Analysis of Mortgage Application

Executive Summary

This document presents an analysis of data concerning mortgage application. The analysis is based on 500,000 observations of HMDA Loan Application Register each containing specific characteristics of necessary features required to get a loan originated (approved/accepted).

After the exploration of the data by calculating summary and descriptive statistics, and by creating visualizations of data, I identified several potential relationships between the characteristics of getting a loan originated (approved) and disapprove. After the data exploration, I created a predictive model to classify the mortgage application into two categories i.e. Origination of a loan (approval) and disapproval from the features in the data.

Some of the significant features I found in this analysis were:

Loan amount: Size of the requested loan in thousands of dollars

County code: A categorical with no ordering indicating the county

Applicant income: applicant income in thousands of dollars

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

lender: indicating which of the lenders was the authority in approving or denying this loan

Applicant Ethnicity: Ethnicity of the applicant, available values are 'Hispanic or Latino'. An applicant who is not Hispanic or Latino has a higher chance of getting a loan.

Applicant Race: the race of the applicant. An applicant who is Native of Hawaiian or other Pacific Islander tends to get a loan application originated.

Loan Type: Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured. Federal Housing Administration Insured (FHA-insured) or Veterans Administration type of loan will get its loan application approved.

Property Type: Indicates whether the loan or application was for a one-to-four-family dwelling, manufactured housing, or multifamily dwelling. A loan applicant with manufactured housing as property tends to get the loan originated.

Preapproval: Indicate whether the application or loan involved a request for a pre-approval of a home purchase loan. It turns out that if a loan application does not request pre-approval there is a higher tendency of originating the loan.

Applicant Sex: the gender of the applicant i.e. male or female

Loan Purpose: Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing. Loan Applicant with the purpose of purchasing a house or home improvement has a high chance of getting the approval of the loan application.

