Capstone Project - Analysis of rate spread for loan applications

- Microsoft Professional Program in Data Science

Oguzhan San

November 2019

Summary

This report explains the steps in the analysis of data from 200 000 loan applications in the US and the making of a machine learning model to predict the rate spread (difference between the offered and standard mortgage rate for the loan) for other applications. Of 21 given features the following 16 features were found to contribute to the prediction of rate spread: Applicant ethnicity, applicant race, applicant sex, loan purpose, loan type, occupancy, preapproval, property type, co-applicant, lender, metropolitan statistical area, U.S. state code, county code, loan amount and applicant income.

First visualizations and summary statistics of the dataset were created to get a better understanding of the data. Then missing values and outliers were treated, feature selection and engineering performed, and a suitable model was selected. The evaluation measure used was the coefficient of determination (R^2) which in the final model was 0.715.

Table of Contents

Table of Figures

1 Introduction

This report involved the analysis of 200,000 mortgage applications from the United States spanning one year. Each record contained a range of data related to personal and demographic characteristics, along with features of the loan itself.

The analysis sought to understand what factors influence the mortgage rates offered to individuals. In addition, the analysis aimed to calculate the rate spread from a test dataset, and to compare the results with the actual rate spread that resulted from the applications.

Data exploration was performed using Python 3 within a Jupyter Notebook to help understand and summaries the data. Next, Python was used to identify features and relationships within the data. The results of the exploration helped to inform the subsequent data cleaning and modeling steps.

A predictive rate spread machine learning model was built by using training data and labels. After the model was iteratively refined, a test dataset was processed. Results from the test dataset were submitted to an independent website for scoring the accuracy of algorithmic predictions.

Several findings resulted from this analysis, including:

- **The largest influence on the rate spread was the lender utilized.** This is a significant concern because the product that lenders offer, mortgage approval, is a standardized commodity, and should not be a significant influence on the rate spread. From a business perspective, mortgage rate spreads would be expected to vary according to risk variables associated with the loan, and not the lender making the loan.
- **Four of the leading five variables influencing rate spread make business sense.** As stated in the previous finding, features affecting the risk of the loan would be expected to affect the rate spread. This was the evident in our analysis where the loan amount, property type, loan type, and property purpose were ranked two through five in order of influence over the rate spread.
- **Important individual and demographic features had negligible influence on the rate spread.** An applicant's ethnicity, race, sex, and the minority population percentage making up their neighborhood of residence had little influence over the rate spread. This is a positive finding, indicating that credit is offered at unbiased interest rates to applicants irrespective of these features. However, the total number of all non-white applicants combined is far less than the total number of white applicants, suggesting that non-white candidates are not applying for mortgages at a level commensurate with their total population proportion.

2 Data Preprocessing

The original dataset was not properly settled, since some data is unknown or not applicable in some mortgage applications and labeled as "-1" or "NA". Imported data needed to be preprocessed.

msa/md: A categorical variable with 409 distinct values. It represents the Metropolitan Division for the mortgage property. There is no missing value for this variable, however, after the importance analysis, msa/md is removed since it has low importance value on feature importance metric.

state code: A categorical variable with 52 distinct values. It represents the U.S. state for the mortgage property. There are 1338 missing values indicated with "-1" for this variable. With the importance analysis, state code feature is removed since the importance metric is 0.07 and less than the threshold 0.1.

county code: A categorical variable with 306 distinct values. It represents the county for the mortgage property. There is no missing value in county code column, but it is removed by the importance analysis because of it has low importance value on feature importance metric.

lender: A categorical variable with 3893 distinct values. It indicates the loan-issuing institution for the mortgage property. There is no missing value for lender.

loan amount: A integer variable with minimum 1, maximum 11104, median 116 and average 142.6. It shows the mortgage size in thousands of dollars. There is no missing value for loan amount.

loan type: A categorical variable with 4 distinct values. It represents the type of the mortgage for the property is either conventional, FHA-insured, VA-guaranteed or FSA/RHS. There is no missing value for loan type.

property type: A categorical variable with 3 distinct values. It indicates the type of property is either 1 to 4 family, manufactured housing or multifamily. There is no missing value for property type feature.

loan purpose: A categorical variable with 3 distinct values. It represents the purpose of mortgage is either home purchase, home improvement or refinancing. There is no missing value for loan purpose variable.

occupancy: A categorical variable with 3 distinct values. It is indicating if the property to which the mortgage application relates will be either owner'-occupied as a principal dwelling, not owner-occupied or not applicable. There is no missing value in column occupancy.

preapproval: A categorical variable with 3 distinct values. It shows the property's mortgage application involved either preapproval was requested, preapproval was not request, or not applicable. There is no missing value for preapproval.

applicant income: A integer variable with minimum 1, maximum 10042, median 56 and average 73.62. It represents the amount of income of loan applicant in thousands of dollars. There are 10708 missing values indicated with "NA" for this variable. Applicant income records with missing income will be eliminated for analysis purpose.

applicant ethnicity: A categorical variable with 4 distinct values. It shows the loan applicant is either Hispanic or Latino, Not Hispanic or Latino, Information not provided by applicant or not applicable. There is no missing value for applicant ethnicity.

