

# Homework 3 (Group 5)

## Binary Logistic Regression

Maria A Ginorio

3/30/2022

## Contents

Overview . . . . .	1
Dataset . . . . .	2
1. Data Exploration . . . . .	3
Objective . . . . .	3
Data Overview . . . . .	3
Distributions . . . . .	4
Outliers . . . . .	5
Relationships . . . . .	6
2. Data Preparation . . . . .	7
3. Building Models . . . . .	8
Model # 1 . . . . .	8
Model #2 . . . . .	9
Model #3 . . . . .	10
4. Select Models . . . . .	11
Model_1 Testing . . . . .	11
Model_2 Testing . . . . .	12
Model_3 Testing . . . . .	13
Final Selection . . . . .	14
References . . . . .	14
Apendix . . . . .	14

## 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:

## Dataset

- zn: proportion of residential land zoned for large lots (over 25000 square feet) (predictor variable)
- indus: 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: 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)
- black:  $1000(B_k - 0.63)^2$  where  $B_k$  is the proportion of blacks 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)

# 1. Data Exploration

## Objective

- Understand the variables provided
- Build a binary logistic regression model on the training data
- Predict the whether the neighborhood will be at risk for high crime.
- Provide classifications and probabilities for the evaluation data set using logistic regression.

## Data Overview

Lets first look at the raw data values by using the skim package

Table 1: Data summary

Name	crime_train
Number of rows	466
Number of columns	13
Column type frequency: numeric	13
Group variables	None

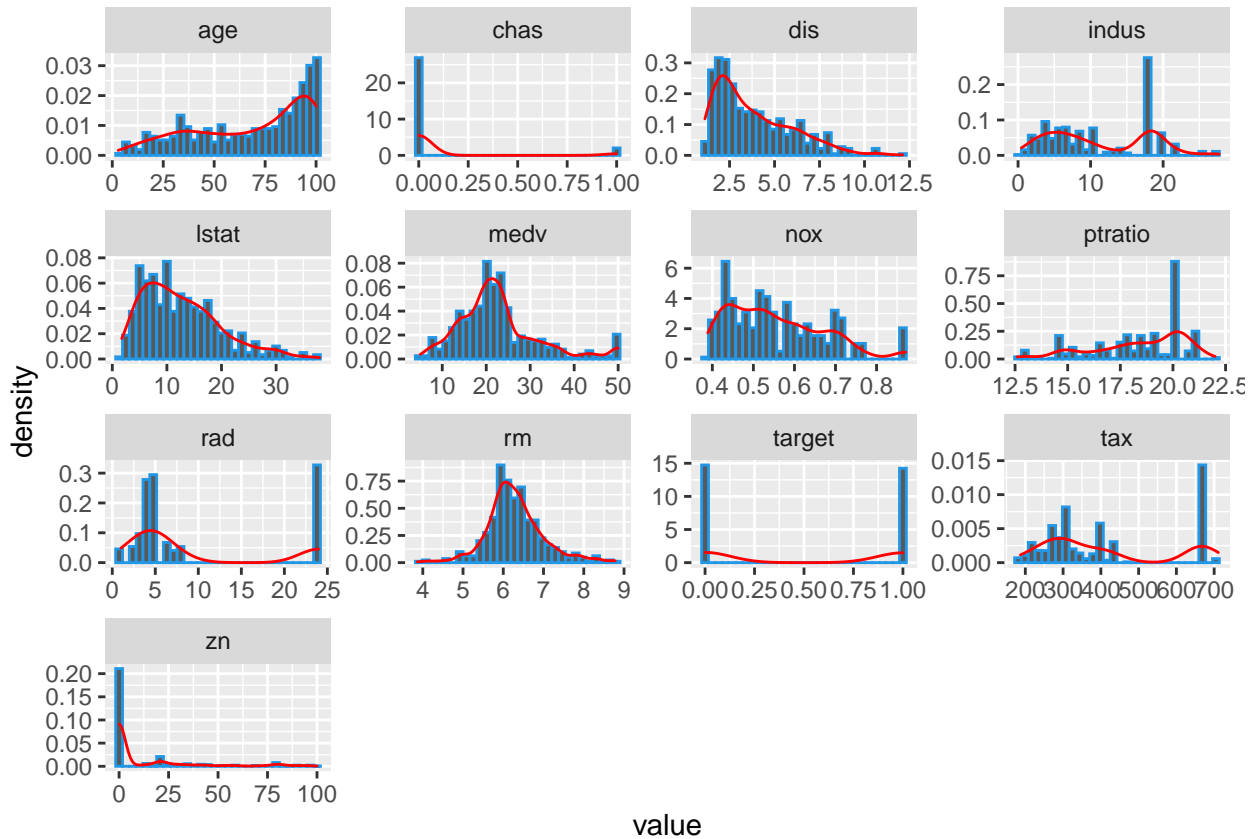
## Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
zn	0	1	11.58	23.36	0.00	0.00	0.00	16.25	100.00	
indus	0	1	11.11	6.85	0.46	5.15	9.69	18.10	27.74	
chas	0	1	0.07	0.26	0.00	0.00	0.00	0.00	1.00	
nox	0	1	0.55	0.12	0.39	0.45	0.54	0.62	0.87	
rm	0	1	6.29	0.70	3.86	5.89	6.21	6.63	8.78	
age	0	1	68.37	28.32	2.90	43.88	77.15	94.10	100.00	
dis	0	1	3.80	2.11	1.13	2.10	3.19	5.21	12.13	
rad	0	1	9.53	8.69	1.00	4.00	5.00	24.00	24.00	
tax	0	1	409.50	167.90	187.00	281.00	334.50	666.00	711.00	
ptratio	0	1	18.40	2.20	12.60	16.90	18.90	20.20	22.00	
lstat	0	1	12.63	7.10	1.73	7.04	11.35	16.93	37.97	
medv	0	1	22.59	9.24	5.00	17.02	21.20	25.00	50.00	
target	0	1	0.49	0.50	0.00	0.00	0.00	1.00	1.00	

From the description seen by the skim package we can observe we have two variables that should be transformed into factors since they have (1) or (0) values. **chas** & **target**.

