ANALYSIS OF THE HOME MORTGAGE DISCLOSURE ACT (HDMA) DATASET

BY

RAHEEM RIDWAN LEKAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MICROSOFT PROFESSIONAL PROGRAM IN DATA SCIENCE (MPPDS)

JULY, 2019

Executive Summary

This report gives the analysis on the Home Mortgage Disclosure Act (HMDA) dataset and the result obtained from the machine learning model (binary classifier) used to predict whether a mortgage application was accepted or denied.

The dataset consists of 500,000 mortgage applications having 21 features each. After performing exploratory data analysis, data visualization, data cleaning, data processing and feature engineering, I was able to identify the features that are highly important for predicting the outcome of the target variable, i.e. acceptance/rejection of a mortgage application. These features were further used in several machine learning binary classifiers to predict the acceptance or rejection of a mortgage application. The catboost machine learning binary classifier proved to be the best by predicting the acceptance/rejection of a mortgage application with an accuracy of 73%.

The most important features in the dataset are:

- loan_amount: size of the requested loan in thousands of dollars.
- applicant_income: in thousands of dollars.
- tract_family_income: the tract median family income in dollars, a new feature created for this analysis.
- msa_md: a category with no ordering indicating Metropolitan Statistical Area/Metropolitan Division.
- minority_population: number of people that belongs to a minority group, another new feature created for this analysis.
- number_of_owner_occupied_units: Number of dwellings, including individual condominiums, that are lived in by the owner.
- ffiecmedian_family_income: FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC).
- co_applicant: indicates whether there is a co-applicant (often a spouse) or not.

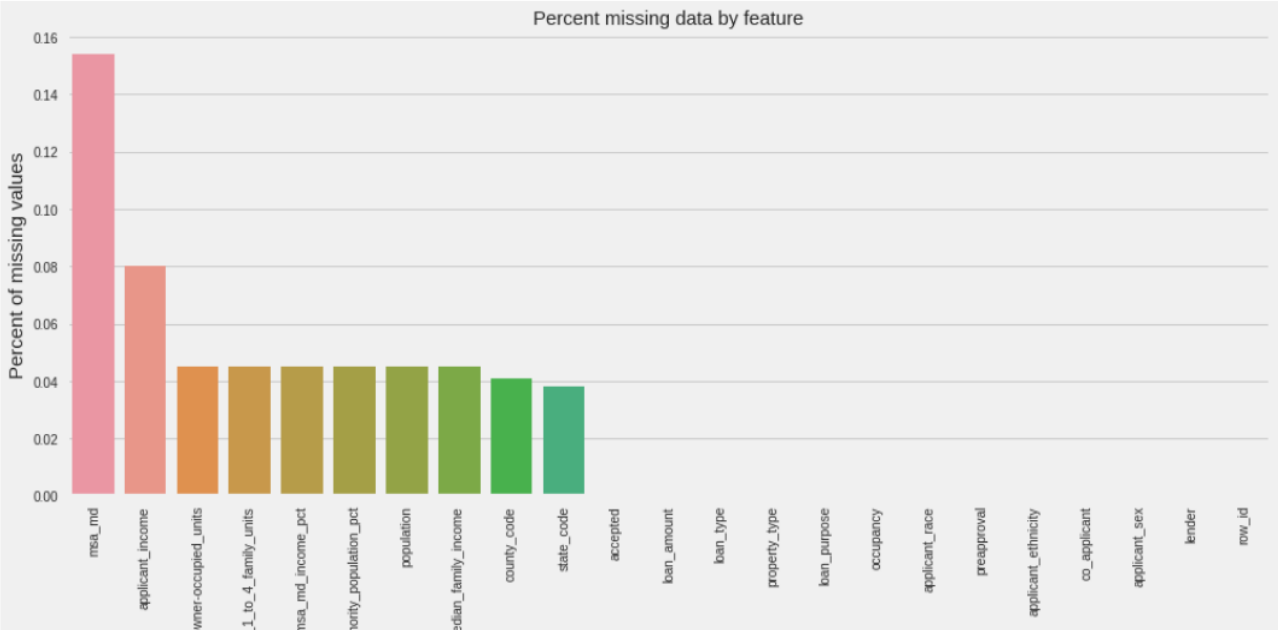
Exploratory Data Analysis

Firstly, I checked for the presence of missing values. This is important and needs to be addressed if any, because while some of the machine learning algorithms do not know how to handle missing data (values), others perform poorly in the presence of missing values.

```
row_id                                0
loan_type                             0
property_type                         0
loan_purpose                          0
occupancy                             0
loan_amount                           0
preapproval                           0
msa_md                            76982
state_code                        19132
county_code                       20466
applicant_ethnicity                   0
applicant_race                        0
applicant_sex                         0
applicant_income                  39948
population                        22465
minority_population_pct           22466
ffiecmedian_family_income         22440
tract_to_msa_md_income_pct        22514
number_of_owner-occupied_units    22565
number_of_1_to_4_family_units     22530
lender                                0
co_applicant                          0
accepted                              0
dtype: int64
```

