Mortgage Loan Default Prediction

Capstone Project I

Springboard Data Science Career Track | Nikki Seegars

DATA SOURCES

Fannie Mae has made available on their website a subset of the single-family loans that were acquired by the agency since 2000.

The dataset is available in .txt format by quarter and each quarter has the following two types of datasets:

- Acquisition Data (identifying data i.e. loan type, borrower credit score, original interest rate)
- Performance Data (monthly data i.e. current loan balance, delinquency status, loan age)

DATA EXPLORATION

Once the data was cleaned and saved, the summary data was explored in search of discovering trends and insights about the dataset.

Reading through the supporting documents on the Fannie Mae website resulted in concentrating on exploring the following fields:

- Borrower Credit Score, Co Borrower Credit Score
- Zero Balance Code
- Original Interest Rate
- Original Unpaid Balance
- Original Loan to Value
- Original Debt-to-Income Ratio

DATASET WRANGLING

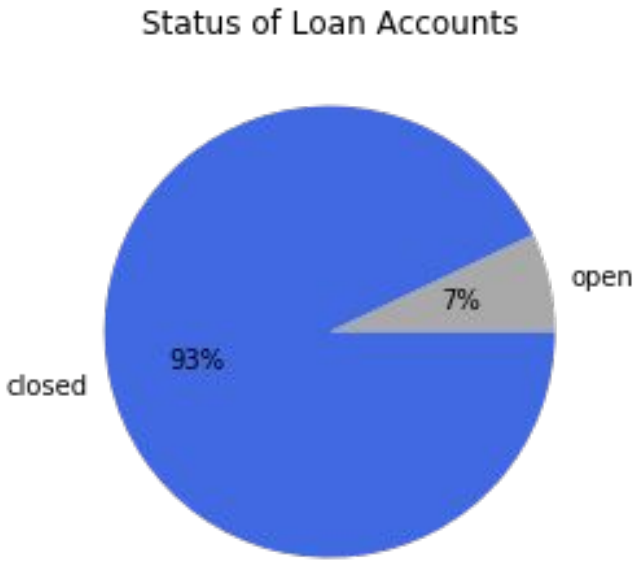
The dataset is pretty well organized in the text files. However there were a few steps that were required to clean the data including:

- Adding header names
- Removing rows with missing data for borrower credit score
- Filling in number of borrowers with either 1 or 2 depending on credit score data
- Combining the two dataset files into DataFrame and saving to a text file for later use

DESCRIPTIVE STATISTICS

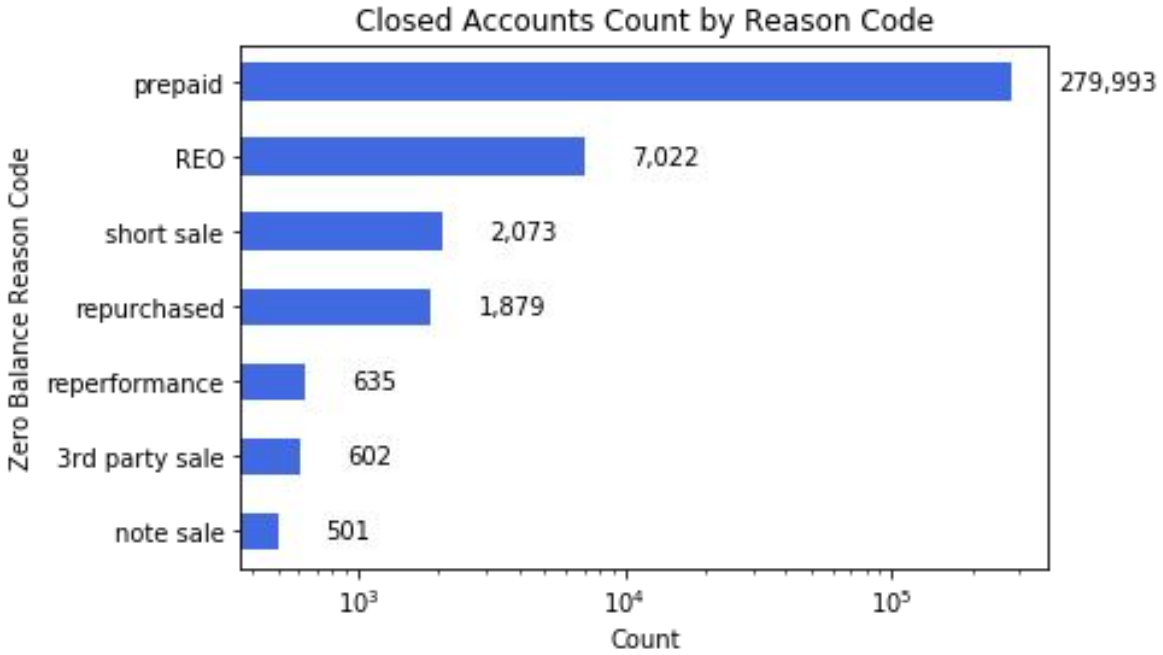
The 4th quarter of 2008 data set has nearly 10 years worth of records. Approximately 93% of those accounts are currently closed.

