Predicting Mortgage Rates From Government Data

Dieuwke Doesburg
November 2019

1. Executive Summary

This document presents an analysis of data concerning mortgage applications and rate spread.

Key definitions

- **Rate spread (%)**
  The difference between the offered mortgage rate for the applicant and the standard rate for a comparative mortgage

- **Training dataset**
  Characteristics of 200.000 mortgage applications including rate spread

- **Test dataset**
  Characteristics of 200.000 mortgage applications for which the true rate spread is unknown

Data was first explored with summary and descriptive statistics. Subsequently, data was visualized to identify potential relationships between mortgage application characteristics and rate spread.

A regression model was created to predict the rate spread based on the mortgage application characteristics. After performing these analyses, the most important conclusions are the following:

- Rate spread generally has a low value of 1.0 to 2.0%, however high percentages of up to 99% are also possible
- The most common loan purpose is for home purchase, followed by refinancing and home improvement
- Most mortgage applications were done by men
- Applications for home improvements have the highest (log) rate spread
- Similarly a conventional loan type has the highest (log) rate spread
- There do not appear to be obvious differences in (log) rate spread for different applicant ethnicities, applicant races or applicant sexes

2. Initial Data Exploration

Each row of the dataset represents a HMDA-reported loan application, in one particular year.

2.1. Variable definitions

There are 21 variables in the dataset

2.1.1. Numeric variables

- **Loan amount**
  Size of the requested loan in thousands of dollars

- **Applicant income**
  Income in thousands of dollars

- **Population**
  Total population in tract

- **Minority population (%)**
  Percentage of minority population to total population for tract

- **Median family income MSA/MD**
  FFIEC median family income in dollars for the MSA/MD in which the tract is located

- **Tract to MSA/MD median family income (%)**
  Percentage of tract median family income compared to the MSA/MD median family income

- **Number of owner-occupied units**
  Number of dwellings, including individual condominiums, that are lived in by the owner

- **Number of 1-4 family units**
  Number of dwelling that are built to house fewer than 5 families

- **Rate spread (%)**
  The difference between the offered mortgage rate for the applicant and the standard rate for a comparative mortgage

2.1.2. Categorical and Boolean variables

- **Row id**
  A unique identifier with no intrinsic meaning

- **Loan type**
  Four categories, indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured

- **Property type**
  Three categories, indicates whether the loan or application was for a one-to-four-family dwelling (other than manufactured housing), manufactured housing, or multifamily dwelling

- **Loan purpose**
  Three categories, indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing

- **Occupancy**
  Three categories, indicates whether the property to which the loan application relates will be the owner's principal dwelling

- **Preapproval**
  Three categories, indicates whether the application or loan involved a request for a pre-approval of a home purchase loan

- **MSA/MD**
  Categorical with no ordering indicating Metropolitan Statistical Area/Metropolitan Division where a value of -1 indicates a missing value, 409 possible values

- **State code**
  Categorical with no ordering indicating the U.S. State where a value of -1 indicates a missing value, 53 possible values

- **County code**
  Categorical with no ordering indicating the county where a value of -1 indicates a missing value, 306 possible values

- **Applicant ethnicity**
  Five categories, ethnicity of the applicant.

- **Applicant race**
  Eight categories, race of the applicant.

- **Applicant sex**
  Five categories, sex of the applicant.

- **Lender**
  A categorical with no ordering indicating which of the lenders was the authority in approving or denying this loan, 3893 possible values

- **Co applicant**
  Boolean variable, indicates whether there is a co-applicant (often a spouse) or not

