Assignment 3

Tina Roha

3/5/2021

Classification

Chapter 4 (page 168): Questions 10, 11, 13

Question-10

This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1, 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

a. Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

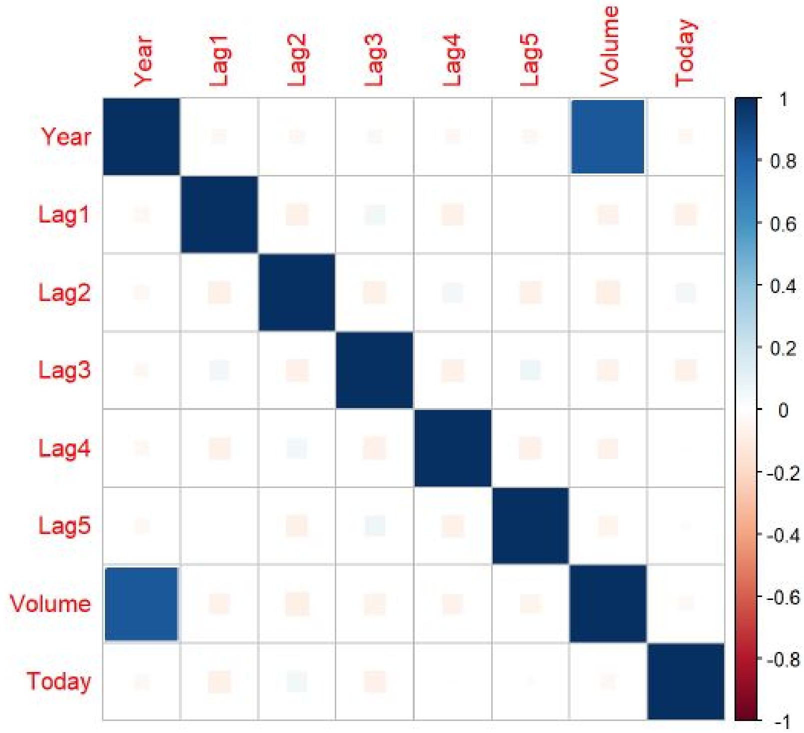
```
library(ISLR)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
summary(Weekly)
```

```
##       Year           Lag1               Lag2               Lag3         
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410  
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472  
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090  
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260  
##       Lag4               Lag5              Volume       
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747  
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202  
##  Median :  0.2380   Median :  0.2340   Median :1.00268  
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821  
##      Today          Direction 
##  Min.   :-18.1950   Down:484  
##  1st Qu.: -1.1540   Up  :605  
##  Median :  0.2410             
##  Mean   :  0.1499             
##  3rd Qu.:  1.4050             
##  Max.   : 12.0260             
```

```
corrplot(cor(Weekly[,-9]), method="square")
```



* The only variables that appear to

In the numerical and graphical summaries of the Weekly data it can be concluded that the variables Year and Volume express significance with evidence of a linear relationship. Furthermore, no other sign of linearity can be detected from the produced output.

b. Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

```
attach(Weekly)
Weekly.fit<-glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data=Weekly,family=binomial)
summary(Weekly.fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

The predictor Lag2 does appear to be statistically significant due to its p-value, 0.0296, being less than 0.05. Therefore, Lag2 has sufficient evidence to reject the null hypothesis and assume the alternative hypothesis (HA: Bi does not equal zero). On the other hand, since all of the other predictors show p-values over 0.05, they do not show significance and accept the null hypothesis (Bi=0).

c. Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

```
logWeekly.prob= predict(Weekly.fit, type='response')
logWeekly.pred =rep("Down", length(logWeekly.prob))
logWeekly.pred[logWeekly.prob > 0.5] = "Up"
table(logWeekly.pred, Direction)
```

```
##               Direction
## logWeekly.pred Down  Up
##           Down   54  48
##           Up    430 557
```

The confusion matrix and overall fraction of correct predictions produced 0.5611 as the result. Furthermore, this value expresses that this final model makes 56.11% correct predictions about the weekly market trend. Additionally, with the overall fraction calculation of the the Up and Down weekly trends we can see that the Up is 97.07% correct whereas Down is only 11.15% correct.

d. Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