applicant race: A categorical variable with 7 distinct values. It shows the loan applicant is either American Indian or Alaska Native, Asian Black or African American, Native Hawaiian or Other Pacific Islander, White, information not provided by

applicant or not applicable. There is no missing value for this variable, however, after the importance analysis, applicant race is removed since its importance metric is 0.01 and less than the threshold 0.1.

applicant sex: A categorical variable with 4 distinct values. It shows the loan applicant is either Male, Female, information not provided by applicant or not applicable. With the importance analysis, applicant sex is eliminated since it has low importance value on feature importance metric

co applicant: A Boolean variable determines if applicant is a co-applicant or not. Co applicant is removed by the importance analysis because of its low importance value on feature importance metric.

population: A integer variable with minimum 7, maximum 34126, median 4959 and average 5391. It represents the amount of total population in applicant's census tract. There are 1995 missing values indicated with "NA" for this variable. After the importance analysis, population is eliminated since its importance metric is 0.02 and less than the threshold 0.1

minority population pct.: A numeric variable in percentage with minimum 0.326, maximum 100.000, median 25.996 and average 34.239. It represents the percentage of minority population to total population in applicant's census tract. There are 1995 missing values indicated with "NA" for this variable. This column is removed by the importance analysis since its importance metric is 0.06 and less than the threshold 0.1.

FFIEC median family income: A integer variable with minimum 17860, maximum 125095, median 63485 and average 64595. It shows the median family income in dollars in applicant's census tract within the msa/md, adjusted by FFIEC. There are 1985 missing values indicated with "NA" for this variable. Records with missing population will not be kept for analysis purpose.

tract to msa/md income pct.: A numeric variable in percentage with minimum 6.193, maximum 100.000, median 98.959 and average 89.283. It indicates the percentage of tract compared to msa/md for the median family income. There are 2023 missing values indicated with "NA" for this variable. This column is eliminated by the importance analysis because of its low importance value on feature importance metric

number of owners occupied units: A integer variable with minimum 3, maximum 8747, median 1304 and average 1403. It represents the number of households lived in by the owner. There are 2012 missing values indicated with "NA" for this variable. This column is removed by the importance analysis since it has low importance value on feature importance metric

number of 1 to 4 family units: A integer variable with minimum 6, maximum 13615, median 1799 and average 1927. It represents the number of households built for 1 to 4 families. There are 2016 missing values indicated with "NA" for this variable. After the importance analysis, this feature is eliminated since its importance metric is 0.003 and less than the threshold 0.1

loan to income: An added numerical hyperparameter dividing loan amount by the corresponding applicant income in the mortgage application. Minimum 0.0037, maximum 462.6667, median 2.1475 and average 2.2669. No missing value since any missing applicant income record has been removed.

Log of loan to income: A numerical hyperparameter taking log of the newly added feature loan to income. Minimum -5.6005, maximum 6.1370, median 0.7643 and average 0.5828. There is not any missing value for log of loan to income.

log of applicant income: A added numerical hyperparameter taking log of applicant income. Minimum 0, maximum 9.215, median 4.025 and average 4.072. There is no missing value for log of applicant income.

log of loan amount: A numerical hyperparameter taking log of loan amount. Minimum 0, maximum 9.315, median 4.754 and average 4.655. There is not any missing value for log of loan amount.

log of FFIEC median family income: A new numerical hyperparameter taking log of the FFIEC median family income. Minimum 9.79, maximum 11.74, median 11.06 and average 11.06. There is no missing value for log of FFIEC median family income.

3 Analysis of Data

3.1 Numeric Relationships

Regarding the numeric variables, the correlation plot shows that there is very high positive correlation between population and number of owner-occupied units, as well as population and number of 1-4 family units. There is also a high positive correlation between the number of 1-4 family units and the number of occupied units. A positive correlation means, that if one variable increases the both others will increase too, as well as if the variable decreases the positive correlated others will also decrease. Because there is such a strong correlation, only one variable is needed because the other both do not add additional information.

The scatter plots of the numeric variables regarding the spread rate shows that the spread rate de- creases with the **loan amoun**t and also decreases with the **applicant's income**. Both variables should be used in the prediction model.