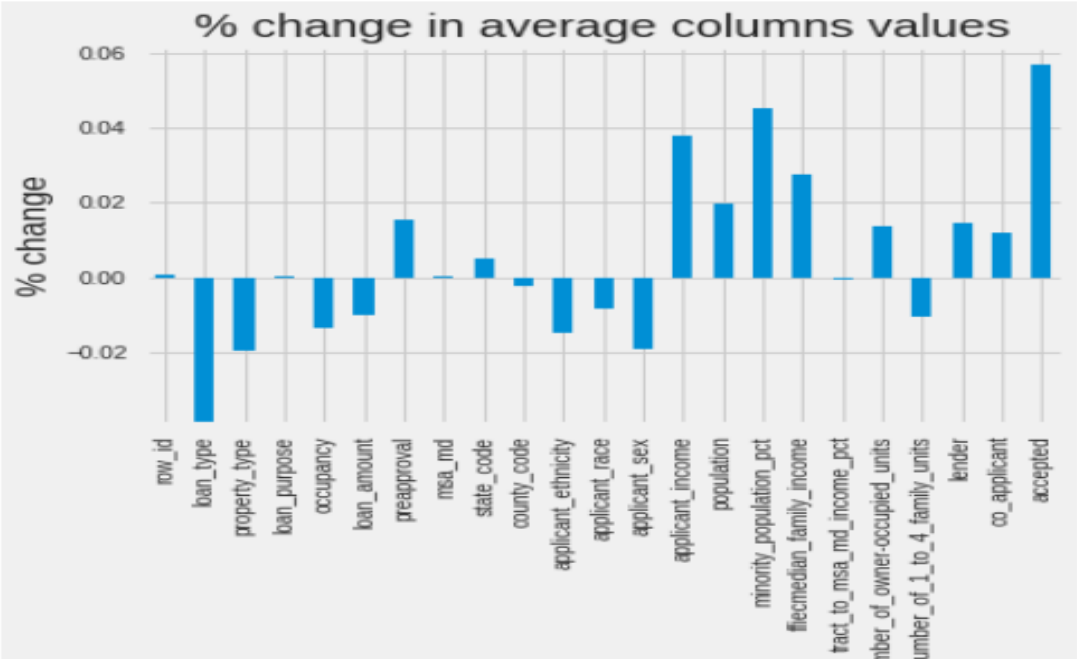
We can see that while the features applicant_income and msa_md have 39,948 and 76,982 missing values respectively, loan_type and property_type have no missing values. Three of the features namely msa_md, state_code and county_code have missing values, denoted with -1 as specified in the dataset description.

	Total	Percent
msa_md	76982	0.153964
applicant_income	39948	0.079896
number_of_owner-occupied_units	22565	0.045130
number_of_1_to_4_family_units	22530	0.045060
tract_to_msa_md_income_pct	22514	0.045028



Bar charts were created to show the percentage of missing data by feature. The table and the bar chart show the percentage of missing values from each feature. One of the most important features *msa_mda* has *76,982* missing values, which is equivalent to 15%, this is a big leap.

loan_type	-0.039328
property_type	-0.019788
loan_purpose	0.000176
occupancy	-0.013737
loan_amount	-0.009940
preapproval	0.015580
msa_md	0.000144
state_code	0.005136
county_code	-0.002400
applicant_ethnicity	-0.015081
applicant_race	-0.008558
applicant_sex	-0.019201
applicant_income	0.037813
population	0.019950
minority_population_pct	0.045296
ffiecmedian_family_income	0.027602
tract_to_msa_md_income_pct	-0.000547
number_of_owner-occupied_units	0.013753
number_of_1_to_4_family_units	-0.010543
lender	0.014707
co_applicant	0.011865
accepted	0.056897



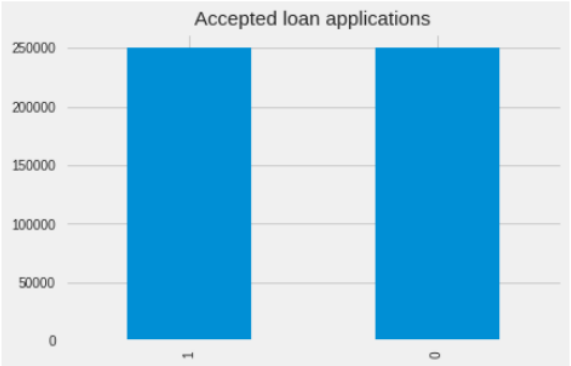
To further investigate the effect of dropping (removing) features with missing values, the percentage change in means of the features were plotted using a bar chart. This was obtained by finding the difference between the mean values of each feature in the dataset including the missing values, and the mean values of each feature in the dataset excluding the missing values.

