Gregory Morneault

General Assembly

Data Science - Fall 2014

Predicting Loan Defaults on Lending Club Borrowers

# Introduction

- Size of the industry

- Disappearance of traditional credit sources post-2008 financial crisis

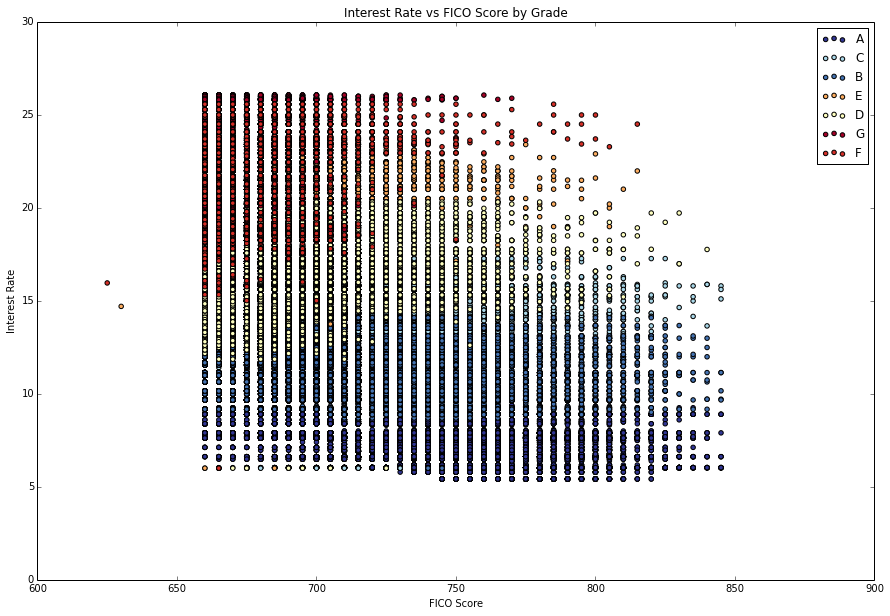
- peer-to-peer lending comes in to fill the gap

- comparison of institutional investors to peer-to-peer lending

# Problem and Hypothesis

Lending Club has their own proprietary credit risk pricing algorithm. Similar to FICO scores, the grades are designed to reliably predict the likelihood of a borrower defaulting given a set of borrower and loan characteristics. Unfortunately, the borrower grading algorithm is not openly shared with potential investors. Moreover, there appear to be some anomalies with respect to how the grades translate into the interest rate assigned to each borrower as evidenced by the scatter plot of interest rate by FICO score with an overlay of the grade assigned to the borrower (Figure 1). It is expected that the grading algorithm will incorporate other factors than just borrower FICO and that there need not be a direct correlation between these variables. However, it is interesting to see that interest rates applied to each grade are not consistent across each band. In other words, some borrowers assigned a higher grade actually incur a higher interest rate.

Figure 1 - Scatter Plot of Interest Rate by FICO Score



This is just one example where the grading system is not entirely transparent or consistent with expectations. In addition to lack of transparency, Lending Club is incentivized to originate as many loans as possible because the company earns income from origination and servicing fees and whereas the investor earns income from interest. This incentive may come at the cost of originating loans to less creditworthy borrowers as lower interest rates than would otherwise be underwritten had the originating company had "skin in the game".

Therefore, there is an opportunity for independent third-party risk pricing platforms to identify mispricing in Lending Club's credit risk grading algorithm. The intention of this paper is to demonstrate some machine learning methods that might be used reliably predict default and may eventually productionalized to independently price borrower credit risk. Assuming such models are successful, it would then be possible to take advantage of such mispricings or sell the independent pricing service to other willing investors.

# Description of Dataset

According to the Lending Club site , as of June 30, 2014, the average borrower has the following characteristics:

* 700 FICO score
* 16.7% debt-to-income ratio (excluding mortgage)
* 15.7 years of credit history
* $72,751 personal income (top 10% of US population)
* Average Loan Size: $14,056

The dataset was pulled from Lending Club's site and is freely available and continuously updated. As of August 30th, 2014, when the data was pulled, there were 376, 308 unique records with 100 fields related to details of the loan being issued as well as borrower characteristics and risk metrics. The data fields can broadly be classified into the following catagories:

* Loan Details - Issue Date, Amount, Maturity, Purpose, Interest Rate, Grade
* Borrower Characteristics - Occupation, Years of Employment, Annual Income, Housing Status, Location of Residence
* Borrower Credit Metrics - FICO Score, Debt-to-Income Ratio, Credit History, Public Records
* Performance - Current Status, Payment History

Some highlights from Lending Club's site :

* Over $5 billion has been originated since Lending Club first started in 2007
* The majority (83%) of borrowers take out a loan for the purposes of refinancing an existing loan or to pay of credit cards.

