DAT102x: Microsoft Professional Capstone - Data Science:

Predicting Mortgage Rates from Government Data

Jonathan Yeh, November 2019

Executive Summary

This report aims to present an analysis of data to convey the insights gained from dataset provided by Federal Financial Institutions Examination Council's (FFIEC).

With initial data exploration, we will have visualized images and statistics summary of these data. Following up with the skills we've learnt from MPP data science course enrolled, we shall then take Microsoft Algorithm Cheat Sheet into consideration. By leveraging resource of Azure Studio, we can further create a machine learning model to predict the rate spread for loan applications across the United States.

The goal is to consider how demographics, location, property type, lender, and other factors are related to the mortgage rate offered to applicants.

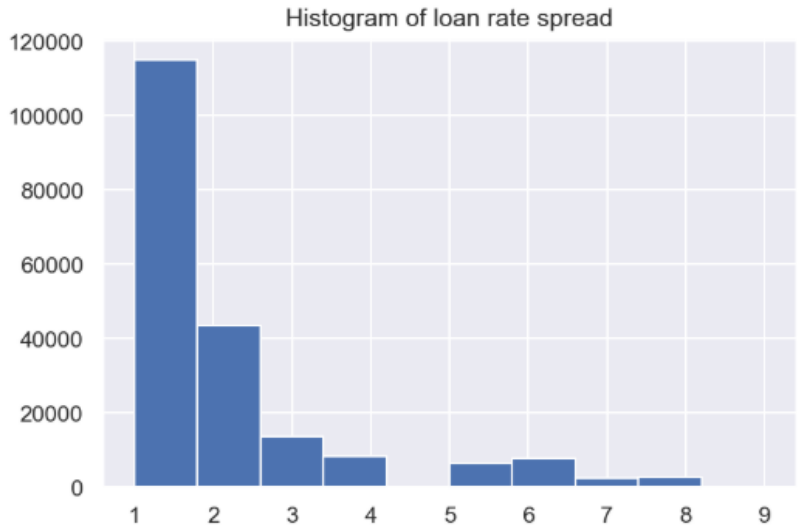
Initial Data Exploration

The given train inputs with joined train labels consist of 200000 rows and 23 columns. Initial exploration of the data begins with summary statistics of train values for minimum, maximum, mean, median, standard deviation shown as below:

Column	Mean	Median	Min	Max	Std Dev
row_id	99999.5	99999.5	0	199999	57735.1713
loan_type	1.5709	2	1	4	0.5594
property_type	1.1549	1	1	3	0.3651
loan_purpose	1.4826	1	1	3	0.8222
occupancy	1.0614	1	1	3	0.246
loan_amount	142.5749	116	1	11104	142.5595
preapproval	2.7029	3	1	3	0.5457
msa_md	226.975	261	0	408	106.6553
state_code	28.202	30	-1	52	15.5934
county_code	166.3352	181	0	316	92.8525
applicant_ethnicity	1.9153	2	1	4	0.5133
applicant_race	4.7627	5	1	7	0.8873

applicant_sex	1.4175	1	1	4	0.5771
applicant_income	73.6179	56	1	10042	105.6969
population	5391.0991	4959	7	34126	2669.0288
minority_population_pct	34.2386	25.996	0.326	100	27.9309
ffiecmedian_family_income	64595.3558	63485	17860	125095	12724.5145
tract_to_msa_md_income_pct	89.283	98.959	6.193	100	15.0592
number_of_owner-occupied_units	1402.8724	1304	3	8747	706.8804
number_of_1_to_4_family_units	1927.3366	1799	6	13615	886.5766
lender	2001.3115	1834	0	4283	1271.1342
co_applicant	Unique Values, true or false				
rate_spread	1.9791	1	1	99	1.6568

Note the parameter ‘rate_spread’ is our interest in this analysis. By visualizing it, we can come out the histogram of the of loan rate spread are right-skewed. In other words, most loan spreads are at the lower end of the loan rate spread range, the amount of loan rate spread outside of the main range from 1 to 8 as shown below:



In addition to the numerical values, the data include categorical features. These are:

- msa_md - A categorical with no ordering indicating Metropolitan Statistical Area/Metropolitan Division where a value of -1 indicates a missing value.
- state_code - A categorical with no ordering indicating the U.S. state where a value of -1 indicates a missing value.
- county_code - A categorical with no ordering indicating the county where a value of -1 indicates a missing value.
- lender - A categorical with no ordering indicating which of the lenders was the

authority in approving or denying this loan.