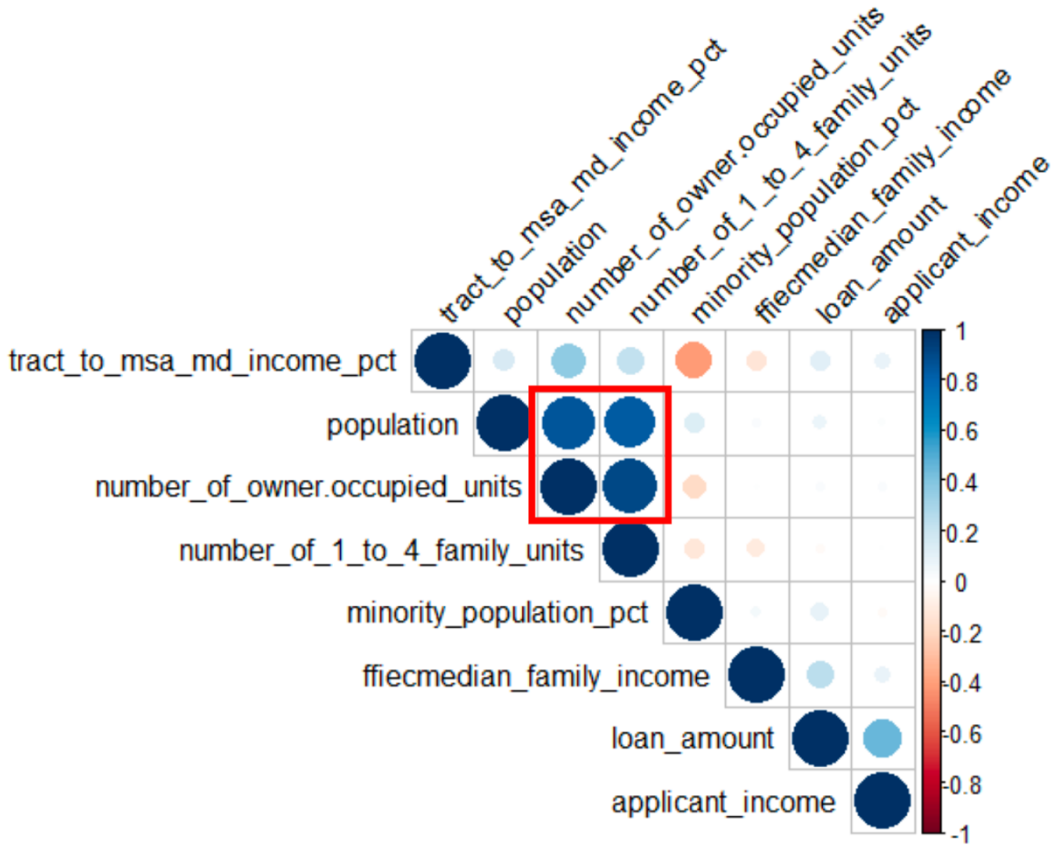

Figure 1 Data features correlations

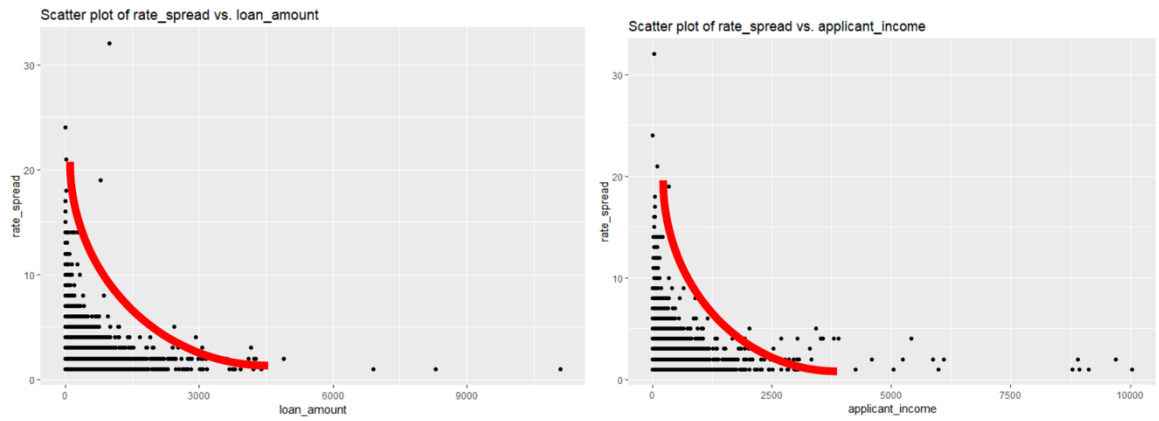

Figure 2 Applicant income and loan amount correlations with rate spread

The **minority population** percentage and median **family income** don't seem to have any impact on the spread rate. For almost every level along the variable, the same spread rates can be found. These numeric variables should **not be used** in the prediction model.

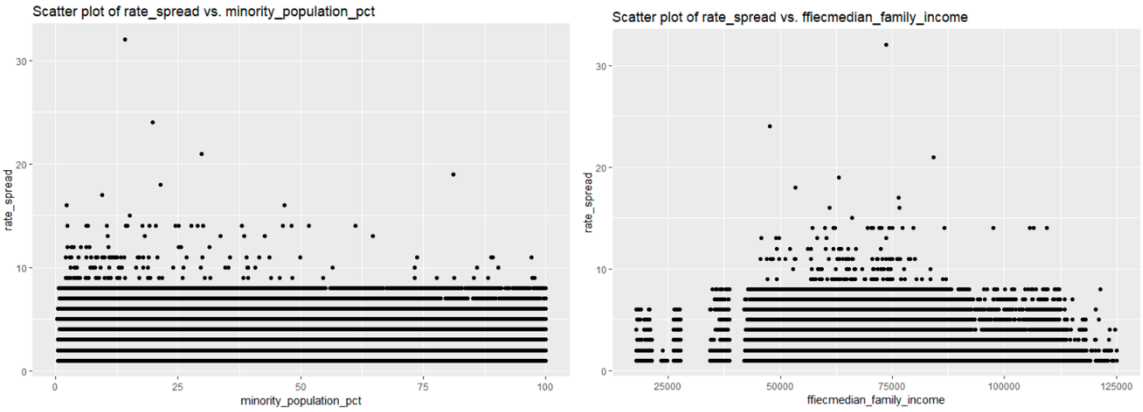

Figure 3 Minority population percentages and median family income distribution for rate spread values

