Report: Prediction of Mortgage Applications Approval From Government

July,2019

Executive Summary

This document presents analysis of data collected from mortgage applicants. This analysis is performed on distinct entries of 500,000 applicants primarily to determine whether a mortgage application was accepted (meaning the loan was originated) or denied according to the given dataset, which is adapted from the Federal Financial Institutions Examination Council's (FFIEC).

The analysis for the given data was approached in three ways. Firstly, descriptive statistics was calculated for the numerical features in the data. Secondly, data visualizations were created to explore, understand and infer the relationship of the independent features and the target variable. Lastly, a classification model was built with the features selection option to predict the possibility of a customer's mortgage application being accepted. The model's hyper-parameters were also tuned to optimize the model.

The following conclusions were drawn from the analysis:

- Several algorithms such as Random forest, XGBoost and decision trees classification models were performed but the prediction accuracies were low. The Catboost Model that achieved an accuracy of 73.25% was successfully performed.
- A feature importance plot showed that the 5 most important features are: county_code, loan_purpose, state_code, Ffiecmedian_faimily_ income, and minority-population_pct. Categorical features accounted for the most important four features out of the first five features.

Given Data Insight

The provided data are training data, test data and a training label. The training and test data had 21 features ranging from:

- Property Location data
- Loan information data
- Applicant Information
- Census information
- Index and Target Variable

An immediate observation shows that the datasets contain more categorical features (13 features) than numerical features (7 features). It was also noticed that all the categorical data were presented in a nominal categorical format (numerical format).

In addition, there were some null values that needed to be handled in the preprocessing stage.

Data Preparation, Data Cleaning and Data Exploration

The summary statistics for the numeric features was calculated and the results shown in the table below.

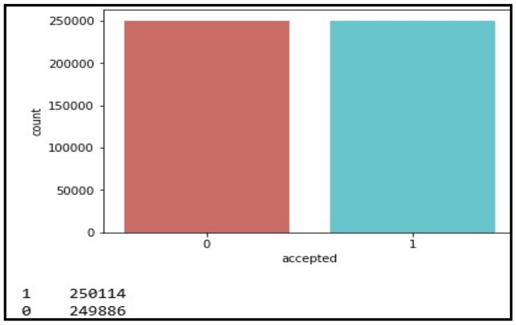
Columns	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Loan_amount	500000	221.75	590.64	1	93.00	162.00	266.00	100878.0
Applicant_income	460052	102.39	153.53	1.00	47.00	74.00	117.00	10139.0
Minority_population_pct	477534	31.62	26.33	0.534	10.70	22.90	46.02	100.00
Poplation	477535	5416.83	2728.14	14.00	3744.00	4975.00	6467.00	37097.00
Ffiecmedian_faimily_income	477560	69235.6	14810.0	17858.0	59731.0	67526.0	75351.0	125248.0
Tract_to_msa_md_Income_pct	477486	91.83	14.21	3.98	88.07	100.00	100.00	100.00
Number_of_owner_occupied_units	477435	1427.72	737.56	4.00	944.00	1327.00	178000	8771.00
Number_of_1_to_4_family_units	477470	1886.15	914.12	1.00	1301.00	1753.00	2309.00	13623.0

In addition, the following data preprocessing was done:

- The numeric null values were replaced with zeroes.
- The categorical data null values represented as -1 in the data as replaced with -999 for optimum performance of the catboost model.

Since the project was a classification problem, it is important to check for class imbalance. In case, class imbalance was observed, a class balancing procedure has to be taken. The two classes were plotted and it was observed that the two classes were very close in count, hence no need for class imbalance processing of the data.

