PREDICTING MORTGAGE APPROVALS FROM GOVERNMENT DATA

*Fabiyi Opeyemi (*Username: Opiano), *June 2019.*

Executive Summary

This work presents an analysis of data which goal is to predict 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 considered how demographics, location, property type, lender, and other factors are related to whether a mortgage application was accepted or denied and to use your skills to build a model for predicting acceptance of loan application across the United States**.**

The framework and pipeline taken to proffer solution to the problem statement is as follows;

1. Exploratory Data Analysis (EDA)
2. Data Modelling
3. Model Evaluation

1. Exploratory Data Analysis (EDA)

This stage is also known has data preprocessing or data munging to get insights and patterns from the data as well as prepare it for the modelling phase. In exploring and analyzing the data to gain insights the following approach was adopted;

i. Univariant Analysis
ii. Bi/Multi-Variant Analysis

After exploring the data by calculating summary, descriptive statistics, and by creating visualizations of the data, several potential relationships between applicant characteristics and the target which is the 'acceptance' was observed. After exploring the data, a predictive model to classify the mortgage approval into rejected or accepted through thorough insights into the analytics of the data.

There are 23 variables in this dataset for 500,000 observations where each row in the dataset represents a HMDA-reported loan application, and the dataset covers one particular year. The 23 variables/features for the given data is shown below;

```
row_id                              154463 non-null int64
loan_type                           154463 non-null int64
property_type                       154463 non-null int64
loan_purpose                        154463 non-null int64
occupancy                           154463 non-null int64
loan_amount                         154463 non-null float64
preapproval                         154463 non-null int64
msa_md                              154463 non-null int64
state_code                          154463 non-null int64
county_code                         154463 non-null int64
applicant_ethnicity                 154463 non-null int64
applicant_race                      154463 non-null int64
applicant_sex                       154463 non-null int64
applicant_income                    142261 non-null float64
population                          153062 non-null float64
minority_population_pct             153061 non-null float64
ffiecmedian_family_income           153072 non-null float64
tract_to_msa_md_income_pct          153056 non-null float64
number_of_owner-occupied_units      153055 non-null float64
number_of_1_to_4_family_units       153055 non-null float64
lender                              154463 non-null int64
co_applicant                        154463 non-null bool
accepted                            154463 non-null int64
```

