Analysis of Mortgage Approval from Government Data

JUNE, 2019

Analysis of Mortgage Approval from Government Data

June, 2019

Executive Summary

This report presents an analysis of data with regards to mortgage approvals from a given dataset, adapted from the Federal Financial Institutions Examination Council's (FFIEC). The analysis is based on 500,000 observations of mortgage data with 21 features and 1 Label, each containing specific characteristics of HMDA-reported loan application, which covers one particular year.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between mortgage features and the status of an application were identified. After data exploration, data cleaning and some transformations were performed on the features after which a predictive model to classify mortgage loan application into two categories: accepted and not accepted was created.

After performing the analysis, the author presents the following conclusions:

While many factors can help indicate whether a mortgage loan approval gets accepted or not, significant features found in this analysis were:

1. Applicants Income: The income of most applicants is between 0 to 100, 000 dollars.
2. Loan Amount: An integer variable that represents the size of the requested loan in thousands of dollars. The amount of loan requested by most applications is below 400, 000 dollars.
3. Loan Purpose: Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing.
4. Number of Family units.

Exploratory Data Analysis

The initial data exploration began with some summary and descriptive statistics.

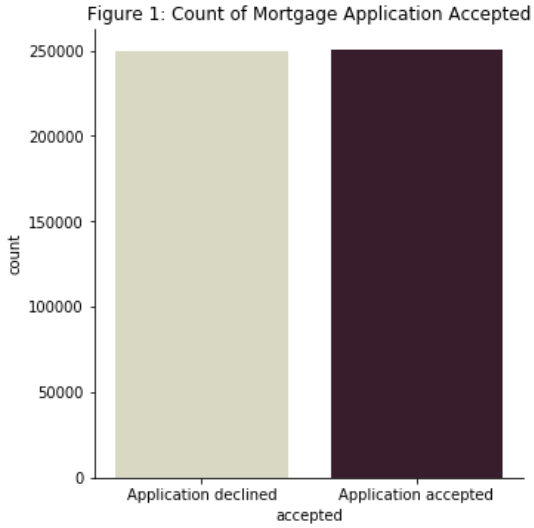
Individual Feature Statistics

Summary Statistics for minimum, maximum, mean, median, standard deviation and distinct count were calculated for numeric columns, and the results taken from 500,000 observations are shown in Table 1.

Table 1: Descriptive Statistics

Column	Count	Mean	Std Dev	Min	Median	Max
loan_amount	500000	221.7532	590.6416	1	162	100878
applicant_income	460052	102.3895	153.5345	1	74	10139
Population	477535	5416.8340	2728.1450	14	4975	37097
minority_population_pct	477534	31.6173	26.3339	0.534	22.901	100
ffiecmedian_family_income	477560	69235.6033	14810.0588	17858	67526	125248
tract_to_msa_md_income_pct	477486	91.8326	14.2109	3.981	100	100
number_of_owner-occupied_units	477435	1427.7183	737.5595	4	1327	8771
number_of_1_to_4_family_units	477470	1886.1471	914.1237	1	1753	13623

Since mortgage application acceptance is of interest in this analysis, it was noted that the percentage of mortgage application accepted is equivalent to the percentage of mortgage application declined. A count plot of the accepted column shows that there is a balance between the number of records of mortgage application accepted and declined with a ratio of 1:1, as shown in figure 1:



In addition to the 8 numeric values, the mortgage approval observations include 13 categorical features with description as follows:

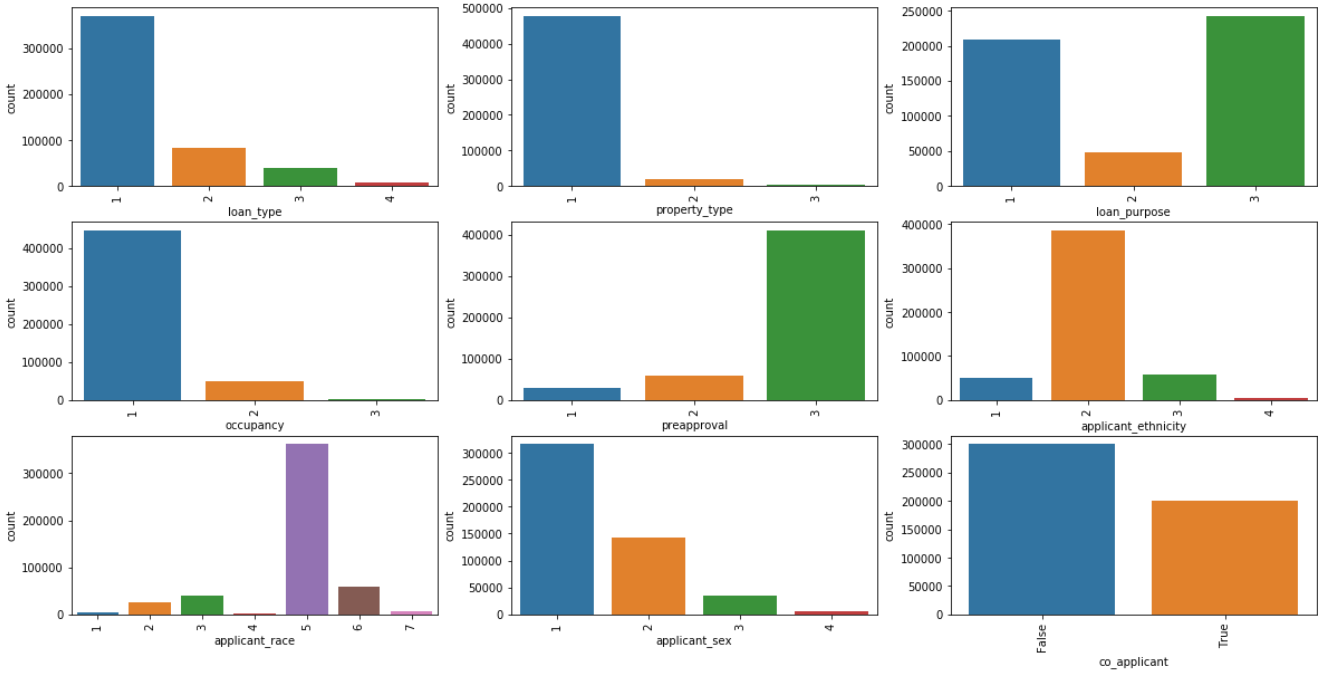
- **msa-md** – Metropolitan Statistical Area/Metropolitan Division with 408 unique observations and -1 indicating a missing value.
- **state_code** – US State code with 52 unique codes and -1 indicating a missing value.
- **county_code** – county code with 317 unique codes and -1 indicating a missing value.
- **lender** – Indicates which lender was in authority in approving or denying the loan. It has 6111 unique lenders.
- **loan_type** – 4 categories with 1 = Conventional, 2 = Federal Housing Administration-insured, 3 = Veteran Administration- guaranteed, 4 = FSA/RHS (i.e. Farm Service Agency/ Rural Housing Service)
- **property_type** – 3 categories indicating whether the loan/ application was 1= one to four-family (other than manufactured housing), 2 = Manufactured housing, 3 = Multifamily.
- **loan_purpose** – 3 categories indicating whether the purpose of loan is 1= Home purchase, 2= Home Improvement, 3 = Refinancing.
- **occupancy** – 3 categories indicating whether a property to which loan application relates is 1 = Owner-occupied as a principal dwelling, 2= Not owner-occupied, 3 = Not applicable.
- **preapproval** – 3 categories indicating whether an application/ loan involves a request is 1 = Preapproval was requested, 2 = Preapproval was not requested, 3= Not applicable.
- **Applicant_ethnicity** – 5 categories indicating ethnicity of applicant with 1 = Hispanic or Latino, 2 = Not Hispanic or Latino, 3 = Information not provided by applicant in mail, Internet, or telephone application, 4 = Not applicable, 5= No co-applicant.
- **Applicant_race** - 8 categories indicating race of applicant with available values 1 = American Indian or Alaska Native, 2 = Asian, 3 = Black or African American, 4 = Native Hawaiian or Other Pacific Islander, 5 = White, 6 = Information not provided by applicant in mail, Internet, or telephone application, 7 = Not applicable, 8 = No co-applicant.
- **Applicant_sex** – 5 categories indicating the sex of the applicant with applicable values: 1 = Male, 2= Female, 3 = Information not provided by applicant in mail, Internet, or telephone application and 4 or 5 = Not applicable
- **Co_applicant** – a Boolean value indicating true or false, which indicates whether there is a co-applicant (often a spouse) or not.

Bar charts were created to show frequency of these features as shown in figure 2, and indicate the following:

- Conventional Loan type are the most common of the four loan types.
- 95.6 percent of the mortgage application was for a one-to-four-family dwelling, followed by Manufactured housing which is approximately 3.9 percent of the loan application.
- The most common purpose for obtaining a loan was for refinancing, followed by home purchase and lastly home improvement.
- A property to which a loan application relates is most of an Owner's principal dwelling, and less of not owner occupied.

