Predicting Mortgage Rates from Government Data

Oguzhan San, November 2019

Executive Summary

The document presents a predictive analysis of mortgage rates according to government dataset from the Federal Financial Institutions Examination Council's (FFIEC). The paper conducts the analysis how demographics, location, property type, lender, and other factors are related to forecast the rate spread of mortgage applications across the United States.

Data contains 21 features and a target to indicate the rate spread value with 200.000 samples.

Analysis consists of 3 steps:

1. Exploratory data analysis of rate spread by summaries and descriptive statistics.
2. Visualization of several potential relationships of factors with rate spread.
3. Predictive analysis of data by using a machine learning classifier.

After conducting the analysis, 5 factors are found more significantly important to compared to others:

- **Loan Type:** Indicates whether the loan granted, applied for, or purchased was conventional government-guaranteed, or government-insured.
- **Property Type:** The loan or application was for a one-to-four-family dwelling (other than manufactured housing), manufactured housing, or multifamily dwelling.
- **Loan Amount:** Size of the requested loan in thousands of dollars
- **Loan Purpose: T**he 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.

Data Exploration

A descriptive statistic is used here to have better insight into the data. There are 21 features in the dataset, which are categorized as categorical features and numeric features, and listed below:

Categorical features:

- msa_md: Metropolitan Statistical Area/Metropolitan Division
- state_code: the U.S. states
- county_code: County of the U.S.
- lender: Indicates 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: The loan or application was for a one-to-four-family dwelling (other than manufactured housing), manufactured housing, or multifamily dwelling.
- loan_purpose: 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: The application or loan involved a request for a pre-approval of a home purchase loan
- applicant_ethnicity: Ethnicity of the applicant
- applicant_race: Race of the applicant
- applicant_sex: Sex of the applicant
- co_applicant: Indicates whether there is a co-applicant or not

Numeric Features:

- applicant_income: In thousands of dollars
- 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
- tract_to_msa_md_income_pct: % 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

Descriptive Statistic and Data Visualization

The count, mean, standard deviation, minimum, quartile (25%, 50%, 75%), and maximum are giving for each numeric column in Table 1.

	applicant income	loan amount	population	minority population	ffiecmedian family	tract_to_msa_md income_pct	number_of_owner	number_of_1_to 4 family_units
Count	189292	200000	198005	198005	198015	197977	197988	197984
Mean	73.617	142.57	5381.09	34.23	64595.36	89.283	1402.87	1927.33
Standard Deviation	105.696	142.56	2669.03	27.93	12725	15.059	706.88	886.58
Min	1	1	7	0.32	17860	6.193	3	6
25%	39	67	3717	Oct 92	56654	81.648	932	1344
50%	56	116	4959	25.996	63485	98.959	1304	1799
75%	83	179	6470	52	71238	100	1742	2353
Max	1042	11104	34126	100	125095	100	8747	13615

Table 1. Descriptive Statistics of Numeric Features

The loan amount varies from $1,000 to $11,104,000 while the average and the standard division loan amount is 143.000 which most of the values are below $500.000. The distribution of the loan amount is left-skewed as can be seen in the graph below.

