Prosper Loan Data Analysis Conclusion

This Prosper loan data set roughly contains four kinds of loan data information — Loan Status, Borrower Data, Loan Information Data and Credit Risk Matrices during 2005 to 2014. I started by understanding some stories about Prosper and exploring the individual variables related to my initial question of interest — What factors are connected to defaulted/completed loan? And then I found Prosper have improve their credit measurement system by Prosper Rating from only using Credit Score before. And I continued to make observations on plots, then I make assumption that maybe Prosper put more weight of Prosper Prosper Score than Credit Score in their Prosper Rating. I kept exploring features related to Prosper Score and relationship between Prosper Score and these features and Loan Status. Eventually, I pick features that I observed that having some trends with Loan Status and Prosper Score. I created a Logistic model to predict the likelihood of a loan would be high risk or completed using these features.

I found Bank Card Utilization and Debt To Income Ratio have strongest inverse trend with Prosper Score. ScoreX Change At Time Of Listing and Stated Monthly Income have strongest positive trend with Prosper Score. And after investigating the relationship between status of loans and each related factor across each Prosepr Score, I found high risk loans tend to have higher BankCardUtilization, DebtToIncomeRatio, ProsperPrincipalOutstanding, and have lower ScorexChangeAtTimeOfListing, IncomeRange and StatedMonthlyIncome, holding Prosper Score constant.

Finally, I picked all these features above except IncomeRange and StatedMonthlyIncome to build a Logistic Regression model to predict likelihood of one loan would be completed or high risk. And I also built two other models: one only contains CreditScoreAverage as predictor variable; the other use Prosper Score as predictor variable. These three models present some predictive power for loan quality, but the power is weak. For model with four features, BankCardUtilization, DebtToIncomeRatio and ScorexChangeAtTimeOfListing are the significant variables. The results suggest that they have association of higher Bank Card Utilization, higher Debt To Income Ratio, and lower Scorex Change of the borrower with the probability of being high risk. The other two models only consider the predictor of credit matrices–CreditScoreAverage and Prosper Score, and the results suggest that both predictors are significant and have higher AUC ratio than previous model. Seems like directly using these two risk metrices is enough than using other four metrices at a time to predict a loan be high risk or not.

I think there are still some of limitations that cause such a not good model:

Data duration: The duration of the data set is from 2006 to 2014, and it covers the duration of financial crisis, which may cause the quality of loans and borrower condition be unstable. And this

may cause the predict model be more inaccurate if I want to evaluate the loan status of today's market.

Data subset and cleaning: Since the data contains so many categories of data type like borrower quality or time series data, it is appropriate to conduct a data subset to have more precise data content. For example, we know there is a financial crisis in 2008 which affects the whole borrower condition and quality of loans, maybe it is more appropriate to subset this data set by post-crisis years. Besides, I also didn't consider the outlier data in each feature to avoid irrelevant information, I just filled NA values with median in these three models.