We can see that the mean of the feature loan_type fell by about 4% after removing the missing values, which is a big change. In contrast, the applicant_income, minority_population_pct and ffiecmedian_family_income rose by about 4%, 5% and 3% respectively, which is also a big leap.
We can see how dropping observations (rows) affect the shape of the data. Hence, we want to try to retain as much data as possible.

```
row_id                                0
loan_type                             0
property_type                         0
loan_purpose                          0
occupancy                             0
loan_amount                           0
preapproval                           0
msa_md                                0
state_code                            0
county_code                           0
applicant_ethnicity                   0
applicant_race                        0
applicant_sex                         0
applicant_income                      0
population                            0
minority_population_pct               0
ffiecmedian_family_income             0
tract_to_msa_md_income_pct            0
number_of_owner-occupied_units        0
number_of_1_to_4_family_units         0
lender                                0
co_applicant                          0
accepted                              0
dtype: int64
```

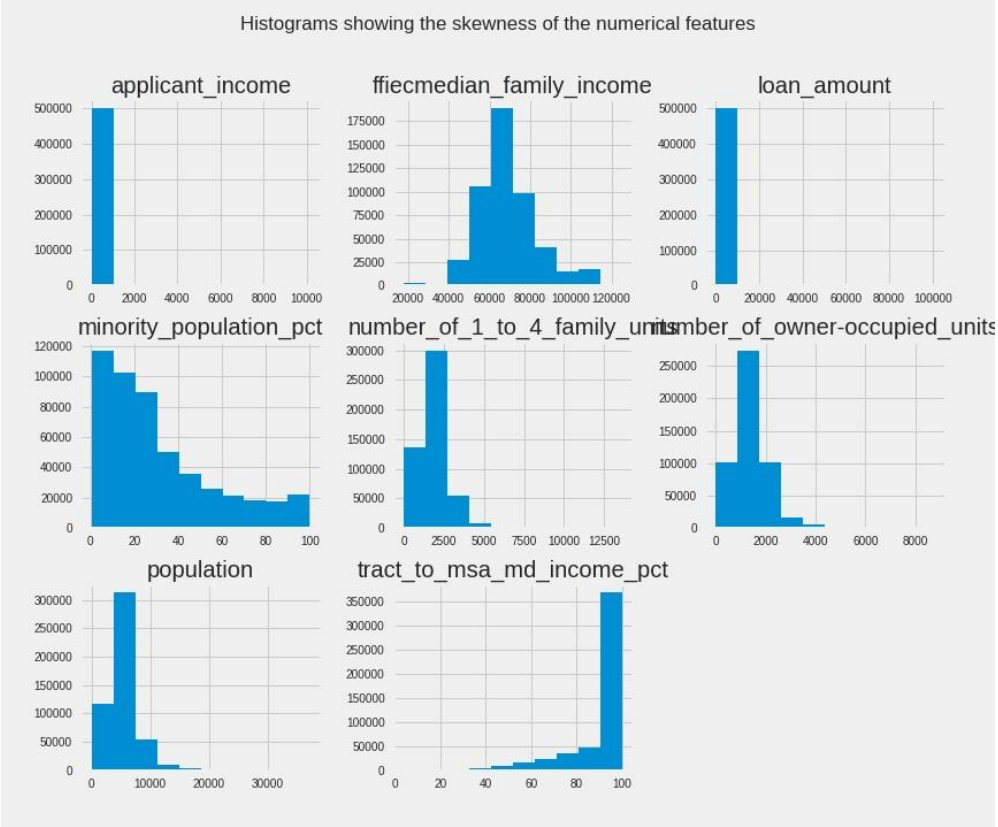
To resolve the problem of the missing values, I filled the categorical features missing values with the mode of each feature, and the continuous (numeric) features with the median of each feature.

```
1    0.500228
0    0.499772
Name: accepted, dtype: float64
```



