Analysis of Mortgage Approvals from Government Data

Microsoft Professional Capstone Project: Data Science

June, 2019

1.0 Selective Summary

This document presents an analysis of data concerning Mortgages applications and the approval status which could be accepted (meaning the loan was originated) or denied according to the dataset adapted from Federal Financial Examination Council's (FFIEC). The analysis is based on 500,000 observations of mortgages approval dataset, each containing specific features of Mortgage application and its approval status.

The number of features were 21 excluding the target variable (accepted) and other features were generated from the existing ones by the author in the process of analysis and optimization of the model performance.

From the description of the problem, these features were sub divided into 5 groups for clarity which are

- Property location
- Loan information
- Applicant information
- Census information
- Index and target variable

After extensively exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several essential relationships between mortgages applications and approval status were identified. A predictive model to classify approval status into two categories was created.

Below are the conclusions the author made from the analysis:

While many factors can help indicate the approval status of a mortgage application, significant features found in this analysis were:

- **Lender**: A categorical feature with no ordering indicating which of the lenders was the authority in approving or denying this loan. The number of mortgage applicants for some lenders are extremely high compare to others who barely have applicants thereby increasing the chances of getting mortgages approved.

- **Applicant income**: the rate of acceptance for applicants with their income greater than applied mortgage is more than those with lesser income to applied mortgage.

- **Loan amount**: the rate of acceptance for applicants with loan amount lesser than the mean loan amount in their ethnicity grouped by population tracts is higher than those with loan amount exceeding the mean loan in their ethnicity grouped by population tracts.

2.0 Initial Data Exploration

The initial exploration of the data began with some summary and descriptive statistics. Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for the continuous numerical features and the results taken from 500,000 observations are shown in the figure below:

	count	mean	std	min	25%	50%	75%	max
loan_amount	500000	221.753	590.642	1	93	162	266	100878
applicant_income	460052	102.39	153.534	1	47	74	117	10139
population	477535	5416.83	2728.14	14	3744	4975	6467	37097
minority_population_pct	477534	31.6173	26.3339	0.534	10.7	22.901	46.02	100
ffiecmedian_family_income	477560	69235.6	14810.1	17858	59731	67526	75351	125248
tract_to_msa_md_income_pct	477486	91.8326	14.2109	3.981	88.0673	100	100	100
number_of_owner-occupied_units	477435	1427.72	737.56	4	944	1327	1780	8771
number_of_1_to_4_family_units	477470	1886.15	914.124	1	1301	1753	2309	13623
accepted	500000	0.500228	0.5	0	0	1	1	1

Fig 1. Statistical summary

Apart from the numerical features described above, below are the categorical features which to some degree also helped in the classification problem.

- **Row id**: A unique identifier with no intrinsic meaning. This was used for segregation analysis for investigations.
- **Loan type**: Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured; available values are:
  1 -- Conventional (any loan other than FHA, VA, FSA, or RHS loans)
  2 -- FHA-insured (Federal Housing Administration)
  3 -- VA-guaranteed (Veterans Administration)
  4 -- FSA/RHS (Farm Service Agency or Rural Housing Service)
- **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; available values are:
  1 -- One to four-family (other than manufactured housing)
  2 -- Manufactured housing
  3 -- Multifamily
- **loan purpose**: Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing; available values are:
  1 -- Home purchase
  2 -- Home improvement
  3 -- Refinancing

- **occupancy**: Indicates whether the property to which the loan application relates will be the owner's principal dwelling; available values are:
  1 -- Owner-occupied as a principal dwelling
  2 -- Not owner-occupied
  3 -- Not applicable
- **preapproval**: Indicate whether the application or loan involved a request for a pre-approval of a home purchase loan; available values are:
  1 -- Preapproval was requested
  2 -- Preapproval was not requested
  3 -- Not applicable
- **msa md**: A categorical with no ordering indicating Metropolitan Statistical Area/Metropolitan Division where a value of -1 indicates a missing value
- **state code**: A categorical with no ordering indicating the U.S. state where a value of -1 indicates a missing value
- **county code**: A categorical with no ordering indicating the county where a value of -1 indicates a missing value
- **applicant ethnicity:** A categorical feature indicating the ethnicity of each applicant.
- **applicant race**: Race of the applicant; available values are:

  1 -- American Indian or Alaska Native

  2 -- Asian
  3 -- Black or African American
  4 -- Native Hawaiian or Another Pacific Islander
  5 -- White
  6 -- Information not provided by applicant in mail, Internet, or telephone application
  7 -- Not applicable
  8 -- No co-applicant
- **applicant sex**: Sex of the applicant; available values are:
  1 – Male
  2 -- Female
  3 -- Information not provided by applicant in mail, Internet, or telephone application
  4 or 5 -- Not applicable
- **lender**: A categorical with no ordering indicating which of the lenders was the authority in approving or denying this loan
- **co applicant**: Indicates whether there is a co-applicant (often a spouse) or not
- **accepted**: Indicates whether the mortgage application was accepted (successfully originated) with a value of 1 or denied with a value of 0

2.1 imbalance data checking

To check if the data is imbalance or otherwise, it will be great to start with the target variable which is a categorical variable. Below is a bar and a violin plot of the frequency vs the target variable.

