Analysis of Mortgage approvals

Tomas Podprocky, April 2019

Executive summary

This document presents an analysis of the government data regarding mortgage approvals. The data analysis and results are based on the 500000 observations in 21 different features.

Analysis of the data and their relationships has been performed. After performing data imputation and feature engineering, a predictive model to classify the mortgage approval was created.

Below are the important findings regarding the features of the dataset:

- The numerical features are not normally distributed – most probably better fit with an extreme large value distribution.
- State code information does not follow the FIPS conventions
- When a state code is missing, it is also missing all the census information and in ~97% of cases the mortgage application is rejected in the training dataset
- The number of rejected applications for minorities is rising together with the increased minority percentage in a tract
- Each unique lender in the dataset can have an assigned acceptance (or rejection) ratio. This information can be used in addition to the existing features to predict whether the mortgage will be approved or rejected. This engineered feature has the highest importance for the model accuracy.
- The features 'occupancy' and 'loan type' play no role in influencing the 'accepted' label

Data exploration

The data exploration process first started with evaluating the descriptive statistics of the numerical features, understanding the number of missing values and then continued with understanding the relationships in the dataset.

Numerical feature statistics

The following table shows the descriptive statistic information for numerical features:

Table 1: Descriptive statistic of the numerical features of the test dataset.

	count	mean	median	std	min	max
loan_amount	500000	221.75	162	590.64	1	100878
applicant_income	460052	102.39	74	153.53	1	10139
population	477535	5416.83	4975	2728.14	14	37097
minority_population_pct	477534	31.62	22.901	26.33	0.534	100
ffiecmedian_family_income	477560	69235.60	67526	14810.06	17858	125248
tract_to_msa_md_income_pct	477486	91.83	100	14.21	3.981	100
number_of_owner-occupied_units	477435	1427.72	1327	737.56	4	8771
number_of_1_to_4_family_units	477470	1886.15	1753	914.12	1	13623

Correlations

It can be seen from Table 1, there are several numerical features with missing values, one of them being 'applicant income'. As this feature is part of applicant information, it can be assumed it may be an important feature for accepting or rejecting a mortgage application.

As a next step, a grid of scatter plots was created to compare the relationships between various numerical parameters. The scatter plots on Figure 1 were created from a random sample of 10000 from the training dataset and illustrate a difference between accepted and rejected mortgage application.

Figure 1: Scatter plots representing relationships between numerical features in the dataset based on a sample of 10000.

Viewing the scatter plots and distribution plots the following can be seen:

- There is no clear separation of data in cases when mortgage application was accepted, and when was rejected. Most of the data overlays on top of each other (see the overlayed distributions on the diagonal axis)
- The distributions for loan amount and applicant income are having a very long tail with high values. It would probably make sense to apply logarithmic translation on these data
- There seem to be a positive correlation between values from the census data. This can be further explored through a correlation matrix.

The matrix on Figure 3 shows a correlation value/ratio between all features (numerical and categorical) in the test dataset (code used from dpython package: https://github.com/shakedzy/dython). There are several interesting as well as expected relationships, amongst which the ones that stand out are:

- There is a significant correlation between “number of owner-occupied units” and “number of 1 to 4 family units”
- There is a significant correlation between “population” and “number of owner-occupied units”
- There is a correlation between ‘msa md’ and ‘ffiecmedian family income’
- There is a very weak correlation between “loan amount” and “applicant income”. It shows that there are many applicants with low income asking for high loans and vice versa.