The variable **population** and **tract to msa md income** both have limited relevance for the model. The rate spread decreases from a population level of 15.000. The rate spread increases from an income percent of 0 to 45 percent. Both variables can be **used in the prediction model.**

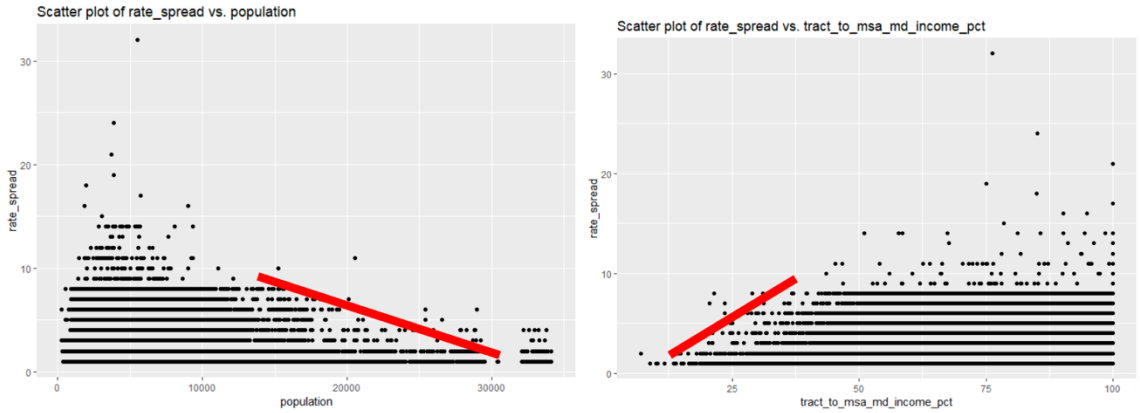

Figure 4 Population and tract median family income percentages distribution for rate spread values

3.2 Multidimensional Analysis

The analysis of spread rate, loan type and loan amount show that the higher loan amounts are al- most conventional loans. The analysis of spread rate, applicant income and loan type show that applicants with higher income up from 2500 have conventional loans.

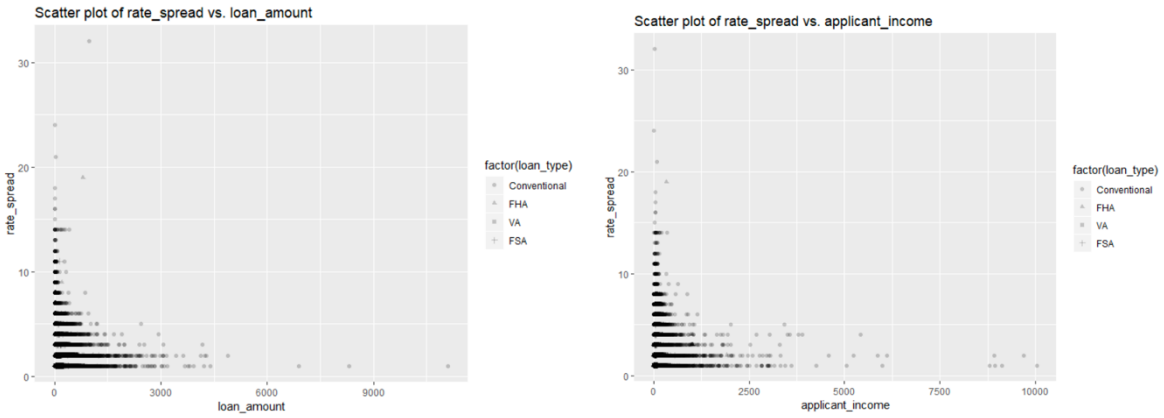

Figure 5 Loan amount and applicant income distribution for rate spread values

The analysis of spread rate, applicant income and loan purpose shows that applicants with higher income (approx. 1250) mainly use loans for purchase and refinancing. Improvement loans are mainly taken by applicants with lower income at a higher spread rate. There is also peak of applicants with very low income that take refinancing loans at high spread rate.

3.3 Rate Spread

Since rate spread is the target variable and interest of the entire analysis, let's investigate this column first.

The distribution of distinct rates is shown above, with 115,091 rates 1. It is the majority value, accounting for 57.5% of the 200,000 observations. Few rates are greater than 11, and no rates between 39 and 99. There are 3 rates of 99 that describe extreme value of outliers. The pie char of the rate spread column shows that the greater rate spread, the lower the amount of rate. More than half of the rates spread are 1, as shown below.

