Anlaysis of Prosper Loan Data

Introduction

The data being analyzed as part of this project is a loan dataset from Prosper. The dataset loan data with 81 variables for each loan, including loan amount, interest rate, loan status, borrower income, employment status, credit history, and the latest payment information.

The main goal of this project is to summarize the characteristics of variables that affect loan status and to gain an understanding of the relationships among multiple variables, using summary statistics, exploratory data visualization, and explanatory data visualization. Univariate, Bivariate, and Multivariate analysis was carried out on this dataset.

Univariate Exploration

Univariate exploration of this loan data includes the following analysis:

1. Number of listings by Loan Status

Most of the listings in this dataset are for current loans. Over 50,000 loans listed are still current and active. Approximately 40,000 listings show loans that are in a completed state.

2. Number of listings by Borrower State

The bar graph indicated that the highest listing counts were from the states California, Texas, and New York.These states are also the top US states by population. The least number of listings were from the states Wyoming, Maine, and North Dakota.

3. Analysis of Stated Monthly Income

From the histogram plot, it was observed that the monthly income of most borrowers peaks at around $5000, with a frequency of almost 10,000.

4. Breakdown of listings by Credit Grade

The maximum number of listings in this dataset have a credit grade C, followed by a credit grade D. This analysis only applies to loans taken out before 2009, since the credit grade value has been populated only during the pre-2009 period.

Bivariate Exploration

In this section, the following bivariate analysis was carried out:

1. Changes in Borrower Rate by year

The plot showed that the rate of interest increased from 2005-2006 and then fell from 2006-2007. It then steadily increased up for few years and reached its peak in 2011. It decreased constantly from 2011 onwards.

2. Relationship between available bank card credit and borrower rate

The 2D histogram plot indicates that the larger the amount in a borrower's bank card, the less loans they take and the borrower rate is also less in most cases.

3. Employment Status vs Loan Term

The bargraph plotted showed that most listings in this dataset are borrowers who are employed. Most borrowers who are employed full-time have a loan term of 36 months.

Multivariate Exploration

In the multivariate analysis of the dataset, the following dimensions have been analyzed:

1. Comparison of relationships between Loan Term, Borrower APR, and Loan Amount

Loan term doesn't seem to have significant impact on the relationship between APR and loan amount. Most listings in this dataset have a loan term of 36 months.

2. Comparison of relationships between Loan Term, Available Bank Card Credit, and Loan Amount

The loan amount and available bank card credit appear to have a slight postive correlation. It can be seen that the majority of the loans in this dataset have a term of 36 months.

3. Effect of homeowner status on Borrower Rate over the years

The boxplot showed that the borrower rate increased for some years for 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.

Key Insights

- Of the 81 variables related to the loan listings that were provided in this dataset, it was observed that employment status, available bank card credit, homeowner status, and credit grade were all significant factors in determining the borrower interest rate and APR.

- Most of the listings in this dataset included borrowers who were employed full-time, and a small portion of self-employed individuals. The most frequently occurring credit grade in this dataset was the grade C. The largest number of listings came from the most populated states in the US - California, New York, and Texas.

- There is a great deal of variation in the borrower rate from 2005 to 2014, and it is likely that the financial crisis of 2008 had a significant impact. It was also observed that individuals with greater bank card credit were more likely to have lower rate of interest.Most loans in this dataset have a loan status Current, and were still active. The most frequently occurring loan term was for 36 months.

- Further analysis on this data can be carried out by considering the Prosper Rating (Alpha) for data recorded in the post-2009 period, bank card utilization, and its impact on payments, APR, and interest rate.