Predicting Mortgage Approvals From Government Data

Francis Anokye, June 2019

Executive Summary

This document presents findings from the final capstone project for the Microsoft Professional Program in Data Science which emphasizes the full cycle of exploratory data analysis, statistical and machine learning modeling, and communicating the final results for actionable impact. We were challenged to consider how demographics, location, property type, lender, and other factors influenced the decisions on whether to accept or deny mortgage applications across the United States. By applying machine learning tools to this problem, the author hopes that the created model would be able to generalize to predict whether a mortgage loan application was accepted or denied according to the given dataset adapted from the Federal Financial Institutions Examination Council, FFIEC which is provided by DrivenData.

Data exploration was done by calculating summary and descriptive statistics and visualizations to examine the relationship between features that were available in the data. The training dataset consisted of 500,000 observations across 23 features, with a unique row identifier, and a labeled acceptance outcome. The findings led us to know the features to exclude from the model as well as suggested what would likely help us to predict whether a mortgage loan application was accepted or denied. Seven (7) different predictive models for the binary classification task were built using accuracy (classification rate) as the performance metric. The Catboost package from Yandex, which uses boosted gradient decision trees performed best with an accuracy score of 0.7309 using 20 features which involved an engineered feature called income_to_loan_ratio on our test data and further achieved an accuracy score of 0.7297 on the 500,000 test set hosted on DrivenData.

After the modeling, we were also interested in applying the feature importance method as a way to interpret how the given features contributed to the performance of the model. Using an arbitrary threshold of 5 for our top feature selection, we observed that the under listed features were the most influential contributors in the model's performance:

- lender
- county code
- state code
- applicant income
- income to loan ratio
- loan purpose
- preapproval

We can confidently conclude that even without the standard key features such as FICO (Fair Isaac Corporation) score, occupation, total household income, DTI (debt to income) ratio, loan to value ratio (LTV) and credit profiles, we can efficiently predict with an accuracy of ~73.1% whether to accept or deny a mortgage loan application using data from the applicant's demographic location, readily available information from census and the applicant's and loan information.

Data Description

The training dataset is made of 500,000 observations with 21 unique features and a single target variable which excluded the row_id in the train values and train labels which subsequently was assigned as the index for our data frame. The train labels were in a separate file which was merged together with our train values using the row_id as the key. Below is a short description of each feature in the data set separated into Boolean, numeric and categorical data types:

Boolean:

- **Co_applicant** - Indicates whether there is a co-applicant (often a spouse) or not.

Numeric:

- **Loan_amount** - Size of the requested loan in thousands of dollars.
- **Population**- Total population in tract.
- **Minority_population_pct**- Percentage of minority population to total population for tract.
- **Ffiecmedian_family_income**- FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC).
- **Tract_to_msa_md_income_pct**- percentage of tract median family income compared to MSA/ MD median family income.
- **Number_of_owner-occupied_units**- Number of dwellings, including individual condominiums, that are lived in by the owner.
- **Number_of_1_to_4_family_units** - Dwellings that are built to house fewer than 5 families.

Categorical:

- **Msa_md** - Indicates Metropolitan Statistical Area/Metropolitan Division where a value of -1 indicates a missing value.
- **State_code** - Indicates the US state where a value of -1 indicates a missing value.
- **County_code** - Indicates the county where a value of -1 indicates a missing value.
- **Lender** (categorical) - Indicates which of the lenders was the authority in approving or denying this loan.
- **Loan_type** - Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured.
- **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.
- **Loan_purpose** - Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing.
- **Occupancy** - Indicates whether the property to which the loan application relates will be the owner's principal dwelling.
- **Preapproval** - Indicates whether the application or loan involved a request for a preapproval of a home purchase loan.
- **Applicant_income** - In thousands of dollars.
- **Applicant_ethnicity** - Ethnicity of the applicant.
- **Applicant_race** - Race of the applicant.

- **Applicant_sex** - Sex of the applicant.
- **Accepted** - Indicates whether the mortgage application was accepted (successfully originated) with a value of 1 or denied with a value of 0.

