Microsoft DAT102x: Predicting Mortgage Rates

Luka Paradiz Udovc, November 2019

Executive Summary

The goal of this Microsoft Professional Capstone project on Data Science was to predict the rate spread of mortgage applications from the US Government data. The analysis presented in this report is based on one year's worth of HMDA-reported loan application, as provided by the Federal Financial Institutions Examination Council (FFIEC).

After performing the initial data exploration and visualizations, several relationships between rate spread and application characteristics were observed. Since rate_spread is float value, this was a regression problem, and *CatBoostRegressor* from the catboost library was used for predicting the target label. Overall, the public score of 0.75 was achieved for the R squared coefficient of determination.

While the dataset contained 21 variables per loan application, it was concluded that the most relevant features for rate spread prediction were (in order of importance):

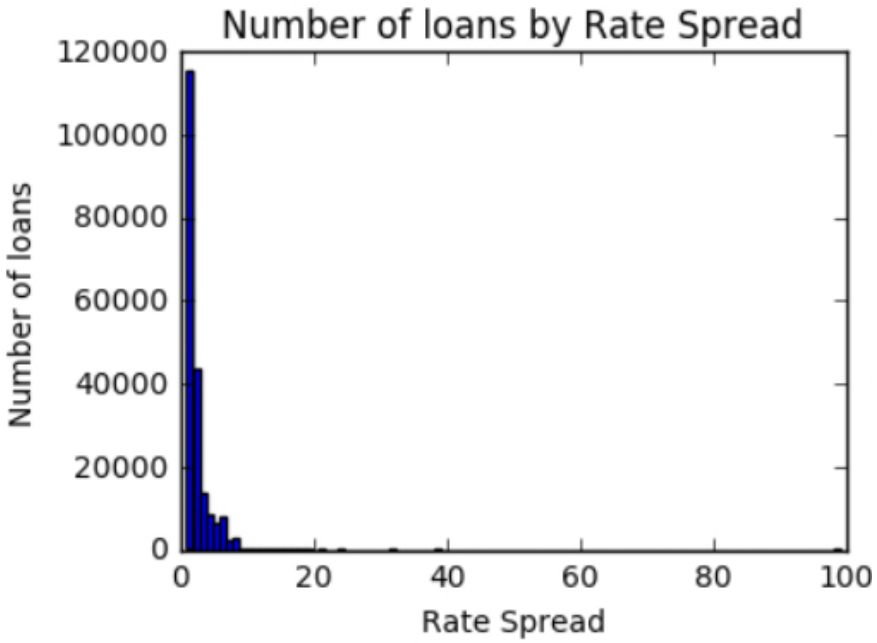
- Lender – containing the information on which of the lenders was the authority in approving or denying the loan
- Loan Amount - containing the information on size of the requested loan in thousands of dollars
- Property Type - containing the information on the type of dwelling the mortgage was applied for; either one-to-four-family housing (other than manufactured housing), manufactured housing, or multifamily dwelling
- Loan Type - containing the information on the mortgage type applied for; either conventional, government-guaranteed, or government-insured
- Applicant income - containing the information on applicant's income in thousands of dollars