- Most loan/application has a request for a preapproval of a home purchase loan as not applicable, few others do not involve a preapproval request, while preapproval was requested for just little of the remaining application.
- 77.2 percent of the applicants are not Hispanic or Latino, 11.5 percent did not provide any information by mail or telephone, only 10.2 percent are Hispanic or Latino and the remaining applicants have their ethnicity as not applicable.
- Most of the applicants are White, followed by Black or African American with very small frequencies of the other values. However, it is worthy of note that very few of the applicants are Native Hawaiian or Other Pacific Islander.
- The vast majority of applicants are males, followed by few females.
- Most applications do not have a co-applicant (often a spouse) involved.

Figure 2: Countplots for each categorical variables



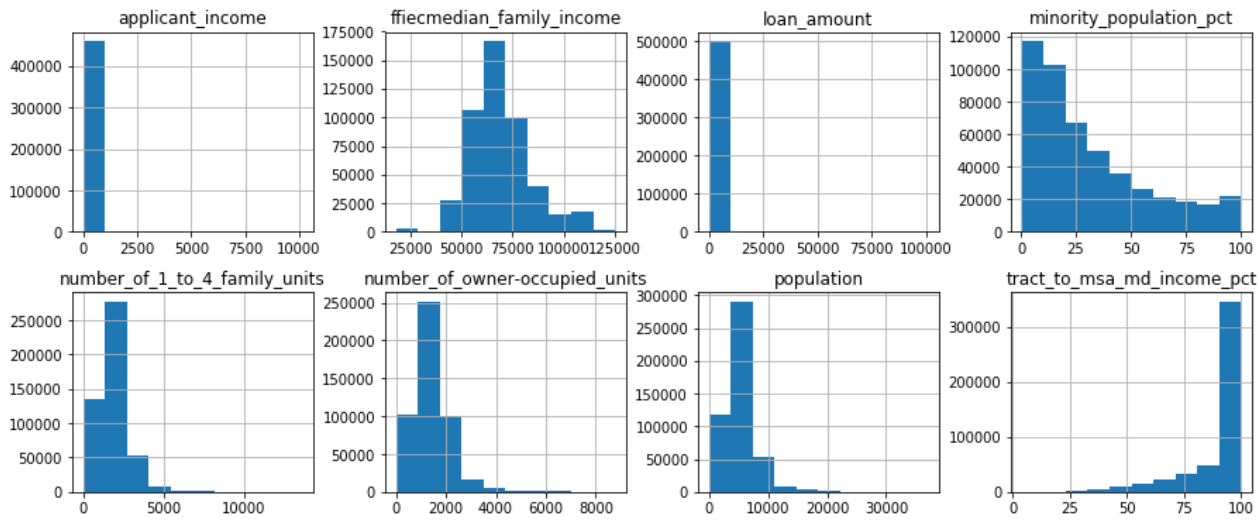
Correlation and Apparent Relationships

After exploring the individual features, an attempt was made to identify relationships between features in the data – in particular, between Accepted and the other features.

Numeric Relationships

The histogram in figure 3 was generated to see the distribution for each of the numeric features. The key features in this matrix are shown here:

Figure 3: Distributions for each of our numerical variables



From this visualization, we get a lot of information. We can deduce the following insight:

- Number_of 1_to_4_family units, number_of_owner-occupied_units and population are heavily skewed right. This implies there are less dwellings that are built to houses fewer than 5 families, also there are smaller number of dwellings occupied by the owners, and the total population in tract is less than 20000 for most of the applications.
- Ffiecmedian_family_income looks normally distributed with majority of the applications having their FFIEC Median family income between 50000 to 75000 dollars.
- Most application have their % of tract median family income compared to MSA/MD median family income between 75 to 100% and Percentage of minority population to total population for tract between 0 to 25%.

However, in a mortgage application, loan amount and the applicant's income is of great interest. But in the joint histogram plot in figure 3, it is difficult to decipher any information about them because of one or more outliers that may need to be removed. We present below a distribution plot of the two:

In Figure 4, we can say that a higher applicant income is associated with a higher loan amount, on average as there is a linear relationship between the two.