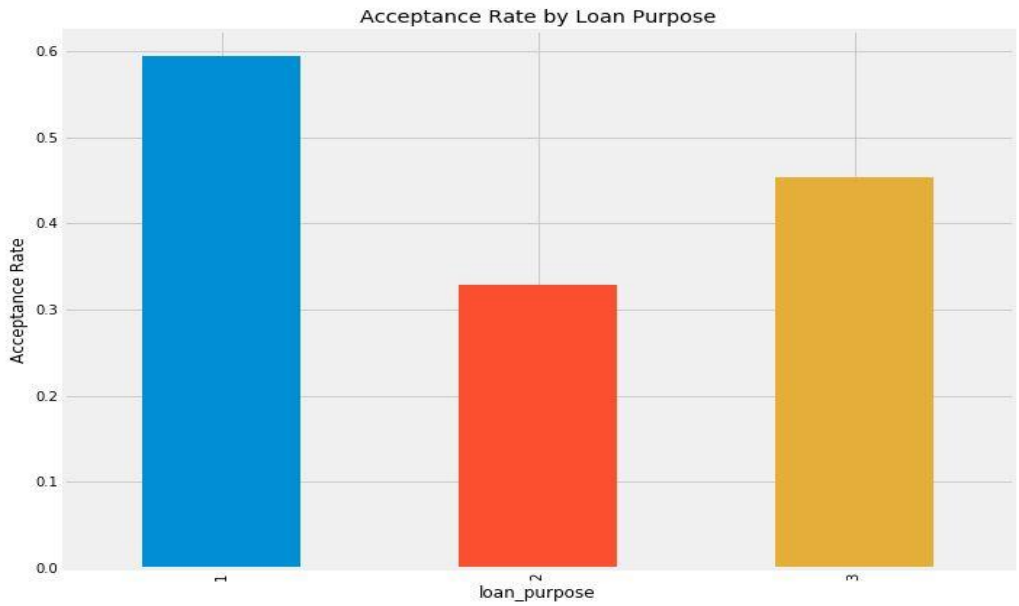
Figure 1: Dataset features

1. **Univariant Analysis**

Categorical

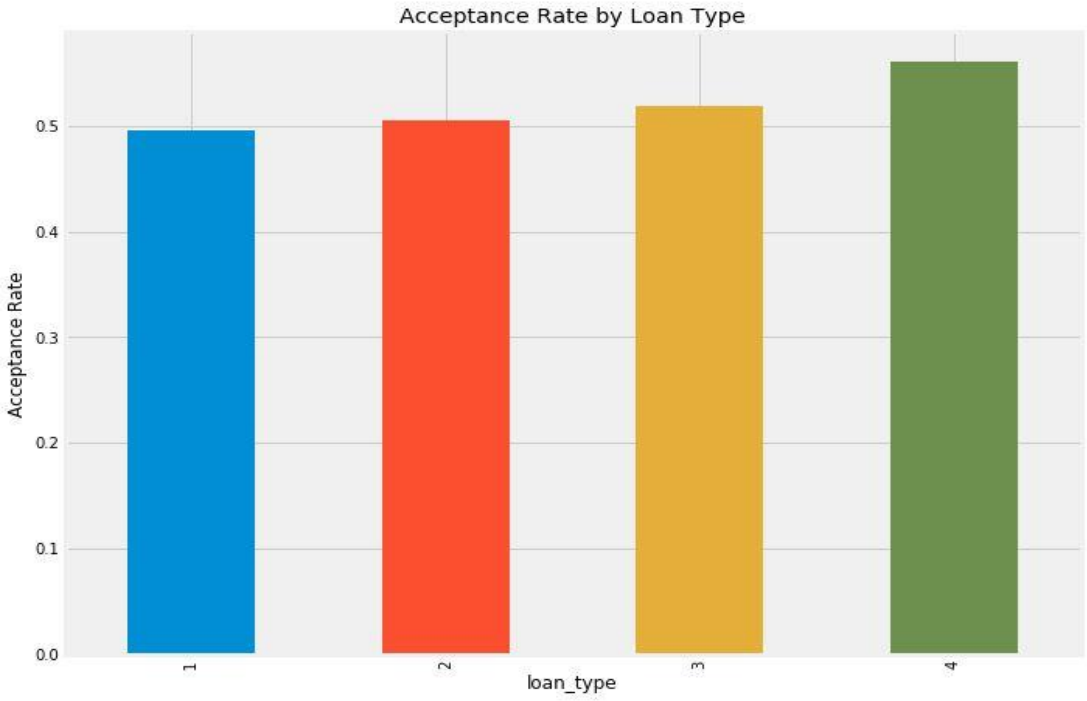
The categorical feature in the data includes; loan_type, property_type, loan_purpose, occupancy, applicant_race etc. which were analyzed individually in reference to the target ‘acceptance’ to discover hidden insights with these features. After exploring these features, the following insights were deduced.

- **Loan Purpose**



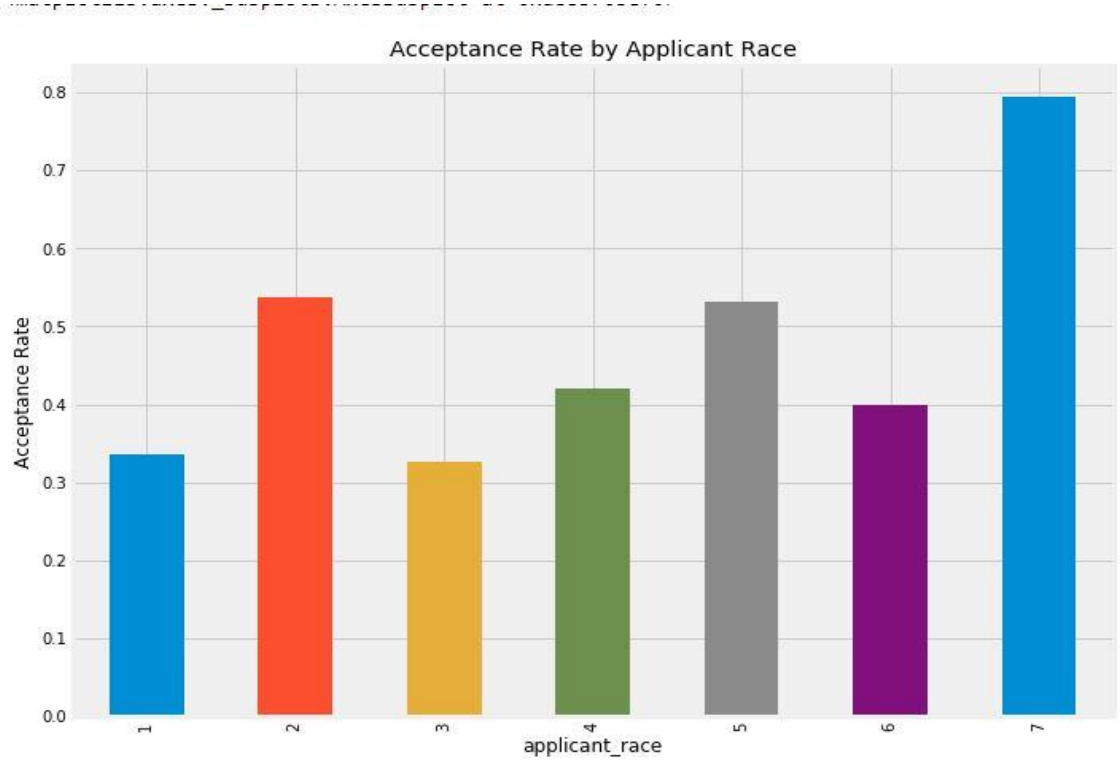
It is observed that loan application for the purpose of home purchase has higher acceptance rate than loan application for the purpose of Refinancing which is also higher acceptance rate than that of Home improvement.