Data Exploration and Transformation

Numeric Features

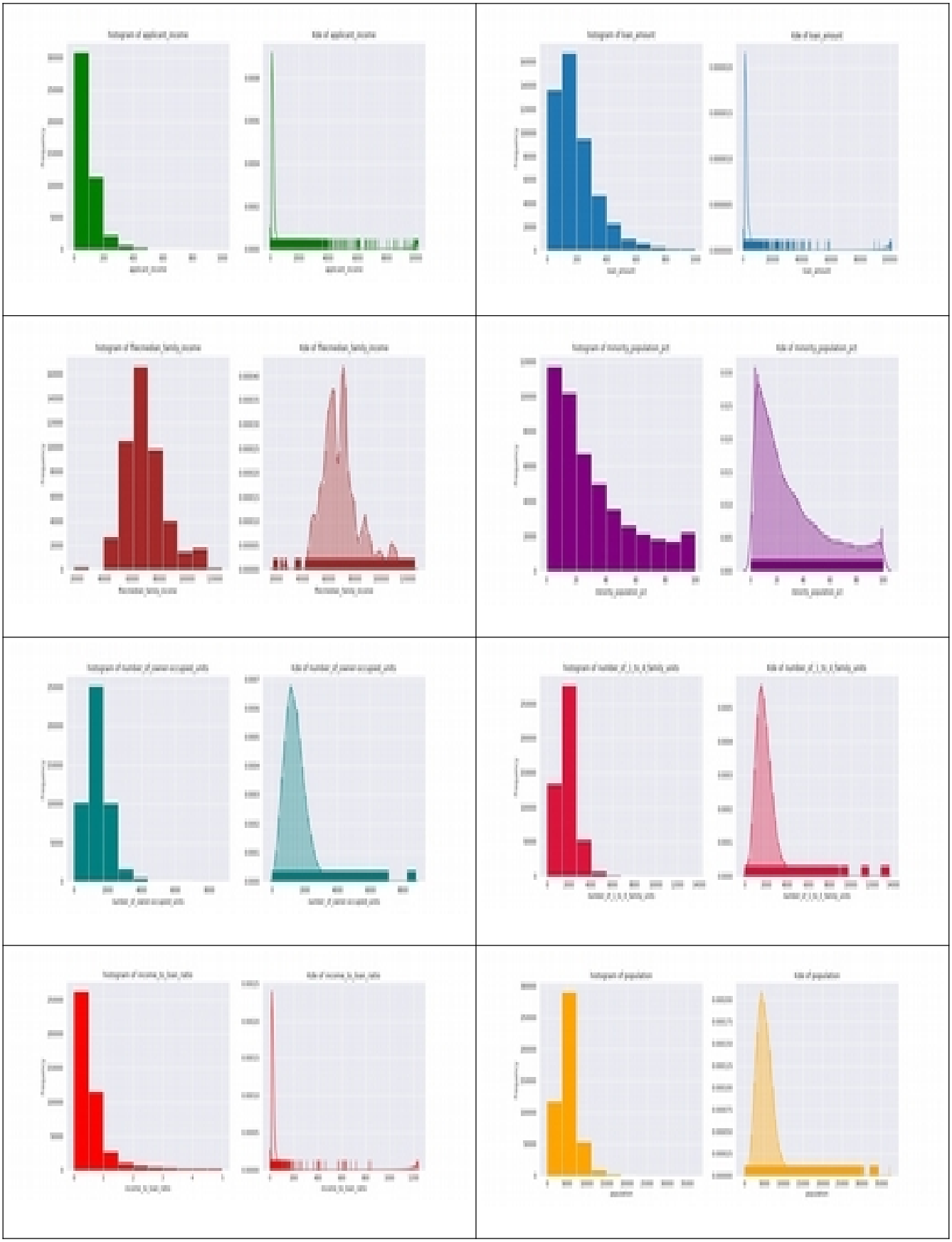
Summary statistics were examined for all the numeric features. It was observed that for all but tract_to_msa_md_income_pct column had mean values greater than their median values (50%), which indicated that the feature distributions were skewed to the right. The tract_to_msa_md_income_pct column was the only numeric feature skewed to the left. As depicted below, loan_amount had a mean value of ~222, but its median value was 162, indicating that its distribution was right skewed. Count, mean, standard deviation, minimum value, 25% quartile, 50% quartile (median), 75% quartile, and maximum value were included within the summary statistics as shown below:

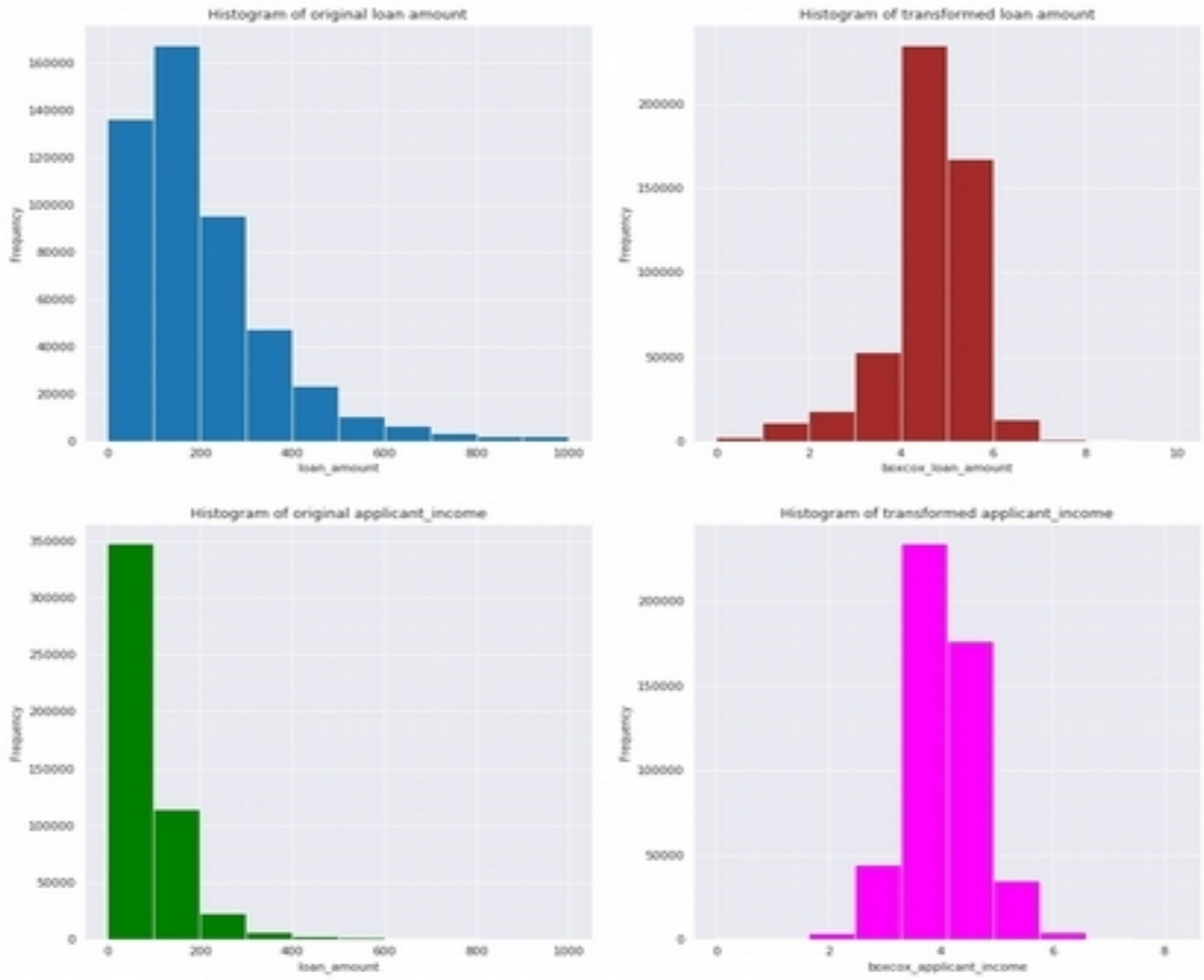
```
In [93]: # Descriptive statistics of the train_values, transposed to prevent horizontal scrolling
         train_values.describe().T
```