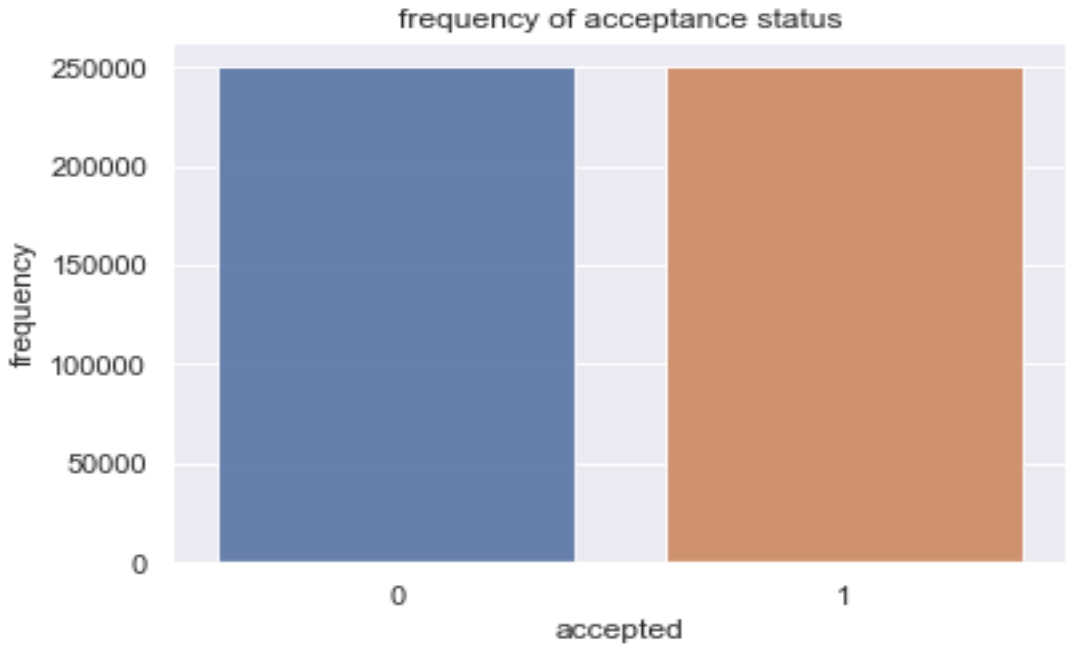


Fig 2.1.1 Acceptance status counts

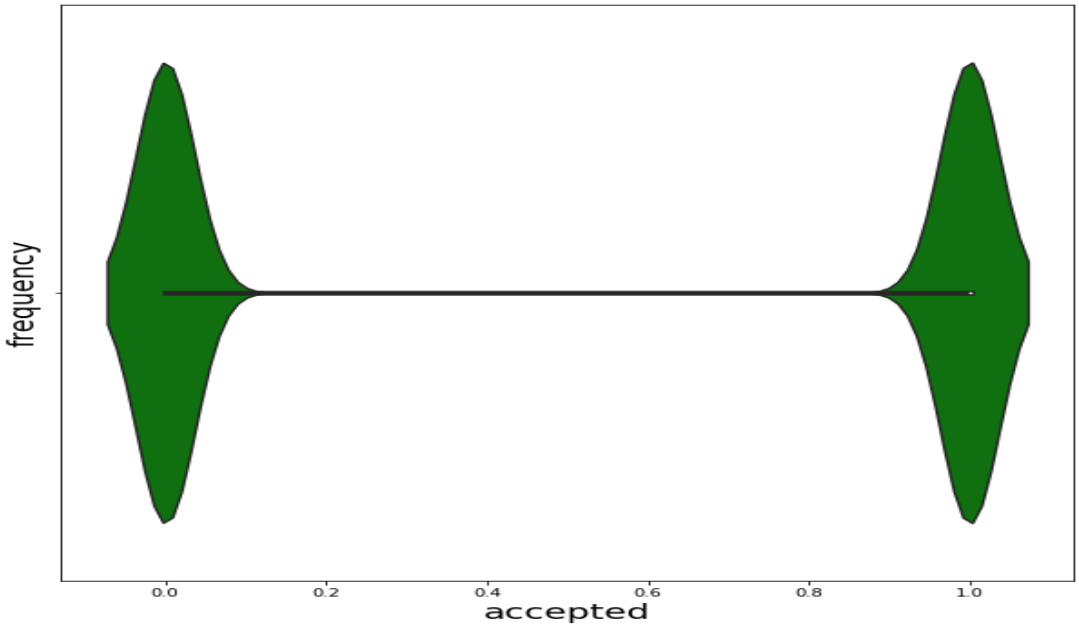


Fig 2.1.2 Acceptance status counts

From the plots above, it is clearly seen that the data is balanced and therefore, the model bias level is approximately zero from this perspective.

2.2 Complete Dataset validation

In other to have a meaningful exploratory data analysis, one has to check if there are missing values in the features. Below is a simple plot showing all features with the horizontal white stripes indicating the presence of missing values.

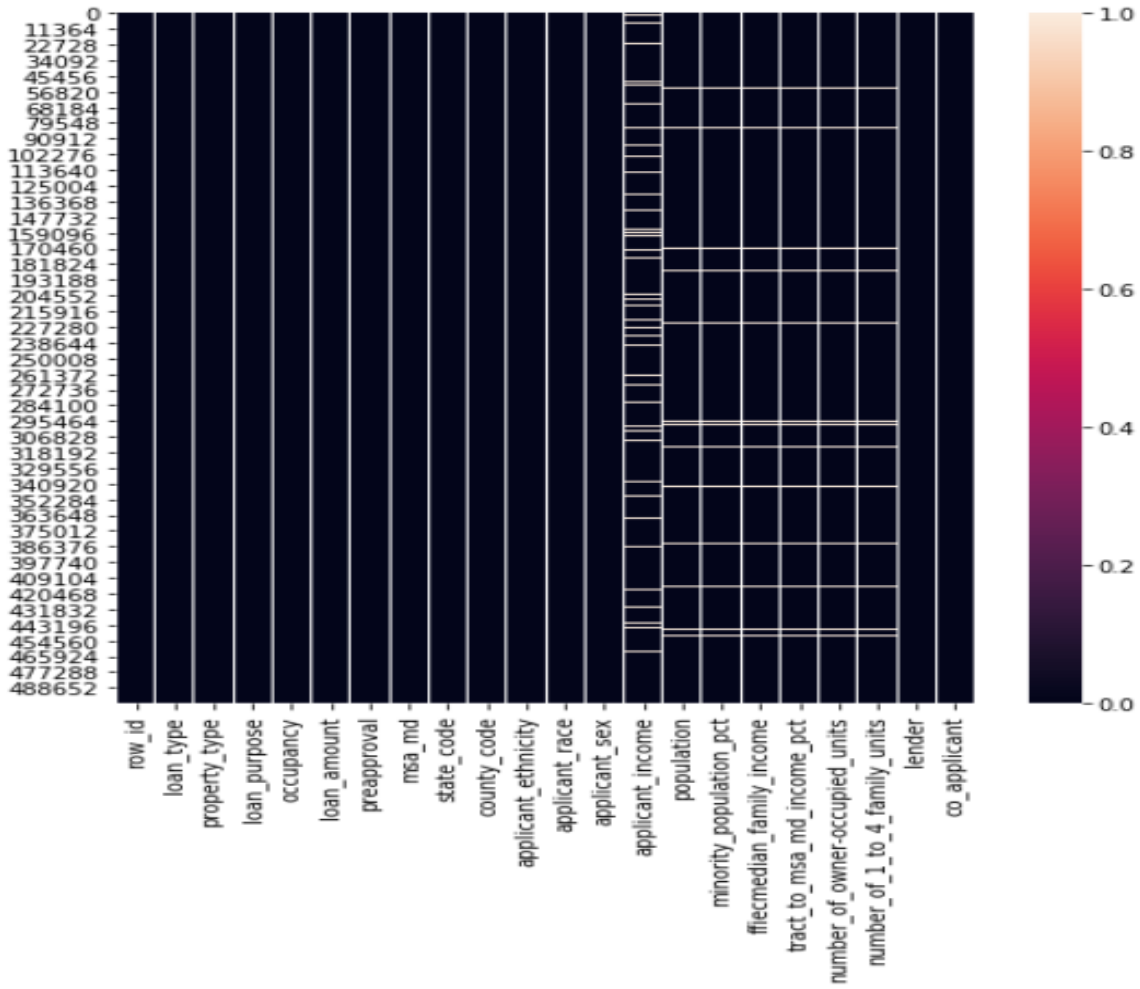


Fig 2.2 Missing values plot

From the figure above, percentages of missing values were further deduced in the following features:

- applicant income = 7.98%
- population = 4.49%
- minority population pct = 4.49%
- ffiecmedian family income = 4.48%
- tract to msa md income pct = 4.50%
- number of owner-occupied units= 4.51%
- number_of_1_to_4_family_units= 4.50%

These missing values were filled with either the feature's mean or median depending on the effect on the model performance and the contextual meaning of the feature.