Initial Data Exploration

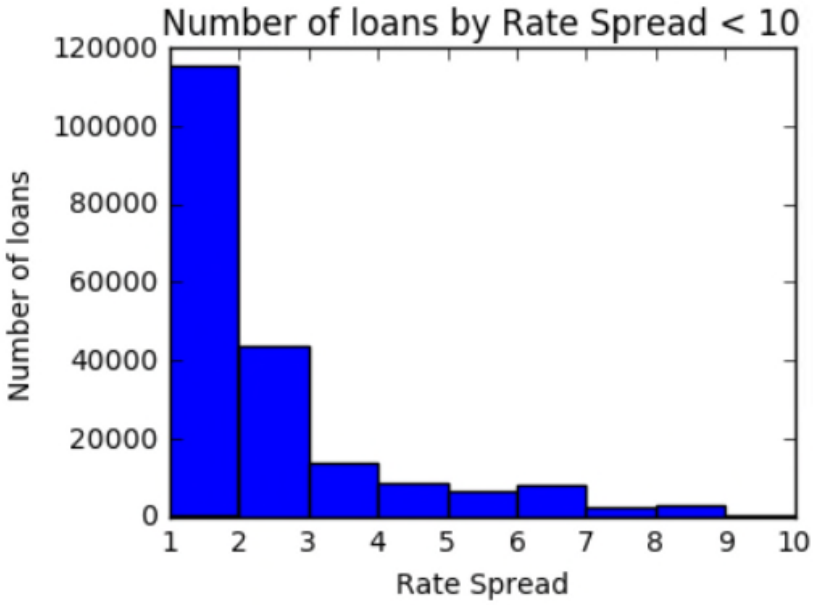
First, some potential features were converted into more appropriate data types (to correspond to numeric or categorical entries). Afterwards, in order to start exploring the dataset, descriptive statistics were calculated on the numerical columns.

Column	Min	Max	Mean	Median	Std Dev	D Count
loan_amount	1.00	11104.00	142.57	116.00	142.56	200000
applicant_income	1.00	10042.00	73.62	56.00	105.70	189292
population	7.00	34126.00	5391.10	4959.00	2669.03	198005
minority_population_pct	0.33	100.00	34.24	26.00	27.93	198005
ffiecmedian_family_income	17860.00	125095.00	64595.36	63485.00	12724.51	198015
tract_to_msa_md_income_pct	6.19	100.00	89.28	98.96	15.06	197977
number_of_owner-occupied_units	3.00	8747.00	1402.87	1304.00	706.88	197988
number_of_1_to_4_family_units	6.00	13615.00	1927.34	1799.00	886.58	197984
rate_spread	1.0	99.00	1.98	1.00	1.66	200000

The values of **rate_spread** is the label that what we were trying to predict in this project. From the summary statistics above and the histogram below, we can see that the minimum value of 1 is also the most frequent value, the mean is close to 2, but the max value is as large as 99.



An unfiltered histogram of the **rate_spread** shows an extreme right-skewedness. In order to inspect the distribution in more detail, a filtered histogram was constructed as shown below.



It can now be seen that the most applicants were offered a mortgage at rates close to the standard competitive rate (rate_spread=1). Fewer and fewer mortgages are offered at increased spread rates, where past the rate of 9, the cumulative number of mortgages if as low as 1% of the total applications.

In addition to the numeric values, the mortgage application records include categorical features, with the following possible values:

- loan type: conventional, government-guaranteed, or government-insured;
- property type: one-to-four-family dwelling, manufactured housing, or multifamily dwelling;
- loan purpose: home purchase, home improvement, or refinancing;
- occupancy: owner occupied, not owner occupied or N/A
- preapproval: requested, not requested or N/A
- applicant ethnicity: Hispanic/Latino, not Hispanic/Latino, Not provided, N/A, no co-applicant
- applicant race: American Indian/Alaska Native, Asian, Black/African American, Native Hawaiian or Other Pacific Islander, White, not provided, N/A, no co-applicant
- applicant sex: male, female, not provided, N/A
- lender: indicating the loan approval authority (one of 3893 approvers in the dataset)

as well as some additional features on the property locations, such as the area, state and county.

Bar charts were created to indicate the frequency of the above features, showing the following:

- Most the loans were either FHA-insured or conventional, with only a few VA-guaranteed or FSA/RHS making up the rest
- Overwhelming majority of the loans were in relation to 1-4 family home, owned-occupied categories
- About 70% of the loans were issued for a home purchase, 20% on refinancing and the rest on home improvement
- Pre-Approval was not a metric that was applicable to most applicants, and so it is unlikely that it will be of use for label prediction
- 2/3 of the applicants were male, overwhelmingly white or at least of non-Latino ethnicity

