The capstone Project for the Microsoft Professional Program in Data Science: Predicting Mortgage Rates From Government Data

Autor: Christian Lennin Rua Giraldo. Date: November 2019

The executive summary report.

The present analysis show how lender, loan amount, property type, loan type and family median income and other features are related to rate spread of mortgage applications according to the Federal Financial Institutions Examination Council's (FFIEC). The rate spread data has 21 features and a target Reat_Spread feature, which indicates the difference between the offered mortgage rate for the applicant and the standard rate for a comparative mortgage, the data set is about 200.000 loans applications.

Catboost Regressor was used to perform the model, and after the training process, the next 5 features was identificated as the most important ones:

Feature	type	% relevance
lender	object	32.10630031
loan_amount	float64	23.17930374
property_type	object	10.93395925
loan_type	object	6.48635725
ffiecmedian_family_income	float64	6.05969682

The definition of this features are:

lender: A categorical with no ordering indicating which of the lenders was the authority in approving or denying this loan

Loan_amount: Size of the requested loan in thousands of dollars

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

Loan_type: indicates whether the loan or application was for a one-to-four-family dwelling (other than manufactured housing), manufactured housing, or multifamily dwelling

ffiecmedian_family_income: FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC).

The conclusion of the project was to determinate which features has most relevance, and find, select and train the best model to get the highest R^2 possible, for this case was the CatBoost Regressor and the features already mentioned.

Data Cleansing

At the beginning all the categorical features, except co_applicant feature, where wrong codify, because all of them was read like int64, so the next step was to convert them to object type.

All the numeric features, excluding the ffiecmedian_family_income, has a left skewness. some of the categorical features like county_code (306 unique values), state_code (53 unique values), msa_md (409 unique values) and lender (3893 unique values), has more categories than the others categorical features, so this situation complexes a few the analysis of the data.

In the missing data, it was found that the following :

Feature	Num Missing	% Missing
state_code	1338	0.669
applicant_income	10708	5.354
population	1995	0.9975
minority_population_pct	1995	0.9975
ffiecmedian_family_income	1985	0.9925
tract_to_msa_md_income_pct	2023	1.0115
number_of_owner-occupied_units	2012	1.006
number_of_1_to_4_family_units	2016	1.008

All the missing data was filled whit the value -999, with this value out of distributions of the features, the catboost model would be able to easily distinguish between them and take it into account.

Data Exploration

1. The most common loan type was FHA-insured (Federal Housing Administration) with 53.15%, followed by the type Conventional (any loan other than FHA, VA, FSA, or RHS loans) with 45.35%.
2. In the property type we can find that One to four-family (other than manufactured housing) (type 1), has the 84.63%, and the type Manufactured housing has the 15.25%
3. In the loan purpose the most common type is home purchase with 73.05% of the total loans, followed by the refinancing purpose whit 21.31%.
4. Seeing occupancy feature, with 94.01% the Owner-occupied as a principal dwelling, is the most relevant followed by not owner occupied with 5.84%
5. Meanwhile in the pre-approval feature the most common is to see that the 74.74% of the loans does not apply this type of procedure, only the 20.81% pre approval was not requested.
6. Lender is explained in the exploratory data analysis section.
7. In the applicant ethnicity, with 74.01% of the total data, not Hispanic or latino are the most common, followed by Hispanic or latino with 17.40.
8. Now, in the race feature, the most common race is the white race with 78.80, followed by black or African American with 10.37%.
9. In the sex feature, the male sex have 62.49%, and female sex has 33.48%
10. In the co applicant feature, 61.64% do not have co applicant, and the 38.35% has co applicant.
11. The top Five state with most loans are:

state_code	count	mean	median	rate_spread %
48	25593	2.108506	1	12.797396
37	16455	1.709085	1	8.228076
30	15789	1.848819	1	7.895053
33	8129	1.600812	1	4.064785
50	7861	2.040707	1	3.930775

12. The top five county code are:

state_code	county_code	count	mean	median	rate_spread %
33	**245**	5448	1.544787	1	2.724191
48	**20**	3316	1.679433	1	1.658116
30	**46**	2603	1.938148	1	1.301591
24	**101**	2495	1.346293	1	1.247587
12	**83**	2189	1.511192	1	1.094577