2.3.1 Ordinal Categorical variable

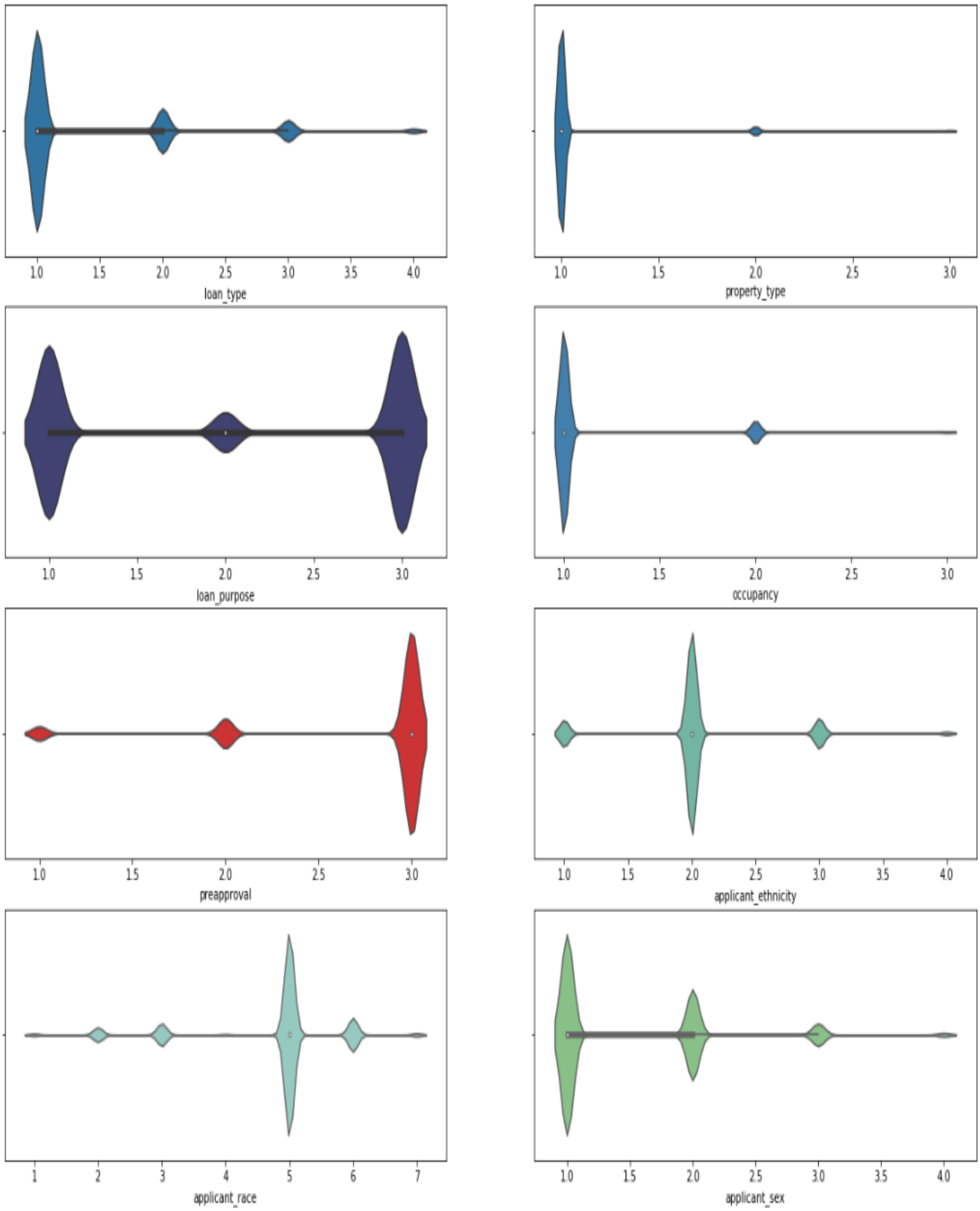


Fig 2.3 Ordinary categorical features counts plots
The violin plots above show the frequency of each category in each feature.

Key insights:

- “Conventional loans” are the most common loan type, followed by a marginal difference in “Federal Housing Administration” and “Veterans Administration”. The least common loan is the “Farm service Agency or Rural Housing Service”.
- “One to four-family” is the most common of property types while “manufactured housing” and “multifamily” types barely occur.
- Most of the loans are either used for “Home purchase” or refinancing but sometimes used on “Home improvement”.
- “Owner-occupied” as a principal dwelling is the most common among applicants from the perspective of occupancy while only few are “Not owner-occupied” or “Not applicable”.
- Most of the applications or loans were found not applicable in terms of preapproval request while there is a marginal difference between applications or loans that were not based on preapproval request.
- It can be deduced that majority of the applicants are Latino but not Hispanic in terms of their ethnicity
- The most common of the applicant’s races is the white, while few spreads across other races
- Majority of the applicants turn out to be male while few percentages were reported female. Others were inconclusive

Also, from the ordinal categorical plots, some features would have been reduced to smaller number of categories based on their frequency but since this involve individual mortgage application, it will be unfair to limit or enhance the chances of an applicant only on group behavioral characteristics.

2.3.2 Nominal Categorical variable

The given nominal categorical variables that shew some level of importance were:

- **Lender**: the lender feature is the major predictor of this project which caught my attention to further generate features from it.

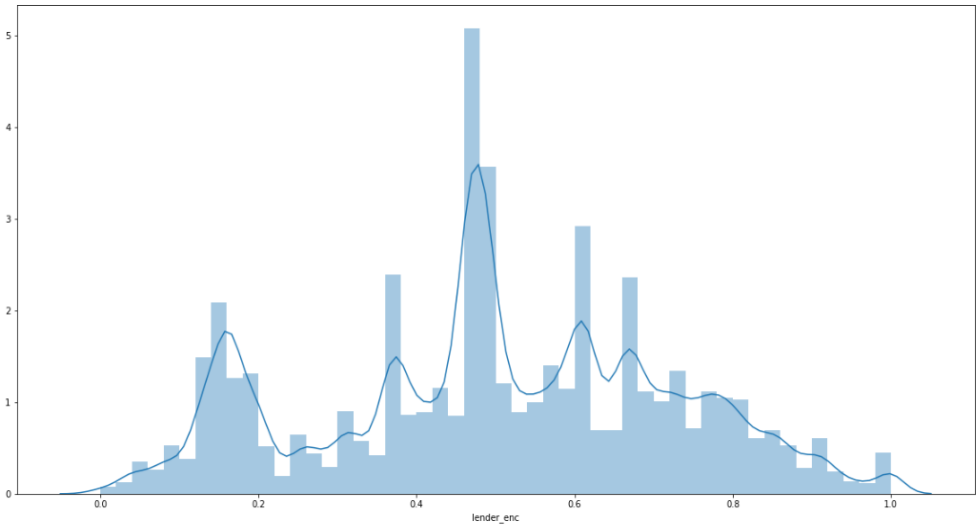


Fig 2.3.1 Distribution of lender encoded with respect to the target

- **State code**: this feature was not really contributing to the predictive power of the model because it has a negligible correlation with the target variable but can be put to use in the nearest future

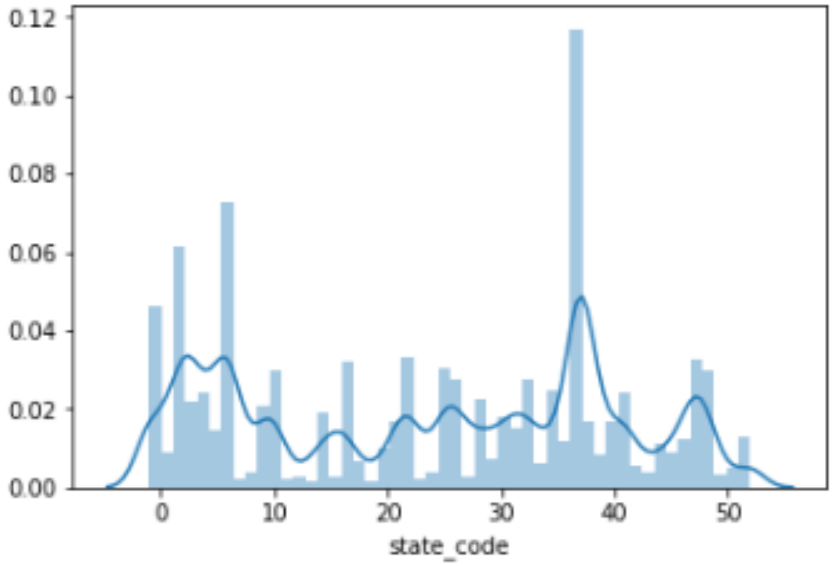


Fig 2.3.2 Distribution of state codes with respect to the target

- **County code**: this also have negligible correlation with the target variable and was not included in the model

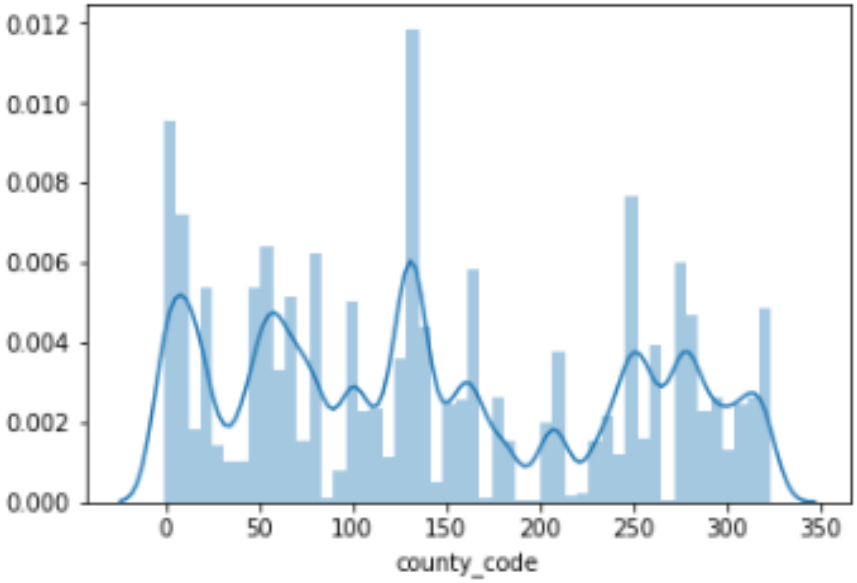


Fig 2.3.3 Distribution of county codes with respect to the target

2.4 Correlation

Building on established inferences, below are correlation pair plots of key numerical features split into 2 different sets due to the large size of the data.