## Distributions

We will first explore the data looking for issues or challenges (i.e. missing data, outliers, possible coding errors, multicollinearity, etc). Once we have a handle on the data, we will apply any necessary cleaning steps. Once we have a reasonable dataset to work with, we will build and evaluate three different Logistic models that predict seasonal wins.



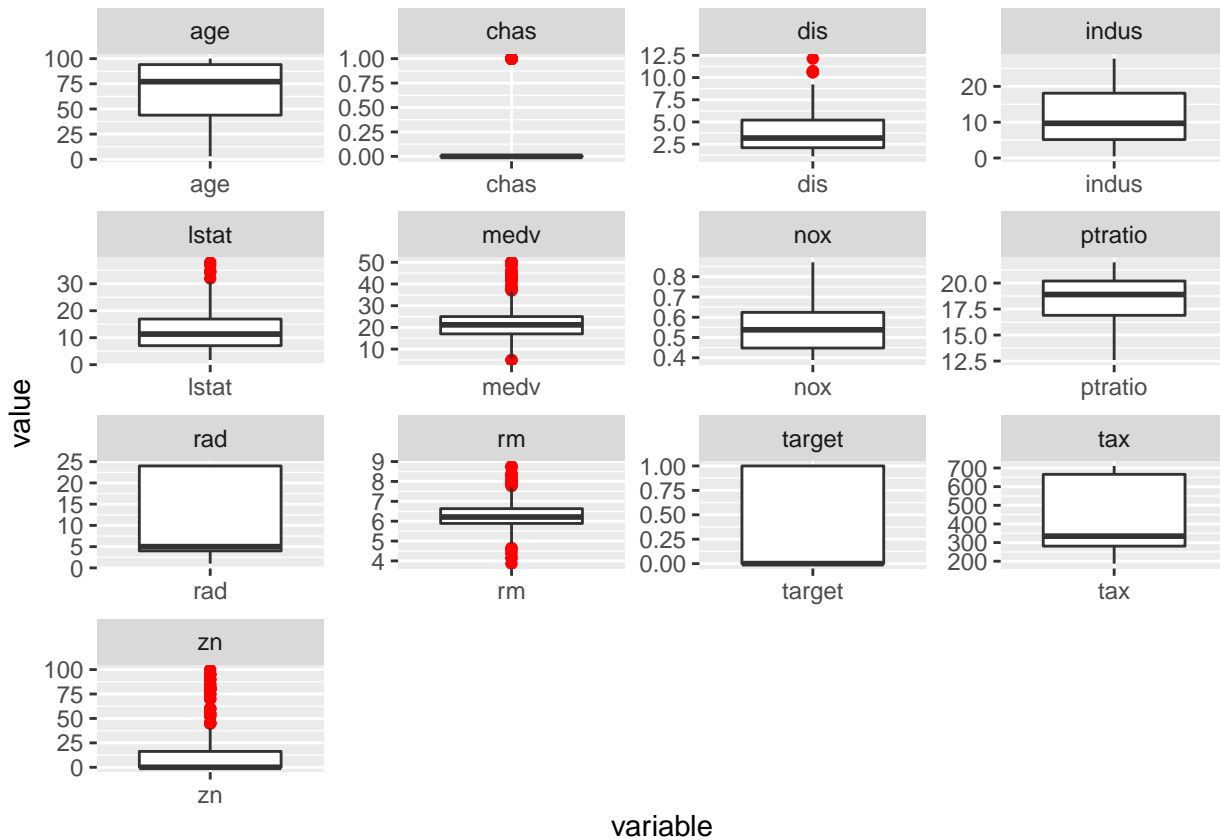
The distribution of our variables can also alert us of unusual patterns, in this case we have observed the prevalence of kurtosis for certain variables like: **nox**, **lstat**, **rad**, **zn** are skewed to the right. In addition, **ptratio** and **age** are left skewed.

After creating independent histograms for each variable we have found 2 variables that appear to be bi-modal. We notice that the graphs of this variables have two distinct humps or peaks with a valley separating them. We could attribute this observations to possibly different groups. We find that **rad** and **tax** are bi-modal.

## Outliers

In addition to histogram graph of our variable we thought it was pertinent to take a look at our variables using a boxplot. It will help us quickly visualize the distribution of the values in the dataset and see where the five number summary values are located.

In addition, we will be able to create a clear picture of the median values and the spreads across all the distributions. One of the most important observation we will obtain from this graph however, is outlier detection.



## Relationships

## 2. Data Preparation

### 3. Building Models

Model # 1



Model #2

Model #3

## 4. Select Models

Model\_1 Testing

## Model\_2 Testing

## Model\_3 Testing

**Final Selection**

**References**

**Apendix**