Even though there is a weak correlation between “loan amount” and “applicant income”, this relation can be used to fill the missing data in the “applicant income” feature (Figure 2). Even though the R^2 statistic of such a regression is going to be very small, it is probably better than other methods of filling for missing data. Using linear regression:

- Intercept = 30.29
- Coefficient = 0.345
- R^2 ~ 23%

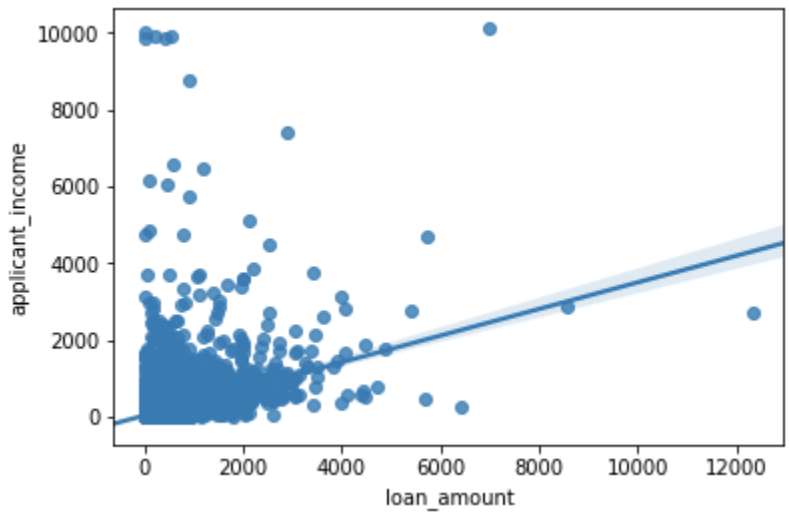


Figure 2: Visualizing the weak relationship between loan_amount and applicant_income.

