Predicting Mortgage Approvals From Government Data

Executive Summary of Analysis

This report presents an analysis of mortgage application. The analysis is based on a dataset which includes 500,000 mortgage applications, each containing specific characterisitics.

After performing basic exploratory data analysis, data visualization and some data preprocessing, several key features of the dataset, relationships between these features, were identified. These features were subsequently used to create a predictive classifier model with the functionality of predicting whether an application will be accepted or rejected. The predictive model yielded an accuracy of 73%.

After performing the analysis, significant features found were:

- **Lender**
- **Loan amount**
- **Applicant Income**
- **State Code**
- **Loan Purpose**

Exploratory Data Analysis

The initial exploration of the data began with some summary and descriptive statistics.

Dataset Info

- Number of Variable 23
- Number of Observations 500,000
- Missing Values 174,928 (35%)

The features of the dataset are:

1. row_id: a unique identifier
2. loan_type: a categorical feature indicating the type of loan
3. property_type: a categorical feature indicating the type of property
4. loan_purpose: a categorical feature indicating the purpose of the loan
5. occupancy: a categorical feature indicating whether the property will be the owner's principle dwelling

6. loan_amount: an int indicating the requested loan in thousands of dollars
7. preapproval: a categorical feature indicating if the loan involves a request for pre-approval
8. msd: a categorical feature indicating Metropolitan Statistical Area/Metropolitan Division
9. state_code: a categorical feature indicating U.S. state
10. county_code: a categorical feature indicating the county
11. applicant_ethnicity: a categorical feature indicating the ethnicity of the applicant
12. applicant_race: a categorical feature indicating the race of the applicant
13. applicant_sex: a categorical feature indicating the sex of the applicant
14. applicant_income: an int indicating the income in thousands of dollars
15. population: total population
16. minority_population_pct: Percentage of minority population to total population
17. ffiecmedian_family_income: FFIEC Median family income in dollars for the MSA/MD
18. tract_to_msa_md_income_pct: % of tract median family income compared to MSA/MD median family income
19. number_of_owner-occupied_units: Number of dwellings that are lived in by owner
20. number_of_1_to_4_family_units: Dwellings that are built to house fewer than 5 families
21. co_applicant: a bool feature indicating whether there's a co-applicant
22. lender: a categorical feature indicating authority in approving or denying a loan
23. accepted: the Target Variable, which indicates if the loan was accepted (1) or denied (0)

The dataset consists of 12 categorical features, 8 numerical features, 1 boolean feature and the target variable.

Individual Feature Statistics

Summary statistics for minimum, maximum, mean, standard deviation, and count were calculated for numeric columns, and the results taken from 500,000 observations are shown in the table below.

Columns	count	mean	std	min	max
loan_type	500000	1.36628	0.69056	1	4
property_type	500000	1.04765	0.23140	1	3
loan_purpose	500000	2.06681	0.94837	1	3
occupancy	500000	1.10959	0.32609	1	3
loan_amount	500000	221.75316	590.64165	1	100878
preapproval	500000	2.76472	0.54306	1	3
msa_md	500000	181.60697	138.46417	-1	408
state_code	500000	23.72692	15.98277	-1	52
county_code	500000	144.54206	100.24361	-1	324
applicant_ethnicity	500000	2.03623	0.51135	1	4
applicant_race	500000	4.78659	1.02493	1	7
applicant_sex	500000	1.46237	0.67769	1	4
applicant_income	460052	102.38952	153.53450	1	10139
population	477535	5416.83396	2728.14500	14	37097
minority_population_pct	477534	31.61731	26.33394	0.534	100
ffiecmedian_family_income	477560	69235.60330	14810.05879	17858	125248
tract_to_msa_md_income_pct	477486	91.83262	14.21092	3.981	100
number_of_owner-occupied_units	477435	1427.71828	737.55951	4	8771
number_of_1_to_4_family_units	477470	1886.14706	914.12374	1	13623
lender	500000	3720.12134	1838.31317	0	6508