Class Imbalance check



```
1    250114
0    249886
```

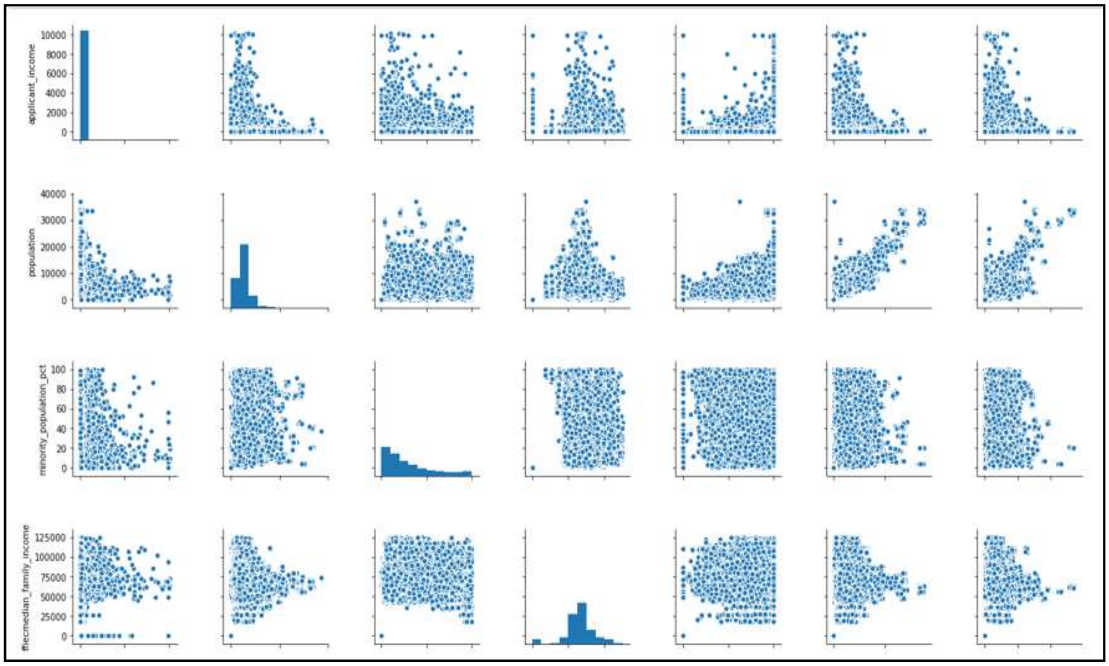
percentage of denied applications 49.98

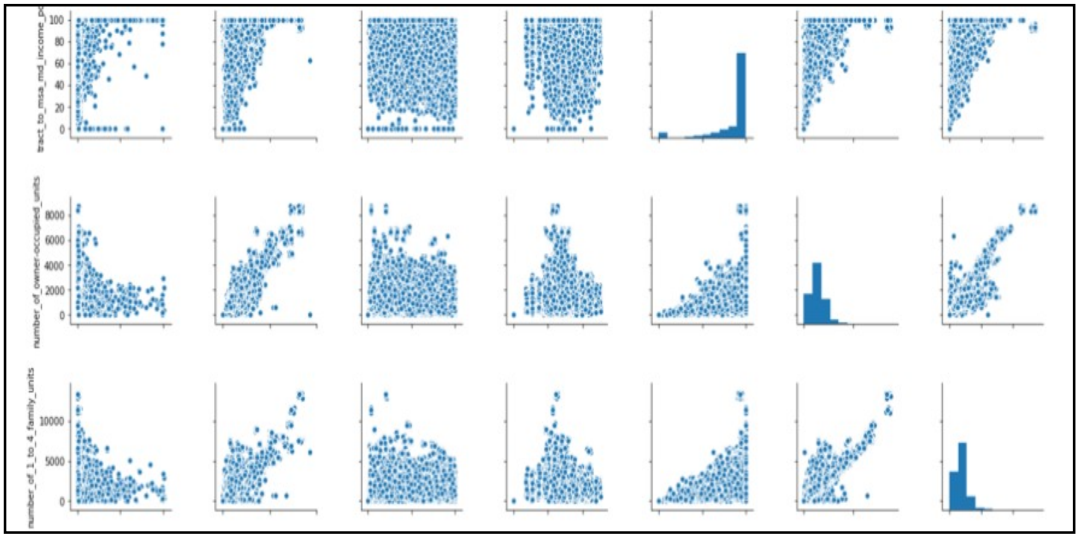
percentage of accepted applications 50.02

Numeric Features Analysis

Relationships and correlations

A scatter-plot matrix was generated to compare numeric features with one another and the target column ‘accepted’.





