Analysis of Mortgage Loan Approval

Frederick Emile Bondzie-Arthur 2018

Executive Summary

This document presents an analysis of data concerning house loan applications and if they were accepted or not. The analysis was based on 500000 observations of loan application data, with each including specific features of the application and whether it was accepted or not.

After cleaning and performing exploratory analysis such as summary and descriptive statistics, several possible relationships was identified between the loan application features and whether it was accepted or not.

After exploratory analysis of the data, a classification model to classify loan applications into two categories, accepted or not accepted was created from the features and other features created.

After performing the analysis, the following conclusion were made:

Whiles many factors contributes to whether a house loan application is accepted or not, significant features found in the analysis were:

- Lender - the lending institution in-charge of approving or denying the loan of the applicant. Approval for some lenders tends to be more than others.
- Applicant Income - the income of the loan applicant

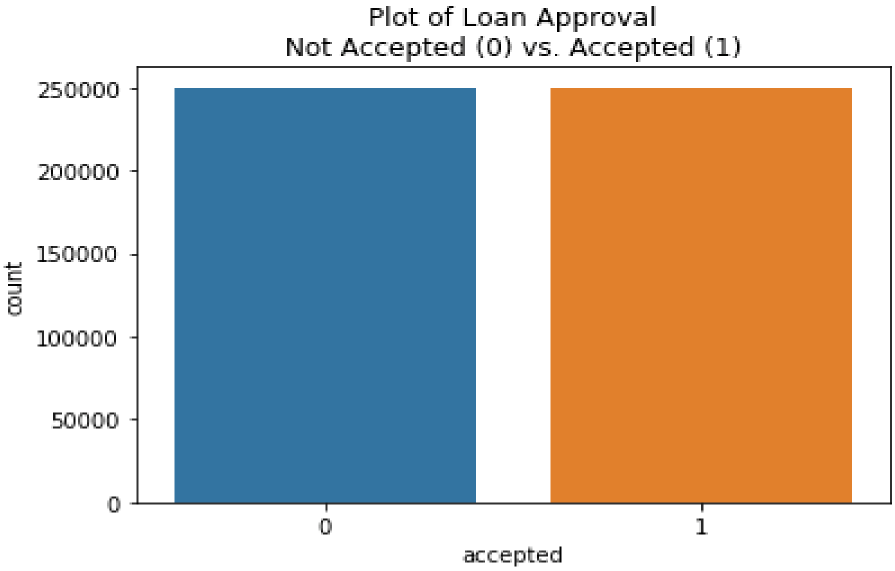
- County Code – a categorical data indicating the U.S. state the applicant is in.
- Loan Amount – the amount of loan the applicant requested. It’s in thousands of dollars.
- Loan purpose - Indicates the purpose of the loan. There are three categories. These are one-to-four-family (other than manufactured housing), manufactured housing and multifamily.

Exploratory Data Analysis

Performing some exploratory analysis, we begin with summary and descriptive statistics numeric columns.

Column	Min	Max	Mean	Median	Std Dev	DCount
Loan Amount	1	1000878	221.75	162.00	590.65	2997
Applicant Income	1	10139	102.39	74.00	153.53	1897
Population	14	37097	5416.83	4975.00	2728.14	18202
Minority Population PCT	0.53	100	31.62	22.90	26.33	91923
Ffiecmedian Family Income	17858	125248	69235.60	67526.00	14810.06	688868
Tract to MSA MD INCOME PCT	3.98	100	91.83	100	14.21	54535
Number of Owner Occupied Units	4.0	8771	1427.72	1327	737.56	6088
Number of 1 to 4 family units	1.0	13623	1886.15	1753	914.12	7374

In this analysis accepted (i.e. loan was approved or denied) is of interest, due to this we plot a chart to view the count of mortgage loans approved or denied.



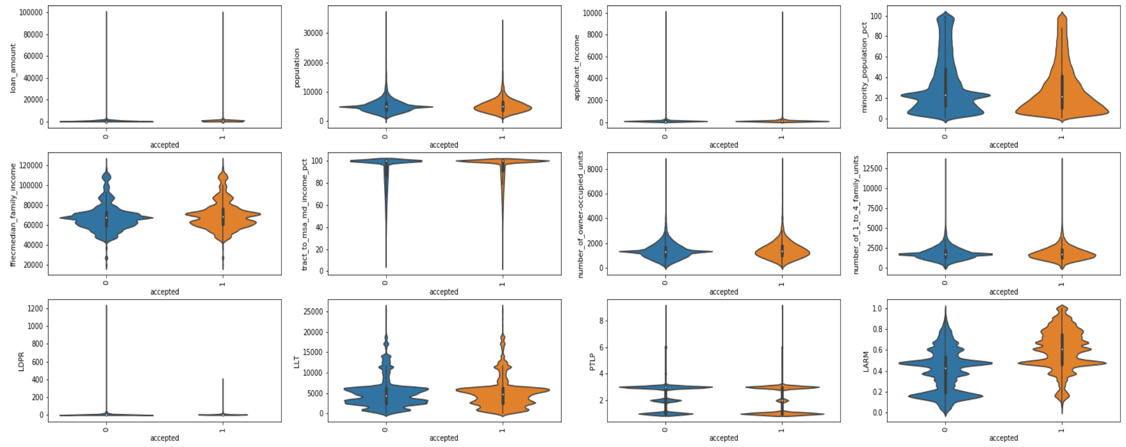
From the chart, it can be seen that the frequency of loan accepted is almost the same as the frequency of loans declined. In addition to numeric values, the loan application data include categorical variables such as:

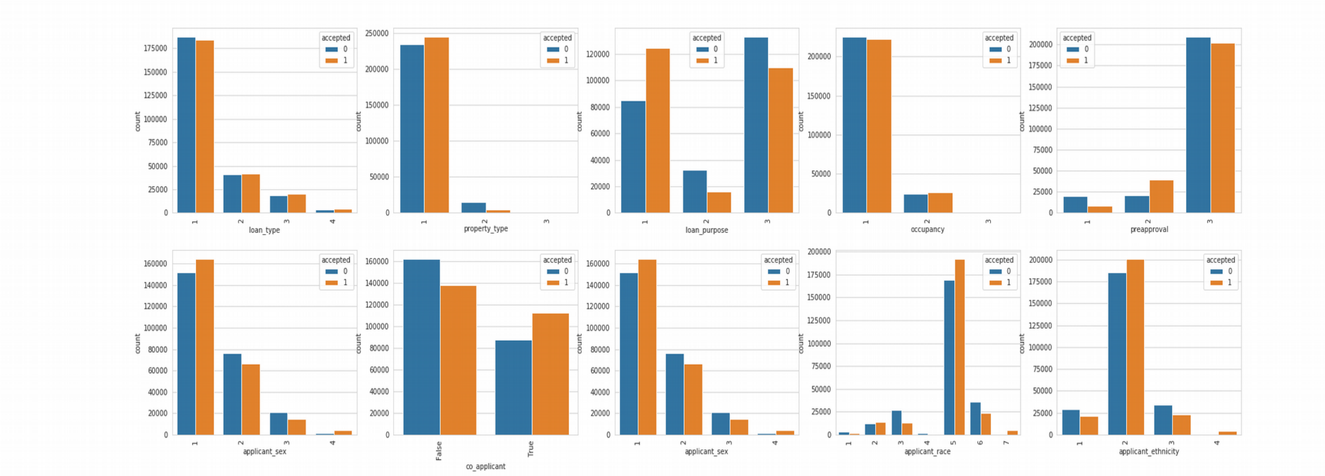
- Loan Type
- Property Type
- Loan Purpose
- Occupation
- Pre-approval
- Metropolitan Statistical/Metropolitan Division
- State Code
- County Code
- Applicant Ethnicity
- Applicant Race
- Applicant Sex
- Lender
- Co-Applicant

Bar charts were created to show frequency of these features and the following was established.:

- Conventional loans were applied more than other loan types.

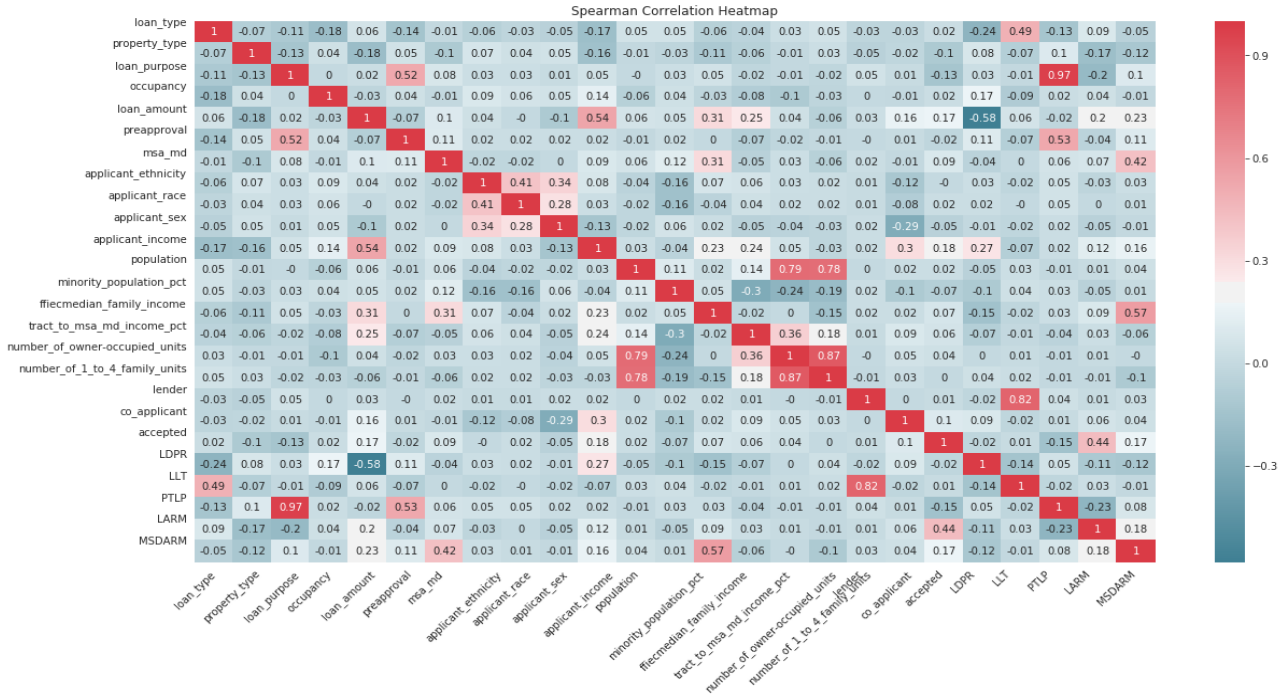
- Most applicants choose One to four-family properties than manufactured housing properties and Multifamily properties.
- Most applicant took loans for refinancing than for home purchasing and home improvement.
- Most applicants owner occupied as a principal dwelling
- The majority of applicant applying for mortgage loan are whites compared to other races.

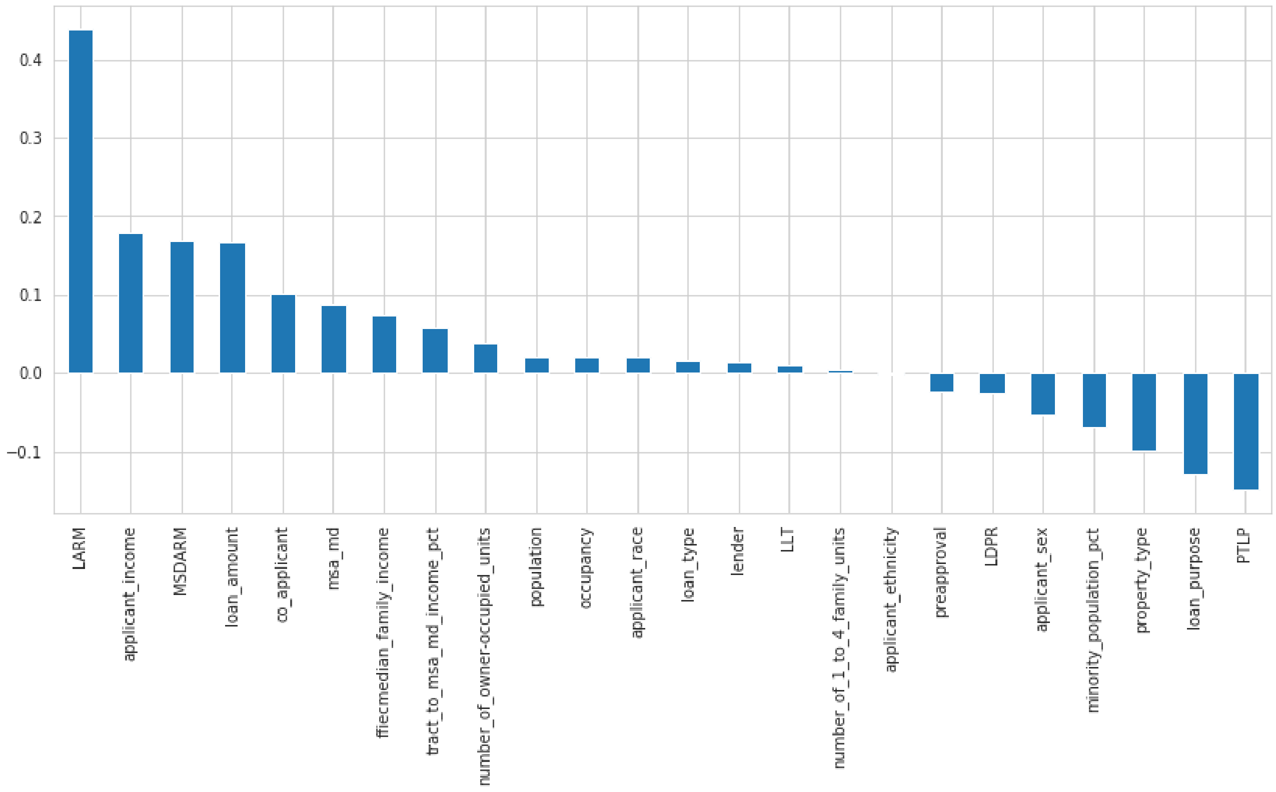




Correlation and Apparent Relationships

After exploratory of both numerical and categorical features, the relationships between features in the data was established. The emphasis was made between accept and the other features.



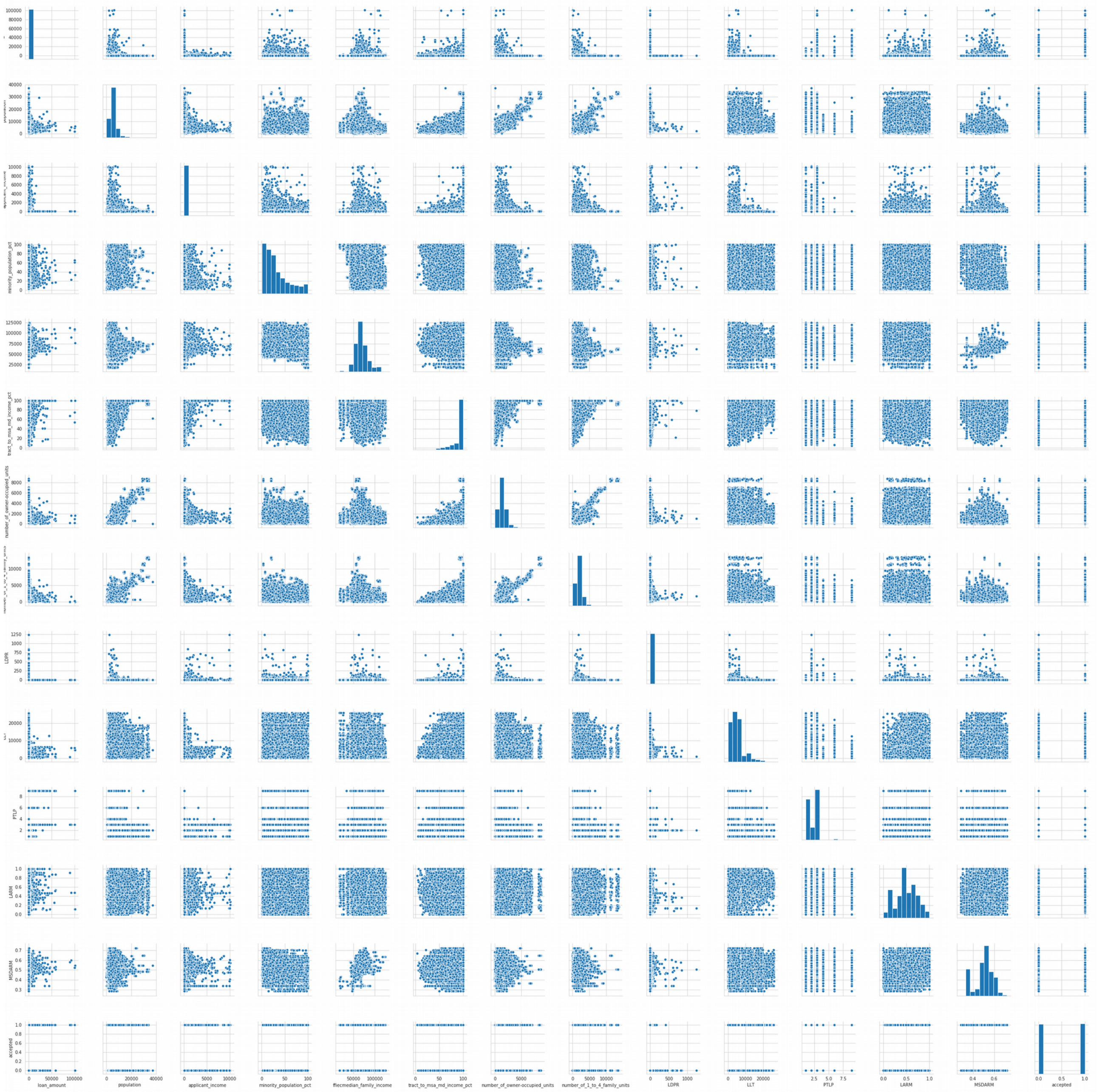


From both chart above, it can be seen that features that have negative correlation include LDPR (loan amount per applicant ration), applicant_sex, minority_population_pct,

loan_type, loan_purpose and PTLP(property type by loan type) and accepted.
Also, features such as number_of_1_to_4_family_units almost no effect. The rest of the features has positive correlation between them and accepted.

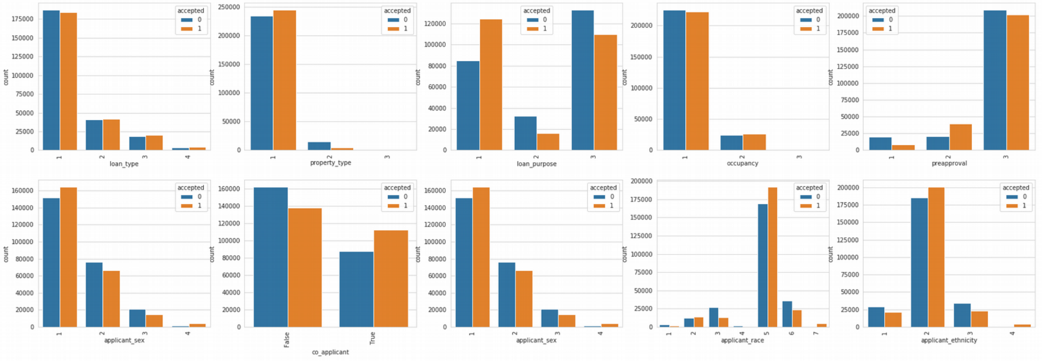
Numeric Relationships

The follow pairwise plot was generated initially to compare numeric features with each other. The key features in the are show below. Viewing the plots it can be seen that there is a relationship between accepted and other numeric features. However, because accepted is in two categories (i.e. 1= accepted and 0= denied), we see two different trends. It can be seen from the plot the relationship between accepted and other numeric features exhibits a “straight” nature.
Also it could be seen that most numeric features are skewed to the extreme left.



Categorical Relationships

After exploring numeric features, an attempt was made to explore categorical features. The following bar-plots show the categorical features and their relationship with accepted.



From the bar-plot was can establish that:

- Most loans taken where conventional loans
- Most properties chosen for the application were one to four family properties.
- Most purpose of the loans were for refinancing, followed by home improvement.
- Most loan applicant were male and also majority of loan applicant were whites.
- Most loan applicant had co-applicants.

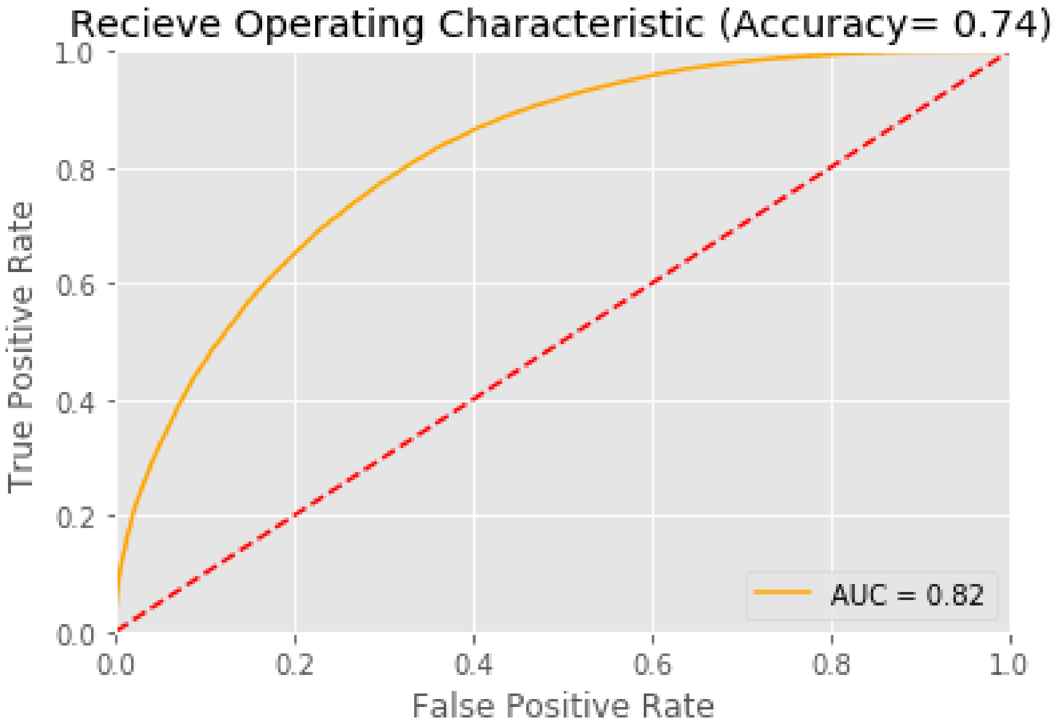
Classification of Loan Approval

After exploratory data analysis of the mortgage application data, features were created from the test datasets to match the reduced data in the train datasets and a classification model to predict if a loan will be accepted or denied was created.

Catboost classifier from yandex was used to create the model and trained with 85% of the data. Testing the model with the remaining 15% of the data yield the following results:

- True Positives: 25700
- True Negatives: 11643
- False Positives: 7987
- False Negatives: 29673

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



This interprets into the following standard performance metrics for classification:

- Accuracy: 0.74
- AUC: 0.82
- Precision: 0.74
- Recall: 0.74
- F1 score:0.74

Conclusion

This analysis has clearly show that mortgage approval can be confidently be predicted with ~ 74% accuracy without standard features, such as, credit score, debt to income ratios and other but instead can be predicted pieces of data from loan application, geographical data and census information.