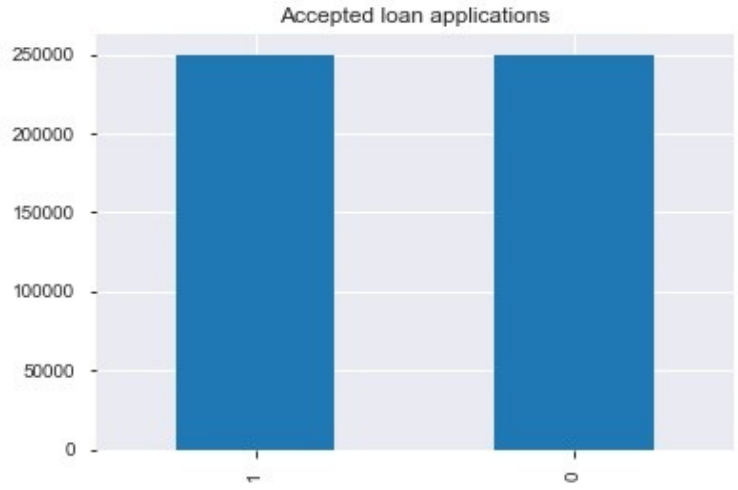
From the table, six(6) features have missing values;

- applicant_income – 39948 missing values
- population – 22465 missing values
- minority_population_pct – 22466 missing values
- ffiecmedian_family_income – 22440 missing values
- tract_to_msa_md_income_pct – 22514 missing values
- number_of_owner-occupied_units – 22565 missing values
- number_of_1_to_4_family_units – 22530 missing values

These missing values will be treated as this report progresses.

