Prosper Loans Data Exploration

Dataset

This Data Exploration document explores a dataset from Prosper Loan. The dataset contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, and many others. The dataset can be found in the repository [here](https://www.google.com/url?q=https://s3.amazonaws.com/udacity-hosted-downloads/ud651/prosperLoanData.csv&sa=D&ust=1592069459914000), with feature documentation available [here](https://www.google.com/url?q=https://docs.google.com/spreadsheet/ccc?key%3D0AllIqIyvWZdadDd5NTlqZ1pBMHlsUjdrOTZHaVBuSlE%26usp%3Dsharing&sa=D&ust=1592069459915000).

Summary of Findings

In the exploration, I found that there was a strong relationship between the Loan Status Outcome and Features of the Loan and the Borrower. For Features of the Loan, I identified the five features: Term, BorrowerAPR, LoanOriginalAmount, MonthlyLoanPayment, ListingCategory (numeric). For Feature of the Borrower, I identified the eight features in five categories:

1. Earnings-related Features: EmploymentStatus, EmploymentStatusDuration, StatedMonthlyIncome
2. Debt-related Features: DebtToIncomeRatio
3. Credit-related Features: CurrentCreditLines, AvailableBankcardCredit
4. Asset-related Features: IsBorrowerHomeowner
5. Aggregated Features: ProsperScore

I developed a hypothesis for how each feature will affect the Loan Status Outcome, and attempted to find relationships to strengthen my hypothesis.

Using IMF’s definition of Non-Performing Loan, I classified the sub-categories Loan Status into two categories: Performing Loans and Non-Performing Loans.

The feature with the strongest relationship to Loan Status Outcome is BorrowerAPR. Other features MonthlyLoanPayment, EmploymentStatusDuration, StatedMonthlyIncome, CurrentCreditLines, AvailableBankcardCredit, ProsperScore and IsBorrowerHomeowner were also identified to have a moderately strong positive relationship with Loan Status.

Outside of the main variable of interest, I identified interesting distributions and interactions. The former will be log-normal distributions for both EmploymentStatusDuration and AvailableBankcardCredit. The latter will be interaction between MonthlyLoanPayment, BorrowerAPR and ProsperScore.

Key Insights for Presentation

For the presentation, I focus on the key Features of Loan and Borrower, and left out correlated features and intermediate plots. I start by introducing the Loan Status Outcome variable, the IMF definition, and its distribution.

Afterwards, I introduce the Feature of Loan and Borrower variables one by one. To start, I use the bar charts and histograms to illustrate the distributions. Next, I used violin plots and bar plots to illustrate the relationship between numerical variables and Loan Status Outcome (a categorical variable). I used clustered bar charts and point plots to illustrate the relationship between categorical variables and Loan Status Outcome. I also used a correlation matrix to illustrate the correlation between the numerical variables. I made use of color encodings as much as possible to differentiate between the variables.