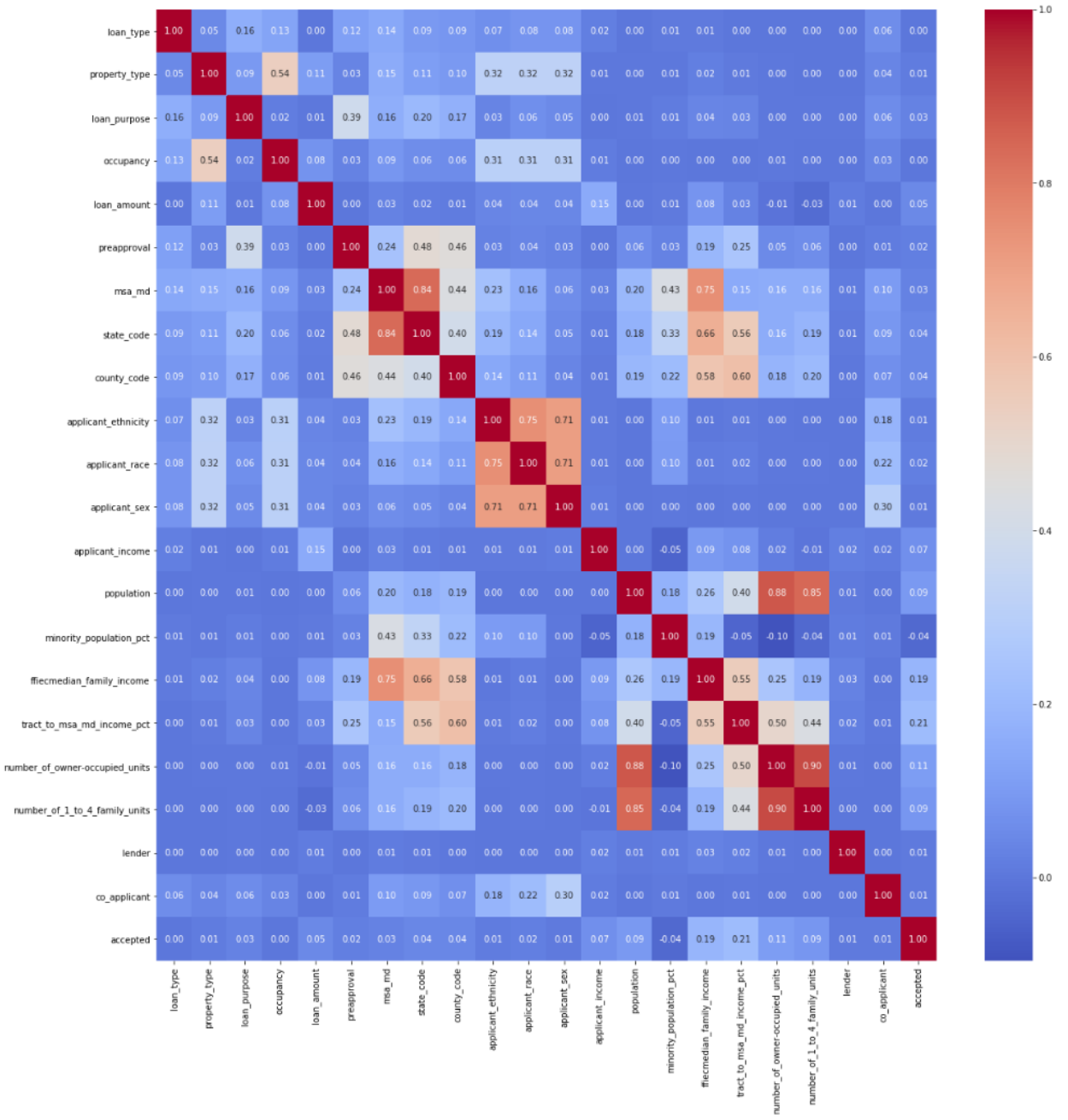


Figure 3: Correlation matrix between all features in the training dataset.

The correlation matrix also shows the relationship between the 'accepted' status and the other features in the dataset. It can be observed that there are two features which have a zero (to decimal places) correlation ratio to the 'accepted' label – 'loan type' and 'occupancy'.

Census information analysis

Each line of the dataset contains the census information related to a tract. The dataset also contains the race information and the relationship to census can be made using 'minority population pct' feature. From Table 1 it is clear, that the 'minority population pct' feature contains missing values. However, let's look at this feature with the missing values removed.

The applicant race information can be tied with the “minority population pct” through a new feature, which can simply distinguish whether an applicant belongs to a minority or not. Figure 4 below summarizes the number of accepted and rejected mortgage application of minority applicants, considering the minority percentage in the tract.

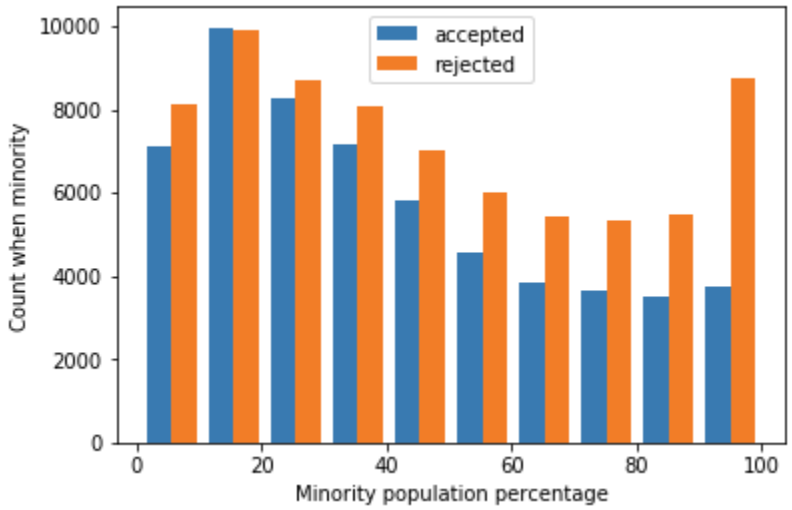


Figure 4: Relationship between minority population percentage and the number of accepted and rejected mortgage applications.

As the minority population percentage rises, so it the number of rejected mortgage applications. This negative relationship is also shown in the correlation matrix (Figure 3). However, any feature related to sex or race should be used with caution, as when used in a model it may cause discrimination.

Property location analysis

The training dataset contains categorical information regarding the property location. The property location contains 3 main features, which are categorical.