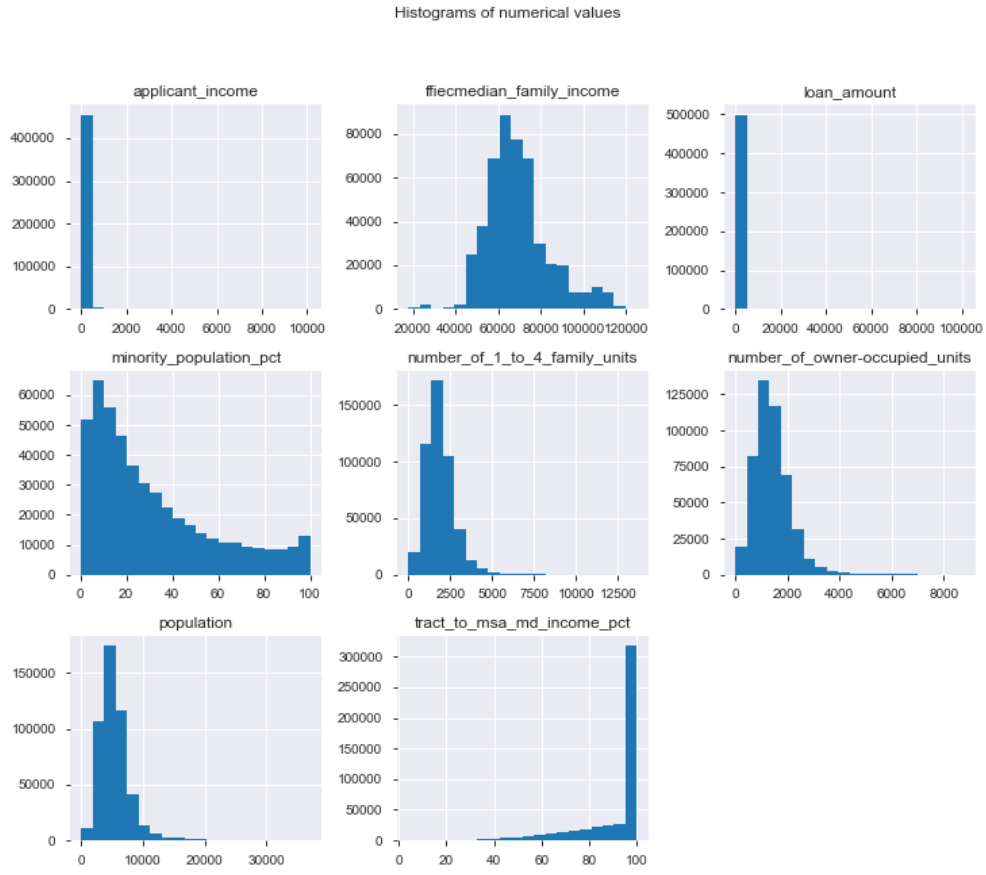
Data Visual Analysis

Since 'accepted' is of interest in this analysis, a bar chart was created to see the distribution of the target variable.

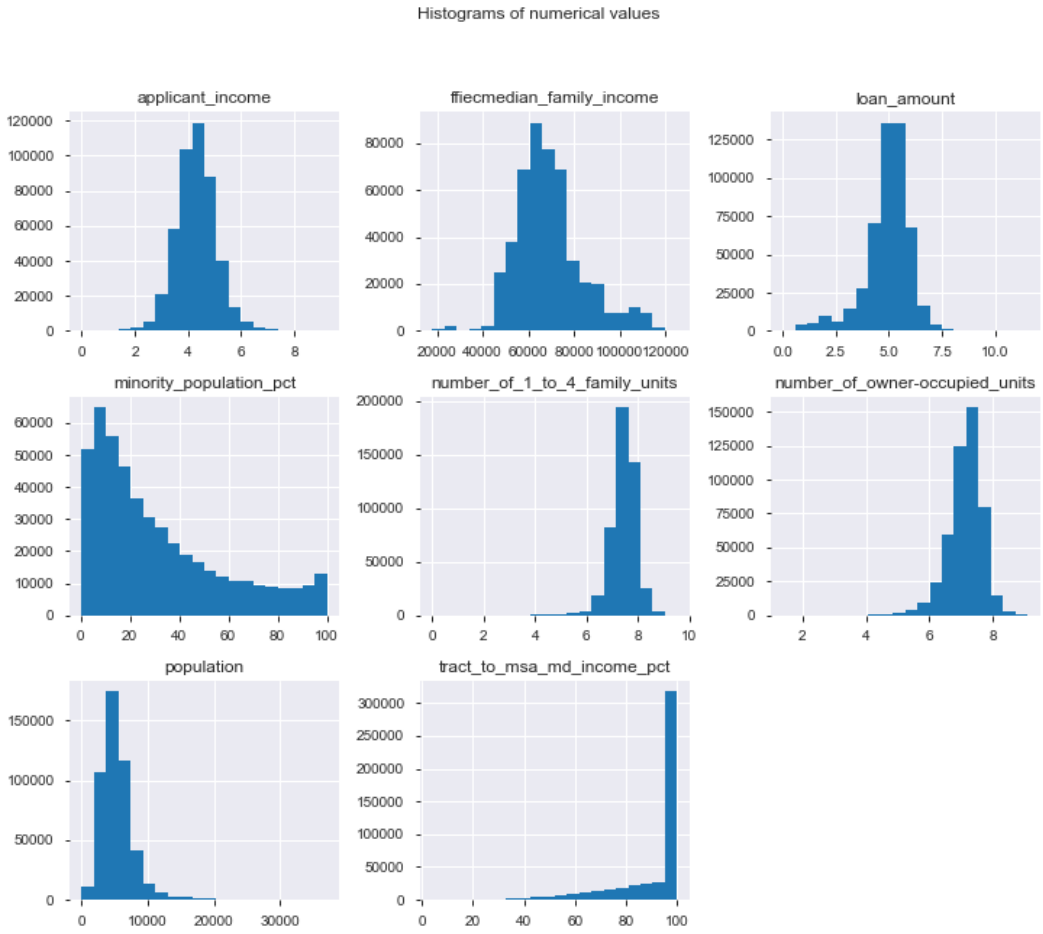


It can be seen from the chart that the dataset is almost balanced, with 250114 accepted applications and 249886 rejected applications.

