LOAN DATA FROM PROSPER BY AJOKE YUSUF

Dataset Overview

The main dataset contains 113937 rows and 81 columns. After thoroughly cleaning my data, the dataset contains 83507 records and 18 features. The variables include:

- ListingNumber: The number that uniquely identifies the listing to the public as displayed on the website.
- Term: The length of the loan expressed in months.
- LoanStatus: The current status of the loan: Cancelled, Chargedoff, Completed, Current, Defaulted, FinalPaymentInProgress.
- ProsperRating: The Prosper Rating assigned at the time the listing was created between AA - HR. Applicable for loans originated after July 2009.
- BorrowerAPR: The Borrower's Annual Percentage Rate (APR) for the loan.
- BorrowerRate: The Borrower's interest rate for this loan.
- ProsperScore: A custom risk score built using historical Prosper data. The score ranges from 1-10, with 10 being the best, or lowest risk score. Applicable for loans originated after July 2009.
- Listing_category: The category of the listing that the borrower selected when posting their listing
- Occupation: The Occupation selected by the Borrower at the time they created the listing.
- EmploymentStatus: The employment status of the borrower at the time they posted the listing.
- EmploymentStatusDuration: The length in months of the employment status at the time the listing was created.
- IsBorrowerHomeowner: A Borrower will be classified as a homeowner if they have a mortgage on their credit profile or provide documentation confirming they are a homeowner.
- IncomeRange: The income range of the borrower at the time the listing was created.
- LoanOriginalAmount: The origination amount of the loan.
- BorrowerState: The two-letter abbreviation of the state of the address of the borrower at the time the Listing was created.
- StatedMonthlyIncome: The monthly income the borrower stated at the time the listing was created.
- LoanOriginationYear: The date the loan was originated
- LoanMonth: The Month the loan was originated.

DATA WRANGLING

- Imported the data and went through the metadata to understand the meaning of each features/variable.
- Subset the relevant columns needed for analysis
- Dropped null values
- Renamed ProsperRating (Alpha) column to ProsperRating and Listing Category (numeric) column Listing Category

- Corrected columns with incorrect data types.
- Slitted Loan Origination Date into Year and Month
- Dropped Loan Origination Date column
- Converted categorical variables to categorical data types
- Renamed all the values in Loan Origination Month

Summary of Finding

I carried out the three types of explorations namely: univariate, bivariate and multivariate explorations

In the **Univariate Exploration,** I used two visualization plots namely: histogram and countplot. In univariate, I am only interested in the distribution of individual variables of my major features of interest and predictor variable.

During my analysis, there exist to be a strong positive relationship between the borrowers APR and borrowersRate, they had the same plot pattern. The distribution of borrowersAPR deduce that the distribution is unimodal with the peak values around 0.35%. The distribution of loan amount explains that the distribution is Tri-modal having peak values around 5k,10k,15k ($), this clearly explains that most loan collected were around the ranges of the peak value

Most of the borrowers are employed, earn averagely and working in top three best occupation

In the **Bivariate Exploration**, I explored various visualization plots ranging across bar plot, box plot, heat map.

Conducted a visualization plot with BorrowerAPR and Year using a bar plot., it depicts that **2011** has the highest % of APR which is 0.25% and 2014 has the least APR which is about 0.17%. There is a gradual fall process from 2012 to 2014.Used a bar plot to depict the relationship between Loan Amount and prosper risk rating and my finding was that there is an inversely proportional relationship (i.e. higher the amount of loan, the lower the risk score attached). In the graph, it can be deduced that 14,000($) which is the highest has lowest risk score because the lower the risk of a borrower, the higher the chance of getting high amount of loan.
Used a box plot to depict the relationship between prosper score and prosper rating in and my finding was that there is an inversely proportional relationship (i.e the lower the prosper ratings and scores, the higher the borrower APR).

Using a correlation heatmap to show relationship between numeric category (Term, Borrower APR, BorrowerRate, Listing_Category, Employment Status Donation, Loan amount and StatedMonthlyIncome) and my finding was there is a strong positive correlation between BorrowerAPR and BorrowerRate with a coefficient value of 0.993 and also, there is a weak positive correlation between Term and Loan Original Amount with a coefficient value of 0.341

In the **Multivariate Exploration,** I used point plot for my visualization, the plot shows that there is a 95% confidence interval between low-risk rating (AA) and the loan amount, which deduce that borrowers who has a home has low risk rating and access to larger amount of loan

Key Insights for Presentation

My key insight focuses more on the main features of interest: BorrowerAPR & Loan Original Amount and the predictor variables. The distribution for the main features of interest: BorrowerAPR is unimodal with the peak values around 0.35% and Loan Original Amount is Trimodal having peak values around $5k, $10k, $15k. This clearly explains that most loan collected were around the ranges of the peak value.

Comparing my one of main features of interest variable with Loan Year, BorrowerAPR distribution shows that 2011 has the highest % of APR which is 0.25% and 2014 has the least APR which is about 0.17%. There is a gradual fall process from 2012 to 2014