Analysis of Rate Spread between Mortgage Applications

James Loffler, November 2019

Executive Summary

This document presents an analysis of mortgage applications with an overall target of predicting an individuals rate spread. The FFIEC website defines rate spread as: *"the spread between the Annual Percentage Rate (APR) and a survey-based estimate of APRs currently offered on prime mortgage loans of a comparable type utilizing the "Average Prime Offer Rates" fixed table or adjustable table, action taken, amortization type, lock-in date, APR, fixed term (loan maturity) or variable term (initial fixed-rate period), and reverse mortgage."*

The dataset used for training the predictive model consisted of 200,000 observations and 21 variables relating to mortgage applications and census data. After initial data exploration with summary statistics, data analysis techniques were used to identify relationships in the data that may contribute to the predictive model. After identifying a number of predictive features, a regression model was deployed to another 200,000 observations.

After performing the analysis, the following conclusions were presented:
Although a number of features were found to contribute to the overall model performance, the variables below were found by the analysis to be particularly predictive:

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 granted, applied for, or purchased was conventional, government-guaranteed, or government-insured.

loan_amount - Size of the requested loan in thousands of dollars

preapproval - Indicate whether the application or loan involved a request for a pre-approval of a home purchase loan.

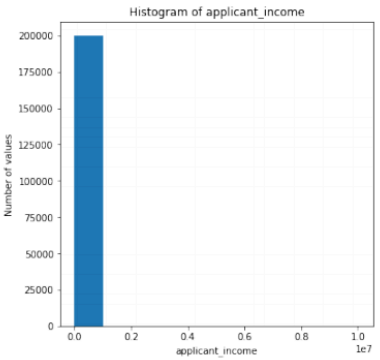
Initial Data Exploration

The dataset was initially explored to assess data quality and imputation was applied to treat null values, generally with the median for each variable. As all the variables contained numeric data, the categorical values were recoded to help understand the properties of the data during analysis. After the initial data preparation, summary statistics were performed on the numeric variables as an early indication of distribution.

	loan_amount	applicant_income	population	minority_population_pct	ffiecmedian_family_income	tract_to_msa_md_income_pct	number_of_owner_occupied_units	number_of_1_to_4_family_units	rate_spread
Count	200000	200000	200000	200000	200000	200000	200000	200000	200000
Mean	142.57494	73239.83528	5386.78891	34.15642004	64770.9562	89.38089451	1401.877745	1926.042985	1.97911
Median	116	58000	4959	25.996	63460	98.959	1304	1799	1
Mode	25	41000	4959	25.996	63460	100	1304	1799	1
Minimum	1	1000	7	0.326	6000	6.193	3	6	1
Maximum	11104	10042000	34126	100	2409000	100	8747	13615	99
Standard Deviation	142.559487	102893.0944	2656.030727	27.80329552	18667.57401	15.01411828	703.3850047	882.1900531	1.656809383
Sample Variance	20323.20733	10586988878	7054499.222	773.0232415	348478319.5	225.4237476	494750.4648	778259.2898	2.745017333
Kurtosis	425.9639343	2396.701201	15.45571695	-0.44749706	3180.517441	1.559689793	9.241377762	12.74505994	182.2942744
Skewness	11.59070453	35.02142566	2.662396134	0.823490821	36.36544893	-1.48494176	1.881192178	2.066849407	5.021600555
Range	11103	10041000	34119	99.674	2403000	93.807	8744	13609	98

The summary statistics seemed to show that a number of the features followed a skewed distribution. A large variance between the mean and median values will often indicate a skewed distribution which is also confirmed by the value of '*Skewness*'. To understand the distribution fully, histograms were created for all the variables and a right-tailed skew was confirmed for the variables shown in the diagram below. This analysis led to the conclusion that min/max scaling would be the appropriate choice for any transformation. An understanding of the rate spread distribution is important to note for the rest of this analysis and the frequency count in fig. 3.1 may enhance this.