In addition, a histogram was plotted to see the distribution of the numerical columns.



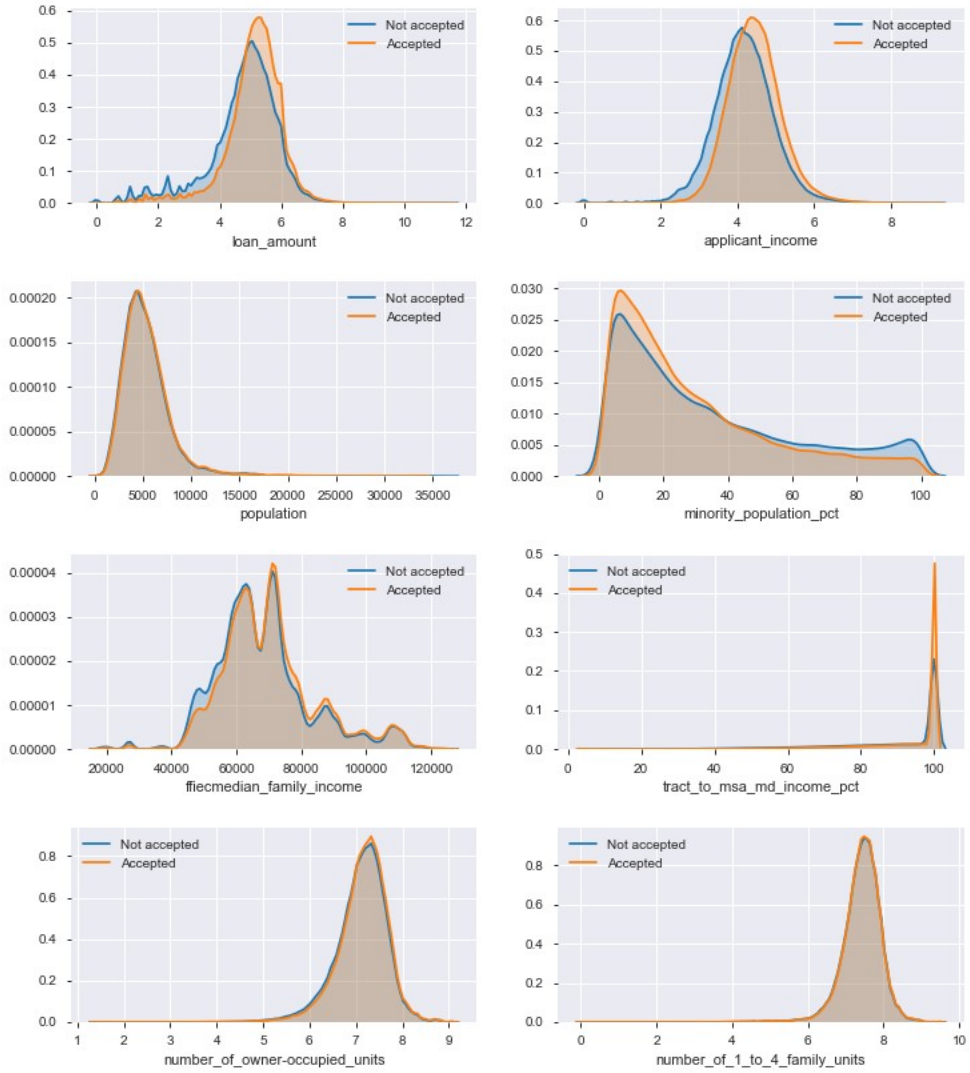
From the above chart, loan_amount and applicant_income are highly skewed which can make our machine learning model perform badly. By transforming the features using the logarithmic function, we can get a distribution close to normal.



After applying the logarithmic function, applicant_income and loan_amount distribution are now looking close to normal distribution.

Next, we make a KDE plot of the numerical features, to see the seperation of the target variable of each numerical features.

KDE plots



As shown in the plot above, the numerical features population, number_of_1_to_4_family_units, number_of_owner-occupied_units doesn't show much separation of the target variable and as such these features were dropped.