- **Loan Type**



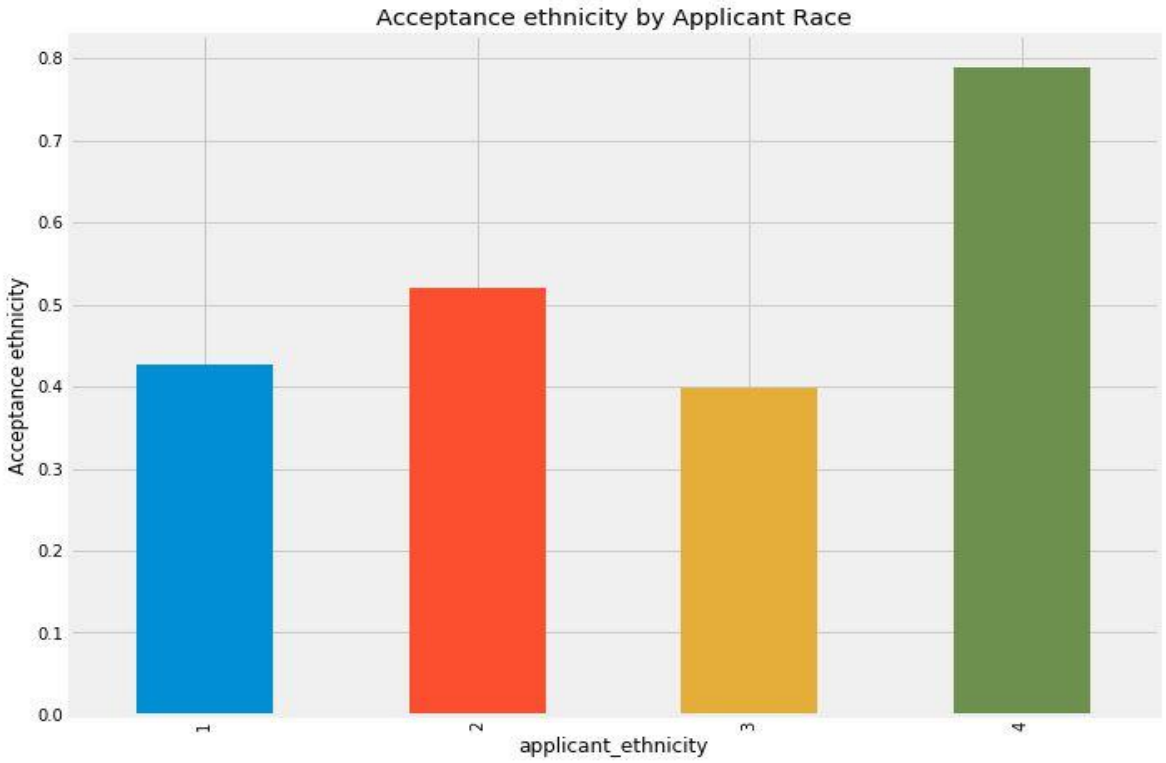
There seems to be no clear pattern or insights to be derived from this feature, hence, it is not a good use for feature generation which will help improve the model.

- **Applicant Race**



From the visualization shown above, we see that Applicant has a high acceptance rate in relation to others. Hence, we can conclude that Applicant race will play a key role in determine if a mortgage will be accepted or not.

