Springboard Data Science Career Track
Capstone Project #1

Proposal

Problem

Credit is essential to our economy. It is utilized by individuals and businesses alike to drive consumer spending and capital investment. Regardless of the loan type, borrower profile, or lender size, all credit decisions boil down to the core problem of managing default risk. Excessive loan defaults can bankrupt a lender, or in the case of the 2008 mortgage crisis, bring the global economy to the brink of financial collapse.

This proposal focuses on a specific type of loan - peer-to-peer consumer loans. These loans are typically used for debt consolidation, home improvement, and major purchases. Since peer-to-peer consumer loans are often not collateralized, the risk to the lender can be higher.

Client

The client is a hypothetical peer-to-peer lending company. Peer-to-peer lending companies act as intermediaries between individual borrowers on one side and individual and institutional investors on the other side. Appropriately managing default risk is essential to building loan portfolios with attractive returns for investors.

Based on this analysis, the client will refine its process for determining which borrowers to lend money to. Additionally, with a better idea of each loan's default risk, the client can construct its loan portfolios to better manage returns for its investors.

Data

This project will utilize loan data from Lending Club, which will be downloaded as CSV files from their website here. This dataset includes complete loan data for all loans issued from 2007 through 2018. Both loan and borrower attributes are included in this dataset. The former includes loan amount, term, interest rate, payment amount, status, purpose, etc. The latter includes employment, income, zip code, credit history length, etc.

Approach

This project requires supervised learning of a classification problem and aims to predict the loan status. There are multiple loan statuses in this dataset, but for this analysis, I will summarize them into two classifications: (1) In Payment, which includes the *current*, *in grace period*, and *fully paid* statuses in the dataset; and (2) Delinquent, which includes the *late (16-30 days)*, *late (31-120 days)*, *default*, and *charged off* statuses in the dataset.

I will test multiple variables in the dataset as predictors, such as employment length, annual income, DTI (monthly debt payments as a percentage of monthly income), credit history length, utilization of revolving credit lines, delinquencies, and derogatory credit remarks/bankruptcies, among others.

Additionally, I will test the relationship between different variables as predictors, such as the loan amount relative to annual income, DTI and credit utilization relative to credit history length, revolving credit lines relative to annual income and credit history length. I will also cross-reference US Census data to analyze the borrower's annual income relative to the average for his/her zip code to obtain a proxy of how well the borrower lives within his or her means.

Deliverables

The deliverables for this project will be a slide deck with an executive summary, followed by detailed visualizations and analysis, as well as a paper with narrative discussion and the underlying code.