Interest rates are set by loan grades and span from 6% to 26%

# Feature Selection and Exploration

An initial set of features were chosen based on a industry experience and understanding of generally what drives loan performance. These features were then examined for collinearity and ensemble classification techniques were applied to identify relative importance when attempting to identify "Good Loans" vs "Bad Loans". Good loans are defined as loans that have fully paid-off either through reaching maturity or through prepayment in full prior to maturity. Bad loans are defined as loans that are either in default status or have already been charged off. All other loans were excluded from the dataset when determining relative importance of the feature set through application of the following three ensemble classification techniques: Random Forest, Extra Tree and Ada Boost. The average importance factors from these three algorithms were taken and a ranking was then generated. Table 1 highlights this ranking.

Table 1 - Feature Importance



From the table we can see that some important features that should be leveraged in any default-payoff classification algorithm are the borrower's annual income, their debt-to-income ratio as well as the current revolving utilization rate. Features that have less of an importance are home ownership and income verification.

It should be noted that while the interest rate has high importance and grade has low importance, these features are highly correlated (0.86). It is likely that each of these ensemble classification algorithms favored interest rate over the numerically encoded grade due to a greater granularity for splitting at each node. However, since both of these factors are being set by Lending Club, if we were to take a truly independent approach, both of these features would be removed from any classification approach.

Nonetheless, after a bit of trial and error, it was decided to keep all features when comparing the logistic regression approach to the random forest approach. This is decision is not optimal and will likely lead to over-fitting, however due to a running short on time and a desire to keep same features in both models for comparison purposes, no features from this initial selection we dropped.

Figure 2 shows distributions of some of the continuous variables within the feature set. Debt-to-income, interest rate, and utilization rate are largely normal while FICO score, funded amount, annual income, and total accounts have largely positively skewed distributions. It is interesting to note that employment length is bimodal with the majority of borrowers either just started employment or employed for 10+ years.

Figure 2 - Histograms of Selected Features

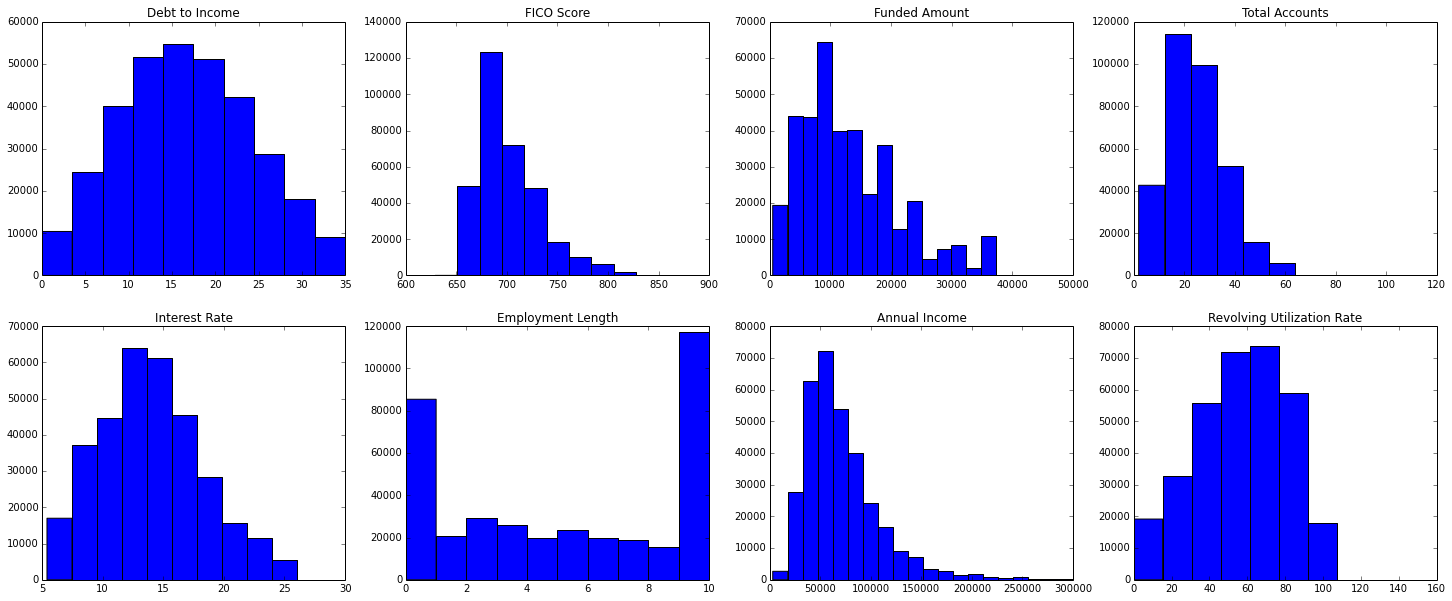
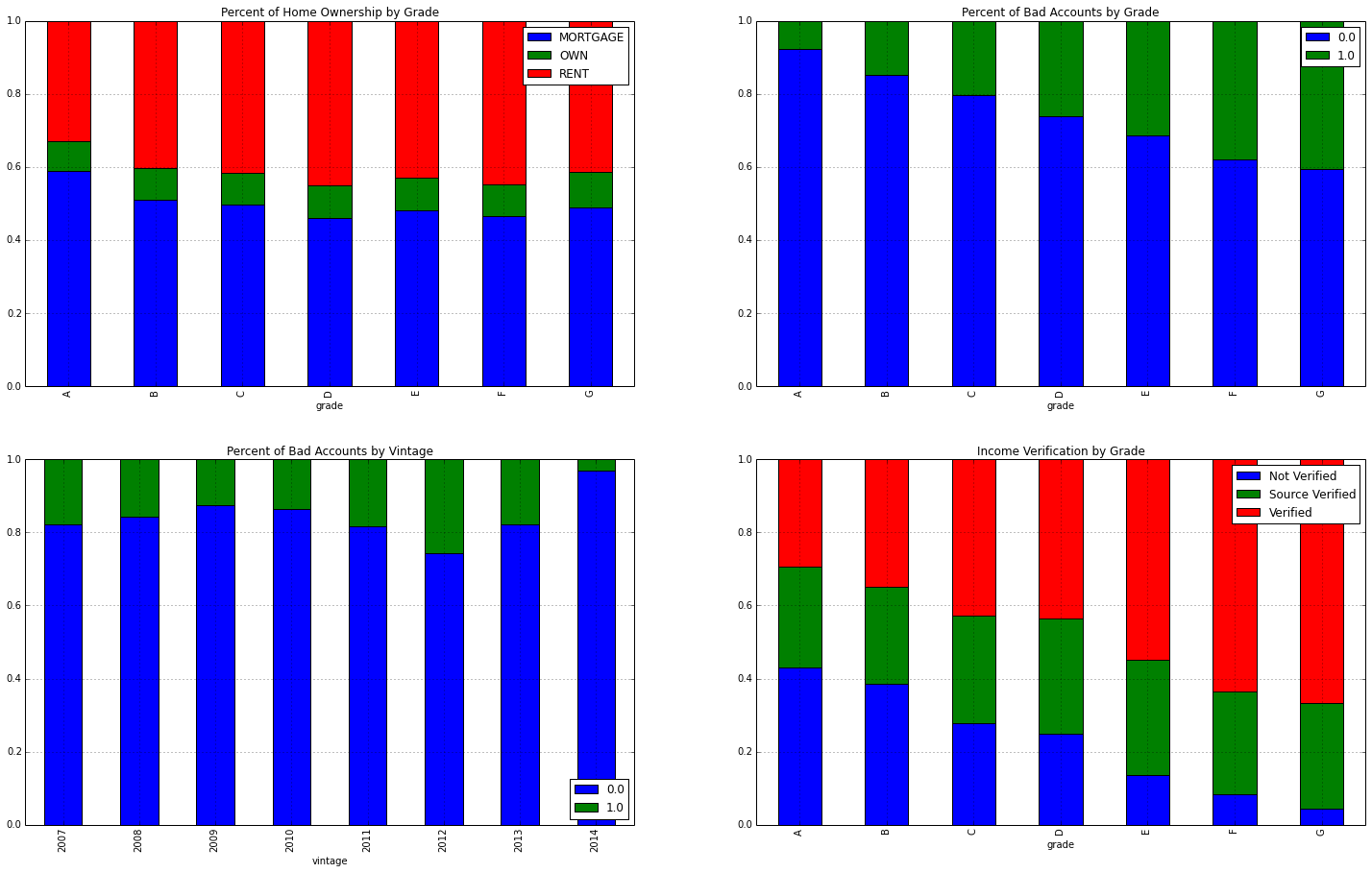