- State code – should be a two-digit FIPS state identifier (missing 19132 values, 3.8%)
- County code – should be a three-digit FIPS county identifier (missing 20466 values, 4%)
- msa_md – metropolitan statistical area/metropolitan division (missing 76982 values, 15.4%)

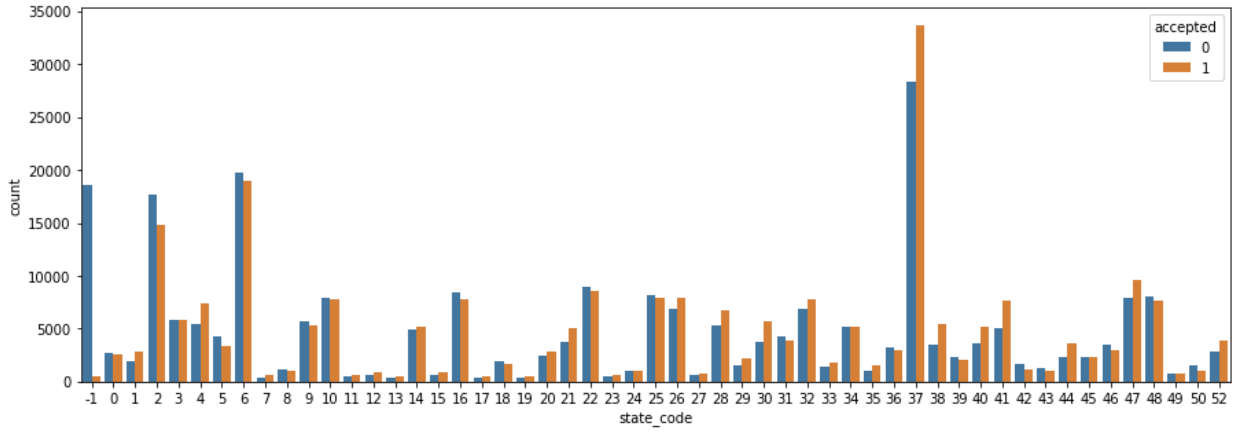
Because the data should be based on the FIPS codes, let’s first analyse if this data makes sense based on the official 2017 FIPS code database, which can be found on the following url https://www.census.gov/geographies/reference-files/2017/demo/popest/2017-fips.html.

The property location data analysis results in the following conclusions:

- Missing state code: 19132
- Max state code value: 52 – this however implies some states are missing completely from the dataset (e.g. as 53 is Washington, 54, West Virginia, etc.)
- State codes contain the following values – 3, 7, 14, 43, 52. These values are not part of the FIPS code database. In the past these were reserved codes.

Regarding the state code information included in the dataset – evidence shows it does not follow the FIPS code conventions, therefore splitting the data into larger categories as to different US regions may be not practical.

To continue the analysis, the figure below shows the number of accepted and rejected mortgage applications per state code



When the state code information is missing in the dataset, the application seems to be rejected almost in all cases (approximately in 97% it is labelled as rejected). Further analysis also shows, that when a state code is missing in the dataset, the all the census information is also missing.

Also, state code and county code are not independent categories. The same county code is assigned to counties in different states. Therefore, there it would be possible to have a single 'state county' code which would simply be a concatenated state code and county code. This approach would however result in an extremely high number of categories.

Lender column analysis

The dataset contains a categorical column 'lender' which is a categorical variable containing a unique lender code – institution considering the mortgage application.

There are 6111 unique lender codes in the dataset, only a single value has a lender code of 0. It is not feasible to use this data as categorical due to the high number of categories. A different approach is needed.

Let's investigate the number of accepted and rejected mortgage applications by each lender. This has been done only graphically and the resulting figure is split into two parts to achieve some level of plotting clarity.

