Communicate Data Findings

Udacity Project By Qingting.Song

Data Set: Prosper Loan Data

Prosper is San Francisco based peer-to-peer(P2P) lending company. PSP lending is a industry that matching and facilitating individuals borrowers and lenders. This kind of financial service is generally cheaper than traditional established financial institutions. For lender, this kind of platform generate higher return than saving or long-term investment in banks, and they also provide credit screening services. For borrowers, it's easier to lend loan from a P2P platform, the administration process is less exhausting and less time consuming.

This dataset is provided by Udacity. It contains 113936 data entries and 81 variables. Each entry contains all the necessary information about a loan. You can access the data dictionary via the link:
https://docs.google.com/spreadsheets/d/1gDyi_L4UvIrLTEC6Wri5nbaMmkGmLQBk-Yx3z0XDEtI/edit#gid=0

The main feature(s) of interest

In this analysis, I mainly focus on these questions:
What influence the Borrower Rate.
What is the demography of borrowers?
Do their personal financial situation influence their borrower rates.

To answer this question, the following variables might be helpful. I subtracted a subset of dataset and investigated further.

Features in the dataset

- ProsperRating (numeric)
- ProsperRating (Alpha)
- ProsperScore
- CreditScoreRangeLower
- CreditScoreRangeUpper
- BorrowerRate
- LenderYield
- EstimatedEffectiveYield
- EstimatedLoss
- EstimatedReturn
- BorrowerState

- Occupation
- EmploymentStatus
- EmploymentStatusDuration
- IsBorrowerHomeowner
- CurrentCreditLines
- OpenCreditLines
- TotalCreditLinespast7years
- OpenRevolvingAccounts
- OpenRevolvingMonthlyPayment
- LP_NetPrincipalLoss

Univariate Exploration

In this section, we will inspect individual variables' distribution and characteristics. And perform data cleaning if necessary.

Credit Score

In the original data set, we only have 2 variables: CreditScoreRangeLower and CreditScoreRangeUpper, the credit score we are using for our analysis is the mean of CreditScoreRangeLower and CreditScoreRangeUpper.

After obtained the average credit score, I inspected the distribution of variable: CreditScore.

CreditScore has Mean: 707.91, Mode 669.50 and Median: 709.5 .The mean to the left of the median. The distribution is negative & left-skewed. It seems that our borrowers are creditworthy in general.

Prosper Score

Prosper Score is 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.

Prosper Score has Mean : 5.95,, Mode: 8, Median: 6
The distributions of ProsperScore is approximately normal, since Mode and Median is approximately equal.

I also removed all the Prosper Score >10

StatedMonthlyIncome

Remove all the data entry where income > 13000 (Outliers)

Mode: 4166.6, Mean: 5468.69, Median: 5000
The distribution of Stated Monthly income is positively/right skewed, where the Mean is greater than Median.

EstimatedLoss

The distribution skewed to the right. In more case the estimate loss is 0.15 which is reasonable and not dramatic. The mean: 0.08, the median: 0.0724. In most cases the estimated loss is rather reasonable.

EstimatedReturn

Mean: 0.1, Median:0.0912. The distribution skewed to the right slightly.

CurrentCreditLines

I got rid of the outliers whose CurrentCreditLines is bigger than 30.
The distribution of CurrentCreditLines looks approximately nomral, where mean equals mean 10.37 meidan equals 10, it skews slightly to the right

OpenCreditLines

I got rid of the outliers whose OpenCreditLines > 24
The distribuation has Mean: 9.31, Median: 9. It skews to the right slightly.

DebtToIncomeRatio

I got rid of outliers whose DebtToIncomeRatio < 0.6
The distribution of debt to income ratio is right-skewed, with mean equals to 0.25 and median 0.23.

RevolvingCreditBalance

The distribution of RevolvingCreditBalance is right-skewed with the mode of 0. Looks like that most of our borrowers are in good financial situation.

LoanOriginalAmount

The mode of Loan Original Amount is 4000, most people are not borrowing that much. There are couples of spikes in borrower numbers around the amount 11000, 15000.

BorrowerRate

Median: 0.1845, mean : 0.19
The distribution is positivly/right skewed.

Bivariate Exploration and Key Findings

In this section, I used Correlation matrix and Scatter plot to explore the relationship between variables.

It seems that both Prosper score and Credit score have strong negative relationship with Borrower Rate. Borrowers with higher Prosper score and Credit score are able to borrow with lower cost (borrower rate).

Loan Original amount is negatively correlated with borrower rate. It seems that the bigger the amount a borrower borrow, the lower the rate.
Loan Original amount is positively correlated with Prosper score and credit score. It seems that borrowers with better scores are able to borrow bigger amount.

One thing interesting is that OpenCreditLines has a weak positively correlated relationship with StatedMonthlyIncome and LoanOriginalAmount. In this analysis we didn't run any regression and test it. But it might because people with higher income, in general has higher creditcard limits.

Useful links for me:

A Little Book of Python for Multivariate Analysis
https://python-for-multivariate-analysis.readthedocs.io/a_little_book_of_python_for_multivariate_analysis.html