Figure 3 highlights some interesting relationships between the categorical variables within the feature set, the grade assigned to each borrower, and whether the loan was ultimately determined to be "good" or "bad".

Home ownership is largely consistent between grades which helps to explain for the lack of explanatory power in determining between good and bad loans. It is interesting to note that more borrowers have income verification with a lower grade. This is likely to compensate for other degrading factors in the borrowers profile and may be a part of Lending Club's underwriting policy.

Figure 3 - Stacked Bar Charts for Selected Features



# Classification Methods and Findings

Ultimately, I settled on three different classification techniques in an attempt to predict the liklihood of a borrower defaulting over paying in full.

## Random Forest

Table 2 - Random Forest Confusion Matrix



Table 3 - Random Forest Classification Report



## Logistic Regression

Table 4 - Logistic Regression Confusion Matrix



Table 5 - Logistic Regression Classification Report



## Naive Bayes

Table 6 - Naive Bayes Confusion Matrix



Table 7 - Naive Bayes Classification Report



# Business Application and Future Research

- predictive accuracy sucks but valuable insights into factors that drive credit risk

- portfolio optimization

# Works Cited

*Avergage Borrower Characteristics*. (n.d.). Retrieved from Lending Club: https://www.lendingclub.com/public/steady-returns.action

*Data Download*. (n.d.). Retrieved from Lending Club: https://www.lendingclub.com/info/download-data.action

*Highlights*. (n.d.). Retrieved from Lending Club: https://www.lendingclub.com/info/statistics.action

*How Lending Club sets Interest Rates*. (n.d.). Retrieved from Lending Club: https://www.lendingclub.com/public/how-we-set-interest-rates.action

*Loan Details*. (n.d.). Retrieved from Lending Club: https://www.lendingclub.com/info/demand-and-credit-profile.action