Predicting Mortgage Approvals From Government Data

Ahmed Meshref, June 2019.

Executive Summary:

This document presents a prediction of Loan approval from government data. The analysis is based on 22 features (observations) with 500,000 data points in total with the goal of predicting whether or not a customer's loan is accepted or rejected.

Starting with a deep understanding of the different features have led to their relationship with the boolean variable of accepting the loan request or not by calculating the summary, descriptive statistics, and creating visualizations of the data. After understanding the data, a classification model to classify the values of 0 and 1 for the loan request based on different features was created.

After creating the model and understanding the relationship between different features and each other and with the loan accepted feature:

All of the different features have a weighted correlation and relationships to each other and they all contributed to build a high accurate model:

- msa_md (categorical) - A categorical with no ordering indicating Metropolitan Statistical Area/Metropolitan Division where a value of -1 indicates a missing value
- state_code (categorical) - A categorical with no ordering indicating the U.S. state where a value of -1indicates a missing value
- county_code (categorical) - A categorical with no ordering indicating the county where a value of -1indicates a missing value
- lender (categorical) - A categorical with no ordering indicating which of the lenders was the authority in approving or denying this loan
- loan_amount (int) - Size of the requested loan in thousands of dollars

- loan_type (categorical) - 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 (categorical) - 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 (categorical) - 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 (categorical) - 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 (categorical) - 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
- applicant_income (int) - In thousands of dollars
- applicant_ethnicity (categorical) - Ethnicity of the applicant; available values are:

    - 1 -- Hispanic or Latino
    - 2 -- Not Hispanic or Latino
    - 3 -- Information not provided by applicant in mail, Internet, or telephone pplication
    - 4 -- Not applicable
    - 5 -- No co-applicant
- applicant_race (categorical) - Race of the applicant; available values are:
    - 1 -- American Indian or Alaska Native
    - 2 -- Asian
    - 3 -- Black or African American
    - 4 -- Native Hawaiian or Other 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 (categorical) - 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
- co_applicant (bool) - Indicates whether there is a co-applicant (often a spouse) or not

- 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 (adjusted annually by FFIEC)
- 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
- row_id - A unique identifier with no intrinsic meaning, but the IDs in your submission must match the submission format exactly
- accepted - Indicates whether the mortgage application was accepted (successfully originated) with a value of 1 or denied with a value of 0

Initial Data Exploration:

This step demonstrates a good understanding of the data via some summary and descriptive statistics.

Summary and Feature statistics:

This shows the mean, standard deviation, max, and min of each features to predict it is distribution and the best way to optimize it, as shown here:

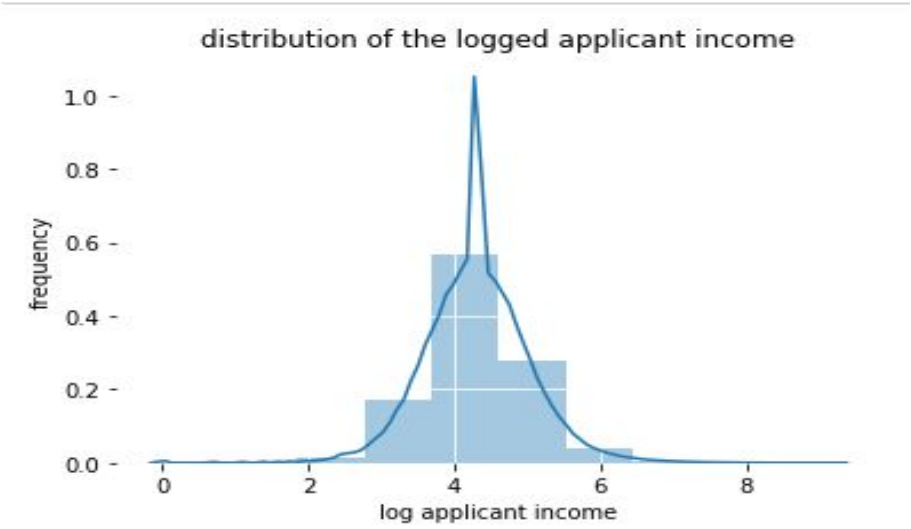
	loan_type	property_type	loan_purpose	occupancy	loan_amount	preapproval	msa_md	state_code	county_code
count	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000
mean	1.366276	1.047650	2.066810	1.109590	221.753158	2.764722	181.606972	23.726924	144.542062
std	0.690555	0.231404	0.948371	0.326092	590.641648	0.543061	138.464169	15.982768	100.243612
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	-1.000000	-1.000000	-1.000000
25%	1.000000	1.000000	1.000000	1.000000	93.000000	3.000000	25.000000	6.000000	57.000000
50%	1.000000	1.000000	2.000000	1.000000	162.000000	3.000000	192.000000	26.000000	131.000000
75%	2.000000	1.000000	3.000000	1.000000	266.000000	3.000000	314.000000	37.000000	246.000000
max	4.000000	3.000000	3.000000	3.000000	100878.000000	3.000000	408.000000	52.000000	324.000000