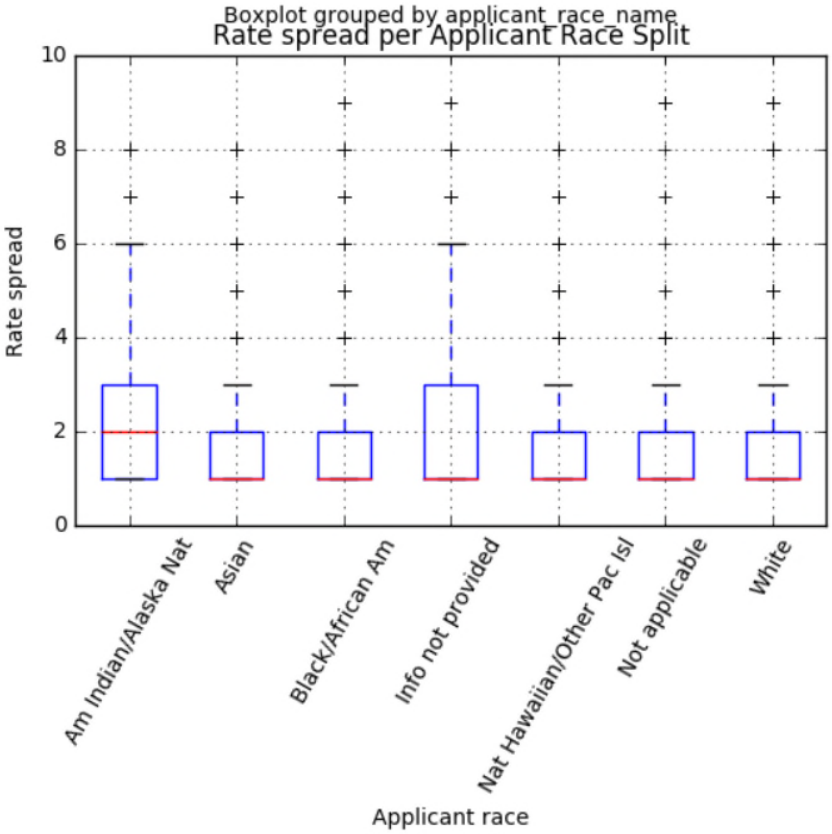
Correlations and Apparent Relationships

After exploring the numerical and categorical features in themselves, an effort was made to observe any outliers and establish trends in relation to rate_spread that could help us in training the machine learning model.

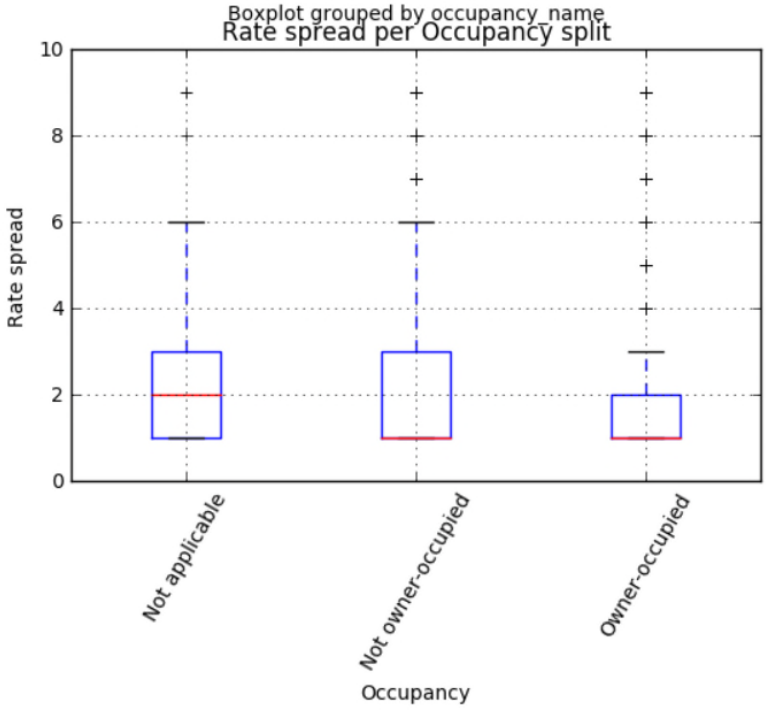
Categorical Relationships

The following box plots show the most interesting observations when grouping-by on categorical features.

