Homework 3

Coffy Andrews-Guo, Krutika Patel, Alec McCabe, Ahmed Elsaeyed, Peter Phung

2022-11-01

Problem Statement and Goals

In this report, we generate a binary logistic regression model that is able to predict whether or not the crime rate for a neighborhood is above the median crime rate (1) or not (0). The independent and dependent variables that are used in order to generate this model use data from various neighborhoods of a major city. The analysis detailed in this report shows the testing of several models from which a best model was selected based on model performance and various metrics.

Data Exploration

The following is a summary of the variables provided within the data to generate the binary logistic regression model:

- `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)

A summary of the variables is shown below. See that within the summary, there does not seem to be any extremely high or extremely low values relative to the medians and means for each of the continuous predictor variables. The single binary predictor variable `chas` has reasonable values as well.

```
       zn             indus        chas         nox               rm       
 Min.   :  0.00   Min.   : 0.460   0:433   Min.   :0.3890   Min.   :3.863  
 1st Qu.:  0.00   1st Qu.: 5.145   1: 33   1st Qu.:0.4480   1st Qu.:5.887  
 Median :  0.00   Median : 9.690           Median :0.5380   Median :6.210  
 Mean   : 11.58   Mean   :11.105           Mean   :0.5543   Mean   :6.291  
 3rd Qu.: 16.25   3rd Qu.:18.100           3rd Qu.:0.6240   3rd Qu.:6.630  
 Max.   :100.00   Max.   :27.740           Max.   :0.8710   Max.   :8.780  
      age              dis              rad             tax       
 Min.   :  2.90   Min.   : 1.130   Min.   : 1.00   Min.   :187.0  
 1st Qu.: 43.88   1st Qu.: 2.101   1st Qu.: 4.00   1st Qu.:281.0  
 Median : 77.15   Median : 3.191   Median : 5.00   Median :334.5  
 Mean   : 68.37   Mean   : 3.796   Mean   : 9.53   Mean   :409.5  
```

```
 3rd Qu.: 94.10   3rd Qu.: 5.215   3rd Qu.:24.00   3rd Qu.:666.0
 Max.   :100.00   Max.   :12.127   Max.   :24.00   Max.   :711.0
    ptratio          lstat             medv        target
 Min.   :12.6   Min.   : 1.730   Min.   : 5.00   0:237
 1st Qu.:16.9   1st Qu.: 7.043   1st Qu.:17.02   1:229
 Median :18.9   Median :11.350   Median :21.20
 Mean   :18.4   Mean   :12.631   Mean   :22.59
 3rd Qu.:20.2   3rd Qu.:16.930   3rd Qu.:25.00
 Max.   :22.0   Max.   :37.970   Max.   :50.00
```

The multivariate plot distribution focus on the dependent variable, `target`, against other independent variables. All points are colored by `target`.

Figure 1: Binomial Distribution plots for each of the predictor variables in the dataset

`nox`, `tax`, and `zn` look separated between the factors. Therefore, we reasoned that these were good variables to add in a logistic regression model.

Figure 2 reveals that there are no missing values within the dataset. Therefore, no imputing is required for this dataset.

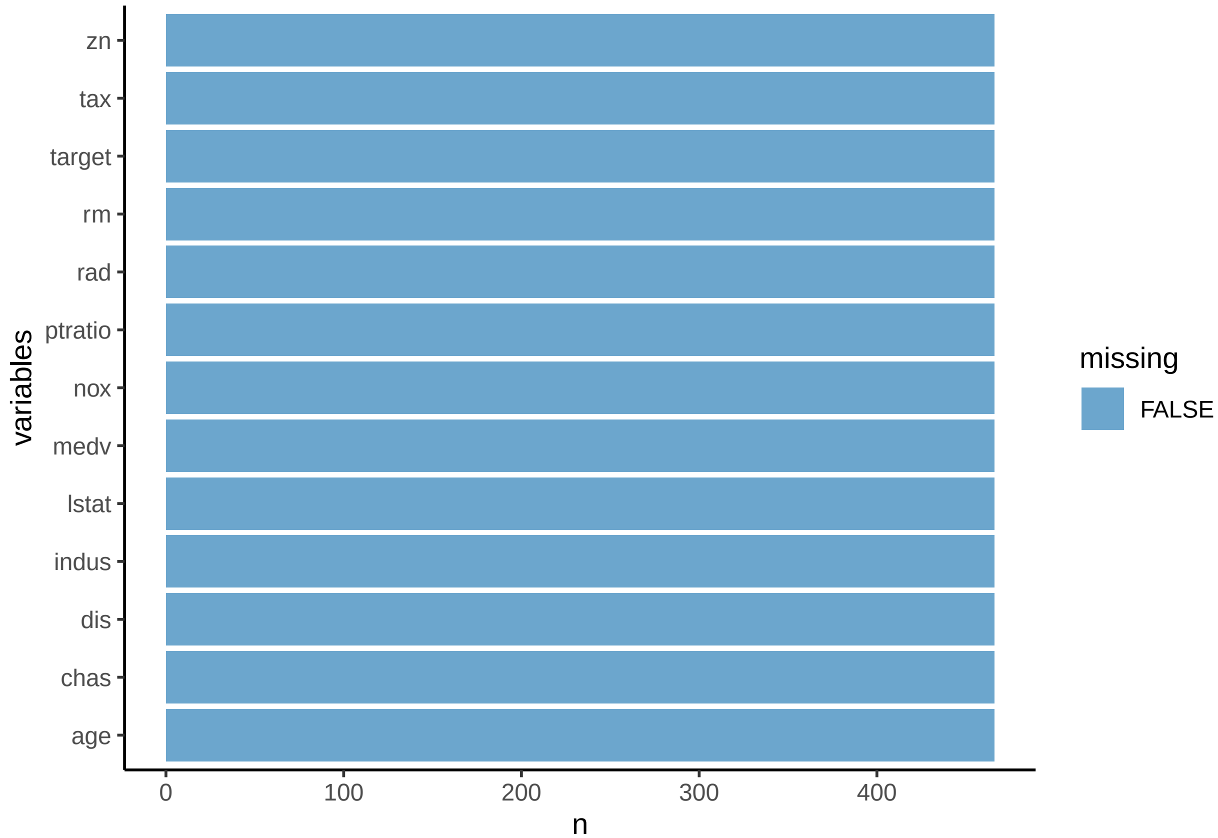

Figure 2: Chart showing the count of missing values for each of the variables in the dataset. Note that since there are no missing values, the legend only shows one item.

Outliers

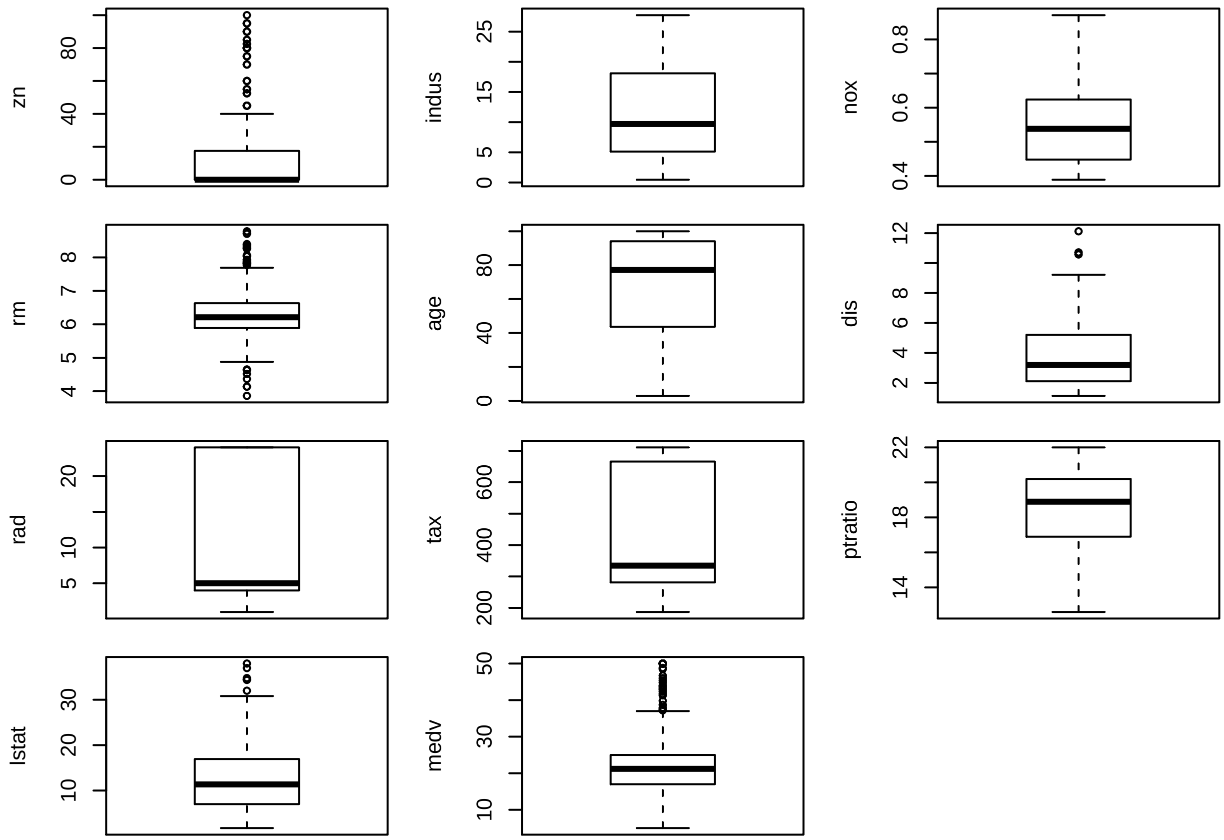

Figure 3: Box plots for each of the variables in the dataset.

Figure 3 shows boxplots for the continuous variables. While `zn`, `rm`, `dis`, `lstat` and `medv` contain outliers, the outliers in general do not seem to be any significant enough to affect the model greatly. However. note that `rad` and `tax` have significantly large interquartile ranges which indicates skewness.

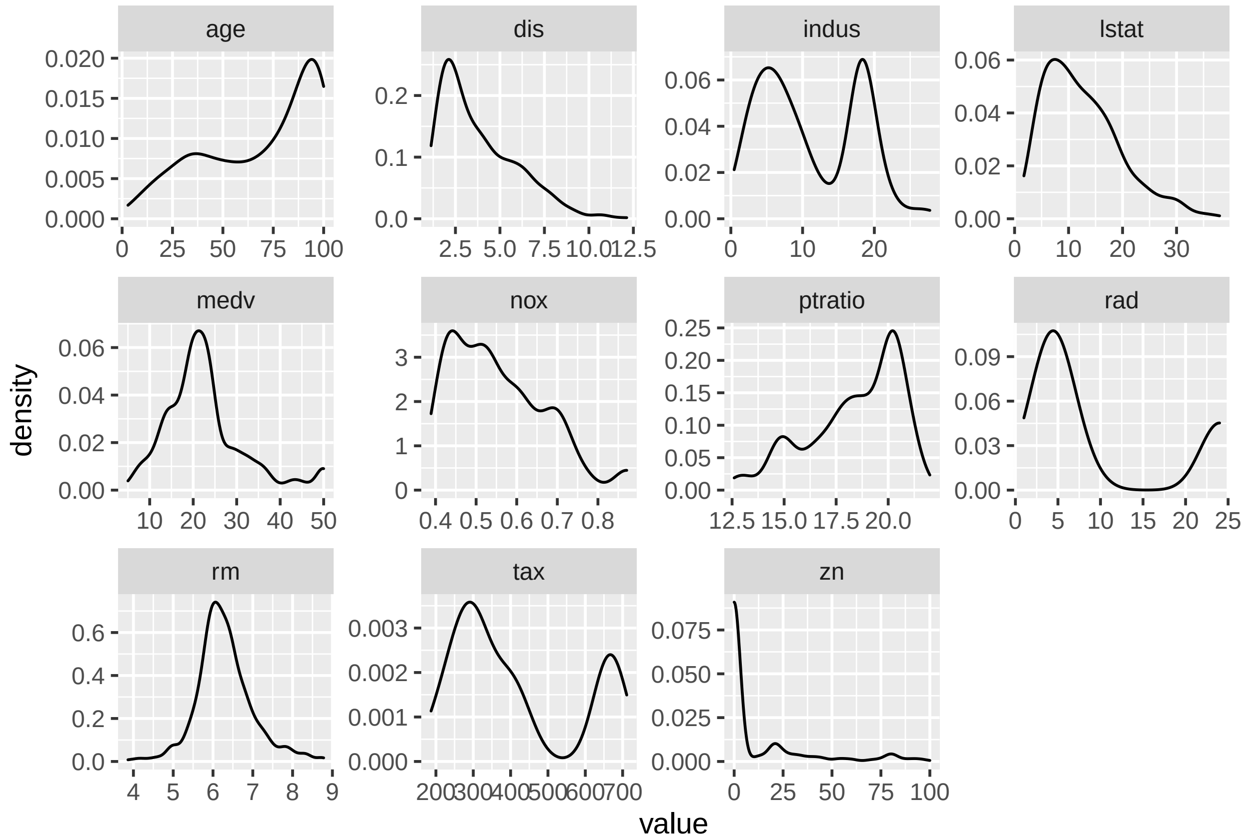

Figure 4: Density plots for continuous variables

Figure 4 reveals that `tax`, `indus`, and `rad` have bimodality. `age` appears to have bimodality as well but it is not as pronounced as the others. `rm` is relatively normally distributed while all of the other variables possess skewness, with `zn` possessing extreme skewness. Dummy variables for each of the bimodal variables were created and are given an explanation in the "Dealing with Bimodal Variables" section.

Examining Feature Multicollinearity

Finally, it is imperative to understand which features are correlated with each other in order to address and avoid multicollinearity within our models. By using a correlation plot, we can visualize the relationships between certain features. The correlation plot is only able to determine the correlation for continuous variables. There are methodologies to determine correlations for categorical variables (tetrachoric correlation). However there is only one binary predictor variable which is why the multicollinearity will only be considered for the continuous variables.

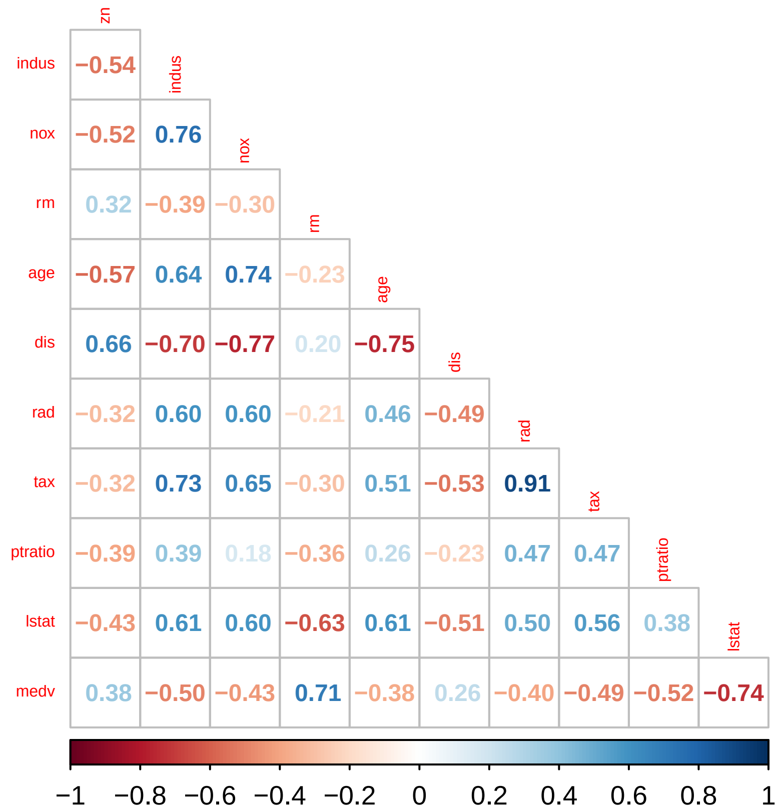

Figure 5: Multicollinearity plot for continuous predictor variables

Figure 5 reveals that `rad` and `tax` have an extremely high correlation of 0.91. What this indicates that there is a significant correlation between access to radial highways and property taxes. Therefore, the `tax` variable should be removed from the dataset because it has a higher p-value in the simple model to mitigate this high degree of correlation.

Data Preparation

Bimodal distributions in data are interesting, in that they represent features which actually contain multiple (2) inherent systems resulting in separated distributional peaks. Our approach to solving this is to create dummy variables representing which side of the local minimum each datapoint falls with respect to it's original bimodal distribution. Two new dummy variables were created for the two bimodal variables (`bi_indus` and `bi_rad`). The algorithm that was written to determine the local minimum was able to determine the local minimum for `indus` to be 12.70692. The algorithm was unable to detect a local minimum for `rad`. There is probably not enough information for the right peak for the algorithm to work properly. Nevetheless, we determined that a cutoff value of 15 for this variable would suffice. To summarize:

- `bi_indus`: 1 if `indus` is greater than 12.70692, 0 otherwise.
- `bi_rad`: 1 if `rad` is greater than 15, 0 otherwise.

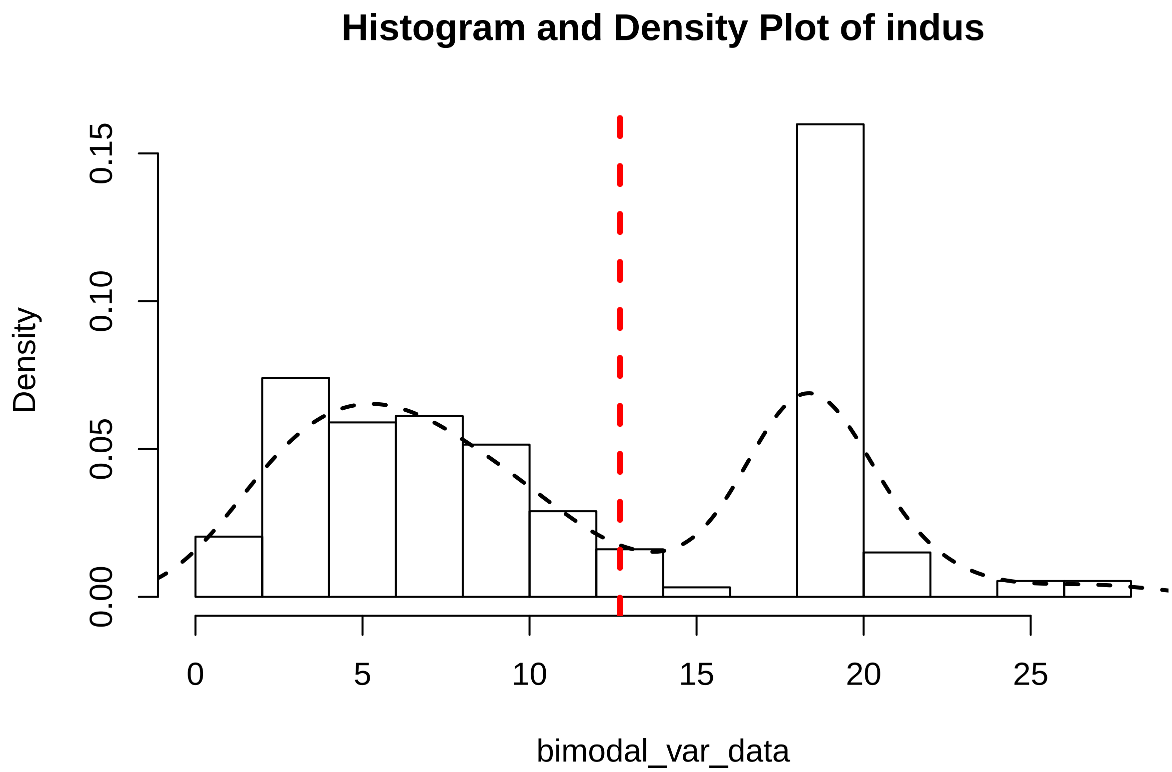

Figure 6: Histogram of `indus`. The dashed red line indicates the intersection point between two fitted histograms for this bimodal variable.

Skewed Variables

A Modern Approach to Regression with R explains the following:

When conducting a binary regression with a skewed predictor, it is often easiest to assess the need for x and log(x) by including them both in the model so that their relative contributions can be assessed directly.

The variables, `lstat`, `dis`, `age`, `nox`, and `ptratio` all exhibit skewness. Therefore, the logs of these variables were added into the dataset.

Split Data Into Testing and Training

The data was into testing and training subsets such that 60% of it will be used to train, and 40% to test. The first row shows the split for the testing data while the second row shows the split for the training data.

```
 0  1 
47 46 
```

```
  0   1 
190 183 
```

Build Models

Simple Model

A simple model was generated using all of the predictors and served as a baseline to which the other models were compared against.

```
Call:
glm(formula = target ~ ., family = binomial(link = "logit"),
    data = original_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9389  -0.1626  -0.0020   0.0045   3.5339

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -39.665958   7.495540  -5.292 1.21e-07 ***
zn           -0.066746   0.038842  -1.718 0.085721 .
indus        -0.056650   0.053192  -1.065 0.286877
chas1         1.517743   0.989835   1.533 0.125195
nox          47.423044   8.837988   5.366 8.06e-08 ***
rm           -0.594410   0.794213  -0.748 0.454203
age           0.042836   0.015437   2.775 0.005521 **
dis           0.800623   0.258283   3.100 0.001937 **
rad           0.604468   0.177767   3.400 0.000673 ***
tax          -0.005072   0.003296  -1.539 0.123786
ptratio       0.383569   0.141267   2.715 0.006624 **
lstat        -0.004275   0.059251  -0.072 0.942482
medv          0.166771   0.075104   2.221 0.026382 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 516.96  on 372  degrees of freedom
Residual deviance: 156.76  on 360  degrees of freedom
AIC: 182.76

Number of Fisher Scoring iterations: 9

Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

`nox` has an extremely low P-value. The U.S. Department of Housing indicates that low-income communities are much more likely than others to experience crime. The National Institute of Environmental Health Sciences indicates that poor communities are exposed to elevated pollution levels, which probably explains why there `nox` is statistically significant. Note that the `lstat`, `indus`, `rm` variables have a p-value greater than 0.05. We reasoned that if the skewed variables were transformed to a normal distribution, than the p-values could decrease, but the p-values actually increased further, thus negating the need to transform the variables.

Classification Matrix for Simple Model

The classification matrix for the simple model is provided below.

```
      Predicted
Actual  0  1
     0 46  1
     1  4 42
```

Model with Bimodal Variables and Log Transformed Variables

The following model includes the bimodal and log transformed variables in addition to the original variables that were used in the simple model.

```
Call:
glm(formula = target ~ ., family = binomial(link = "logit"),
    data = modified_train)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-7.928e-05  -2.100e-08  -2.100e-08   2.100e-08   9.849e-05

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.324e+03  1.063e+07   0.001    1.000
zn           1.244e+00  1.479e+03   0.001    0.999
indus       -4.711e+00  1.239e+04   0.000    1.000
chas1        5.673e+01  6.885e+05   0.000    1.000
nox         -9.678e+03  1.213e+07  -0.001    0.999
age         -6.422e+00  3.098e+03  -0.002    0.998
dis         -2.220e+02  9.932e+04  -0.002    0.998
rad          3.093e+01  1.034e+04   0.003    0.998
ptratio     -1.056e+02  1.102e+05  -0.001    0.999
lstat        4.663e+01  2.920e+04   0.002    0.999
bi_indus1    7.300e+01  8.555e+04   0.001    0.999
bi_rad1     -3.546e+02  1.943e+05  -0.002    0.999
log_lstat   -5.451e+02  3.313e+05  -0.002    0.999
log_dis      9.742e+02  3.450e+05   0.003    0.998
log_age      2.397e+02  1.499e+05   0.002    0.999
log_nox      7.679e+03  7.003e+06   0.001    0.999
log_ptratio  2.172e+03  2.136e+06   0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.2891e+02  on 92  degrees of freedom
Residual deviance: 4.8809e-08  on 76  degrees of freedom
AIC: 34

Number of Fisher Scoring iterations: 25
```

```
Setting levels: control = 0, case = 1
```

```
Setting direction: controls < cases
```

Classification Matrix for Model with Bimodal Variables and Log Transformed Variables

The classification matrix for the model with bimodal and log-transformed variables is provided below.

```
      Predicted
```

```
Actual   0   1
     0 155  35
     1  15 168
```

Negative Bimodal Model

We fitted a negaative binomial generalized linear model to the original dataset with bimodal and log transformed variables. The output of the model is shown below.

```
Call:
glm.nb(formula = as.numeric(target) ~ ., data = modified_train,
    init.theta = 637212.4506, link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.59269  -0.10125   0.01833   0.11233   0.42618

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) 20.8266020 22.2509813   0.936    0.349
zn           0.0001627  0.0078757   0.021    0.984
indus        0.0247362  0.0598662   0.413    0.679
chas1        0.0639719  0.4060549   0.158    0.875
nox         -7.4759301 13.1922271  -0.567    0.571
age          0.0034761  0.0179202   0.194    0.846
dis         -0.0268861  0.2338379  -0.115    0.908
rad          0.0690781  0.0818333   0.844    0.399
ptratio      0.3905143  0.8481515   0.460    0.645
lstat        0.0047497  0.0449358   0.106    0.916
bi_indus1   -0.2518111  0.7551382  -0.333    0.739
bi_rad1     -1.2036747  1.6410732  -0.733    0.463
log_lstat   -0.0177751  0.5704887  -0.031    0.975
log_dis      0.2506813  0.8780460   0.285    0.775
log_age     -0.1940858  0.8031829  -0.242    0.809
log_nox      5.4109855  8.7478082   0.619    0.536
log_ptratio -7.0051716 14.7082305  -0.476    0.634

(Dispersion parameter for Negative Binomial(637212.5) family taken to be 1)

    Null deviance: 15.8180  on 92  degrees of freedom
Residual deviance:  3.8296  on 76  degrees of freedom
AIC: 254.06

Number of Fisher Scoring iterations: 1

              Theta:  637212
          Std. Err.:  23976342
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -218.06

Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

Confusion Matrix for Negative Bimodal Model

The confusion matrix for the negative bimodal model is provided below.

```
      Predicted
Actual   1
     0 190
     1 183
```

Model with P-Values below 0.05

For this model, the predictor variables with p-values below 0.05 from the second model (Model with Bimodal Variables and Log Transformed Variables) were used and the output is shown below.

```
Call:
glm(formula = target ~ chas + nox + dis + age + rad, family = binomial(link = "logit"),
    data = modified_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.18588  -0.07468  -0.01053   0.00617   2.46007

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -60.12506   20.42928  -2.943  0.00325 **
chas1         1.90917    4.55293   0.419  0.67498
nox         100.42633   33.47472   3.000  0.00270 **
dis           0.95020    0.62362   1.524  0.12759
age          -0.02453    0.02356  -1.041  0.29779
rad           1.21797    0.46920   2.596  0.00944 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 128.91  on 92  degrees of freedom
Residual deviance:  27.88  on 87  degrees of freedom
AIC: 39.88

Number of Fisher Scoring iterations: 10

Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

Confusion Matrix for Model with P-Values below 0.05

The confusion matrix for the model with p-values below 0.05 is provided below.

```
      Predicted
Actual   0   1
     0 142  48
     1  10 173
```

Step AIC Model

Step AIC works by deselecting features that negatively affect the AIC. It selects the model with not only the best AIC score but also a model with less predictors than the full model, since the full model may have predictors that do not contribute or negatively contribute to the model's performance. The direction for the Step AIC algorithm was set to `both`, because this implements both forward and backward elimination in order to decide if a predictor negatively affects the model's performance. The original model was used in order to use the Step AIC algorithm in R and the output is shown below.

```
Call:
glm(formula = target ~ zn + chas + nox + age + dis + rad + tax +
    ptratio + medv, family = binomial(link = "logit"), data = original_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0211  -0.1945  -0.0019   0.0048   3.4965

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -38.861239   7.176433  -5.415 6.12e-08 ***
zn           -0.072028   0.036827  -1.956 0.050481 .
chas1         1.385959   0.952632   1.455 0.145704
nox          42.650611   7.541520   5.655 1.55e-08 ***
age           0.037653   0.012668   2.972 0.002955 **
dis           0.749390   0.246035   3.046 0.002320 **
rad           0.646840   0.168664   3.835 0.000126 ***
tax          -0.006429   0.002903  -2.215 0.026772 *
ptratio       0.345394   0.129844   2.660 0.007812 **
medv          0.121621   0.040889   2.974 0.002935 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 516.96  on 372  degrees of freedom
Residual deviance: 158.47  on 363  degrees of freedom
AIC: 178.47

Number of Fisher Scoring iterations: 9

Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

Confusion Matrix for Step AIC Model

The confusion matrix for the Step AIC model is provided below.

```
      Predicted
Actual  0  1
     0 39  8
     1  3 43
```

Model Selection

Model	Precision	Recall	AIC	AUC	F-score	Accuracy	Error
Simple	0.98	0.91	182.76	0.97	0.944	0.95	0.05
Transformed	0.83	0.92	34	0.94	0.87	0.87	0.13
Negative Bimodal	0.49	1	254.06	0.94	0.658	0.49	0.51
Reduced Transformed	0.78	0.95	39.88	0.95	0.856	0.84	0.16
Step AIC	0.84	0.93	178.47	0.97	0.887	0.88	0.12

Table 1: Model metrics

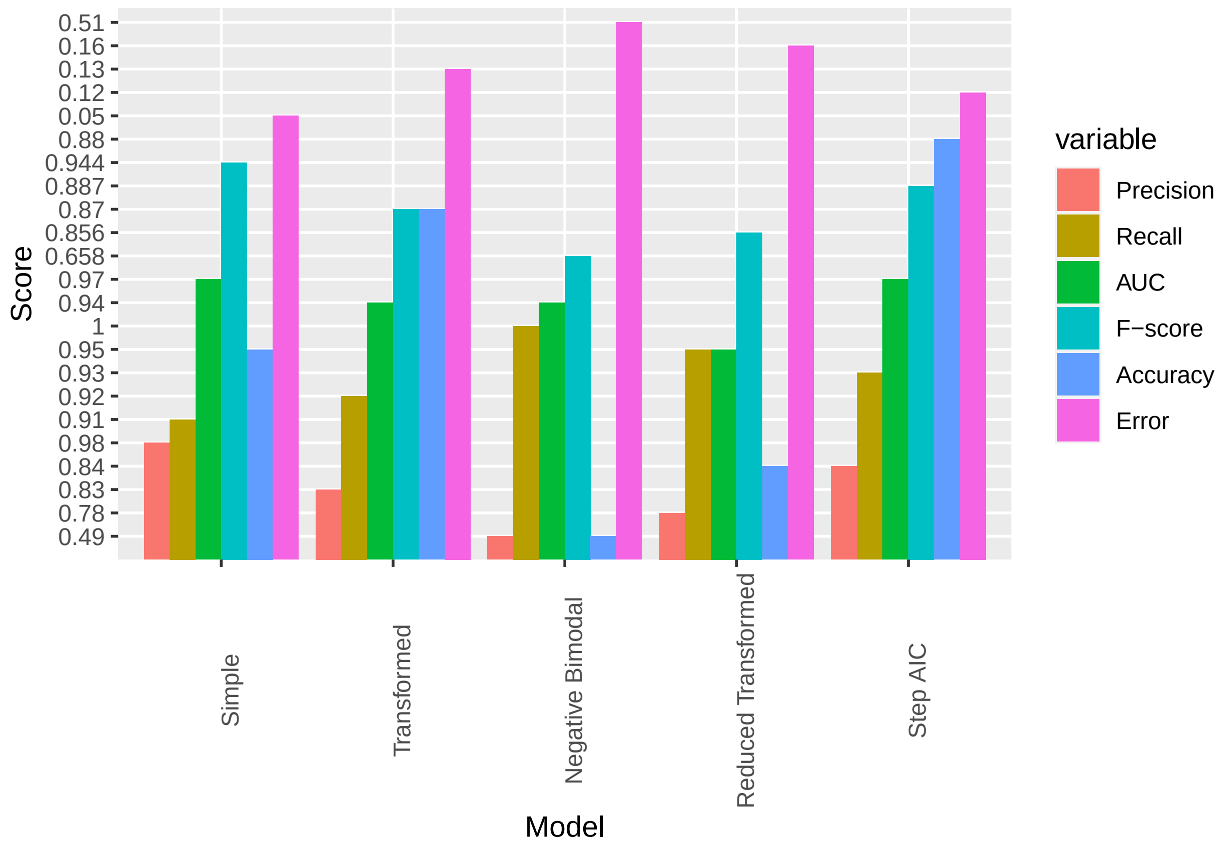

Figure 7: Bar chart of metrics for all 5 models

For this assignment, we will be choosing Model 2, which utilizes a bimodal distribution flag, as well as transformed data to address skewness. Based on Figure 7, we can also see that this model outperforms the rest in terms of one important metric: Recall (excluding model 3). When addressing a nation's city crime rates, it is important that whichever detection model is used classifies as many at-risk cities correctly as possible. It could be a major issue if an at-risk city was left unattended, and without aid.

Additionally, this model performs roughly as well in terms of precision and F-score to the simple model, while also using less predictor features. This will naturally reduce the cost of performing such an investigation, justifying the reduced precision. It is possible as well that, due to the small size of the dataset, that these values may not reflect the true predictive power of the discussed models.

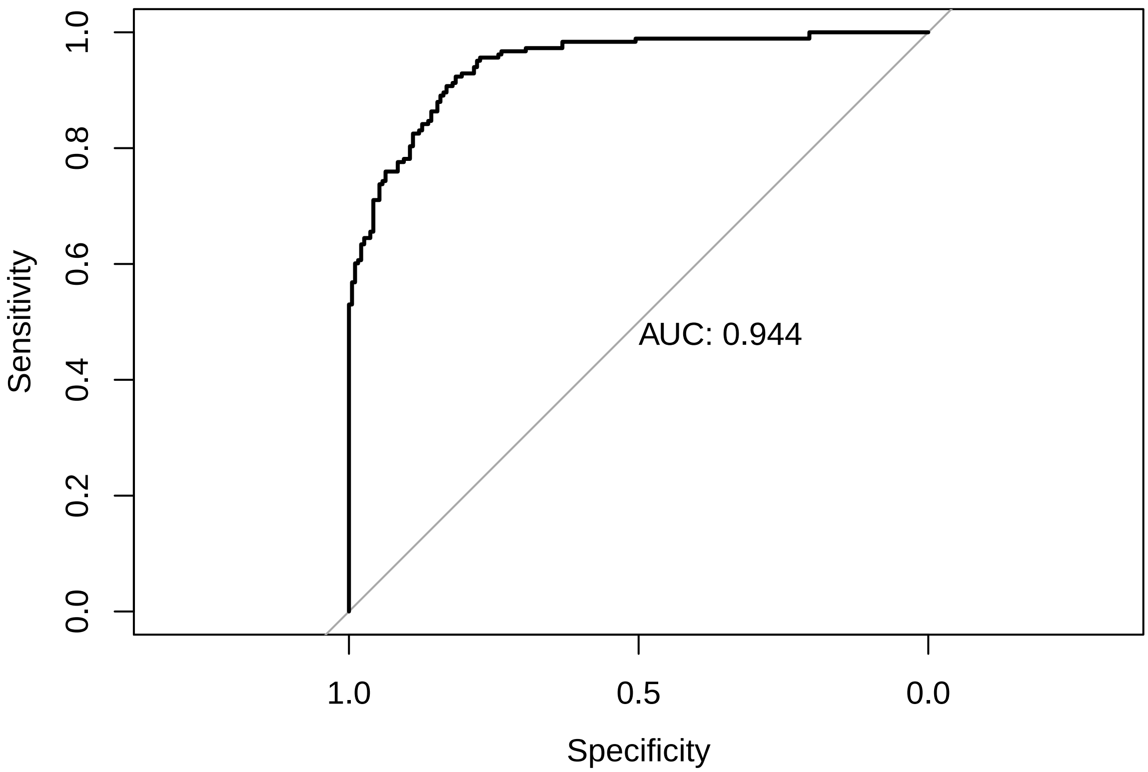

Figure 8: ROC Curve for selected model (Model with Bimodal Variables and Log Transformed Variables)

As we see on Figure 8, our model performs really well with an AUC of 0.947.

Works Cited

[1] What is Cook's Distance? (StatisticsHowTo): https://www.statisticshowto.com/cooks-distance/

[2] How to Calculate Correlation Between Categorical Variables (Statology): https://www.statology.org/correlation-between-categorical-variables/

[3] The 6 Assumptions of Logistic Regression (Statology): https://www.statology.org/assumptions-of-logistic-regression/

[4] Neighborhoods and Violent Crime (U.S. Department of Housing) https://www.huduser.gov/portal/periodicals/em/summer16/highlight2.html

[5] Poor Communities Exposed to Elevated Air Pollution Levels https://www.niehs.nih.gov/research/programs/geh/geh_newsletter/2016/4/spotlight/poor_communities_exposed_to_elevated_air_pollution_levels.cfm

[6] Logistic Regression Assumptions (Kenneth Leung): https://github.com/kennethleungty/Logistic-Regression-Assumptions/blob/main/Box-Tidwell-Test-in-R.ipynb

[7] Logistic Regression Assumptions and Diagnostics in R (STHDA): http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/