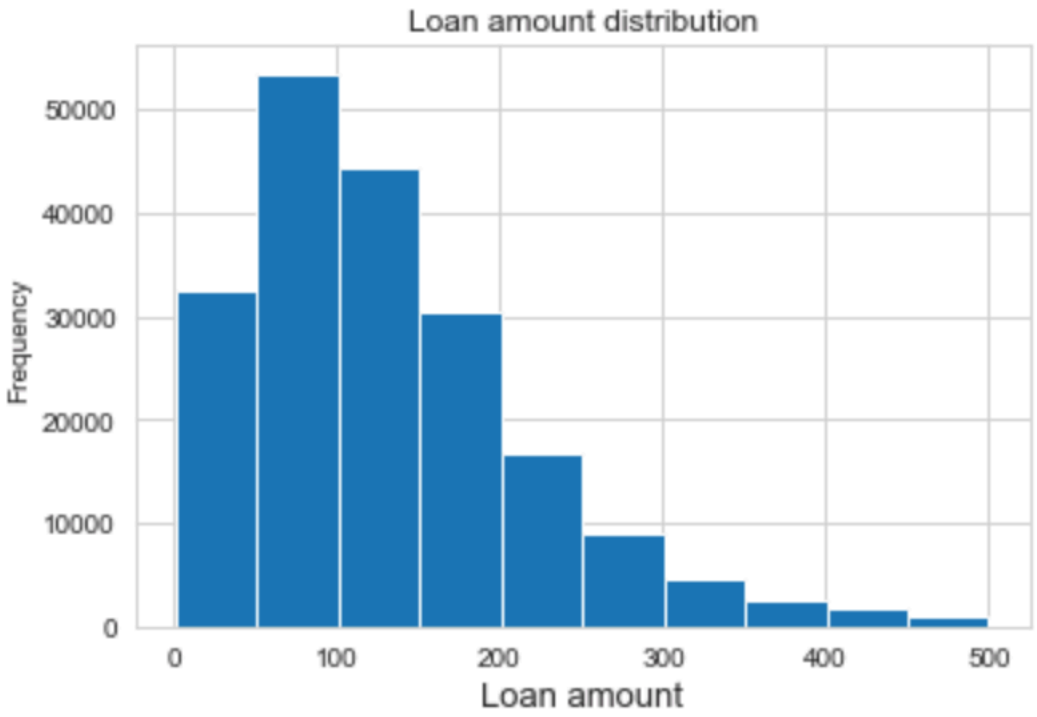

Figure 1. Loan distribution

In the graph below, the relation applicant ethnicity with rate spread is given. Numeric values on x axis stands for Hispanic, Not Hispanic, No information, No Co-applicant respectively.

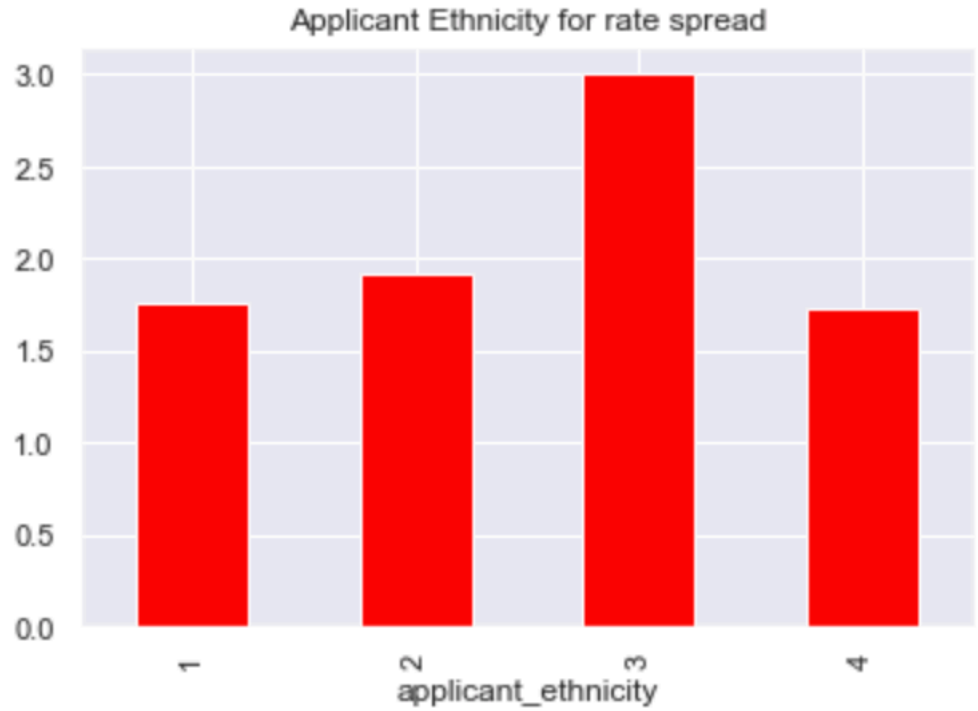

Figure 2. Applicant Ethnicity distribution for rate spread

In figure 3, correlations heatmap of Numeric features are given. Population and number of owners occupied, and number of family units are quite correlated. The majority of the features are categorical, and it would be also possible to compare categorical features to each numeric feature, however for the purpose of the report requires to keep it briefly.

