**Prosper Loan Data Exploration**

**By Monika Pare**

## **Dataset:**

The dataset consisted of borrower APRs and attributes of 113,937 loans. The attributes included original loan amount, borrower's Prosper rating, loan term, borrower's stated monthly income, as well as many other features such as borrower's employment status, debt to income ratio, current loan status etc. The dataset can be found [here](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/prosperLoanData.csv), with feature documentation available [here](https://docs.google.com/spreadsheets/d/1gDyi_L4UvIrLTEC6Wri5nbaMmkGmLQBk-Yx3z0XDEtI/edit#gid=0)

## **Summary of Findings:**

I successfully analyzed This Prosper Dataset. My analysis was primarily focused at identifying variables affecting Borrower APR, and Prosper Rating and Loan Status. Before starting my analysis I have performed cleaning of the data set as there were lot of missing values in the data, I fixed those value either by replacing them with their median values or by simply deleting them .I have also renamed some of the columns like prosper Rating alpha , also dropped some of the column which are of no use for our further analysis. I also created new column out of existing ones like Listing creation date, Loan origination date.

After cleaning data, I performed univariate exploratory data analysis to observe how data is distributed. Since the dataset has 81 variables, I decided to choose some important ones for detail analysis. I narrowed down the predictive variables by excluding, all-unique, all-same, and more than 50% missing value, these three kinds.

**Univariate explanatory data analysis:**

For univariate analysis my variable of interest were Loan original Amount, Borrower Rate, Borrower APR, Occupation, Income Range, Employment status, Monthly Loan payment, Prosper score ,Borrower state ,Loan Original amount ,Lander Yield.

**Loan Original amount:** The distribution is expected to be right skewed, which is obvious from the histogram, other than couple of peaks. Moreover, we notice that it is quite rare for borrowers to ask for huge amount of loans through prosper.

**Borrower APR:** As expected the distribution is roughly normal, except couple of peaks on right side of mean.

**Occupation:** Apart from others and professional, Computer programmers and Administrative assistant were the occupation who have payoff their loan.

**Income Range** : Income range from 25,000−50,000 was the Highest income range for largest number of borrowers.

**Employment Status:** Full time employed and Employed were the highest in frequency.

**Borrower State:** California, Florida, Illinois are the top three states with highest number of borrower while South Dakota, Wisconsin, North Dakota are state with least number of borrowers.

**Loan Original amount:** Most of the people borrowed loan amount under $10,000.

**Lander Yield:** Lender yield distribution is skewed towards the right direction with highest peak between 0.1 and 0.2.

### **Some Important variables that I explored in Bivariate Data Exploration:**

**Borrower APR and its Predictors :**

**Prosper Rating Alpha and its Predictors**:

**ProsperRatingAlpha VS Inquries last 6 Months:** We observe following trends from the boxplot. First, on an average the borrowers who have made frequent inquiries to the bureau have lower prosper rating, hence are more risky. Second, we see significant outliers for every value of ProsperRating. But the variation for safe borrowers, i.e. ProsperRating=AA is lower than other grades. It is expected, since borrowers in excellent credit rating tend not to make several inquiries to the bureau. Next, we also observe skewness in the distribution for InquiriesLast6Months. The mean is significantly higher than median in most of the cases. This can be attributed to the presence of significant number of outliers corresponding to each ProsperRating.

**Prosper Rating and Estimated Loss:** From the plot we can say that high estimated loss is expected from risky borrowers while less estimated loss is expected from A and AA prosper rating holding.

While conducting bivariate analysis between ProsperRating and its predictors, few interesting facts were discovered. First, EstimatedLoss for a given borrower significantly affects ProsperRating. This is expected, since borrowers who have been prompt in repaying entire loan amount on time ought to have higher ProsperRating. Second, borrowers who have made frequent inquiries to the Bureau tend to be desperate for money, hence their ProsperRating is affected adversely. On the other hand, variables such as BankCardUtilization and DebtToIncomeRatio appear not to affect ProsperRating significantly. We can’t be so confident in this conclusion, because our results are based on this small dataset, hence, might not reflect the true figures.

**Loan Status and its predictors:**

1 **LoanOriginalAmount:** We can see from the above plot that borrowers who borrow small amount of LoanOriginalAmount are more likely to complete their loan on time as compare to the borrowers who borrow large amount .If we see the box plot median value of the Loan original Amount under the borrowers who have successfully completed their loan is 5000 which give us a idea that lesser the loan amount there is high probability to complete the loan on time.

2.**ProsperRatingAlpha:** Borrowers having prosper Rating A and AA are most likely to pay off their loans, and E and D prosper rating rating are most likely to be the defaulters.

3.**Employment status**: On observing the Employment status plot we can say that more than half of borrowers who are employed and full time employee pay off their loans but for other categories we don’t have sufficient data to come to any conclusion.

4.**Inquries Last 6 Months** :The plot shows that borrowers who made lot of inquiries paid of the loan as compared to the ones who made lesser or nearly no inquiries. On the other hand, the effect of InquiriesMadeWithinLast6Months and Inquries last 6 months on LoanStatus is not remarkable.