13. The top five ms mda:

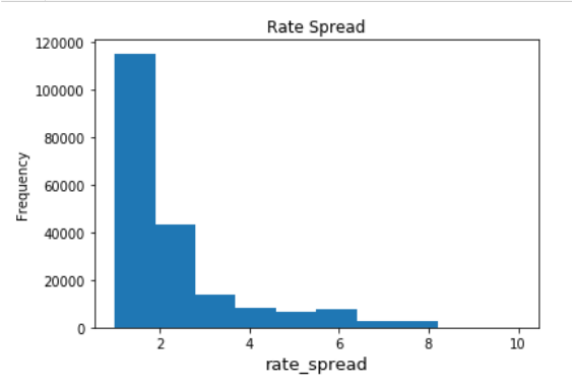
msa_md	count	mean	median	rate_spread %
261	35650	2.482272	2	17.826248
154	6237	1.551066	1	3.118718
352	5651	1.825341	1	2.825698
215	4367	1.415617	1	2.183653
345	4016	1.883466	1	2.008141

Exploratory Data Analysis (EDA)

Numeric features

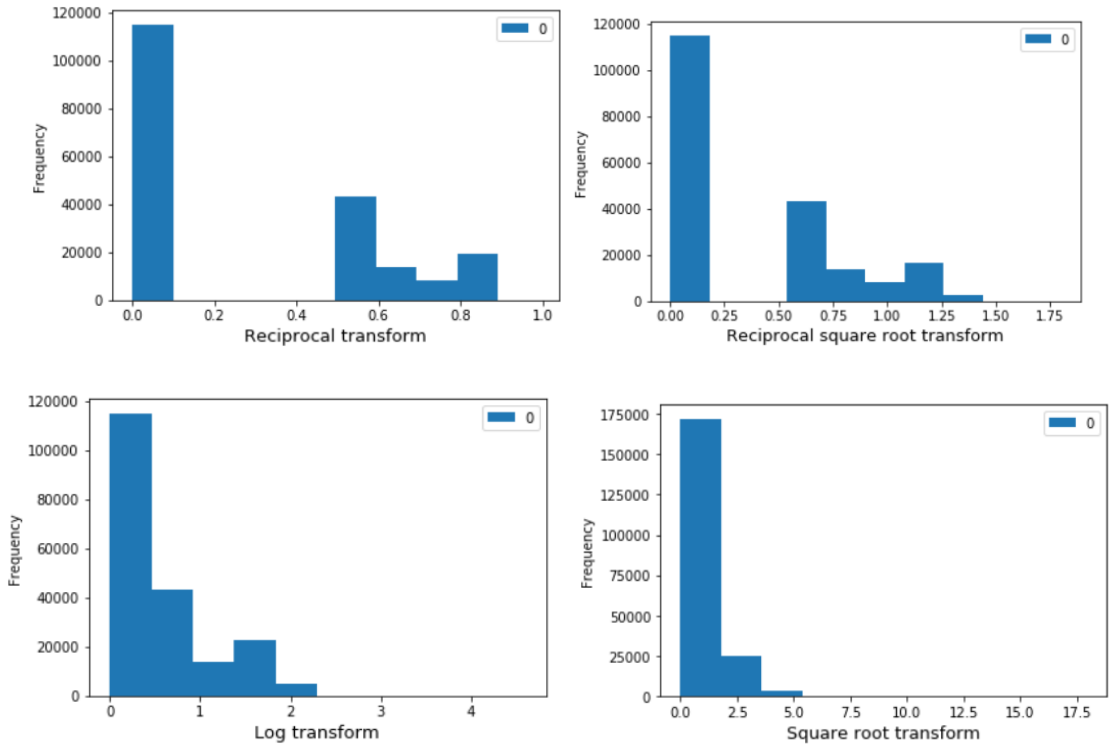
The exploratory data analysis allows us to see the main characteristics of the data, the shape, the distribution and the summary statistical data.