2.2. Individual Feature Statistics

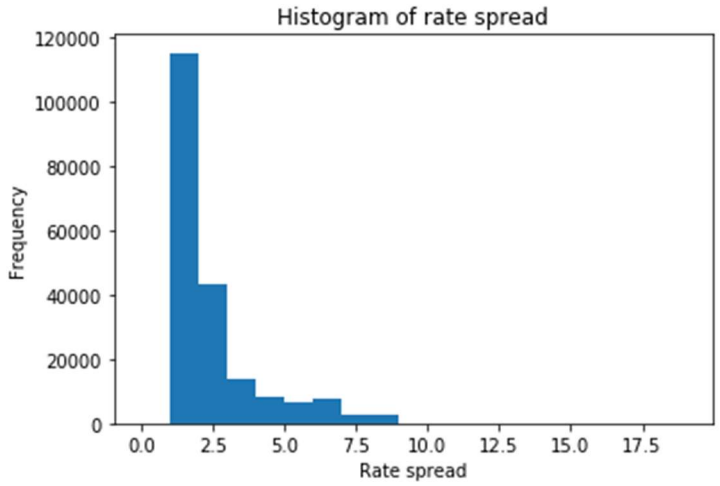
The dataset was first explored with summary and descriptive statistics.

2.2.1. Numeric variables

Summary statistics for minimum, maximum, mean, median, standard deviation and count, were calculated for numeric variables of the training dataset.

Variable	Min	Max	Mean	Median	Std Dev	Count
Loan amount	1	11 104	142.57	116	142.56	200 000
Applicant income	1	10 042	73.62	56	105.70	189 292
Population	7	34 126	5391.10	4959	2669.03	198 005
Minority population (%)	0.33	100	34.24	26.00	27.93	198 005
Median family income MSA/MD	17 860	125 095	64 595.36	63 485	12724.51	198 015
Tract to MSA/MD median family income (%)	6.19	100	89.28	98.96	15.06	197 977
Number of owner-occupied units	3	8747	1402.87	1304	706.88	197 988
Number of 1-4 family units	6	13 615	1927.34	1799	886.58	197 984
Rate spread	1.0	99.0	1.98	1.0	1.66	200 000

Rate spread is the target variable for the regression model. The mean rate spread is larger than the median, with a comparatively large standard deviation. This indicates that the rate spread values are right-skewed, as can be visualized with a histogram of rate spread (note: outliers of rate spread larger than 20% are not included in this histogram for visualization purposes. Most rate spreads are at the lower end of the rate spread range.



2.2.2. Categorical and Boolean variables

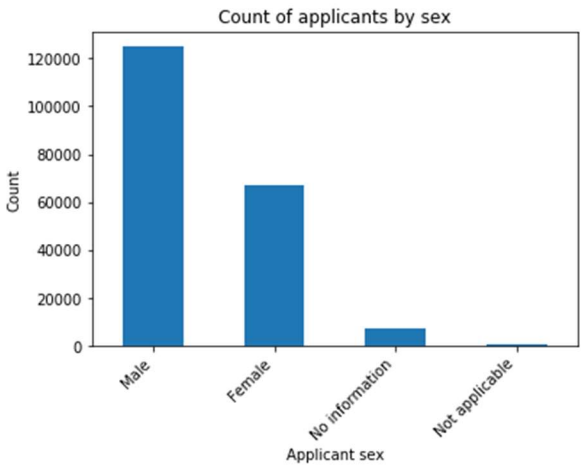
Bar charts were created to assess the frequency of categorical features, with the exception of variables MSA/MD, state code, county code and lender due to the high number of categories for these variables (resp. 409, 53, 306 and 3893 categories).

The following could be deducted from the bar charts:

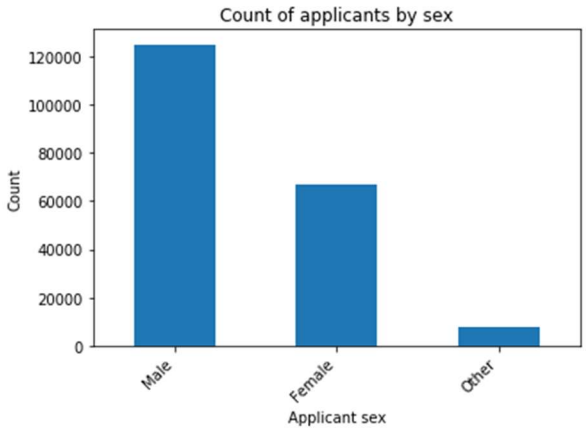
- Most mortgage application have a conventional or FHA-insured **loan type**
- Most properties are either one-to-four-family dwellings or manufactured housing (**property type**)
- The most common **loan purpose** is for home purchase, followed by refinancing and home improvement
- Most properties will be owner-occupied as a principal dwelling (**occupancy**)
- For most properties, **preapproval** was not applicable
- The most common **applicant ethnicity** is: Not Hispanic or Latino, followed by 2. Hispanic or Latino, 3. Information not provided and 4. Not applicable
- The most common **applicant race** is white, followed by Black or African American and Information not provided
- Most mortgage applications were done by men (**applicant sex**)