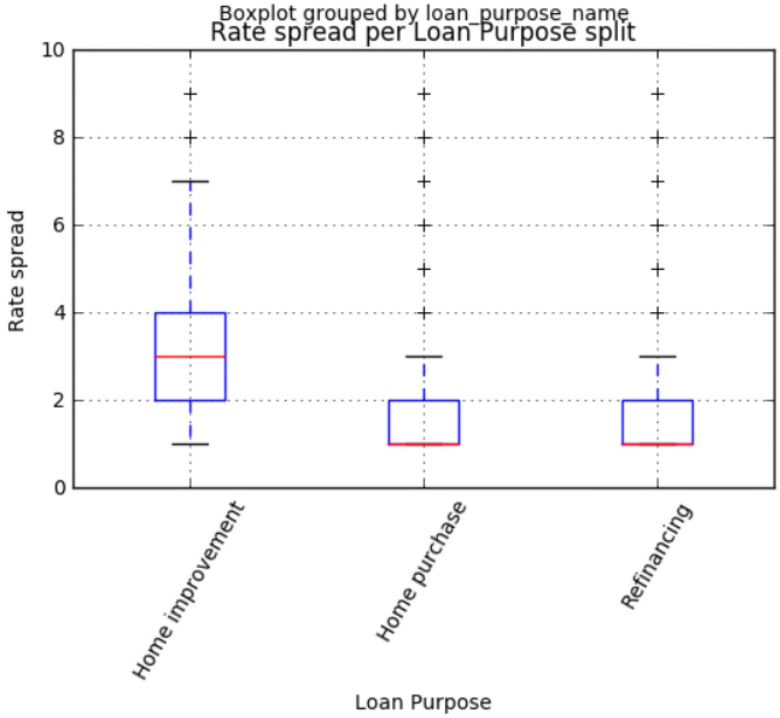
APPLICANT RACE: We can see below that whereas the different race categories show the same medians and upper quartiles, except for the American Indian or Alaska native group – which exhibits an overall less favorable rate spread distribution.



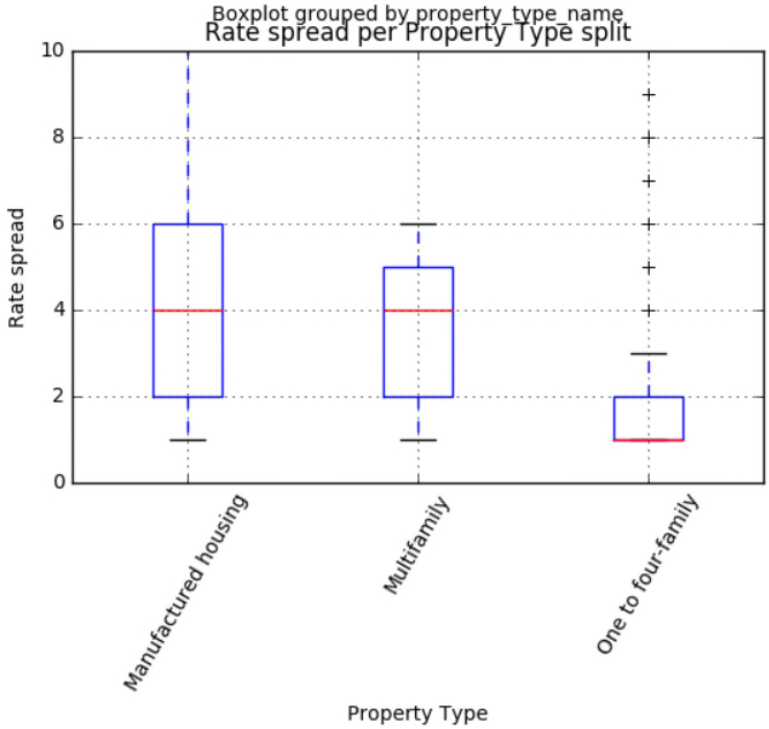
OCCUPANCY: We can see below that the mortgages for the owner-occupied dwellings received better spread rates than the not owner occupied ones (compare the differences in the extents of 'whiskers').



