Capstone Project Proposal – Intermediate Data Science
David Ermer

I propose to analyze the LendingClub (LC) data available on Kaggle.com[1] that includes data from 887379 loans made by LC.

There loans are sorted by LC into 7 grades and 35 sub-grades and approximatel 5% of the loans have been charged off or are in default.

The first step will be to evaluate LC's grading system by determining the relationship between the sub-grade assigned and the percent of loans that are bad, i.e. in default or charged off.

It should also be informative to examine the statistical relevance of the various factors reported in the data. For example *Debt Income Ratio*, *Number of Open Credit Accounts*, *Earliest Credit Obtained* (month and year) are some of the factors included in the data. Calculations of the distribution of these factors as a function of grade should reveal how relevant the factor is to the performance of the loan.

Finally a model that predicts the performance of loans will be developed. The provided data suffers from selection bias, i.e. we only know the data for loans that were issued. This means that we are trying to do a second selection process using data with a limited (by the loan selection process) distribution. This needs to be kept in mind while analyzing the data and the results of the modeling process.

1 https://www.kaggle.com/wendykan/lending-club-loan-data/downloads/lending-club-loan-data.zip