Analysis of Home Loan Application Acceptances

Kevin Malis, April 2019

**Executive Summary**

This analysis looks into the driving factors behind whether a mortgage loan application might be accepted or rejected. 500000 different applications with 21 different characteristics and a label of whether or not the application was accepted were analyzed in order to determine how accurately one could use these characteristics to predict whether an application would be accepted.

The first step in this process was examining the relationships between home loan application characteristics as well as how each characteristic correlated to loan application acceptance frequency. By visualizing these relationships as well as using various data exploration and model selection methods, I was able to determine which characteristics were most important in determining whether a mortgage loan application was accepted. After using these visualizations and initial statistical analysis to determine which features in the dataset had the most predictive power for acceptance rate, I used extensive feature engineering to create new characteristics that helped more accurately predict whether a loan was accepted.

Based on my analysis, I was able to conclude that the following factors are some of the most significant indicators of mortgage application acceptance.

**Applicant Income-** The higher an applicant’s income, the more likely that application is going to be accepted, even accounting for the loan amount likely being higher. For this analysis, a logarithmic transformation was applied to the applicant income data to minimize the effect of outliers at the upper end. The same correlation was found when using the logarithm in the model.

**Loan Amount –** Somewhat counterintuitively, as the loan amount increases, application acceptance tended to increase as well. As in the case for the applicant income characteristic, a logarithmic transformation was applied to the loan amount information to account for outliers.

**Loan Purpose -** Loan purpose also seemed to influence loan acceptance. Applications for loans that were intended for home purchase were accepted at a much higher rate than applications intended for home improvement or for loans that were for refinancing, and this strong correlation became clearer when engineering a loan purpose acceptance rate feature.

**Property Type**- Applications intended for loans on multifamily homes tended to be accepted at a higher rate than applications for loans for manufactured housing, though both of these categories made up a small amount of the total applications.

**Lender –** Across the 500,000 applications analyzed, there were over 6,000 unique lenders, whose specific acceptance tendencies had by far the strongest effects on whether a loan was accepted.

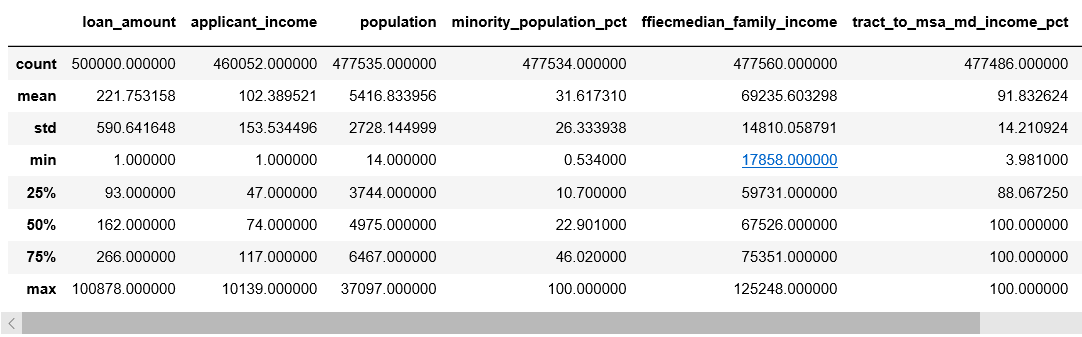
**State –** There were 52 state codes, with acceptance rates ranging from .38 to .62, with applications that were coded with a missing value for state having an acceptance rate of .03. These fluctuations had a small but significant influence on application acceptance.

Although my initial data exploration and analysis demonstrated that the above characteristics were some of the most crucial in predicting loan power, they were not the features that ended up being included in my classification model. In order to improve the accuracy of my model, I used feature engineering to create rates for the most relevant categorical variables. The actual factors included in my model to maximize the prediction of acceptance of loan rates were: log(loan\_amount), log(applicant\_income), lender acceptance rate, state acceptance rate, loan type acceptance rate, loan purpose acceptance rate, and property type acceptance. This process was validated when comparing the correlation values between the engineered feature rates and the original training data features, which showed much stronger correlations across the board between acceptance and the engineered features. Specifically, the acceptance rate of a lender tended to be by far the stronger predictor of overall application acceptance rate.

