# DATA 621: BUSINESS ANALYTICS AND DATA MINING HOMEWORK#3: LOGISTIC REGRESSION

Group 2 - Gabriel Campos, Melissa Bowman, Alexander Khaykin, & Jennifer Abinette

Last edited October 26, 2023

## 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: 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)
- 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 0.5 threshold. Include your R statistical programming code in an Appendix.

## Write Up:

**1. DATA EXPLORATION (25 Points)** Describe the size and the variables in the crime training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas. a. Mean / Standard Deviation / Median b. Bar Chart or Box Plot of the data c. Is the data correlated to the target variable (or to other variables?) d. Are any of the variables missing and need to be imputed/"fixed"?

**2. DATA PREPARATION (25 Points)** Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations. a. Fix missing values (maybe with a Mean or Median value) b. Create flags to suggest if a variable was missing c. Transform data by putting it into buckets d. Mathematical transforms such as log or square root (or, use Box-Cox) e. Combine variables (such as ratios or adding or multiplying) to create new variables

**3. BUILD MODELS (25 Points)** Using the training data, build at least three different binary logistic regression models, using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Be sure to explain how you can make inferences from the model, as well as discuss other relevant model output. Discuss the coefficients in the models, do they make sense? Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

**4. SELECT MODELS (25 Points)** Decide on the criteria for selecting the best binary logistic regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model. * For the binary logistic regression model, will you use a metric such as log likelihood, AIC, ROC curve, etc.? Using the training data set, evaluate the binary logistic regression model based on (a) accuracy, (b) classification error rate, (c) precision, (d) sensitivity, (e) specificity, (f) F1 score, (g) AUC, and (h) confusion matrix. Make predictions using the evaluation data set

# Data Exploration

## Load the data

```
git_url<-
  "https://raw.githubusercontent.com/GitableGabe/Data621_Data/main/"
```

```
df_crime_eval <-
  read.csv(paste0(git_url,"crime-evaluation-data_modified.csv"))
head(df_crime_eval,n=10)
```

```
##    zn indus chas   nox    rm  age    dis rad tax ptratio lstat medv
## 1   0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8  4.03 34.7
## 2   0  8.14    0 0.538 6.096 84.5 4.4619   4 307    21.0 10.26 18.2
## 3   0  8.14    0 0.538 6.495 94.4 4.4547   4 307    21.0 12.80 18.4
## 4   0  8.14    0 0.538 5.950 82.0 3.9900   4 307    21.0 27.71 13.2
## 5   0  5.96    0 0.499 5.850 41.5 3.9342   5 279    19.2  8.77 21.0
## 6  25  5.13    0 0.453 5.741 66.2 7.2254   8 284    19.7 13.15 18.7
## 7  25  5.13    0 0.453 5.966 93.4 6.8185   8 284    19.7 14.44 16.0
## 8   0  4.49    0 0.449 6.630 56.1 4.4377   3 247    18.5  6.53 26.6
## 9   0  4.49    0 0.449 6.121 56.8 3.7476   3 247    18.5  8.44 22.2
## 10  0  2.89    0 0.445 6.163 69.6 3.4952   2 276    18.0 11.34 21.4
```