initial Data Exploration

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

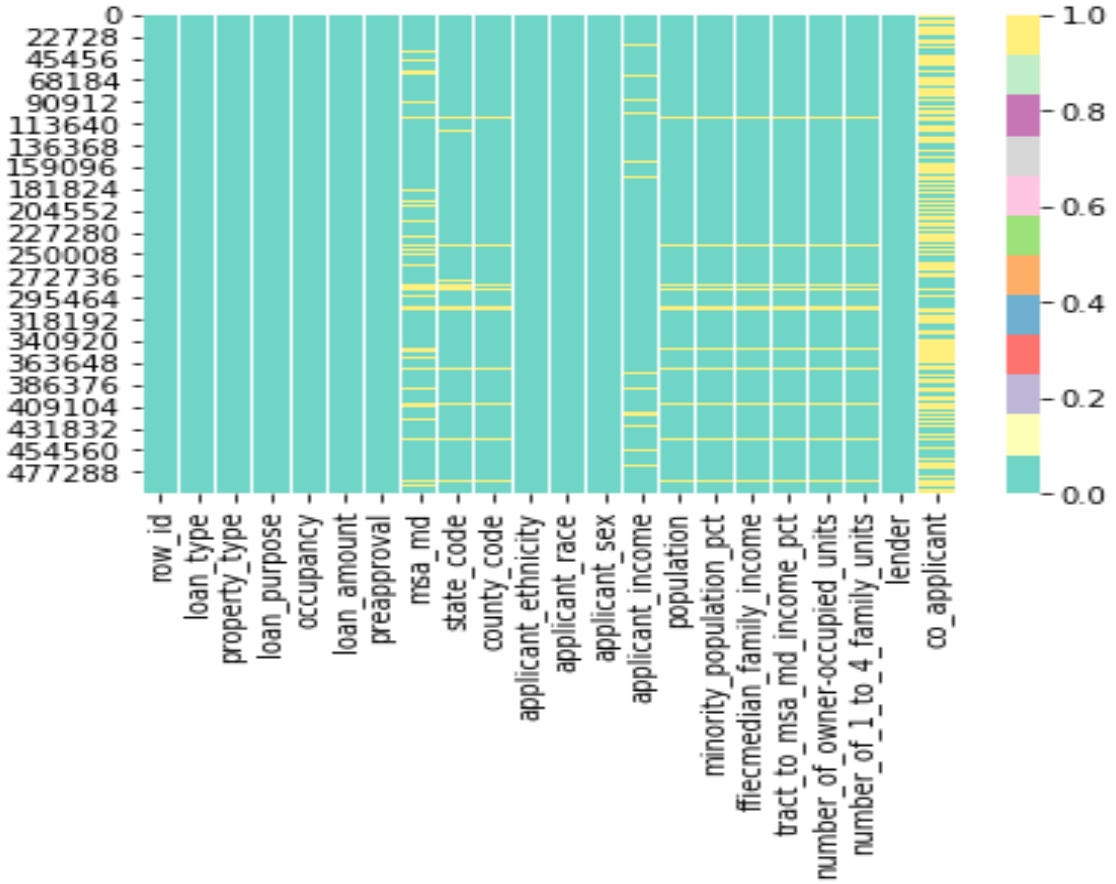
Individual Feature Statistics

I took the summary statistics for minimum, maximum, mean, median, standard deviation, and I as well calculated distinct count for numeric columns, and the results taken from 500,000 observations are shown here:

Column	Count	Mean	STD	Min	25%	50%	75%	Max
Loan_amount	500000	221.753158	590.641648187931	1	93	162	266	100878

Applicant_income	460052	102.389521184562	153.534495563816	1	47	74	177	10139
Population	477535	5416.83395562629	2728.14499871113	14	3744	4975	6467	37097
minority_population_pct	477534	31.617310254348	26.3339380710524	0.534	10.7	22.901	46.02	100
Ffiecmedian family income	477560	69235.6032980149	14810.0587907783	17858	59731	67526	75351	125248
tract_to_msa_md_income_pct	477486	91.8326239470876	14.2109242866298	3.981	88.06725	100	100	100
number_of_owner-occupied_units	477435	1427.71828206981	737.559511379416	4	944	1327	1780	8771
number_of_1_to_4_family_units	477470	1886.14706473705	914.123744384939	1	1301	1753	2309	13623
Accepted	500000	0.500228	0.500000448017627	0	0	1	1	1

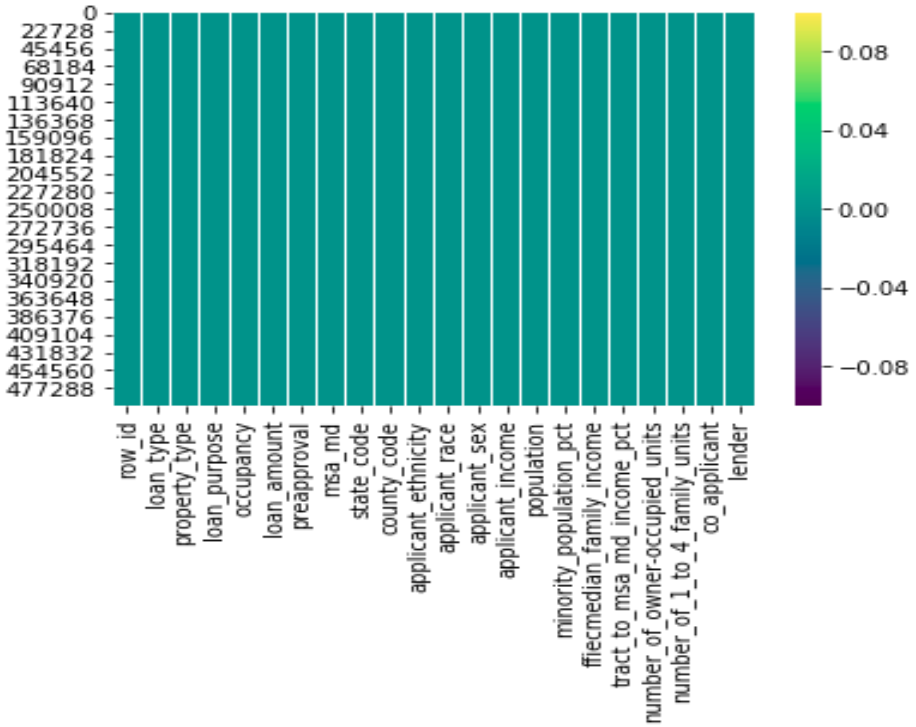
I did some exploration of the data and found out that the data contains some missing values.