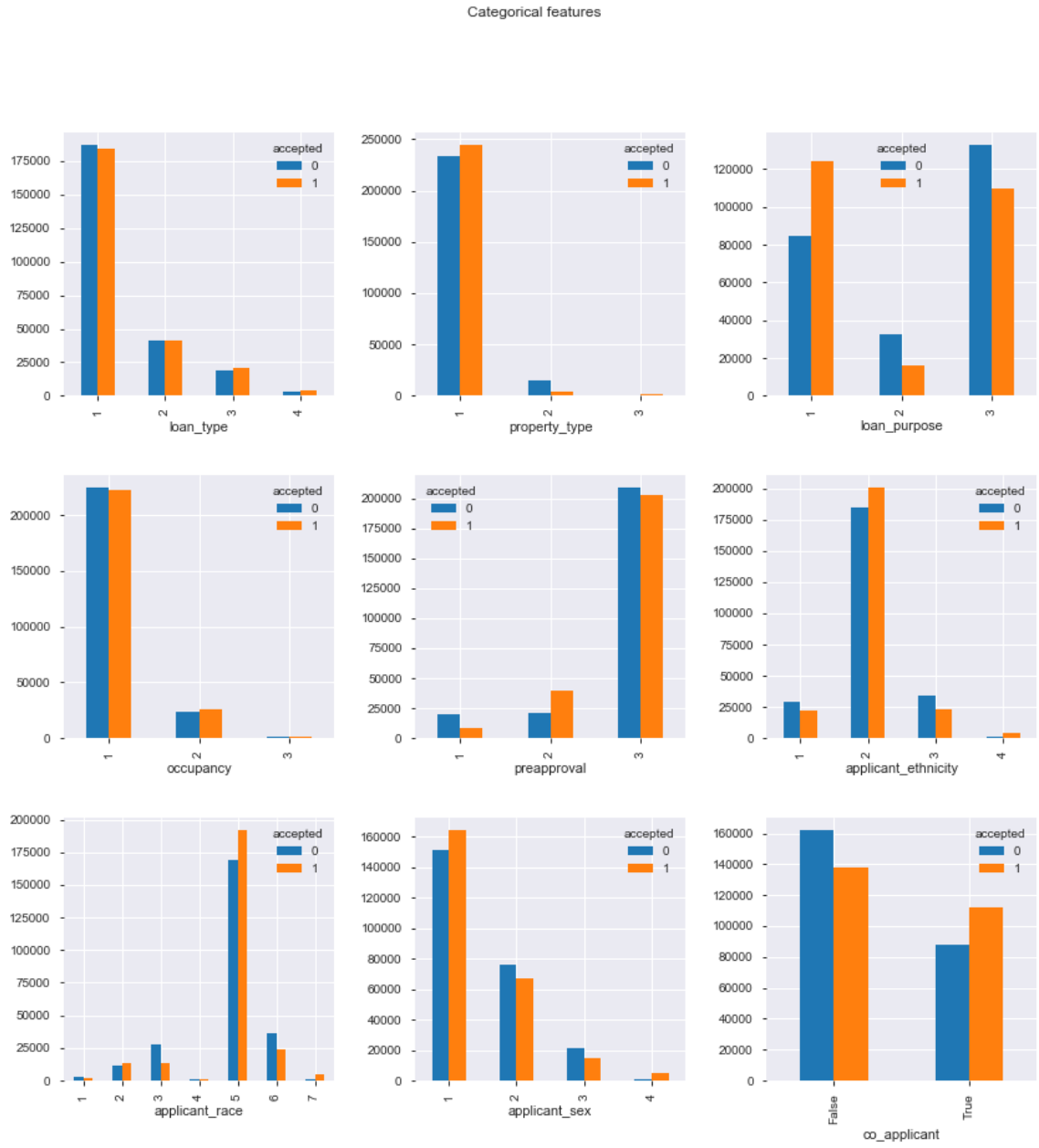
Relationship between Target Variable 'accepted' and some Categorical Features