```
train = (Year<2009)
Weekly.0910 <-Weekly[!train,]
Weekly.fit<-glm(Direction~Lag2, data=Weekly,family=binomial, subset=train)
logWeekly.prob= predict(Weekly.fit, Weekly.0910, type = "response")
logWeekly.pred = rep("Down", length(logWeekly.prob))
logWeekly.pred[logWeekly.prob > 0.5] = "Up"
Direction.0910 = Direction[!train]
table(logWeekly.pred, Direction.0910)
```

```
##               Direction.0910
## logWeekly.pred Down Up
##           Down    9  5
##           Up     34 56
```

```
mean(logWeekly.pred == Direction.0910)
```

```
## [1] 0.625
```

The confusion matrix and the overall fraction of correct predictions for the held out data from 2009 and 2010, the produced result is 0.625. Furthermore, this value expresses that the model correctly predicts weekly trends 62.5% of the time. Additionally, for the Up and Down weekly trends, the model correctly predicted the Up trends 91.80% and correctly predicted the Down trends 20.93%. Overall, this specific model correctly predicted weekly trends better than when the entire data set was included.

e. Repeat (d) using LDA.

```
library(MASS)
Weeklylda.fit<-lda(Direction~Lag2, data=Weekly,family=binomial, subset=train)
Weeklylda.pred<-predict(Weeklylda.fit, Weekly.0910)
table(Weeklylda.pred$class, Direction.0910)
```

```
##       Direction.0910
##        Down Up
##   Down    9  5
##   Up     34 56
```

```
mean(Weeklylda.pred$class==Direction.0910)
```

```
## [1] 0.625
```

Repeating part d and utilizing a Linear Discriminant Analysis model produced similar results to the logistic regression model.

f. Repeat (d) using QDA.

```
Weeklyqda.fit = qda(Direction ~ Lag2, data = Weekly, subset = train)
Weeklyqda.pred = predict(Weeklyqda.fit, Weekly.0910)$class
table(Weeklyqda.pred, Direction.0910)
```

```
##               Direction.0910
## Weeklyqda.pred Down Up
##           Down    0  0
##           Up     43 61
```

```
mean(Weeklyqda.pred==Direction.0910)
```

```
## [1] 0.5865385
```

Repeating part d and utilizing Quadratic Linear Analysis, the model produced shows it correctly predicts the weekly trends 58.65% of the time. In addition, the model shows the correctly predicted weekly upward trends but did not produce any results for the weekly downward trends.

g. Repeat (d) using KNN with K = 1.

```
library(class)
Week.train=as.matrix(Lag2[train])
Week.test=as.matrix(Lag2[!train])
train.Direction =Direction[train]
set.seed(1)
Weekknn.pred=knn(Week.train,Week.test,train.Direction,k=1)
table(Weekknn.pred,Direction.0910)
```

```
##             Direction.0910
## Weekknn.pred Down Up
##         Down   21 30
##         Up     22 31
```

```
mean(Weekknn.pred == Direction.0910)
```

```
## [1] 0.5
```

Repeating part d and utilizing KNN with K=1 produced a model that correctly predicts weekly trends only 50% of the time.

h. Which of these methods appears to provide the best results on this data?

Based on the previous outputs, it is clear that the methods that provide the best results for this specific data includes the Logistic Regression and Linear Discriminant Analysis due to the fact that they produced the highest rates at 62.5%.

i. Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

Post experimentation with different combinations of predictors for each of the methods concludes that both logistic regression and Linear Discriminant Analysis produce the best overall accuracy rates.

Question-11

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

