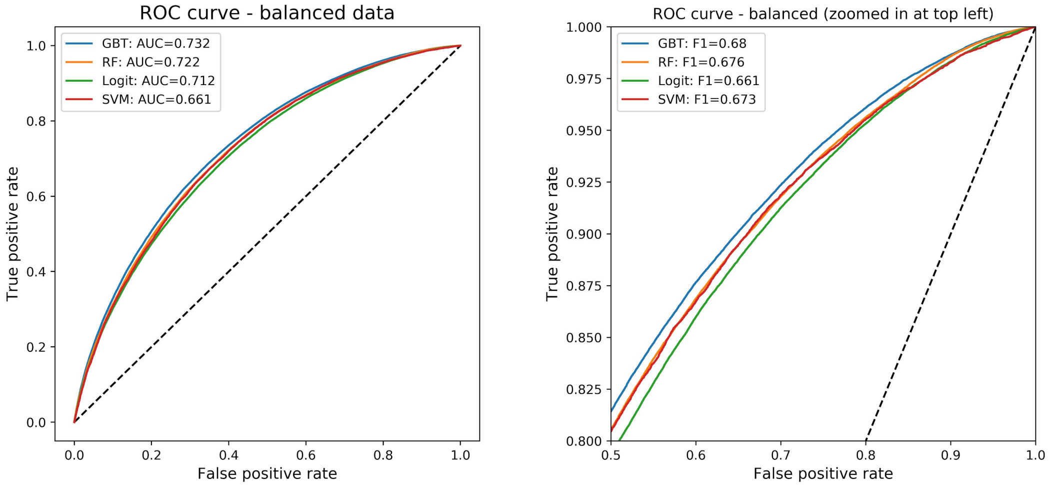

Figure 2. Molde performance based on balanced data.

Table 1. Feature importance in predicting defaulted loans

Feature	**Importance**
Average FICO score	0.092
Charge off times within 12 months	0.074
Total credit revolving balance	0.064
Debt to income ratio	0.061
Total number of credit lines	0.058
Interest rate	0.057
Loan description length	0.056
Times of delinquency in 2 years	0.056
Number of derogatory records	0.055
Loan application year	0.044

Conclusions

Based on data from Lending Club, I developed a predictive model of loan status using GBT model, which could efficiently identify default loans by leveraging information available at the time of loan application. Based on my results, applicant's credit health and income were most important factors to consider when evaluating risk of default. However, the accuracy of prediction was hampered by the imbalanced distribution of good and bad loans. Nevertheless, my project demonstrates great potential of applying machine learning in risk prediction for P2P lending. By integrating the predictive modeling on their investment shopping interface, Lending Club could easily flag loans at high risk of default and can adjust interests rate to offset the risk of default.

Tools

- Jupyter notebook, Pandas, Numpy: Ingest, organize, and process data
- Matplotlib, Seaborn: Data visualization
- Scikit-learn, statsmodels: Fit regression model

Data

Data were downloaded from Lending Club website (https://www.lendingclub.com/info/download-data.action), which contains 1,059,979 complete loans issued through 2007-2018. The dataset comes with 145 features, of which 28 features that are available at the time of loan application and correlated with loan status were selected. These features included loan information, application type and borrower's financial and demographic information. My dependent variable is the final status of the loan, which is defined as good if the loan is paid off on time or bad if the loan is charged off, paid off late, or default. My analytical goal is to fit a classification model that best predicts the loan status (good or bad).

Models

- K nearest neighbors (KNN)
- Logistic Regression
- Naive Bayes
- Support vector machine (SVM)
- Decision tree
- Random forest
- Gradient boosting

What would I do next time?

Some information in the dataset have not been used in my model but worth further examination. For example, loan description and job title are written as free text, which can potentially provide sights into the motivation and socioeconomic status of the loan applicant. In my future work, I can extract relevant features from these texts using natural language processing (NLP).

Second, many features were correlated with each other. Therefore, it is ideal to apply dimension reduction technique, such as principal component analysis, to simplify the feature matrix and to reduce computational complexity.