Prosper Loans Data

Introduction

This project is on a data set from Prosper, which is America's first marketplace lending platform, with over $7 billion in funded loans. This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, borrower employment status, borrower credit history, and the latest payment information.

The main purpose of this project is to summarize the characteristics of variables that can affect the loan status and to get some ideas about the relationships among multiple variables using summary statistics and data visualizations. As per the specs of the project, there are multiple visualizations carried out with univariate, bivariate, multivariate plots

Univariate Analysis

Incase of univariate analysis, I chose to analyze

- The number of listings in each state
- Number of listings under each loan status type
- Income range distribution of listings.

Because all these variables frequency is to be analyzed, a bar chart is used. Bar chart is one of the best relevant options for univariate analysis. The insights produced by these plots are as below -

1. **The number of listings in each state** - The data says the highest listing counts in CA, TX, NY, FL and IL, which tracks with the rankings of top US states by population and the least in WY, ME, ND. In other words, the Prosper listings are distributed similarly to the population of the US.
2. **Number of listings under each loan status type -** The number of listings under each LoanStatus type are analyzed and the plot shows that 'Current' status has got the highest number of listings followed by 'Completed' and 'Chargedoff'. Other statuses include dues for various intervals and cancelled whose frequency is too less
3. **Income range distribution of listings. -** The income rate distribution is analyzed to know the average income rate of listings. The plot says that most

listings fall under 50000 dollars, followed by 50000-74999 dollars and so on. This also concludes that the more the income rate the less the listings count.

Bivariate Analysis

For bivariate analysis, two scenarios are considered and analyzed -

- Available bank card credit vs. the rate of interest
- Income range vs the amount borrowed from Prosper
- BorrowerRate pattern over years

For each of the mentioned scenario, different plots are used to depict the results. Let's see them in detail -

1. **Available bank card credit vs. the rate of interest -** In this case, a hist2d plot is used because most of the listings have same range of bank card credit, there needs to be an intense segregation to see the diffusion between them and the plot says the more the amount in their bank card, the less the loans they take and less the borrower rate in most cases.
2. **Income range vs the amount borrowed from Prosper -** In this scenario, a boxplot is used to depict the results. The insights from the boxplot are the medians are in increasing order i.e., the more the income range the more the principal amount borrowed from the Prosper except for the not employed or 0 dollar salaried.
3. **Borrower rate pattern over years -** Point plot is used to analyze this and it suites the analysis because this graph shows the nature of the variable from one xfactor to another. For this scenario, it shows that rate of interest increased from 2005-2006 and then fell off from 2006-2007 and then it constantly picked up for few years, reached its peak in 2011 which then started decreasing constantly.

Multivariate Analysis

To exhibit the multivariate analysis, one particular scenario is taken i.e., analysing how the rate of interest changed over years for Homeowners vs Non-Homeowners and to depict this scenario, a boxplot is used because it can be easily analyzed with the advantage of having only two cases in the hue column. This boxplot shows that the borrower rate increased for some years in case of non homeowners, however it constantly decreased around 2010. As regards of homeowners, there wasn't much difference till 2009 but after 2009, the graph went up till 2011 and started falling off from then.