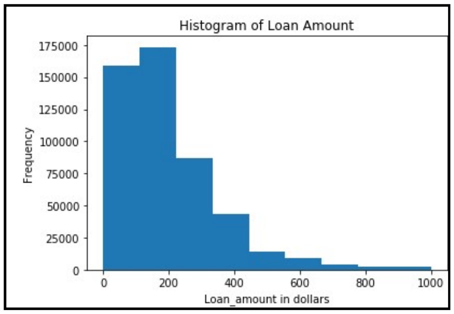
The plots in the bottom row or the right-most column of this matrix shows the apparent relationship Between the accepted column and the numeric features. Specifically, the population and number of owner occupied units columns increase shows a positive correlation with the accepted column. No distinct relationship as detected with the other numeric columns, hence a correlation table was created to give a better presentation.

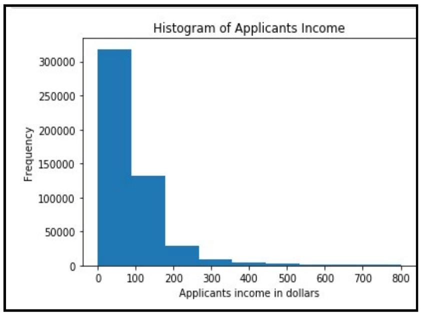
The correlation between the numeric columns was then calculated with the following results:

	applicant_ income	population	minority_ population_ pct	Ffiecmedian family_ income	tract_to_msa _md_income_ pct	number_of_ owneroccupied _units	number_of1t 4_family _units	accepted
applicant_ income	1	0	-0.05	0.09	0.08	0.02	-0.01	0.07
population	0	1	0.17	0.26	0.39	0.88	0.85	0.09
minority_popula tion_pct	-0.05	0.17	1	0.19	-0.05	-0.1	-0.04	-0.04
ffiecmedian_fam ily_income	0.09	0.26	0.19	1	0.54	0.25	0.19	0.19
tract_to_msa_ md_income_pct	0.08	0.39	-0.05	0.54	1	0.5	0.44	0.21
number_of_ owner-occupied _units	0.02	0.88	-0.1	0.25	0.5	1	0.9	0.11
number_of_1_4_ family_units	-0.01	0.85	-0.04	0.19	0.44	0.9	1	0.09
accepted	0.07	0.09	-0.04	0.19	0.21	0.11	0.09	1

These correlations supports the plots by showing a negative correlation between minority_population_pct and accepted, and moderate to strong positive correlations for the other numeric features.

Histogram plots of Loan Amount and Applicants' Income.

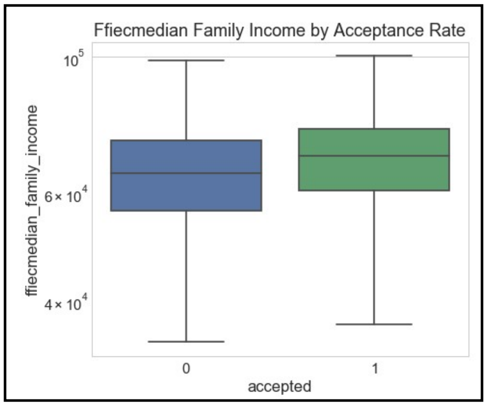


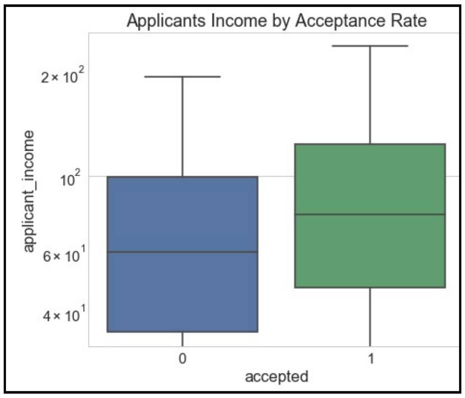


The histogram plots shows that majority of the applicants applied for loans below 800 dollars. Apparently, most applicants' monthly income is below 300 dollars.