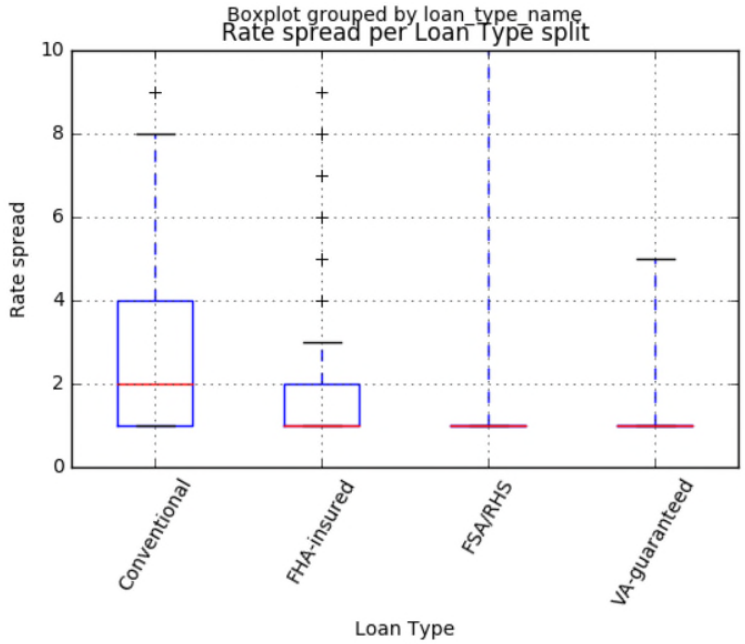
LOAN PURPOSE: We can see below that the loans used for home improvements received considerably worse mortgage conditions than the ones for home purchases or refinancing (3-1 ratio for medians).



PROPERTY TYPE: We can see below that the loans used for the one to four family dwellings received much more favorable spread rates than the multifamily or manufactures housing categories.



LOAN TYPE: We can see below that the rate spread behavior per loan type can be split into 2 categories: the conventional + FHA-insured applications, and the FSA/RHS + VA-guaranteed ones. The latter exhibit a very suppressed median/quartiles ratio but a huge spread of outliers. The former behave much more predictably, but we can clearly see that FHA-insured loans overall fare more favorably to the conventional ones.



Overall, it was also observed that a large proportion of all entries across most categories corresponds to Information nor provided or not applicable, which unfortunately were not very useful rubrics for making predictions.