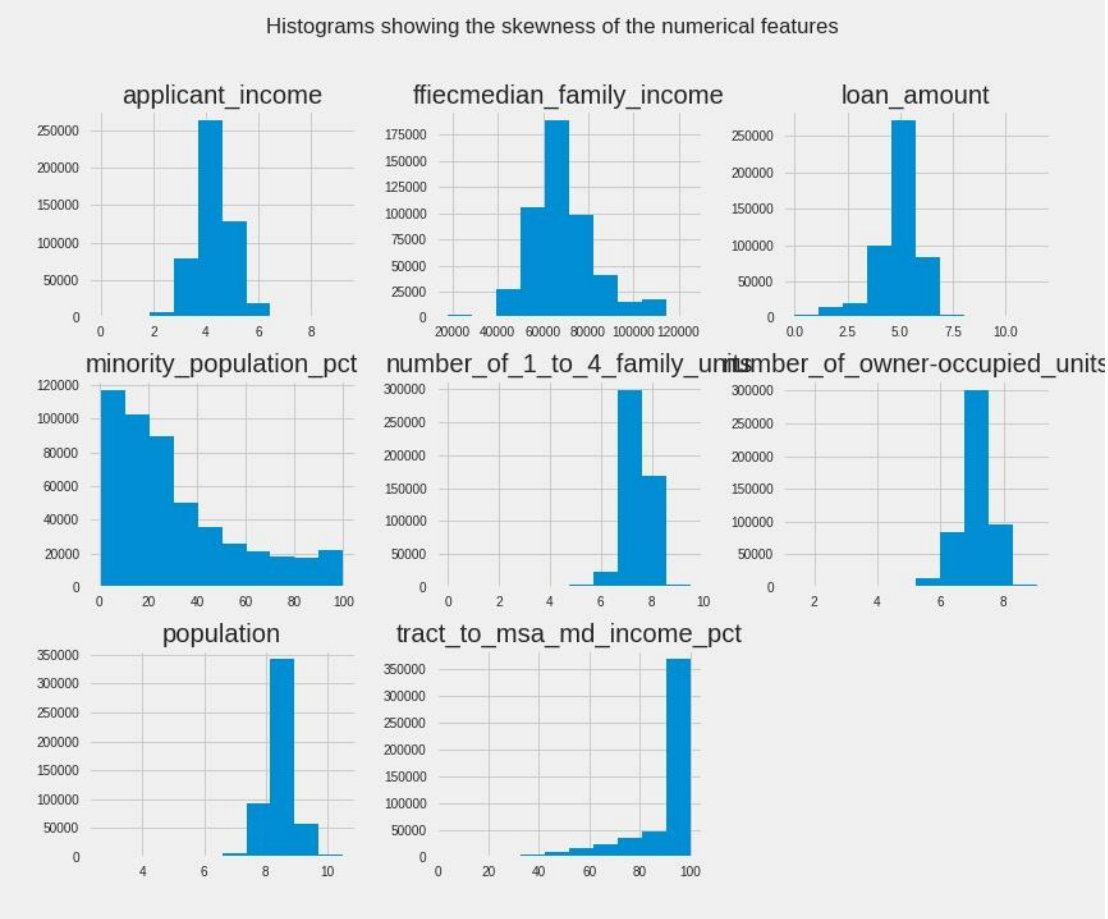
Next, a check on the balance between the classes (acceptance denoted by 1, and rejection denoted by 0) of the dataset shows that it is almost perfectly balance. This will improve the performance of our machine learning model.

```
Skewness of the numerical features

loan_amount                          76.552786
applicant_income                     23.174985
population                            2.947782
minority_population_pct               1.068839
ffiecmedian_family_income             0.806355
tract_to_msa_md_income_pct           -2.035543
number_of_owner-occupied_units        1.942059
number_of_1_to_4_family_units         2.080321
dtype: float64
```



Ideally, for a machine learning model to perform optimally, the dataset given must be close or follow a normal distribution. From the histograms and table, the numeric features *applicant_income* and *loan_amount* are highly skewed to the right, and this would affect the performance of our machine algorithm. To reduce the skewness, and have a better visualization, logarithmic function is applied to the dataset.

```
Skewness of the numerical features

loan_amount                    -1.190548
applicant_income                0.029289
population                     -0.159924
minority_population_pct         1.068839
ffiecmedian_family_income       0.806355
tract_to_msa_md_income_pct     -2.035543
number_of_owner-occupied_units -1.116117
number_of_1_to_4_family_units  -1.615771
dtype: float64
```



Now, all the numeric features almost follow the normal distribution. In addition, the categorical features were explored. These categories include:

- loan_type: 1 -- Conventional, 2 -- FHA-insured, 3 -- VA-guaranteed, 4 -- FSA/RHS
- property_type: 1 -- One to four-family, 2 -- Manufactured housing, 3 -- Multifamily
- loan_purpose: 1 -- Home purchase, 2 -- Home improvement, 3 -- Refinancing
- occupancy: 1 -- Owner-occupied as a principal dwelling, 2 -- Not owner-occupied, 3 -- Not applicable.
- preapproval: 1 -- Preapproval was requested, 2 -- Preapproval was not requested, 3 -- Not applicable
- applicant_ethnicity: 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: One of 8 races
- applicant_sex: 1 – Male, 2 – Female, 3 -- Information not provided by applicant in mail, Internet, or telephone application, 4 or 5 -- Not applicable
- co_applicant: True or False
- msa_md: One of 408 Metropolitan Statistical Area/Metropolitan Division
- state_code: One of 52 state codes
- county_code: One of 317 county codes
- lender: One of 6111 of the authorities in accepting or denying a loan

Bar charts were plotted to show the frequency of each of the categorical features, and we observed the following:

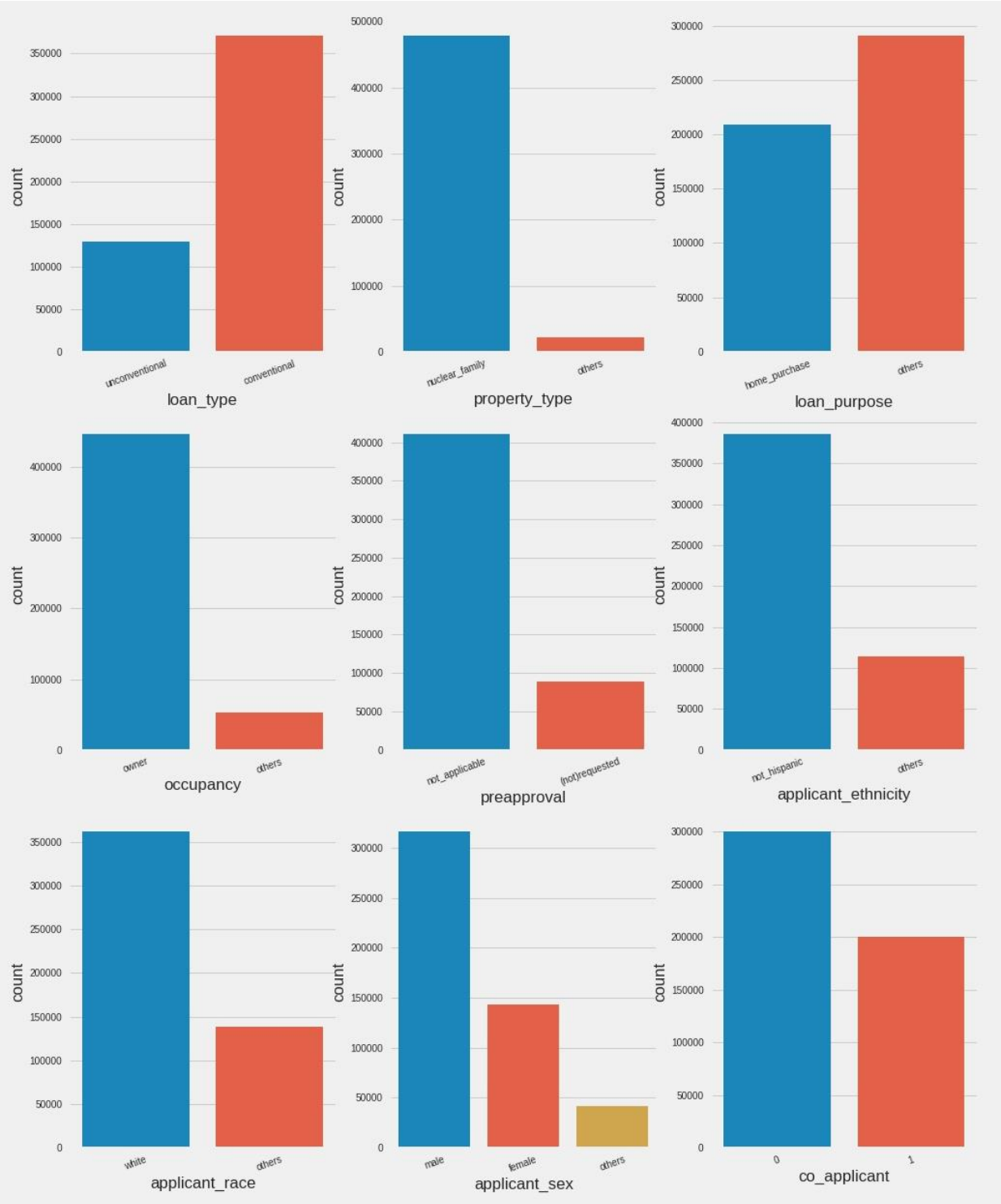
- Conventional loan type was the most common loan_type, followed by FHA-insured, then VA-guaranteed, while FSA/RHS was not common.
- The property_type, one-to-four-family was the most common, followed by Manufactured housing, and then Multifamily.
- Majority of the loan applicants specified Refinancing as their loan_purpose, followed by home purchase, while home improvement few applicants specified home improvement as the reason for the loan_purpose.
- Almost all the applicants are occupy/live in their own dwelling.
- Majority of the applicants did not request for a preapproval of a home purchase loan.
- The most common ethnic group among applicants is Latino, while the most uncommon applicants did not belong to any of the outlined ethnic groups.
- For the race of the applicants, majority of them are white, followed by those who did not disclose their race, then the Black or African American, then the applicants who do not fall in any of the categories, then the American Indian, followed by the Native Hawaiian or other pacific Islander who are the most uncommon among the races.
- In terms of sex of the applicants, the majority of them are males, followed by females, then those who did not disclose their gender (sex) during the application, and then those that do not fall in any other specified sex categories.
- Majority of the applicants applied individually, i.e. they applied without a co-applied (without a spouse).

It the categorical features, it was observed some of the unique values have very high frequency, while others have very low frequency. This can be seen in the bar charts below

To have a better visualization, the categorical features with few values having high frequency and many values with low frequency were regrouped to have a smaller group, by combining smaller categories as specified below.

- loan_type: 1- conventional, 2- unconventional, 3- unconventional, 4 - unconventional
- applicant_race: 5 - white, 6 - others, 3 - others, 2 - others, 7 - others, 1 - others, 4 - others
- applicant_sex: 1- male, 2 – female, 3 – others

