Predicting Mortgage Approvals from Government Data

Pedro Galinha, April 2019

Executive Summary

This Document presents an analysis of data concerning government mortgage approvals. The analysis is based on 500000 observations of government mortgage approval data, each containing specific features of a mortgage approval loan.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between mortgage characteristics and their potential approvals were identified. After exploring the data, a predictive model to classify mortgages into two approval categories was created.

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

While many factors can help indicate the approval of a mortgage loan, the most significant features in the dataset used in this analysis were:

- **lender** - Indicates which of the lenders was the authority in approving or denying the loan.
- **applicant_income** - Income of the applicant in thousands of dollars.
- **loan_amount** – Size of the requested loan in thousands of dollars.
- **loan_purpose** – indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing.

Key Findings:

- the author theorizes that the income of the applicant influences the attribution of the loans. More loans were attributed to applicants with incomes lower than 250
- loans for lower amounts seem to have higher acceptance rate
- loans for Home purchase and Refinancing have higher acceptance rate
- the approval of loans is influenced by the lender authority
- The overall accuracy achieved by the model is sufficiently generalizable as to be reliable when deployed in Federal Financial Institutions

Initial Data Exploration

The initial exploration of the data began with some summary and descriptive statistics. The majority of features in the dataset are categorical, as the author is trying to predict acceptance or denial of loans, the relationship of categorical and numerical features with the label will be explored.

Feature Importance

Using the Permutation Feature Importance algorithm, it was possible to obtain the contribution of the features to the performance of the model in terms of how much a chosen evaluation metric deviates after permuting the values of that feature. After analyzing the scores, the features row_id and number_of_owner_occupied_units were removed.

Feature	Score
lender	0.09844
applicant_income	0.0526
loan_purpose	0.02634
loan_amount	0.0224
loan_type	0.0067
state_code	0.00524
applicant_race	0.00514
property_type	0.00432
tract_to_msa_md_income_pct	0.00402
applicant_ethnicity	0.00321
county_code	0.00288
preapproval	0.00272
ffiecmedian_family_income	0.00272
occupancy	0.00269
applicant_sex	0.00198
minority_population_pct	0.00148
co_applicant	0.00134
population	0.00107
number_of_1_to_4_family_units	0.00097
msa_md	0.0007
number_of_owner-occupied_units	0.0004
row_id	0.00029

By removing these two features the model improved in performance which allowed better accuracy in the predictions.

Cleaning Missing Values

The dataset has a significant number of missing values in categorical and numerical features. To correct this problem, the values missing from categorical features were calculated using a method described as “Multivariate Imputation by Chained Equations” and in the numerical features the missing values were replaced by the mean, which calculates the column mean and uses the mean as the replacement value for each missing value in the column.

columns
23

Feature	Count	Unique Value Count	Missing Value Count
row_id	500000	500000	0
loan_type	500000	4	0
property_type	500000	3	0
loan_purpose	500000	3	0
occupancy	500000	3	0
loan_amount	500000	2997	0
preapproval	500000	3	0
msa_md	423018	408	76982
state_code	480868	52	19132
county_code	479534	317	20466
applicant_ethnicity	500000	4	0
applicant_race	500000	7	0
applicant_sex	500000	4	0
applicant_income	460052	1897	39948
population	477535	18202	22465
minority_population_pct	477534	91923	22466
ffiecmedian_family_income	477560	68868	22440
tract_to_msa_md_income_pct	477486	54535	22514
number_of_owner-occupied_units	477435	6088	22565
number_of_1_to_4_family_units	477470	7374	22530
lender	500000	6111	0
co_applicant	500000	2	0
accepted	500000	2	0

