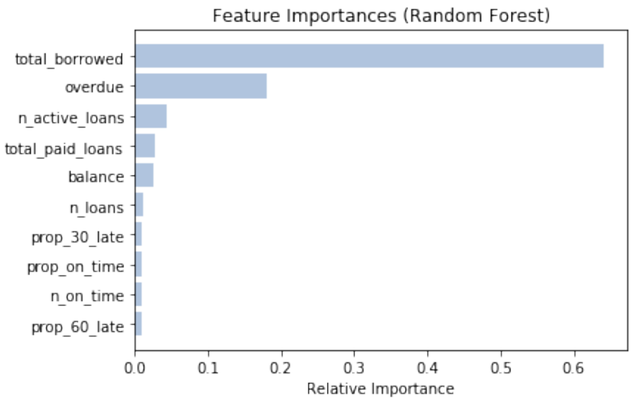
Overview:

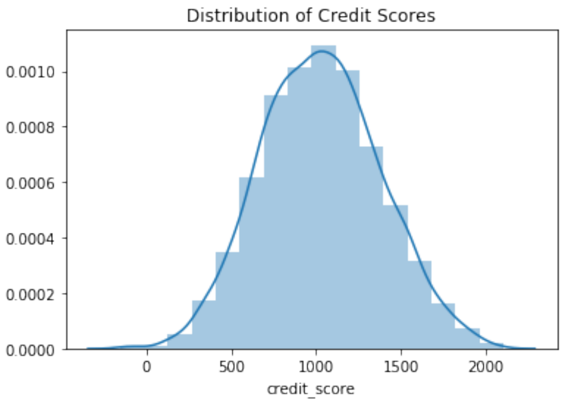
Here's how I formatted the data:

Each row in the dataset represents a single customer. It will contain their SSN and ID (to uniquely identify the customer), credit score, as well as a few features that I will create to summarize important information about their loan history. Here is a list of features I created (12 in total):

1. The total number of loans and number of active loans
2. The current balance owed and the total amount borrowed (across all loans)
3. The year the borrower started borrowing
4. The total number of months of loans
5. Counts and proportions of on time, 30 days late, 60+ days late payments (across all loans).

After creating the dataset, I used 2 models to measure performance. I used linear regression as a baseline model and random forest (tuned via 5-fold cv) as the final model. I won't be interpreting the linear regression coefficients because some of these features are collinear (ex. proportion of on time payments and number of on time payments). However, tree-based models like random forests can also give me a sense of feature importance (via mean decrease impurity). So, from the random forest, I was be able to extrapolate the most important features (as deemed by the model) for determining one's credit score.





The OLS model yielded an RMSE of about 104 while the random forest yielded an RMSE of about 123. Given the scale of the distribution of credit scores (mean around 1000 and sd around 350), the initial models performed moderately well, though there is certainly much room for improvement.

From the random forest, we can see that the most important features for determining credit score are the total amount of money borrowed across all loans and the amount (currently) overdue. Other important features include the total number of loans paid off, the number of active loans, and the current balance of those loans.

So, if each person has $1,000 to improve their credit score, they should use it to pay off some of the overdue balance, which affects two of the top 5 most important features (i.e. overdue, balance).