Out[93]:

	count	mean	std	min	25%	50%	75%	max
row_id	500000.0	249999.500000	144337.711634	0.000	124999.75000	249999.500	374999.25	499999.0
loan_type	500000.0	1.366276	0.690555	1.000	1.00000	1.000	2.00	4.0
property_type	500000.0	1.047650	0.231404	1.000	1.00000	1.000	1.00	3.0
loan_purpose	500000.0	2.066810	0.948371	1.000	1.00000	2.000	3.00	3.0
occupancy	500000.0	1.109590	0.326092	1.000	1.00000	1.000	1.00	3.0
loan_amount	500000.0	221.753158	590.641648	1.000	93.00000	162.000	266.00	100878.0
preapproval	500000.0	2.764722	0.543061	1.000	3.00000	3.000	3.00	3.0
msa_md	500000.0	181.606972	138.464169	-1.000	25.00000	192.000	314.00	408.0
state_code	500000.0	23.726924	15.982768	-1.000	6.00000	26.000	37.00	52.0
county_code	500000.0	144.542062	100.243612	-1.000	57.00000	131.000	246.00	324.0
applicant_ethnicity	500000.0	2.036228	0.511351	1.000	2.00000	2.000	2.00	4.0
applicant_race	500000.0	4.786586	1.024927	1.000	5.00000	5.000	5.00	7.0
applicant_sex	500000.0	1.462374	0.677685	1.000	1.00000	1.000	2.00	4.0
applicant_income	460052.0	102.389521	153.534496	1.000	47.00000	74.000	117.00	10139.0
population	477535.0	5416.833956	2728.144999	14.000	3744.00000	4975.000	6467.00	37097.0
minority_population_pct	477534.0	31.617310	26.333938	0.534	10.70000	22.901	46.02	100.0
ffiecmedian_family_income	477560.0	69235.603298	14810.058791	17858.000	59731.00000	67526.000	75351.00	125248.0
tract_to_msa_md_income_pct	477486.0	91.832624	14.210924	3.981	88.06725	100.000	100.00	100.0
number_of_owner-occupied_units	477435.0	1427.718282	737.559511	4.000	944.00000	1327.000	1780.00	8771.0
number_of_1_to_4_family_units	477470.0	1886.147065	914.123744	1.000	1301.00000	1753.000	2309.00	13623.0
lender	500000.0	3720.121344	1838.313175	0.000	2442.00000	3731.000	5436.00	6508.0

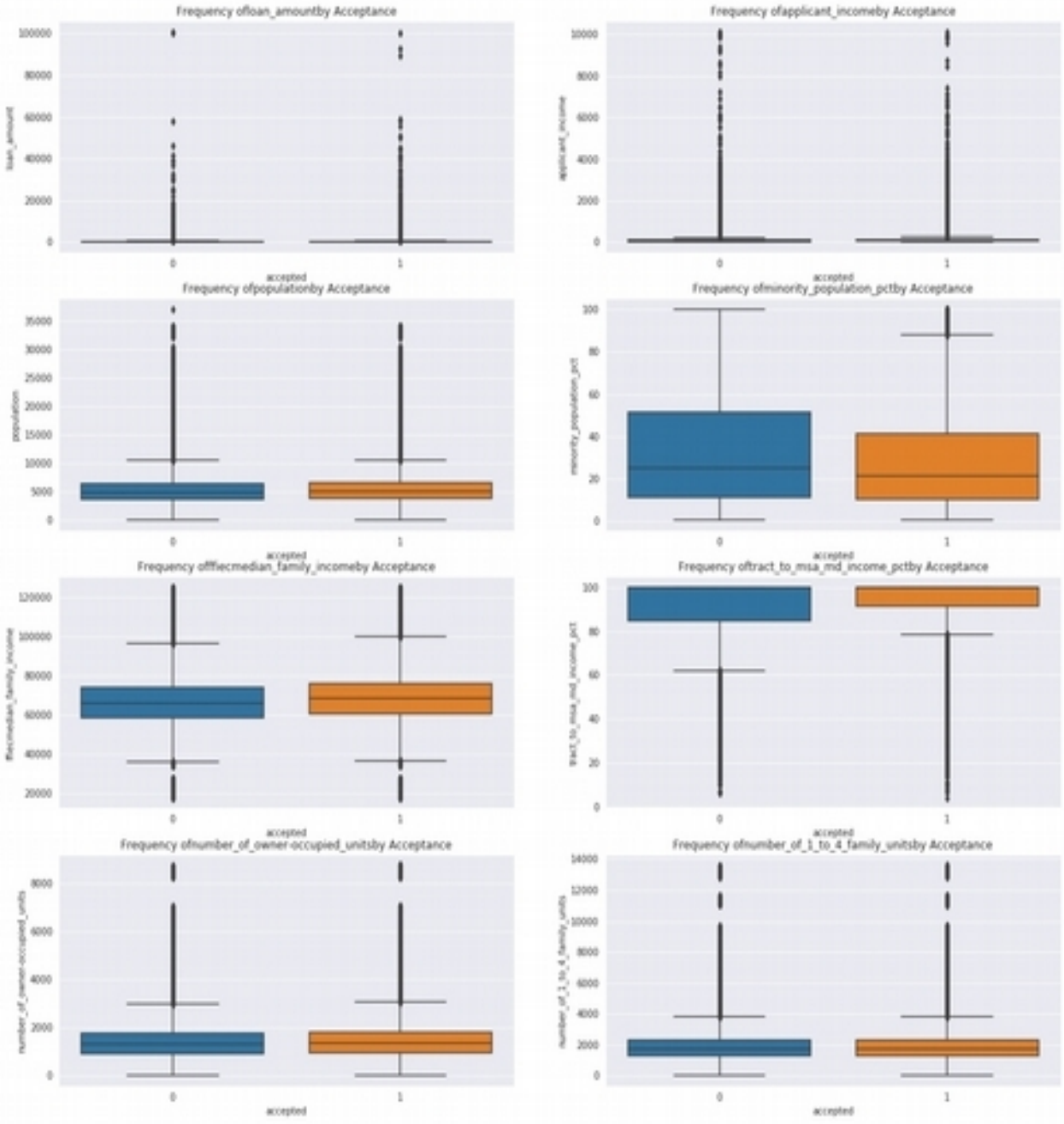
The skewed nature of the numeric features indicate the presence of extreme values (outliers) that require further investigation. Below is the visualization of our numeric features:


Histogram of applicant_income
KDE of applicant_income
Histogram of loan_amount
KDE of loan_amount
Histogram of Fbarmedian_family_income
KDE of Fbarmedian_family_income
Histogram of minority_population_pct
KDE of minority_population_pct
Histogram of number_of_owner-occupied_units
KDE of number_of_owner-occupied_units
Histogram of number_of_1_to_4_family_units
KDE of number_of_1_to_4_family_units
Histogram of income_to_loan_ratio
KDE of income_to_loan_ratio
Histogram of population
KDE of population

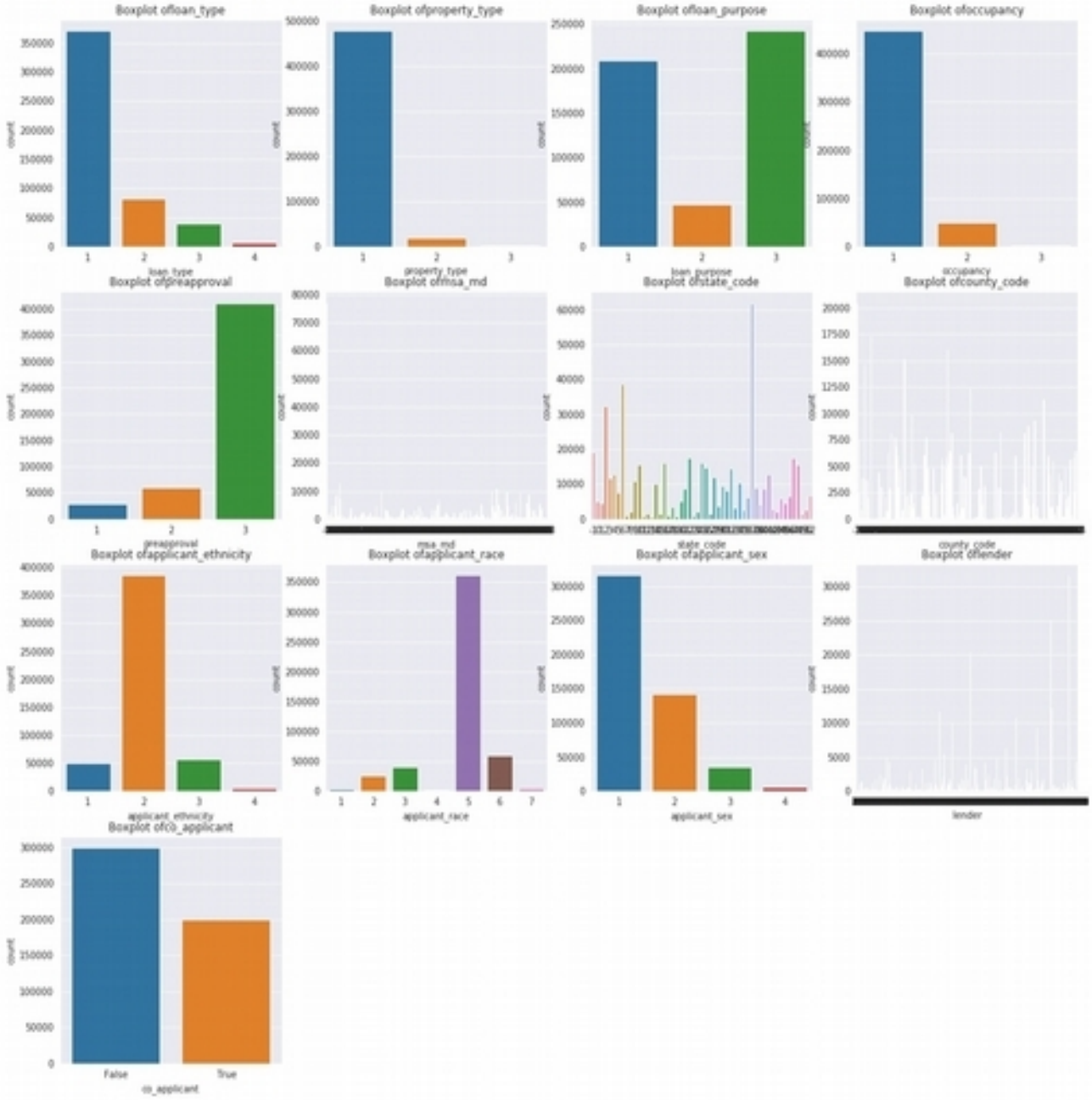


A new numerical feature, income_to_loan_ratio, was created by dividing the applicant income feature value by the loan amount feature value for each row of data. This feature was generated as an effort to mimic the more traditional industry standard of debt-to-income ratio. All the eight numeric features were initially transformed due to their high levels of skewness using the box-cox method from the stats module in Scipy, but none was used in our final model due to their negligible impact as compared to their default forms. This could be associated with their massive loss of information, hence, the notion of intentional outliers was considered as our motive for resorting to the default features. See below loan amount and applicant income with their respective transformed visualizations.