```
library(ISLR)
attach(Auto)
summary(Auto)
```

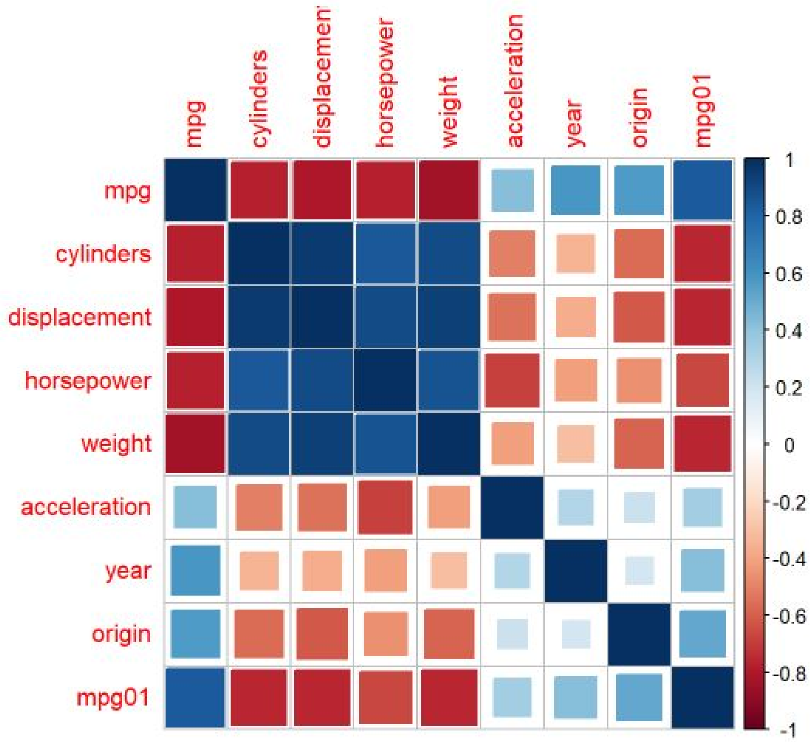
```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##
##                  name
##  amc matador       :  5
##  ford pinto        :  5
##  toyota corolla    :  5
##  amc gremlin       :  4
##  amc hornet        :  4
##  chevrolet chevette:  4
##  (Other)           :365
```

a. Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.

```
mpg01 <- rep(0, length(mpg))
mpg01[mpg > median(mpg)] <- 1
Auto = data.frame(Auto, mpg01)
```

b. Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

```
corrplot(cor(Auto[,-9]), method="square")
```



After exploring the data graphically, it can be concluded that the features most likely to be useful in predicting mpg01 include Cylinders, Displacement and Weight due to each of them having a negative correlation with mpg01.

c. Split the data into a training set and a test set.

```
train <- (year %% 2 == 0)
train.auto <- Auto[train,]
test.auto <- Auto[-train,]
```

d. Perform LDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

```
autolda.fit <- lda(mpg01~displacement+horsepower+weight+year+cylinders+origin, data=train.auto)
autolda.pred <- predict(autolda.fit, test.auto)
table(autolda.pred$class, test.auto$mpg01)
```

```
##
##       0   1
##   0 169   7
##   1  26 189
```

```
mean(autolda.pred$class != test.auto$mpg01)
```

```
## [1] 0.08439898
```

After performing LDA on the training data in order to predict mpg01, using the variables that seemed most associated with mpg01 in part b, the test error of the model obtained is 0.08439898 or 8.44%.

e. Perform QDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

```
autoqda.fit <- qda(mpg01~displacement+horsepower+weight+year+cylinders+origin, data=train.auto)
autoqda.pred <- predict(autoqda.fit, test.auto)
table(autoqda.pred$class, test.auto$mpg01)
```

```
##
##       0   1
##   0 176  20
##   1  19 176
```

```
mean(autoqda.pred$class != test.auto$mpg01)
```

```
## [1] 0.09974425
```

After performing QDA on the training data in order to predict mpg01, using the variables that seemed most associated with mpg01 in part b, the test error of the model obtained is 0.09974425 or 9.97%.

f. Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

```
auto.fit<-glm(mpg01~displacement+horsepower+weight+year+cylinders+origin, data=train.auto,family=binomial)
auto.probs = predict(auto.fit, test.auto, type = "response")
auto.pred = rep(0, length(auto.probs))
auto.pred[auto.probs > 0.5] = 1
table(auto.pred, test.auto$mpg01)
```

