DATA 621 Homework 4 - Binary Logistic Regression

Erik Nylander

March 23, 2016

1 Data Exploration

In this analysis we will be looking at crime data for various neighborhoods of Boston Massachusetts . Each of 466 observations of consists of the measurements of 14 various predictor values for neighborhoods in the city. The data set also contains and evaluation set that contains 40 observations with the target variable removed. To facilitate the model selection we will be diving the training data set into a training set (70%) and a test set (40%).

1.1 Explanation of the Variables

- Target Variable
  - target - a variable that takes on (1) if the crime rate is above the median or a (0) if it's not.
- Predictor Variables
  - zn: proportion of residential land zoned for large lots (over 25000 ft^2)
  - indus: proportion of non-retail business acres
  - chas: dummy variable for bordering the Charles River (1) or not (0)
  - nox: nitrogen oxides concentration (parts per 10 million)
  - rm: average number of rooms per dwelling
  - age: proportion of owner-occupied units built before 1940
  - dis: weighted mean of distances to five Boston employment centers
  - rad: index of accessibility to radial highways
  - tax: full-value property-tax rate per $10,000
  - ptratio: pupil-teacher ratio by town
  - black: $1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town
  - lstat: lower status of the population (percent)
  - medv: median value of the owner occupied homes in $1000s

1.2 Exploring the *target* Variable

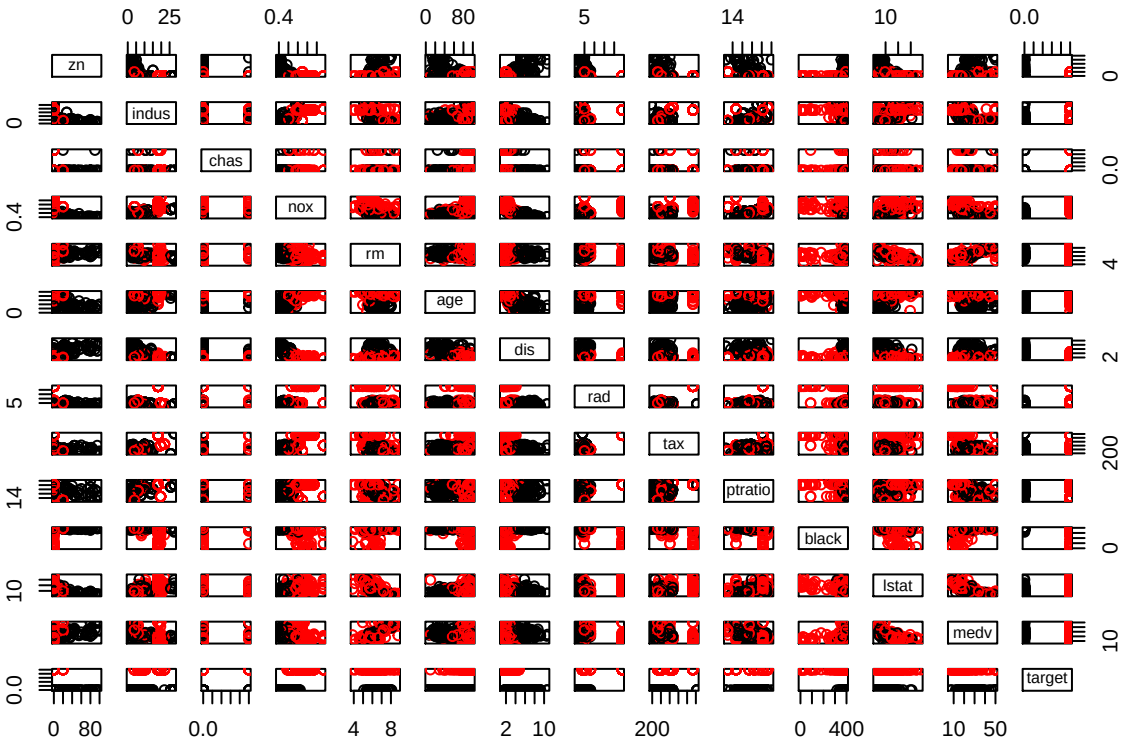
In the logistic regression model we have a binary target variable that we would like to predict. For this data set that target variable is a 1 if the crime rate in a neighborhood is above the median crime rate and a 0 if the crime rate is below the median crime rate. Before we along with the analysis we first want to take a look at the prevalence of the two values of the target variable. Table 1 contains the results. We see that the number of lower crime neighborhoods and higher crime neighborhoods are fairly equally distributed which is what we would expect using a median measure.

Table 1: Prevalence of the target Variable in the Data

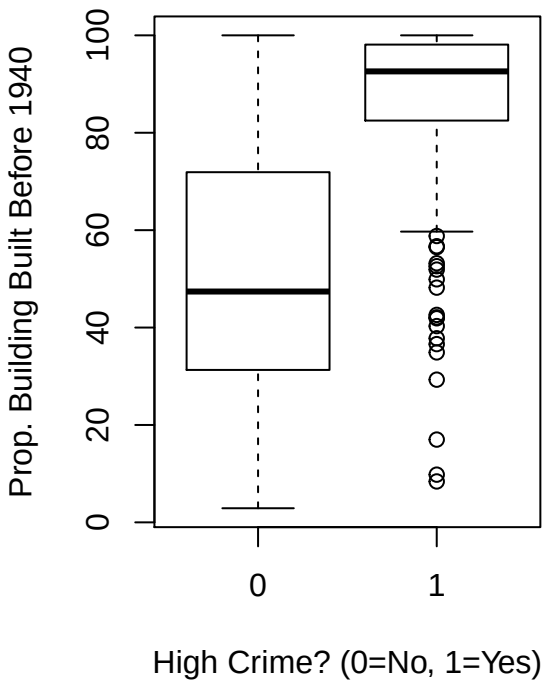
	Low Crime	High Crime
Frequency	237	229
Proportion	0.5086	0.4914

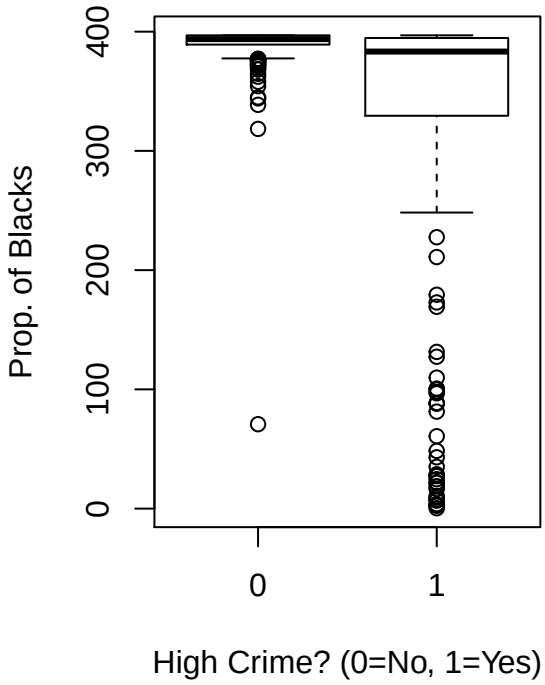
1.3 Exploring the Predictor Variable

In this data set there are a few features that we would like to look for. The first is too look at the different predictor variables graphed against each other. The figure below shows the pairs plot of the different elements. We have elected to color the plot by our target variable. The plot is rather hectic but we note that there are some groupings that appear to be correlated and we also notice that for some of our predictor variables we see value of the target is correlated with the value of the predictor variable.



The next concern that we have with a logistic relation is that we want to determine if any of the predictor variables are highly skewed. If they are we will discuss how to handle these in the section 2. To investigate this we elected to create box-plots of all of the variables against the target variable. This allows us to further explore differences in the predictor variables that are related to the target variable and allows us to explore the skew. In the figure below we have included the resulting boxplots for two of our variables that we will need to address. The remainder of the boxplots can be seen in Appendix A. We do note that there are some distinct difference between various predictor variables based on the target variable of crime rate. This indicates that these predictor variables may be important in predicting the target variable. The most notable of these is the Nitrogen Oxide Concentration which may be vary important for our final model.





2 Data Preperation

The data set that we have is in very good shape and for the most part it contains values for all of the predictor variables for each of the neighborhoods. We also do not notice any observations that are outside of the norm. Given that many of the predictor variables are proportions we would expect that they are fairly well behaved. The only issues that we found in the data are noted in the two boxplots above. The proportion of of buildings built before 1940 shows a high skew in the higher crime neighborhoods. We also note that the calculation for the number of blacks in a given neighborhood also show a strong skew for both of the low crime and high crime neighborhoods. To work with this we will include a log(age) variable and log(black) variable in our first model.

3 Building the Models

Now that we have explored our predictor and target variables we will now construct our models. For this analysis we will be constructing three different models. The first model will be a model including all of the variable and the log(age) and log(black) variables. Our second model will be the model including all of the given predictor variables. Our last model will include a subset of the predictor variables that have larger p-values as determined by the 1st model. We will perform a quick assessment of these models based on their confusion matrices and AIC values. We will also perform a Chi-squared ANOVA test to check and see if each of the new models is has the potential to be true.

3.1 Full Model with logs

In our first model we will build a model that contains all of the predictor variables along with the log(age) and log(black) variables. In a logistic regression we would expect that this model will have the largest predictive power. The results of fitting this model are in Table 2 below. Starred values indicate that the variable was significant at the $\alpha = 0.05$ level.

Table 2: Full Logistic Model with Logs

Variable	Coefficient	Factor of Odds
Intercept	-47.835	0
zn	-0.132	0.876
indus	-0.203*	0.816
chas	2.159	8.663
nox	67.538*	$2.14 \text{x} 10^{29}$
rm	-1.061	0.346
age	0.135*	1.145
dis	0.901*	2.461
rad	0.864*	2.372
tax	-0.012*	0.988
ptratio	0.641*	1.899
black	-0.074*	0.928
lstat	0.026	1.026
medv	0.2*	1.222
log(age)	-4.126*	0.016
log(black)	6.002	404.3

From Table 2, we can interpret the results as follows, if we take a look tax value, it has a coefficient of -0.012, we can then interpret this as having a multiplicative effect on the odds. If all other variables are held constant then a one unit increase in the tax rate ($10,00 increase) will change the odds of being in a high crime neighborhood by $e^{-0.012}$ or 0.988. We can interpret this as decrease of 0.12% in the probability.

We do notice some interesting results in this model. The presence of nitrous oxide (nox) in the environment seems to have a massive effect on the probability of a neighborhood being included in the high crime group. We also see some unusual things like the large coefficients for the the location relative to the Charles River (chas) and the log(black) that are not significant. There may be a case of this model over fitting our data and leading to some of these odd results.
We will do a quick assessment of the model on the training data to see if we should even test it further. In Table 3 we see the confusion matrix for the model on the training data. From the confusion matrix we see that we have an accuracy rate of 0.9264. We also note that our model generated an AIC of 138.46 which we can use to compare to out other two models. This model seems to do a very good job on our training data and we will continue to evaluate it further.

Table 3: Confusion Matrix of Predicted Values (log model)

Class	Pred. Low	Pred. High
True Low	154	14
True High	10	148

3.2 Full Model

The second model that we will construct is the full model including all of the given predictor variables. We would expect that this model will perform slightly worse than the first model given that it includes less variables. Since this model is a nested version of the first model we can use an ANOVA test to compare these two models at we determine its viability on the training data set. The results of fitting this model are in Table 4 below. Starred values indicate that the variable was significant at the $\alpha = 0.05$ level.

Table 4: Full Logistic Model

Variable	Coefficient	Factor of Odds
Intercept	-28.685*	~0
zn	-0.129	0.879
indus	-0.184*	0.832
chas	1.925	6.859
nox	63.683*	$4.54 \text{x} 10^{27}$
rm	-0.719	0.487
age	0.054*	1.056
dis	0.853*	2.347
rad	0.773*	2.166
tax	-0.012*	0.989
ptratio	0.588*	1.801
black	-0.059*	0.943
lstat	0.067	1.069
medv	0.19	1.208

From Table 4 we note that many of the coefficients and odds values that we saw in the first model. We also note that many of the coefficients stayed relatively stable with reduction of the two models. We also note that we still have a very large impact of nitrous oxide (nox) and that we have a still have large coefficients for the proximity to the Charles River but it is still not a significant predictor.

We will do a quick assessment of the model as well using the training data to see if we should test it further. In Table 5 we see the confusion matrix for the model on the training data. From the confusion matrix we see that we have an accuracy rate of 0.9264, the same as the first model. We also note that our model generated an AIC of 139.6 which compares very closely to the previous model and implies that the two models may be indistinguishable. This model seems to do a very good job on our training data and we will continue to evaluate it further.

Table 5: Confusion Matrix of Predicted Values (full model)

Class	Pred. Low	Pred. High
True Low	153	13
True High	11	149

3.3 Reduced Model

For our final model we elected to remove the variables that were not significant in the log and full models or showed a potentially high level of correlation as seen in the pairs plot. Using these criteria we decided to remove the following variables. The proportion of residential land zoned for large lots (zn), the average number of rooms per dwelling (rm), the lower status of the population (lstat), and the log(black) variable. We did elect to leave in the dummy variable for the proximity for the Charles River (chas) in the model. This variable was significant at the 0.1 level for both models and it was one of the few geographic variables

that we had in the model. The results of running this reduced model can be seen in Table 6. Starred values indicate that the variable was significant at the $\alpha = 0.05$ level.

Table 6: Full Logistic Model with Logs

Variable	Coefficient	Factor of Odds
Intercept	-17.335	~0
indus	-0.196*	0.822
chas	2.784*	16.619
nox	63.081*	$2.49 \text{x} 10^{27}$
age	0.12*	1.128
dis	0.533*	1.704
rad	0.749*	2.115
tax	-0.012*	0.988
ptratio	0.663*	1.941
black	-0.062*	0.94
medv	0.01*	1.105
log(age)	-3.933*	0.019

Once again we see that the coefficients stay very consistent and we can make the same interpretations from this smaller model that we made from the larger models. We still have a very large effect from nitrous oxide (nox) concentrations and we also now see a large value from the proximity to the Charles River.

We will do one more quick assessment of the model using the training data to see if we should test it further. In Table 7 we see the confusion matrix for the model on the training data. From the confusion matrix we see that we have an accuracy rate of 0.9202, just slightly worse then the larger models. We also note that our model generated an AIC of 137.56 which compares very closely to the previous model and implies that the three models may be indistinguishable. This model seems to do a very good job on our training data and we will continue to evaluate it further.

Table 7: Confusion Matrix of Predicted Values (reduced model)

Class	Pred. Low	Pred. High
True Low	155	17
True High	9	145

4 Selecting the Models

Given that we have three models with high degrees of accuracy on the training data we will now evaluate all three of the models using the test data that we generated from the training data set. We will evaluate these models by calculating the log likelihood, performing an an ANOVA on the reduced models against the full model including logs, calculating the area under the curve (AUC) and measuring the models accuracy, classification error rate, precision, sensitivity, specificity, F1 score, and confusion matrix. Once we have these results we will then select one of the models and predict the probability of of being in the higher crime classification on the evaluation data set.

4.1 Full Model with Logs

The first model that we will be evaluating is the full model including the logs. Table 8 shows the confusion matrix generated on the test data. Table 9 contains all of the measurements that we will use to evaluate the models. We notice several things in the evaluation of the model. The first is that the confusion matrix does a

very nice job of predicting the true lower crime and higher crime neighborhoods in the test data set. The second is that model has a high level of sensitivity and specificity, couple that with a high accuracy and good AUC, and we are favor this model. Given the solid performance of this model on the test data we will use this model as our benchmark for the other two models.

Table 8: Confusion Matrix of Predicted Values (full-log model)

Class	Pred. Low	Pred. High
True Low	69	10
True High	4	57

Table 9: Summary of Classification Metrics (full-log model)

Metric	Value
Accuracy	0.9
CER	0.1
Precision	0.9344
Sensitivity	0.8507
Specificity	0.9452
F1 Score	0.8906
AUC	0.9706
Log Likelihood	-53.23

4.2 Full Model

The next model that we will evaluate is our full model with all of the given predictor values. Table 10 shows the confusion matrix generated on the test data. Table 11 contains all of the measurements that we will use to evaluate the models. We notice that our full model also does an excellent job of predicting the values. We also notice the that full model is more accurate than the full-log model but it does sacrifice slightly in the other measures.

Table 10: Confusion Matrix of Predicted Values (full model)

Class	Pred. Low	Pred. High
True Low	68	8
True High	5	59

Table 11: Summary of Classification Metrics (full model)

Metric	Value
Accuracy	0.9071
CER	0.0929
Precision	0.9219
Sensitivity	0.8806
Specificity	0.9315
F1 Score	0.9008
AUC	0.9677
Log Likelihood	-55.8

Given how similar the two models are to each other in their performance on the test data we now conduct a Chi-Squared ANOVA hypothesis test to check for differences in the two models with $\alpha = 0.05$. We set up the hypothesis test as follows.
H_0: The full model is true.
H_A: The full model is not true and we prefer the full-log model.

The results of the hypothesis test are $p = 0.07634$ on df=2. We would then fail to reject the null hypothesis and an assume that our full model is true and could be used to predict on our evaluation data set. Next we will check our reduced model.

4.3 Reduced Model

The final model that we will testing is our reduced model. Table 12 shows the confusion matrix generated on the test data. Table 13 contains all of the measurements that we will use to evaluate the models. One of the things that we note from the confusion matrix on both the training and test data is that this model has a high level of specificity and is much less likely to predict that a neighborhood is lower crime when it is actually higher crime. Depending on the use case for the model we may prefer this even over the accuracy of the model. We also note that once again the model does a very good job of predicting the true lower and higher crime neighborhoods. The model also has very good performance on all of our measures and compares favorably to our other two models.

Table 12: Confusion Matrix of Predicted Values (reduced model)

Class	Pred. Low	Pred. High
True Low	70	11
True High	3	56

Table 13: Summary of Classification Metrics (reduced model)

Metric	Value
Accuracy	0.9
CER	0.1
Precision	0.9492
Sensitivity	0.8358
Specificity	0.9589
F1 Score	0.8889
AUC	0.9681
Log Likelihood	-56.78

Once again, given how similar this model is to the previous two models we will construct a hypothesis test against our current favored model, the full model, against our reduced model. We set up the Chi-squared ANOVA hypothesis at an $\alpha = 0.05$ level as follows.
H_0: The reduced model is true.
H_A: The reduced model is not true and we prefer the full model.

The results of the hypothesis test are $p = 0.3768$ at df $= 2$. We do not have enough evidence to reject the null hypothesis and conclude that the reduced model is a true model for our data set.

4.4 Selecting our Model and Evaluating the Data

Given the results from the above analysis we have decided to select our reduced model for predicting the probability of a neighborhood being in the lower crime or higher crime category. Ultimately this model compares favorably to the full-log and full models. In each of the evaluation criterion that we look at this model performs similarly to the other two models and therefore get selected based on it being more parsimonious and our bias to having a model that has a high specificity.

Now that we have selected our model we will make our predictions on the 40 neighborhoods that were included in the evaluation data set. Table 14 shows the first 6 values for the predicted neighborhoods. In Table 15 we include a summary table of the number of predicted lower crime neighborhoods and higher crime neighborhoods.

Table 14: First 6 Predicted Neighborhoods(0 = Low Crime, 1 = High Crime)

Neighborhood	1	2	3	4	6	7
Probability	0.003	0.847	0.881	0.999	0.021	0.215
Category	0	1	1	1	0	0

Table 15: Prediction of the target Variable in the Data

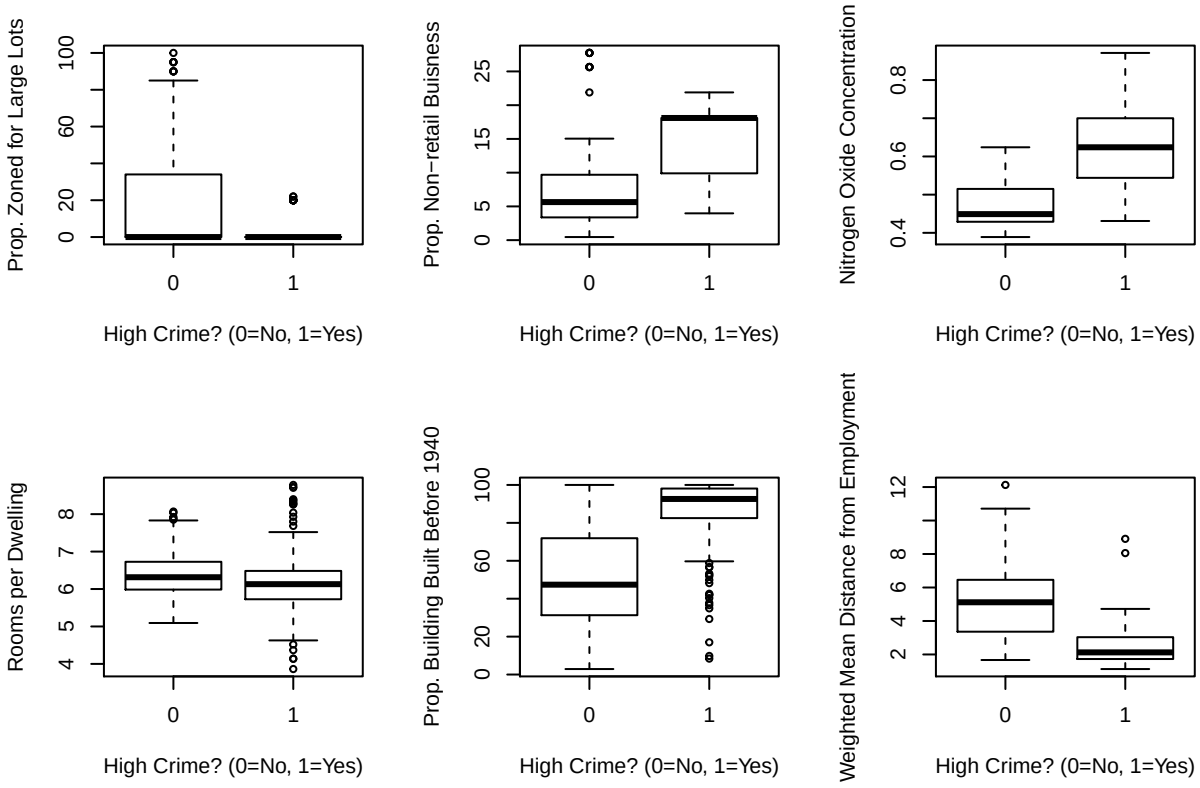
	Low Crime	High Crime
Frequency	17	23
Proportion	0.425	0.575

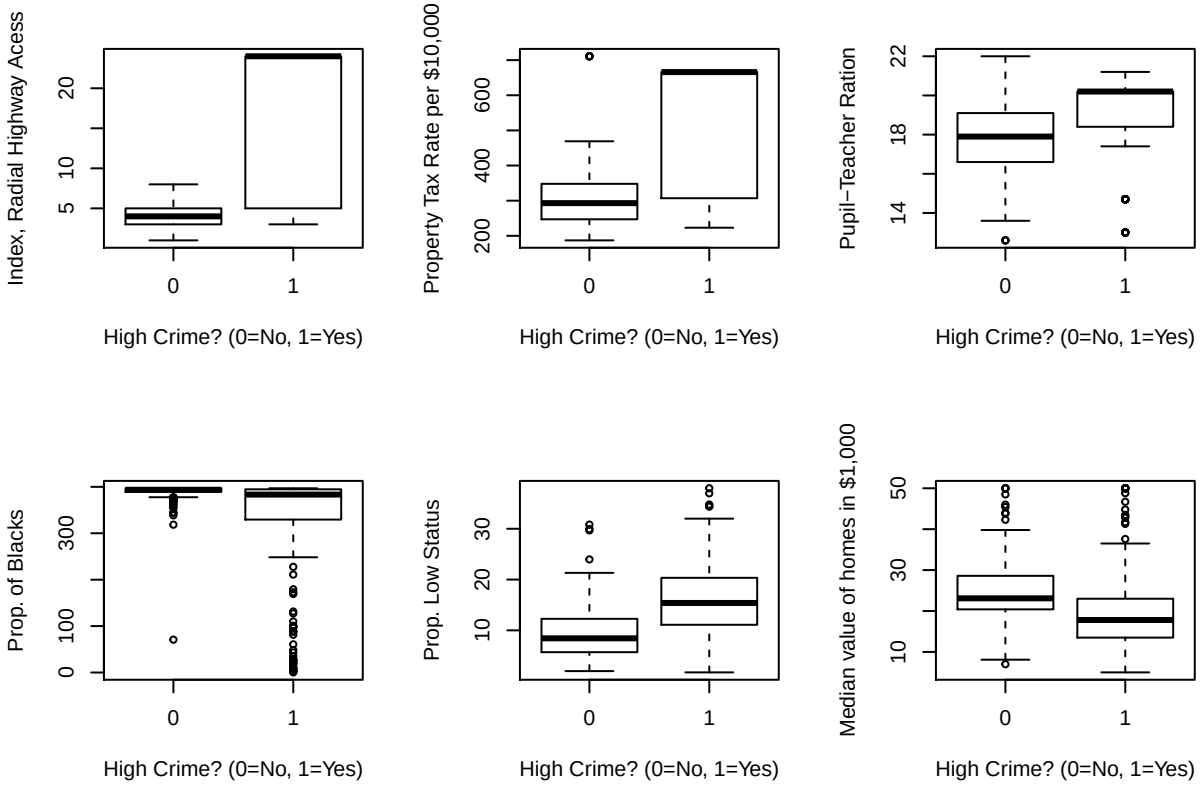
5 Concluding Remarks

Given the results of all of our tests and measurements of these three models we conclude that we have a trustworthy model that appears to do a good job predicting the category for a given neighborhood. We see that the number of low crime neighborhoods is smaller in the predicted data then what we see in the training data set but when we look at the predictions for training data we see that this same pattern. Overall this method of logistic regression gives us some very interesting results and will likely be the direction that my group uses in our final project.

Apendix A: Boxplots of the Predictor Variables

The following are the boxplots of the predictor variables that were used to evaluate the need for including the log(variable) for our second model.




Index, Radial Highway Acess
5
10
20
0
1
High Crime? (0=No, 1=Yes)
Property Tax Rate per $10,000
200
400
600
0
1
High Crime? (0=No, 1=Yes)
Pupil−Teacher Ration
14
18
22
0
1
High Crime? (0=No, 1=Yes)
Prop. of Blacks
0
100
300
0
1
High Crime? (0=No, 1=Yes)
Prop. Low Status
10
20
30
0
1
High Crime? (0=No, 1=Yes)
Median value of homes in $1,000
10
30
50
0
1
High Crime? (0=No, 1=Yes)

Appendix B: R Code

The following R code was used to generate the model construction and evaluation in that we preformed in the report.

```
#### Loading the Data ####
crime <- read.csv("DATA621/crime-training-data.csv")
crime.eval <- read.csv("DATA621/crime-evaluation-data.csv")

#### Lodaing Packages ####
library(caret)
library(pROC)

#### Data Exploration ###
summary(crime)
#49.14% of the neighbor hoods are high crime
table(crime$target)
table(crime$target)/sum(table(crime$target))

# Generating the Pairs Plot
pairs(crime.train, col=as.factor(crime.train$target))

# Generating all the boxplots for the predictor values.
boxplot(zn ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Prop. Zoned for Large Lots") # High skew include log(zn)
boxplot(indus ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Prop. Non-retail Buisness") # Good, median high crime higher
# Charles River Dummy varaible not boxplot
boxplot(nox ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Nitrogen Oxide Concentration") # Good, median high crime higher
boxplot(rm ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Rooms per Dwelling") # Good, Medain slightly larger for low crime
boxplot(age ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Prop. Building Built Before 1940") # May be skewed, high crime is larger
boxplot(dis ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Weighted Mean Distance from Employment")# Good, low crime higher (suburbs?)
boxplot(rad ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Index, Radial Highway Acess") #High index high crime
boxplot(tax ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Property Tax Rate per $10,000")# Good, High tax rate higher crime?
boxplot(ptratio ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Pupil-Teacher Ration") # Good, median higher for high crime
boxplot(black ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Prop. of Blacks") # High skew include log(black)
boxplot(lstat ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Prop. Low Status") # Good, high crime larger median
boxplot(medv ~ target, data = crime, xlab = "High Crime? (0=No, 1=Yes)",
        ylab = "Median value of homes in $1,000") # higher value lower crime


#### Bulding 3 Models ####

# Breaking the data into train and test sets
smp_size <- floor(0.7*nrow(crime))
```

```r
set.seed(42)
train_index <- sample(seq_len(nrow(crime)), size = smp_size)

# Subseting the data
crime.train <- crime[train_index,]
crime.test <- crime[-train_index,]


### Model with Logs
crime.log <- glm(target~ . + log(age) + log(black), data = crime.train,
                 family = binomial(link = "logit"))
summary(crime.log)
logLik(crime.log) # -53.23

# Logit model odds ratios
coefficients(crime.log)
exp(crime.log$coefficients)

# Assessing the model
log.probs <- predict(crime.log, type="response")
log.pred <- ifelse(log.probs > 0.5, 1, 0)
table(log.pred, crime.train$target)
mean(log.pred == crime.train$target)


### Full Model with all predictors
crime.full <- glm(target ~ ., data = crime.train,
                  family = binomial(link = "logit"))
summary(crime.full)
logLik(crime.full)


# Full model odds ratios
coef(crime.full)
exp(crime.full$coefficients)

# Assessing the Full Model
full.probs <- predict(crime.full, type="response")
full.pred <- ifelse(full.probs > 0.5, 1, 0)
table(full.pred, crime.train$target)
mean(full.pred == crime.train$target)

# Comparing the log and full models
# H_0 = crime.full (the reduced model) is true
# H_A = The reduced model is not true and we favor the log model.
anova(crime.full, crime.log, test = "Chisq") # Fail to Reject H_0


# Reduced Model
# -lstat, -rm, - zn(correlates to other variables), log(black)
crime.small <- glm(target ~  chas + nox + indus + age + dis +
                     rad + tax + ptratio + black + medv +
                     log(age),
```

```
                    data = crime.train, family = binomial(link = "logit"))
summary(crime.small)

# Testing the reduced model vs the full model
# H_0: The reduced model is true
# H_A: The reduced model is not true and we favor the log model
anova(crime.small, crime.full, test = "Chisq") # Fail to reject H_0

# Testing the reduced model vs the log model
# H_0: The reduced model is true
# H_A: The reduced model is not true and we favor the log model
anova(crime.small, crime.log, test = "Chisq") # Fail to reject H_0

# Reduced Model Odds
exp(crime.small$coefficients)

# Assessing the Reduced Model
small.probs <- predict(crime.small, type="response")
small.pred <- ifelse(small.probs > 0.5, 1, 0)
table(small.pred, crime.train$target)
mean(small.pred == crime.train$target)
logLik(crime.small)

# Very Simillar AIC's and Very simmilar Accuarcy lets check the
# accuarcy on the test data.


#### Checking vs. The Test Data ####

# Checking the logs model vs the test data.
log.probs_test <- predict(crime.log, newdata=crime.test,type = "response")
log.pred_test <- ifelse(log.probs_test > 0.5, 1, 0)

# Generating the Confusion Matrix and metrics
confusionMatrix(data = as.factor(log.pred_test),
                reference = as.factor(crime.test$target),
                positive = "1")

# Classification Error Rate
1-.9 # 0.1

# F1 Score
(2*0.9344*0.8507)/(0.9344 + 0.8507) #0.8905877

# Generating the ROC curve and the AUC
roc_log <- roc(response = crime.test$target,
               predictor = log.probs_test)
auc(roc_log) #0.9706
plot(roc_log, legacy.axes = TRUE) # Not included in the report


# Checking the full model vs the test data.
full.probs_test <- predict(crime.full, newdata=crime.test, type = "response")
```

```r
full.pred_test <- ifelse(full.probs_test > 0.5, 1, 0)

# Generating the confusion matrix
confusionMatrix(data = as.factor(full.pred_test),
                reference = as.factor(crime.test$target),
                positive = "1")

# Classification Error Rate
1-.9071 # 0.0929

# F1 Score
(2*0.9219*0.8806)/(0.9219 + 0.8806) # 0.9008

# generating the ROC curve and the AUC
roc_full <- roc(response = crime.test$target,
               predictor = full.probs_test)
auc(roc_full) #0.9677
plot(roc_full, legacy.axes = TRUE) # not included in the report


# Checking the small model vs the test data.
small.probs_test <- predict(crime.small, newdata=crime.test, type = "response")
small.pred_test <- ifelse(small.probs_test > 0.5, 1, 0)

# Generating the classification matrix
confusionMatrix(data = as.factor(small.pred_test),
                reference = as.factor(crime.test$target),
                positive = "1")

# Classification Error Rate
1-.9 # 0.1

# F1 Score
(2*0.9492*0.8358)/(0.9492 + 0.8358) # 0.8889

# Generating the ROC Curve and the AUC
roc_small <- roc(response = crime.test$target,
                predictor = small.probs_test)
auc(roc_small) #0.9681
plot(roc_small, legacy.axes = TRUE) # Not included

### Predicting Values
crime.eval$probs <- predict(crime.small, newdata=crime.eval, type = "response")
crime.eval$predicted <- ifelse(crime.eval$probs > 0.5, 1, 0)
head(crime.eval)

# Generating the table of predicted values.
table(crime.eval$predicted)
table(crime.eval$predicted)/sum(table(crime.eval$predicted))
```