DAT102x: Predicting Mortgage Rates From Government Data  
HOSTED BY MICROSOFT

Problem Description

* [About the Data](https://www.datasciencecapstone.org/competitions/18/mortgage-rates-from-government-data/page/56/#about-the-data)
* [Target Variable](https://www.datasciencecapstone.org/competitions/18/mortgage-rates-from-government-data/page/56/#target-variable)
  + [Submission Format](https://www.datasciencecapstone.org/competitions/18/mortgage-rates-from-government-data/page/56/#submission-format)
  + [Performance Metric](https://www.datasciencecapstone.org/competitions/18/mortgage-rates-from-government-data/page/56/#performance-metric)
* [Features](https://www.datasciencecapstone.org/competitions/18/mortgage-rates-from-government-data/page/56/#features)
  + [Example Row](https://www.datasciencecapstone.org/competitions/18/mortgage-rates-from-government-data/page/56/#example-row)
* [References](https://www.datasciencecapstone.org/competitions/18/mortgage-rates-from-government-data/page/56/#references)

About the Data

Your goal is to predict the [rate spread of mortgage applications](https://ffiec.cfpb.gov/tools/rate-spread) according to the given dataset, which is adapted from the [Federal Financial Institutions Examination Council's (FFIEC)](https://www.ffiec.gov/hmda/hmdaflat.htm).

Target Variable

We're trying to predict the variable rate\_spread for each row of the test data set.

Your job is to:

1. Train a model using the inputs in train\_values.csv and the labels train\_labels.csv
2. Predict value for each row in test\_values.csv for which you don't know the true value of rate\_spread.
3. Output your predictions in a format that matches submission\_format.csv **exactly**.
4. Upload your predictions to this competition in order to get a score.
5. Export your grading token (click the "Export Score for EdX" tab) and paste it into the assignment grader on edX to get your course grade.

Submission Format

The format for the submission file is two columns with row\_id and rate\_spread. The data type of rate\_spread is a float value.

If you predicted 1.0 as the rate spread for each respondent, the .csv file that you submit would look like:

row\_id,rate\_spread

0,1.0

1,1.0

2,1.0

3,1.0

4,1.0

⁝

Performance Metric

We're predicting a float value, so this is a regression problem. To measure regression, we'll use a metric known as the "coefficient of determination" or "R squared" (sometimes denoted R2 or r2). R-squared is one metric for evaluating regression models. Informally, r-squared is the proportion of the variance in the dependent variable that is explained by the model. R-squared can be calculated:

R2=1−∑i(yi−y^i)2/∑i(yi−y¯)2R2=1−∑i(yi−y^i)2/∑i(yi−y¯)2

Features

There are 21 variables in this dataset. Each row in the dataset represents a HMDA-reported loan application, and the dataset we are working with covers one particular year.

We provide a unique identifier called lender for each individual loan-making institution.

The variables are as follows:

PROPERTY LOCATION

* msa\_md (categorical) - A categorical with no ordering indicating Metropolitan Statistical Area/Metropolitan Division where a value of -1 indicates a missing value
* state\_code (categorical) - A categorical with no ordering indicating the U.S. state where a value of -1 indicates a missing value
* county\_code (categorical) - A categorical with no ordering indicating the county where a value of -1 indicates a missing value

LOAN INFORMATION

* lender (categorical) - A categorical with no ordering indicating which of the lenders was the authority in approving or denying this loan
* loan\_amount (int) - Size of the requested loan in thousands of dollars
* loan\_type (categorical) - Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured; available values are:
* 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\_type (categorical) - 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:
* 1 -- One to four-family (other than manufactured housing)
* 2 -- Manufactured housing
* 3 -- Multifamily
* loan\_purpose (categorical) - Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing; available values are:
* 1 -- Home purchase
* 2 -- Home improvement
* 3 -- Refinancing
* occupancy (categorical) - Indicates whether the property to which the loan application relates will be the owner's principal dwelling; available values are:
* 1 -- Owner-occupied as a principal dwelling
* 2 -- Not owner-occupied
* 3 -- Not applicable
* preapproval (categorical) - Indicate whether the application or loan involved a request for a pre-approval of a home purchase loan; available values are:
* 1 -- Preapproval was requested
* 2 -- Preapproval was not requested
* 3 -- Not applicable

APPLICANT INFORMATION

* applicant\_income (int) - In thousands of dollars
* applicant\_ethnicity (categorical) - Ethnicity of the applicant; available values are:
* 1 -- Hispanic or Latino
* 2 -- Not Hispanic or Latino
* 3 -- Information not provided by applicant in mail, Internet, or telephone pplication
* 4 -- Not applicable
* 5 -- No co-applicant
* applicant\_race (categorical) - Race of the applicant; available values are:
* 1 -- American Indian or Alaska Native
* 2 -- Asian
* 3 -- Black or African American
* 4 -- Native Hawaiian or Other Pacific Islander
* 5 -- White
* 6 -- Information not provided by applicant in mail, Internet, or telephone application
* 7 -- Not applicable
* 8 -- No co-applicant
* applicant\_sex (categorical) - Sex of the applicant; available values are:
* 1 -- Male
* 2 -- Female
* 3 -- Information not provided by applicant in mail, Internet, or telephone application
* 4 or 5 -- Not applicable
* co\_applicant (bool) - Indicates whether there is a co-applicant (often a spouse) or not

CENSUS INFORMATION

* population - Total population in tract
* minority\_population\_pct - Percentage of minority population to total population for tract
* ffiecmedian\_family\_income - FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC)
* tract\_to\_msa\_md\_income\_pct - % of tract median family income compared to 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\_to\_4\_family\_units - Dwellings that are built to house fewer than 5 families

INDEX AND TARGET VARIABLE

* row\_id - A unique identifier with no intrinsic meaning, but the IDs in your submission must match the submission format exactly
* rate\_spread - Indicates the difference between the offered mortgage rate for the applicant and the standard rate for a comparative mortgage.

Example Row

Here's an example of one of the rows in the dataset so that you can see the kinds of values you might expect in the dataset. Most are categorical, a few are numerical, and there can be missing values.

| **row\_id** | **0** |
| --- | --- |
| **loan\_type** | 1 |
| **property\_type** | 1 |
| **loan\_purpose** | 3 |
| **occupancy** | 1 |
| **loan\_amount** | 116 |
| **preapproval** | 3 |
| **msa\_md** | 24 |
| **state\_code** | 4 |
| **county\_code** | 106 |
| **applicant\_ethnicity** | 2 |
| **applicant\_race** | 3 |
| **applicant\_sex** | 2 |
| **applicant\_income** | 66 |
| **population** | 3263 |
| **minority\_population\_pct** | 52.815 |
| **ffiecmedian\_family\_income** | 71852 |
| **tract\_to\_msa\_md\_income\_pct** | 81.198 |
| **number\_of\_owner-occupied\_units** | 786 |
| **number\_of\_1\_to\_4\_family\_units** | 1067 |
| **lender** | 494 |
| **co\_applicant** | False |

References

* FFIEC. Home Mortgage Disclosure Act [https://www.ffiec.gov/hmda/default.htm](https://www.datasciencecapstone.org/competitions/18/mortgage-rates-from-government-data/page/56/).
* HMDA Loan Application Register [code sheet](https://www.ffiec.gov/hmdarawdata/FORMATS/2015HMDACodeSheet.pdf).