1	2	3	4	5	6	7	8	9	10
115091	43464	13663	8292	6353	7854	2455	2611	77	41
11	12	13	14	15	16	17	18	19	21
41	16	6	20	2	2	1	1	2	1
24	32	39	99						
1	2	1	3						

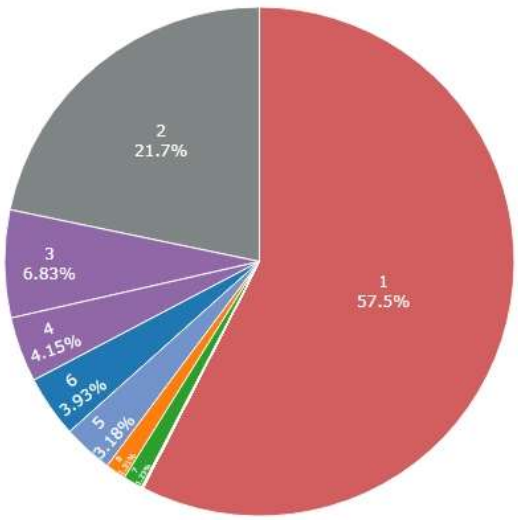

Figure 6 Rate spread values distribution

3.4 Loan Amount and Applicant Income

The loan amount is an integer column representing the size of the loan requested in thousands of dollars, and applicant income is another numerical indicator that describes the applicant's income in thousands of dollars. The comparison between these two variables at the same scale can be considerable and thought-provoking.

Before making the plot, it is important to exclude those outliers from the data. For loan amounts, the first quartile is $67k and the third quartile is $179k. So, with an IQR (interquartile range) of $112k, the upper bound of the outlier could be calculated as Q3 + 1.5 * IQR = $347k. Therefore, any loan amount within $347k will be used to make the plot.

For the applicant income, the first quartile is $39k and the third quartile is $83k. Therefore, the IQR is $44k and the upper bound of the outliers can be calculated as Q3 + 1.5 * IQR = $149k. Hence, any applicant income above $149k will not be used to draw the plot below.

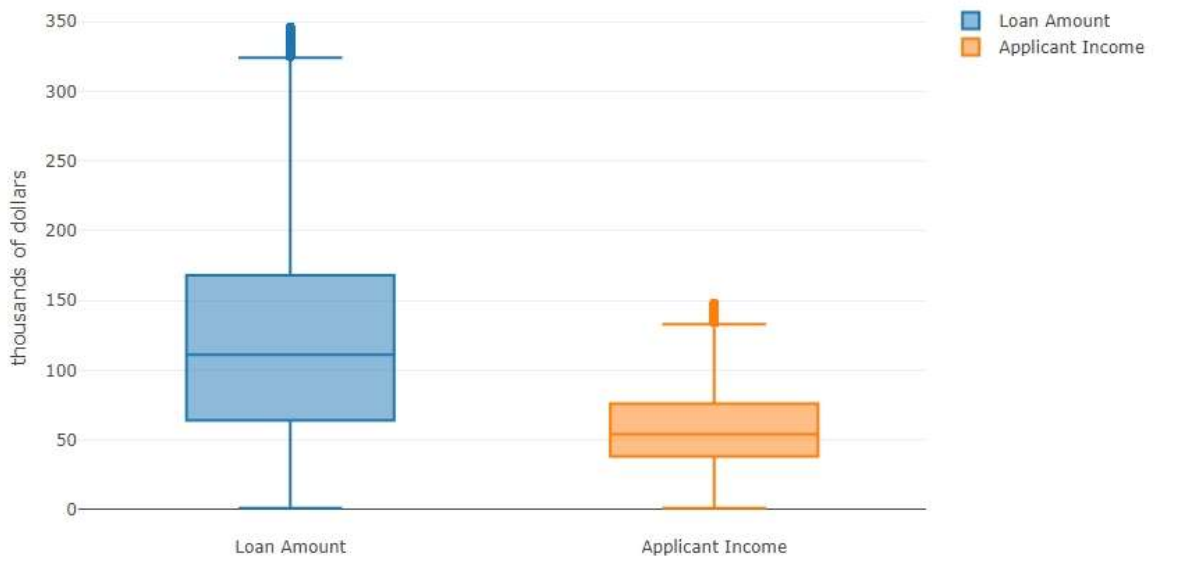

Figure 7 Loan amount and Applicant income box plots

The essentially spread of the loan amount and the applicant income were examined through the above combined box plot, and the results showed that the least loan and income were both $1k.

The box plot of the loan amount shows that $67k to $179k hold the middle 50% of the loan, and the thick line within the box shows the median loan is $116k. The box plot of the applicant income shows that $39k to $83k hold the middle 50% of the income, and the thick line within the box shows that the applicant's median income is about $74.

The combined box plot shows a comparison of two related variables, where applicant income seems to be generally lower than the loans, which is a common situation for each type of mortgage application.

The histogram below is another great visual representation of the loan and income comparison. It shows the frequency of two features. The graph explicitly indicates that more than 2,000 which is the peak quantity were founded around $100k loan. It applies common sense, where $100k is the general case for a loan application.

In addition to loans, the histogram also shows that more than 3,000 were found near the applicant income $50k, which is the majority. It applies general knowledge, where $50k is roughly the median income of a U.S resident.

The figure below also shows the shape of the data is skewed to the left, where the tails of loans and income distribution are long. Combined with the box plot results in the previous section, it turns out that small loans are more common than large loans. Median income is the most common situation for mortgage applicants.