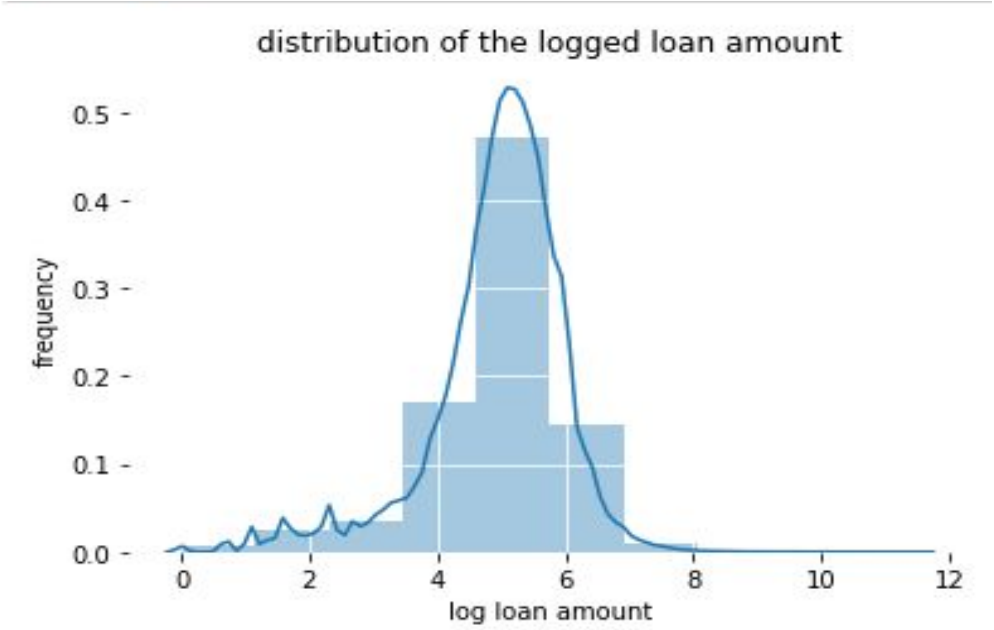
applicant_ethnicity	applicant_race	applicant_sex	applicant_income	population	minority_population_pct	ffiecmedian_family_income
500000.000000	500000.000000	500000.000000	460052.000000	477535.000000	477534.000000	477560.000000
2.036228	4.786586	1.462374	102.389521	5416.833956	31.617310	69235.603298
0.511351	1.024927	0.677685	153.534496	2728.144999	26.333938	14810.058791
1.000000	1.000000	1.000000	1.000000	14.000000	0.534000	17858.000000
2.000000	5.000000	1.000000	47.000000	3744.000000	10.700000	59731.000000
2.000000	5.000000	1.000000	74.000000	4975.000000	22.901000	67526.000000
2.000000	5.000000	2.000000	117.000000	6467.000000	46.020000	75351.000000
4.000000	7.000000	4.000000	10139.000000	37097.000000	100.000000	125248.000000

tract_to_msa_md_income_pct	number_of_owner-occupied_units	number_of_1_to_4_family_units	lender
477486.000000	477435.000000	477470.000000	500000.000000
91.832624	1427.718282	1886.147065	3720.121344
14.210924	737.559511	914.123744	1838.313175
3.981000	4.000000	1.000000	0.000000
88.067250	944.000000	1301.000000	2442.000000
100.000000	1327.000000	1753.000000	3731.000000
100.000000	1780.000000	2309.000000	5436.000000
100.000000	8771.000000	13623.000000	6508.000000

The first thing we can highly notice here is the outliers in most of the numeric features such as the loan_amount, applicant_income, population, minority_population_pct, ffiecmedian_family_income, tract_to_msa_md_income_pct, number_of_owner-occupied_units, and number_of_1_to_4_family_units which demands normalization in the analysis phase.

For the loan amount and applicant_income, getting the log of all data points is an appropriate solution to treat the outliers since it makes the distribution normal as shown here.





Going from the distribution to the missing values report and treatment:

```
loan_type                          0
property_type                      0
loan_purpose                       0
occupancy                          0
loan_amount                        0
preapproval                        0
msa_md                         76982
state_code                     19132
county_code                    20466
applicant_ethnicity                0
applicant_race                     0
applicant_sex                      0
applicant_income               39948
population                     22465
minority_population_pct        22466
ffiecmedian_family_income      22440
tract_to_msa_md_income_pct     22514
number_of_owner-occupied_units 22565
number_of_1_to_4_family_units  22530
lender                             0
co_applicant                       0
dtype: int64
```

So, starting with the numeric missing values and replace each missing value with the median of the feature such as applicant_income, population, minority_population_pct, ffiecmedian_family_income, tract_to_msa_md_income_pct, number_of_owner-occupied_units, and number_of_1_to_4_family_units.

For the categorical features, replacing the missing values was appropriate since the missing values was the dominant and the most frequent data points such as county_code, state_code, msa_md.