```
##
## auto.pred   0   1
##         0 174  12
##         1  21 184
```

```
mean(auto.pred != test.auto$mpg01)
```

```
## [1] 0.08439898
```

After performing logistic regression on the training data in order to predict mpg01, using the variables that seemed most associated with mpg01 in part b, the test error of the model obtained is 0.08439898 or 8.44%.

g. Perform KNN on the training data, with several values of K, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

```
#K=1
train.K= cbind(displacement,horsepower,weight,cylinders,year, origin)[train,]
test.K=cbind(displacement,horsepower,weight,cylinders, year, origin)[-train,]
set.seed(1)
autok.pred=knn(train.K,test.K,train.auto$mpg01,k=1)
mean(autok.pred != test.auto$mpg01)
```

```
## [1] 0.07161125
```

```
#K=5
autok.pred=knn(train.K,test.K,train.auto$mpg01,k=5)
mean(autok.pred != test.auto$mpg01)
```

```
## [1] 0.112532
```

```
#K=10
autok.pred=knn(train.K,test.K,train.auto$mpg01,k=10)
mean(autok.pred != test.auto$mpg01)
```

```
## [1] 0.1253197
```

```
detach(Auto)
```

After performing KNN on the training data, with several values of K, in order to predict mpg01, using the variables that seemed most associated with mpg01 in part b, the test errors of the model obtained are 0.07161125 or 7.16%, 0.112532 or 11.25% and 0.1253197 or 12.53%. Furthermore, K=1 seems to perform best due to it having the lowest error rate of 7.16%.

Question-13

Using the Boston data set, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, LDA, and KNN models using various subsets of the predictors. Describe your findings.

```
summary(Boston)
```

```
##       crim                zn             indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       nox               rm             age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio          black
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

```
attach(Boston)
```

Creating binary *crim* variable.

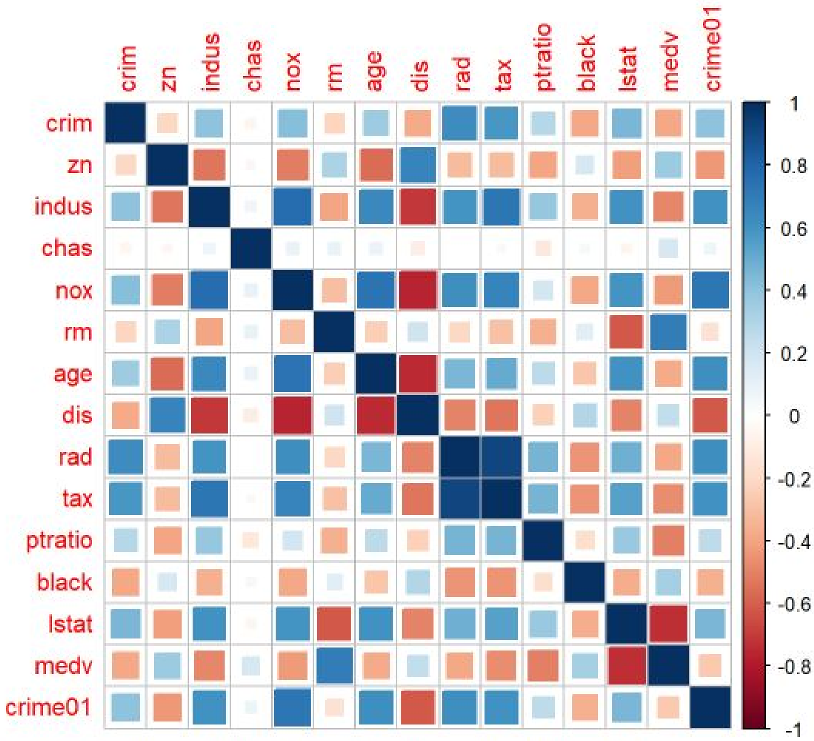
```
crime01 <- rep(0, length(crim))
crime01[crim > median(crim)] <- 1
Boston= data.frame(Boston,crime01)
```

Splitting the dataset

```
train = 1:(dim(Boston)[1]/2)
test = (dim(Boston)[1]/2 + 1):dim(Boston)[1]
Boston.train = Boston[train, ]
Boston.test = Boston[test, ]
crime01.test = crime01[test]
```

Determination of any associations to *crime01*

```
corrplot(cor(Boston), method="square")
```



- It appears that the variables *indus*, *nox*,*age*,*dis*, *rad* and *tax* have the strongest association with the desired variable.

Logistic Regression

```
set.seed(1)
Boston.fit <-glm(crime01~ indus+nox+age+dis+rad+tax, data=Boston.train,family=binomial)
Boston.probs = predict(Boston.fit, Boston.test, type = "response")
Boston.pred = rep(0, length(Boston.probs))
Boston.pred[Boston.probs > 0.5] = 1
table(Boston.pred, crime01.test)
```

```
##            crime01.test
## Boston.pred   0   1
##           0  75   8
##           1  15 155
```

```
mean(Boston.pred != crime01.test)
```

```
## [1] 0.09090909
```

```
summary(Boston.fit)
```

```
##
## Call:
## glm(formula = crime01 ~ indus + nox + age + dis + rad + tax,
##     family = binomial, data = Boston.train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.97810  -0.21406  -0.03454   0.47107   3.04502
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -42.214032   7.617440  -5.542 2.99e-08 ***
## indus        -0.213126   0.073236  -2.910  0.00361 **
## nox          80.868029  16.066473   5.033 4.82e-07 ***
```

```
## age           0.003397   0.012032   0.282  0.77772
## dis           0.307145   0.190502   1.612  0.10690
## rad           0.847236   0.183767   4.610 4.02e-06 ***
## tax          -0.013760   0.004956  -2.777  0.00549 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 329.37  on 252  degrees of freedom
## Residual deviance: 144.44  on 246  degrees of freedom
## AIC: 158.44
##
## Number of Fisher Scoring iterations: 8
```

Linear Discriminat Analysis

```
Boston.ldafit <-lda(crime01~ indus+nox+age+dis+rad+tax, data=Boston.train,family=binomial)
Bostonlda.pred = predict(Boston.ldafit, Boston.test)
table(Bostonlda.pred$class, crime01.test)
```

```
##    crime01.test
##       0   1
##   0  81  18
##   1   9 145
```

```
mean(Bostonlda.pred$class != crime01.test)
```

```
## [1] 0.1067194
```

K Nearest Neighbors

```
#K=1
train.K=cbind(indus,nox,age,dis,rad,tax)[train,]
test.K=cbind(indus,nox,age,dis,rad,tax)[test,]
Bosknn.pred=knn(train.K, test.K, crime01.test, k=1)
table(Bosknn.pred,crime01.test)
```

```
##            crime01.test
## Bosknn.pred   0   1
##           0  31 155
##           1  59   8
```

```
mean(Bosknn.pred !=crime01.test)
```

```
## [1] 0.8458498
```

```
#K=100
train.K=cbind(indus,nox,age,dis,rad,tax)[train,]
test.K=cbind(indus,nox,age,dis,rad,tax)[test,]
Bosknn.pred=knn(train.K, test.K, crime01.test, k=100)
table(Bosknn.pred,crime01.test)
```

```
##            crime01.test
## Bosknn.pred   0   1
##           0  21   6
##           1  69 157
```

```
mean(Bosknn.pred !=crime01.test)
```

```
## [1] 0.2964427
```

Using the Boston data set and fitting classification models in order to predict whether a given suburb has a crime rate above or below the median, it is concluded that logistic regression produced the lowest test error rate at 9.09%. Furthermore, in the logistic regression model the variables indus, nox, rad and tax all expressed statistical significance with p-values 0.00361, 4.82e-07, 4.02e-06 and 0.00549 respectively. Furthermore, these were the only variables to best associate with crime01 which can be seen graphically. Out of all the models, the KNN model is the only one ineffective in classification due to the error rate being the largest at 84.58%. Moreover, even though the error rate went down as the K value grew, the logistic regression and LDA error rates are still much lower.