Total # of Accounts ~ 315,000



DESCRIPTIVE STATISTICS

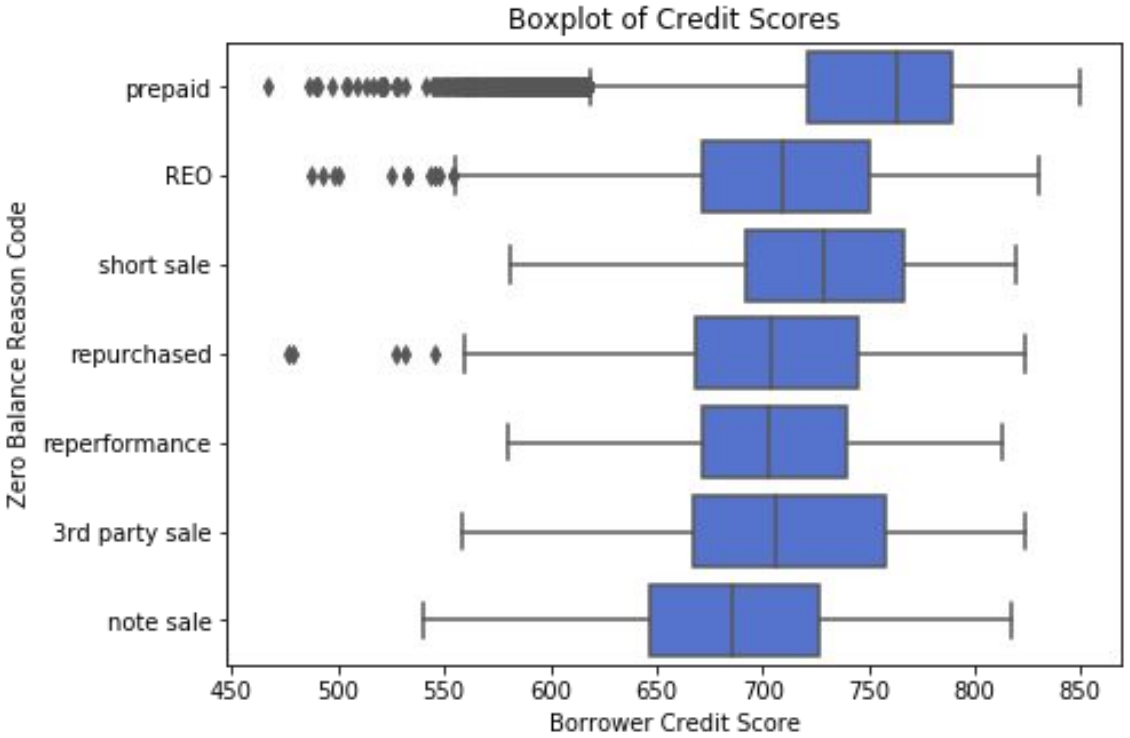
Of the 93% of accounts that are closed, most were prepaid (or matured).



DESCRIPTIVE STATISTICS

- With a larger portion of the data, the prepaid accounts have a credit scores with a larger variance and higher medians.

- Accounts closed coded as REO had significantly lower mean.