For **applicant sex**, there appear to be 4 different possibilities. Because of the relatively small number of applicants in category No information and Not applicable, these categories were combined into category Other to decrease the number of categories.

Before

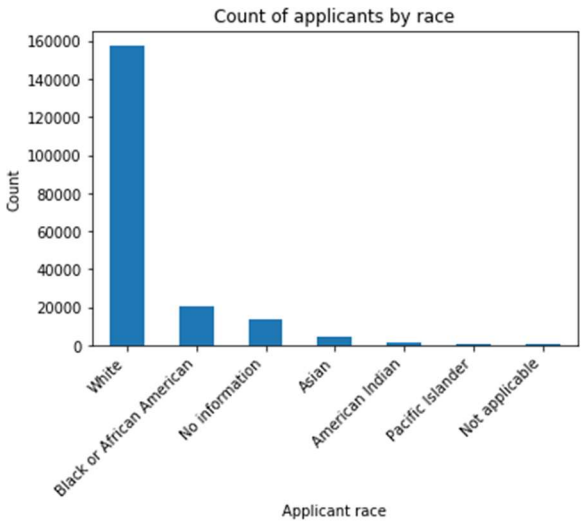


After

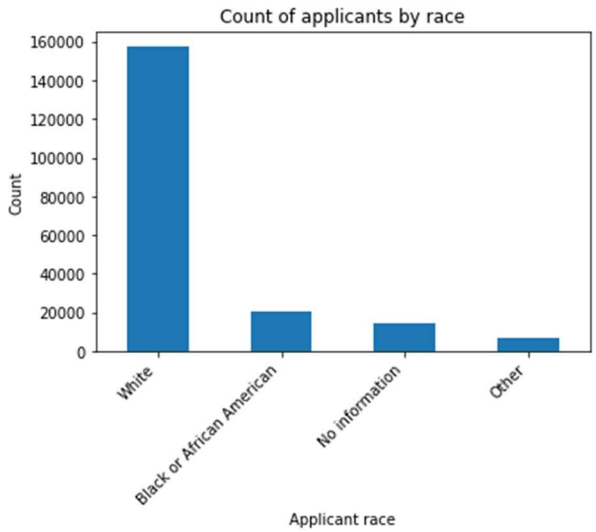


For variable **applicant race** there also appear to be many categories with low frequencies. Therefore, it was decided to combine the categories No information and Not applicable into category Unknown, and all categories except White and Black or African American into Other.

Before



After



2.3. Missing data and duplicates

Both the **training dataset** and **test dataset** were tested for missing values. The following amount of missing values were found for each dataset

Variable with missing data	Training	% training missing	Test	% test missing
Applicant income	10 708	5.4%	10 371	5.8%
Population	1995	1.0%	1918	1.0%
Minority population (%)	1995	1.0%	1920	1.0%
Median family income MSA/MD	1985	1.0%	1905	1,0%
Tract to MSA/MD median family income (%)	2023	1.0%	1946	1.0%
Number of owner-occupied units	2012	1.0%	1933	1.0%
Number of 1-4 family units	2016	1.0%	1933	1.0%
State code	1338	0.7%	0	0%

2.3.1. Handling of missing data

Dropping rows with missing values was not an option, since the assignment is to give a prediction for all 200.000 rows in the test dataset. Therefore, imputation was used to cope with missing values in both datasets.

