Analysis of Rate Spread Values for Predicting Mortgage Rates

Injung Ahn, November 2019

Executive Summary

The following report provides an analysis of home loans mortgage rate spreads. The rate spread is defined 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 data set analyzed had 200,000 entries of loans. Each loan had 21 variables and a rate spread value.

After exploring the data by observing the descriptive statistics, visualizations and summary of the data, a number of relationships between the features of the loans and rate spread were discovered. Once the observations were made, a regression model was created to predict the rate spread of a home loan from selected features of the loan data set.

The following conclusions have been made after analysis of the data set.

The most significant features that helped predict the rate spread of the loan were:

- **Property Type** - for example, “one to four-family” and “manufactured housing” property types the rate spread generally ranges from 1.0 - 6.0 and 1.0 - 8.0 respectively
- **Loan Type** - whether the loan was granted, applied for, or purchased was conventional, government-guaranteed, or government-insured. Most of the loans fall in the first two categories Conventional and FHA-insured. The vast majority of the FHA-insured loans have a rate spread value of 1.0 or 2.0.
- **Occupancy** - whether the property to which the loan application relates will be the owner’s principal dwelling. The majority of the loans were for houses that were occupied by the owners.
- **Applicant Ethnicity** - ethnicity of the loan applicant. Majority of the applicants fell under the category not Hispanic or Latino.
- **Pre-approval** - indicates whether the application or loan involved a request for a pre-approval of a home purchase loan. Most of the loans fall under “not requested” or “not applicable” and the majority of the “not requested” have a rate spread of range 1.0 - 3.0.

- **Loan Amount** - as the loan amount lowers, the rate spread lowers begins to lower on average.

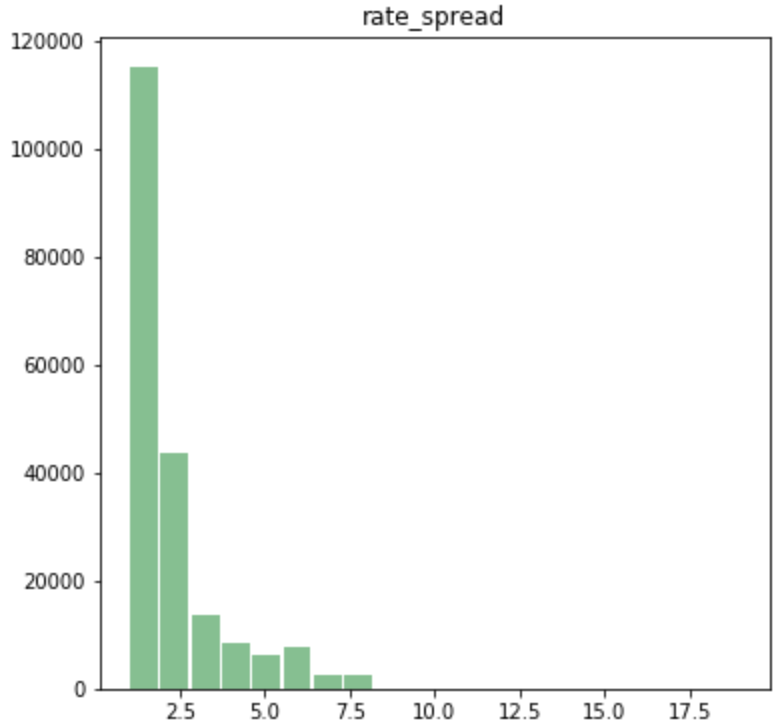
Initial Data Exploration

Individual Feature Statistics

Summary statistics for minimum, maximum, mean, median, and standard deviation were calculated for numeric columns, and the results taken from 200,000 observations are shown here:

Column	Min	Max	Mean	Median	Std Dev
loan_amount	1	11104	142.575	116	142.559
preapproval	1	3	2.703	3	0.546
msa_md	0	408	226.975	261	106.655
applicant_ethnicity	1	4	1.915	2	0.513
applicant_race	1	7	4.763	5	0.887
applicant_sex	1	4	1.418	1	0.577
applicant_income	1	10042	73.618	56	105.697
population	7	34126	5391.099	4959	2669.029
minority_population_pct	0.326	100	34.239	25.996	27.931
ffiecmedian_family_income	17860	125095	64595.356	63485	12724.514
tract_to_msa_md_income_pct	6.193	100	89.283	98.959	15.059
number_of_owner-occupied_units	3	8747	1402.872	1304	706.880
number_of_1_to_4_family_units	6	13615	1927.337	1799	886.577
rate_spread	1	99	1.979	1	1.657

Since rate spread is the feature being predicted, it's important to note that the mean, min and median are quite close to each other. As shown in the following *rate_spread* histogram the values of rate spread are **right skewed**, meaning most loans have a lower rate spread and are generally within a range of 1.0 - 8.0.