Correlation and Apparent Relationships:

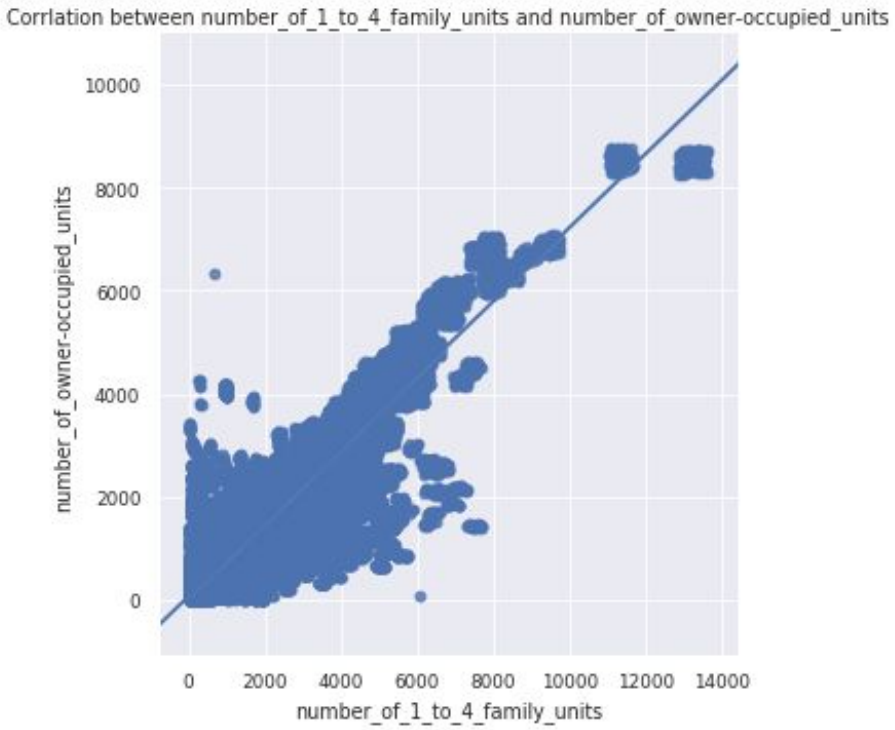
When we get the correlation between all of the 22 features and the target loan approval feature as shown:

```
loan_type                          0.018589
property_type                     -0.080603
loan_purpose                      -0.131595
occupancy                          0.022043
loan_amount                        0.169416
preapproval                        0.017209
msa_md                             0.080046
state_code                         0.088597
county_code                        0.052264
applicant_ethnicity                0.009777
applicant_race                     0.045361
applicant_sex                     -0.038391
applicant_income                   0.178756
population                         0.025540
minority_population_pct           -0.076546
ffiecmedian_family_income          0.070197
tract_to_msa_md_income_pct         0.064809
number_of_owner-occupied_units     0.040934
number_of_1_to_4_family_units      0.012038
lender                             0.008494
co_applicant                       0.101116
dtype: float64
```

We can notice that there is a good correlation with the loan_prepose, loan_amount, applicant_income, co-applicant. But, why did we use all other features? Why not using only the mentioned ones?

To answer this question, we need to see whether or not the other features correlated with each other because that play a very important part of our final accuracy.

Let’s see the correlation between the number_of_owner-occupied_units and number_of_1_to_4_family_units which both doesn’t seem to have good correlation with the loan approval feature:



Classification of the Loan Approval:

Based on the analysis of all features which included understand and preparing the data, a model was built to classify the loan approval process into accepted indicated by 1 and not accepted which indicated by 0.

A CatBoostClassifier model is built using catboost trained with 75% of the training data. Testing the model with the remaining 25% of the data yielded the following results:

- True Positives: 48935
- True Negatives: 42356

- False Positives: 20129
- False Negatives: 13580

With a total accuracy of 73%.

Classification:

The CatBoostClassifier model was created to predict the loan approval whether getting 0 for “not accepted” or 1 for “accepted”. After identifying the relationship between different variables and getting the log of some categorical features to get the target feature of loan approval. The model was trained with 75% of the trained data, and tested with the remaining 25% of it to get a good accuracy of 73%. Using hyperparameter tuning with the model have helped to increase the accuracy by identifying the categorical features again to the model and sitting the accuracy rate to 0.23 with depth of 6. Building such a model can predict whether or not to accept the loan request of any new customer if we feed the model with the needed data which can be used by banked to automate the loan request and fasten it.

Conclusion

Many bank customers ask for loans, but they wait for a long time to get a response back from the bank after submitting all what's needed. With the classification model produced in this document, banks can easily check whether or not to accept a loan request by feeding the needed data to this model. Analysis have done for 22 features from the applicant income to loan amount to fix the missing values and treating the outliers of the model. Then, CatBoostClassifier was created to train 75% of the data and test on the 25% remaining. The model reached an accuracy of 73% which is quite good

with the ability to learn for feature customers and improve this accuracy. In conclusion, with this model, we can be 73% sure about whether or not to accept any customer loan request.