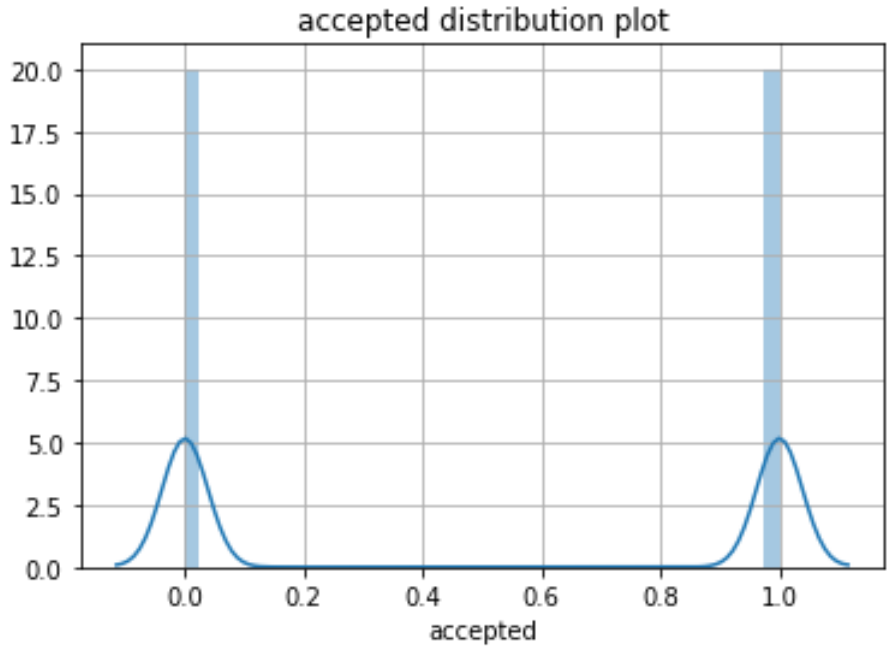
where columns with yellow strips imply the presence of missing values.

Treatment

I carefully treated the columns with missing values and I did this were carefully with domain knowledge in respect to the dataset, below is a visualization of the data after the treatment of missing values.



Now all columns have been treated of missing value. Since the target label **Accepted** is of interest in this analysis, I created histogram with a distribution plot of the column **accepted**, the plot shows that it follows a normal distribution.



The chart above shows that there is more accepted loan application than the loan application that was rejected. (accepted :250115, rejected: 249886)

In addition to the numeric values, the mortgage application observations include categorical features, including.

- Msa_md: indicating Metropolitan Statistical Area/Metropolitan Division.
- State code: indicates states in United States
- County code: indicates county
- Loan type: - Conventional, FHA-insured (Federal Housing Administration), VA-guaranteed (Veterans Administration), FSA/RHS (Farm Service Agency or Rural Housing Service)
- Property type: - One to four-family, Manufactured housing, Multifamily
- Loan purpose: - Home purchase, Home Improvement, Refinancing
- Occupancy: - Owner occupied as a principal dwelling place I.e. (the applicant owns a house)
- Pre-approval: - pre-approval was requested, pre-approval was not requested
- Applicant race: - American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or another pacific islander, White.
- Applicant sex: Male, Female.

I created count plots to show the frequency of each of these features and indicate the following:

loan Type: conventional type of loan is more common than Federal Housing Administration insured, Veterans Administration guaranteed and Farm Service Agency or Rural Housing Service.

Property Type: one-to-four-family dwelling applicants are more than Manufactured Housing. The applicants with Manufactured Housing type of property are more than the applicant with

Multifamily dwelling property type. This means that the applicants with a home that accommodates a family of four sizes are more than the applicant with Manufactured Housing and Multifamily dwelling.

Loan Purpose: Applicants with the purpose of getting a loan for Refinancing are much more than applicants with the purpose of purchasing a house, while applicants with the purpose of Purchasing house are more than applicants with the purpose of Home Improvement.

Occupancy: Applicants who own a principal dwelling place are more than those who don't own a principal dwelling place.

Pre-approval: Loan application that doesn't request for a pre-approval are more than those who requested for a pre-approval.

Applicant Race: White applicants are more than all other race.

Applicant Sex: there are more males than females.