- loan_type - Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured; available values are: Conventional, FHA-insured, VA-guaranteed and FSA/RHS.
- 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: One to four-family, Manufactured housing and Multifamily.
- loan_purpose - Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing; available values are: Home purchase, Home improvement and Refinancing.
- occupancy - Indicates whether the property to which the loan application relates will be the owner's principal dwelling; available values are: Owner-occupied as a principal dwelling, Not owner-occupied, and Not applicable.
- preapproval - Indicate whether the application or loan involved a request for a pre-approval of a home purchase loan; available values are: Preapproval was requested, Preapproval was not requested, and Not applicable.
- applicant_ethnicity - Ethnicity of the applicant; available values are Hispanic or Latino, Not Hispanic or Latino, Information not provided by applicant in mail, Internet, or telephone application, Not applicable and No co-applicant.
- applicant_race - Race of the applicant; available values are American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Information not provided by applicant in mail, Internet, or telephone application, Not applicable and No co-applicant.
- applicant_sex - Sex of the applicant; available values are: Male, Female, Information not provided by applicant in mail, Internet, or telephone application, and Not applicable.

Data cleaning

Before starting to build a prediction model, we should be careful if there are any missing values existed. Checking out the data in detail, we are able to list these columns including missing values:

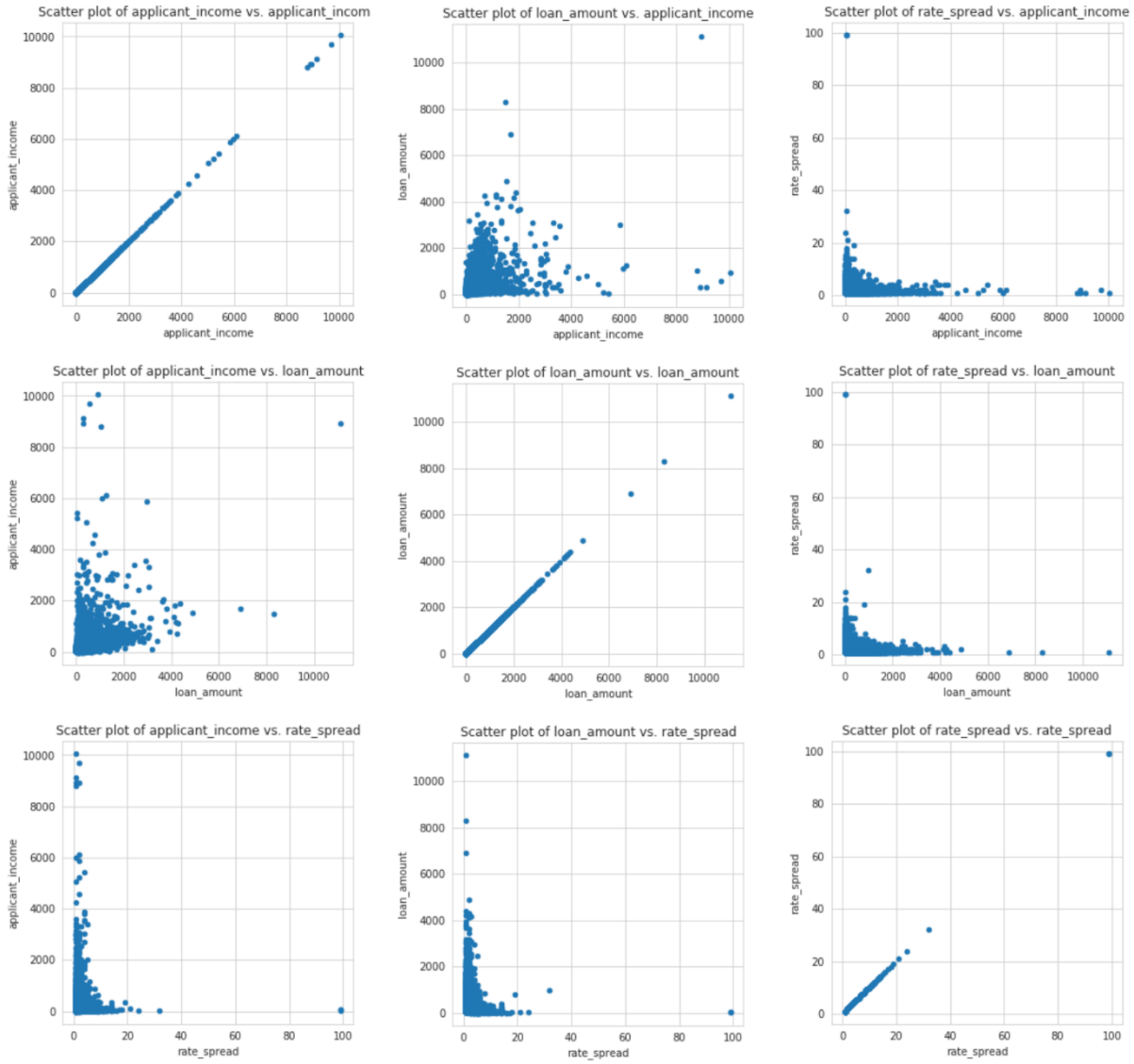
- applicant_income
- population
- minority_population_pct
- ffiecmedian_family_income
- tract_to_msa_md_income_pct
- number_of_owner-occupied_units