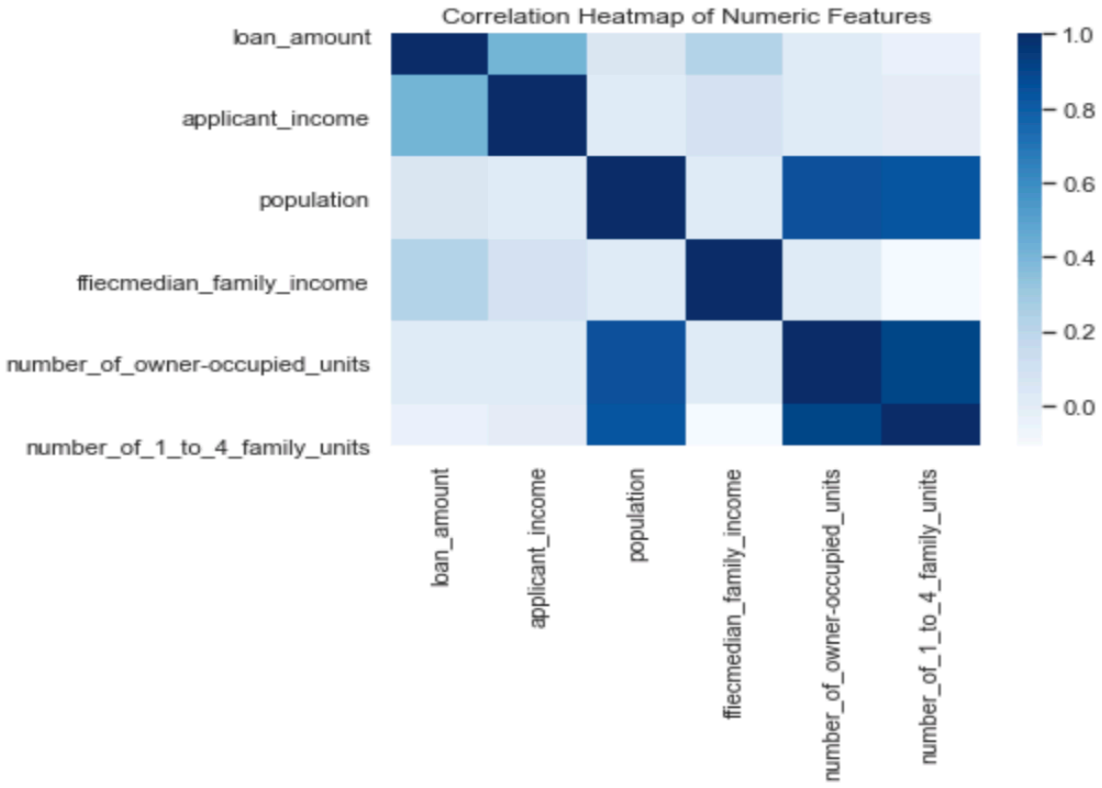

Figure 3. Correlations heatmap of numeric features.

Applicant income and loan amount distribution for mortgage rates are given in the figure below. For any rate spread, loan amount is higher when applicant income is high, so they are directly proportional.