The target feature is rate_spread, the distribution and the summary data is:



count	200000
mean	1.97911
std	1.656809
min	1
25%	1
50%	1
75%	2
max	99

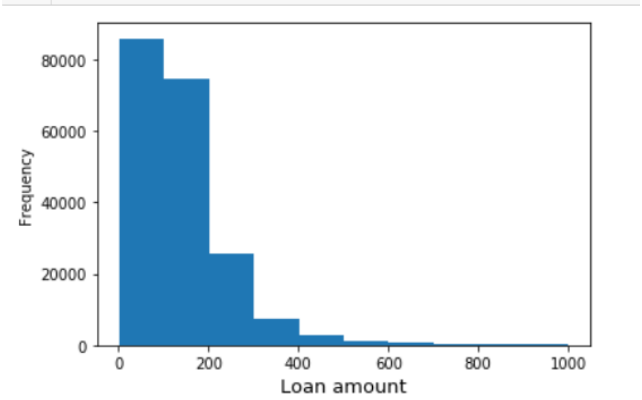
The target feature is very left skewed, and has outliers, because the third percentile is 2, and the maximum value is 99. If we tray corrected the skewness by applying different type of transformations, we have the next:

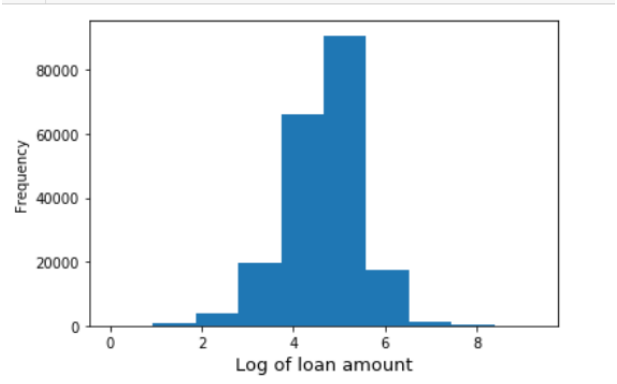


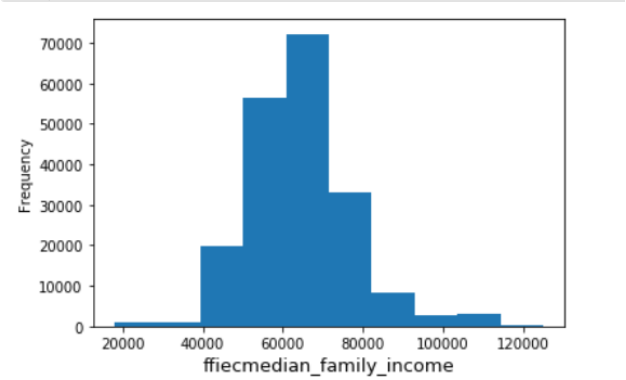
The other two numeric features has the next summary statistics:

	loan_amount	ffiecmedian_family_income
count	200000.000000	198015.000000
mean	142.574940	64595.355801
std	142.559487	12724.514485
min	1.000000	17860.000000
25%	67.000000	56654.000000
50%	116.000000	63485.000000
75%	179.000000	71238.000000
max	11104.000000	125095.000000

The distribution of this numeric features are:







has we can see, the loan_amount feature has some left skewness, but, this feature allows us to correct this by applying a log transformation, the ffiecmedian_family_income, has a distribution more closer to the normal distribution.

If we check the correlation index between this features vs the target feature, we have this:

```
Correlation rate spread log vs loan amount log
 [[ 1.         -0.47221972]
 [-0.47221972  1.        ]]
Correlation rate spread vs loan amount
 [[ 1.         -0.21816759]
 [-0.21816759  1.        ]]
```

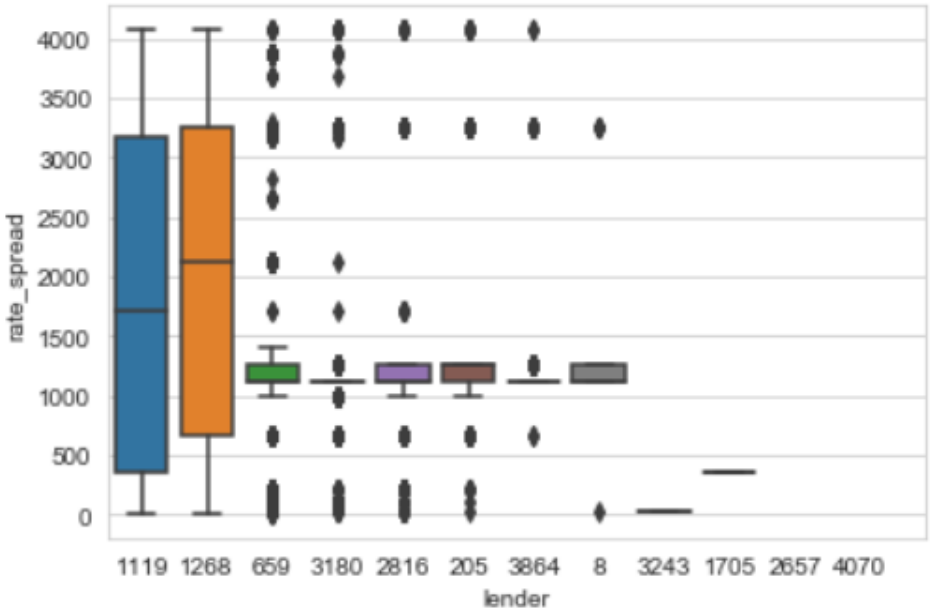
```
Correlation rate spread vs ffiecmedian_family_income
 [[ 1.         -0.1034093]
 [-0.1034093  1.       ]]
Correlation rate spread log vs ffiecmedian_family_income
 [[ 1.          -0.11728164]
 [-0.11728164  1.        ]]
```