- number_of_1_to_4_family_units

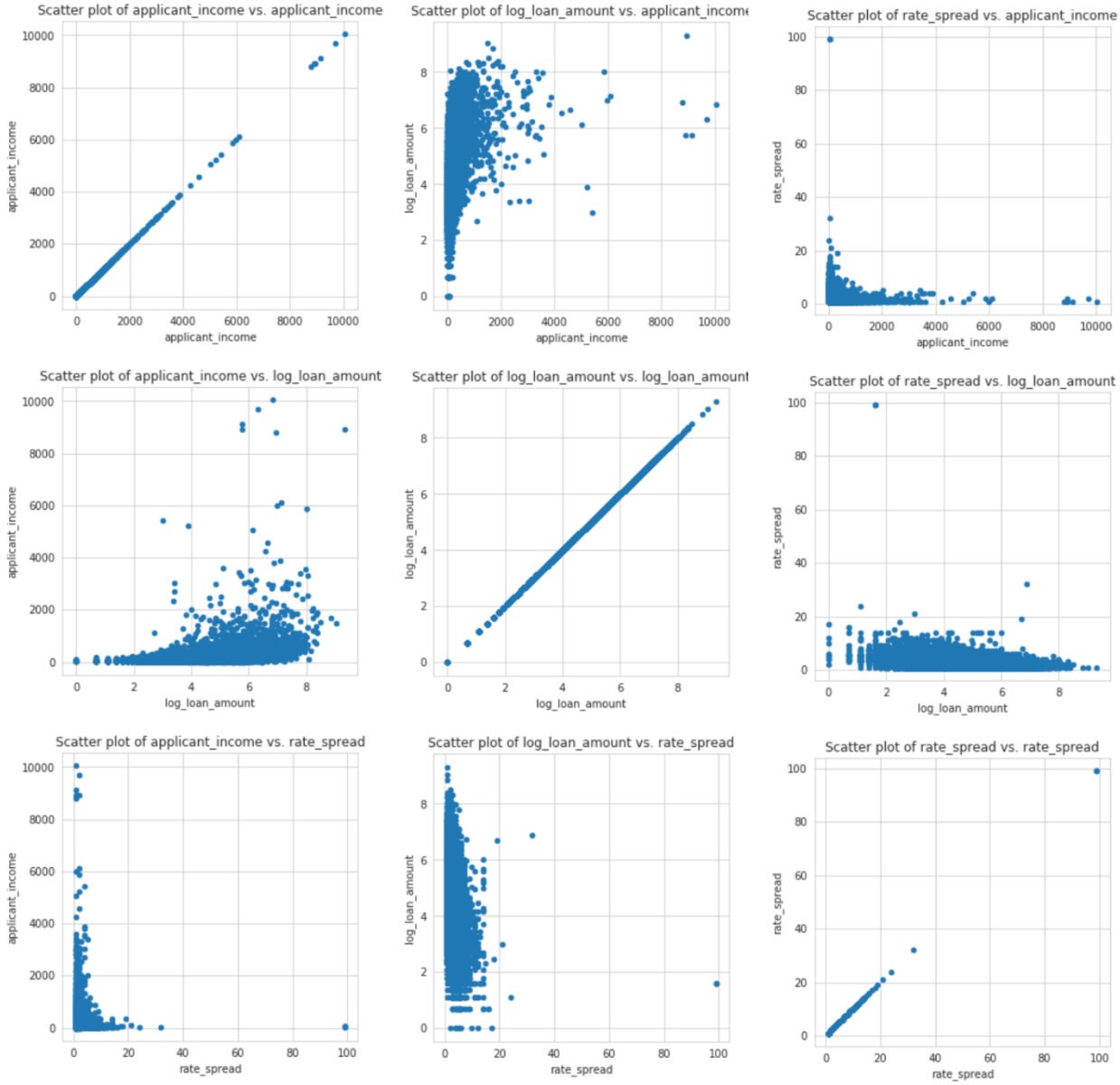
Since there are two types of feature which are numeric and string, we may apply different cleaning data mode respectively, median and mode.