For numeric variables with missing data (all of the above except state code), the skewedness of the data was assessed by comparing the mean and median, and by plotting histograms of each variable (data not shown). Most variables turned out to be either right or left-skewed. Because of the relatively low percentage of missing data for each variable (between 5.8% to 1.0%), the median value of each variable was used to substitute for missing values.

For the categorical variable **state code**, it was not possible to replace the missing values (state code -1) with another reasonable value (seeing as a categorical variable has no median). The missing values were left as -1.

This did not raise any concern, mainly due to the reason that the variable will not be included in the regression model, due to the large number of categories (53).

2.3.2. Duplicates

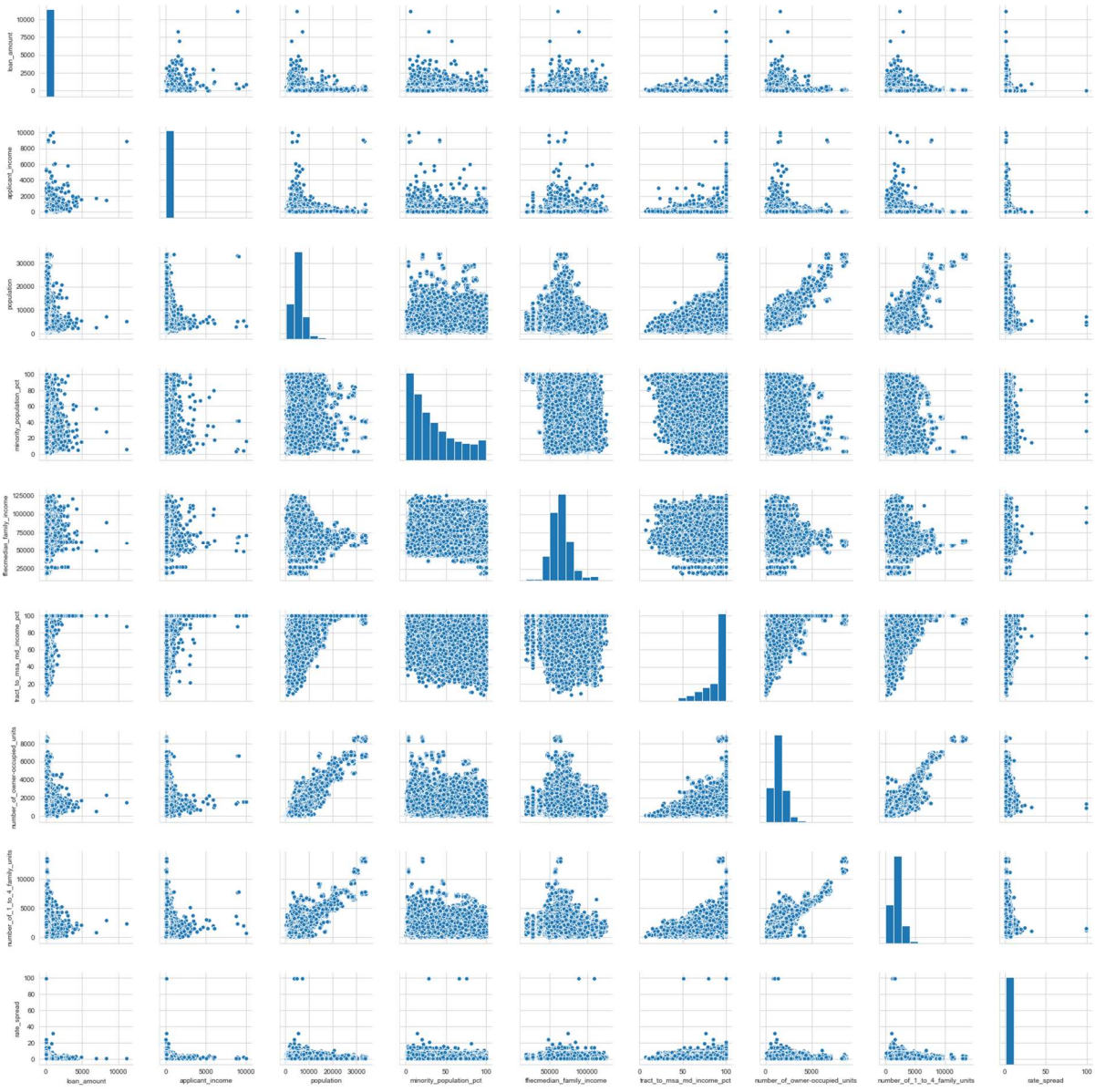
Both datasets were checked for duplicate rows, which were not present.