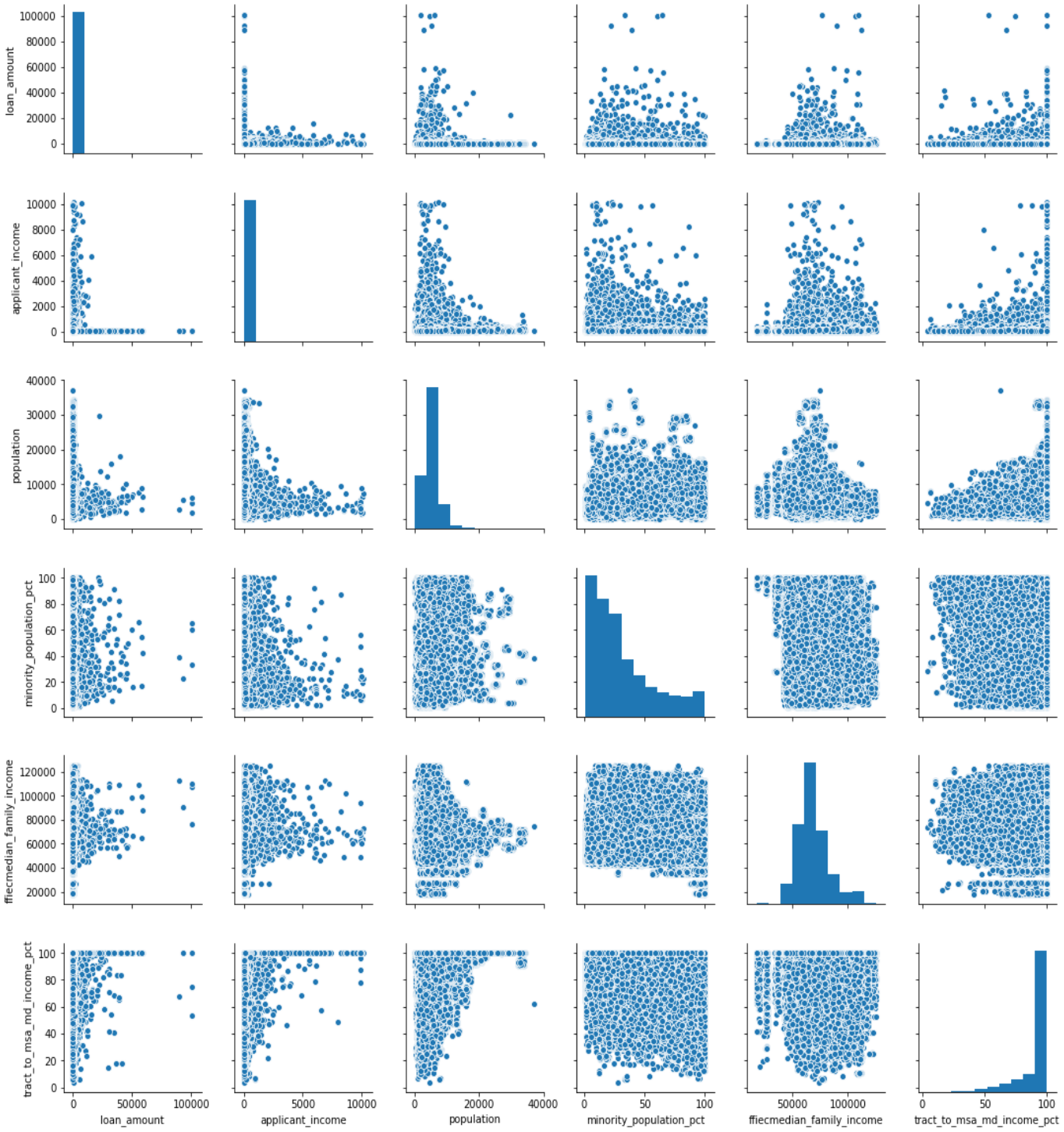
Applicant Ethnicity: Hispanic or Latino Applicants are less than those who are not Hispanic or Latino.