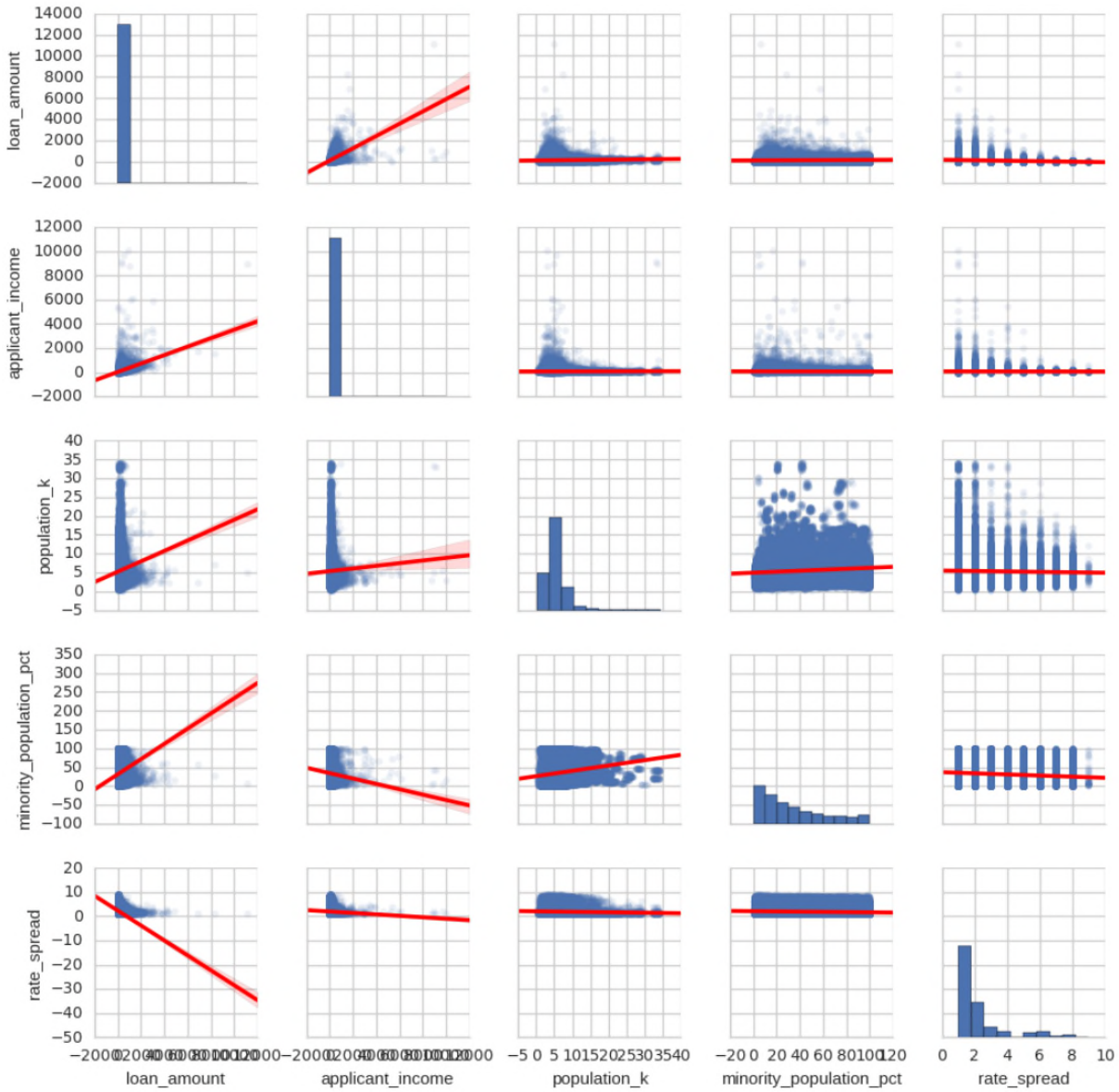
Numerical Relationships

First, correlation between the numeric columns and rate_spread was calculated to identify the most promising features to investigate further.

Feature	**Pearson Correlation with rate_spread variable**
loan_amount	-0.218168
applicant_income	-0.020662
population	-0.034157
minority_population_pct	-0.076955
ffiecmedian_family_income	-0.084964
tract_to_msa_md_income_pct	0.010798
number_of_owner-occupied_units	0.004927
number_of_1_to_4_family_units	0.020661
co_applicant	0.042659