2.4. Correlation and Apparent Relationships

Next, an attempt was made to identify relationships between the variables in the dataset, in particular between **rate spread** and the other variables.

2.4.1. Numeric Relationships

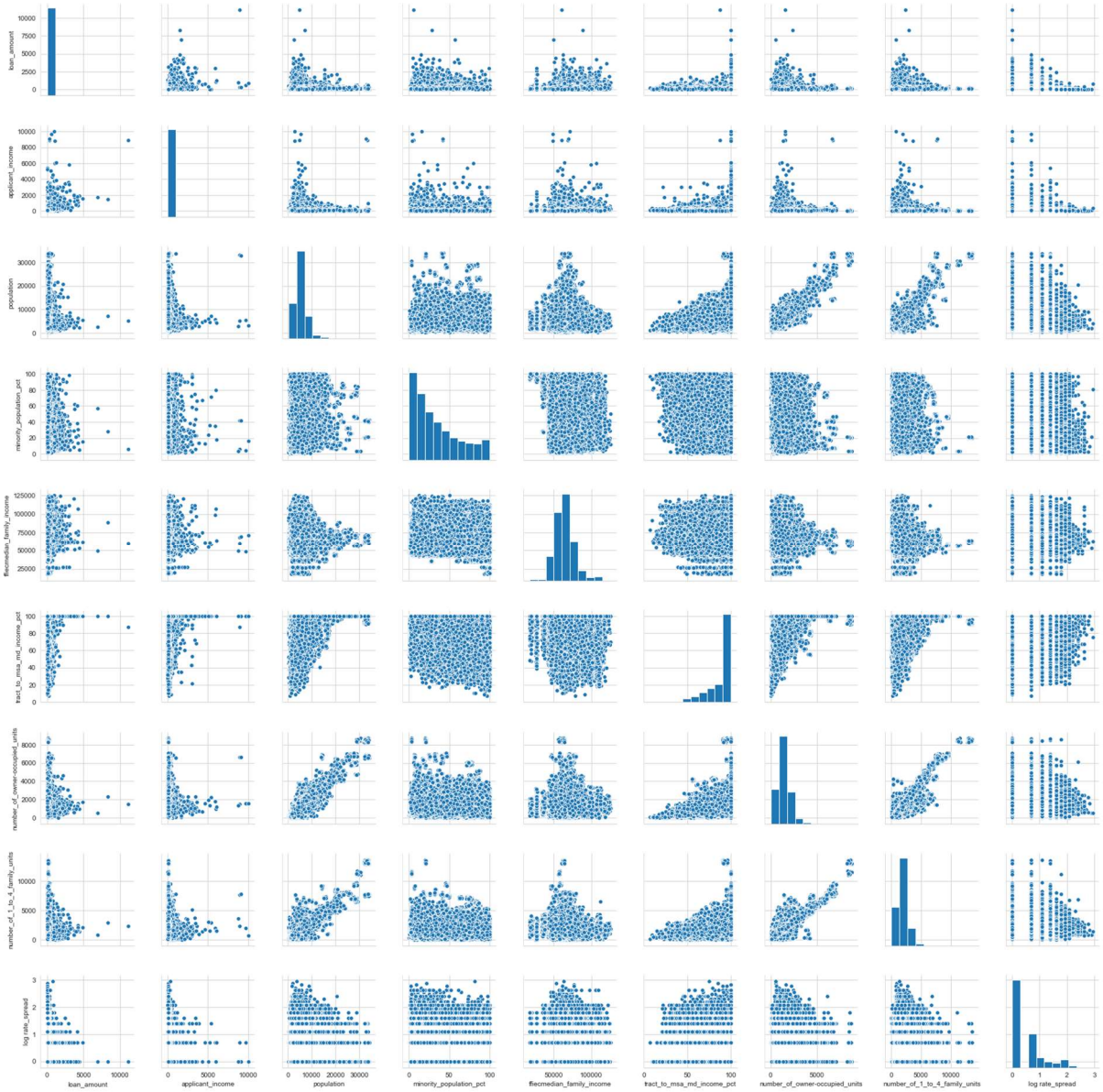
To explore the relationship between numeric features, a scatterplot matrix was generated.