Box Plot

Box plots were also used to show the relationship between accepted and Applicants' Income Applicant Income and ffiecmedian_family_income are two determinant columns on whether a loan is granted and how much loan is granted.



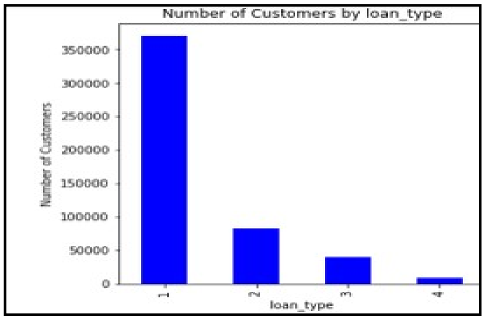


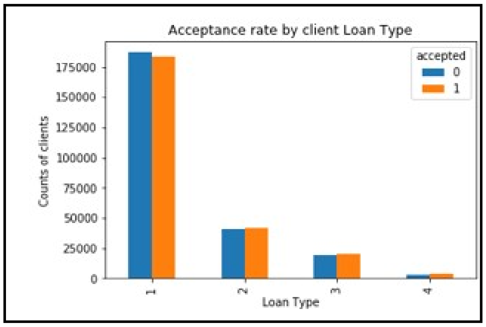
The two plots above show the 5 summary statistics of the box plot(min, 25th percentile, median, 75th percentile and max) of the applicants based on Ffiecmedian family income and applicants income. It can be observed that the income of those granted loan is higher than those rejected. Hence, they are both important numeric features in this project.

Categorical Features Analysis

The relationship between target column ‘accepted’ and categorical features was also explored. The following bar plots show the relationship between the categorical columns and the accepted column.

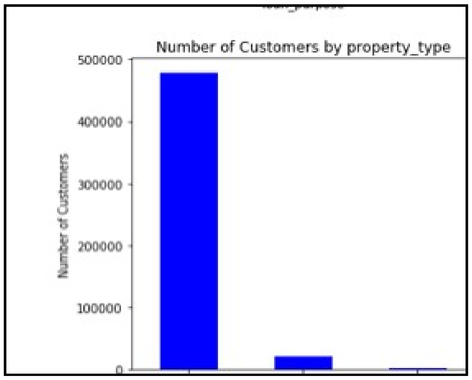
Loan Type

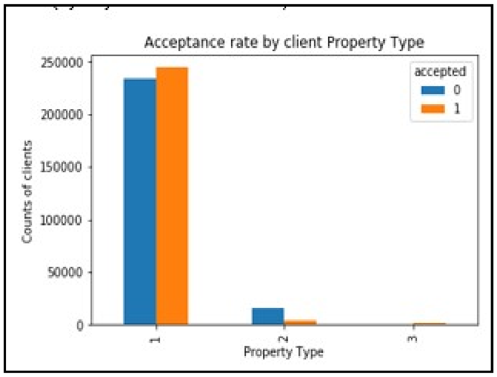




Loan_type indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured. The first bar plot shows that loan type 1 which is the Conventional type is the highest applied loan type. The second bar plot further shows that denied applications were slightly higher than approved applications for the loan type1.

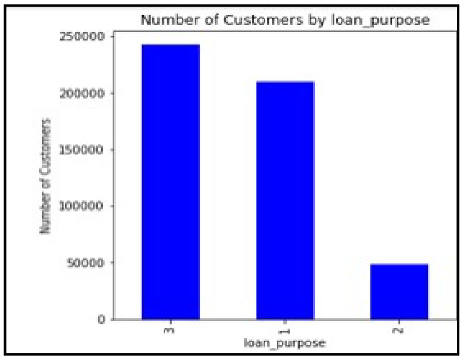
Property Type

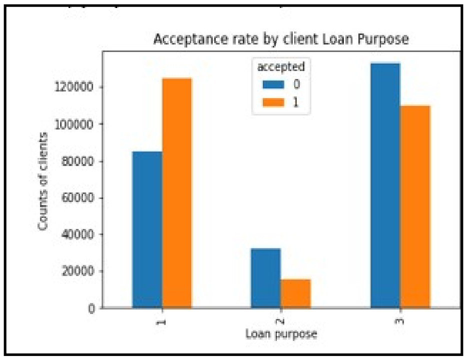




Property_type indicates whether the loan or application was for a one-to-four-family dwelling, manufactured housing, or multifamily dwelling. The first bar plot shows that the property type one(One to four-family) had the highest count of applicants while the second plot shows that more applications were approved for the type one .

