Homework 3

Crime Logistic Regression

Group 2

4/11/2021

Contents

Group 2 members: *Diego Correa, Jagdish Chhabria, Orli Khaimova, Richard Zheng, Stephen Haslett.*

Assignment Overview

In this homework assignment, you will explore, analyze, and model a data set containing information on crime for various neighborhoods of a major city. Each record has a response variable indicating whether or not the crime rate is above the median crime rate (1) or not (0).

Your objective is to build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. You will provide classifications and probabilities for the evaluation data set using your binary logistic regression model. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

- `zn`: the proportion of residential land zoned for large lots (over 25000 square feet) (predictor variable)
- `indus`: the proportion of non-retail business acres per suburb (predictor variable)
- `chas`: a dummy var. for whether the suburb borders the Charles River (1) or not (0) (predictor variable)
- `nox`: nitrogen oxides concentration (parts per 10 million) (predictor variable)
- `rm`: average number of rooms per dwelling (predictor variable)
- `age`: the proportion of owner-occupied units built prior to 1940 (predictor variable)
- `dis`: weighted mean of distances to five Boston employment centers (predictor variable)
- `rad`: index of accessibility to radial highways (predictor variable)
- `tax`: full-value property-tax rate per $10,000 (predictor variable)
- `ptratio`: pupil-teacher ratio by town (predictor variable)
- `lstat`: lower status of the population (percent) (predictor variable)
- `medv`: median value of owner-occupied homes in $1000s (predictor variable)
- `target`: whether the crime rate is above the median crime rate (1) or not (0) (response variable)

Deliverables

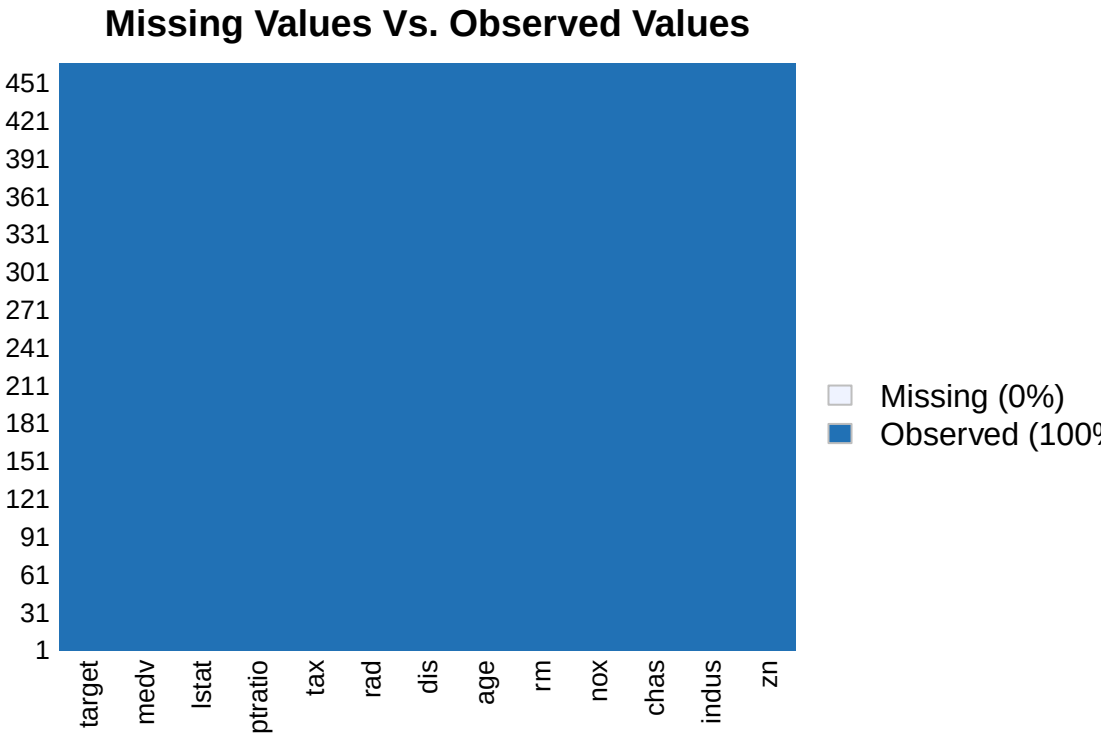
- A write-up submitted in PDF format. Your write-up should have four sections. Each one is described below. You may assume you are addressing me as a fellow data scientist, so do not need to shy away from technical details.
- Assigned prediction (probabilities, classifications) for the evaluation data set. Use a 0.5 threshold.
- Include your R statistical programming code in an Appendix.

Task 1: Data Exploration

Describe the size and the variables in the crime training data set.

```
##        zn             indus             chas              nox        
##  Min.   :  0.00   Min.   : 0.460   Min.   :0.00000   Min.   :0.3890  
##  1st Qu.:  0.00   1st Qu.: 5.145   1st Qu.:0.00000   1st Qu.:0.4480  
##  Median :  0.00   Median : 9.690   Median :0.00000   Median :0.5380  
##  Mean   : 11.58   Mean   :11.105   Mean   :0.07082   Mean   :0.5543  
##  3rd Qu.: 16.25   3rd Qu.:18.100   3rd Qu.:0.00000   3rd Qu.:0.6240  
##  Max.   :100.00   Max.   :27.740   Max.   :1.00000   Max.   :0.8710  
##        rm             age              dis              rad        
##  Min.   :3.863   Min.   :  2.90   Min.   : 1.130   Min.   : 1.00  
##  1st Qu.:5.887   1st Qu.: 43.88   1st Qu.: 2.101   1st Qu.: 4.00  
##  Median :6.210   Median : 77.15   Median : 3.191   Median : 5.00  
##  Mean   :6.291   Mean   : 68.37   Mean   : 3.796   Mean   : 9.53  
##  3rd Qu.:6.630   3rd Qu.: 94.10   3rd Qu.: 5.215   3rd Qu.:24.00  
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.00  
##       tax           ptratio          lstat             medv      
##  Min.   :187.0   Min.   :12.6   Min.   : 1.730   Min.   : 5.00  
##  1st Qu.:281.0   1st Qu.:16.9   1st Qu.: 7.043   1st Qu.:17.02  
##  Median :334.5   Median :18.9   Median :11.350   Median :21.20  
##  Mean   :409.5   Mean   :18.4   Mean   :12.631   Mean   :22.59  
##  3rd Qu.:666.0   3rd Qu.:20.2   3rd Qu.:16.930   3rd Qu.:25.00  
##  Max.   :711.0   Max.   :22.0   Max.   :37.970   Max.   :50.00  
##      target      
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0000  
##  Mean   :0.4914  
##  3rd Qu.:1.0000  
##  Max.   :1.0000
```

Check for missing values in the dataset.



Task 2: Data Preparation

Describe how you have transformed the data by changing the original variables or creating new variables.

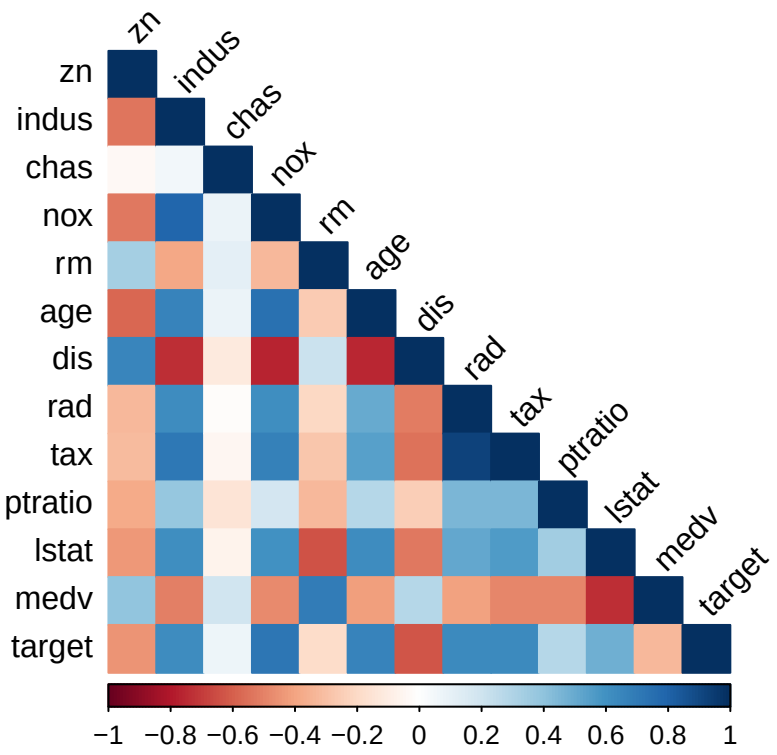
There are no missing values in the dataset so there is no need to impute values. Instead, we will split the training dataset (crime-training-data_modified.csv) into training and test sets using a 70/30 split respectively.

Summary of predictor variables in the training dataset.

```
##        zn             indus            chas              nox        
##  Min.   : 0.00   Min.   : 0.74   Min.   :0.00000   Min.   :0.3920  
##  1st Qu.: 0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4470  
##  Median : 0.00   Median : 8.56   Median :0.00000   Median :0.5200  
##  Mean   :12.55   Mean   :10.80   Mean   :0.08563   Mean   :0.5495  
##  3rd Qu.:20.00   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240  
##  Max.   :95.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710  
##        rm              age              dis              rad        
##  Min.   :4.138   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000  
##  1st Qu.:5.875   1st Qu.: 40.15   1st Qu.: 2.096   1st Qu.: 4.000  
##  Median :6.185   Median : 74.30   Median : 3.414   Median : 5.000  
##  Mean   :6.279   Mean   : 66.22   Mean   : 3.931   Mean   : 9.349  
##  3rd Qu.:6.633   3rd Qu.: 93.55   3rd Qu.: 5.344   3rd Qu.:24.000  
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000  
##       tax           ptratio          lstat            medv      
##  Min.   :187.0   Min.   :12.60   Min.   : 1.980   Min.   : 5.00  
##  1st Qu.:277.0   1st Qu.:17.40   1st Qu.: 6.885   1st Qu.:17.15  
##  Median :329.0   Median :18.90   Median :10.870   Median :21.20  
##  Mean   :402.5   Mean   :18.48   Mean   :12.487   Mean   :22.57  
##  3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.490   3rd Qu.:25.00
```

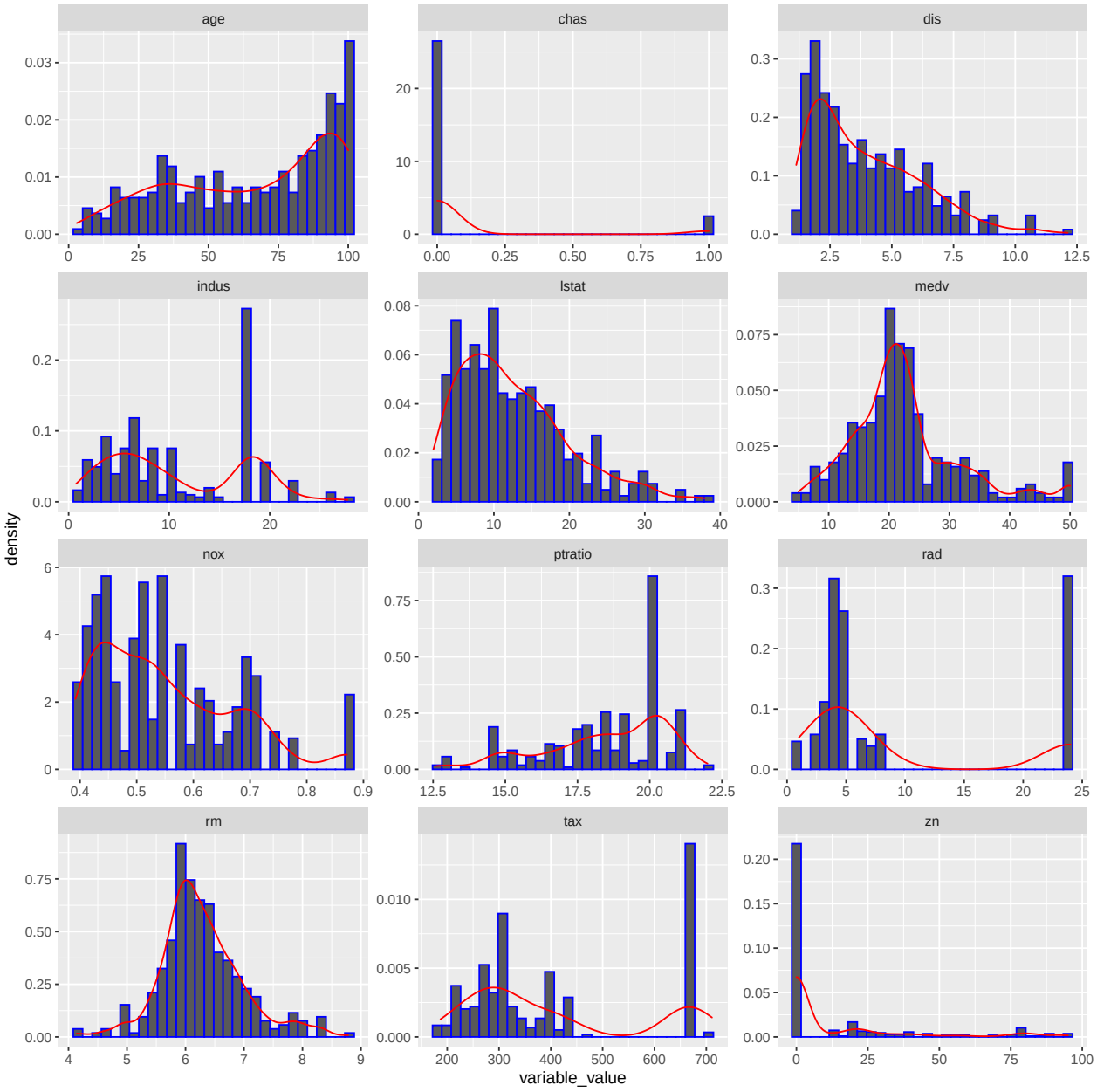
```
##  Max.   :711.0   Max.   :22.00   Max.   :37.970   Max.   :50.00
```

Correlation Matrix of Training Set Predictor Variables



Distributions

take a look at the distribution profiles for each of the predictor variables.



Task 3: Build Models

Using the training data, build at least three different binary logistic regression models, using different variables (or the same variables with different transformations).

Task 4: Select Models

Decide on the criteria for selecting the best binary logistic regression model.

Appendix