GA Final Project Part 1: Proposal

Predicting Loan Defaults

Problem Statement:  **I will build a model to predict which loans will be paid back in full vs. loans that will default based on historical loan data. The purpose of this analysis is to help investors avoid picking riskier loans and thus maximize their returns. I will use loan data from Lending Club, an online peer-to-peer lending platform where borrowers can obtain loans and investors can purchase notes backed by payments based on loans. The dataset provided contains data for loans issued through 2007–2015. I will use the loan and borrower characteristics and corresponding loan outcomes/status included in this dataset to train a loan default prediction model. I intend to use a combination of (supervised) classification techniques to predict defaults and inform a superior loan picking strategy.**

**Data:** The Lending Club files provided contain complete loan data including the current loan status (current, late, fully paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, address including zip codes and state, and collections among others. The file is a matrix of about 890,000 observations and 75 variables.

The loan characteristics provided in this dataset can largely be divided into two groups: features of the loan, and features of the borrower. The loan features include loan amount, the interest rate, and the term of the loan. The borrower features include employment length, credit history, and income.

**Hypothesis:** I expect that borrower risk is positively correlated with likelihood of default. Specifically, I expect that (lower) borrower credit or FICO score, (higher) delinquency rate and (higher) debt-to-income (“DTI”) ratio are most predictive of default. I expect that loan characteristics that are more predictive of default will include loan amount (higher loan amounts will more likely lead to default), (lower) loan grade, and (higher) interest rate.

**Data challenges:** The Lending Club loan dataset contains a substantial number of variables (75), so I will need to develop a robust method of selecting a subset of those variables to predict default. Some of the specific loan data challenges include the following:

* Mix of categorical (e.g. borrower state) and numerical variables that need to be treated consistently
* Some irrelevant variables (e.g. member id, url)
* Potentially several variables with poorly documented data
* Potential collinearity between sets of variables (e.g. borrower risk characteristics)
* Potentially several variables with substantial portion of missing values
* Potential outliers for some data fields that need to be trimmed

Another main challenge is the potentially very low number of historical defaults vs. non-defaulted loans in this dataset, which would create an imbalanced sample. This would render the default prediction accuracy rate from a model trained on this dataset meaningless.

**Success metric:** As described above, the default prediction accuracy rate is likely not the best success metric with potentially imbalanced data (namely, very low number of defaults vs. non-defaulted loans). I instead propose gauging the success of my prediction models based on their sensitivity (i.e. the fraction of loans that are actually non-default that were predicted as non-default) and specificity (the fraction of loans that are actually default that were predicted as default). I will aim for a sensitivity and specificity score of at least 90%.