Fig. 1.1, 1.2



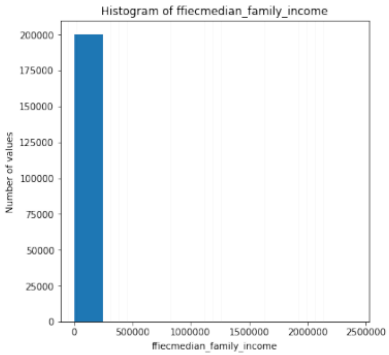
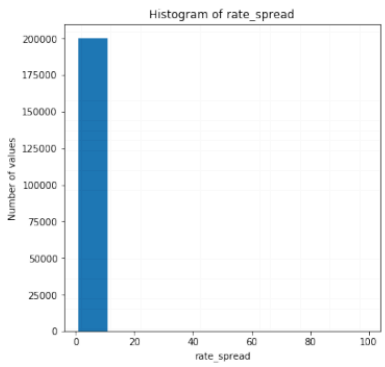
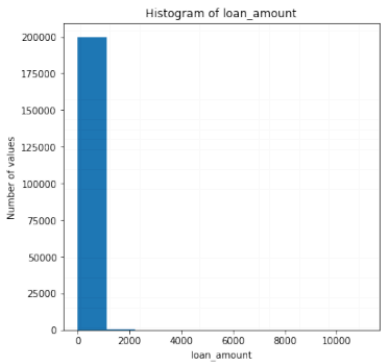


Fig. 1.3, 1.4





The dataset also included a number of categorical variables that were coded as numeric values. The categorical data was influential in the overall model performance so it was important to understand the relationship between these characteristics and rate spread. A number of boxplots were created to understand the distribution between rate spread and each of the categorical values. To make the visualisations easier to read, rate spread values of greater than 25 were removed from the boxplots. A brief description of each variable is below:

loan_type - 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 loan)*
2. *FHA-insured (Federal Housing Administration)*
3. *VA-guaranteed (Veterans Administration)*
4. *FSA/RHS (Farm Service Agency or Rural Housing Service)*

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; available values are:

1. *One to four-family (other than manufactured housing)*
2. *Manufactured housing*
3. *Multifamily*

loan_purpose- 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- 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- 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_ethnicity- 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 application*
4. *Not applicable*
5. *No co-applicant*

applicant_race - Race of the applicant; available values are:

1. *American Indian or Alaska Native*
2. *Asian*

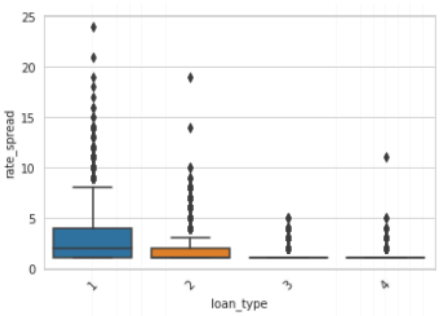
3. *Black or African American*
4. *Native Hawaiian or other Pacific Islander*
5. *White*
6. *Information not provided by applicant in mail, Internet, or telephone application*
7. *Not applicable*
8. *No co-applicant*

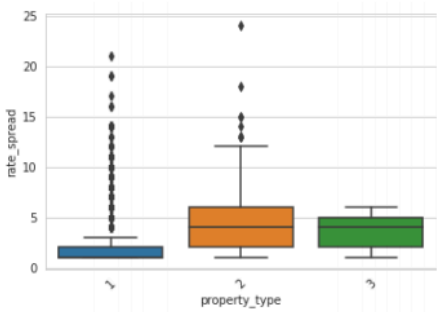
applicant_sex - Sex of the applicant; available values are:

1. *Male*
2. *Female*
3. *Information not provided by applicant in mail, Internet, or telephone application*
4. *Not applicable (4 and 5)*

co_applicant - Indicates whether there is a co-applicant (often a spouse) or no

Fig. 2.1, 2.2, 2.3





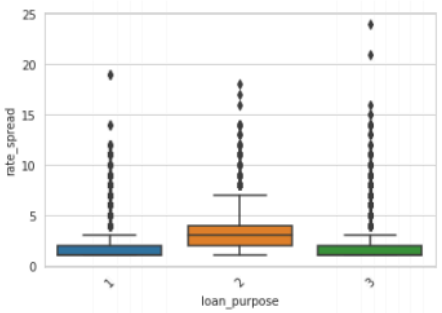
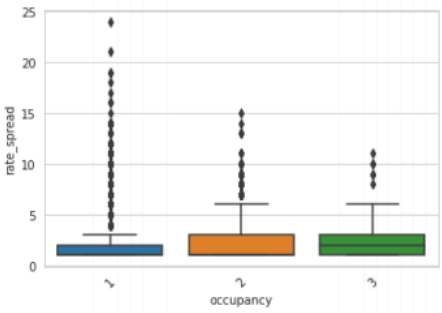
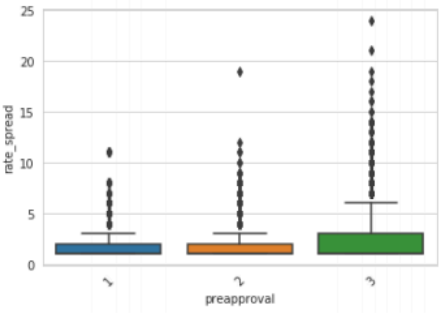


Fig. 2.4, 2.5, 2.6





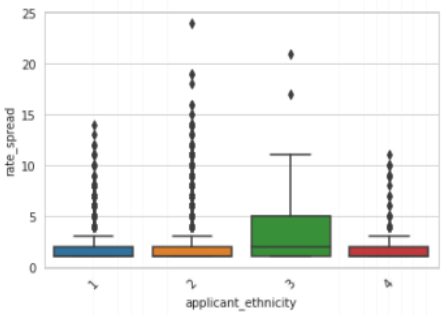
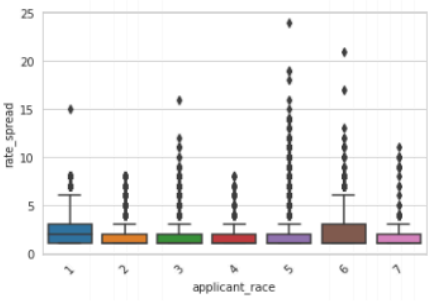
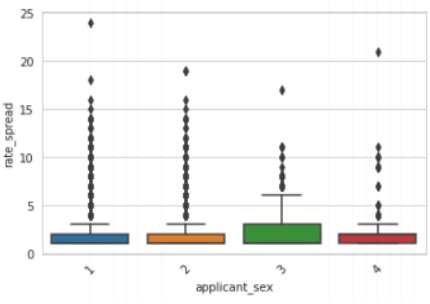
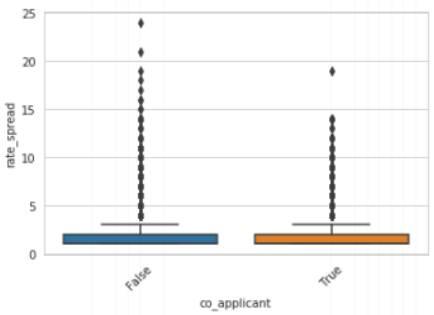