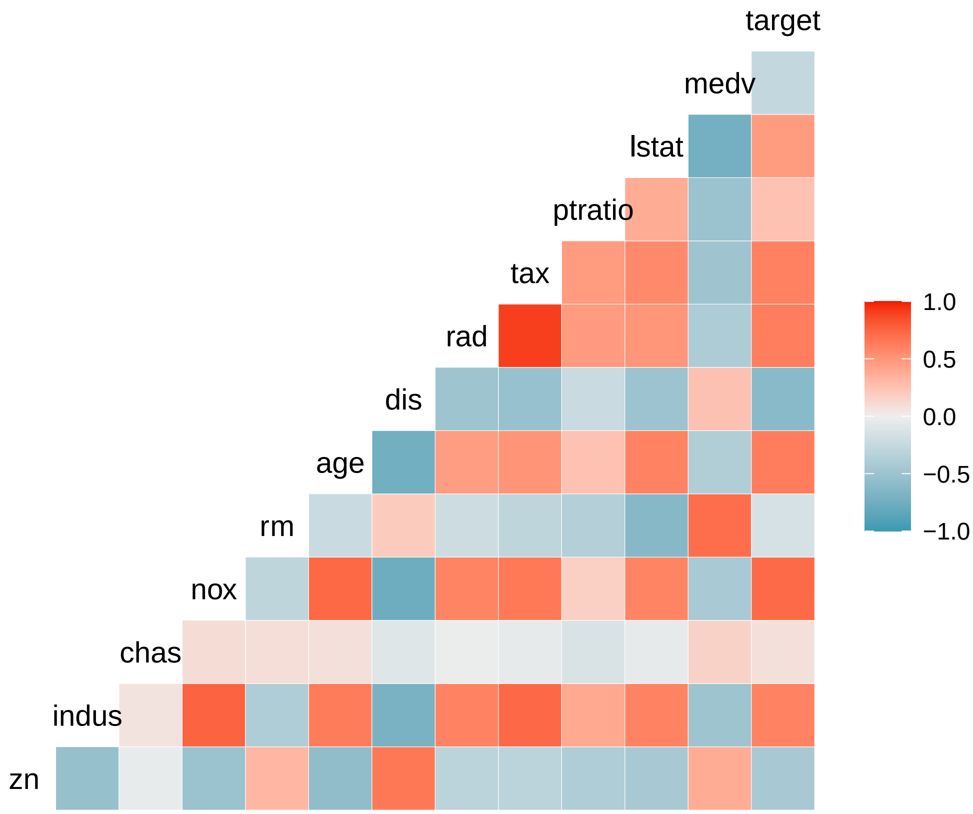
```
df_crime_train <-
  read.csv(paste0(git_url,"crime-training-data_modified.csv"))
head(df_crime_train,n=10)
```

```
##    zn indus chas   nox    rm   age    dis rad tax ptratio lstat medv target
## 1   0 19.58    0 0.605 7.929  96.2 2.0459   5 403    14.7  3.70 50.0      1
## 2   0 19.58    1 0.871 5.403 100.0 1.3216   5 403    14.7 26.82 13.4      1
## 3   0 18.10    0 0.740 6.485 100.0 1.9784  24 666    20.2 18.85 15.4      1
## 4  30  4.93    0 0.428 6.393   7.8 7.0355   6 300    16.6  5.19 23.7      0
## 5   0  2.46    0 0.488 7.155  92.2 2.7006   3 193    17.8  4.82 37.9      0
## 6   0  8.56    0 0.520 6.781  71.3 2.8561   5 384    20.9  7.67 26.5      0
## 7   0 18.10    0 0.693 5.453 100.0 1.4896  24 666    20.2 30.59  5.0      1
## 8   0 18.10    0 0.693 4.519 100.0 1.6582  24 666    20.2 36.98  7.0      1
## 9   0  5.19    0 0.515 6.316  38.1 6.4584   5 224    20.2  5.68 22.2      0
## 10 80  3.64    0 0.392 5.876  19.1 9.2203   1 315    16.4  9.25 20.9      0
```

## Pairwise correlation

Because all of the variable in the dataset are numeric, I can perform pairwise correlations to measure the strength of linearity among the variables in the training set.