An exploratory data analysis was carried out to see the relationship between the target variable and some categorical features such as:

- property_type

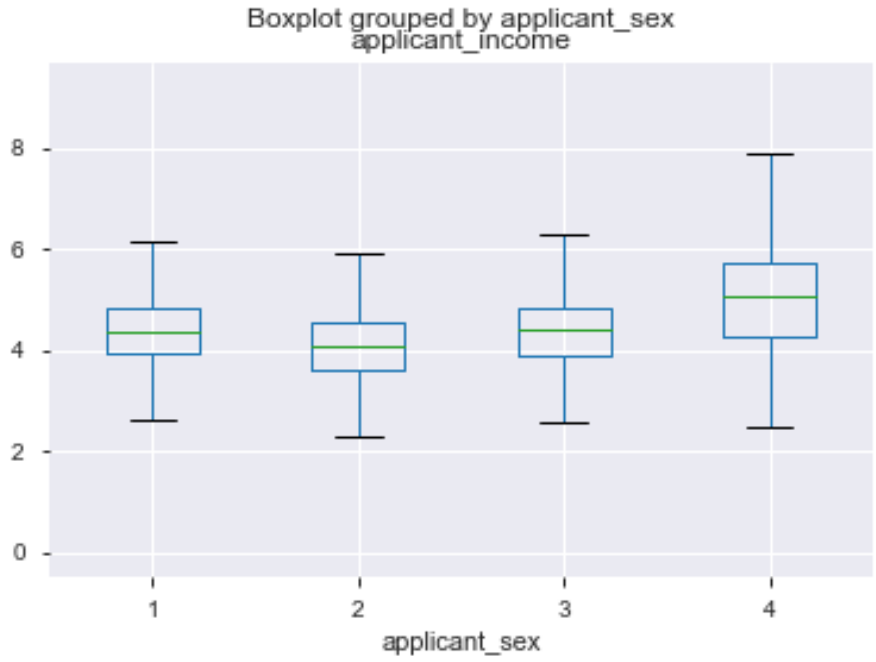
- loan_type
- preapproval
- loan_purpose
- applicant_sex

Charts were created to show the frequencies of these features against the target variable.



From the bar chart above, the following inference were made:

- It can be seen that loans for a one-to-four family properties were accepted more than the other two property types.
- Conventional loan types were approved more.
- Applicants whose purpose of loan where either, home purchasing or refinancing were granted more.
- Preapproval of a loan does not greatly affect the acceptance of loan application, as a large part of accepted applications did not request for preapproval or were not applicable.
- Women had a lower loan approval rate compared to men.
- The loan applications of white people had a higher approval rate compared to other races.
- Another key observation was that women earned less compared to men.



A new feature, ‘loan-income ratio’, was created. This shows the capacity of an applicant to pay back a loan.

