data621 HW3

jim lung

April 5, 2018

1. Data Exploration:

Analyzing the overall data to see if there is any discrepancies there as missing data or there is any need for data transformation

```
names(crime)

##  [1] "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"
##  [8] "rad"     "tax"     "ptratio" "lstat"   "medv"    "target"

str(crime)

## 'data.frame':    466 obs. of  13 variables:
##  $ zn     : num  0 0 0 30 0 0 0 0 0 80 ...
##  $ indus  : num  19.58 19.58 18.1 4.93 2.46 ...
##  $ chas   : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.605 0.871 0.74 0.428 0.488 0.52 0.693 0.693 0.515 0.392
...
##  $ rm     : num  7.93 5.4 6.49 6.39 7.16 ...
##  $ age    : num  96.2 100 100 7.8 92.2 71.3 100 100 38.1 19.1 ...
##  $ dis    : num  2.05 1.32 1.98 7.04 2.7 ...
##  $ rad    : int  5 5 24 6 3 5 24 24 5 1 ...
##  $ tax    : int  403 403 666 300 193 384 666 666 224 315 ...
##  $ ptratio: num  14.7 14.7 20.2 16.6 17.8 20.9 20.2 20.2 20.2 16.4 ...
##  $ lstat  : num  3.7 26.82 18.85 5.19 4.82 ...
##  $ medv   : num  50 13.4 15.4 23.7 37.9 26.5 5 7 22.2 20.9 ...
##  $ target : int  1 1 1 0 0 0 1 1 0 0 ...

dim(crime)

## [1] 466  13

kable(summary(crime))
```

zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv	target
Min. : 0.00	Min. : 0.460	Min. :0.00000	Min. :0.3890	Min. :3.863	Min. : 2.90	Min. : 1.130	Min. : 1.00	Min. :187.0	Min. :12.6	Min. : 1.730	Min. : 5.00	Min. :0.0000
1st Qu.	1st Qu.:	1st Qu.:0.	1st Qu.:0	1st Qu.:	1st Qu.	1st Qu.	1st Qu.:	1st Qu.:	1st Qu.:	1st Qu.:	1st Qu.:	1st Qu.:0

: 0.00	5.145	00000	.4480	5.887	: 43.88	: 2.101	4.00	281.0	16.9	7.043	17.02	.0000
Median : 0.00	Median : 9.690	Median :0.00000	Median :0.5380	Median :6.210	Median : 77.15	Median : 3.191	Median : 5.00	Median :334.5	Median :18.9	Median :11.350	Median :21.20	Median :0.0000
Mean : 11.58	Mean :11.105	Mean :0.07082	Mean :0.5543	Mean :6.291	Mean : 68.37	Mean : 3.796	Mean : 9.53	Mean :409.5	Mean :18.4	Mean :12.631	Mean :22.59	Mean :0.4914
3rd Qu. : 16.25	3rd Qu.:18.100	3rd Qu.:0.00000	3rd Qu.:0.6240	3rd Qu.: 6.630	3rd Qu. : 94.10	3rd Qu. : 5.215	3rd Qu.: 24.00	3rd Qu.: 666.0	3rd Qu.: 20.2	3rd Qu.:16.930	3rd Qu.: 25.00	3rd Qu.:1.0000
Max. :100.00	Max. :27.740	Max. :1.00000	Max. :0.8710	Max. :8.780	Max. :100.00	Max. :12.127	Max. :24.00	Max. :711.0	Max. :22.0	Max. :37.970	Max. :50.00	Max. :1.0000

We observed that:

- The crime dataset contains 13 variables, with 466 observations
- There are no missing values.
- The Minimum, Quatiles and Maximum values.
- Since this is logistic regression we don’t have to worry about the normal distribution of data and no transformation is needed

2. Data Preparation

There is no major data preparation effort is needed as this is a logistic regression and more over there is no missing data in the dataset.