Fig. 2.7, 2.8, 2.9







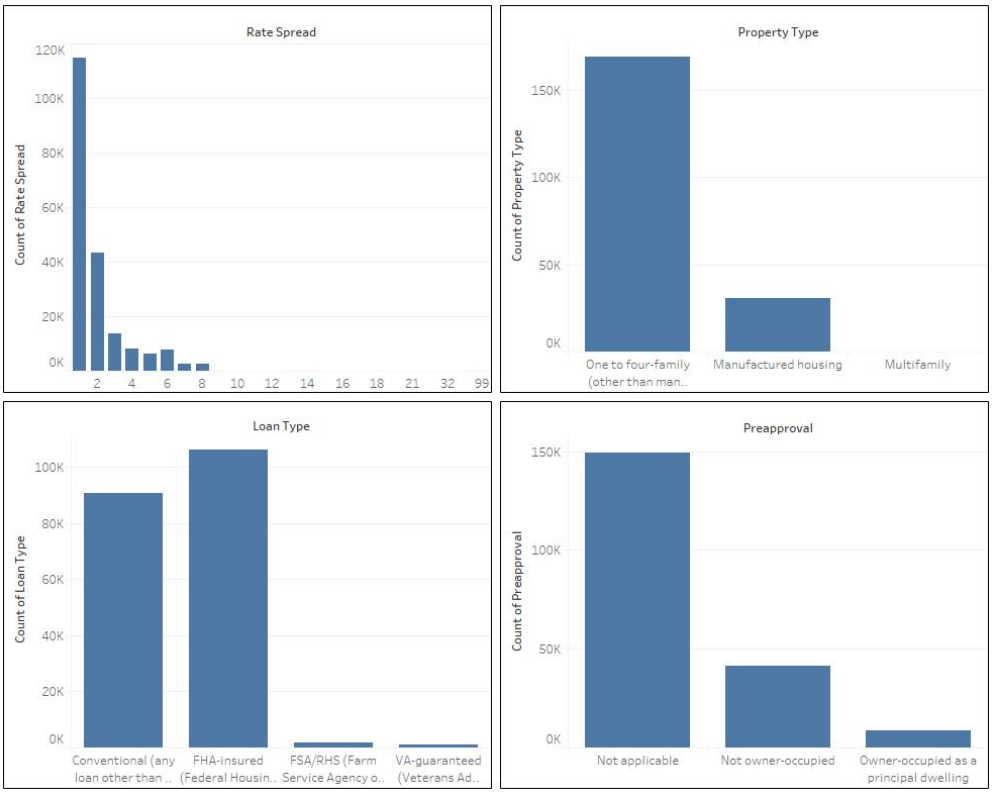
The boxplots presented an interesting set of results that contributed to the feature selection. Although the results were subtle, we can see the distinct trends with loan type, property type and loan purpose - these were key predictors in the model. Occupancy and preapproval appeared

initially to provide limited power, but with the use of trial and error it was discovered that the addition of pre-approval improved the overall model performance.

Applicant ethnicity was an interesting result as higher rates spreads were observed with applicants that didn't respond. This was used successfully in the current model but caution may be required with later iterations. Applicant race and applicant sex displayed minimal variance, although when added to the model, slight improvement was noted. It may be advantageous to exclude these variables for ethical reasons.

After examining the relationship between rate spread and the categorical variables in the data set, frequency counts were performed on a number of the variables to gain a broader understanding of the observations. These were noted and used for multi-faceted analysis later in the process. At this level of detail, the distribution of rate spread became very apparent with a defined right skew. Approximately 75% of the values are less than or equal to 2 with a very small number of values greater than 8. Initially, it was considered that the value of 99 may be an outlier but with closer inspection it was decided that this may be a legitimate value and should be left in the model.

Fig 3.1, 3.2, 3.3



The next step of the analysis was to investigate any relationship between the numeric variables. A collection of scatter matrixes was created, firstly using the individual observation data (*fig. 4.1*). From the plots below, trends were identified between rate spread and applicant income/loan_amount.
This style of analysis was then extended to the census data to identify any trends based on location (*fig. 4.2*). The only variable that required further analysis was ffiec_median_family_income.