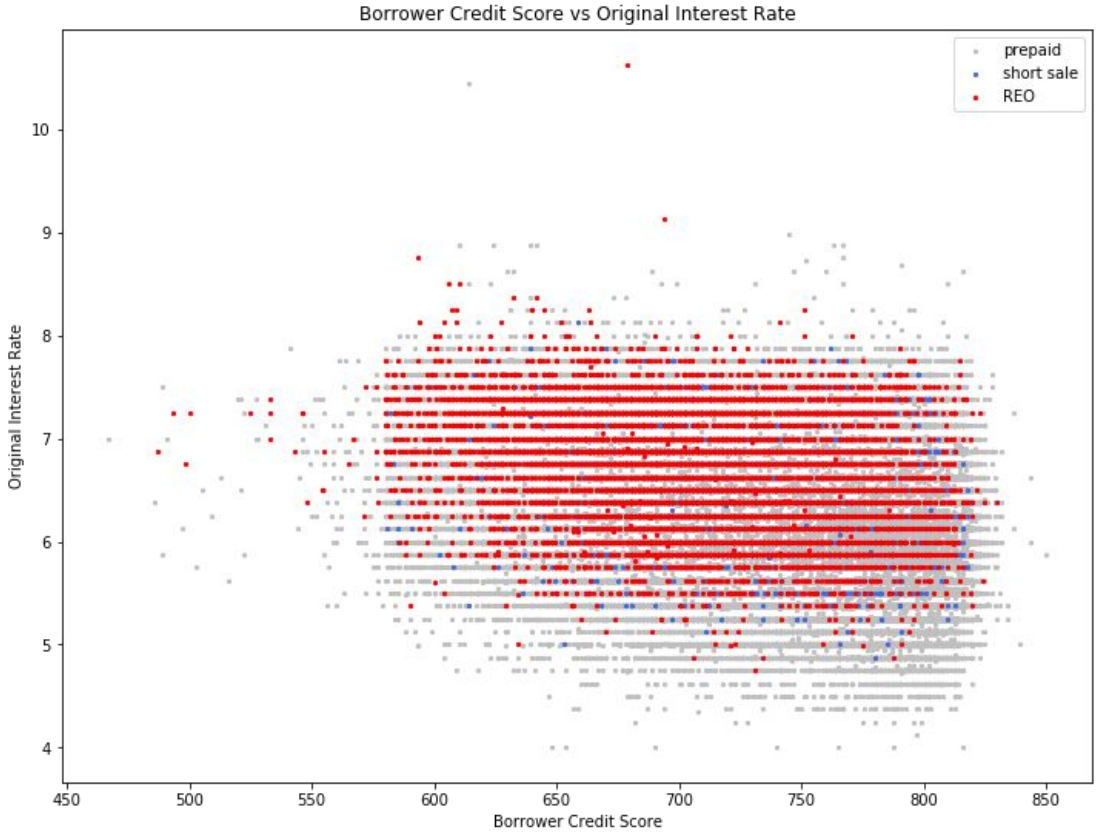
DESCRIPTIVE STATISTICS

Loan Closure Type	Median	Mean	Standard Deviation
All Closed Loans	761.0	749.3	48.8
Prepaid	763.0	751.1	47.9
REO	710.0	708.9	52.9

DESCRIPTIVE STATISTICS

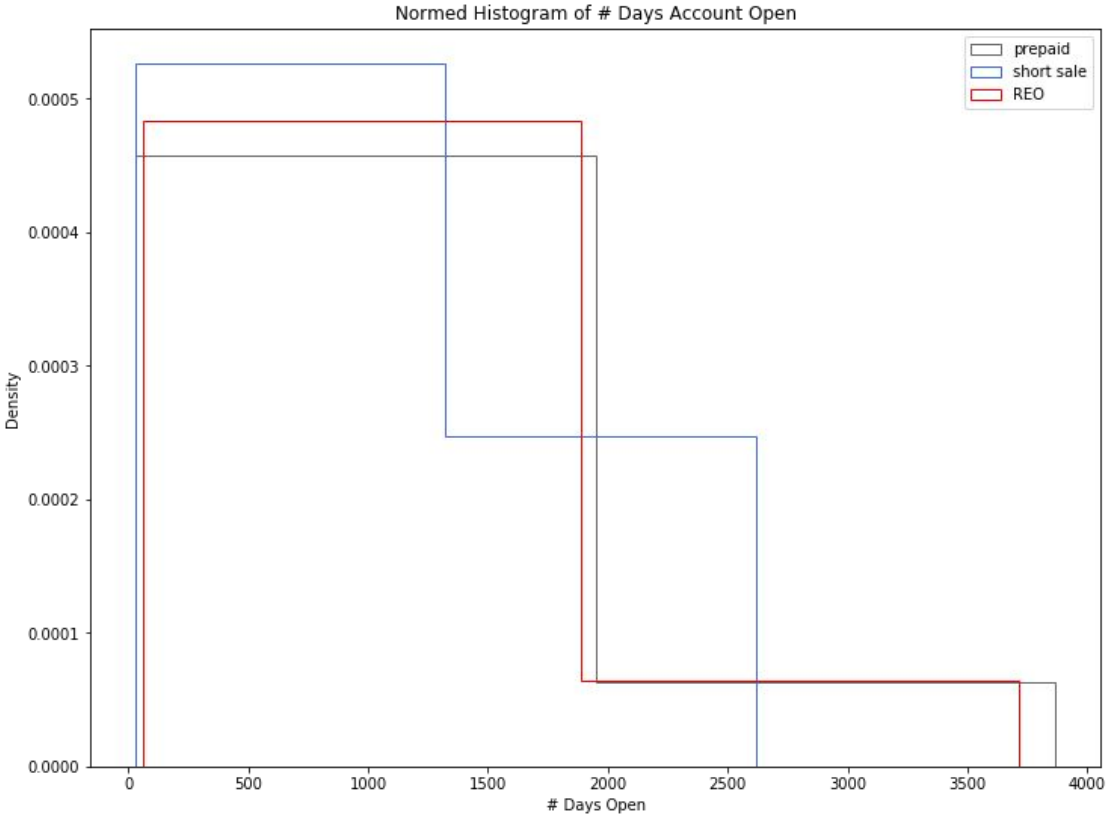
Narrowing down the data's focus to those accounts that were either in default by borrower or sold short, comparison can be made to the larger subset of prepaid loans.