The presence of the extreme values in the numeric features were confirmed during our attempt to visualize the relationship between the numeric features and the target variable in the data .



Categorical Features

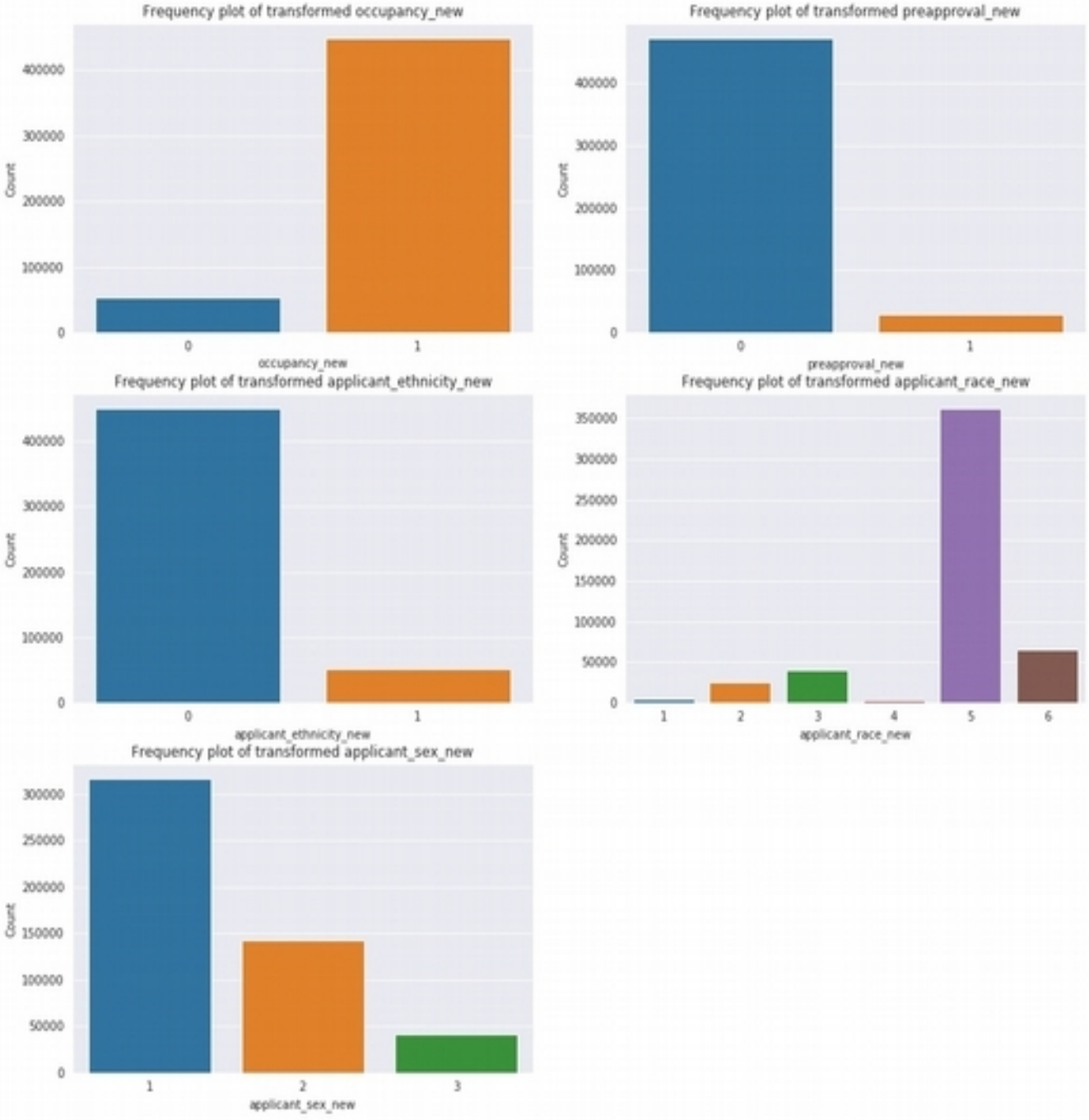


Default Categorical Features

Bar charts above were created to show the frequencies of the categorical features. Some useful observations were made which have been summarized in the following:

- Males were the majority representing 63.2% of total loan applicants and Females form about 28.6% and the remaining 8.2% did not reveal their gender status during the application process.
- Whites were the obvious majority in loan application representing 72.3% of the total loan applications. Blacks or African Americans followed with about 8.1% of the loan applications. Asians also constituted about 5.2% and native Hawaiian or other pacific Islanders forming about 0.5% of the total applicants. About 12.0% of loan applicants did not provide information about their race for reasons best known to them.
- Only 10.2% of loan applicants were either Hispanic or Latino. The remaining were neither Hispanic nor Latino as indicated in the applications.
- About 48.5% of loan applicants did apply for the mortgage loans to refinance other existing loans. And 41.8% of the applicants did so for home purchasing, and a little below 9.6% applied to improve their homes.
- It was observed that only 5.7% of loan applicants for home purchase loan did request for preapproval. The remainder had no preapproval.
- About 95.6% of the loan applications were for One-to-four-family dwelling.
- About 89.5% of the loan applications related to properties in which the applicants occupied as their principal dwelling. 10.1% of the properties were not owner-occupied and the remaining did not relate with the properties.
- Approximately 74.2% of the loan applications constituted conventional loans i.e any loan other than FHA, VA,FSA, or RHS loans. The Federal Housing Administration (FHA-insured) loan represented a little above 16.5% and Veterans Administration (VA-guaranteed) loans, also formed about 7.9% of the total loan types. The Farm service Agency or Rural Housing Service (FSA/RHS) formed a little above 1.4% of the total loan applications.
- Total missing values for county_code column was observed to be 20466 representing 4.1% of the total county_code data. County code contains 318 unique county codes.
- Total missing values for state_code column was observed to be 19132 also representing 3.8% of the total state_code data. State code contains 52 unique state codes.
- Total missing values for msa_md column was observed to be 76982 also representing 15.4% of the total county_code data.