Fig. 4.1

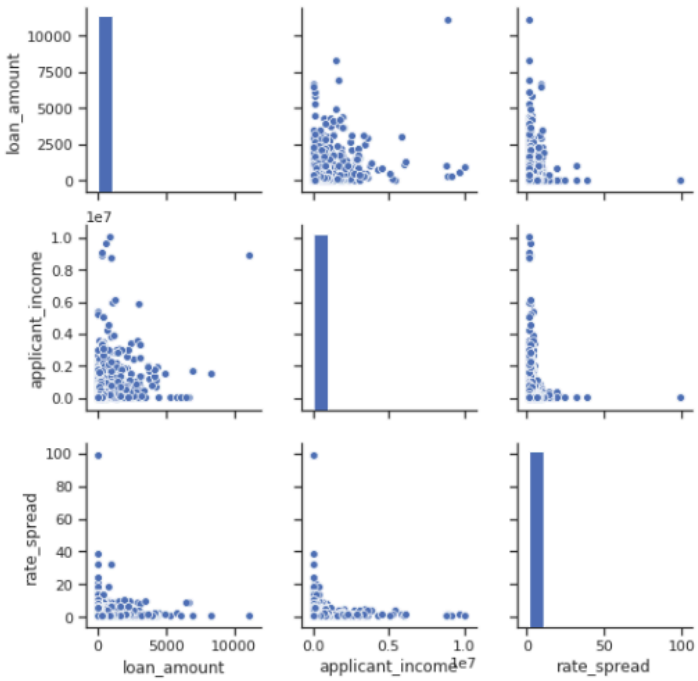
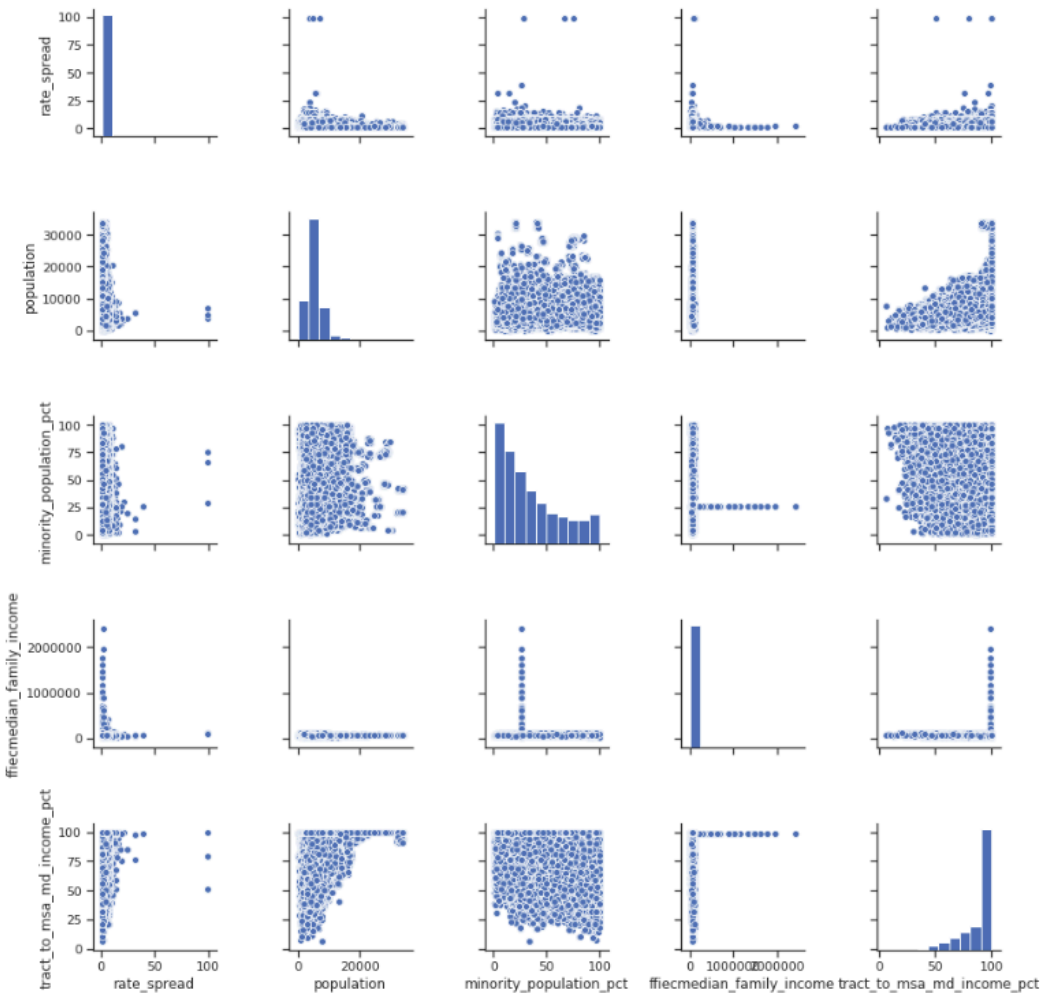