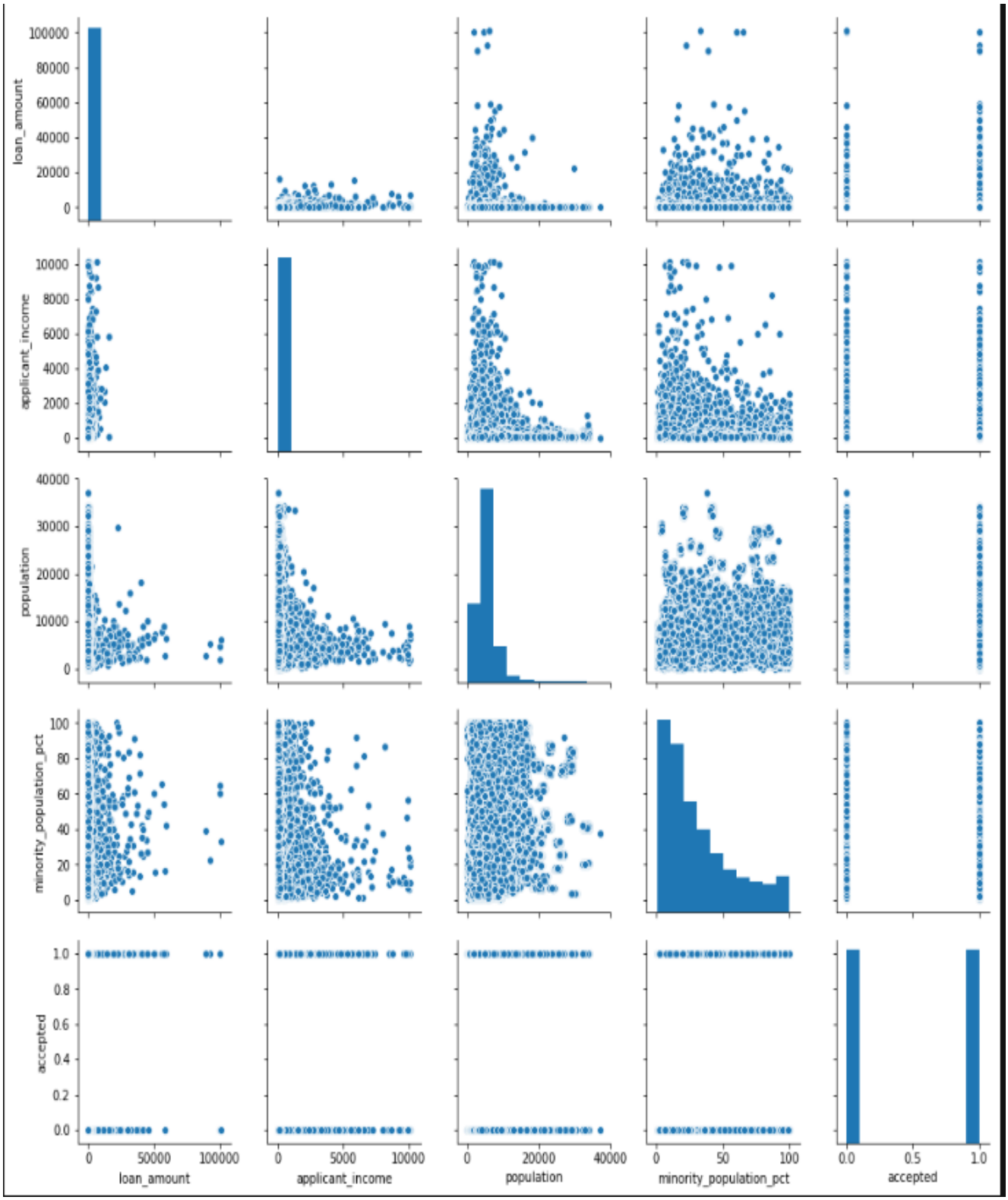


Fig 2.4.1 first set of correlation pair plots for numerical features

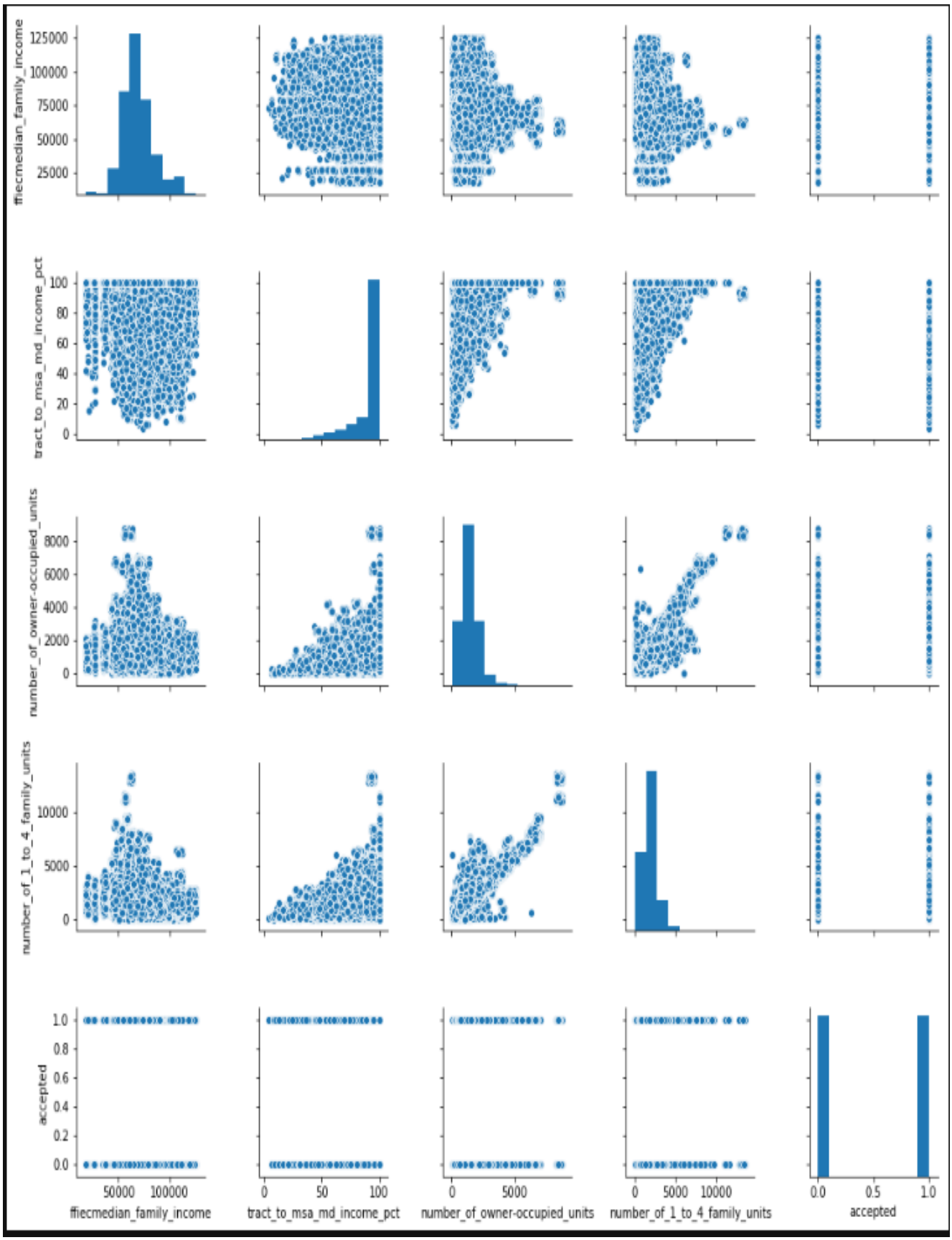


Fig 2.4.2 second set of correlation pair plots for numerical features

Since this is a classification problem, correlation of the features from regression perspective might not really be considered as a major factor but still needs attention for good predictive power.