The relation is more stronger when we have the log of the rate_spread and the log of loan_amount, the level of the correlation is near to -0.4722

Categorical features

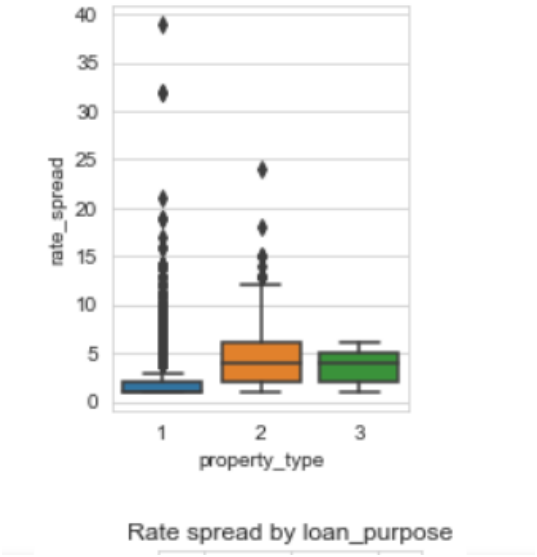
The most important categorical feature is lender, in this feature we can find the code of the different lenders in the country, and the relation with the target feature is the next one:

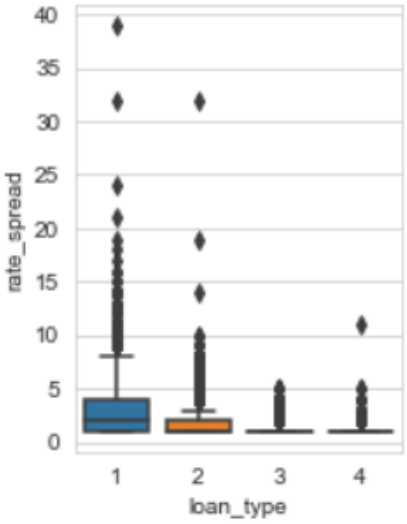
lender	count	mean	median	rate_spread %
1119	9235	5.696156	6.0	4.617823
1268	5445	5.340312	6.0	2.722691
659	5217	1.400230	1.0	2.608683
3180	3174	1.176749	1.0	1.587111
2816	2973	1.030609	1.0	1.486604
205	2558	1.394840	1.0	1.279090
3864	2376	1.343434	1.0	1.188083
8	2353	1.146621	1.0	1.176582
3243	2287	2.264539	2.0	1.143580
1705	2100	1.190476	1.0	1.050074
2657	2064	1.148740	1.0	1.032072
4070	1991	1.221497	1.0	0.995570



As we can see the median of the most relevant lenders are different, when is compared with the target feature, for this reason the variability of the lenders has most relevance on the model.

The same phenome is present in the features propery_type and loan_type.





Rate spread by loan_purpose

Training the model

CatBoost is a machine learning algorithm that handle categorical (CAT) data automatically. The model derives its name from the word **Category** and **Boosting**. It is a model that works well with multiple categories of learning problem such as classification and regression.