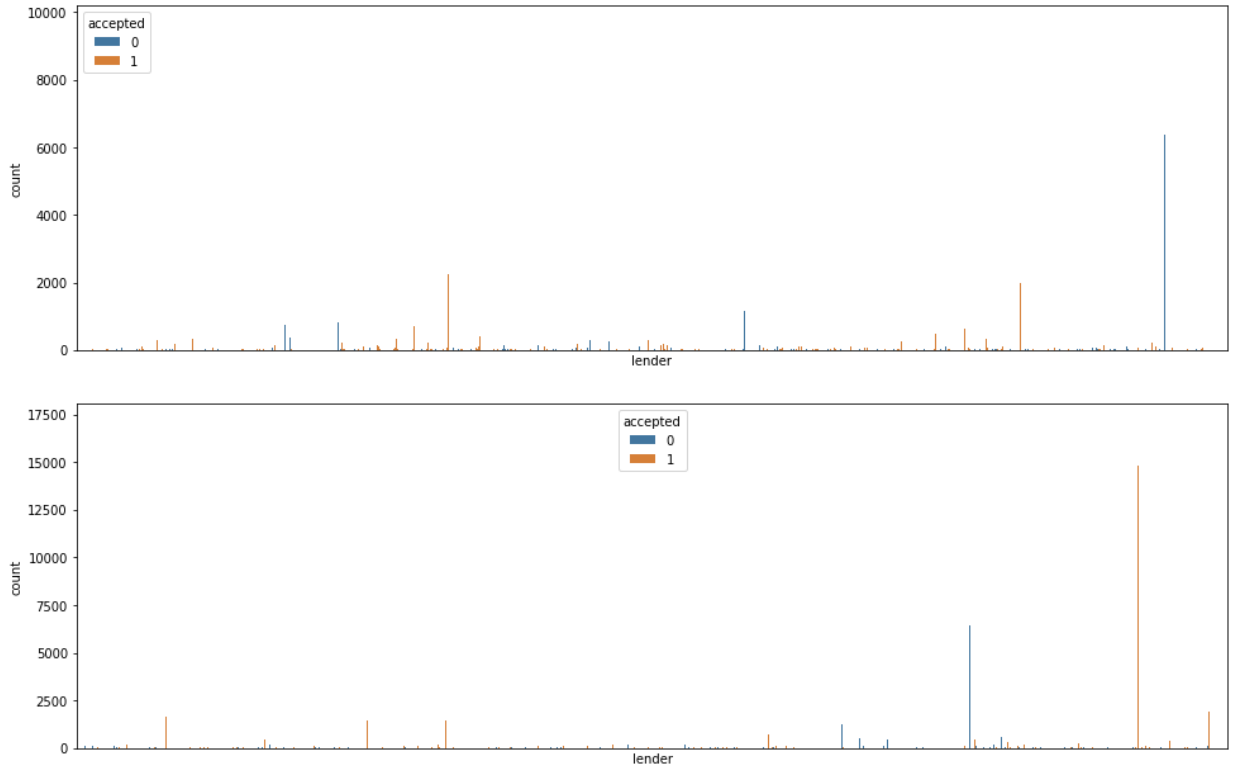


Figure 5: Number of approved and rejected application by lender code (The top figure shows lenders with codes up to 3000, the bottom figure is for lenders with a code above 3000).

Figure 5 shows that there are some lenders who have significantly higher acceptance rate than other lenders, which on the other hand have a significant reject rate. Therefore, one possibility is assigning to each lender an acceptance rate, based on the training data and training labels. The acceptance rate for each lender was saved in a separate csv file for use on the test dataset. The figure below illustrates the assigned "acceptance ratio" value with the frequency of a accepting or rejecting an application.