Focusing on the last row in each set of the plots which compare each of the numerical feature with the target variable (accepted), the features are marginally distributed between the “2” classes in consideration (1 or 0). Further analysis is needed in order to establish the importance of each numerical feature relative to the target variable.

	loan_amount	applicant_income	population	minority_population_pct	ffiecmedian_family_income	tract_to_msa_md_income_pct	number_of_owner-occupied_units	number_of_1_to_4_family_units	accepted
loan_amount	1	0.483951	0.000100062	0.00722687	0.105924	0.0438108	-0.0136604	-0.0366437	0.0463698
applicant_income	0.483951	1	-0.00694757	-0.0537948	0.114988	0.102667	0.00454127	-0.0197484	0.0747216
population	0.000100062	-0.00694757	1	0.0873831	-0.0143772	0.149677	0.858732	0.816952	0.0191627
minority_population_pct	0.00722687	-0.0537948	0.0873831	1	0.0210592	-0.4428	-0.21441	-0.157976	-0.0929223
ffiecmedian_family_income	0.105924	0.114988	-0.0143772	0.0210592	1	-0.0545001	-0.0213896	-0.148235	0.0669194
tract_to_msa_md_income_pct	0.0438108	0.102667	0.149677	-0.4428	-0.0545001	1	0.360774	0.210613	0.0917657
number_of_owner-occupied_units	-0.0136604	0.00454127	0.858732	-0.21441	-0.0213896	0.360774	1	0.887591	0.0360289
number_of_1_to_4_family_units	-0.0366437	-0.0197484	0.816952	-0.157976	-0.148235	0.210613	0.887591	1	0.00602746
accepted	0.0463698	0.0747216	0.0191627	-0.0929223	0.0669194	0.0917657	0.0360289	0.00602746	1

Fig 2.4.3 Correlation table for numerical features

From the correlation table above, one could see that majority of the numerical features if not all have a very small correlation with the target variable.

Key insights:

- The distributions of predictors vary in shape
- Some predictors displayed different distributions when the response is positive than when the response is negative.
- Many variables appeared to be weakly correlated which in turn resulted in creation of new features from the existing ones.

Further exploration in correlations:

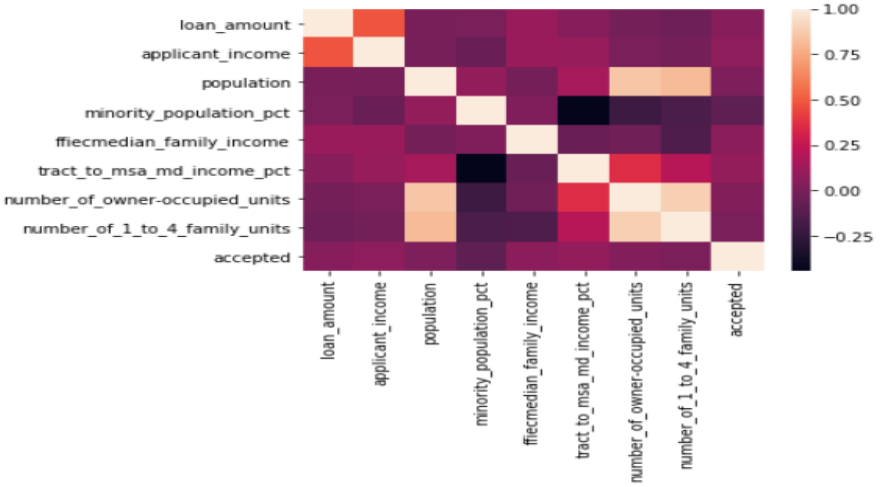


Fig 2.4.3 Correlation plot for numerical features

3.0 Data manipulation and generated features

- **Ethinicity_acceptance rate**: this helped in describing the rate at which mortgages are accepted within each ethnicity.

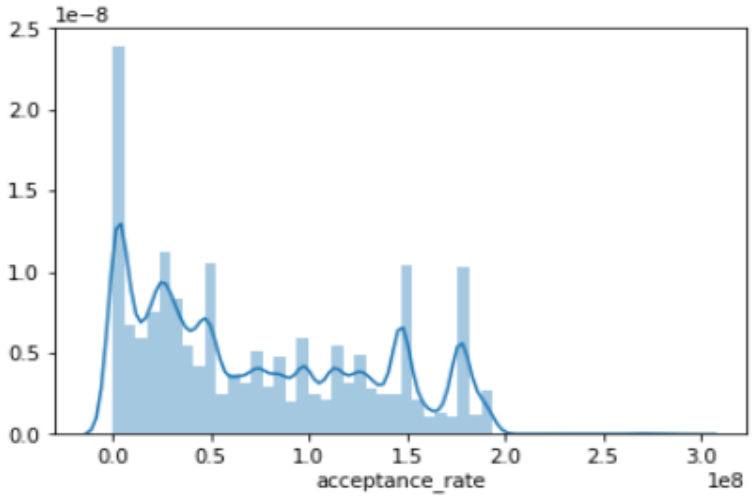


Fig 3.1.1 Ethnicity acceptance rate distribution plot

- **Ethnicity_loan_deviation**: this shows how individual loan amount deviate from the mean loan amount per ethnicity.

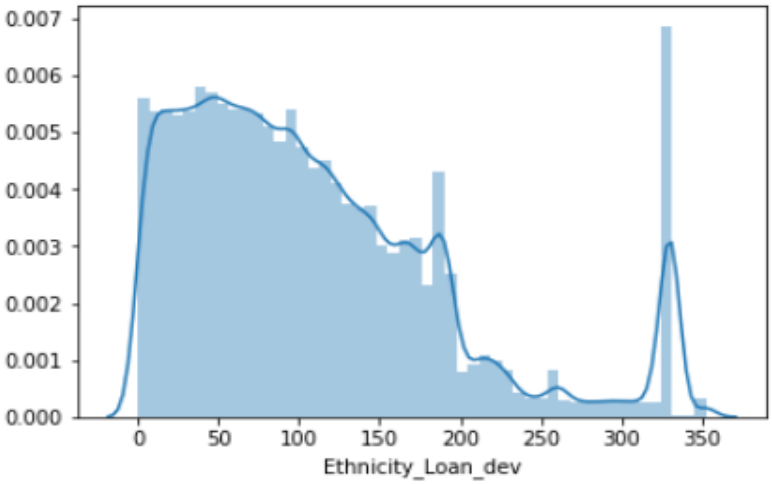


Fig 3.1.2 Ethnicity loan deviation distribution plot

Other features generated are:

- leverage_ratio
- credit worthiness score.
- Amortization
- County_state_ratio

3.2 Categorical and Numerical features Relationships

Having explored the relationships between accepted status and numeric features, an attempt was made to discern any apparent relationship between categorical feature values and accepted status. The following are the analysis and insights deduced from the relationships.