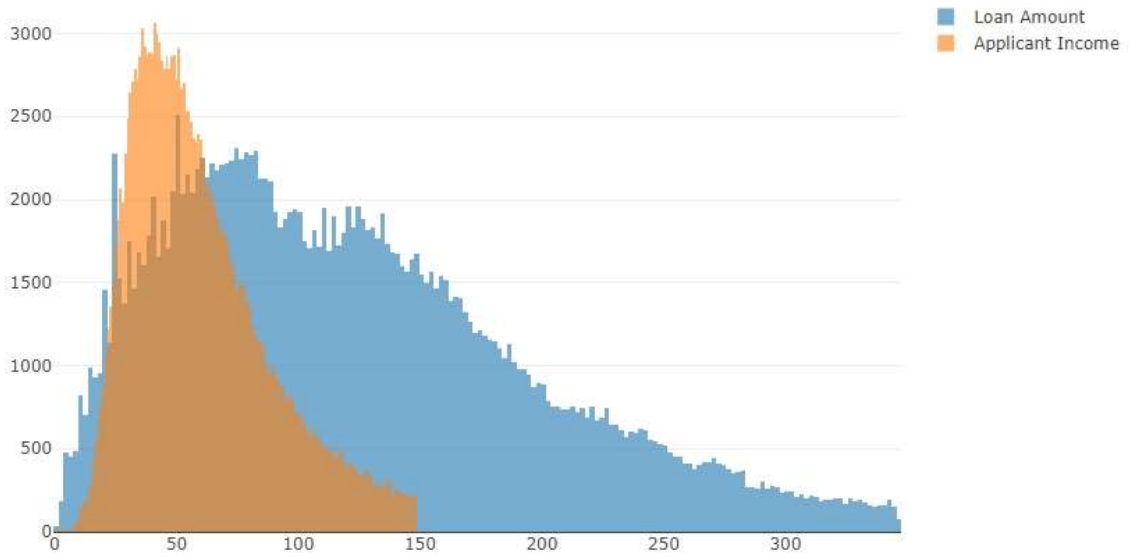

Figure 8 Loan amount' and Applicant income' values distribution plots

3.5 Rate Spread vs Loan Type

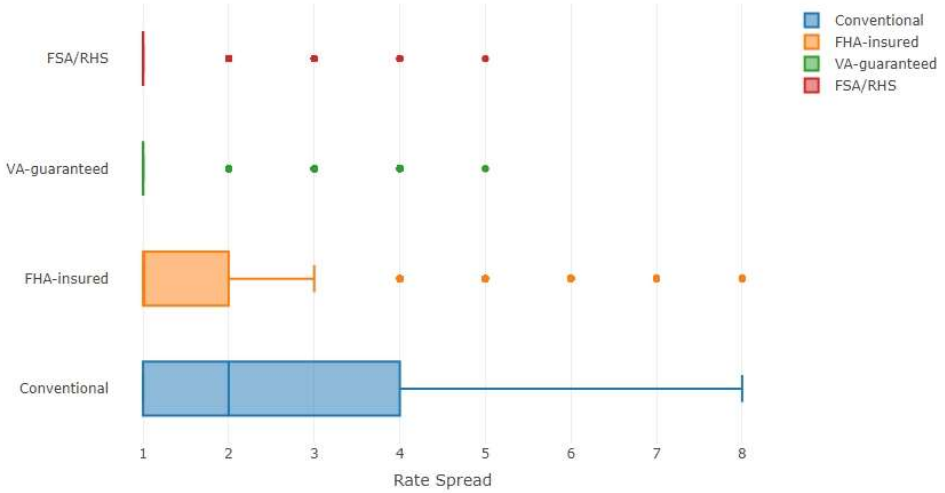

Figure 9 Compression box plots of rate spread for different loan types

As shown in the grouped box plot above, the average and median rate spread of conventional loan are the largest, and the interquartile range is the wildest. Not surprisingly, the conventional is a general loan type other than FHA-insured which is administrated by Federal Housing, it should keep rate spread low.

The VA-guaranteed and FSA/RHS types retain the smallest interquartile range, indicating that they have fewer application cases and a narrow range of rate spread.

3.6 Rate Spread vs Loan Purpose

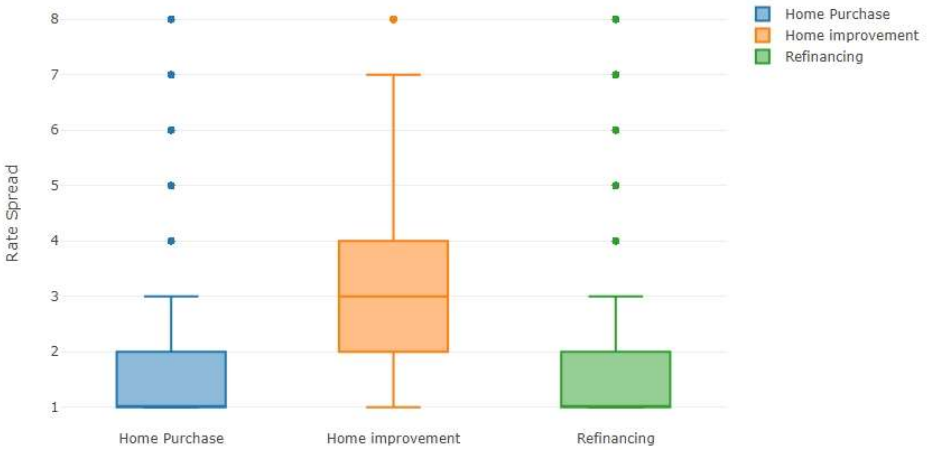

Figure 10 Compression box plots of rate spread for different loan purposes

As shown in the grouped box plot above, the home improvement has the largest mean and median rate spread as well as the widest interquartile range. The financial institutions consider such kind of loan purpose to be the highest risk cases, so they usually have higher rate spread than the other two loan purposes. A very interesting finding is that, home purchases and refinancing appear to share the same level of risk and rate spread.