- **Applicant Ethnicity**

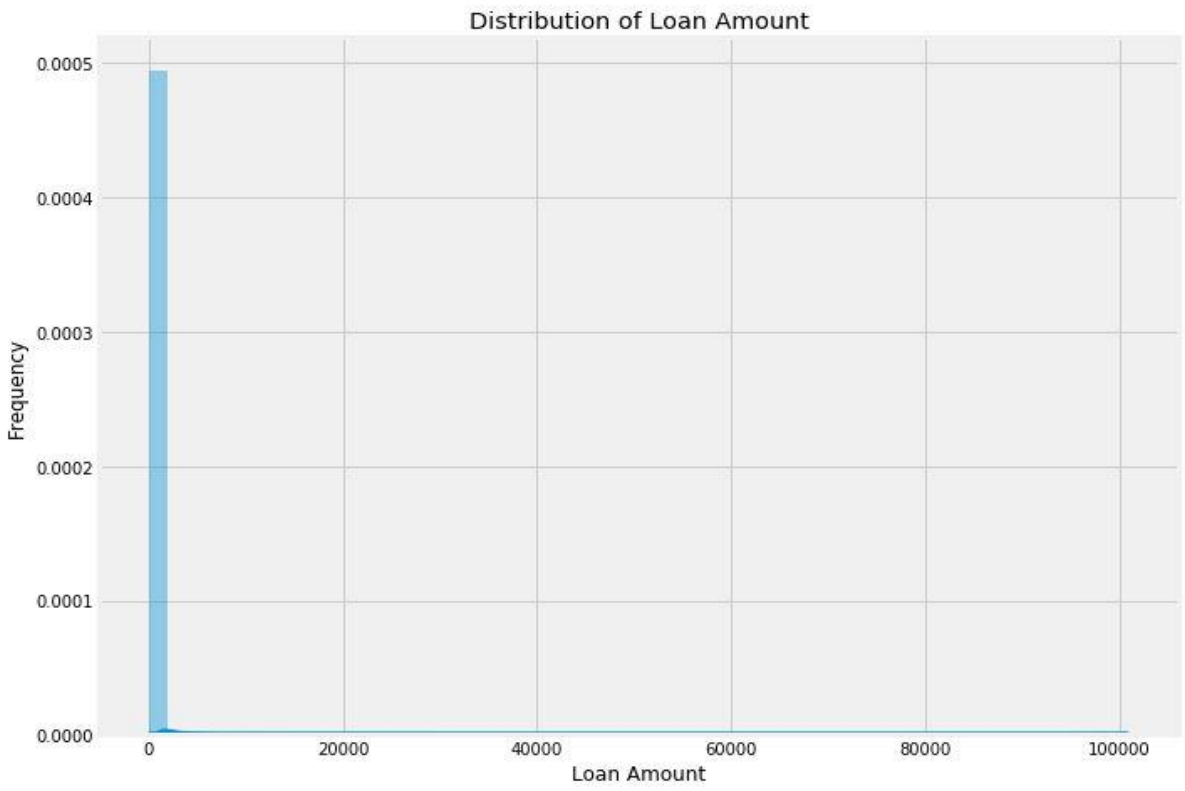


The applicant ethnicity also has an important role to play in determining if a mortgage will be granted or not. Applicant ethnicity 4 has the highest acceptance rate.

This analysis was performed for other categorical features to get an insight into them.

