Prosper Loans: An Exploration of Loss and Defaults
by Summer Cook

DATASET

> Provide basic information about your dataset in this section. If you selected your own dataset, make sure you note the source of your data and summarize any data wrangling steps that you performed before you started your exploration.

I chose the Prosper Loan data set. There are 81 columns and 113937 rows in the original data set. Each observation is a loan and each column is an associated variable that describes the loan and the borrower of the loan. I chose to exclude loans that were currently running at the time data collection stopped. I only looked at loans that had run their course to either completion or default. There were many outliers in Debt to Income Ratio, Credit Scores and Stated Monthly Income columns. However, I did not have reason to believe that these were invalid values. The final data set after outliers are non-completed loans were removed was 43092. The loans started in 2006 and ended in 2014. There is a period of missing data between 2009 and 2010. I added a new column with boolean values that indicated whether or not a loan had been defaulted or charged off as well as a column with the titles of Listing Names.

Variables of interest are Net Principal Loans and loans with a Loan Status of Default or Charged Off. Variables I expect to be explanatory are: Credit Scores, Income, Debt to Income Ratio, Prosper Rating, Interest Rates, Listing Category (What the loan is used for), Homeownership and Estimated Loss.

SUMMARY OF FINDINGS

> Summarize all of your findings from your exploration here, whether you plan on bringing them into your explanatory presentation or not.

Explanatory variables: The numeric variables that I hoped would help explain Loss and defaults were correlated relatively strongly with each other. For example, estimated loss and credit scores have a correlation coefficient of 0.57 and estimated loss and credit scores have 0.91.

Box plots with Prosper ratings on the x-axis shows that Prosper ratings have a similar relationship to both estimated loss and interest rates.

Net Principal Loss: I found only weak correlations between my numeric variable of interest (Net Principal Loss) and the explanatory variables. All of the relationships were in the direction I assumed: Variables associated with risk such as a low credit score or a high debt to income ratio are positively correlated with Net Principal Loss.

However the strongest correlation was with Prosper Principal Outstanding, which as 0.081. No linear relationships were found.

When the numeric data has visualized a tendency became clear. Loans with high Net Principal Loss were more likely to be described by lower risk variables. Overall, credit scores are negatively correlated with net principal loss, but among the loans where the loss was high, the lowest credit scores are not represented. It became apparent that borrowers with high risk variables were not permitted to borrow large sums or interest rates were set so high that the loans were not desirable.

Defaults and Charged Off loans: Relationships were not easy to identify. However I could identify that the lowest two Prosper ratings have higher rates of defaults than the better Prosper ratings.

The categories with the highest levels of defaults were the 'Not Available' Loan Listing Category (what the loan is used for) and the 'Not Displayed' categories in the Income Range variable.

Not collecting data on these variables seemed to be associated with risk of Loss. When I explored these variables over time, I discovered that the listing category and income data were not collected until 2007 and 2008 respectively. I then explored how defaults have progressed over time, and they have decreased. This led me to speculate that better data collection may have led to fewer
defaults over time.

KEY INSIGHTS FOR THE PRESENTATION

> Select one or two main threads from your exploration to polish up for your presentation. Note any changes in design from your exploration step here.

I chose to focus on Net Principal Loss in my presentation and to exclude the other variable of interest, Defaults and Charged off loans for simplification, as relationships were not clear for either variable of interest. There were weak interactions but no linear relationships with my variable of interest and the chosen explanatory variables. The visual exploration showed a difference between loans with higher loss and loans with lower loss. Borrowers with higher risk variables seem to be excluded from large loans, and they are therefore not represented among the loans that have high levels of loss. A multivariate plot shows that loans with high levels of loss were more likely to be homeowners and have higher levels of credit scores. Low levels of Loss were more likely among non-homeowners with low credit scores.