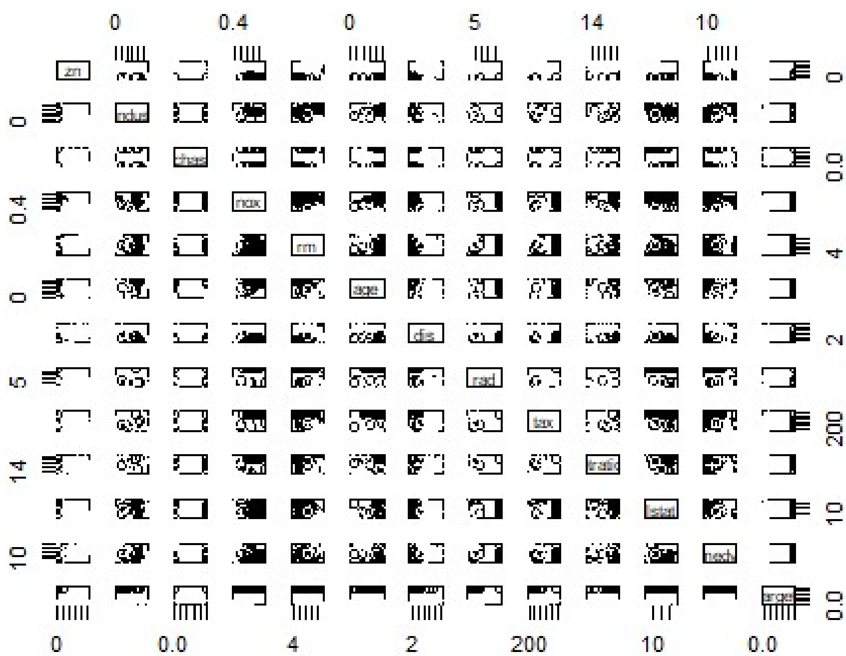
```
## checkin no missing data
sapply(crime, function(x) sum(is.na(x)))

##      zn   indus    chas     nox      rm     age     dis     rad     tax 
##       0       0       0       0       0       0       0       0       0 
## ptratio   lstat    medv  target 
##       0       0       0       0

sapply(crime_evaluation, function(x) sum(is.na(x)))
```

```
##       zn    indus     chas      nox       rm      age      dis      rad      tax 
##        0        0        0        0        0        0        0        0        0 
## ptratio    lstat     medv 
##        0        0        0
```

3. Build Models

Consdering target as a response variable (Independent variable), lets pair it with complete data set and also find the best fit model using GLM package

```
pairs(crime, col=crime$target)
```



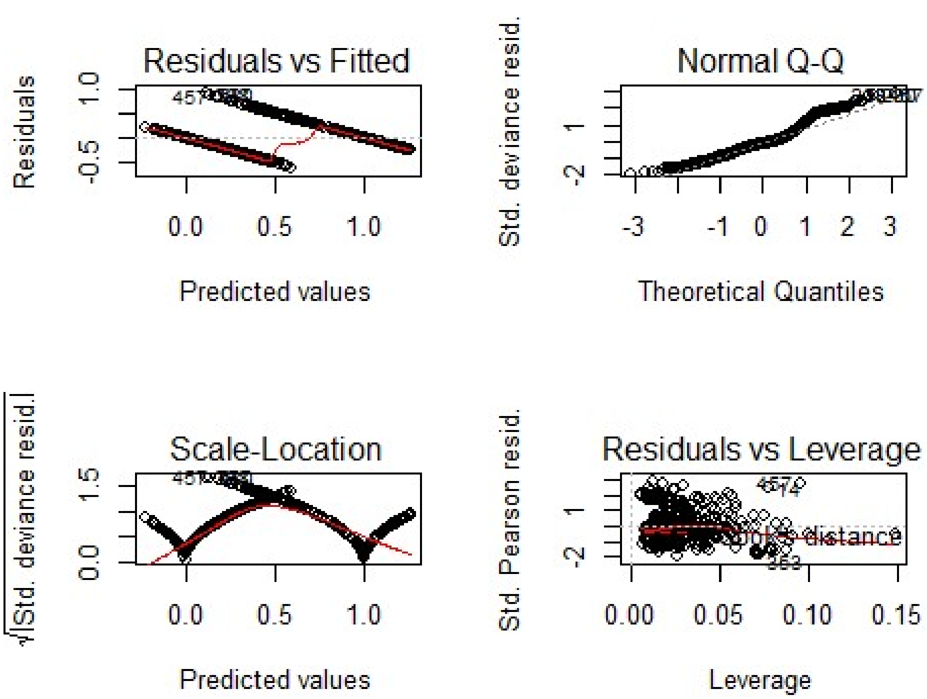
Simple regression model

```
fit <- glm(target ~., data = crime)
summary(fit)
```

```
## 
## Call:
## glm(formula = target ~ ., data = crime)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -0.59701  -0.21505  -0.04691   0.14908   0.88702  
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.6013725  0.3594901  -4.455 1.06e-05 ***
```

```
## zn          -0.0009668  0.0009442  -1.024 0.306432
## indus        0.0031277  0.0042909   0.729 0.466433
## chas         0.0059892  0.0588402   0.102 0.918970
## nox          1.9722476  0.2632648   7.491 3.60e-13 ***
## rm           0.0249823  0.0315042   0.793 0.428202
## age          0.0031738  0.0009045   3.509 0.000495 ***
## dis          0.0125382  0.0141433   0.887 0.375814
## rad          0.0207000  0.0043384   4.771 2.47e-06 ***
## tax         -0.0002787  0.0002617  -1.065 0.287396
## ptratio      0.0115287  0.0093460   1.234 0.218013
## lstat        0.0045124  0.0038923   1.159 0.246935
## medv         0.0089246  0.0029992   2.976 0.003080 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09737169)
##
##     Null deviance: 116.466  on 465  degrees of freedom
## Residual deviance:  44.109  on 453  degrees of freedom
## AIC: 251.85
##
## Number of Fisher Scoring iterations: 2
```

```r
par(mfrow=c(2,2))
plot(fit)
```