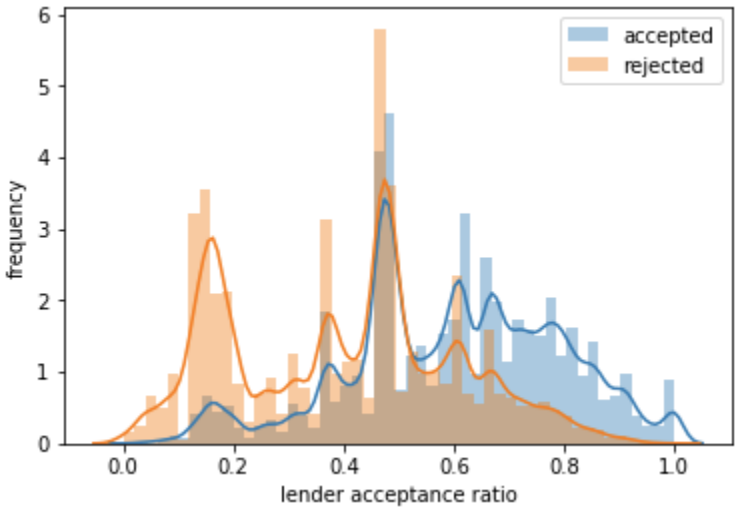
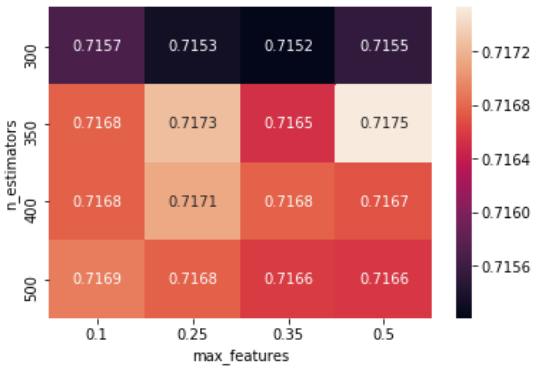


Figure 6: Lender acceptance ratio.

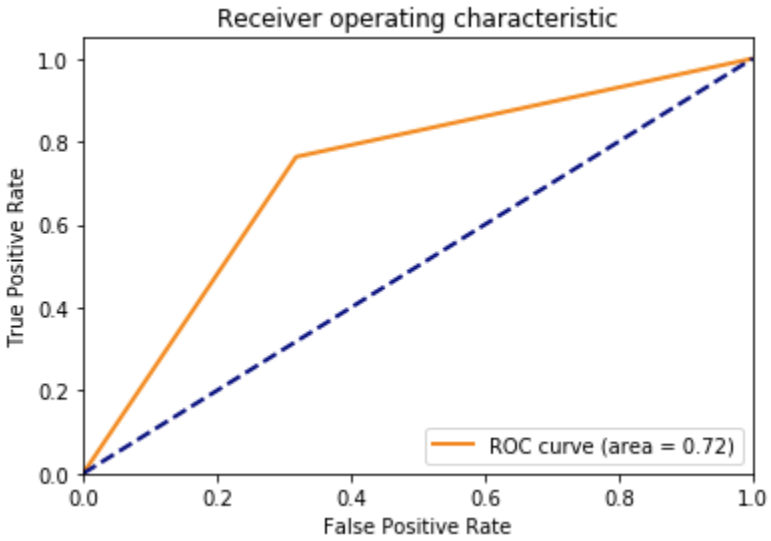
The importance of the lender acceptance ratio is illustrated in the next chapter – in the classification model evaluation.