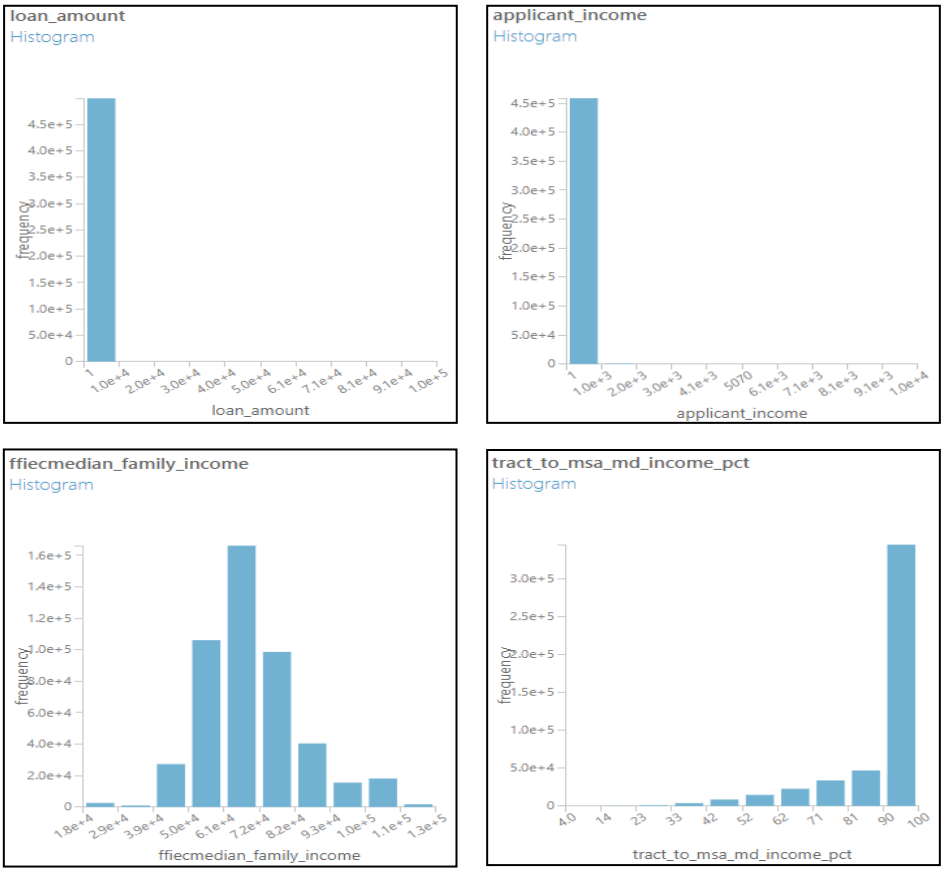
Using these two distinct methods for categorical and numeric columns, the dataset was fixed with the insertion of new values, preventing problems caused by missing data that could occur when training the model.

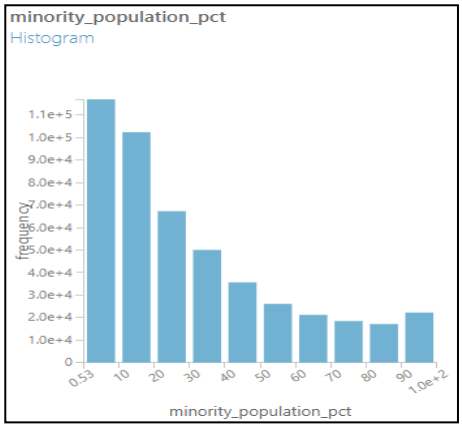
Individual Feature Statistics

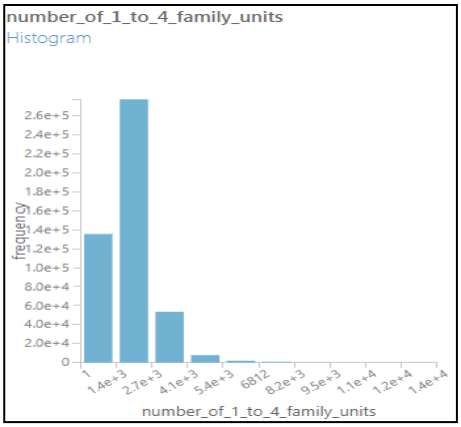
Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for numeric columns, and the results taken from 500000 observations are shown here:

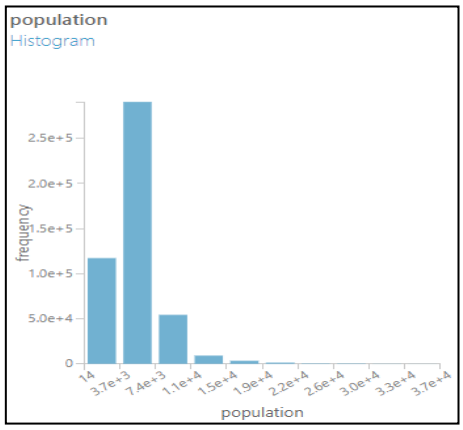
Column	Min	Max	Mean	Median	Std Dev	DCount
loan_amount	1	100878	221.753158	162	590.641648	500000
applicant_income	1	10139	102.389521	74	153.534496	460052
population	14	37097	5416.833956	4975	2728.144999	477535
minority_population_pct	0.534	100	31.617310	22.901	26.333938	477534
ffiecmedian_family_income	17858	125248	69235.603298	67526	14810.058791	477560
tract_to_msa_md_income_pct	3.981	100	91.832624	100	14.210924	477486
number_of_1_to_4_family_units	1	13623	1886.147065	1753	914.123744	477470

Since applicant_income and loan_amount revealed themselves to be important features for the analysis, it was noted that the mean and median of these values were significantly different and that the comparatively Max and Min values indicate the presence of outliers. The histogram of the applicant_income and loan_amount columns showed that the values were right-skewed, which indicated that the loan amounts and the applicant's incomes are in great majority at the lower end of the range of each feature.









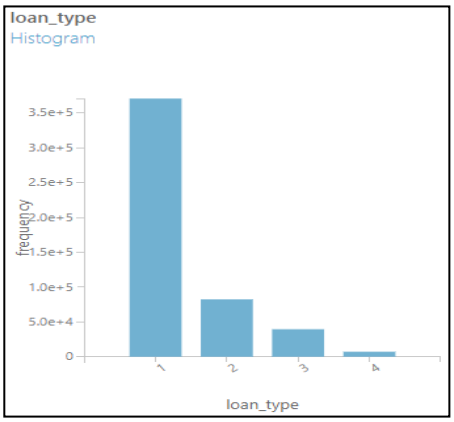
In addition to the numeric values, the mortgage approval observations also include categorical features including:

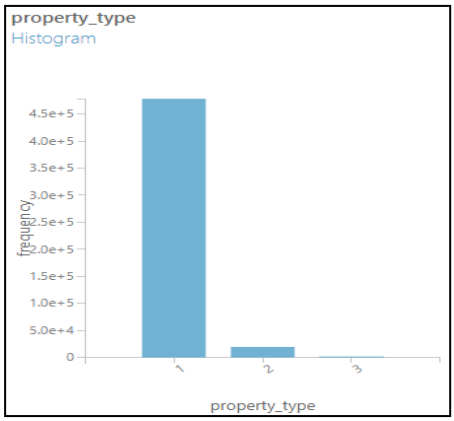
- **msa_md -** Metropolitan Statistical Area/Metropolitan Division (409 unique values)
- **state_code -** U.S. state (53 unique values)
- **county_code –** County (318 unique values)
- **lender -** Indicates which of the lenders was the authority in approving or denying the loan (6111 unique values)
- **loan_type -** Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured (4 unique values)
- **property_type -** Indicates whether the loan or application was for a one-to-four-family dwelling (other than manufactured housing), manufactured housing, or multifamily dwelling (3 unique values)
- **loan_purpose -** Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing (3 unique values)
- **occupancy -** Indicates whether the property to which the loan application relates will be the owner's principal dwelling (3 unique values)
- **preapproval -** Indicate whether the application or loan involved a request for a pre-approval of a home purchase loan (3 unique values)
- **applicant_ethnicity -** Ethnicity of the applicant (4 unique values)
- **applicant_race -** Race of the applicant (7 unique values)

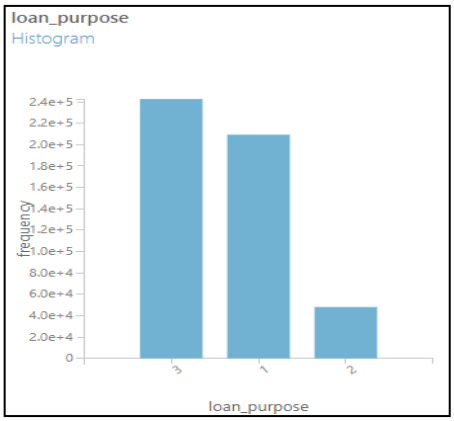
- **applicant_sex -** Sex of the applicant (4 unique values)
- **co_applicant -** Indicates whether there is a co-applicant (often a spouse) or not (Boolean type of data with 2 unique values)

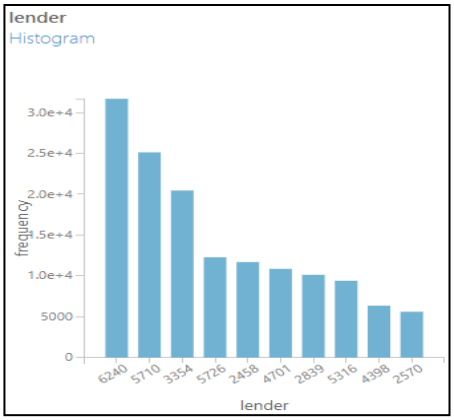
Histograms were created to show frequency of these features, and indicated the following:

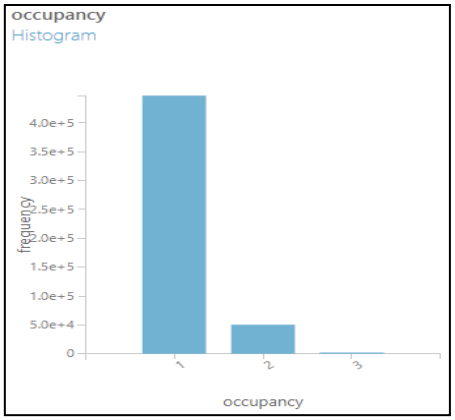
- The majority of the loans was of a Conventional type (any loan other than FHA, VA, FSA, or RHS loans) - type1
- The most common property type was for a one to four-family (other than manufactured housing) – type1
- The majority of the loans had the purpose of Refinancing and Home purchase – type3 and type1
- The most common lenders have ID´s > 5500
- The majority of the loans were for a owner-occupied property as a principal dwelling
- Most part of the loans did not involve a preapproval request – type 3 and type 2
- The majority of the applicants were white
- The majority of the applicants were from the male sex
- Most of the loans didn´t had a co-applicant

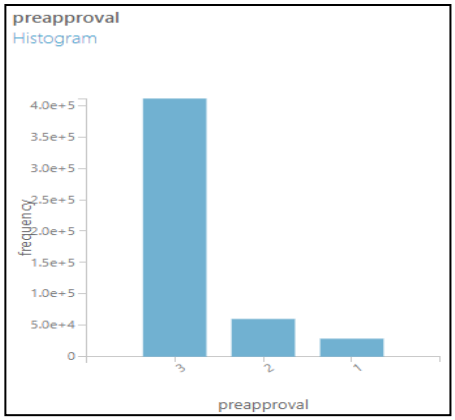


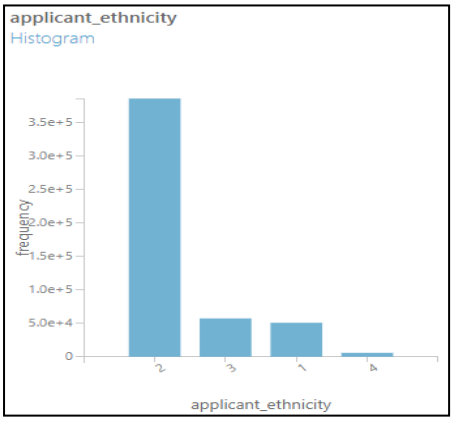


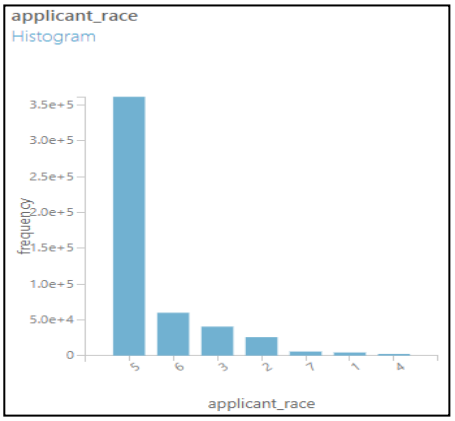


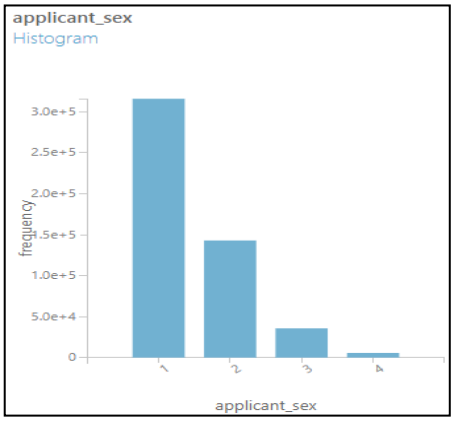

lender
Histogram
3.0e+4
2.5e+4
2.0e+4
1.5e+4
1.0e+4
5000
0
frequency
6240
5710
3354
5726
2458
4701
2839
5316
4398
2570
lender

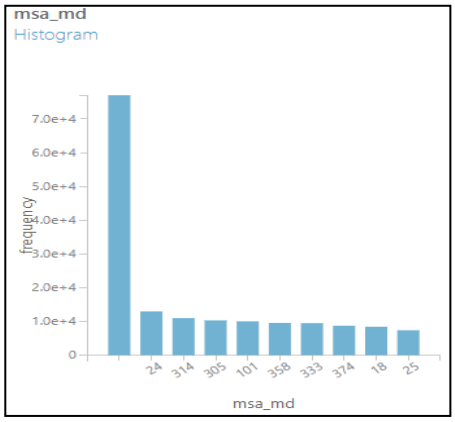

occupancy
Histogram
4.0e+5
3.5e+5
3.0e+5
2.5e+5
2.0e+5
1.5e+5
1.0e+5
5.0e+4
0
frequency
1
2
3
occupancy

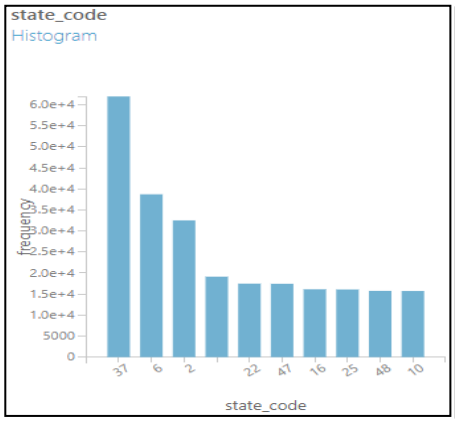

preapproval
Histogram
4.0e+5
3.5e+5
3.0e+5
2.5e+5
2.0e+5
1.5e+5
1.0e+5
5.0e+4
0
frequency
3
2
1
preapproval

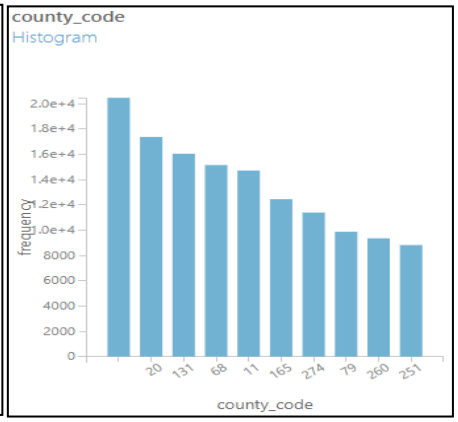

applicant_ethnicity
Histogram
3.5e+5
3.0e+5
2.5e+5
2.0e+5
1.5e+5
1.0e+5
5.0e+4
0
frequency
2
3
1
4
applicant_ethnicity

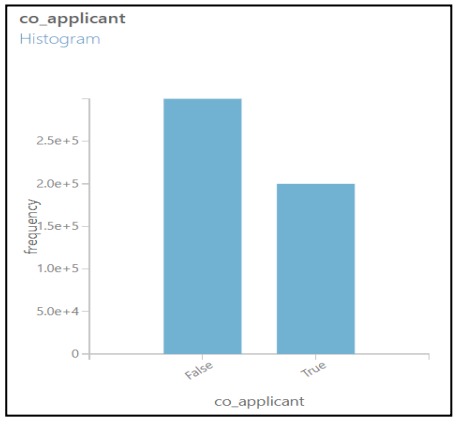

applicant_race
Histogram
3.5e+5
3.0e+5
2.5e+5
2.0e+5
1.5e+5
1.0e+5
5.0e+4
0
frequency
5
6
3
2
7
1
4
applicant_race


applicant_sex
Histogram
3.0e+5
2.5e+5
2.0e+5
1.5e+5
1.0e+5
5.0e+4
0
frequency
1
2
3
4
applicant_sex


msa_md
Histogram
7.0e+4
6.0e+4
5.0e+4
4.0e+4
3.0e+4
2.0e+4
1.0e+4
0
frequency
24
314
305
101
358
333
374
18
25
msa_md


state_code
Histogram
6.0e+4
5.5e+4
5.0e+4
4.5e+4
4.0e+4
3.5e+4
3.0e+4
2.5e+4
2.0e+4
1.5e+4
1.0e+4
5000
0
frequency
37
6
2
22
47
16
25
48
10
state_code


county_code
Histogram
2.0e+4
1.8e+4
1.6e+4
1.4e+4
1.2e+4
1.0e+4
8000
6000
4000
2000
0
frequency
20
131
68
11
165
274
79
260
251
county_code


co_applicant
Histogram
2.5e+5
2.0e+5
1.5e+5
1.0e+5
5.0e+4
0
frequency
False
True
co_applicant

Correlation and Apparent Relationships

After exploring the individual features, an attempt was made to identify relationships between features in the data and the label feature (accepted).

Numeric Relationships

The correlation between the numeric columns and the label column was calculated with the following results:

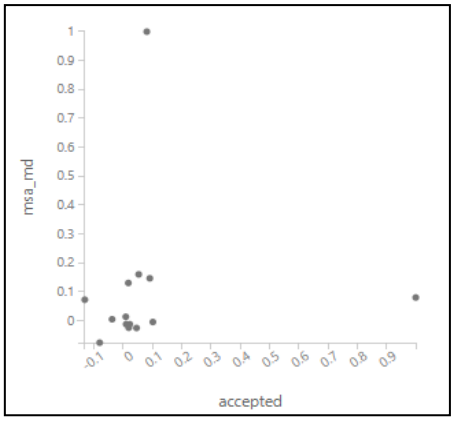
	loan_amount	applicant_income	population	minority_population_pct	ffiecmedian_family_income	tract_to_msa_md_income_pct	number_of_1_to_4_family_units	accepted
loan_amount	**1.000000**	0.483951	0.000100	0.007227	0.105924	0.043811	-0.036644	0.046370
applicant_income	0.483951	**1.000000**	-0.006948	-0.053795	0.114988	0.102667	-0.019748	0.074722
population	0.000100	-0.006948	**1.000000**	0.087383	-0.014377	0.149677	0.816952	0.019163
minority_population_pct	0.007227	-0.053795	0.087383	**1.000000**	0.021059	-0.442800	-0.157976	-0.092922
ffiecmedian_family_income	0.105924	0.114988	-0.014377	0.021059	**1.000000**	-0.054500	-0.148235	0.066919
tract_to_msa_md_income_pct	0.043811	0.102667	0.149677	-0.442800	-0.054500	**1.000000**	0.210613	0.091766
number_of_1_to_4_family_units	-0.036644	-0.019748	0.816952	-0.157976	-0.148235	0.210613	**1.000000**	0.006027
accepted	0.046370	0.074722	0.019163	-0.092922	0.066919	0.091766	0.006027	**1.000000**

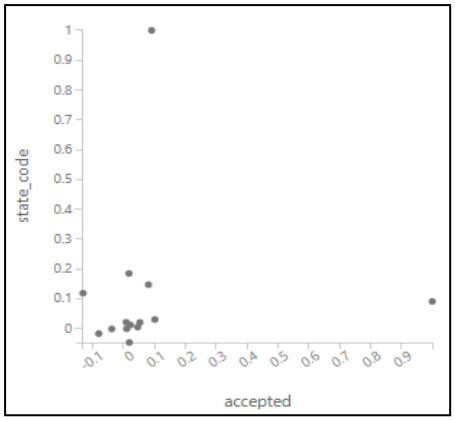
These correlations showed the existence of:

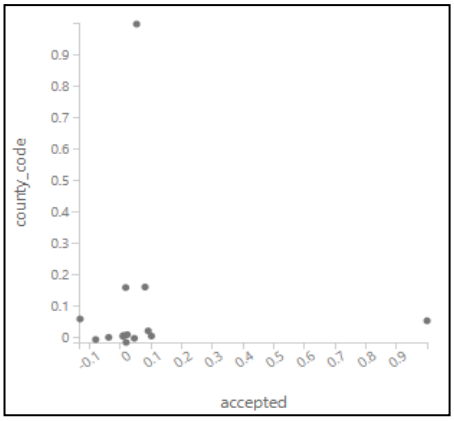
- a moderate negative correlation between minority_population_pct and track_to_msa_md_income_pct
- a moderate positive correlation between loan_amount and applicant_income
- a strong positive correlation between population and number_of_1_to_4_family_units

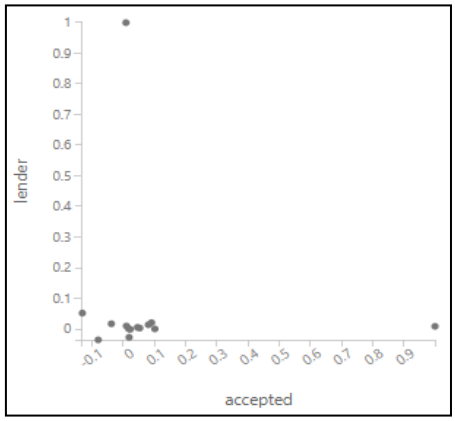
Categorical Relationships

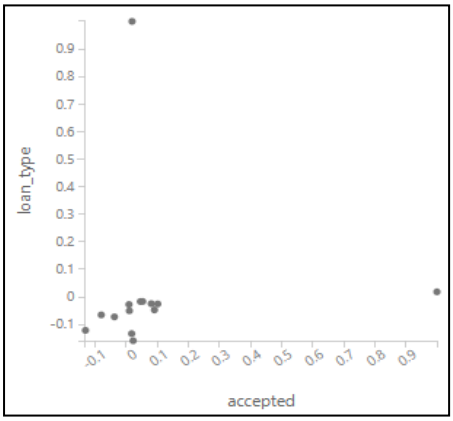
Having explored the relationship between the label and numeric features, an attempt was made to discern any apparent relationship between categorical feature values and the label (accepted). The following scatter-plots show the categorical columns and their correlation with the label feature:

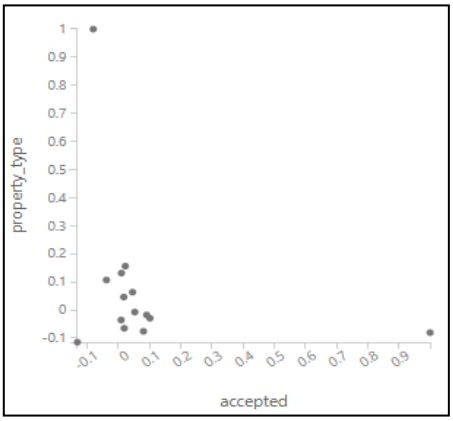


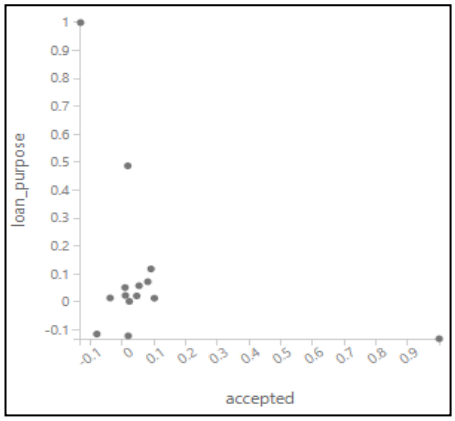


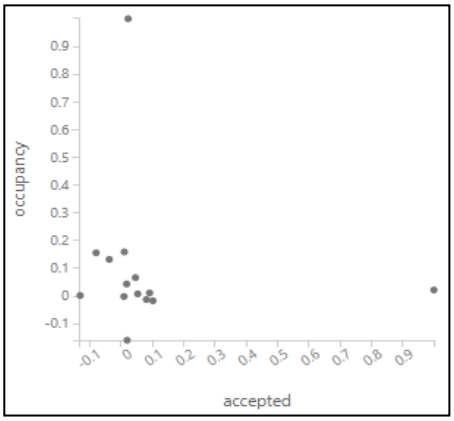


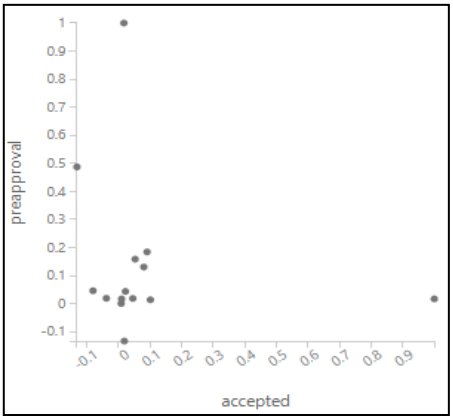


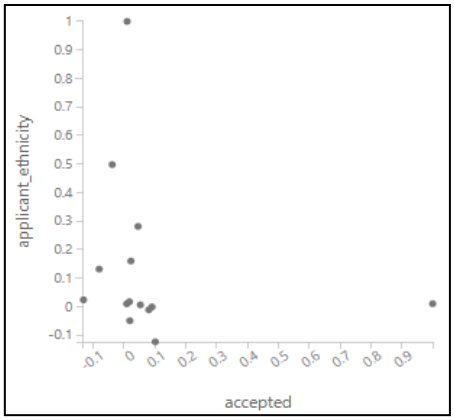


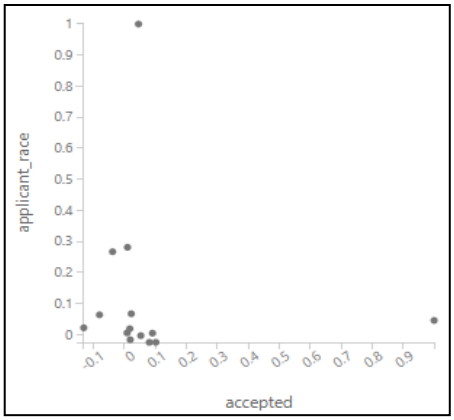


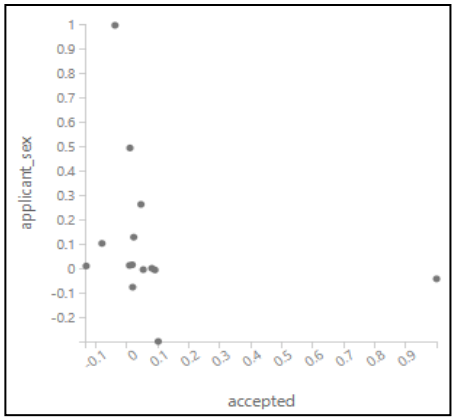


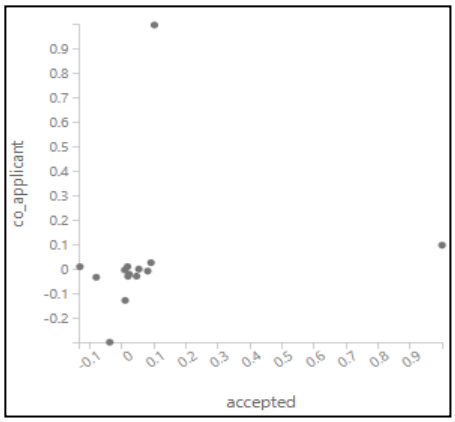




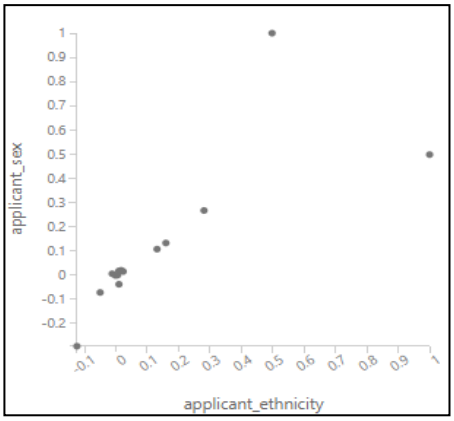


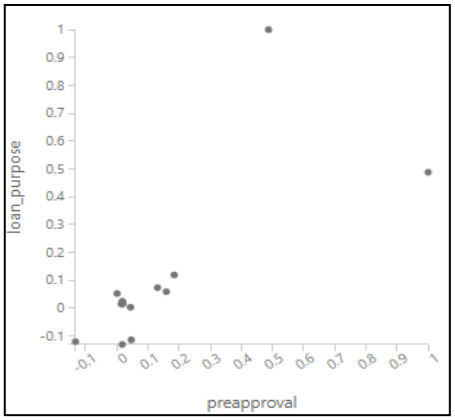






Although the plots failed to show a clear linear relationship between the label and the categorical columns, it was possible to find moderate positive correlation between the feature's applicant_ethnicity and applicant_sex and also between loan_purpose and preapproval features.



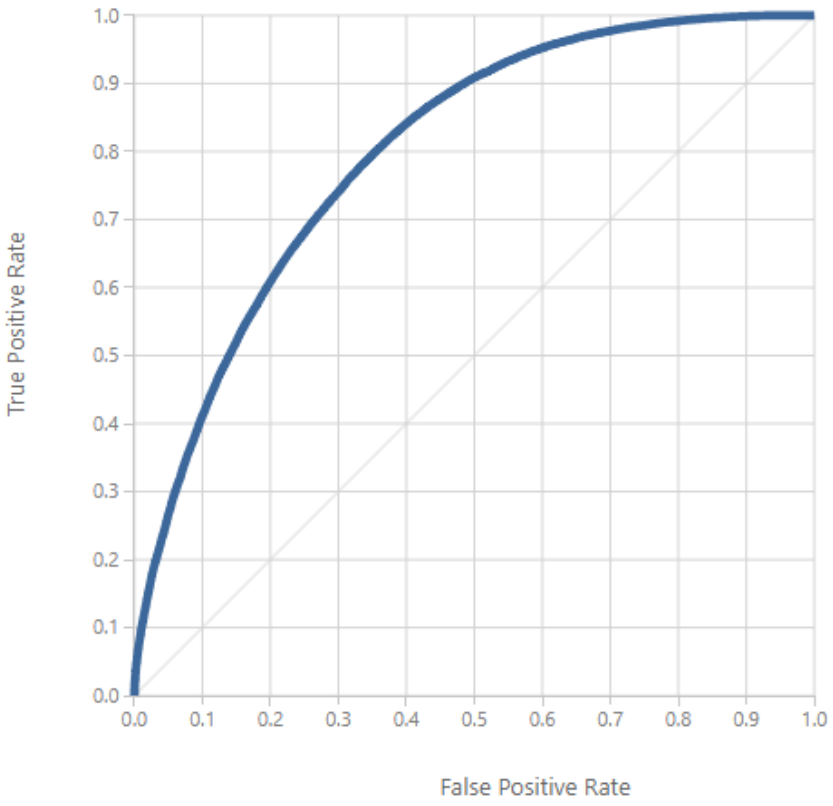


Classification of Mortgage Approvals

Based on the analysis of the mortgage approval data, a predictive model was created.
The model was created using the Two-Class Boosted Decision Trees algorithm and trained with 80% of the data. Testing the model with the remaining 20% of the data yielded the following results:

- True Positives: 38731
- True Negatives: 33530
- False Positives: 16244
- False Negatives: 11495

The Received Operator 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:



This translates in to the following standard performance metrics for classification:

- Accuracy: 72,3%
- Precision: 70,5%
- Recall: 77,1%
- F1 Score: 73,6%

Conclusion

This analysis has shown that the model used is able to predict with good accuracy whether a loan is approved or denied. In this project, several models were used (two-class boosted decision tree, two-class average perceptron, two-class Bayes point machine, two-class decision forest, two-class decision jungle, two-class locally-deep support vector machine, two-class logistic regression, two-class neural network and the two-class support vector machine). The two-class boosted decision tree was the one that showed better accuracy when compared to the other models used.

The most predictive features are lender, applicant income, loan amount and loan purpose. Secondary features, such as loan type, applicant race and property type can help further classify the approval or rejection of a loan.

In addition, feature engineering possibilities exist if more domain knowledge is available for analysis. This would require a dataset with more quality data, categorical and numerical features with less missing data that would allow a better knowledge and a better performance predicting in a classification model.