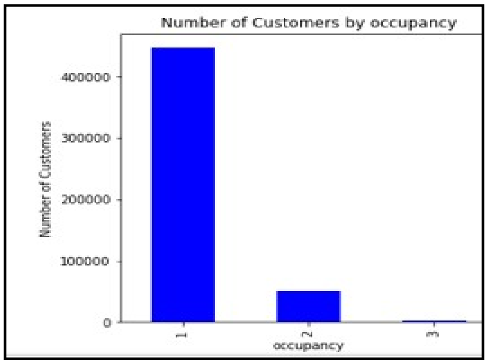
Loan Purpose

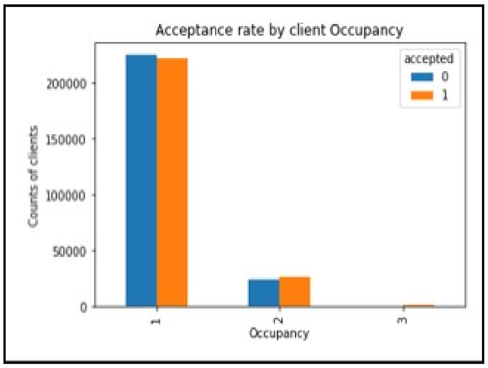




Loan_purpose indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing. The first plot shows that loan purpose group three (refinancing) had the highest number of applicants. The second plot shows that more applicants were denied for loan purpose group three while more were accepted for loan purpose group one.

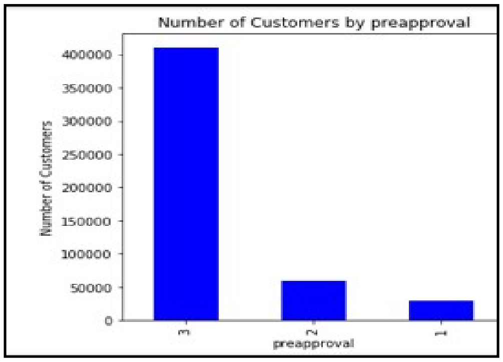
Occupancy

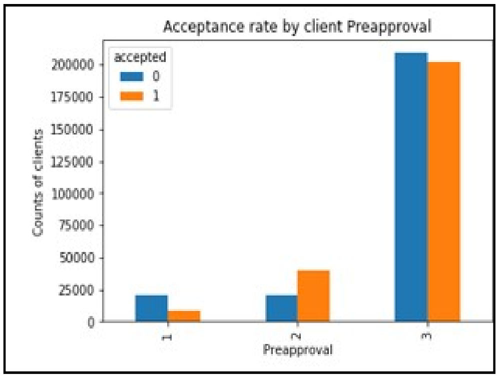




Occupancy indicates whether the property to which the loan application relates will be the owner's principal dwelling or not. The first plot shows that most applicants fall under the occupancy group one (Owner-occupied as a principal dwelling). The second plot shows the counts of applications, denied and accepted applications for the three groups.

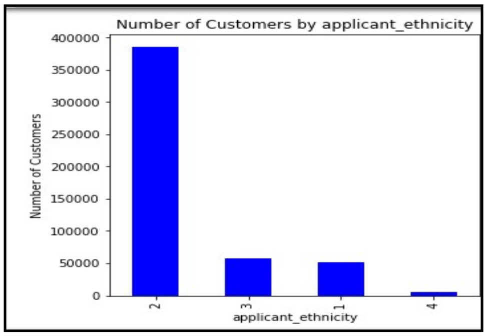
Preapproval

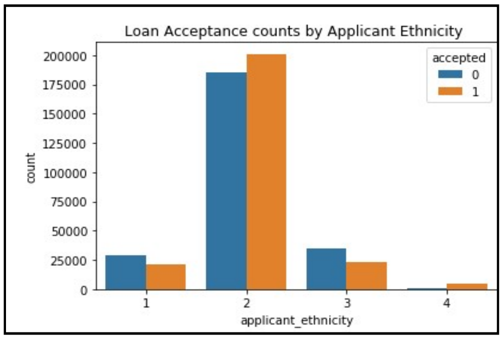




Preapproval indicates whether the application or loan involved a request for a pre-approval of a home purchase loan. Preapproval group three (not applicable) clearly has the highest number of applicants than one and two. This means most applications don't require the preapproval process. The second plot shows the counts of applications denied and accepted for the three groups.