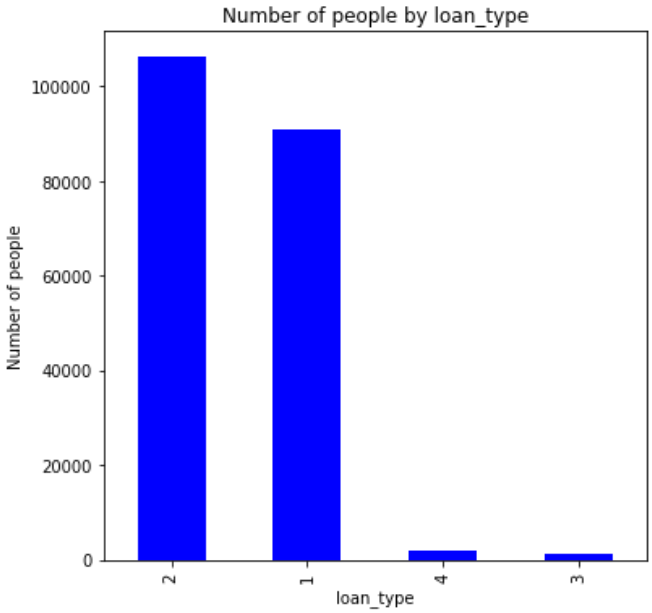


Categorical Features:

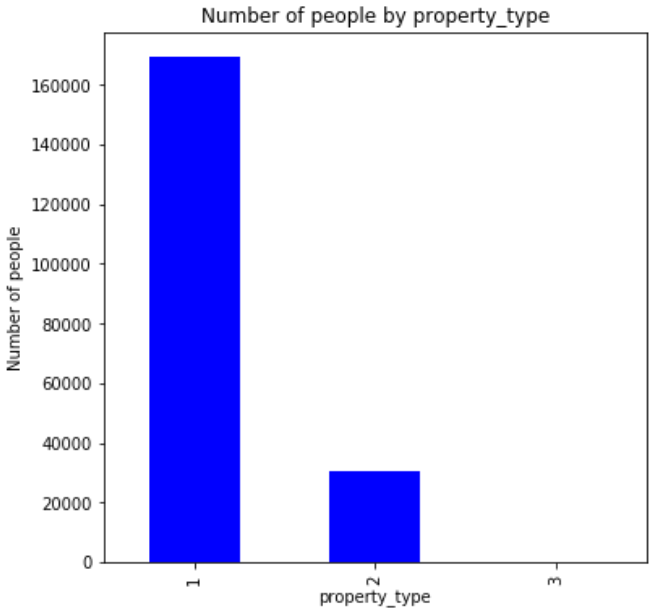
- **Loan type** - conventional, FHA-insured, VA-guaranteed, FSA/RHS
- **Property Type** - one to four-family, manufactured housing, or multifamily
- **Loan Purpose** - Home purchase, home improvement, or refinancing
- **Occupancy** - owner-occupied, not owner-occupied, or NA
- **Pre-approval** - requested, not requested, or NA
- **Applicant Ethnicity** - Hispanic or Latino, not Hispanic or Latino, not provided, NA, no co-applicant
- **Applicant Race** - American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, not provided, NA, no co-applicant
- **Applicant Sex** - male, female, not provided or NA
- **Co Applicant** - whether or not there is a co-applicant

Bar charts were created and analyzed to show frequency of these features, and resulted in the following findings. The most significant bar chart findings are included below.