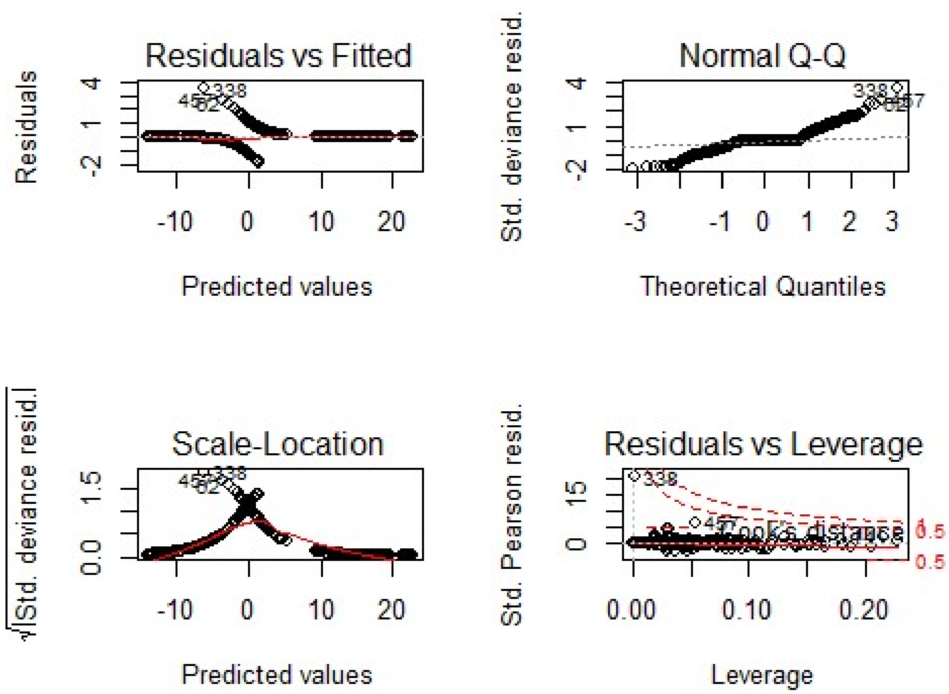


Simple regression model using glm package shows that the p value for zn,indus,chas, rm,dis, tax, ptratio,black ,lstat are more than the significance value of 0.05, so they are not contributing much to the target (independent variable)

So, lets move to the logistic regression for binomial distribution where we can see the variables interdependent on the independent variable target and get teh best fit subset of the crime dataset

Using Logistics regression for a better results as

```
crimetarget <- glm(target~., family=binomial(link='logit'),data=crime)

summary(crimetarget)

##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = crime)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8464  -0.1445  -0.0017   0.0029   3.4665
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -40.822934   6.632913  -6.155 7.53e-10 ***
## zn           -0.065946   0.034656  -1.903  0.05706 .
## indus        -0.064614   0.047622  -1.357  0.17485
## chas          0.910765   0.755546   1.205  0.22803
## nox          49.122297   7.931706   6.193 5.90e-10 ***
## rm           -0.587488   0.722847  -0.813  0.41637
## age           0.034189   0.013814   2.475  0.01333 *
## dis           0.738660   0.230275   3.208  0.00134 **
## rad           0.666366   0.163152   4.084 4.42e-05 ***
## tax          -0.006171   0.002955  -2.089  0.03674 *
## ptratio       0.402566   0.126627   3.179  0.00148 **
## lstat         0.045869   0.054049   0.849  0.39608
## medv          0.180824   0.068294   2.648  0.00810 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 192.05  on 453  degrees of freedom
## AIC: 218.05
##
## Number of Fisher Scoring iterations: 9

par(mfrow=c(2,2))
plot(crimetarget)
```



The variables like zn, indus, chas,rm and lstat are not statistically significant due to their p-value being greater than statiscally accepted p-value of 0.05, So we have a scope to refine the model without these variables and repeat the best fit logistic regression and build a preditive model.

Null deviance is 645.88 to imply if all other parameters are held constant(control or not included), the estimate would be 645.88, while the Residual deviance of 186.15 means with the imclusion of other estimator, we expect the deviance to be 186.14.

AIC is 214,15 and signifies the best fit quality of the model compared to other similar model available. If we are comparing with other models, best model should have lowest deviance and AIC value.

The greater the difference between the Null deviance and Residual deviance, the better.

The Analysis of Variance (ANOVA)

To confirm if we have concluded the significance of varaibles correctly or not

```
anova(crimetarget, test="Chisq")

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: target
##
```

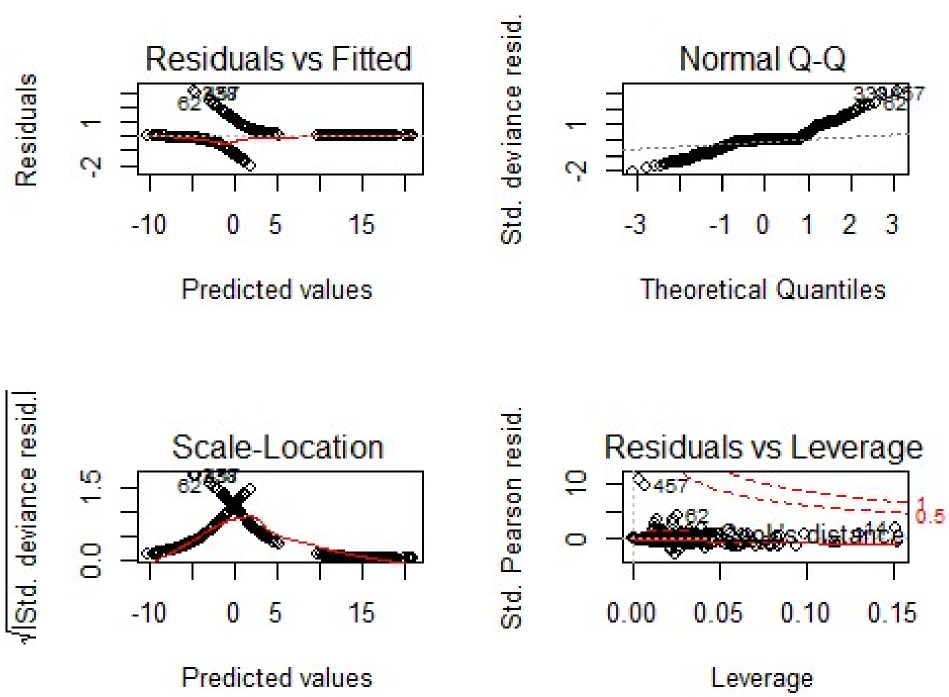
```
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       465     645.88
## zn       1  127.411        464     518.46 < 2.2e-16 ***
## indus    1   86.433        463     432.03 < 2.2e-16 ***
## chas     1    1.274        462     430.76  0.258981
## nox      1  150.804        461     279.95 < 2.2e-16 ***
## rm       1    6.755        460     273.20  0.009349 **
## age      1    0.217        459     272.98  0.641515
## dis      1    7.981        458     265.00  0.004727 **
## rad      1   53.018        457     211.98 3.305e-13 ***
## tax      1    5.562        456     206.42  0.018355 *
## ptratio  1    5.657        455     200.76  0.017388 *
## lstat    1    0.814        454     199.95  0.366872
## medv     1    7.904        453     192.05  0.004933 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It shows that the chas, age and lstat has no significance and rest all are contributing towards target variable. So lets run the best fit model keeping significant variables.

```
crime2 <- subset(crime, select = -c(zn,indus,chas,rm,lstat))

crimetarget2 <- glm(target~., family=binomial(link='logit'),data=crime2)
summary(crimetarget2)

##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = crime2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.01059  -0.19744  -0.01371   0.00402   3.06424
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -36.824228   5.858405  -6.286 3.26e-10 ***
## nox          42.338378   6.639207   6.377 1.81e-10 ***
## age           0.031882   0.010693   2.982 0.002867 **
## dis           0.429555   0.171849   2.500 0.012433 *
## rad           0.701767   0.139426   5.033 4.82e-07 ***
## tax          -0.008237   0.002534  -3.250 0.001153 **
## ptratio       0.376575   0.108912   3.458 0.000545 ***
## medv          0.093653   0.033556   2.791 0.005255 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 203.45  on 458  degrees of freedom
## AIC: 219.45
##
## Number of Fisher Scoring iterations: 9
```

```r
par(mfrow=c(2,2))
plot(crimetarget2)
```



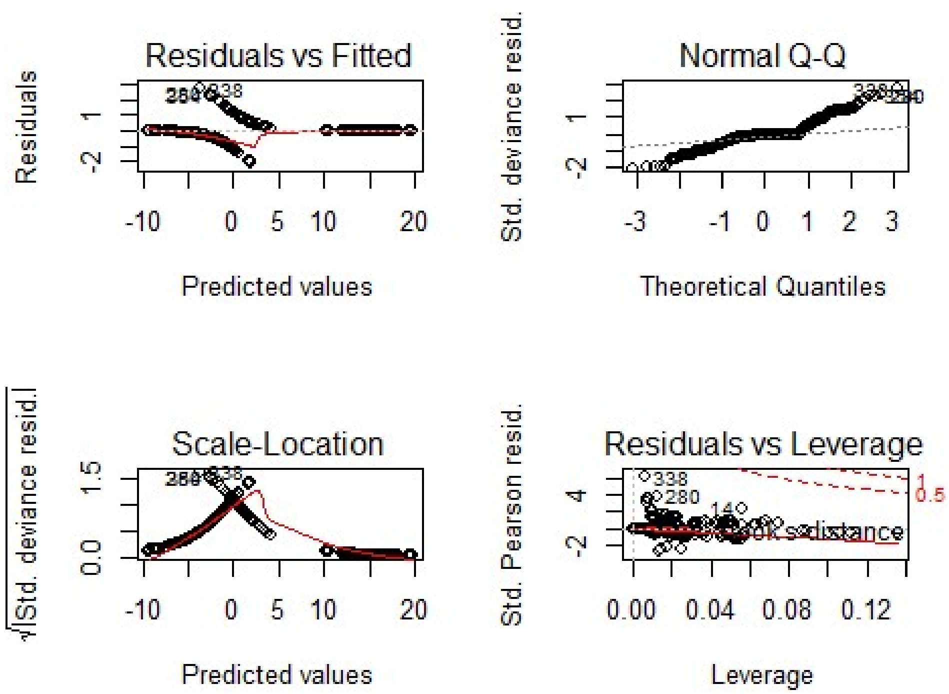

```r
anova(crimetarget2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: target
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      465     645.88
## nox      1   353.86       464     292.01 < 2.2e-16 ***
## age      1     1.39       463     290.63  0.238898
## dis      1     1.94       462     288.68  0.163583
```

```
## rad      1    54.52       461     234.17 1.542e-13 ***
## tax      1    16.00       460     218.17 6.344e-05 ***
## ptratio  1     5.77       459     212.40  0.016304 *
## medv     1     8.95       458     203.45  0.002769 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

age, dis are not significantly contributing to the target variable as it's p value is more than the significance value, so lets remove that from the next iteration

```
crime3 <- subset(crime2, select = -c(age, dis))

crimetarget3 <- glm(target~., family=binomial(link='logit'),data=crime3)
summary(crimetarget3)

##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = crime3)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.05242  -0.25136  -0.01751   0.00330   2.70219
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -28.282949   4.063731  -6.960 3.41e-12 ***
## nox          38.099001   4.900368   7.775 7.56e-15 ***
## rad           0.701410   0.135172   5.189 2.11e-07 ***
## tax          -0.008313   0.002439  -3.408 0.000654 ***
## ptratio       0.304825   0.104419   2.919 0.003509 **
## medv          0.050244   0.027761   1.810 0.070312 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 215.23  on 460  degrees of freedom
## AIC: 227.23
##
## Number of Fisher Scoring iterations: 9

par(mfrow=c(2,2))
plot(crimetarget3)
```



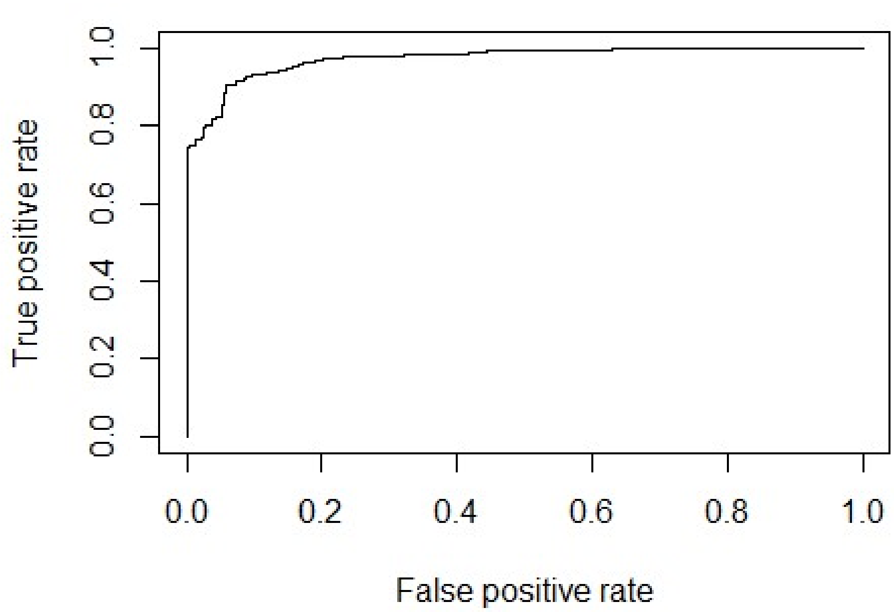

```
anova(crimetarget3, test="Chisq")

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: target
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      465     645.88
## nox      1   353.86       464     292.01 < 2.2e-16 ***
## rad      1    52.50       463     239.51   4.3e-13 ***
## tax      1    15.04       462     224.47 0.0001053 ***
## ptratio  1     5.77       461     218.70 0.0162983 *
## medv     1     3.47       460     215.23 0.0623311 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

crimetarget3 model has nox,black, rad, tax, ptratio and medv as the significant variables and contrbuting to the target as key variable predicting crime in that area

4. Selection Models

Predictive model for crimetarget model

```
pred <- predict(crimetarget, type="response")
pred2 <- prediction(pred, crime$target)
pred3 <- performance(pred2, measure = "tpr", x.measure = "fpr")
plot(pred3)
```



Above is the plot for Sensitivity and Specitivity for the city target, while the value below is it AUC.

```
auc <- performance(pred2, measure = "auc")
auc <- auc@y.values[[1]]
auc

## [1] 0.9737623
```

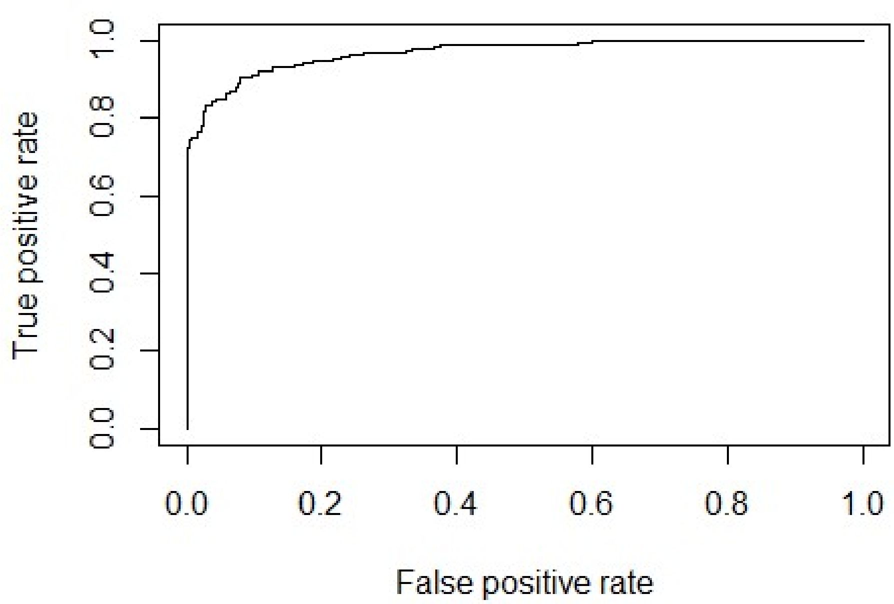
Predictions and Accuracy for crimetarget model

```
target_predicts <-
predict(crimetarget,newdata=subset(crime,select=c(1,2,3,4,5,6,7,8,9,10,11,12,
13)),type='response')
target_predicts <- ifelse(target_predicts > 0.5,1,0)

attach(crime)

CM1<-table(target_predicts, target)
```

```
Pos_Pos=CM1[1,1]
Pos_Neg=CM1[1,2]
Neg_Pos=CM1[2,1]
Neg_Neg=CM1[2,2]

Specificity= Neg_Neg/(Pos_Neg+Neg_Neg)
Sensitivity= Pos_Pos/(Pos_Pos+Neg_Pos)
Pos_Pred_Val= Pos_Pos/(Pos_Pos+Pos_Neg)
Neg_Pred_Val=Neg_Neg/(Neg_Pos+Neg_Neg)

misClasificError <- mean(target_predicts != target)
Accuracy=1-misClasificError

print(paste('Accuracy',1-misClasificError))

## [1] "Accuracy 0.916309012875536"

BestFitModel1<-
data.frame(auc,Specificity,Sensitivity,Accuracy,Pos_Pred_Val,Neg_Pred_Val)
```

Predictive model for crimetarget2 model

```
pred <- predict(crimetarget2, type="response")
pred2 <- prediction(pred, crime$target)
pred3 <- performance(pred2, measure = "tpr", x.measure = "fpr")
plot(pred3)
```



Above is the plot for Sensitivity and Specitivity for the city target, while the value below is it AUC.

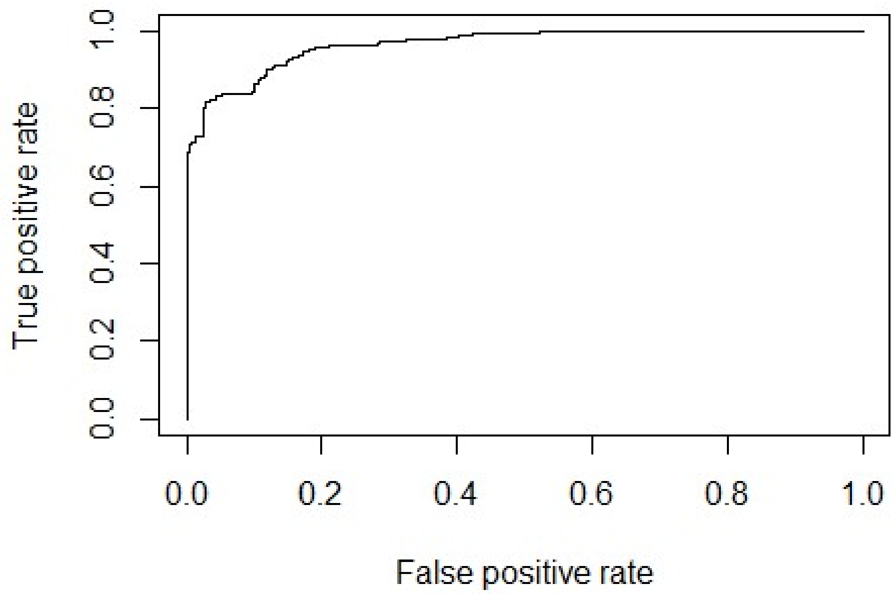
```
auc <- performance(pred2, measure = "auc")
auc <- auc@y.values[[1]]
auc

## [1] 0.9692849
```

Predictions and Accuracy for crimetarget2 model

```
target_predicts <- predict(crimetarget2,newdata=crime,type='response')
target_predicts <- ifelse(target_predicts > 0.5,1,0)

attach(crime)

## The following objects are masked from crime (pos = 3):
##
##     age, chas, dis, indus, lstat, medv, nox, ptratio, rad, rm,
##     target, tax, zn

CM1<-table(target_predicts, target)
Pos_Pos=CM1[1,1]
Pos_Neg=CM1[1,2]
Neg_Pos=CM1[2,1]
Neg_Neg=CM1[2,2]

Specificity= Neg_Neg/(Pos_Neg+Neg_Neg)
Sensitivity= Pos_Pos/(Pos_Pos+Neg_Pos)
Pos_Pred_Val= Pos_Pos/(Pos_Pos+Pos_Neg)
Neg_Pred_Val=Neg_Neg/(Neg_Pos+Neg_Neg)

misClasificError <- mean(target_predicts != target)
Accuracy=1-misClasificError

print(paste('Accuracy',1-misClasificError))

## [1] "Accuracy 0.912017167381974"

BestFitModel2<-
data.frame(auc,Specificity,Sensitivity,Accuracy,Pos_Pred_Val,Neg_Pred_Val)
```

Predictive model for crimetarget3 model

```
pred <- predict(crimetarget3, type="response")
pred2 <- prediction(pred, crime$target)
pred3 <- performance(pred2, measure = "tpr", x.measure = "fpr")
plot(pred3)
```



Above is the plot for Sensitivity and Specitivity for the city target, while the value below is it AUC.

```
auc <- performance(pred2, measure = "auc")
auc <- auc@y.values[[1]]
auc

## [1] 0.9658394
```

Predictions and Accuracy.

```
target_predicts <- predict(crimetarget3,newdata=crime,type='response')
target_predicts <- ifelse(target_predicts > 0.5,1,0)

attach(crime)

## The following objects are masked from crime (pos = 3):
##
##     age, chas, dis, indus, lstat, medv, nox, ptratio, rad, rm,
##     target, tax, zn

## The following objects are masked from crime (pos = 4):
##
##     age, chas, dis, indus, lstat, medv, nox, ptratio, rad, rm,
##     target, tax, zn

CM1<-table(target_predicts, target)
Pos_Pos=CM1[1,1]
```

```
Pos_Neg=CM1[1,2]
Neg_Pos=CM1[2,1]
Neg_Neg=CM1[2,2]

Specificity= Neg_Neg/(Pos_Neg+Neg_Neg)
Sensitivity= Pos_Pos/(Pos_Pos+Neg_Pos)
Pos_Pred_Val= Pos_Pos/(Pos_Pos+Pos_Neg)
Neg_Pred_Val=Neg_Neg/(Neg_Pos+Neg_Neg)

misClasificError <- mean(target_predicts != target)
Accuracy=1-misClasificError

print(paste('Accuracy',1-misClasificError))

## [1] "Accuracy 0.873390557939914"

BestFitModel3<-
data.frame(auc,Specificity,Sensitivity,Accuracy,Pos_Pred_Val,Neg_Pred_Val)
```

Compare the Models to choose the best

```
CompareBestFitModel=rbind(BestFitModel1,BestFitModel2,BestFitModel3)
colnames(CompareBestFitModel)=c("AUC","Specificity","Sensitivity","Accuracy",
"Pos_Pred_Val","Neg_Pred_Val")
rownames(CompareBestFitModel)=c("Model1","Model2","Model3")
CompareBestFitModel

##              AUC Specificity Sensitivity  Accuracy Pos_Pred_Val
## Model1 0.9737623   0.9039301   0.9282700 0.9163090    0.9090909
## Model2 0.9692849   0.9039301   0.9198312 0.9120172    0.9083333
## Model3 0.9658394   0.8427948   0.9029536 0.8733906    0.8560000
##        Neg_Pred_Val
## Model1    0.9241071
## Model2    0.9159292
## Model3    0.8935185
```

Conclusion

From the above analysis, we can deduce that the AUC (Area Under Curve) for all the three models are very close to 1, which indicate that the model 1 is more specificity, sensitivity and accuracy.

And the nox, rad, tax, pratio, black and medv contributed significantly to the increasing crime rate of the city under observation.