The Lending Club Corporation: Predicting and Profiling Customers Who Will Miss a Payment

Sinem Buber Singh, Liang Zhao, and Dan Silitonga

The Graduate Center

sbuber@gc.cuny.edu, ch00226855@gmail.com, dannysilitonga@gmail.com

April 1, 2015

Introduction to Peer to Peer Lending and the Data

- The Lending Club Corporation works as a third party intermediary that bring together investors and borrowers. Borrowers borrow money, primarily to re-finance their high interest credit card debts with a more favorable terms. Investors lend money to gain an attractive rate of return of their investments, assuming that the borrowers do not default on the loans.
- The entire dataset is about 1Gb.
- Data clean up is required. More on this later.

Goals

There are 2 different goals as follows:

- To build a classifier to predict those who will be late in their payments. This is a classification problem, where Logistic Regression and Support Vector Machines can be useful. Alternatively, an assemble model combining results from different models through, for example, the Principal Component Analysis may lead to more accurate classifications.

- To create a classifier to estimate interest rates obtained by Lending Club customers. This is more a traditional regression problem, where Linear Regression, LASSO, Ridge, and the Elastic Net can be helpful.

One research question is whether there is a relationship between the customers who miss a payment and those who are charged a very high interest rates?

Data Clean Up Highlights

This dataset contains symbols, words, ranges that require clean up.

- The column interest rate has a percent sign symbol. This would create issue for plotting and building a regression model. The percent symbol needs to be stripped and the variable converted into a float.
- The column loan length has the word month in each record. The same procedure is required.
- The FICO score column includes a range and should be converted into 2 columns of low and high scores.

Data Exploration Highlights

- The loan amount does not seem normally distributed. The center of the distribution is skewed, with a large tail on the right side of the distribution.
- The amount requested also does not seem normally distributed. One difference between amount requested and loan amount is that some loans requested by the borrowers are rejected by the investors. Consequently, the loan amount is set to 0.
- There is a linear relationship between interest rates and FICO Score and Monthly Income

Linear Model

There is a linear relationship between interest rates and FICO Score and Monthly Income, with an R square of 65 percent.