3.7 Rate Spread vs Property Type and Co-Applicant

The apparent relationship between the rate spread and individual variable is easy to implement and define. Nevertheless, the relationship between multiple variables is more realistic and complex. when multiple variables are considered simultaneously, they will become apparent. Combined and grouped bar charts were created to help identify complex relationships.

Figure 11 Property type compression of Applicant and Co-Applicant

The chart above represents some interesting aspects of property type and whether co applicant or not. The proportion of main applicant from one to four family is often high, while the proportion of co applicant is higher in the case of multifamily loan application.

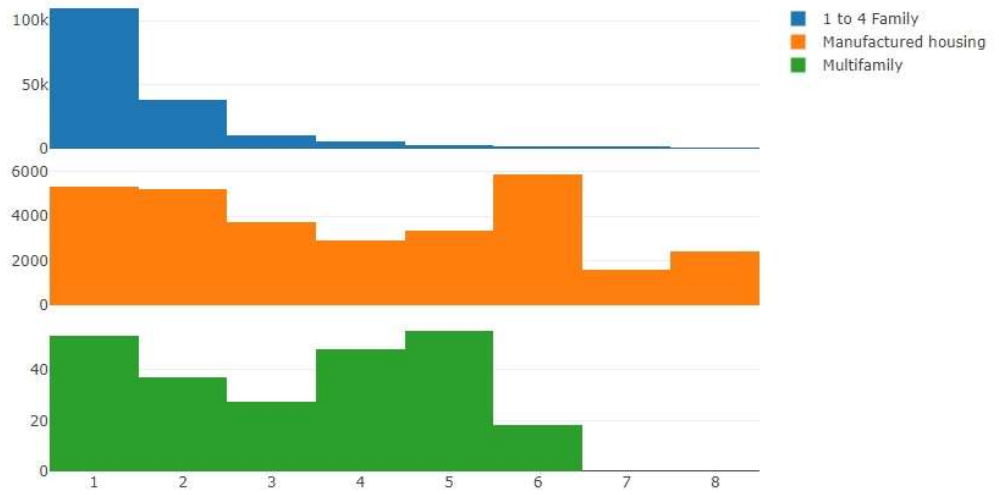

Figure 12 Group distribution of rate spread for property types

The grouped histogram plots above shows rate spread distribution of one to four family as well as manufactured housing and multifamily. The rate spread of those three property type distributions have the different characteristics. For one to four family, the histogram is a right-skewed plot indicating there are a bit lower rate offered in the application. For manufactured housing and multifamily, the histogram is evenly distributed with higher rate spread.

4 Conclusion

4.1 Training Model

After analyzing and determining the different relationships between rate spread and categorical and numerical characteristics, an efficiency prediction models is needed to predict the actual rate spread. According to the relationship analyzed in the previous sections, the CatBoost model is constructed by Catboost package in Python. The model is trained with 90% of the entire dataset and tested with the remaining 10%. There are 187,348 observations after data preprocessing, 131,143 observations for the training dataset and 56,205 test records.

An initial model for CatBoost was created with default parameters, such as iterations = 500, learning rate = 0.03, depth = 6, l2_leaf_reg = 3.0 and so on. The resulting model performance metric is RMSE = 0.8976, R-squared = 0.7085, MAE = 0.5960.

4.2 Optimizing Model

From the feature importance metric in the initial model, a fractional number ranks the importance for each variable. A threshold of 0.1 was used to keep the top importance features for the label variable. As a result 11 less important features were eliminated to optimize the model, such as msa/md, state code, county code, applicant race, applicant sex, population, minority population pct., tract to msa/md income pct., number of owner occupied units, number of 1 to 4 family units and co applicant.

```
> model$feature_importances
                                        [,1]
loan_type                         10.409948368
property_type                     24.923771673
loan_purpose                       7.844230695
occupancy                          0.956804265
loan_amount                       10.798643641
preapproval                        0.249332715
msa_md                             0.000000000
state_code                         0.070871745
county_code                        0.000000000
applicant_ethnicity                0.913205125
applicant_race                     0.013210721
applicant_sex                      0.000000000
applicant_income                   0.317601376
population                         0.024034850
minority_population_pct            0.059926761
ffiecmedian_family_income          1.552725976
tract_to_msa_md_income_pct         0.000000000
number_of_owner.occupied_units     0.000000000
number_of_1_to_4_family_units      0.002522124
lender                            28.836553402
co_applicant                       0.000000000
log_applicant_income               0.162149788
log_loan_amount                    8.102230103
log_ffiecmedian_family_income      0.865281972
loan_to_income                     2.395847581
log_loan_to_income                 1.501107116
>
```

Figure 13 Feature importance table for machine learning model

The updated training and test datasets contain only 15 features. An update model was trained with less features but same number of observations. A set of better performance parameters for the CatBoost model was found as iterations: 500, learning rate: 0.05, depth: 8, l2_leaf_reg: 6.0. The resulting model performance metric (RMSE) is 0.8051, R-squared is 0.7654, MAE is 0.5274.