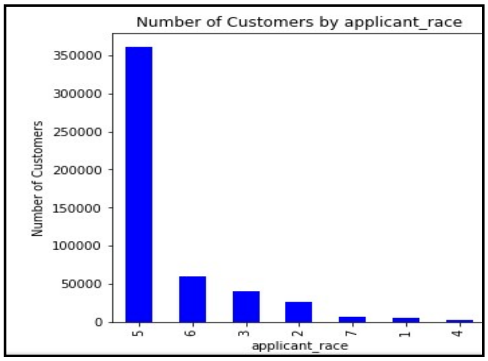
Applicant Ethnicity

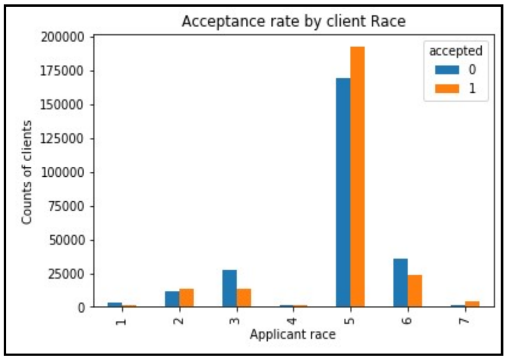




Applicant ethnicity group two (not Hispanic or Latino) accounts for the highest number of applicants. The second plot shows the counts of applications denied and accepted for the four ethnic groups.

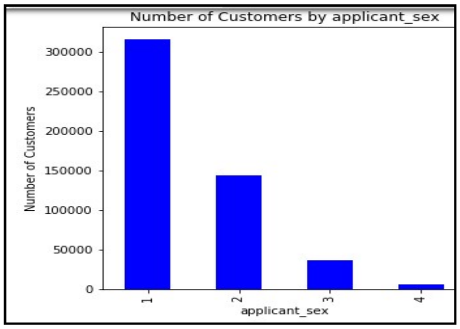
Applicant Race

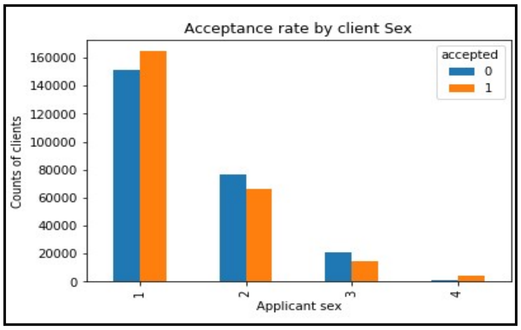




Most applicants were of race group 5(white). This also supports the applicant ethnicity explored above that most applicants were neither latino or Hispanic. The second plot further shows the counts of applications denied and accepted for the seven groups.