Classification of mortgage acceptance

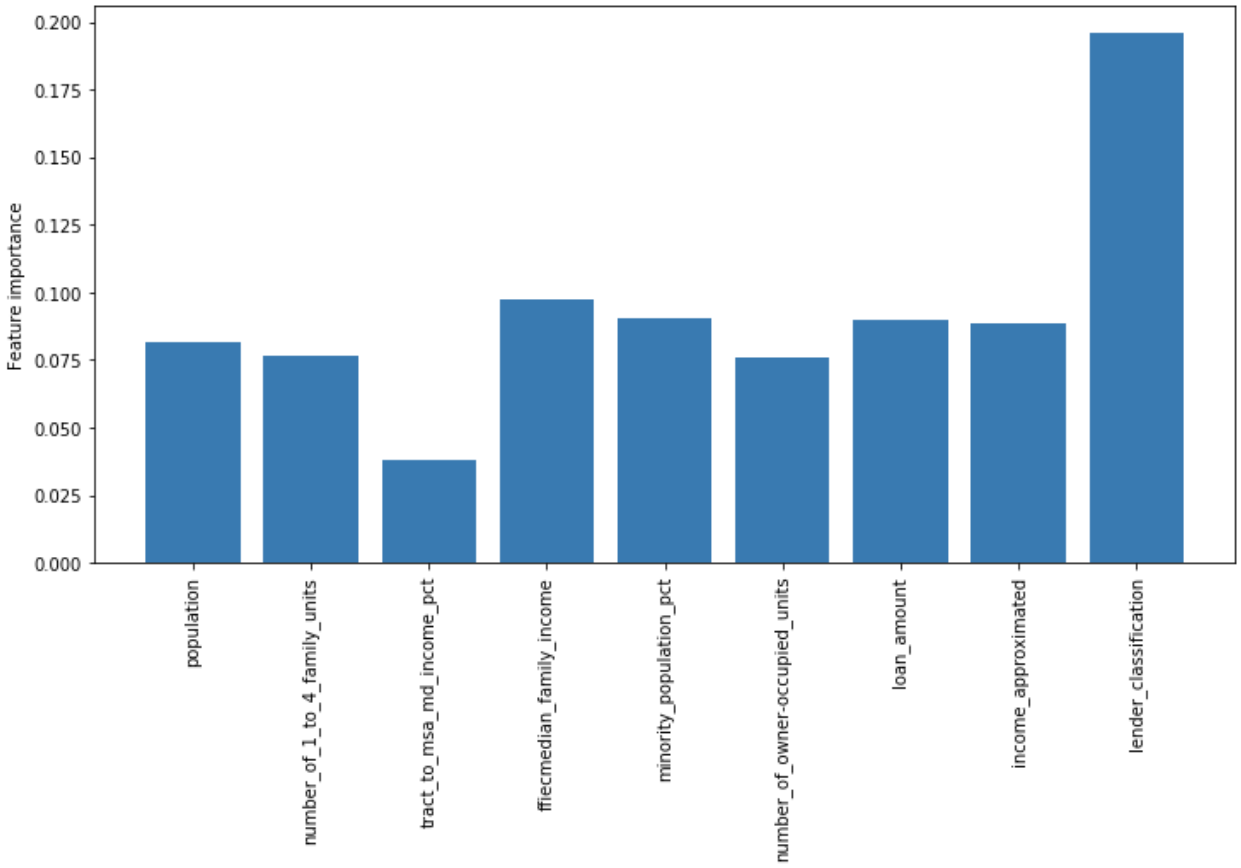
A predictive model was created based on Random Forest Classifier. Model hyperparameters were optimized using grid search with cross validation method using 10 folds. Due to the large number of datapoints in the training dataset, the cross validation was performed on the 20% random sample (100000 data points). The hyperparameter search was focusing on the maximum number of features and number of estimator parameters of the random forest classifier with accuracy being the optimizing parameter. The result of the hyperparameter selection is as follows using mean accuracy score across different folds:



After the hyperparameters were optimized, the model was trained again with the 70% of the training data and tested with the 30% of the remaining data. The following ROC curve was obtained.



In addition, feature importances were evaluated from with the following results:



As can be seen, the highest feature importance is attributed to the engineered feature 'lender classification'. Please note, that the importance of the encoded categorical features are not plotted as they are much lower than the numerical features. This is however expected as the categorical features are encoded in a larger number of columns in the dataset.

Overall, the model translates into the following standard performance classification metrics:

- Accuracy: 0.72238
- Precision: 0.70719
- Recall: 0. 76195
- F1 Score: 0.7335

And the confusion matrix: