Capstone Proposal

Weiwei Liu
February 6, 2018

Domain Background

Lending Club is a peer to peer online lending platform, which publishes the data of the loans it underwrites, including the status (current, fully paid, charged-off, etc). The Lending Club dataset is well-known in the machine learning community, evidenced from the abundant lending club data analysis found through Google search (referenced materials in the end). The availability of hundreds of thousands of loan records and numerous loan characteristics make it promising to apply machine learning to predict bad loans from a pool of loans.

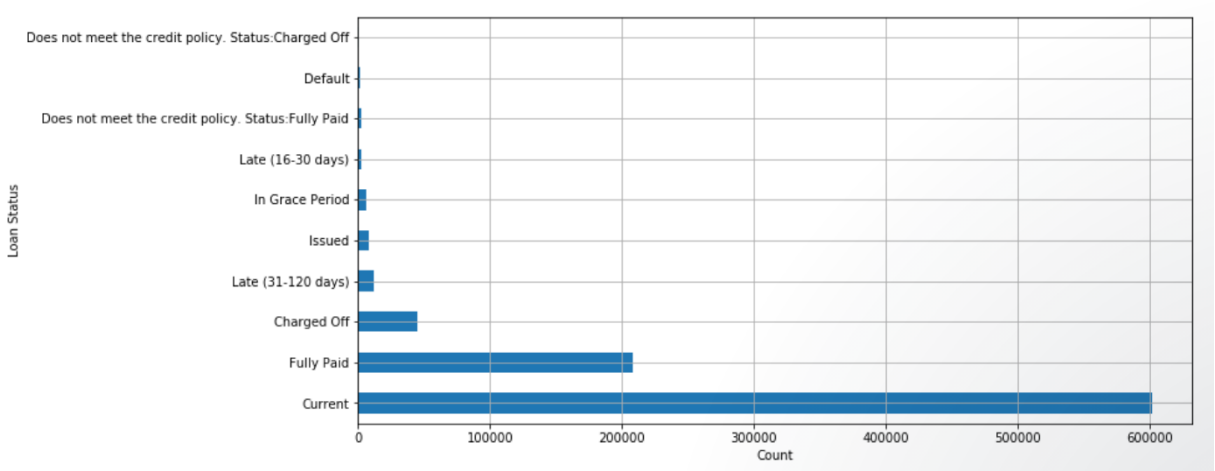
I work at a consumer credit lending company – it will be an interesting exercise to explore the feasibility of using machine learning to make credit decisions. Also, as an individual investor to Lending Club, the machine learning algorithm would help me select and invest in the loans with low likelihood of default.

Problem Statement

Given the Lending Club data, with loan status as the label, I plan to develop supervised-learning algorithms (binary classification) to predict if a loan is 'good' or 'bad' given the characteristics of the loan and its borrower. 'Good' is defined as loans that are current, in grace period, early delinquent and fully paid; 'bad' is defined as loans that are late delinquent, in default and charged-off.

Datasets and Inputs

I will use the lending club loan data provided by Kaggle (https://www.kaggle.com/wendykan/lending-club-loan-data/data). The dataset loan.csv includes ~890K loans underwritten by Lending Club from 2007 to 2015, with data dictionary provided in LCDataDictionary.xlsx. The loan dataset includes the current loan status (Current, Fully Paid, Charged-Off, etc), which will be transformed into either 'good' or 'bad' label defined above. The distribution of loan status is imbalanced in that 93.37% are good loans, which affects the choice of evaluation metric.



Source: https://www.kaggle.com/wsogata/good-or-bad-loan-draft

The dataset also includes borrower and loan characteristics (samples listed below), reflecting borrowers' ability and willingness to pay. A subset of these variables can be used as features to predict the loan status.

- Loan characteristics: loan amount, term, grade, interest rate, purpose
- Borrower characteristics: debt to income ratio, delinquency in last 2 years, inquiries in last six months

Lending Club (https://www.lendingclub.com/info/download-data.action) also provides loan data booked after 2015, which can be used for out of time / out of sample testing.

Solution Statement

Given the labeled data, supervised learning can be used to predict the loan status ('good' or 'bad'). For bad loans, we assign bad_loan = 1, and for good loans, we assign bad_loan = 0. As the output is binary, I will explore several classification algorithms, which include random forest, gradient boosting, etc., and select the model with the best performance (to be discussed in Evaluation Metrics section) as my underwriting engine (the primary model).

Benchmark Model

The benchmark model performance will be compared with the primary model using the same set of evaluation metrics. For benchmark, I will use a simple model, such as Naïve Bayes, to get a baseline score. The primary model should perform better than the benchmark.

Evaluation Metrics

As the goal of the model is to correctly predict if a loan will stay good or go bad, the model should try to maximize the F1 score, which is a suitable metric for a binary classification with imbalanced classes. Somers' D will also be calculated to assess the discriminatory power of the algorithm.

Project Design

Given the purpose, dataset and methodology discussed above, below is the project outline.

1. Explore the datasets through visualization to understand the distribution of potential features and outcome ('good' or 'bad' loan), as well as the relationship between features and outcome.
2. Select and transform features in preparation of training classifiers. This step includes removing obviously irrelevant or hard to use data fields (e.g. member_id, employer_title). Potential sub-steps include:
    a. For data fields with significant missing values, either considering dropping the field or giving proper treatment for missing value.
    b. Categorical variables will be transformed using one hot encoding.
    c. Depending the classifiers to use, numerical features may need to be rescaled.
3. Develop classification algorithms with the train set, and evaluate model performance against test set using F1 score and Somers' D.
    a. Develop a simple model (e.g. Naïve Bayes) as the benchmark model to for a baseline performance score.
    b. Potential classifiers for primary model include Random Forest, Gradient Boosting (xgboost), etc.
    c. Compare the model outcome and select an algorithm that can be used as the primary model.
    d. Use grid search to fine tune the model parameters.
    e. Interpret the model from the business (consumer loan underwriting) context.

References

https://www.kaggle.com/wsogata/good-or-bad-loan-draft

http://cs229.stanford.edu/proj2015/199_report.pdf

https://nycdatascience.com/blog/student-works/project-1-analysis-of-lending-clubs-data/