Loan Data From Prosper

Exploratory and Explanatory Analysis by Monika Pare

This notebook will document my efforts to investigate an interesting dataset from the Prosper peer-to-peer lending platform, and to then apply some machine learning classifiers. There are many questions one could ask of this data, but in this exercise, I will focus on one question specifically: whether one can successfully predict which loans will default. Prosper loans pay hefty interest rates to their creditors.

Other than loan status I will also investigate other variables like Borrower APR and Prosper rating which pay a significant role in predicting loan status.

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, and many others. However, we are not expected to explore all of the variables in the dataset. We will focus our exploration on about 10-15 of them

Questions:

- What Insights would be useful to a bank?
- What factors affect a loan's status?
- How the variable is distributed**?**
- What is impact of type of Occupation on loan Original Amount?
- How Loan status and Loan original amount related?
- How Borrower APR is correlated to prosper rating, number of inquiries, debt to income ratio, Open r evolving account.
- What is the impact of other variables on prosper rating or how prosper rating helps in deciding loan outcome Status?
- In which year or month Bank approved maximum total number of loans to the borrowers.
- From which state there were maximum number of borrowers?

Data Wrangling:

- Firstly, I visually and programmatically assessed the data. I found that there are lot of missing values and some variables have a lot of null values. Total Prosper Loans, for instance, and a lot of other variables associated with the Prosper history of the debtor.
- There were lot of variables with incorrect data types like loan origination
- year, listing creation year, listing closing year. I corrected them
- There were some variables (Employment Status Duration, Credit Score Range Lower, First Recorded Credit Line, Current Credit Lines,

TotalCreditLinespast7years) which were not needed for my analysis, so I simply dropped them.

- There were lot of null values in Debt to income ratio, I fixed them by filling them with new calculated debt to income ratio with the help of Monthly loan payment and stated Monthly income.
- I fixed Borrower APR null entries by substituting them with their median value.
- Also, there were column where there were few null entries, I fixed them by deleting.

Data analysis:

Firstly I explored each variable of my interest separately to observe hoe each and every variable is distributed. These are my reflections:

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 APR: Borrower APR was rightly skewed with highest peak from 0.1 to 0.3

Occupation: Computer Programmer, Executives and Teacher have the largest count other than professional and Others. But these two does not provide useful information.

Employment Status: highest number of borrowers counted were Employed followed by Full time employers while retired, Part-time, other and not employed have least count.

Monthly stated income: Monthly loan distribution was right skewed with highest number of borrowers having less than $400 income.

Term: most of the borrowers opted for 36 of loan term while minimal number of borrowers opted for 12 month of loan term.

Income Range Distribution: Number of borrowers having highest income range are reportedly high.Bank mostly approved the loan of the borrowers having income range from $29000 to $50000.

Prosper Rating: Highest number of borrowers belong to prosper rating C, and lowest number of borrowers belong to prosper rating AA.

BorrowerState: Highest number of borrowers reported were from the state CA(California),NY(NewYork) and Fl(Florida) and least number of borrowers were reported in ND(North Dakota).

Listing Creation Year: Highest Listing was created in year 2013 followed by year 2012.

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.
I also made a heat map to closely analyze how the variables are correlated. from there I observed from above plots open revolving accounts current credit line and Employment status duration are seems to have a positive relationship. Other than that Debt income ratio and loan original amount are significantly correlated, remaining variable does not show any kind of relationship.

Bivariate Data Exploration:

Occupation Loan Original amount: Pharmacists and Doctors borrowed highest loans while student borrowed the l east amount loan

Loan status and loan original amount: As we have got less data it unclear to establish any result but we can say that loan defaulter borrower more amount than borrower who paid off their loans .

Loan Status and term distribution: Borrowers having loan status past due for the term of 40 -60 months

Loan status Borrower rate: Borrowers who were defaulter have higher borrower rate than the borrowers who paid off their loans.

Borrrower APR lender yield: It was difficult to establish any relation by seeing the scatter plot I made. But after closely observing it shows points overlapping which gives a sense that the relation ship in neutral or positive.

borrower APR and lender yield : They positive strong relationship

Prosper rating state monthly income: Stated monthly Income of the borrowers having prosper rating E, D, HR were higher than remaining rating holders.

Credit grade borrower apr distribution: Borrower APR of A,B,AA credit grade holder was low while Borrower APR of F and HR holder was high.

My results:

Borrower APR:

Credit Grade VS Borrower APR : For a given Credit Grade, there is significant variation in Borrower APR, hence boxplot is a suitable choice. The boxplot also helped me to observe the skewness in distribution of Borrower APR across different Credit Grade. As expected, the average APR charged to borrowers with poor credit grades is significantly higher than safe borrowers. It is worth adding that the distribution for Credit Grade =HR, distribution is left skewed, which might hint towards higher APR being charged for risky borrowers.

Borrower APR VS Lender Yield: Borrower and lender yield are highly correlated. Higher the Borrower APR higher the Lender yield.

Borrower APR and Income Range Distribution: It is clear that borrower with higher income range has highest Borrower while borrower who are not employed and have income range under $25,000.

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 riskier. 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: 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 Prosper Rating. 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 cannot be so confident in this conclusion, because our results are based on this small dataset, hence, might not reflect the true figures.

Loan Status:

1 Loan Original Amount : We can see from the above plot that borrowers who borrow small amount of Loan Original Amount 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 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 Inquiries last 6 months on Loan Status is not remarkable.

Adding more, I also investigated loan performance at state level. The goal is to identify the state with highest percentage of defaulters in the given dataset. To this end, I identified top five rich and poor states. Since distribution for monthly income is right skewed, we used median as a performance measure. I found that North Dakota (ND) has highest percentage of defaulters. One possible reason for large number of default cases in ND can be attributed to the fact that average loan amount for ND exceeds the average income of residents. Identification of additional reasons may require thorough investigation of this dataset as well as obtaining more data.

Code Used To Render Slideshow

'jupyter nbconvert Explanatory_Analysis.ipynb --to slides --post serve --template output_toggle'

Feedback Obtained from Others:

The feedback I received from others was that my presentation had a few minor typos and grammatical errors that I could clean up. On top of that, they pointed out how my box plot visual was confusion because it did not show the risk values on the x-axis. The issue was that the plot was too large, and so needed to be resized for the presentation. Additional feedback included a suggestion to make the visualizations stand out from the background. As such, a dark grid was added to create more contrast from visual to slideshow. Lastly, it was advised that I create a cleared label for y column on my stacked plot, as it was confusing at first site. Hence, the label was updated to 'Loan Count by Loan Status'.

List of Resources Used

https://www.renttrack.com/blog/how-tradelines-impact-your-credit-score/

https://seaborn.pydata.org/tutorial/axis_grids.html

https://matplotlib.org/examples/color/named_colors.html

https://stackoverflow.com/questions/40389018/dropping-multiple-columns-from-a-data-frame-using-python/41579847

https://stackoverflow.com/questions/19377969/combine-two-columns-of-text-in-dataframe-in-pandas-python

https://stackoverflow.com/questions/48429156/how-to-resize-matplotlib-scatter-plot

https://stackoverflow.com/questions/52834616/i-used-matplotlib-but-the-error-message-figure-size-720x576-with-0-axes-app