loan_amount
applicant_income
rate_spread
1e7
1e7

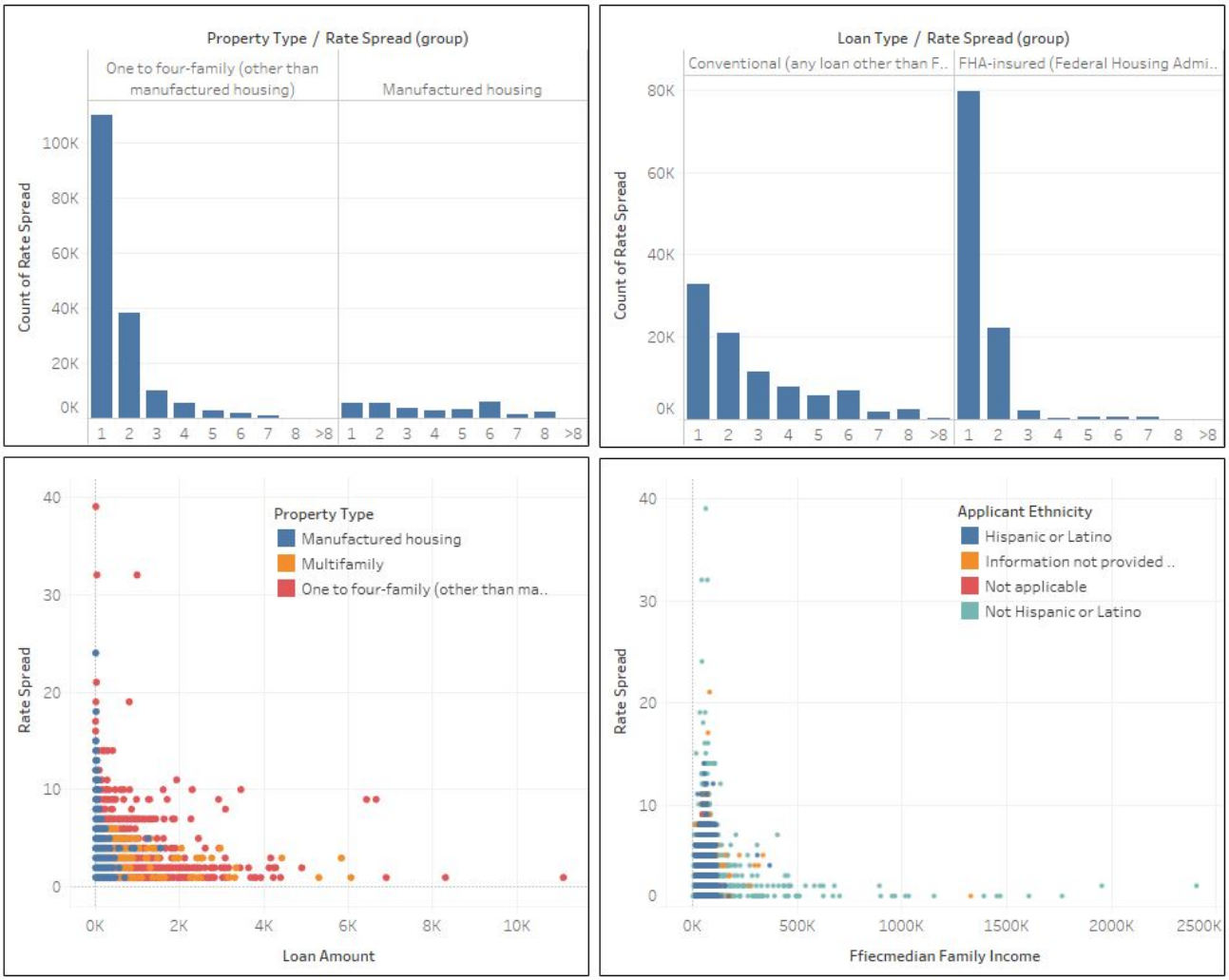
Fig 4.2


rate_spread
population
minority_population_pct
ffiecmedian_family_income
tract_to_msa_md_income_pct
1e6