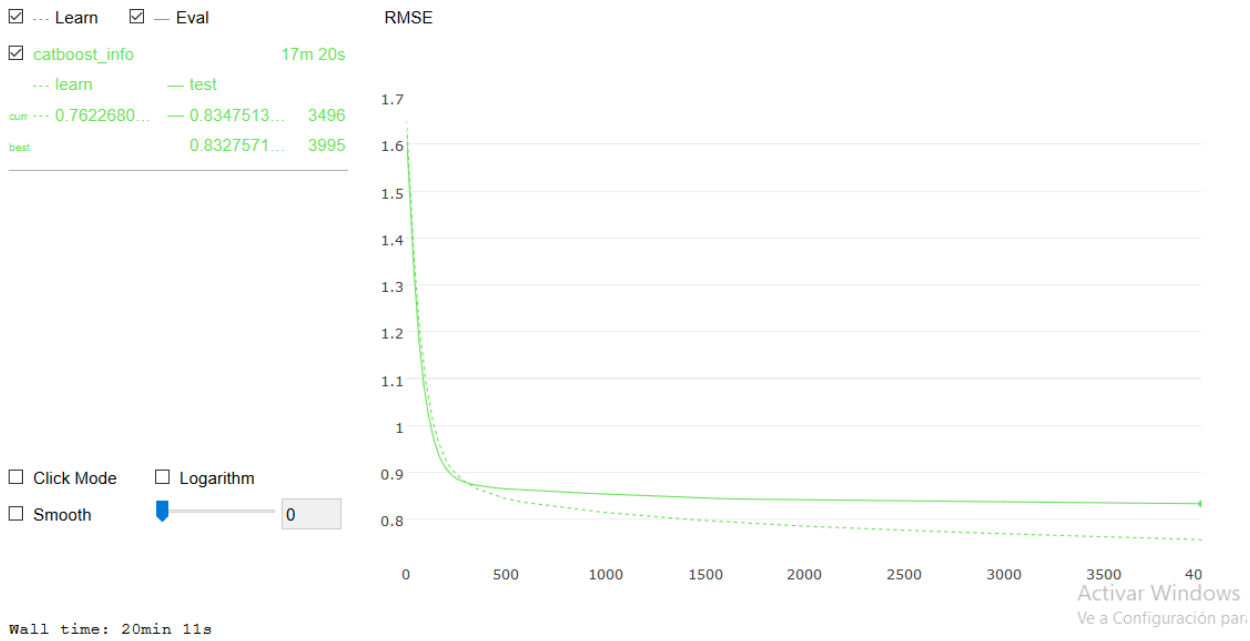
CatBoost provides state of the art results and it is competitive with any leading machine learning algorithm on the performance front.

For the training process, I started with the definition of the categorical features indices, this because the model has a parameter with this kind of information.

After that, I used the train split method to separate the date, in the train and tests packages, the distribution was 85% for the training data and the rest 15% for the test data.

I defined the CatBoostRegressor object with the next parameters:

iterations =4000, depth = 8, learning_rate = 0.01, eval_metric = 'RMSE',
use_best_model = True, random_seed=40, verbose=False



The total time to training was the 20 minutes, 11 seconds.

The features importance was:

Feature	type	% relevance
lender	object	32.10630031
loan_amount	float64	23.17930374
property_type	object	10.93395925
loan_type	object	6.48635725
ffiecmedian_family_income	float64	6.05969682
loan_purpose	object	5.17448925
applicant_income	float64	4.82051844
state_code	object	2.04740916
applicant_ethnicity	object	1.86061596
msa_md	object	1.65411562
occupancy	object	1.18500877
preapproval	object	0.93269786

applicant_race	object	0.77693989
county_code	object	0.73099285
minority_population_pct	float64	0.70575342
tract_to_msa_md_income_pct	float64	0.35725124
applicant_sex	object	0.3262937
population	float64	0.21875734
number_of_1_to_4_family_units	float64	0.16133864
number_of_owner-occupied_units	float64	0.14611424
co_applicant	bool	0.13608625

the metrics for the best score for all the training process was:

```
'learn': {'RMSE': 0.7559964128278903},
'validation': {'RMSE': 0.8327571036530814}
```

The R^2 for the training data was 0.8026, and the R^2 for the test data was 0.7410

Conclusions

The majority of the citizens who request for the loans are white people, mostly of them lives in the state 48, have in average loan amounts of 142.000 US Dollars, request for the loan in the majority of cases for purchase a house, and lived in it.

The mean in the difference between offered mortgage rate for the applicant and the standard rate for a comparative mortgage, was 1.97%.

The most important features to predict this difference are lender, loan amount, property type, loan type and ffiecmedian family income