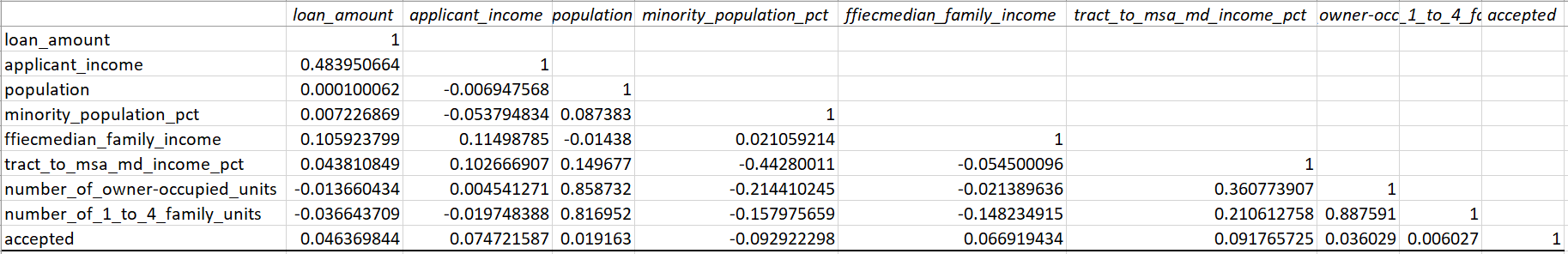
**Initial Data Exploration**

Initial data exploration consisted of determining which characteristics had incomplete data as well as examining their distribution and tendencies to see which must be included in any predictive analysis for loan acceptance.

Below is a table of summary statistics some of the key numeric features in the analysis. I have omitted statistics for a few numeric variables that did not end up being included or considered in the final analysis.



As noted above, the distributions of the loan amount and application income features are skewed at the upper end by a handful of incredibly wealthy loan applicants, resulting in large differences between the mean and median of each. Thus, log transformations were applied to these features with the resulting distributions being approximately normal. Furthermore, it is apparent that the applicant income characteristic is scaled differently than the median family income characteristic, which had to be accounted for in preliminary versions of the model, though median income was eventually excluded.



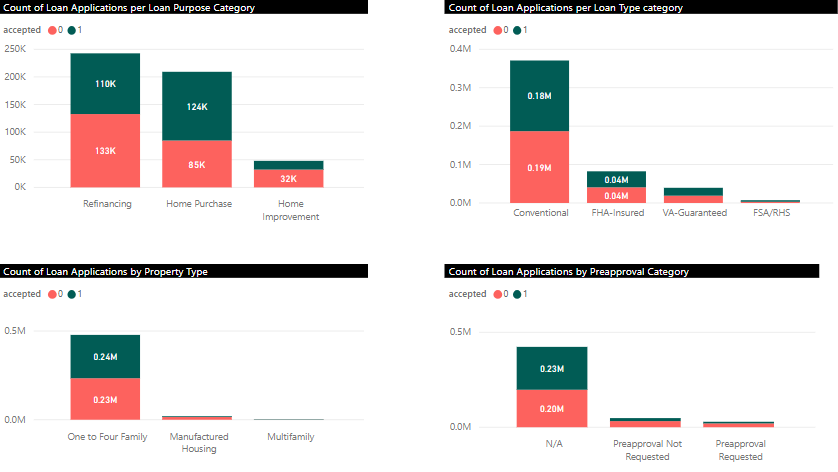
In addition to summary statistics for numeric features, bar charts were created for some of the key categorical variables to determine their relative frequencies within the data set. Here is what was discovered about some of the key categorical features:

**Loan Purpose-** Loans for home purchases and loans for refinancing were applied for at about approximately the same frequency, with home improvement loans being far less common in the sample data.

**Property Type** – Applications for manufactured housing and multifamily loans paled in comparison to the amount of applications for standard 1-4 family housing.