Since the reduced features set was more time efficient and resulting better performance compared to the initial model, the reduced version feature set was determined as the final model for predicting the submission dataset.

4.3 Summary and results

The steps for the final model have been explained for each step in the process but here is a summary:

- **Missing values** were replaced with the median value for numerical features, and the most common value for categories
- Rate spread with a value higher than 16 were seen as **outliers** (which was true for 12 loan applications) and replaced with the median rate spread value
- The **features used** in the model were:
  Numerical: ’loan amount’, ’applicant income’ and ’minority population’ Categorical: All categorical features were kept
- **Feature transformation** that was done:
  - The features ’loan amount’ and ’applicant income’ were transformed to their logarithm
  - The categorical feature ’lender’ was removed and replaced with two numerical features; one with the median loan amount and the other the mean rate spread for the specific lender for the respective loan application
- **Standard scaling** was done
- The chosen **model** was ’CatBoost’

In total the model consisted of 816 features distributed as following:

Feature	Features in the model
’msa md’	409
’state code’	52
’county code’	317
’loan type’	4
’property type’	3
’loan purpose’	3
’occupancy’	3
’preapproval’	3
’applicant ethnicity’	4
’applicant race’	7
’applicant sex’	4
’co-applicant’	2
’loan amount’	1
’applicant income’	1
’minority population pct’	1
’lender ratespread mean’	1
’lender loanamount median’	1

As 10% of the data from the given training dataset was not used for training the model it could be used for testing. The result from the model evaluation (comparing true labels with predicted labels for this test set) is shown in the table below. Note that the result could differ depending on the random seed used when splitting the dataset.

Evaluation measure	Value
Mean Square Error	0.653
Root Mean Square Error	0.808
Mean Absolute Error	0.535
Median Absolute Error	0.341
R^2	0.747
Adjusted R^2	0.743

This model was then used to predict the real test labels using the given test dataset. When handed in, an R^2 value of 0.7149 was achieved.

Appendix

Explanation of the features

Features	*Description*	*Datatype*
U.S. state code	A categorical with no ordering indicating the U.S. state	Categorical
County code	A categorical with no ordering indicating the county	Categorical
Metropolitan Statistical Area	A categorical with no ordering indicating Metropolitan Statistical Area/Metropolitan Division	Categorical
Lender	A categorical with no ordering indicating which of the lenders was the authority in approving or denying this loan	Categorical
Loan Type	Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured	Categorical
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	Categorical
Loan Purpose	Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing	Categorical
Occupancy	Indicates whether the property to which the loan application relates will be the owner's principal dwelling	Categorical
Preapproval	Indicate whether the application or loan involved a request for a pre-approval of a home purchase loan	Categorical
Loan Amount	Size of the requested loan in thousands of dollars	Numerical
Applicant income	In thousands of dollars	Numerical
Applicant ethnicity	Ethnicity of the applicant	Categorical
Applicant race	Race of the applicant	Categorical
Applicant sex	Sex of the applicant	Categorical
Co-applicant	Indicates whether there is a co-applicant or not	Boolean
Population	Total population in tract	Numerical
Minority population %	Percentage of minority population to total population for tract	Numerical
FFIEC Median family income	FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC)	Numerical
Tract median family income %	Percentage of tract median family income compared to MSA/MD median family income	Numerical
Number of owner-occupied units	Number of dwellings, including individual condominiums, that are lived in by the owner	Numerical
Number of 1 to 4 family-units	Dwellings that are built to house fewer than 5 families	Numerical

Label to predict	*Description*	*Datatype*
Rate spread	Indicates the difference between the offered mortgage rate for the applicant and the standard rate for a comparative mortgage	Numerical

Explanations of the numbers representing categorical values

Applicant ethnicity

- Hispanic or Latino
- Not Hispanic or Latino
- Information not provided by applicant in mail, Internet, or telephone application
- Not applicable
- No co-applicant

Applicant race

- 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
- No co-applicant

Applicant sex

- Male
- Female
- Information not provided by applicant in mail, Internet, or telephone application
- Not applicable

Loan purpose

- Home purchase
- Home improvement
- Refinancing

Loan type

- Conventional (any loan other than FHA, VA, FSA, or RHS loans)
- FHA-insured (Federal Housing Administration)
- VA-guaranteed (Veterans Administration)
- FSA/RHS (Farm Service Agency or Rural Housing Service)

Occupancy

- Owner-occupied as a principal dwelling
- Not owner-occupied
- Not applicable

Preapproval

- Preapproval was requested
- Preapproval was not requested
- Not applicable

Property type

- One to four-family (other than manufactured housing)
- Manufactured housing
- Multifamily

Co-applicant

- True
- False