Analysis of home mortgage rate

Yang Jiao

November 19, 2019

1 Executive summary

This report presents an analysis of data concerning how demographics, location, property type, lender and other factors are related to the mortgage rate offered to applicants. The analysis is based on 200,000 observations of home mortgage disclosure act (HMDA) data, each containing specific characteristics of an loan application.

Potential relationships between characteristics and rate spread were identified. A model to predict this rate for loan applications was created.

The author reached the following conclusions:

geo The location

2 Data exploration and analysis

2.1 Individual Feature Statistics

The targeted value is rate spread, which is discrete numerical value. The rate spread distribute from 1.0 to 8.0 and has outliers from 9.0 to 32.0 and then a few cases in 99.0.

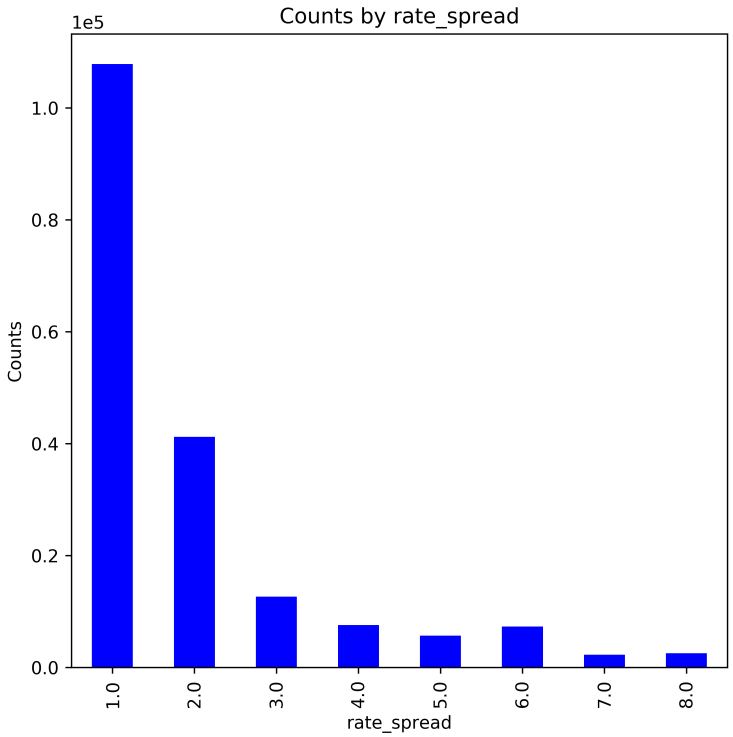

Figure 1: The distribution of rate spread in training dataset.

For numerical columns

- Loan information
  - Loan amount (K$)
- Applicant information
  - Gross Annual Income (K$)
- Census information
  - Population
  - Minority population (%)
  - FFIEC median family income ($) for the MSA/MD
  - Tract to MSA/MD median family income (%)
  - Number of owner occupied units
  - Number of 1- to 4-family units

categorical features,

- Property location, which includes
  - MSA/MD
  - State
  - County
- Loan information
  - Lender
  - Loan type – Conventional, FHA-insured, VA-guaranteed and FSA/RHS
  - Property type – One to four-family, manufactured housing and multifamily
  - Loan purpose – Home purchase, home improvement and refinancing
  - Owner occupancy – Owner-occupied as a principal dwelling, not owner-occupied and not applicable
  - Preapproval – Preapproval was requested, was not requested or not applicable
- Applicant information
  - Ethnicity
  - Race
  - Sex
  - Co-applicant

In the categorical features, the property location features, including MSA/MD, state and county have large amount of unique values, and some values only have a few entries. It the same case for lender feature. From the plot, property location highly related to rate ratio. These features are included and values with few entries will be selected out in feature selection processes. In the following plots, only values with at least 1% frequency are shown.

3 Key findings

Based on the analysis of the home mortgage data, a predictive model to estimate the rate spread was created. Based on the apparent relationships identified when analyzing the data, a random forest regressor model was created to predict the rate spread. Superparameters are validated using nested cross validation method.

The model was trained with .. and tested with the remaining 20,000 observations. A scatter plot shows the predicted rate spread and the actural rate spread. The metrics of this prediction

4 Conclusions and recommendations

From this analysis, home mortgage rate can be predicted

Table 1: Table

RMSE
RAE
R2