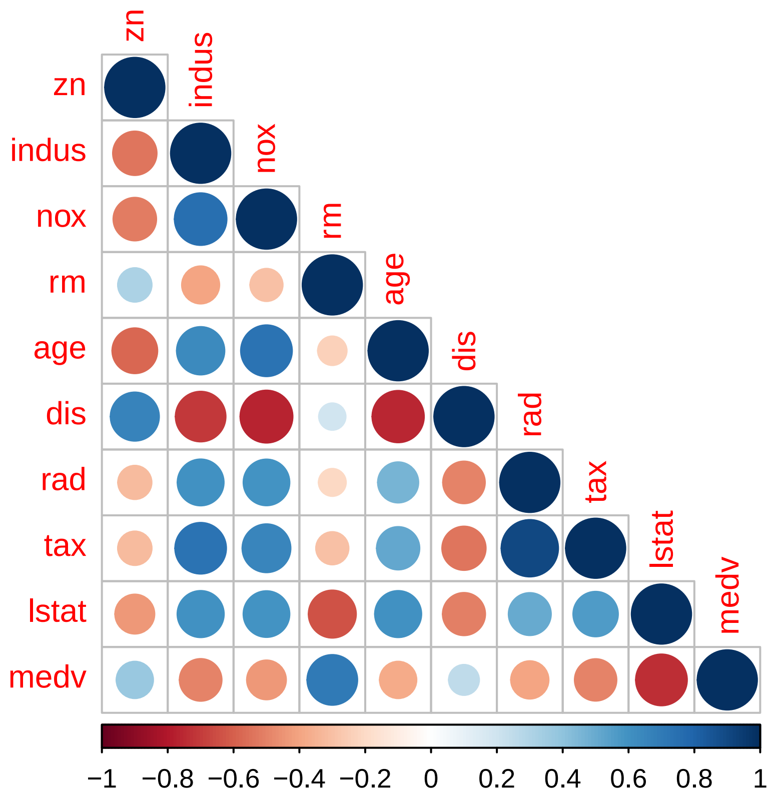
```
ggcorr(df_crime_train)
```




Correlation coefficients range from +1 to -1, where zero indicates no correlation. Initially, there appears to be modest to high correlations between the outcome *target* and *tax*, *rad*, *age*, *dis*, *nox*, and *indus*. There also appears to be some possible collinearity among some of the variables.

## Assessing multicollinearity

From the above correlogram the variable we do not need to worry about for collinearity are *target*, *chas*, *ptratio* and will not include them in the assessment for collinearity.




```
##          zn indus   nox    rm   age   dis   rad   tax lstat  medv
## zn       NA 0.000 0.000 0.016 0.000 0.000 0.001 0.001 0.001 0.006
## indus 0.000    NA 0.000 0.003 0.000 0.000 0.000 0.000 0.000 0.001
## nox   0.000 0.000    NA 0.007 0.000 0.000 0.000 0.000 0.000 0.002
## rm    0.016 0.003 0.007    NA 0.012 0.028 0.008 0.004 0.000 0.000
## age   0.000 0.000 0.000 0.012    NA 0.000 0.001 0.001 0.000 0.004
## dis   0.000 0.000 0.000 0.028 0.000    NA 0.001 0.000 0.001 0.011
## rad   0.001 0.000 0.000 0.008 0.001 0.001    NA 0.000 0.001 0.002
## tax   0.001 0.000 0.000 0.004 0.001 0.000 0.000    NA 0.000 0.001
## lstat 0.001 0.000 0.000 0.000 0.000 0.001 0.001 0.000    NA 0.000
## medv  0.006 0.001 0.002 0.000 0.004 0.011 0.002 0.001 0.000    NA
```