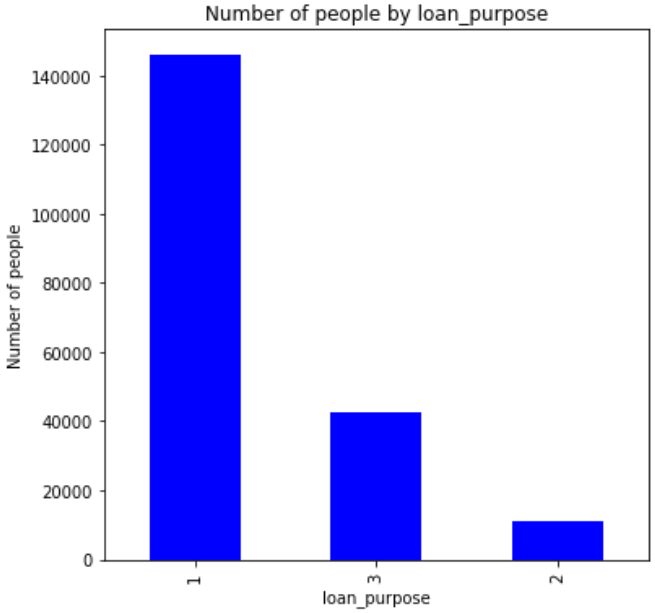
The majority of loans are Conventional (loan_type 1) and FHA-insured (loan_type 2):



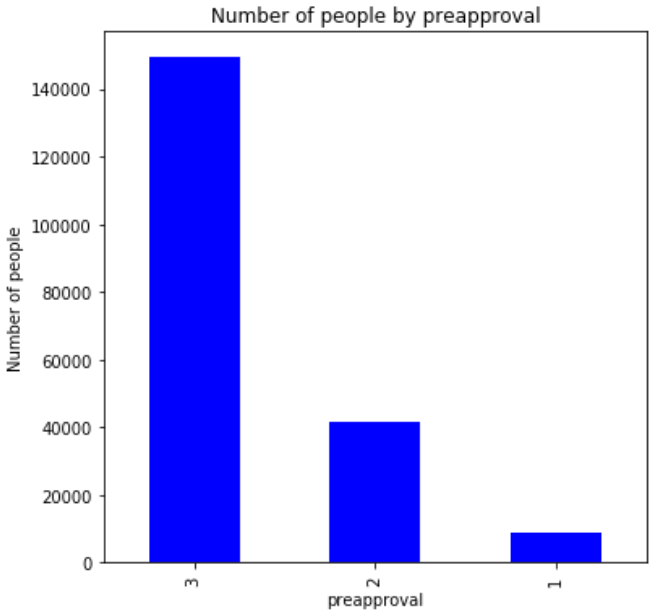
The most common property type is one- to four-family (property_type 1):



The most common loan purpose is Conventional (loan_purpose 1):



The majority of pre-approvals are NA, meaning the pre-approval status was not recorded (preapproval_type 3):



Additional bar chart findings include:

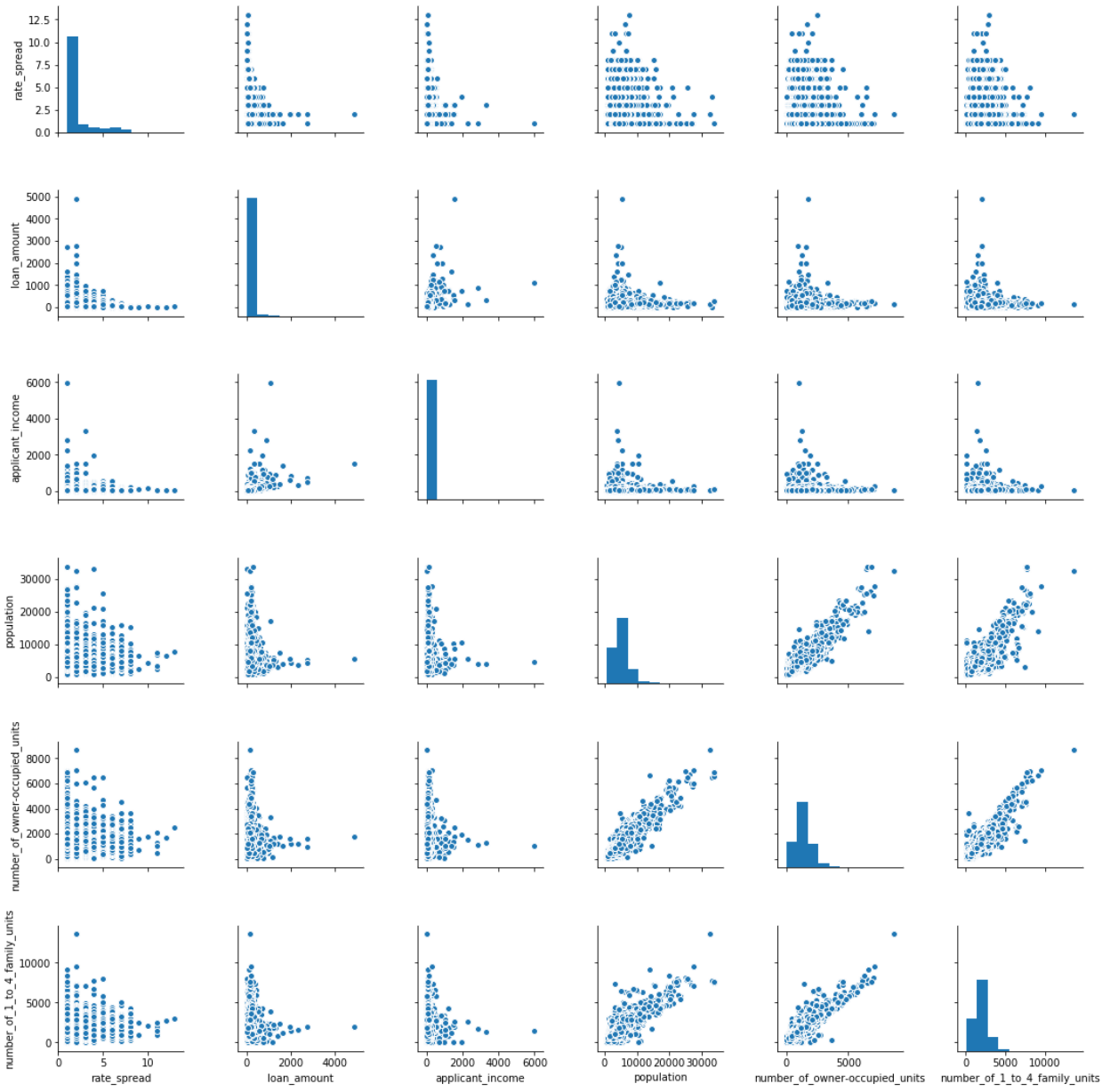
- The vast majority of the homes are occupied by their owners
- The most common ethnicity is non Hispanic or Latino at 74% of the loan applicants
- The most common race is white at 79%
- The percentage of males is almost double females
- The ratio of non co_applicants to co_applicants is 1:1.6

Data Cleaning

Upon reviewing the rate spread feature statistics, the average rate spread is 1.979 while the max rate spread value is 99. This shows that there are clearly outliers within the rate spread feature that could throw off the model. Of the 200,000 observations only 217 had a rate spread over 8.0. Disregarding these values greatly improved the accuracy of the model.

Correlation and Apparent Relationships

A scatter-plot matrix was created to view the relationships between numerical data.



To better understand the data a correlation coefficient table was calculated to get a numerical understanding of the relationships of the numerical data.