Numerical

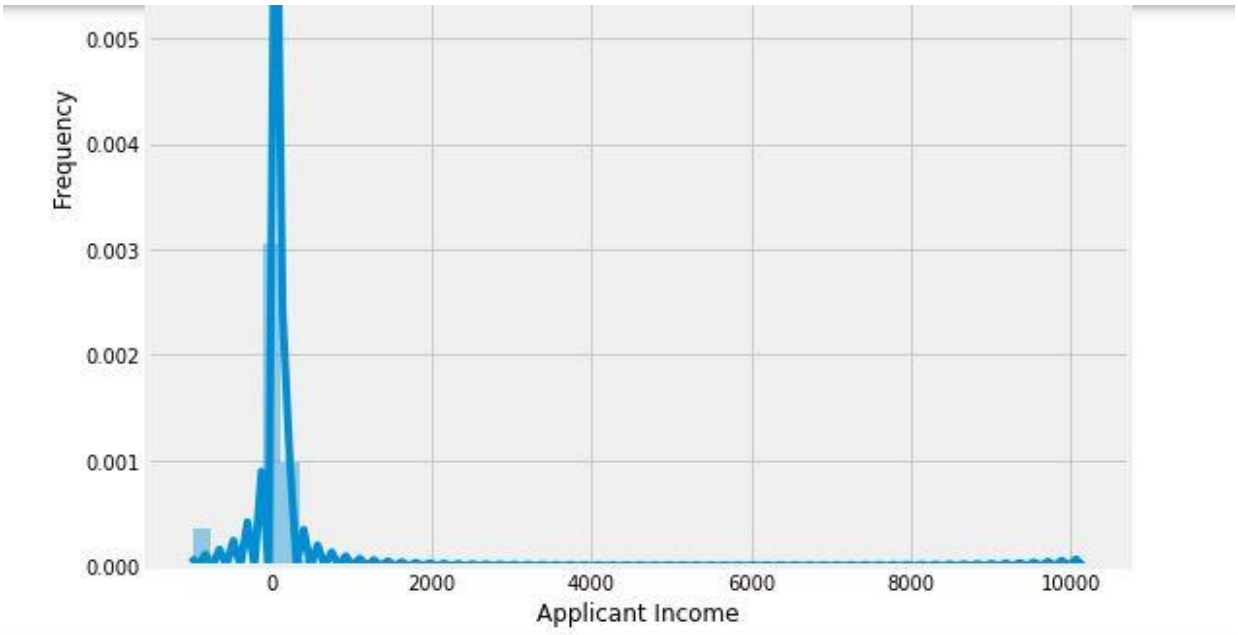
- **Loan Amount**



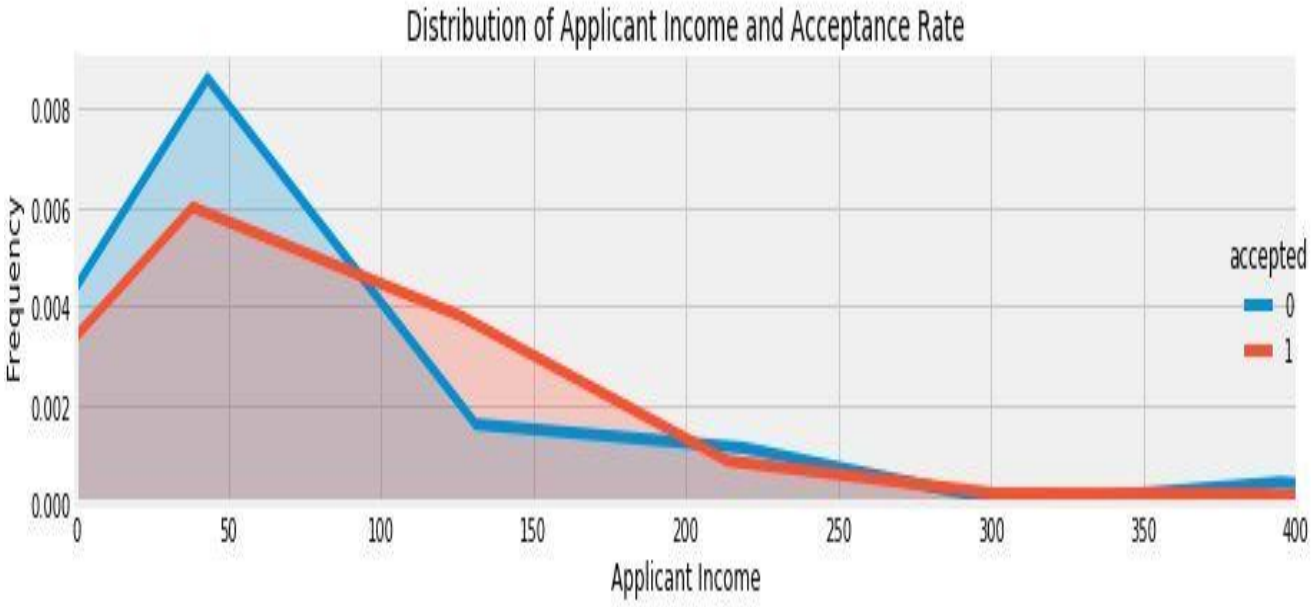
From the distribution display we see that the loan amount is highly skewed which will affect the performance of the model, hence, it is expedient that the feature is transformed so as to help improve the model performance.

- **Applicant Income**

Similarly, the distribution for applicant income is highly skewed which will also affect the performance of the model, hence, the feature is also transformed and scaled for model improvement.