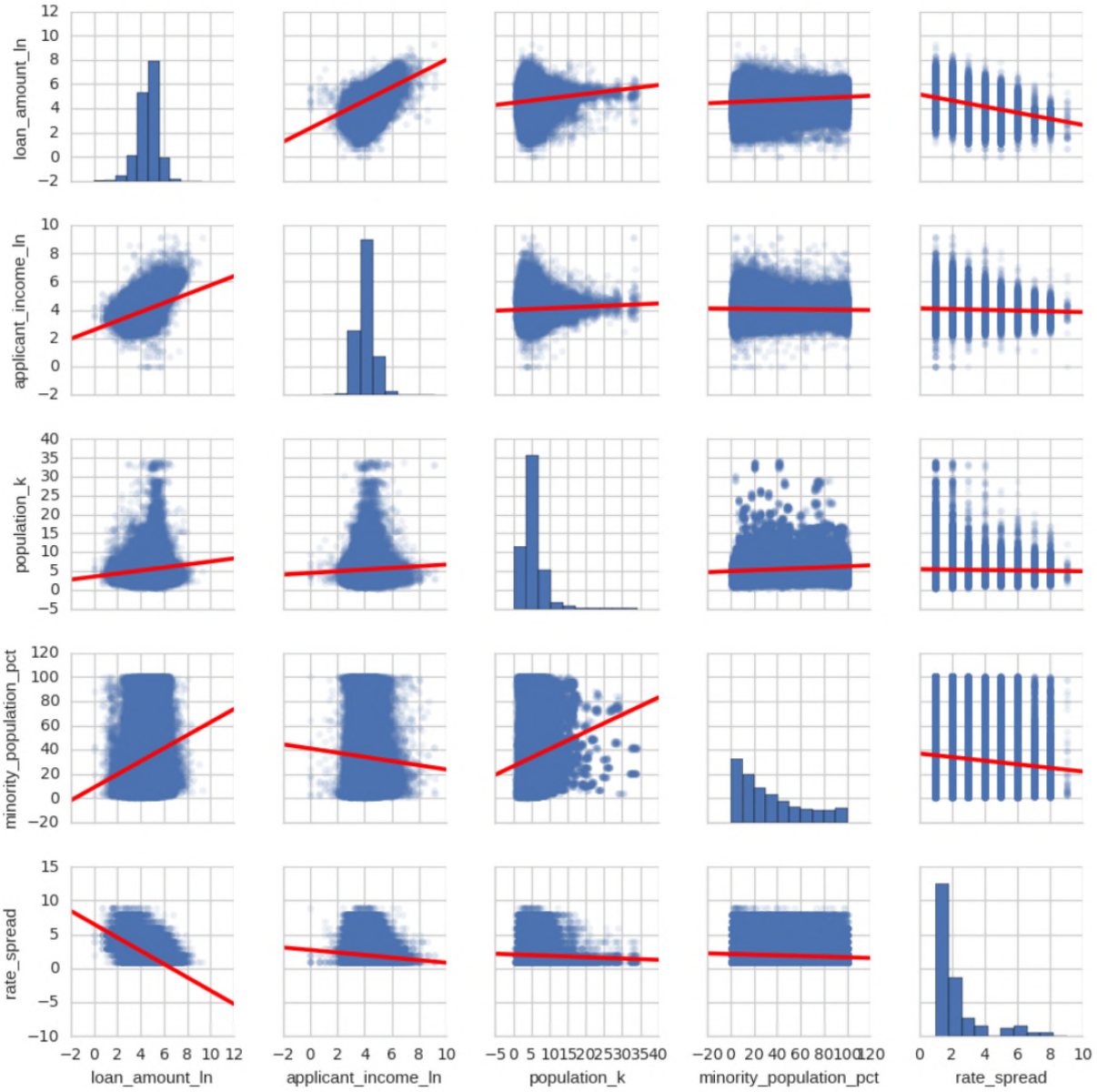
The somewhat baffling result that as the loan amount increases, the rate spread of the mortgage offered decreases (i.e. is more favorable) can be explained by the fact that the loan amount and applicant are strongly positively correlated. In other words, applicants with a large income can afford to borrow more money at better rates, which fits in with our intuitions.

A scatter-plot matrix was generated to compare the these promising numeric features with one another, and especially against the rate_spread variable. A linear regression trend-line was included to indicate the pair-wise behavior of these variables.



Viewing plots in the bottom row or the right-most column of this matrix shows an the relationship between spread rate and other numeric features.

It can be seen that histograms of loan amount (sub-plot 1) and applicant income (sub-plot 7) shows very right-skewed behaviors, and so a natural logarithm of these values was calculated and the scatter-plot matrix re-plotted. With these new logarithmic values, the linear relationships with spread rate are much clearer (steeper trends), which is confirmed also by an increased correlation values (to -0.459279 for loan_amount_ln and -0.070935 for applicant_income_ln).