Correlation and Apparent Relationships

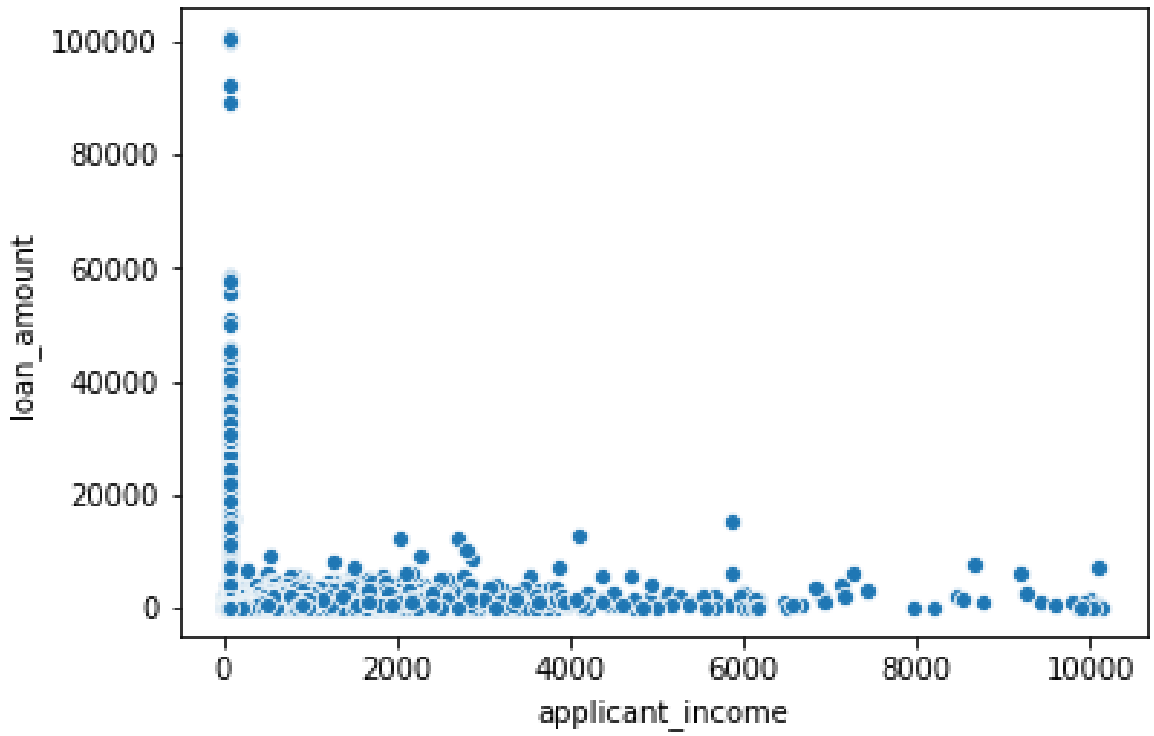
After my exploration of the individual features, I identified the relationships between features in the data especially, between the target label (accepted) and the other features.

Numerical Relationships

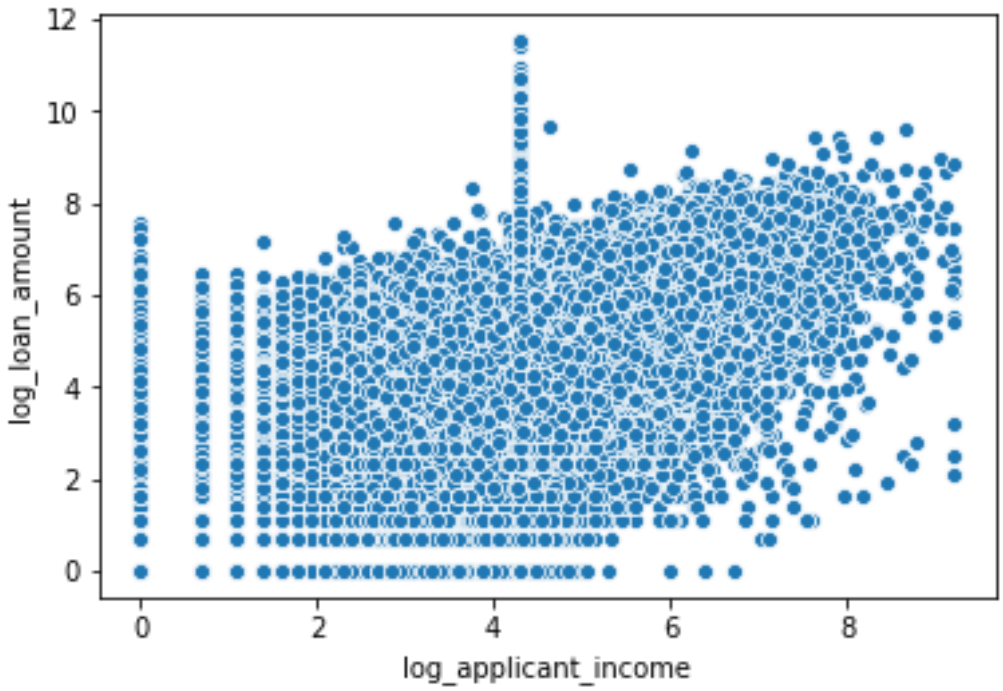
I generated the following scatter-plot initially to compare numeric features with one another. The key features in this matrix are shown here.



It can be seen from the above matrix plot that the relationship between Loan Amount and Applicant Income does not exhibit linearity, which is the chart below.