Exploratory Data Analysis

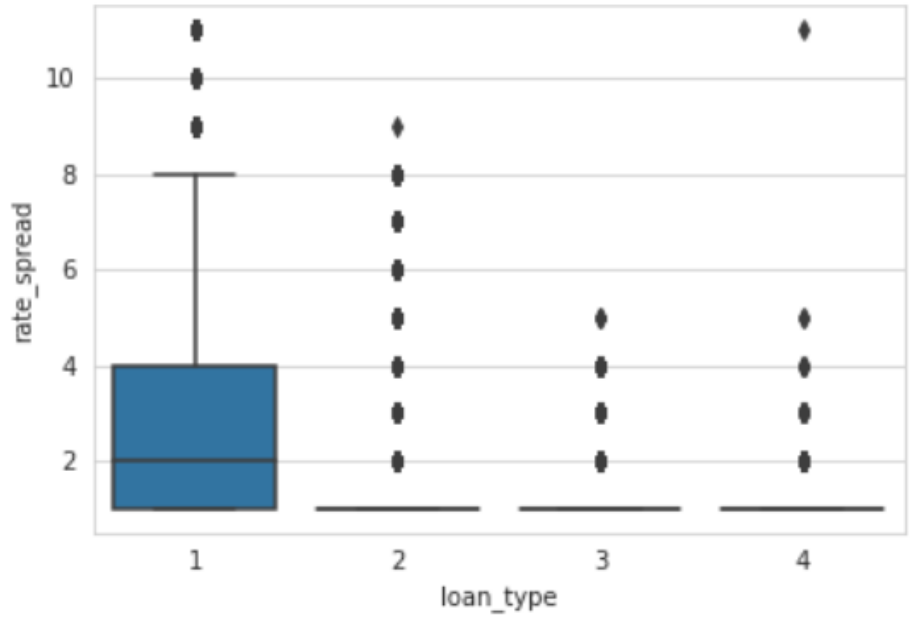
After exploring each of the individual features, an attempt was made to identify relationships between features in the loan data. The following scatter-plot was created to compare numeric features of applicant income, loan amount and so on. The key features in this matrix are shown here is among applicant income, loan amount and rate spread.



An attempt made to improve the fit of the features to rate spread, applicant income and loan amount, the log value for loan amount was calculated. The results show increased linearity in the relationships between loan amount and rate spread.

