Loan Data from Prosper

Author: Praxitelis-Nikolaos Kouroupetroglou

Data Analysis Nanodegree, Udacity

Visualization: Story Version 1, Story Final Version, GitHub

Introduction

In this data analysis project, I have explored the Prosper dataset and used Tableau to create my visualizations. Prosper is a peer-to-peer platform that lends money and its goal is to connect people who need money with those people who have the money to invest.

The Dataset

The Prosper loan data set contains 113,937 loans with 81 variables on each case. The dataset and data-dictionary which the latter explains the variable definitions can be found on the below links.

- Prosper Loan Dataset
- Prosper Loan Dataset – Variable Definitions

Summary

Prosper was founded in 2005 as the first peer-to-peer lending marketplace in the United States. Since then, Prosper has facilitated more than $13 billion in loans to more than 830,000 people.

Through Prosper, people can invest in each other in a way that is financially and socially rewarding. Borrowers apply online for a fixed-rate, fixed-term loan between $2,000 and $40,000. Individuals and institutions can invest in the loans and earn attractive returns. Prosper handles all loan servicing on behalf of the matched borrowers and investors.

Prosper Marketplace is backed by leading investors including Sequoia Capital, Francisco Partners, Institutional Venture Partners, and Credit Suisse NEXT Fund.

In my Tableau story I have done exploration of the Prosper loan data dataset, I did not know from where to start so, I started exploring by asking questions and I was guided by the Prosper Loan dataset variable definitions. at first, my initial question was the following:

1. ***I was wondering for the number of loans per state.***

The result is a distribution of number of loans per US state.

Then subsequently the next question that came to my mind was:

2. ***What is the demand of loans from 2005 to 2014 in respect of each US state***

The result is that there is an increasing demand in all the US states from 2005 to 2013 and California state has most of the loans from Prosper all these years. However, after 2013, in 2014 there is an overall drop in loan demand.

After the distribution of loan demand per US state and during the period of 2005 to 2014, the next question was:

3. ***How much and why do people ask for loans from Prosper***

The result is that most of the borrowers want to get some loans for “Debt Consolidation” and for “Baby and Adoption” reasons.

This question presented to me the reasons why people take a loan, the distribution of the loan amounts per listings and the number of loans per listing. Another question that I was starting wondering based on the previous one was:

4. ***based on the number of loan listings what are the borrowers’ characteristics that may play a vital role for a loan attribution.***

I plotted 2 histograms between the number of loan listings and the income range in respect of 2 features; the Boolean feature “isHomeOwner” and the Boolean feature “hasIncomeVerifiable”. There were surprising results; Regardless a borrower is a home owner or not, the Prosper will give a loan, however, the company will rarely attribute a loan to those who cannot verify their Income.

Guided from the previous question I was questioning myself who the best borrower in terms of risk and default rating may be, hence the following question that I tried to answer is:

5. ***What type of Borrowers are more risky to default***

The result was a scatterplot between Borrower Rate and Default Rate in respect of Borrowers’ occupations. Mostly Students are more prone to default and they are risky borrowers.

Working so much with risk and default, once again another question came to my if there are patterns between Default Rating and other features from Dataset. Two features that I worked with the Default Rating are the Income Range and the Prosper Rating and the question that I tried to answer is the following:

6. ***Does the Income Range and the Prosper Rating relate with Default Rate?***

The result is that the higher the income range the smaller the default rating. Moreover the better the prosper rating the lower the default rating will be.

Moving from Default Rate to Prosper Rate, I investigated over the period of the dataset (2005-2014) which proportion from each of the Income Range and Prosper Ratings was the majority from the loan listings. I stuck exploring in depth the Income Range and the Prosper Rating since they played a vital role in the previous question with their relationship with Default Rate. Hence my question is the following that I tried to visualize:

7. ***Inspecting the Quality of Loans in respect of the majority from each bin of the Income Range and the Prosper Rating from year to year.***

My visualization answered that question. So, from 2007 to 2014, The proportion of loans with lower incomes ranges have decreased while loans with borrowers with higher incomes ranges have increased. Based on the previous question higher income ranges leads to lower default ratings. Moreover, the quality of the loans have also been increasing after 2011. Moreover, the proportion of low Prosper score ("HR" or "E") loans are decreasing in 2013, this means again based on the previous question that overtime the default rating from loans are decreasing and the company becomes more mature.

Finally, I wanted to investigate some interesting fields from the dataset whether they relate with "Prosper Rating" which represent the loan quality. These are the "Estimated Effective Yield", the "Estimated Loss", the "Estimated Return" and the "LP_NetPrincipalLoss" in respect of the Prosper Rating. So, my questing that I visualized is the following:

8. ***Investigating other features in respect of Prosper Rating.***

My visualization revealed an interesting insight between the feature "Prosper Rating" and 4 other features; The "avg. Estimated Effective Yield", the "avg. Estimated Loss", the "avg. Estimated Return" and the "avg Net Principal Loss". If the quality of the loans decreases from "AA" to "HR", then the "average Estimated Yield", "average Estimated Loss" and the "average Estimated Return" increases. This means that loans with rating close to "HR" have high risk. But there is an exception, although loans with rating close to "HR" have the highest risk but do not have the highest "avg. estimated return" like loans with prosper rating "D" and "E". Furthermore, loans with rank "HR", "D, E " have higher actual principal loss. Overall don't invest your money in "E, D and HR" ratings loans.

Designing Plots – Graphs

Based on the questions that I focused I shifted between features that I was interested in exploring. My main goal is my Tableau story to have a goal and we can conclude to results based on the questions that we ask to the dataset and the visualizations can present the answers during the exploration.

Question 1:

1. ***I was wondering for the number of loans per state.***

To answer this question, I have used bubble chart and completed the aliases for the US States from abbreviations to their complete name.

2. ***What is the demand of loans from 2005 to 2014 in respect of each US state***

For this question, I have used line graphs and map charts to illustrate the loan demand over the years.

3. ***How much and why do people ask for loans from Prosper***

To depict this question, I have used 3 different types of graphs, a tree map which shows the average loan amount per listing, a histographs which shows per listing the loan amount distribution and a bar chart which shows the number of loans per listing.

4. ***based on the number of loan listings what are the borrowers' characteristics that may play a vital role for a loan attribution.***

For this question, I mainly used histographs – barcharts to find out if there is a significant difference between groups that I was investigating.

5. ***What type of Borrowers are more risky to default***

For this question, I used a scatteplot between the features Default Rate, and Borrower Rate in respect of the Occupation feature using resizable features shaped as rectangles.

6. ***Does the Income Range and the Prosper Rating relate with Default Rate?***

Here the bar charts played there role again to find out if there is a significant difference between groups.

7. ***Inspecting the Quality of Loans in respect of the majority from each bin of the Income Range and the Prosper Rating from year to year.***

I have used Area Charts to present the change over time between the relations of the features that I was investigating

8. ***Investigating other features in respect of Prosper Rating.***

Finally, once more I used bar charts to explore the difference among different groups between the related features.

Feedbacks

After completing the first story with Tableau, I shared it on Udacity Data Science Slack group. These are the feedback that I got:

- You have shown many stories. I would suggest keeping it concise with 4-5 important stories. You can figure it out on your own which is important. If two stories are similar, then merge them

To be honest I had more 1 tab that it was talking like the 3rd question about loan amount listings in respect of different occupations and it was a bit redundant, thus I erased that tab.

References

- Prosper main site
- Tableau resources