Accounts with lower credit scores tend to have only a slightly higher interest rate if any at all.



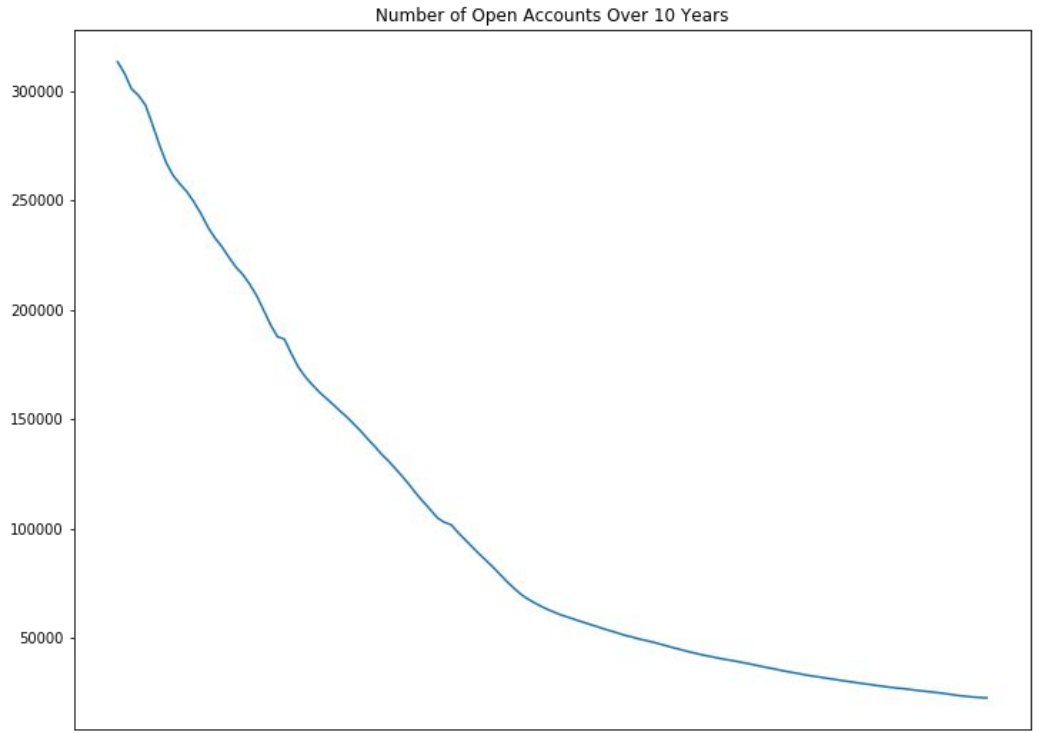
DESCRIPTIVE STATISTICS

The vast majority of all the accounts are closed within 2,000 days ~ 7.5 years of the loan origination date.



DESCRIPTIVE STATISTICS

Over 80% of loans are closed within five years of the origination date.



INFERENTIAL STATISTICS

Bootstrap sampling was performed to test the null hypothesis that there is no difference between borrower credit scores that were foreclosed (REO loans) and those loans that were prepaid.

$\overline{\mu_P}$ = population mean of prepaid loans

$\overline{\mu_R}$ = population mean of REO loans

$\overline{x_P}$ = sample mean of prepaid loans

$\overline{x_R}$ = sample mean of REO loans

$$p\ value = \sum \left[(\overline{\mu_P} - \overline{\mu_R}) >= (\overline{x_P} - \overline{x_R}) \right] / 10,000$$

$$p\ value = 0$$

CONCLUSION

Preliminary data exploration suggests that Fannie Mae mortgage loan accounts

- Close out early, usually within five years of the origination date.
- Borrower credit score is somewhat in line with original loan interest rate, however there is not a strong correlation.
- Most of the loans close with borrower prepaying however there is a small portion of loans that default. These loans that default do tend to have lower median borrower credit scores.
- The null hypothesis was rejected by using bootstrap resampling to test whether there is no difference between borrower credit scores that were foreclosed (REO loans) and those loans that were prepaid.