Through domain knowledge applicant income usually should be a determining factor for the acceptance, hence, this intuition gives a direction to further analyze this feature and the following insights were discovered;



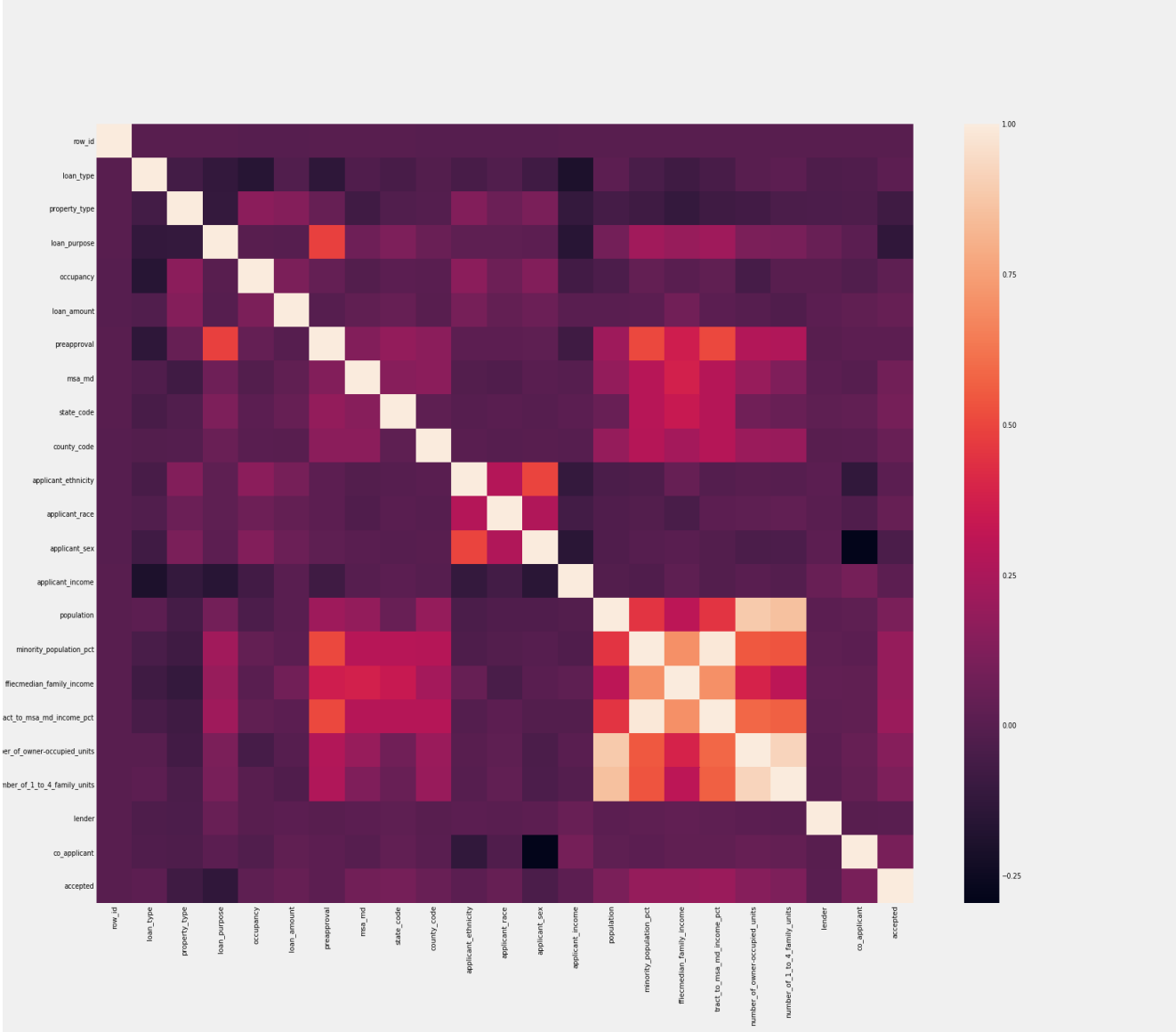
The visualization gives us an insight that for low income earner within the range of 0-100 the probability of loan rejection is high while as it increases between 100-200 there is a high probability of a loan been accepted. This insight will be used to create new feature of **'Applicant**

Income range' into '*Very Low, Low, Medium, High, Very High'* which greatly impact the model performance.

2. Bi/Multivariant Analysis

This analysis was performed to check for correlation and apparent relationships between features that is exploring the individual feature in an attempt to identify relationships between features in the data most especially the target 'accepted'

After exploring the individual features, an attempt was mad to identify relationships between features in the data – in particular between "accepted" and other features