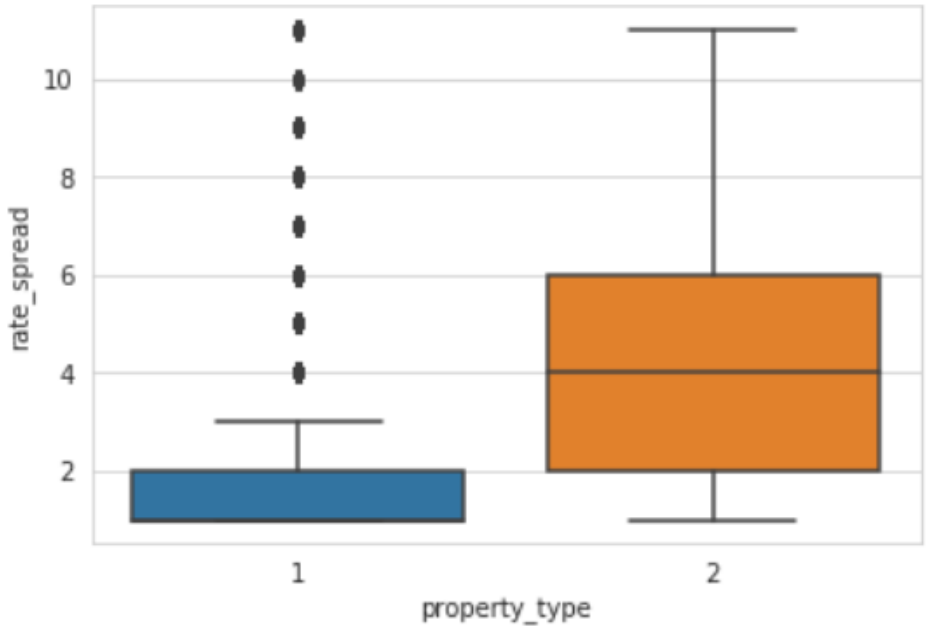


Further, we would look into any apparent relationship between categorical feature values and target variable. The following boxplots show the categorical columns that seem to exist a relationship with the rate spread.



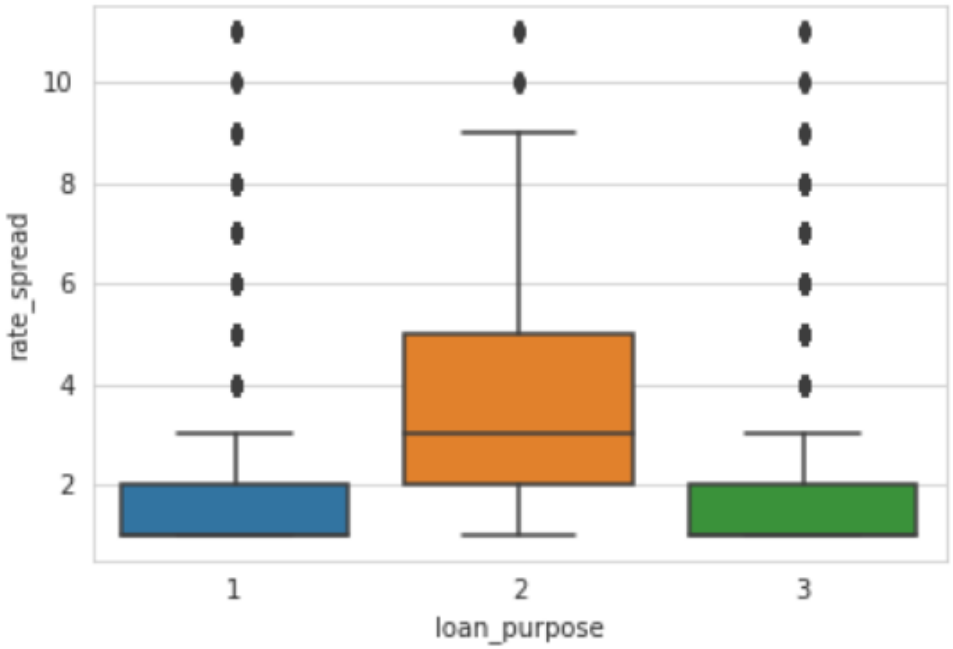
Load types:

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 types:

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



Loan purposes types:

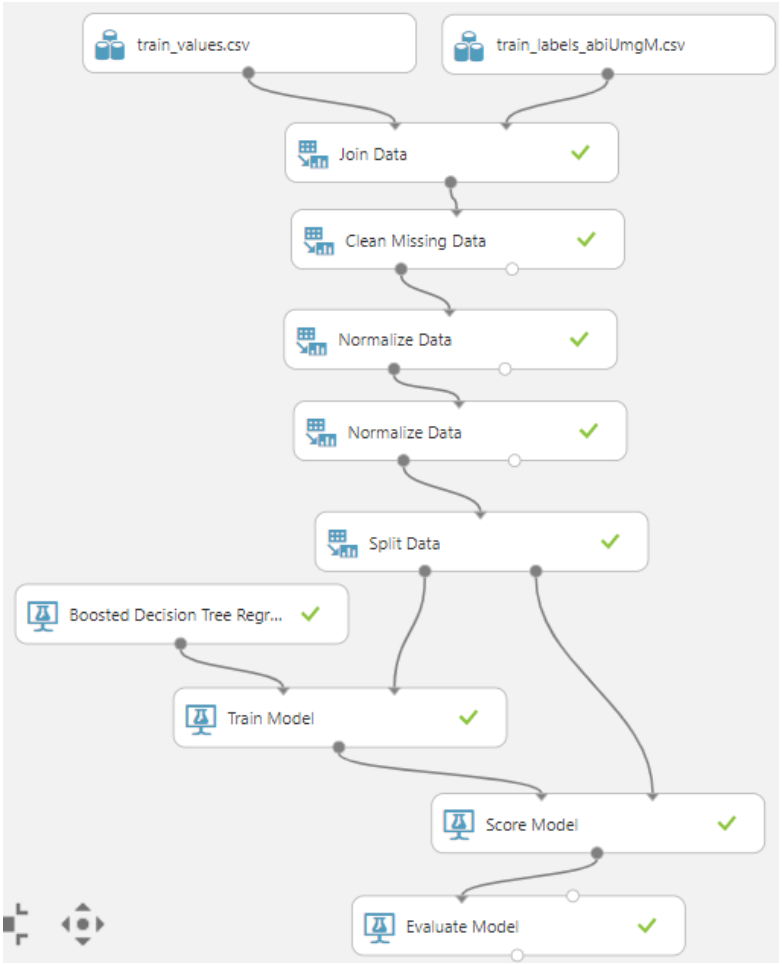
1. Home purchase
2. Home improvement
3. Refinancing

From the data visualization with above box plots, we can have more intuitive information. These info can be addressed below:

- Conventional loans have higher loan spread than any other types of loans. Conventional loan's rate spread is wider range than any other types of loans.
- Manufactured housing has higher loan spread than one to four-family housing loans. They are generally represented to distinguished intervals.
- Home improvement has higher loan spread than home purchase loan and refinancing loan. Home improvement loan spread's median value also much higher than the other two loan purposes. Home purchase loan and refinancing loan have similar rate spread.

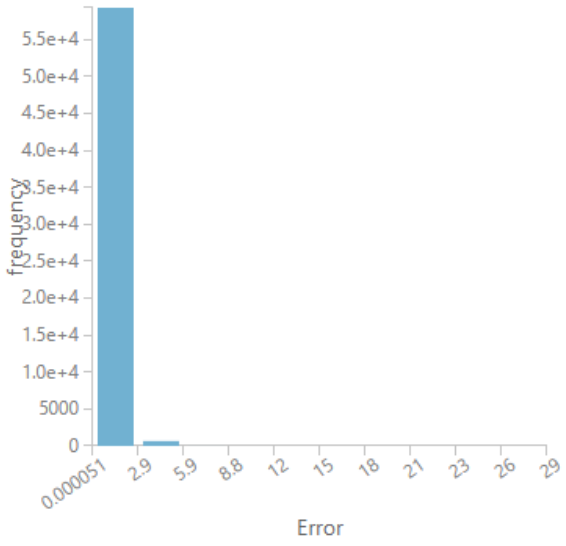
Regression of rate_spread: to implement via Azure machine learning studio

Based on the apparent relationships recognized with data analysis, a linear regression model was created to predict the actual loan rate spread. The model was trained with 70% of the data, and tested with the remaining 30%. The measurement used to gauge the success of the model was the Coefficient of Determination (R2). The test results indicate R^2 value of 0.708033.



- Mean Absolute Error 0.596291
- Root Mean Squared Error 0.86993
- Relative Absolute Error 0.530034
- Relative Squared Error 0.291967
- Coefficient of Determination 0.708033

Error Histogram



Conclusion

According to apparent relationships identified when analyzing the data, a linear regression model was created to predict the value for rate_spread. This analysis has indicated that the desired prediction of rate_spread can be confidently predicted from its characteristics. In particular, loan type, property type, loan purpose, loan amount, state code and lender have significant effects on the loan rate.