Figure 4: A regplot showing the relationship between Applicant's Income and Loan Amount

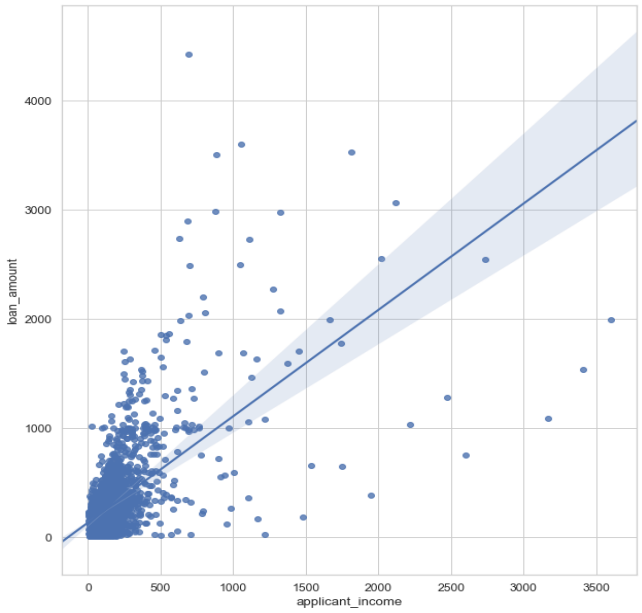
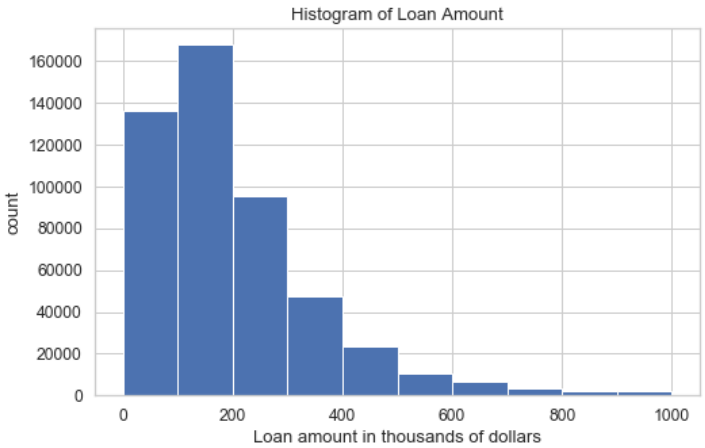
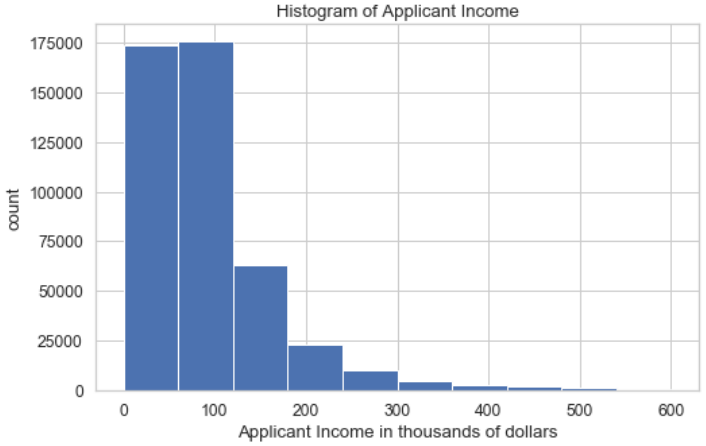


Figure 5: Distributions for Loan Amount and Applicant's Income



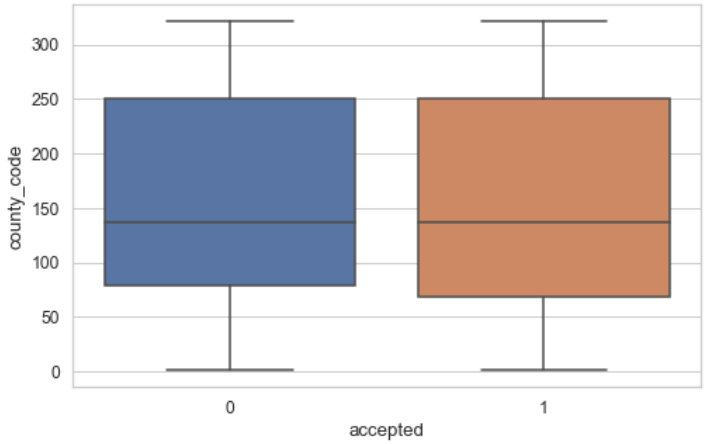


We can infer from the above plots in figure 5 that the amount of loan requested by most applications is below 400, 000 dollars, likewise the income of most applicants is between 0 to 100, 000 dollars. Very few applicants earn above 300,000 dollars.