It was noted that both applicant income and ffiecmedian_family_income followed a similar relationship to rate_spread with a curved trend. As part of the exploration, the rate spread was transformed to a logarithmic scale to try and define a more linear relationship with the income variables. Unfortunately, this didn't increase the model performance so it was removed for simplicity. Again, trial and error was applied on both income variables and surprisingly ffiecmedian_family_income was a more effective overall predictor for the model.

The final stages of the analysis involved creating multi-faceted visualisation. To help display the results, a number of the outlier values were either grouped or removed from the analysis.

Fig. 5.1, 5.2, 5.3, 5.4



From the graphs, clear relationships were confirmed. Manufactured housing is far more likely to have an even distribution of rate spreads, whereas one-to-four family homes are more likely to score a low rate spread value. Unsurprisingly, conventional loans also displayed a more even distribution for rate spread compared to loans that were insured. A higher loan amount seemed to contribute to a lower rate spread with one-to-four and family homes accounting for the highest loan values. Higher ffiec median family income also seemed to contribute to lower rate spreads. The applicant's ethnicity seemed to follow the general trend but it was noted that Hispanic or Latino ethnicity seemed to contribute to a lower median salary.

Regression

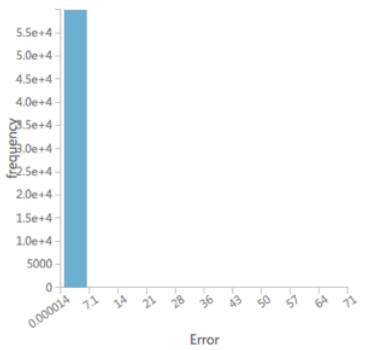
After completing the data exploration to identify predictive features, a regression model was selected to predict the float values for rate spread. After testing a number of algorithms, the Boosted Decision Tree Regression was selected. 70% of the data was used for training the model, with the remaining 30% used for scoring. The following metrics were provided to evaluate the model performance.

Fig. 6.1

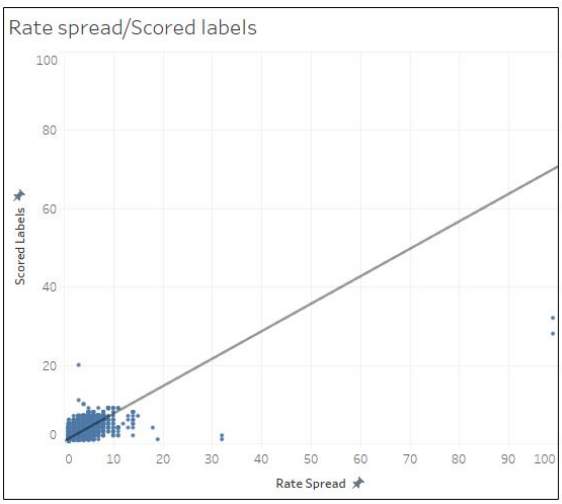
Metrics

Mean Absolute Error	0.556423
Root Mean Squared Error	0.915892
Relative Absolute Error	0.494736
Relative Squared Error	0.287594
Coefficient of Determination	0.712406

Error Histogram



After observing the output metrics, the scored labels were rounded to integer numbers and plotted against the rate spread. A trend line was added to the analysis to highlight the relationship between rate spread and the predicted value.



Conclusion

Overall, rate spreads can be predicted with a satisfactory level of accuracy based on the metrics provided by the model. Observing the results in the graph above, the model has under-predicted the larger values in the dataset. The predictions provided by the model can be used as an indication of the rate spread value that an individual is likely to receive. Further analysis and testing on the model should be undertaken before deploying this in business critical use.