**Loan Type** – Conventional loan applications were more common than the other categories of loans(FHA, VA, FSA/RHS)

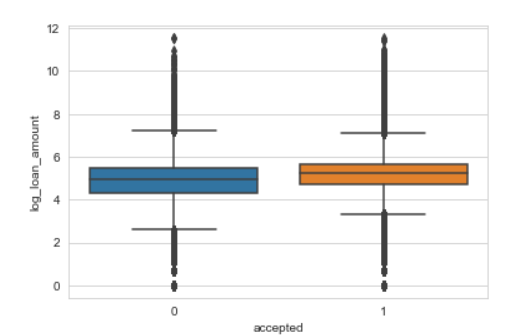
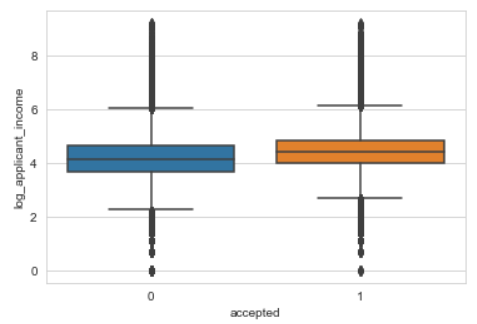
**Preapproval –** Though there were significant differences in the rates of acceptance of applications with preapproval vs. not preapproval, the vast majority of sample applications in the dataset had “not applicable” as a preapproval value.



The same analysis was done for all categorical variables, including geographic and demographic information for each applicant.

**Numeric Relationship Visualizations**

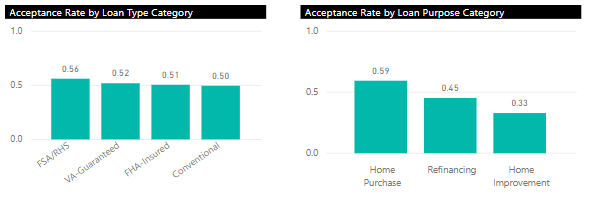
In addition to examining the correlations between numeric features in the chart shown above, Box plots were created in Python to help visualize the distributions of the values of certain numeric features in applications that were accepted compared to applications that weren’t. Initially, I created box plots for the loan amount and applicant income, but the outliers at the upper end rendered the plots uninterpretable. When a log transformation was applied to these two features, the plots were far more revealing and are shown below:

These features are not perfectly well separated by acceptance rate, yet there is a clear correlation between an applicant’s income and the loan amount applied for. This was reinforced by examining the correlations between acceptance and each feature, which will be discussed below.

**Categorical Relationship Visualizations**

For categorical variables, PowerBI was used to examine how acceptance rate varied across categories of the relevant characteristics.



The bar chart for acceptance rate across the loan purpose categories demonstrated that it was likely a strong predictor of acceptance rate, which was confirmed after performing classification analysis and associated feature importance.

Furthermore, two important categorical variables that demonstrated wildly differing acceptance rates for their categories were lender and state. I have left off bar chart visualizations of these because the high amount of categories(6,111 and 52, respectively) made them difficult to visualize. I used Power BI to create a few tables that grouped lender and state by acceptance rate; these tables clearly demonstrated that lender and state both must be taken into account when predicting loan application acceptance.

**Correlation Analysis**

A correlation matrix was created to evaluate some of the correlations apparent in the above visualizations. I have included a snippet of a correlation matrix that shows the correlation of each feature in isolation with the target accepted.



While these coefficients reinforce some of the insights provided by the visualizations in terms of which coefficients are most strongly correlated with acceptance rate(applicant income and loan purpose notably), they are still relatively low absolute in magnitude. However, when I applied some transformations to the categorical variables, which is discussed below in my feature engineering section, the correlations were much stronger.



Specifically, the lender\_rate variable being .445, with the next highest correlation being .178, shows how important the lender was to whether a loan ended up being accepted. The overall magnitude of the engineered correlations being significantly higher than the original correlations demonstrates that these engineered features would be much more helpful in predicting loan acceptance on new application data.

**Feature Engineering**

After using the above visualization and correlation analysis to determine which core characteristics were most important in driving loan application acceptance, feature engineering was employed in order to create new features that increase accuracy and decrease overfitting of the model. Essentially, this process included grouping each category of the different variables by acceptance rate, creating a lookup dictionary with the keys as the different categories and the values as the associated acceptance rate, and then replacing the categorical characteristics with the actual acceptance rate of that category. The features that were created using this methodology were: **loan type acceptance rate**, **loan purpose acceptance rate**, **property type acceptance rate**, **lender acceptance rate**, and **state acceptance rate.**

By replacing categorical variables with these numeric features with more direct relationships to acceptance rate, I was able to drastically improve the accuracy of prediction loans when exposed to new application data.

**Classification Analysis**

After analyzing the descriptive statistics and visualizations for each characteristic, 500,000 sample applications with associated acceptances as well as 500,000 applications to be tested on were loaded in the Azure Machine Learning Studio for rigorous machine learning analysis. Data preparation operations were applied to both sets of data in order to allow a classification to be run. This included the aforementioned feature engineering, logarithmic transformations to loan amount and applicant income, data normalization, and using a statistical method called MICE in Azure Machine Learning to replace missing data with the most appropriate values.

In order to generate predictions for the 500,000 test applications, a boosted two class decision tree was used, with hyperparameter tuning in order to maximize accuracy. I was able to achieve a maximum accuracy of .7134, meaning that the model could accurately predict whether a loan application would be accepted or rejected with a 71.34% accuracy.

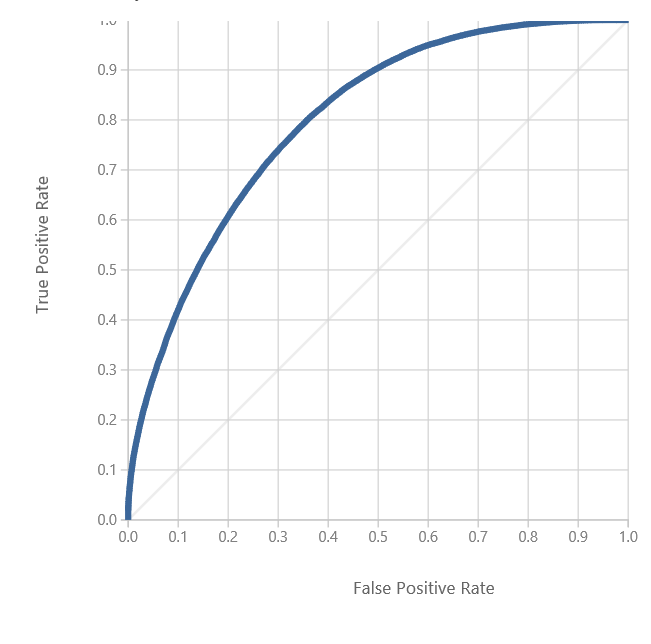
I have included some more detailed statistics around model performance from my training set below:

**Accuracy: .721**

**Precision: .704**

**Recall: .765**

**F1 Score: .733**

**ROC Curve:** 

The above statistics are slightly inflated because my model was slightly overfitted to the test data, as is apparent by the .722 training accuracy compared to my final .7134 accuracy on the test set. However, these statistics can still give some decent insight into how my model performed overall.

**Conclusion**

From this analysis, I was able to conclude that home loan acceptance is able to be predicted by certain characteristics of an application, but only to a certain extent. As you can see above, the highest accuracy I achieved using predictive modeling was .7134. Considering that home loan acceptance across the sample data was nearly exactly 50%, this model did significantly better than a model that just guessed at random, which is valuable. The final analysis showed that the most significant factors in driving home loan application acceptance were lender, applicant income, and loan purpose, with loan type property type, and loan amount being less important but still significant. However, only after transforming the strongest categorical predictors into rates did we see a strong correlation with overall application acceptance rate. These correlations made it clear that the strongest determining factor in whether a given loan application was expected was the lender’s overall tendency to accept rates, with the applicant income and loan purpose being secondary but strong factors.