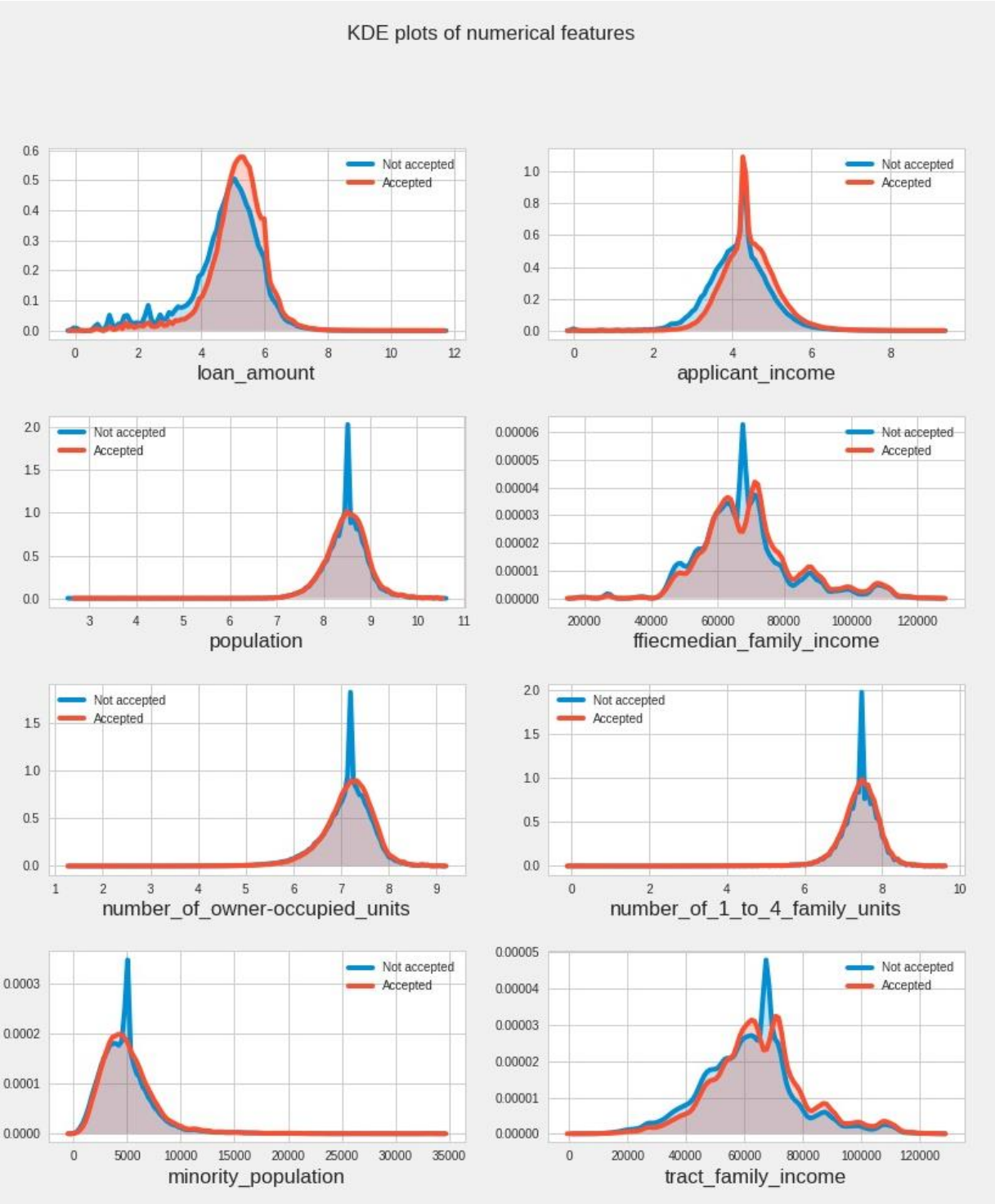
The other categorical features having the same issues were also regrouped in the same manner to have the bar charts shown below:



Correlation and Apparent Relationships

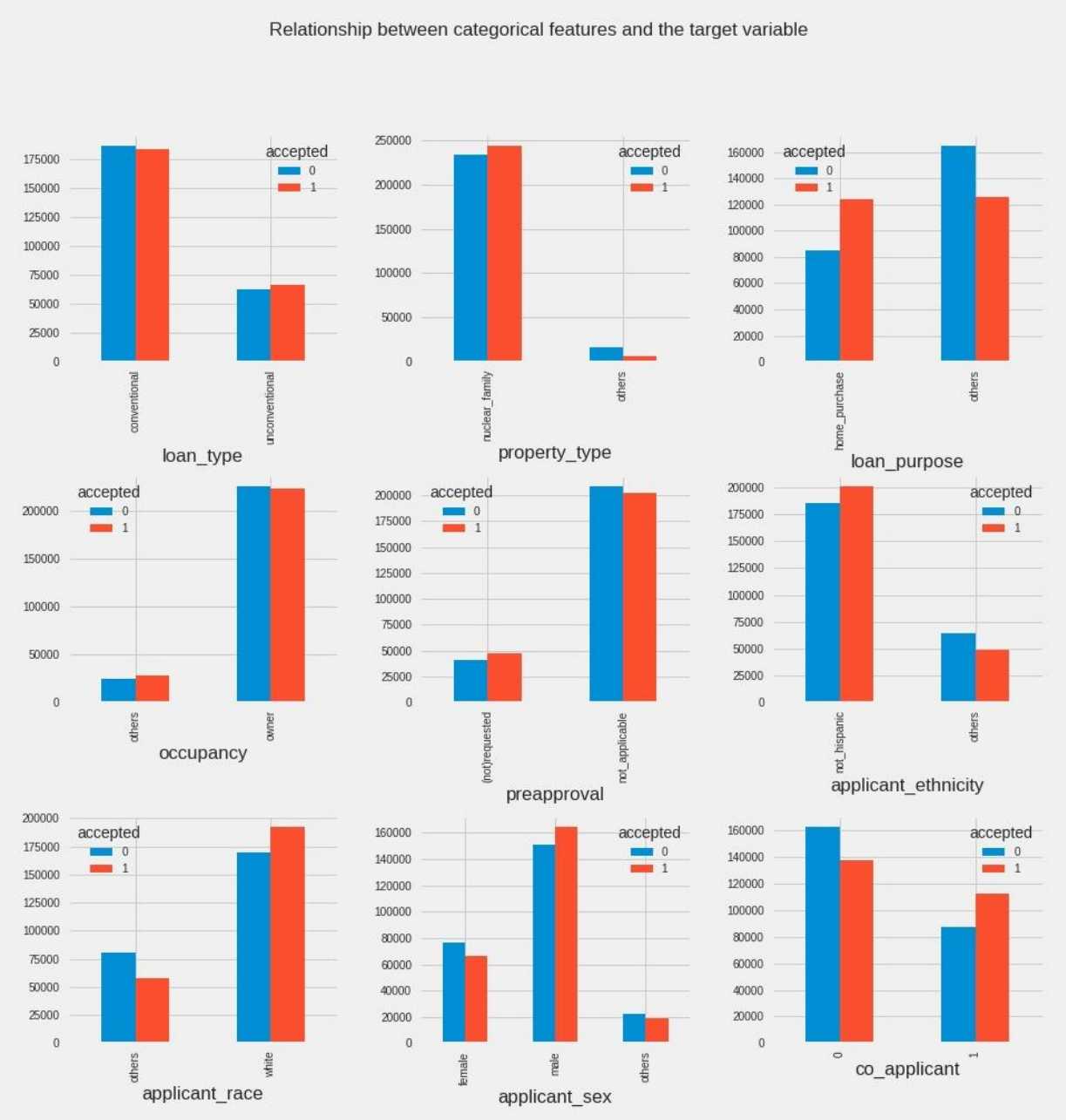
The relationship between the features in the dataset and the target variable (accepted) was plotted. Due to the presence of numeric (continuous) and categorical features in the dataset, the correlation (relationship) plot was divided into two namely; correlation between the numeric features and the target variable (accepted), and the correlation between the categorical features and the target variable (accepted).

Correlation plot between numeric features and the target variable (accepted)