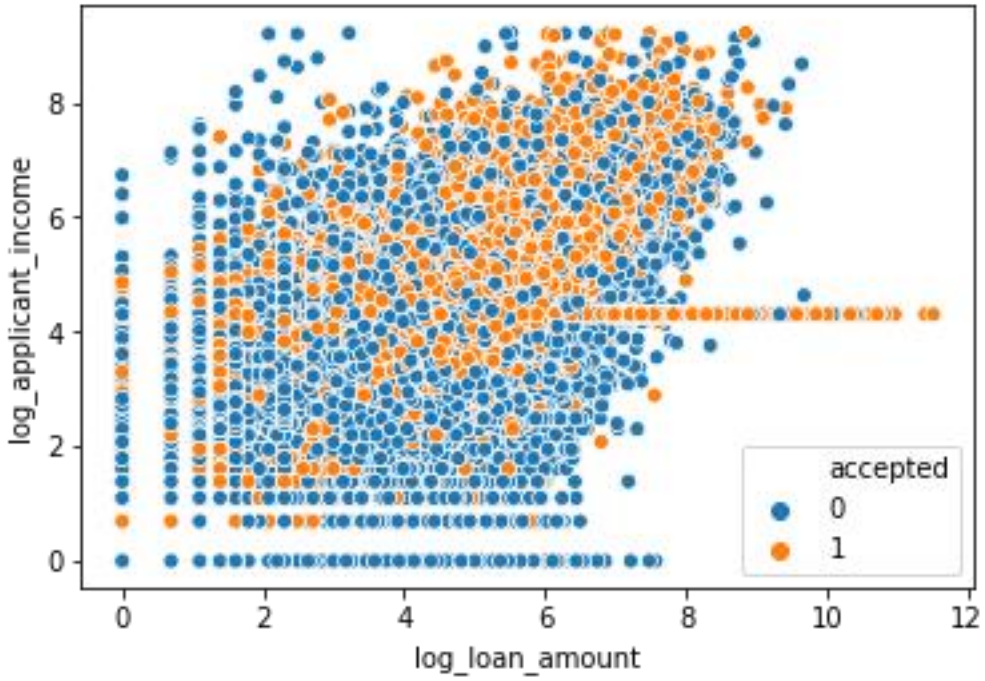
In an attempt to improve the fit of the features, I took the log-normal of both features. After taking the log-normal, it is known in the chart below that there exists a linear relationship between the Log-Loan-Amount and Log-Applicant-Income.



Multi-Faceted Relationships

Apparent relation between the target label **accepted** and individual features are helpful in determining predictive heuristic. Relationships are more complex and may only become apparent when multiple features are considered in combination with one another. To help identify these complex relationships, I created some faceted plots.

It can be seen from the plot below that **accepted** label been indicative of Log-Applicant-Income and Log-Amount both of which are typically predictive of approval (origination) of a loan application.



From the above plot, it can be seen majorly that large income is proportional to loan amount which leads to the origination (acceptation) of a loan application.

Classification of Mortgage Application

Based on the analysis of the mortgage Application data, I created a predictive model of binary classification to classify the application into two categories which are Accepted i.e. **1** and Rejection i.e. **0**.

I created this model using a simple Logistics Regression algorithm and I trained it on 70% of the data, testing the model with the remaining 30% of the data.

To improve the performance of the model, I adopted ensemble method, algorithms includes Random Forest, Light Gradient Boost Method (LightGBM), XGBOOST, CatBoost. I implemented these algorithms and yields the following results:

- True Positive: 46520
- True Negative: 285503
- False Positive: 16275
- False Negatives: 58702

Performance Metric

The metric use in scoring the performance of the predictive model used in this analysis is called Accuracy (Classification Rate).

The mathematical formula of a accuracy = $\frac{(TP + TN)}{TP+TN+FP+FN}$

where

$$\begin{cases} TP & \text{= True Positives} \\ TN & \text{= True Negatives} \\ FP & \text{= False Positives} \\ FN & \text{= False Negatives} \end{cases}$$

I obtained an Accuracy of 0.7730 from the given Test data

Conclusion

The analysis has shown that the granting of a loan application for an applicant can be confidently predicted from its characteristics. in particular, the loan amount, applicant income, loan type, applicant Race, applicant ethnicity, property type, ffiecmedian_family_income.

Recommendation

Although the use of machine learning to automate loans application will raise the issue of some privacy, ethical and legal concerns about being able to predict the behaviour of applicant. After the whole analysis, these factors (Type of Property, Type of Loan, Loan Amount, Purpose of Loan, Applicant and FFIEC Median Family Income) are sufficient and increases the decision of originating a loan.