Categorical Relationships

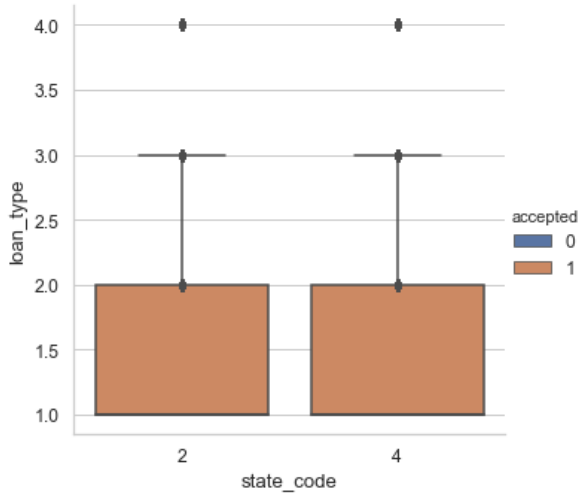
Having explored the relationship between the numeric features, an attempt was made to discern any apparent relationship between categorical feature values and mortgage application acceptance. The following box-plots show the categorical columns that seem to exhibit a relationship with the accepted feature.

Figure 6: Box-and-whisker plot showing relationship between county code and application accepted



In the above snippet, we can see that by limiting the loan acceptance to just state 48 and ignoring where country is missing, the average rate of loan acceptance across counties varies substantially, ranging from around 30% to around 70%.

Figure 7: Box-and-whisker plot showing relationship between state code and loan type



Also in the figure 7 above, for each of the four loan types, the loan acceptance rate in state 2 is lower than the corresponding loan type in state 4.

Classification of Mortgage Application Based on Acceptance

Based on the analysis of the mortgage application data, a predictive model to classify mortgage applications into two categories: application accepted (meaning the loan was originated) or denied was created. During the model development, the following were taken into consideration:

1. Dealing with missing values using the mean and median
2. Z-score normalization was done for the features with normal distribution while min-max normalization was done for features with irregular distribution.
3. Feature selection was done using chi-square feature scoring method

The model was created using the Two-Class Boosted Decision Trees algorithm and trained with 70% of the data. Testing the model with the remaining 30% of the data yielded the following results. Validation was further done with 100% of the test data provided.

True Positives: 58050

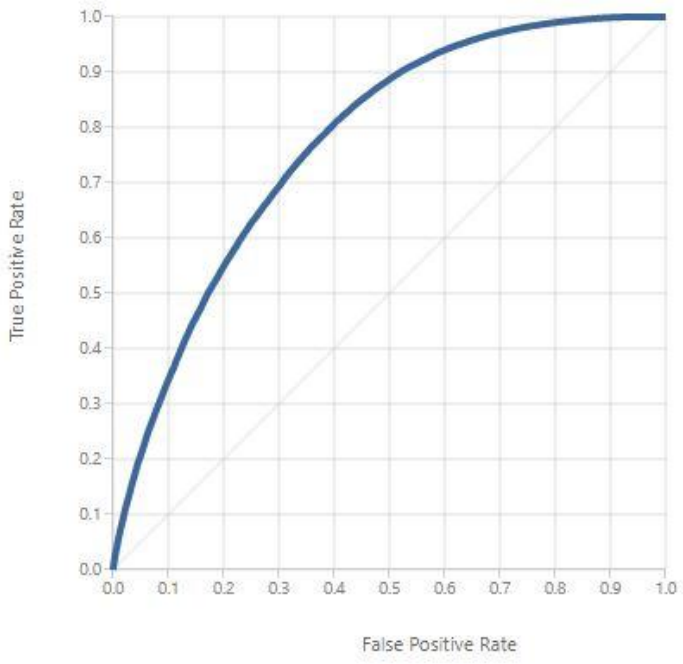
True Negatives: 47428

False Positives: 27274

False Negatives: 17248

The Receiver Operating Characteristic (ROC) curve for the model is shown here, with the blue line indicating the model's performance at varying classification threshold values, and the diagonal line showing the expected results of a random guess:

Figure 8: ROC Curve for mortgage approval classification model



This translates into the following standard performance metrics for classification:

- Accuracy: 71.2%
- Precision: 68%
- Recall: 77.1%
- F1 Score: 72.3%

Conclusion

This analysis has shown that the mortgage applications can be confidently predicted from its characteristics. In particular, the applicant's income, loan amount, loan purpose, and ffiecmedian_family_income has a significant effect on the acceptance or denial of a mortgage application.