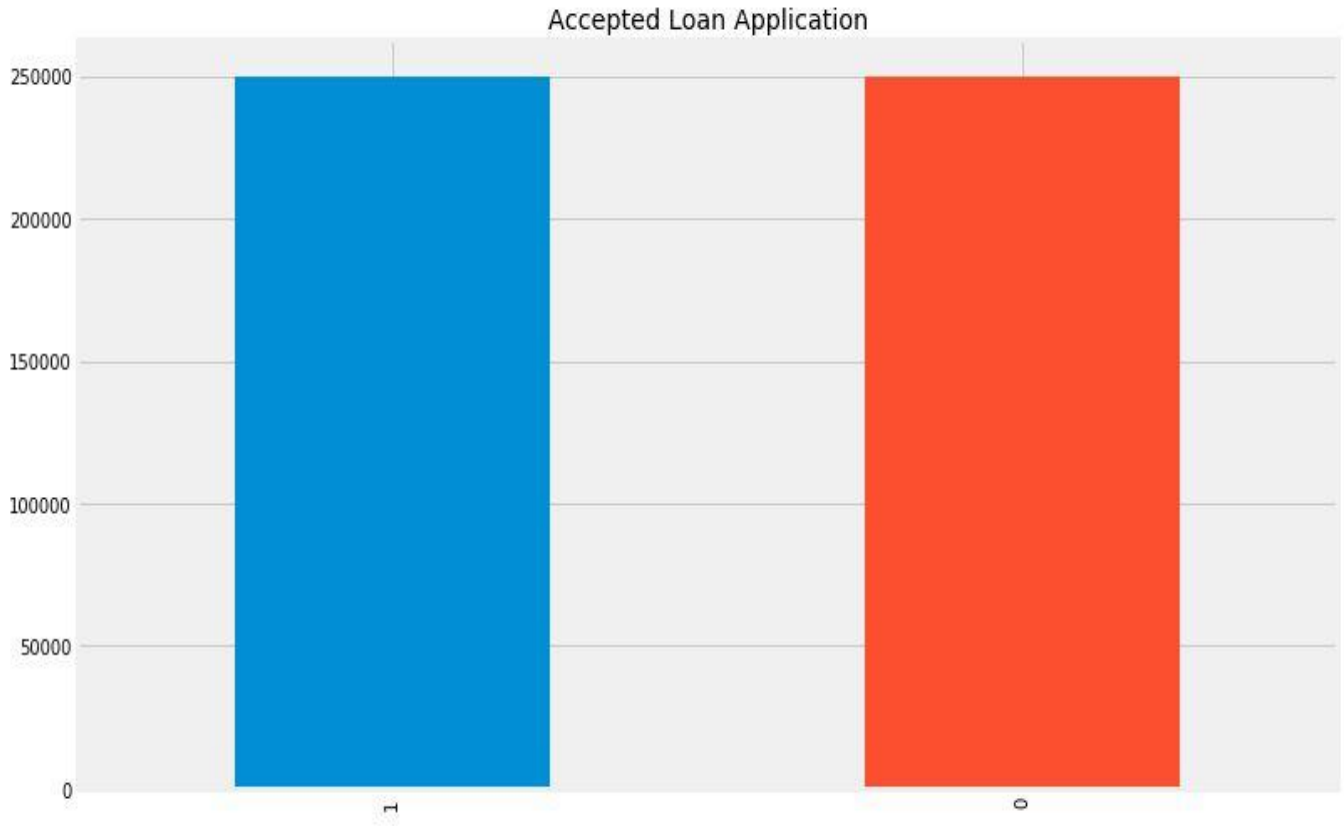
While many factors can help indicate if a mortgage was approved or not from the various analysis performed on the data, however, insights was discovered from some features found to have significant impacts on the target variable which is determinant of a mortgage loan to be accepted or not and they include;

1. **Applicant income:** Applicants with high income tends to have high acceptance rate and applicants with low income tend to have low acceptance rate.
2. **Loan purpose**: The purpose of the loan plays a pivotal role in the mortgage acceptance as loan application for home purchase has higher acceptance rate than the others used for other reasons
3. **Lender**: The institution giving out the mortgage. Some lenders tended to have a higher approval rate than others. Out of all the features, this one was the most impacting.
4. **Applicant race:** The race of an applicant is important as race 7 applicants has higher chance their mortgage being accepted.
5. **Applicant ethnicity:** As it applies with the race of the applicant, the same also applies with the ethnicity. Applicant ethnicity 4 has a higher chance of being accepted
6. **Preapproval:** If preapproval was requested, then the acceptance rate of the loan becomes reduced.

Other insights from the data exploration are;

- Most applications were of the conventional loan type (74 %).
- Most common property type was the one-to-four family dwellings (96 %).
- The loan purpose was mainly refinancing (49 %). See below figure for the Loan Purpose distribution.
- 89 % of applicants would be the owner's principal dwelling.
- Applicant ethnicity was primarily 'non-Hispanic/Latino' (77 %), most common race was 'White' (72 %), and the most common sex was 'male' (63 %). In the ethnicity and race features, combinations of similar distributions were attempted, but resulted in a lower accuracy of the model, thus were eventually left as is. This unfortunately shows the dependence of race on mortgage acceptance rate.
- 40 % of applicants signed with a co-applicant.