Pearson's Correlation Coefficient Table

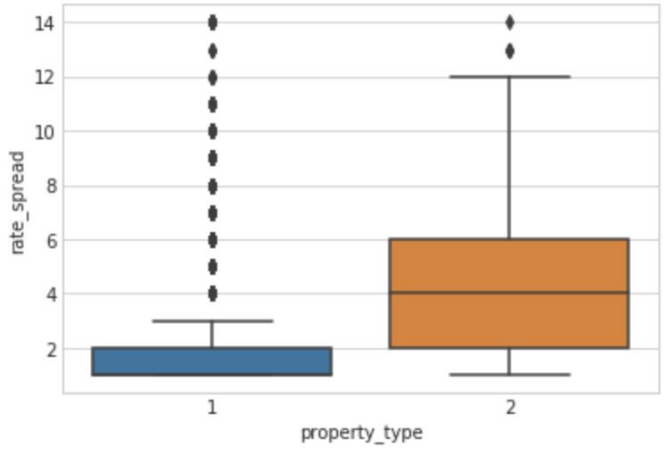
	rate_spread	loan_amount	applicant_income	population	ffiecmedian_family_income	tract_to_msa_md_income_pct	number_of_owner-occupied_units	number_of_1_to_4_family_units
rate_spread	1.0000	-0.2553	-0.0191	-0.0327	-0.0944	0.0122	0.0085	0.0256
loan_amount	-0.2553	1.0000	0.4500	0.0703	0.2451	0.1144	0.0284	-0.0249
applicant_income	-0.0191	0.4500	1.0000	0.0139	0.0886	0.0875	0.0229	0.0009
population	-0.0327	0.0703	0.0139	1.0000	0.0279	0.1558	0.8559	0.8371
ffiecmedian_family_income	-0.0944	0.2451	0.0886	0.0279	1.0000	-0.1323	0.0081	-0.1068
tract_to_msa_md_income_pct	0.0122	0.1144	0.0875	0.1558	-0.1323	1.0000	0.3689	0.2287
number_of_owner-occupied_units	0.0085	0.0284	0.0229	0.8559	0.0081	0.3689	1.0000	0.9052
number_of_1_to_4_family_units	0.0256	-0.0249	0.0009	0.8371	-0.1068	0.2287	0.9052	1.0000

From the table the only notable correlation to rate spread was loan amount having a coefficient of -0.25. The number of owner occupied units and number of 1 - 4 family units were closely correlated with population. Also loan amount and applicant income showed a decent correlation at 0.45. Most of the numerical variables had little to no correlation with rate spread so they were removed from the model fitting process.

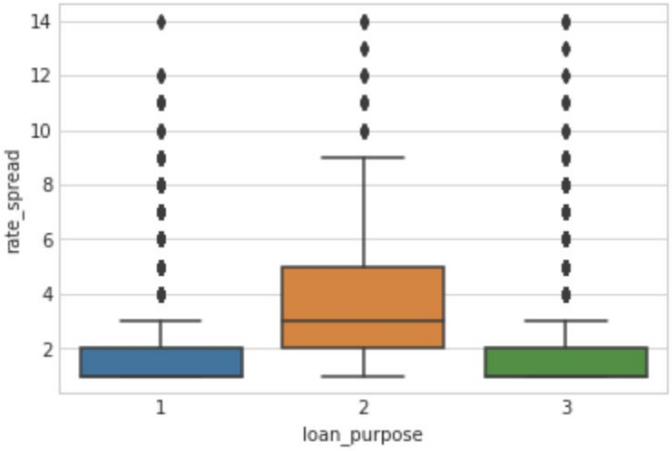
Categorical Relationships

Boxplots were made to help further visualize the categorical data.

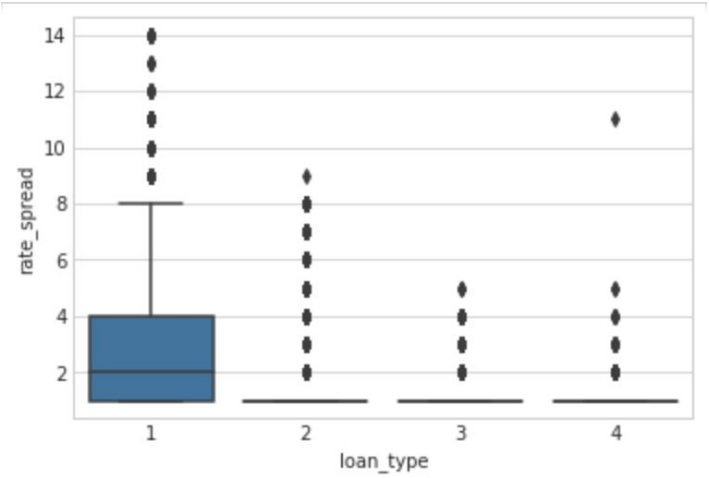
For property type, manufactured housing (2) has a higher rate spread on average than the other types:



In loan purpose, home improvement (2) has a higher rate spread than the others



- In loan type, the Conventional (1) loan has the highest rate spread average among the other types