Prediction of Rate Spread Label

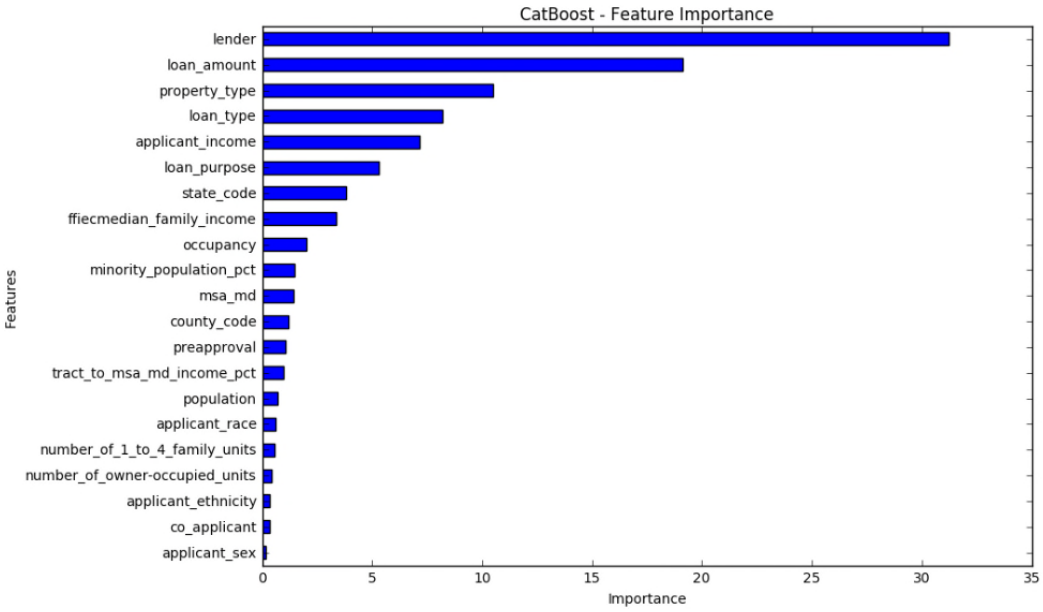
Since our features included a lot of categorical variables, it has been decided to use CatBoostRegressor, which is open-sourced machine learning algorithm from Yandex. The main benefit of using this machine learning algorithm, is that it automatically converts categories into numbers by using various statistics on combinations of categorical features and combinations of categorical and numerical features (source: https://www.analyticsvidhya.com/blog/2017/08/catboost-automated-categorical-data/)

First, the -1 values of msa_md, state_code, county_code variables were replaced with NaN and then interpolation was used to fill in the missing values of all the columns. Finally, the dataset was split 70-30 between the training and the validation portions for the machine learning step.

After several iterations and optimization steps, the CatBoostRegressor was parametrized in the following way: iterations=5000, depth=8, learning_rate=0.01, loss_function='RMSE', use_best_model = True

It can be seen from the above plot that after the first hundred iterations or so the fit to data improved rapidly (large RSM decrease) , and then afterwards the gains were much smaller up to the 5000th iteration. Overall, the RMS to the training portions of the dataset reached the RMS of 0.79 (green line) and 0.95 for the evaluation portion (dotted green line).

The trained model listed the features in the following order of importance:



The trained CatBoostRegressor model was then used to with the test dataset, for which no rate spread values were provided, in order to predict these labels. On this dataset, the model achieved a public score of 0.75 for the R squared coefficient of determination, when submitted on the DataDriven competition website.

Conclusions

Overall, the overall score shows that relationships between the numerical and categorical variables in relation to the label we wanted to predict were well understood and properly handled. Likewise, we did not over-fit the training dataset. CatBoostRegressor model was chosen for its simplicity in handling a large number of categorical features, fast performance and the ease of use. Very limited amount of pre-processing and feature engineering was carried out, which is why there remains much room of improvement