The bar plot is given below. The plot shows that there is no class bias in the provided dataset i.e. half of the applications were accepted and half were denied.



Machine Learning Modelling

For the creation of the binary classifier, I experimented with various algorithms and techniques to compare which gives a perfect accuracy after cross validating it using regular classification algorithms such as Logistic Regression, K-Nearest Neighbors and Support Vector Machines. The accuracy obtained was less than 70% which can be categorized as a poor accuracy base on metrics. The drive my decision to use more sophisticated algorithms and machine learning libraries of ensemble and bagging method like LightGBM, Adaboost, Xgboost, Gradient Boosting. Eventually, I got the highest accuracy (about 73.5%) with the Catboost algorithm. Before modelling was performed on the data, the data was prepared through;

1. Imputing missing information with the median values of the columns as well as replacing some with 0

2. Normalizing the data through necessary feature scaling method
3. Through the insights discovered from the exploratory data analysis phase, new features were created a process known as feature generation to Improve the model performances
4. Features which with distribution that were skewed were transformed.
5. There was no need to perform one hot encoding and label encoding on categorical features because the Catboost algorithm was built to be robust to solve that issue of working with categorical variables without transforming it.

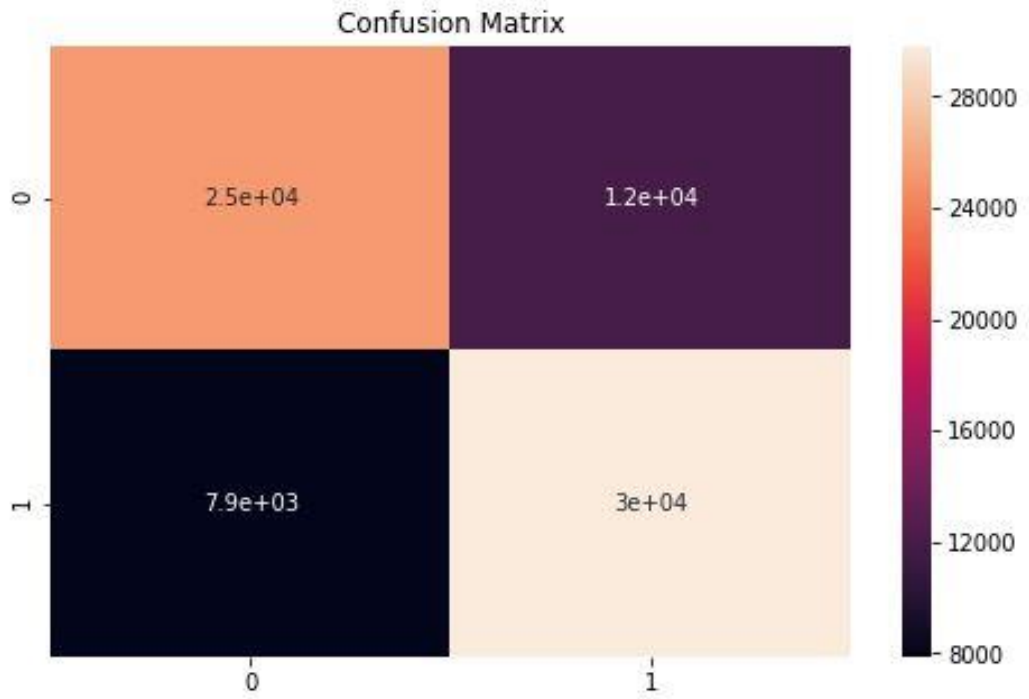
The model was created using the Catboost algorithm with over 1500 iterations. The model was trained on 80% of the data. Testing of the model was done on 20% of the data and it yielded the following results:

True Positives: 25000

True Negative: 30000

False Positives: 12000

False Negatives: 7900



Conclusion

The analysis has shown that mortgage acceptance can be predicted from given characteristics. The biggest factors deciding higher acceptance rate were: which lender the applicant applied to, not skipping information on the application, applying as a company, having your spouse or another co-applicant sign with you, applicant income, applicant race.

The model predicts reasonably well, however the number of false positives and negatives has to be taken into consideration, concluding that the model is not a perfect predictor. The author believes the reason a certain amount of variance/noise in the data is due to the fact that mortgages are partly determined by the lender (human interaction – differing from person to person) and by the lender's local system (automized).