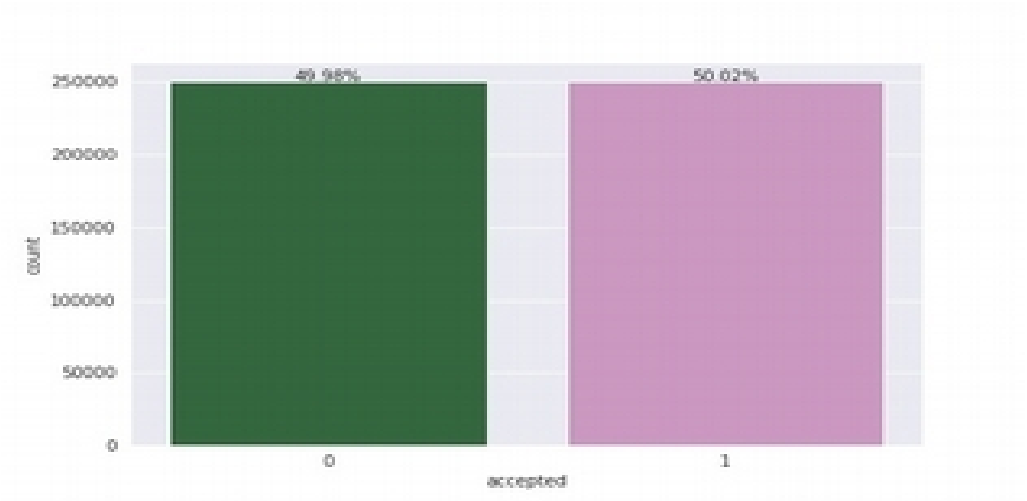
It is conspicuous in the visualization of the categorical features that many contained several levels which had relatively very low frequencies. So it was initially decided during the exploration that occupancy would be reduced from 3 categories to 2, preapproval from 3 to 2, applicant ethnicity from 4 to 2, applicant race from 7 to 6, and applicant sex from 4 to 3. The state code, county code and msa_md features were not altered to reflect the different demographics and location of the loan applicants. However, these transformations had no much information to contribute to the model's performance, so they were dropped and the default features used in the final model. Below is the visualization of the engineered categorical features which were later dropped.



Transformed Categorical Features

Target Variable

The target variable, accepted, showed a balanced split between the mortgage applications that were accepted and those that were denied as illustrated in the visualization below



Dealing With Missing Values

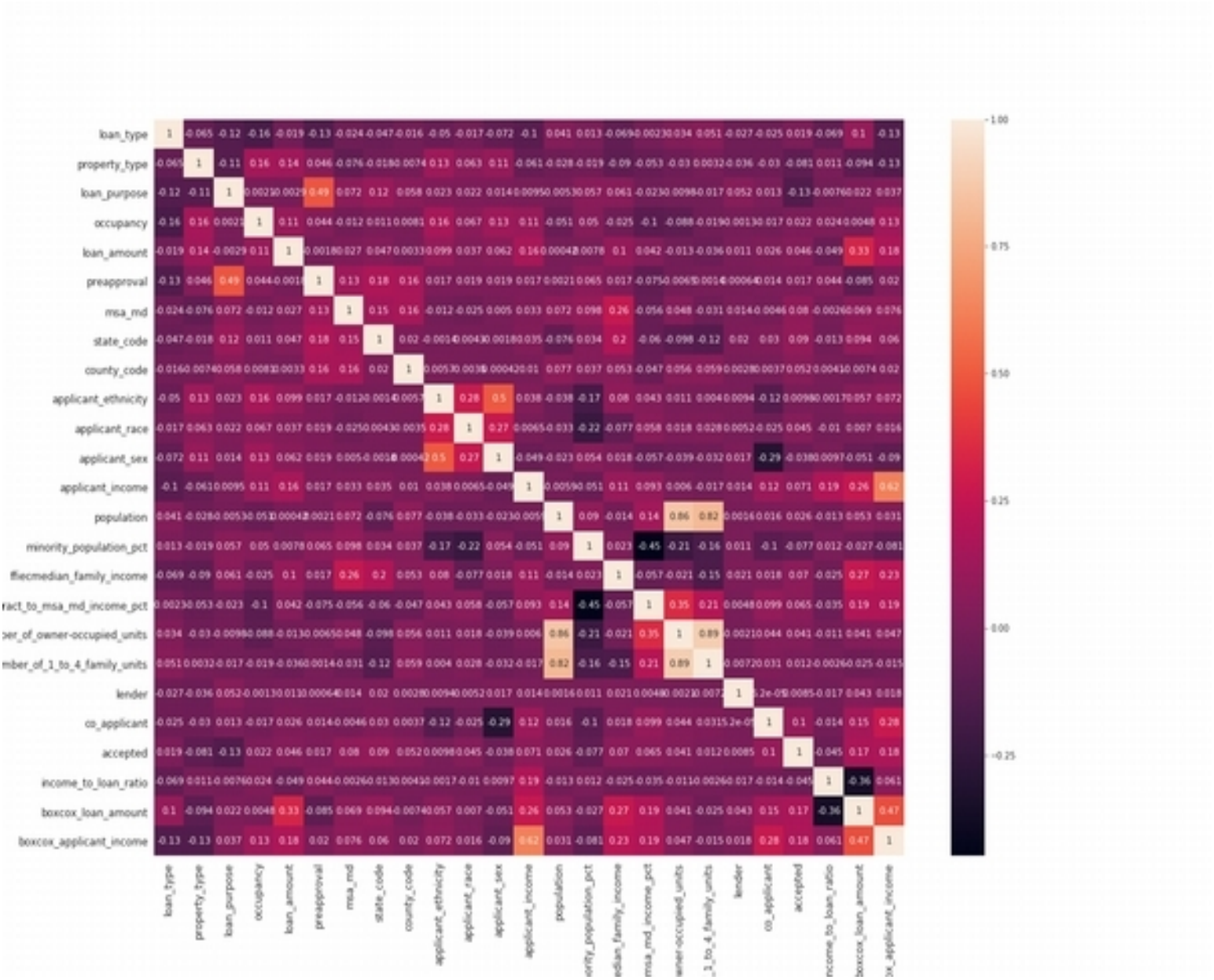
A simple count of missing data within each feature was tabulated and it was decided that the small amount of missing data, under 15.5%, was not large enough to justify removing the entire features. Instead, missing values were imputed using the median value for each of the numerical feature columns. The categorical feature columns indicated by '-1' were left alone and not replaced, since it was thought that the '-1' designation could have meaningful predictive qualities.

Percentage of missing data in the numeric features

Name of Feature	Percentage missing (%)
applicant_income	7.9896
population	4.4930
minority_population_pct	4.4932
ffiecmedian_family_income	4.4880
tract_to_msa_md_income_pct	4.5028
number_of_owner-occupied_units	4.5130
number_of_1_to_4_family_units	4.5060
Total	34.9856

Relationships and Correlations

Using the heat map below, we attempt to explain the presence or absence of relationships between the features and not the nature of their relationships using their correlation coefficients.



The target variable, accepted, positively relates with all the predictor features except for minority population pct, applicant sex and loan purpose. There is a moderate association also between the sex of an applicant and his/her ethnicity. Most visible and noteworthy findings include the very strong positive relationships that exist between the three features: number_of_owner-occupied_units, population and number_of_1_to_4_family_units. Loan purpose had a positive relationship with the preapproval feature, and smaller positive relationships were observed between

ffiecmedian_family_income and the three features msa_md, state_code, and applicant_income. In our attempt to understanding the true relationship between applicant_income and loan_amount, a regplot from the Seaborn package was used to visualize their relationship which indicated that a higher applicant income is associated with a higher loan amount on average.

reg-plot showing relationship between applicant income and loan amount

Model Building

A binary classification model was needed in order to predict the outcome of acceptance or denial for the mortgage loan applications. Seven (7) machine learning classifiers were considered using the classification rate (accuracy) as the performance metric. The Catboost package from Yandex which uses gradient boosting on decision trees emerged as our best model given the above mentioned metric.

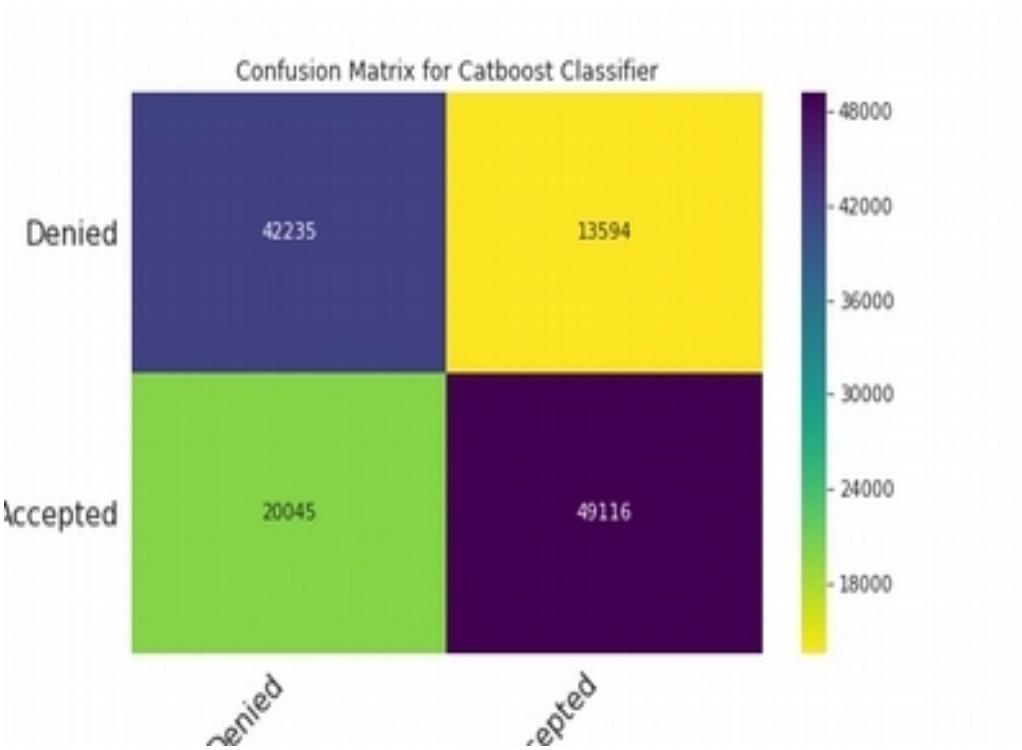
During the model build up, 20 features were used which excludes the population and number_of_owner-occupied_units features. This was as a result of their strong correlation with the number_of_1_to_4_family_units numeric feature. The only new feature which was included in our model was the income_to_loan_ratio. It is to be noted that no form of scaling or normalization was done to the data. Any attempt of that sort resulted in poor performance of our model. 43 duplicates data points were removed and categorical encoding were spared due to the inbuilt capability of Catboost to undertake this required task. The testing dataset that was used for prediction and subsequent submission was prepared in the same manner as the training dataset except for the removal of

duplicates. Missing numerical values were also imputed with their respective median feature values and the missing categorical values identified as '-1' were left unchanged.

The model was created using 75% of the training data for training and the remaining 25% for validation yielding the following results for the respective algorithms:

Summary of model performances

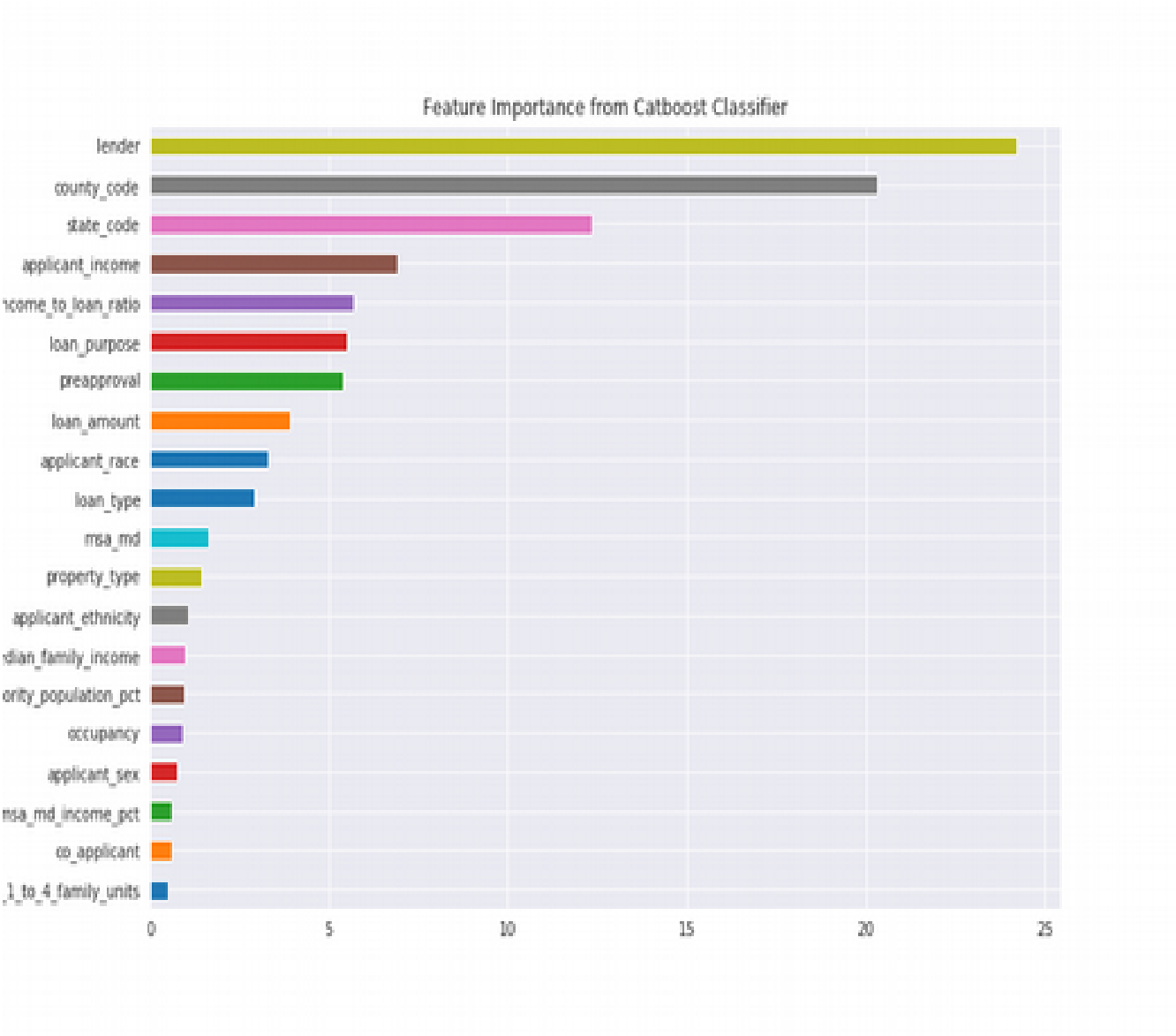
Name of Classifier	Accuracy Score
Logistic Regression	62.68
Decision Tree Classifier	68.68
Random Forest Classifier	70.36
Gradient Boosting Classifier	72.08
Xgboost Classifier	72.04
LightGBM	71.65
Catboost Classifier	**73.09**



The confusion matrix above translates into the following standard performance metrics for classification:

- Accuracy: 73%
- Precision: 74%
- Recall: 73%
- F1 Score : 73%

Below is the visual representation of how each feature column in model contributed to the model's performance.



Conclusion

The findings reported above indicate that the outcome of mortgage loan applications can be predicted with ~73.1% accuracy without having the key industry standard features, such as, credit score, debt to income ratios, the appraised value of the property for a loan to value ratio. Instead, the mortgage loan outcomes can simply be predicted using data from the applicant's demographic location, readily available information from census and the applicant's and loan information.

Limitations of the Model and Data

We were limited in our accuracy by the massive nature of outlier contents in our data, which in our thinking contained a lot of noise. In the business sense, we would likely build separate models for the outliers by collecting more examples or data points. We did attempt to control the noise through the use of classical approaches such as the exclusion of extreme values using the quartile and z-score normalization. However each of the approaches led to a decrease in the accuracy as the model's ability to predict the mortgage loan application degraded significantly.

References

1. Competition Site:
https://www.datasciencecapstone.org/competitions/14/mortgage-approvals-from-government-data/page/43/

2. Source of data:
https://www.datasciencecapstone.org/competitions/14/mortgage-approvals-from-government-data/data/

3. Catboost Package:
https://tech.yandex.com/catboost/

4. Machine Learning Analysis of Mortgage Credit Risk.
http://csis.pace.edu/~aleider/it691-19spring/credit.pdf