The plots in the right-most column of this matrix show the relationship between **rate spread** and the other numeric variables.

Due to the fact that **rate spread** is right skewed, as mentioned earlier in paragraph **2.2.1**, and has some high outliers (eg. rate spread 99.0%), no linear relationships with numeric variables can be easily identified.

To improve the fit of the numerical variables to **rate spread**, the **log** normal value of **rate spread** will be used in the regression model. Therefore, an additional matrix was generated, with the **log rate spread** visible in the bottom row and right-most column. In order to better visualize the potential relationships, all **rate spread** outliers of >20.0% were excluded first.

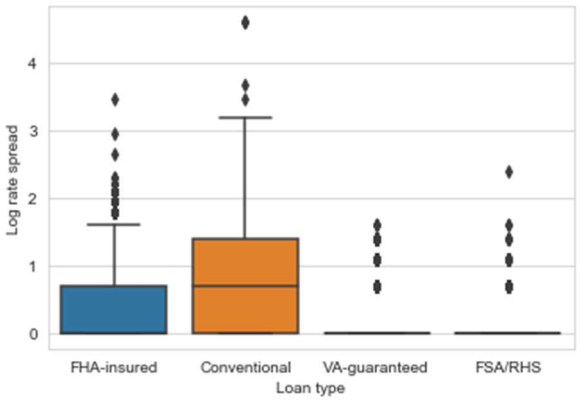


The resulting scatterplot matrix still does not show a clear linear relationship between **rate spread** and the other numeric variables. There does however appear to be a linear relationship between **population** and **number of owner-occupied units, population** and **number of 1-4 family units**, and **number of owner-occupied units** and **number of 1-4 family units**.

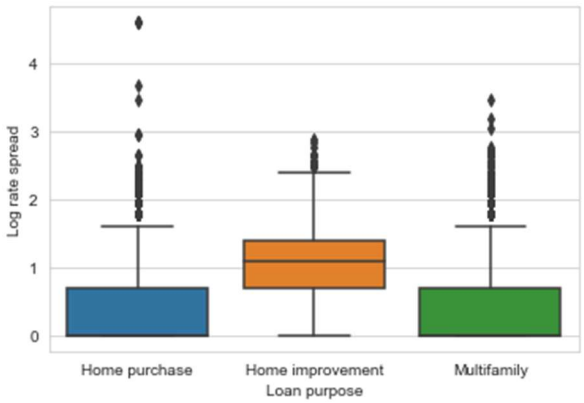
2.4.2. Categorical Relationships

The relationship between **log rate spread** and categorical variables (all but MSA/MD, state code, county code and lender) was assessed with boxplots. Two examples:

Loan type



Loan purpose



The following could be deducted from the boxplots:

- Mortgage applications with a conventional **loan type** appear to have the highest **log rate spread**
- A one-to-four-family **property type** has the lowest **log rate spread**
- Applications with a **loan purpose** for home improvements have the highest **log rate spread**
- An **occupancy** of category Not applicable has the highest **log rate spread**
- There appears to be no difference in mean **log rate spread** for the different categories of preapproval, although the maximum **log rate spread** is highest for the Not applicable category
- There do not appear to be obvious differences in **log rate spread** for different **applicant ethnicities, applicant races** or **applicant sexes**

3. Regression

A regression model (RandomForestRegressor) was trained on the **training dataset** and then used to predict the **log rate spread** of the **test dataset**. From these values the **rate spread** was calculated.