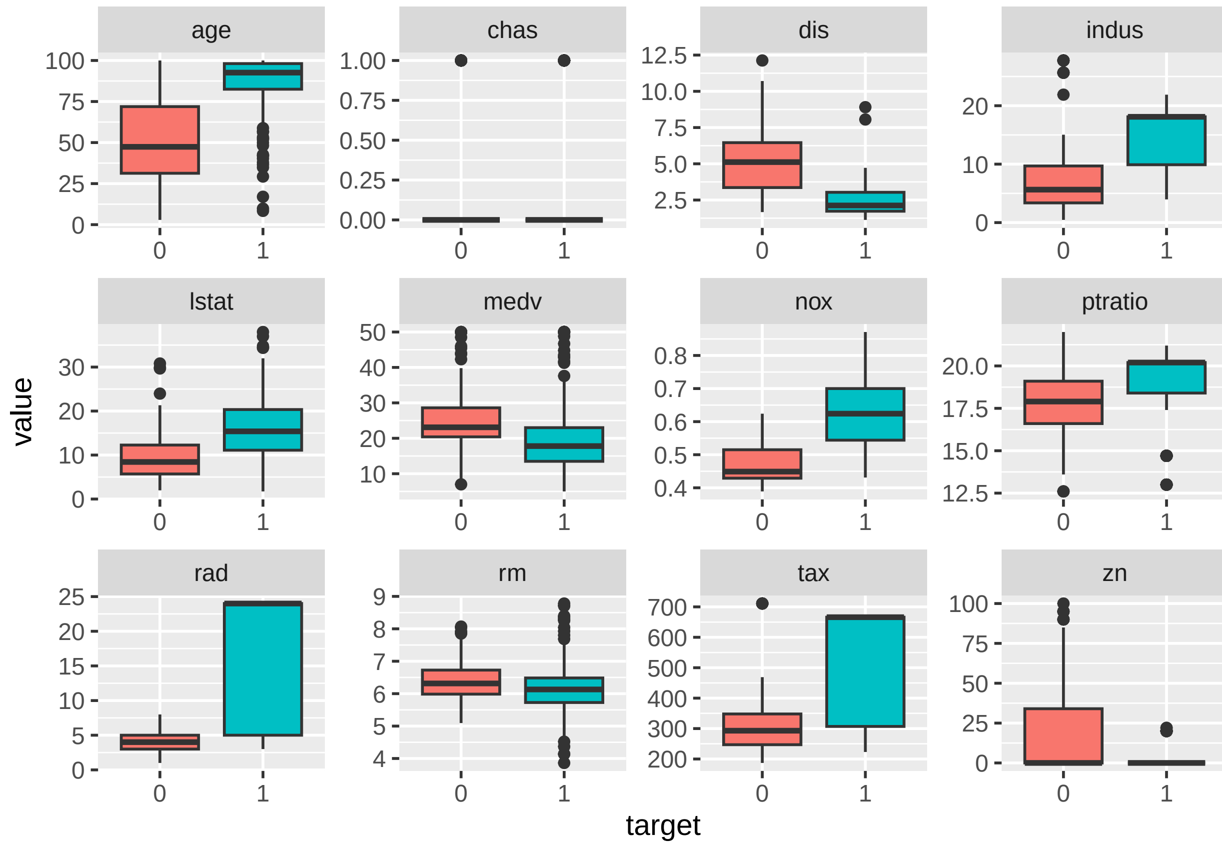
Unfortunately, each of these variable is significantly correlated with every other variable as evidenced of the matrix of p values. The correlogram suggests that *dis* is most highly correlated with with other variables in the dataset followed by *lstat* and *tax*.

### Relationship of each predictor to the *target*.

In order to best assess which predictors are likely to be informative and should thus be included in the full model to be tested we should also compare boxplots to look for predictors with low explanatory values.

```
df_crime_train %>%
  pivot_longer(cols = !target, names_to = "predictor", values_to = "value") %>%
  ggplot(aes(x = as.factor(target), y = value, fill = as.factor(target))) +
```

```r
  geom_boxplot(show.legend = FALSE) +
  xlab("target") +
  facet_wrap(~predictor, scales = "free")
```




The predictors that may have low explanatory values with *target* are *chas* and *zn*. Even though this is the case, we should include both in the model because they are not as highly correlated with other predictors like some of the other.

### Look for sample size differences between the two target groups

```r
df_crime_train %>%
  pivot_longer(cols = !target, names_to = "predictor", values_to = "value") %>%
  group_by(target) %>%
  count()
```

```
## # A tibble: 2 x 2
## # Groups:   target [2]
##   target     n
##    <int> <int>
## 1      0  2844
## 2      1  2748
```

*** Full model should include all variable except lstat *** Swap out lstat if tax in the final model.