- **Loan acceptance rate across each applicant ethnicity**

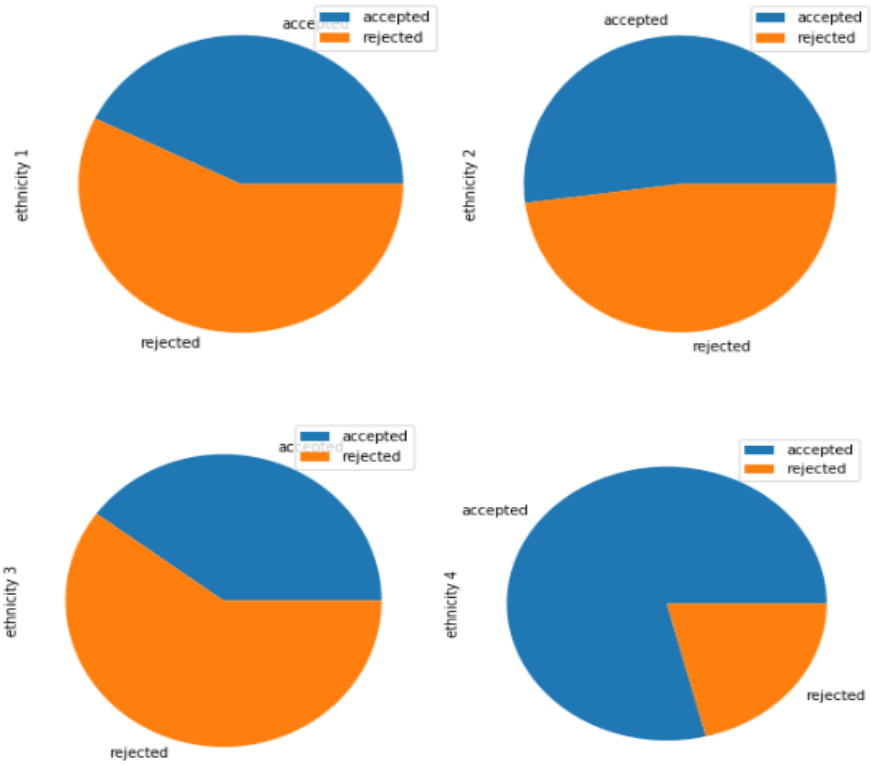
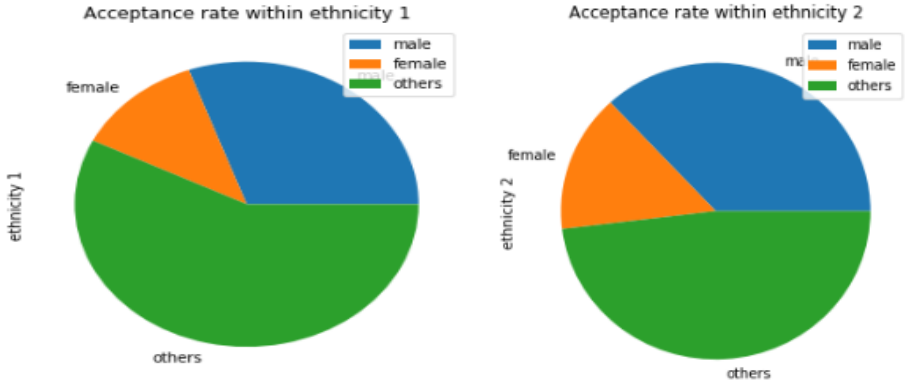


Fig 3.2.1 Percentage acceptance rate across each ethnicity

Key insights:

- From the pie charts above, ethnicity 1 to 3 seem to have marginal rate at which loans are being granted. In contrast to this is the significance increase at the rate at which loans are granted within ethnicity 4.

- **Loan acceptance rate across Ethnicities and Gender difference**



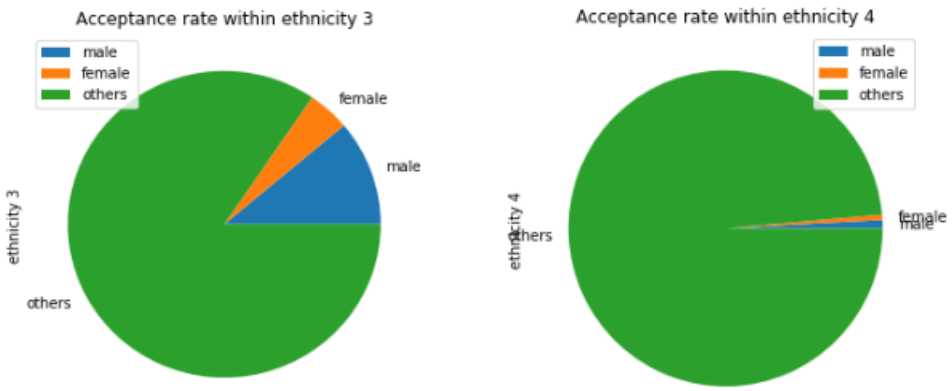


Fig 3.2.2 Percentage acceptance rate across each ethnicity with gender difference clause

Key insights:

- Applicants where applicant_ethnicity= 1,2 or 3 seems to have other gender dominating the rate at which loans are granted. This suggest that, collection of applicant's details should enforce a specific gender for proper screening which is in favor of the government.
- Applicants where applicant_ethnicity = 1 and 2 have loan grant rates marginally distributed.

- **Loan amount and applicant's income analysis across loan purpose, loan type and co applicant**

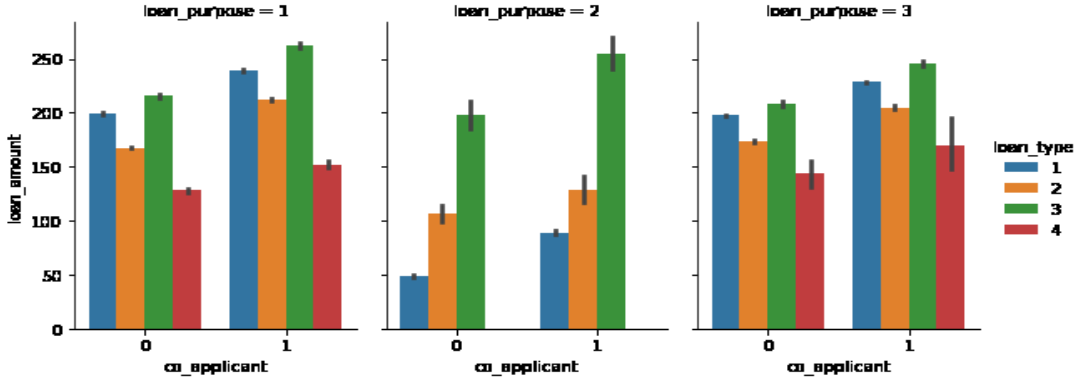


Fig 3.3.3a.

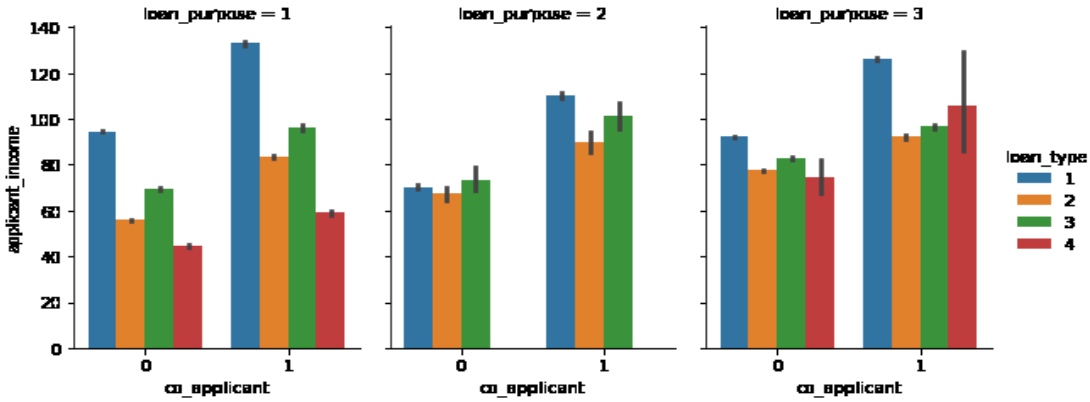


Fig 3.3.3b.