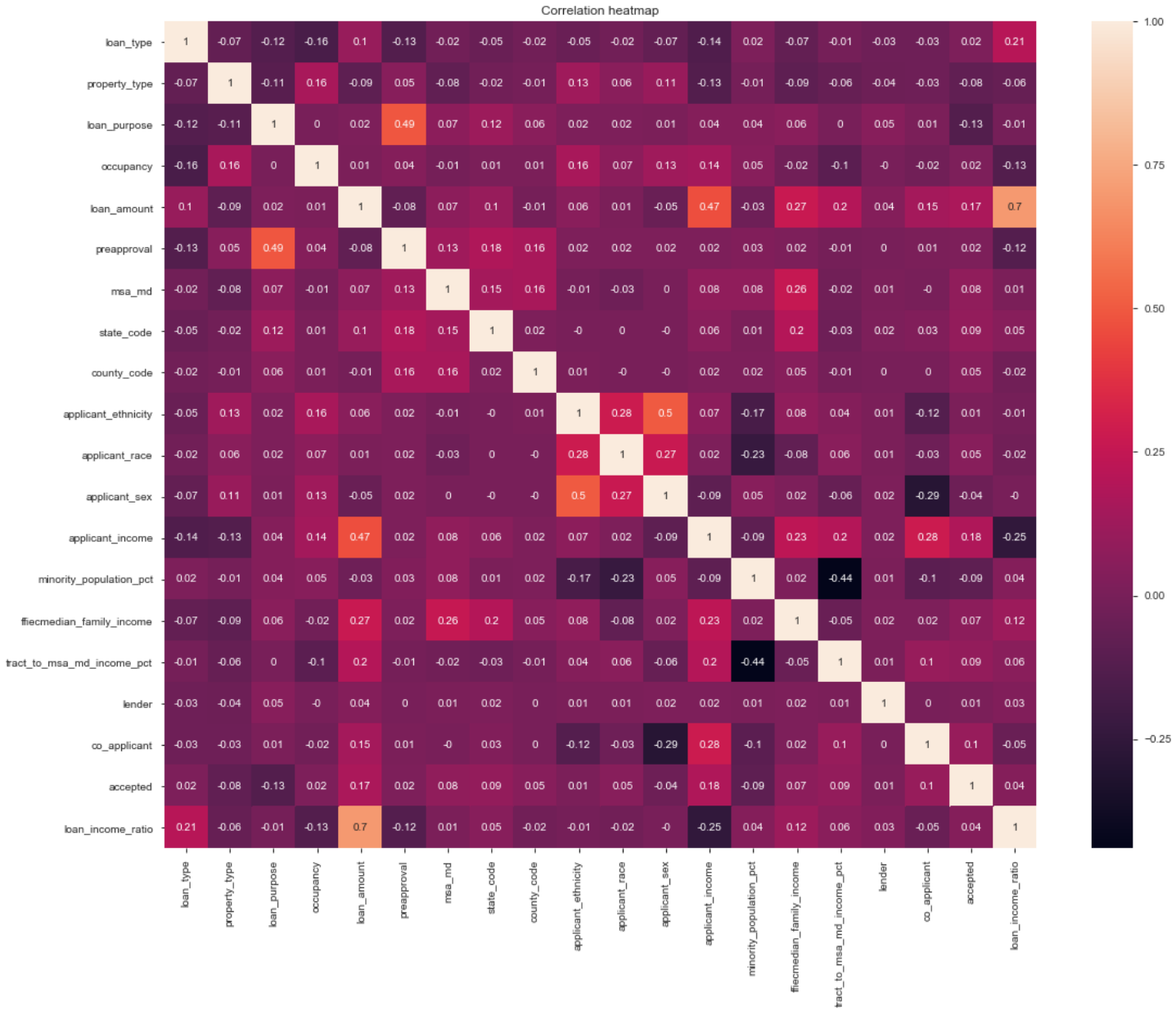
Features with missing values were either replaced by their median or mean.

- missing values in applicant_income were replaced by the median
- missing values in ffiecmedian_family_income were replaced by the median
- missing values in minority_population_pct were replaced by the mean
- missing values in tract_to_msa_md_income_pct were replaced by the mean

- missing values in loan_income_ratio were replaced by the median

Correlations

A correlation heatmap helps visualize features that are correlated.



From the heatmap, some key observations are:

- applicant_income and loan_amount have a correlation factor of ~0.5
- loan_amount and loan_income_ratio have a correlation factro of 0.7

Prediction of Loan Applications

Based on the analysis of the mortgage approval data, a predictive model to predict the acceptance of loan applications was created.

Classification algorithms such as Logistic Regression, Decision Tree Classifier, Random Forest Classifier were used to test. The accuracy of these algorithms were below 70% which was unacceptable.

A more performing model was created using the CatBoost Classifier algorithm and was trained with 85% of the data.

Testing the model with the remaining 25% of the data yielded a **test accuracy of 73.%**.

Conclusion

This analysis has shown that the accuracy of our classifier is satisfactory. Features such as, lender, applicant's income, loan's purpose, state code and loan amount have a significant effect on the approval of a loan.