Feature Selection

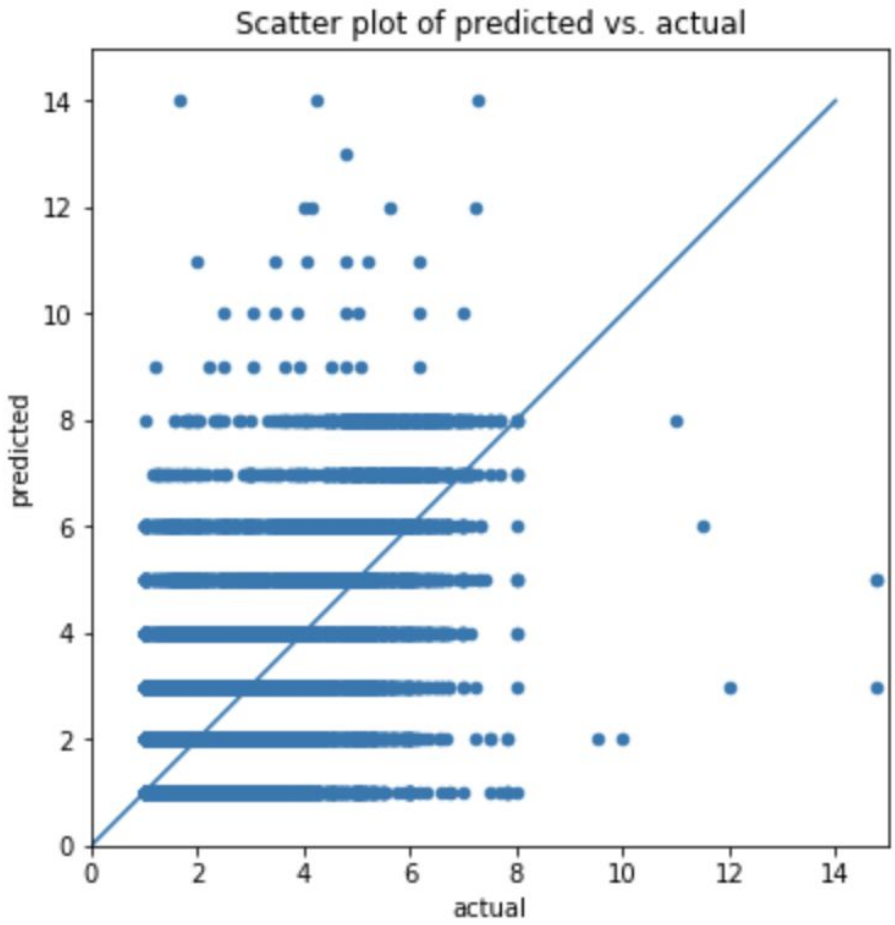
After viewing the data, the categorical features were transformed into numerical values to be included in the model. Then the Select K best features were used to help grade and reduce any less significant features. Below is a chart of the values of the top 10 scoring features from the Select K Best algorithm.

Feature Category	Specific Feature	Score
Property Type	Manufactured Housing	48751.931
Loan Type	Home Purchase	25113.245
Loan Type	Home Improvement	19560.684
Occupancy	Not Owner-occupied	13033.005
Property Type	One to four-family	9277.911
Applicant Ethnicity	Hispanic or Latino	8291.39
Preapproval	Was not requested	6749.913
Loan Amount	(Numerical feature)	4532.358
Preapproval	Not Applicable	2555.562
Occupancy	Owner occupied	2136.295

Regression

After the analysis of the different features, a regression model was made to predict the rate spread of different loans. Using the knowledge of relationships from the analysis a decision tree regression model was created. This model was used because most of the variables used to train the model were categorical.

After removing outliers from the data, the amount of data kept was 95.8%. The model used 79.1% of that data to train and the other 20.9% to test. Below is a scatterplot of the predicted values vs the actual values.



As seen in the chart some of the predicted and actual guesses are correct, but there is a wide margin of error. The Root Mean Square Error was calculated to 1.102 and the R^2 value calculated was 0.541.

Conclusion

This model produced a model R^2 value of 0.541. Property Type, Loan Type, Occupancy, Applicant Ethnicity, Pre-approval and Loan Amount were the characteristics having the greatest impact in predicting the rate spread value in this model.

The majority of the observations had a rate spread value of 1.0 to 2.0 the predictions were skewed in that direction.

Separating the data into two categories of having a rate spread of 1.0 - 2.0 and the other category of rate spreads of greater than 2 could possibly improve the accuracy further by preventing the model from skewing towards the direction of the lower category.