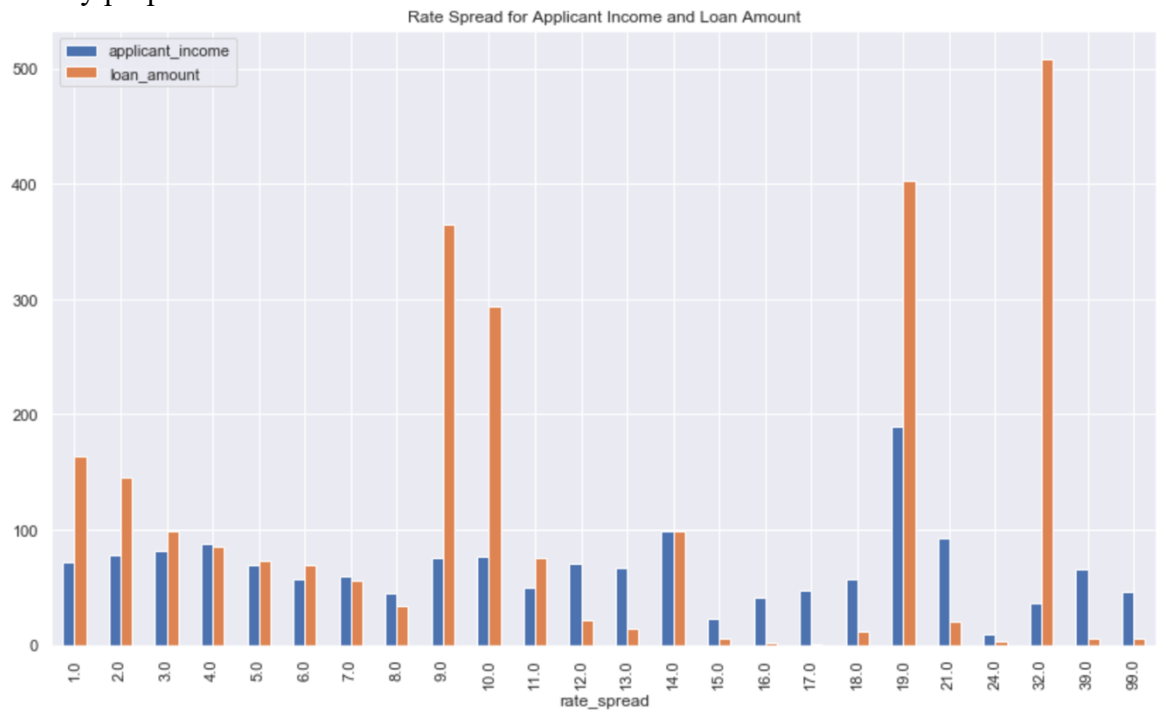

Figure 4. Rate Spread for Applicant Income and Loan Amount

Data Wrangling

Interpretation of these plots and descriptive statistics shows that some features are different distribution as some of them have right skewness or left skewness. Having many categorical data brings its challenge since it is very challenging to interpret it. Missing values of the features are treated based on the type of it. For categorical features, missing values were filled with previous values, and missing values of the numeric features were filled with the mean of each feature. Data has been scaled to increase the precision and the speed of the algorithm

Predictive analysis

Catboost algorithm for the predictive analysis has been adopted and trained. The model has been trained by using 90% and tested %10 of datasets with 500 epochs. The highest score of the model is 69% with learning rate equals 0.05. In addition to score calculation with confusion matrix, feature importance is another metric for the model performance and limitations.

As it can be seen in the table, loan type and property type, loan purpose, occupancy and loan amount are the significant features for the model. Categories with a lot of different values are one of the reasons which increases training time, are msa_md state_code and county. These location features are more important than some of numeric values. Furthermore, they bring more challenges to the model. Lender and co_applicant have no importance for the model which can be dropped to increase the speed of the training. The features which are subtracted from other features such as number_of_1 to_4_family_units or number_of_ owner-occupied_units may be dropped as well since the importance of them are quite low.

	Features	Importances
1	loan_type	11.638246
2	property_type	8.423405
3	loan_purpose	7.781885
4	occupancy	6.312066
5	loan_amount	5.440584
6	preapproval	3.901008
7	msa_md	3.017007
8	state_code	2.464302
9	county_code	2.236660
10	applicant_ethnicity	2.146826
11	applicant_race	1.672571
12	applicant_sex	1.541131
13	applicant_income	1.120769
14	population	1.004773
15	minority_population_pct	0.799553
16	ffiecmedian_family_income	0.727193
17	tract_to_msa_md_income_pct	0.601965
18	number_of_owner-occupied_units	0.558281
19	number_of_1_to_4_family_units	0.551232
20	lender	0.000000
21	co_applicant	0.000000

Table 2. Features importance of the dataset

Conclusion and Recommendations

The analysis has shown that predicting rate spread is quite challenging but possible with given features. The author believes that the current model is able to reliably predict the mortgage rate spread. Feature engineering would increase the overall score of the model, however, this would require a better knowledge of the categorical features.