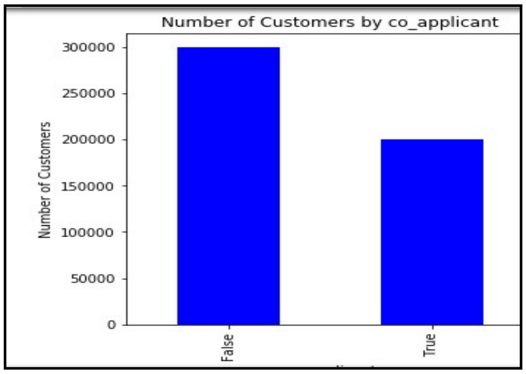
Applicant Sex

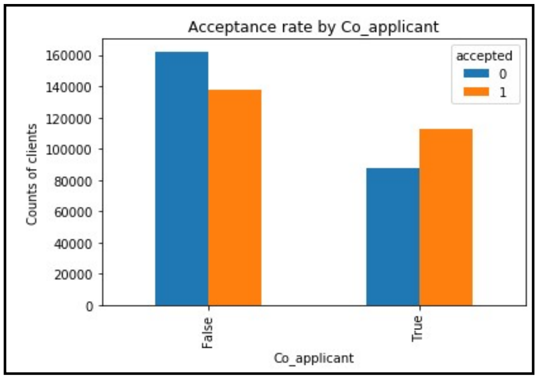




Most applicants were of sex group one(male) followed by the sex group two(female). The second plot shows the counts of applications denied and accepted for the four sex groups. More applications were accepted in the first group (males), unlike the groups two and three.

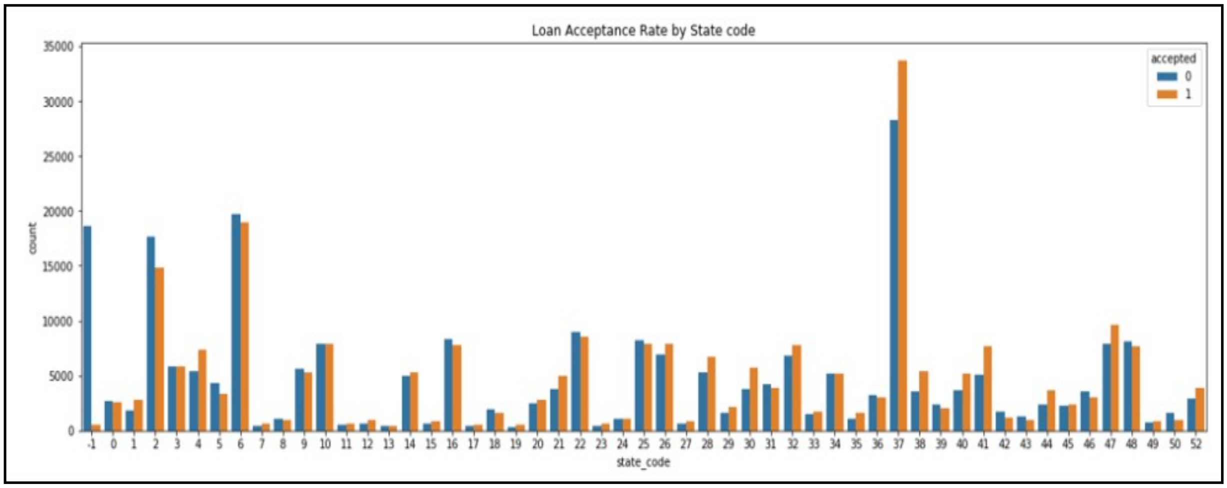
Co_Applicant





It was noticed that most applicants applied alone as the counts of those without co_applicants is more than those with co_applicants. The second plot shows the counts of applications denied and accepted for the two co_applicant groups.

State Code

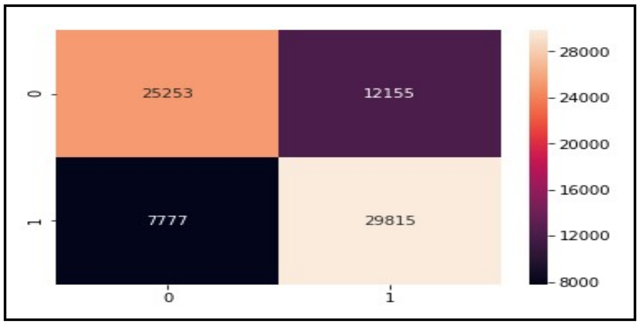


The plot shows the distribution of how applications were denied and accepted in all 52 states. For the property location data, only the state code was plotted because it had 52 categories. The two other columns were not plotted because of their numerous categories. The Msa_md column has 408 categories while county_code has 324 categories.

Classification Model

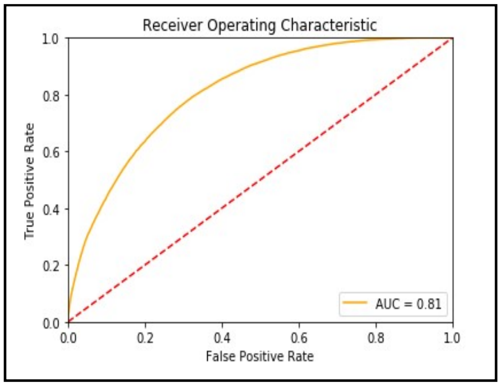
The numeric features account for exactly one-third(7 features), of the total features as opposed 13 categorical features. For this reason, a model that performs well with many numeric features could not give the optimal classification result. Hence, a model that maximizes and leverages its potential with many categorical features was preferred, the 'catboost' algorithm was used to build the classification model. Feature selection option was enabled to allow the Catboost model correctly select the categorical features for its prediction.

The model was trained with 85% of the data while the remaining 15% was used to test the model. The following confusion matrix and results were obtained from the model. :



ROC Curve

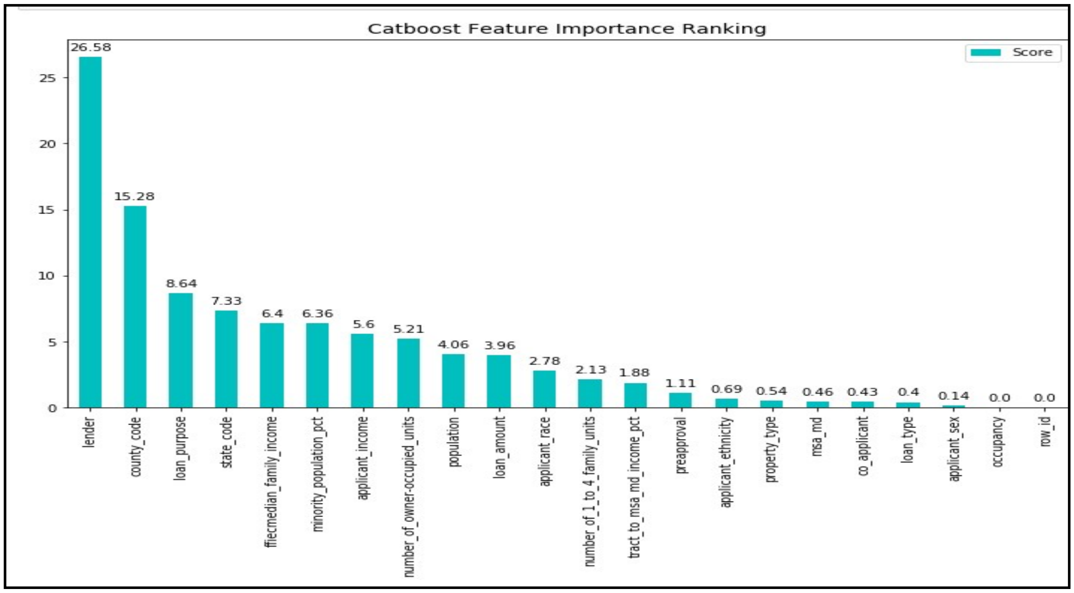
The Received Operator Characteristic (ROC) curve for the model is shown below, with the yellow line indicating the model's performance at varying classification threshold values, and the diagonal line showing the expected results of a random guess:



Statistical summary:

Performance Metrics	Value
True Positive	29815
True Negative	25253
False positive	12155
False Negative	7777
Accuracy	0.73
Precision	0.74
Recall	0.73
F1 Score	0.74
Area Under Curve	0.81

Feature Importance Plot



The feature importance plot simply shows the contribution strength of each feature to the prediction of the target variable(accepted). It is clearly seen that lender showed the higher contribution to the model.

Conclusion

This analysis has demonstrated that acceptance or denial of a mortgage application can be Confidently predicted from different applicants features or characteristics. Specifically, the county_code, loan_purpose, state_code, ffiecmedian_family_income,minority-population_pct features have a significant effect on the acceptance of mortgage application. For the sake of this project the Catboost model excelled in performance than other tried models. Hence, categorical features are very important features when considering mortgage loan acceptance for the sake of this data set.