Before training the regression model the following steps were completed (for both datasets).

3.1. Imputation of missing data

As described in section 2.3.1.

3.2. Check for duplicates

There were no duplicate rows in the datasets.

3.3. Log transformation of rate spread (training dataset)

3.4. Check for columns with zero variance

There were no columns with zero variance in the datasets.

3.5. Feature engineering

In order to potentially optimize the performance of the regression model, the following features were added to the datasets:

- **Ratio** of **loan amount** to **applicant income**
- **Ratio** of **applicant income** to **median family income MSA/MD**
- **Median tract income** (= **median family income MSA/MD** x **Tract to MSA/MD median family income (%)**
- **Ratio** of **applicant income** to **median tract income**

3.6. Deduction of numeric variables due to high correlation

Correlations were calculated between all numeric columns except **log rate spread**. This tables displays the **top 10 highest absolute correlations.**

Variable 1	Variable 2	Absolute correlation
Ratio of applicant income to median family income MSA/MD	Ratio of applicant income to median tract income	0.963680
Applicant income	Ratio of applicant income to median family income MSA/MD	0.959327
Applicant income	Ratio of applicant income to median tract income	0.931013
Number of owner-occupied units	Number of 1-4 family units	0.904284
Population	Number of owner-occupied units	0.856067
Population	Number of 1-4 family units	0.836317
Median family income MSA/MD	Median tract income	0.693583
Tract to MSA/MD median family income (%)	Median tract income	0.610044
Loan amount	Applicant income	0.432428
Minority population (%)	Tract to MSA/MD median family income (%)	0.416285

One of the variables with the strongest correlation was deleted from the potential feature list, until all remaining correlations were <0.45. The following numerical features remain:

- Loan amount
- Minority population (%)
- Median family income MSA/MD
- Tract to MSA/MD median family income (%)
- Number of owner-occupied units
- Ratio of loan amount to applicant income
- Ratio of applicant income to median family income MSA/MD

3.7. Scaling of numerical features

The remaining numerical features were scaled so that they all have a same level of magnitude (mean of 0 with a standard deviation of 1).

3.8. Selection of categorical features + Boolean features

Categorical columns with too many categories were not used for the regression model, since they would add too many dimensions. This concerns the variables MSA/MD, state code, county code and lender (resp. 409, 53, 306 and 3893 categories). Furthermore, **applicant race** and **applicant ethnicity** are closely related, which is why it was decided to drop **applicant ethnicity** from the feature list. The remaining categorical/Boolean features are:

- Loan type
- Property type
- Loan purpose
- Occupancy

- Preapproval
- Applicant race
- Applicant sex
- Co-applicant (Boolean)

3.9. OneHotEncoding of categorical features

All remaining categorical features were recoded using OneHotEncoding, such that each variable is represented by dummy variables with values 0 or 1.

3.10. Creating the final feature set

The remaining numerical variables, recoded categorical variables and the Boolean variable were combined to one final feature set (35 features), on which to train the regression model.

3.11. Splitting the training dataset

The **training dataset** was split in the **'traintrain' dataset** (70% of the data) and the **'traintest' dataset** (remaining 30%).

3.12. Training the model

A RandomForestRegressor model was trained on the **traintrain dataset** including the **log rate spread** of this dataset.

3.13. Testing the model with the traintest dataset

The trained model was used to predict the **log rate spread** of the **traintest dataset**. Predicted log rate spread was compared to the actual log rate spread, which showed an R2 of 0.504.

3.14. Calculating the rate spread of the test dataset

The model was used to predict the **log rate spread** of the **test dataset** and subsequently transformed and rounded to zero decimals to generate the predicted **rate spread** for the **test dataset**.
The R2 in the test dataset is similar to the R2 in the **traintest dataset** = 0.51.

4. Conclusion

This analysis shows that the **rate spread** of and mortgage application can be estimated from its characteristics. In particular, loan type and loan purpose, and ratio of loan amount to applicant income are important factors.