From the compound bar plots above (Fig 2.5.3a and Fig 2.5.3b), below are the insights deduced:

Key insights:

- Applicants who applied for any of the loan type without co-applicants (often spouse) and aiming for “Home purchase” will be needing approximately twice their income to be able to pay back. This makes their application to have about 40% chances of being accepted since a single person bears the whole risk. Validated by the plot below.

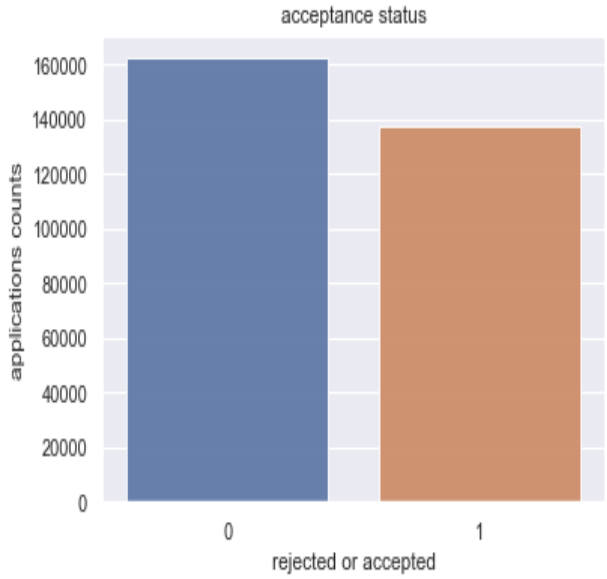


Fig 3.3.4a

- Applicants who applied for any of the loan type alongside a co-applicant (often spouse) and aiming for “Home purchase” will also be needing approximately twice their income to be able to pay back. This makes their application to have about 60% chances of being accepted since the burden is shared among them. Validated by the plot below.

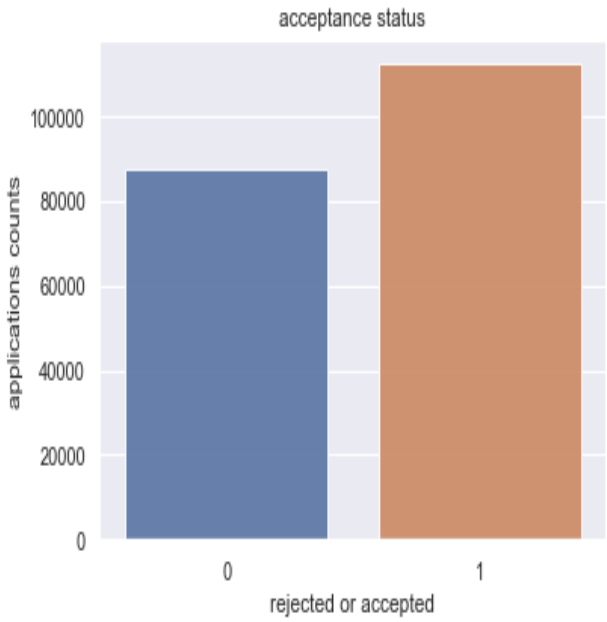


Fig 3.3.4b

- Generally, applicants who applied for loan with or without co-applicants (often spouse) and aiming for “Home improvement” applied for lower loan amount compared to their income because majority of them are veterans. Acceptance rate among this group of people is about 33% due to the level of risk on the loan. Validated by the plot below.

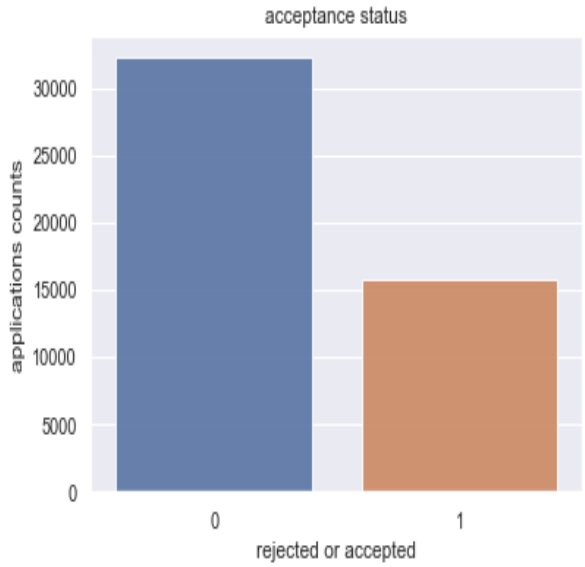


Fig 3.3.4c

- Also, no one is applying for the loan type FSA/RHS (Farm Service Agency or Rural Housing Service) with the aim to improve their home because such combination is not realistic.

4.0 Model prototyping and comparison

A set of models are compared out of the box to identify the most promising ones for further development. These models were trained with 66% of the data while the remaining 34% of the data was use for testing.

The chosen models are:

- AdaBoost
- XGBoost
- Microsoft ligthGBM

Performance is estimated through precision, recall and f-1 score, using each model respective specifications for optimization.

4.1 Classification of Mortgage Applications into Accepted or Rejected

The final predictive model was built based on the analysis of the mortgage approval dataset adapted from the Federal Financial Examination Council’s (FFIEC).
The model was created using the ***Microsoft ligthGBM*** algorithm and trained with 66% of the data. Testing the model with the remaining 34% of the data yielded the following results:

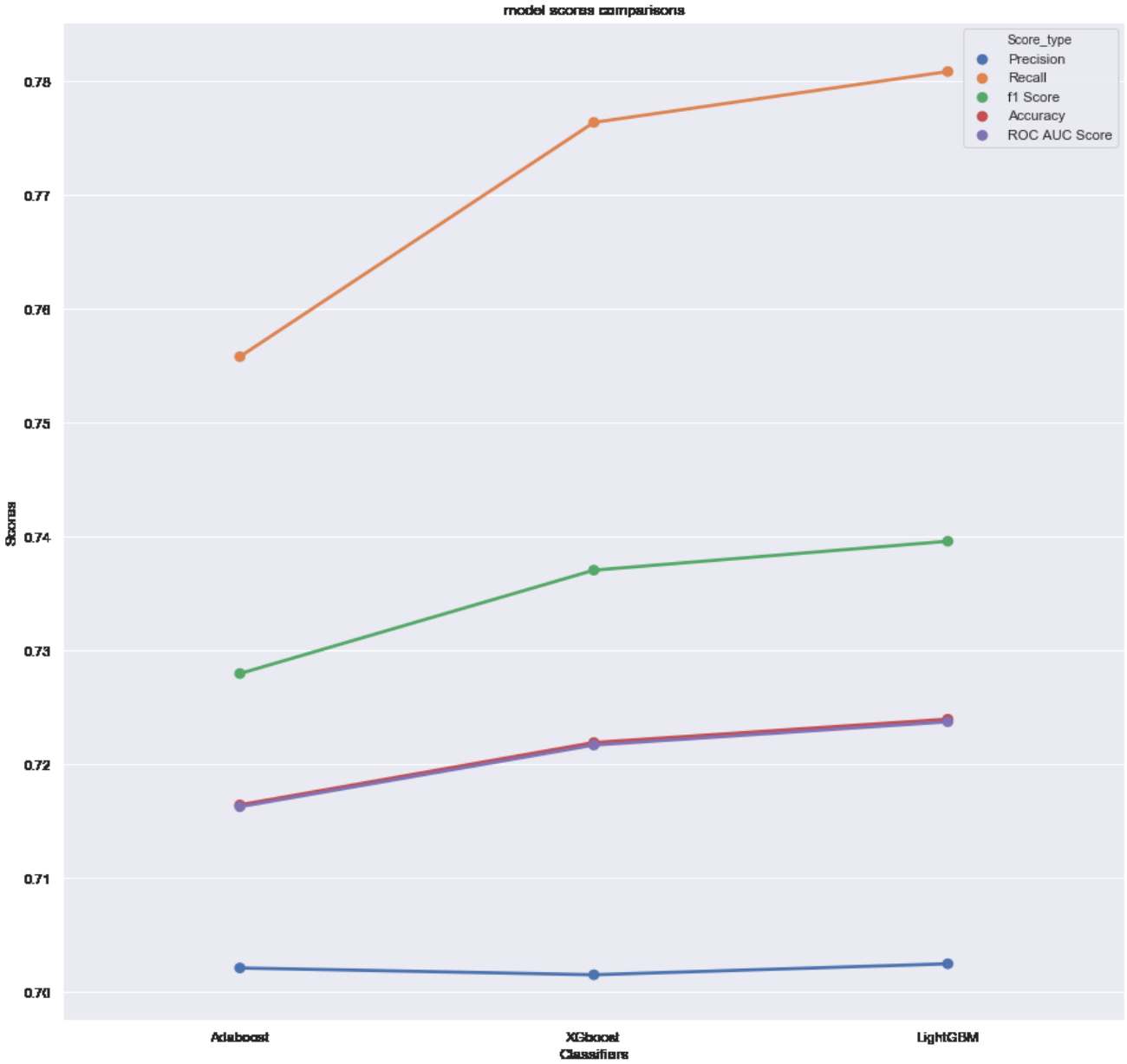


Fig 4.1 Predictive models metrics plot

Key insights:

- Adaboost algorithm appears to be the least performer in all metric measures except in precision where it took lead. The metrics are not outstanding and could be improved by further feature engineering and tuning. Though, ***Microsoft ligthGBM*** will be used as a model of choice for further validation, the choice based on compromise between speed and performance.

Key insights:

This translates into the following standard performance metrics for classification:

- **Accuracy: 72.5%**
  Meaning there is 72.5% confidence in this model leaving out 27.5% for further investigations
- **Precision: 92.3%**
  Meaning that 92.3% of the applications that are predicted as likely to be accepted ultimately will
- **Recall: 70.25%**
  Which implies that about 70.25% of the applications that will be granted will not be considered for any “retention action”
- **F1 Score: 74%**
  Meaning that the weighted average of the precision and recall is 74%

4.2 Validation of chosen model

The Microsoft ligthGBM classifier is further validated by means of:

- Confusion Matrix
- Cross-validated ROC curve

4.2.1 Confusion Matrix

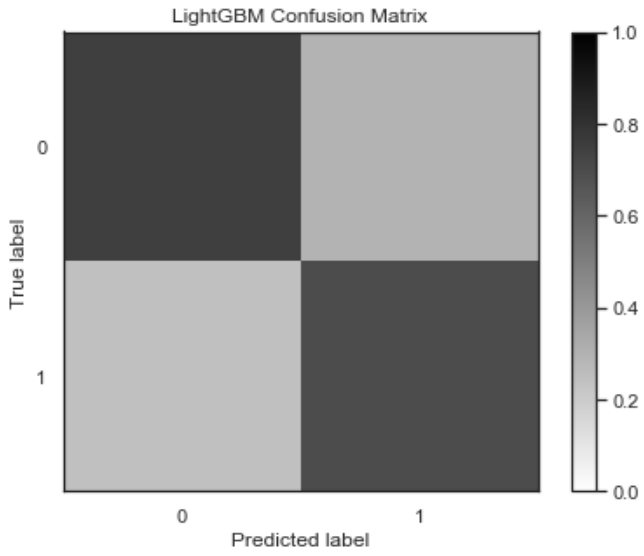


Fig 4.2.1 Confusion matrix plot

- True Positives: 64679
- True Negatives: 54778
- False Positives: 18152
- False Negatives: 27391

4.2.2 5-fold CV (cross validation) ROC (Receiver Operating Characteristics) curve

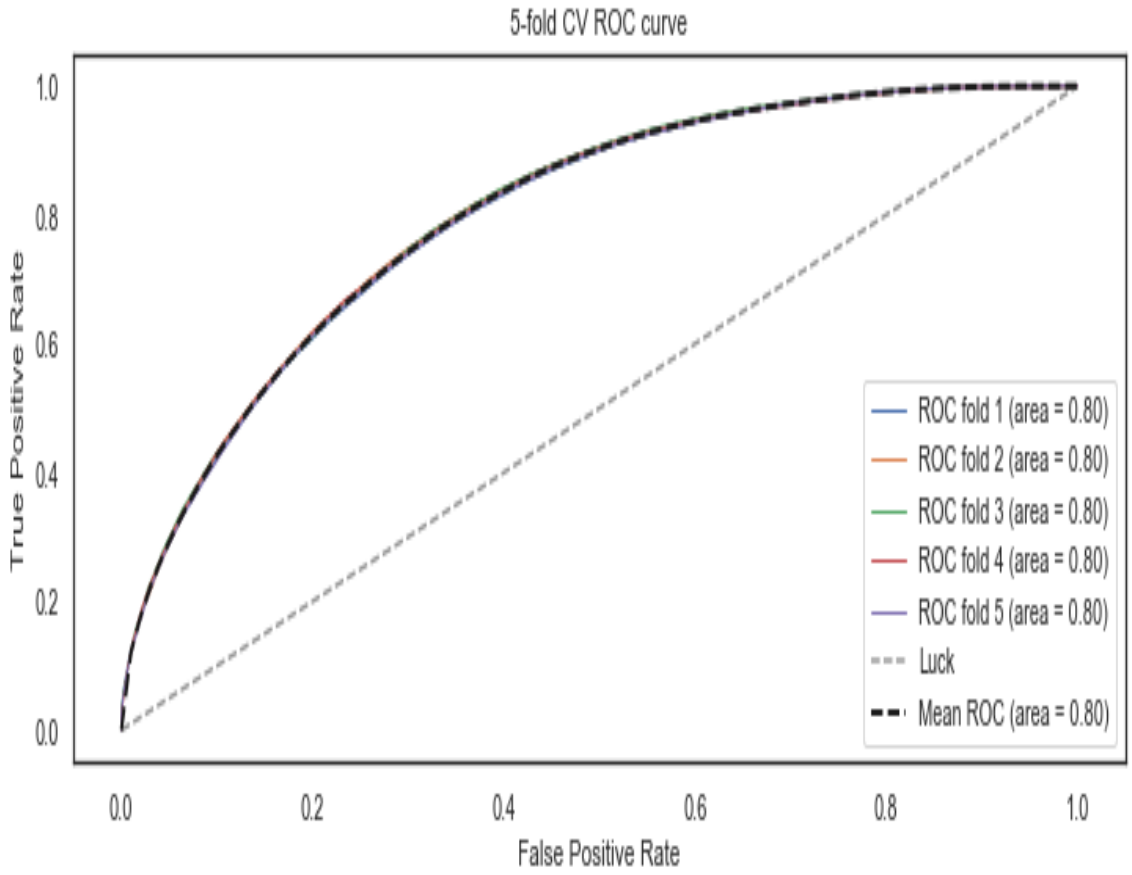


Fig 4.2.2 ROC plot

- From the ROC curve above, we could see that the model is well generalized in the sense that it is not overfitting.

5.0 Conclusion

- This analysis has shown that the status(accepted/rejected) of mortgage applications can be predicted with 78% confidence level from its characteristics.
- While other measures can be put into considerations, the lender, applicant income, loan amount, tract_to_msa_md_income_pct and population have a significant effect on the status(accepted/rejected) of mortgage applications.
- Secondary features, such as applicant sex, loan purpose, loan type, applicant ethnicity, co-applicant and preapproval can help further classify the mortgage applications as needed.