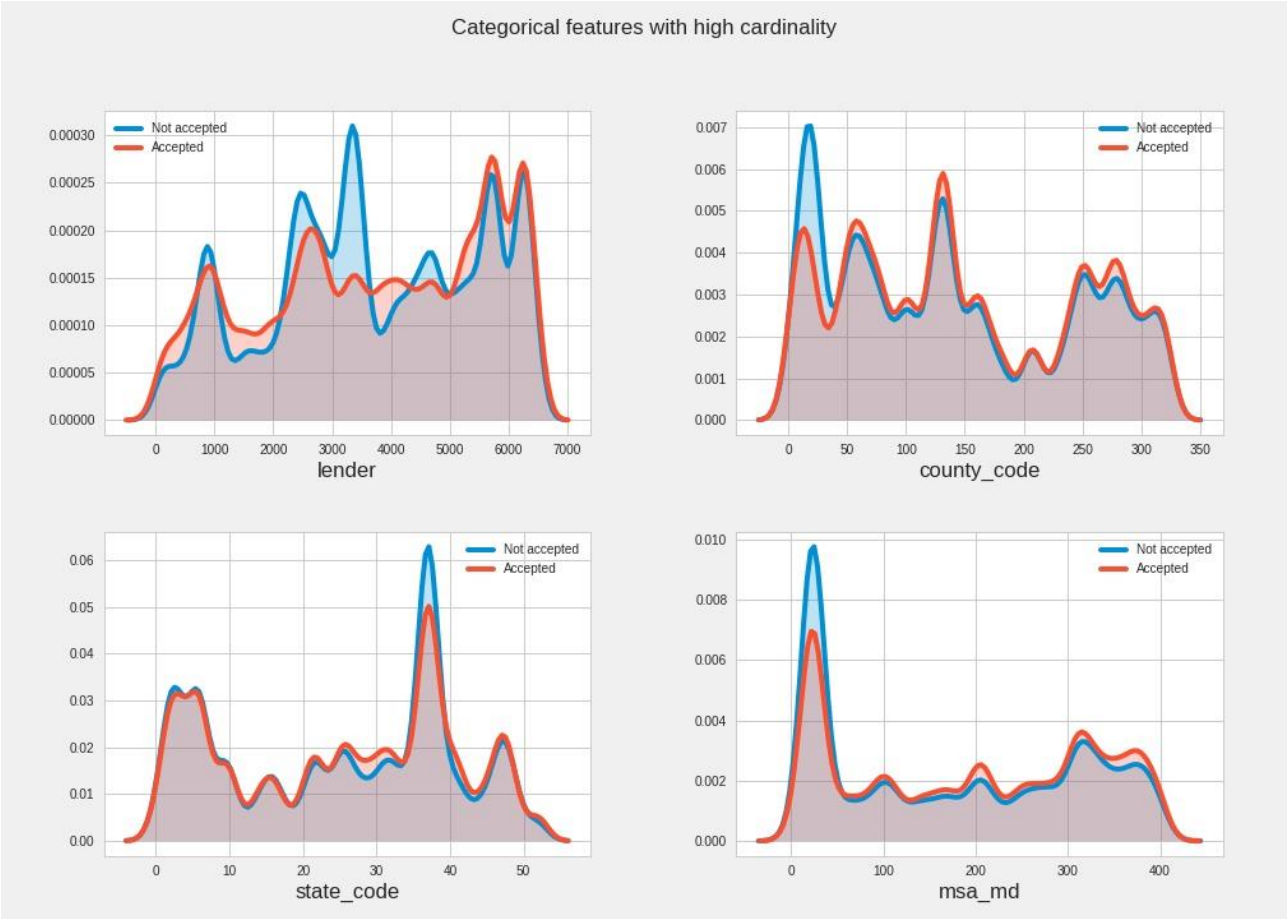


From the plot, we observed that *loan_amount, applicant_income, ffiecmedian_family_imcome, number_of_owner_occupied_units, minority_population and tract_family_income* have a good class separation, which are important features that need to be considered for the prediction of the acceptance or rejection of a loan applicant's application.

Correlation plot between categorical features and the target variable (accepted)



The plots contain only features with low cardinality (i.e. features having few number of unique values.). From our visualization, applicants that are white are given preference when compared with other applicants. Furthermore, single applicant's applications are more accepted than their counterpart that have spouse. In addition, the rate of acceptance is higher for applicants who applied to conventional loans, when compared to the unconventional loan applicants.



loan_amount	0.169416
msa_md	0.073920
state_code	0.004854
county_code	0.045159
applicant_income	0.178756
population	0.020702
minority_population	0.064777
ffiecmedian_family_income	0.070197
tract_family_income	0.097705
number_of_owner-occupied_units	0.035664
number_of_1_to_4_family_units	0.001359
lender	0.008494
accepted	1.000000
loan_type__conventional	-0.014503
loan_type__unconventional	0.014503
property_type__nuclear_family	0.098751
property_type__others	-0.098751
loan_purpose__home_purchase	0.159822
loan_purpose__others	-0.159822
occupancy__others	0.019216
occupancy__owner	-0.019216
preapproval__(not)requested	0.036850
preapproval__not_applicable	-0.036850
applicant_ethnicity__not_hispanic	0.074700
applicant_ethnicity__others	-0.074700
applicant_race__others	-0.102410
applicant_race__white	0.102410
applicant_sex__female	-0.042726
applicant_sex__male	0.053934
applicant_sex__others	-0.024384
co_applicant__0	-0.101116
co_applicant__1	0.101116

KDE plots were used to visualize high cardinality features, that is, features having high number of unique values. As we can see, there is good class separation in all the 4 features. Hence, these features will be useful in prediction the outcome of a loan application. These relationships is also evident as can be seen in the correlation values between the features and the target variable as shown above.

Machine Learning

After the analysis of the HDMA dataset, I tried to predict whether an applicant's loan application was accepted or rejected. Dummy variables were created for the categorical features, which lead to high increase in the size of the dataset. The dataset was then passed into scikit-learn standardScaler to scale the dataset, after which it was split into train data and test data with 70% and 30% respectively. The train data was trained using several machine learning algorithms which included: *Logistic Regression*, *Gradient Boost Classifier, Ada Boost Classifier, MLP Classifier, xgboost and Random Forest Classifier.* The accuracy obtained from these algorithms were below 70%, which were not acceptable. Further research was made on the best algorithm for a dataset having high categorical features, the catboost algorithm proved to be a good choice for the dataset. This algorithm is good for dataset having many categorical features and high cardinality. It was observed that, unlike some algorithms that needed categorical features to be dummified and one-hot-encoded, the catboost algorithm does not need this. The catboost algorithm was trained with 70% of the dataset, and later tested with the remaining 30% of the dataset. The following results were obtained:

	Confusion matrix	
	Score positive	Score negative
Actual positive	25157	12218
Actual negative	7957	29668

Accuracy	0.73
AUC	0.81
Macro precision	0.73
Macro recall	0.73

	Positive	Negative
Num case	37375	37625
Precision	0.76	0.71
Recall	0.67	0.79
F1	0.71	0.75

Recommendation and Conclusion

The goal of this project was achieved, which was to maximize the accuracy of the classifier. I realized some form of nepotism or discrimination in our data. For example, a visual look at the race of the applicants showed that the whites were highly favoured for loan issuance while other races were insignificantly represented for the loan issuance. The same thing applies for the applicants' sex, almost all the applicants granted loans were males. Although, my personal opinion is not needed for this project, but as a data scientist, if I find myself in this kind of organization, I will try to discuss with the management of the company on the importance of treating every person equally, provided they meet the company's requirements for the loan issuance.

Important links:

- About the dataset: https://datasciencecapstone.org/competitions/14/mortgage-approvals-from-government-data/
- GitHub repository for the python code: https://github.com/Ridwanlekan/Final_Project_Microsoft_Professional_Program_in_Data_Science.git