DATA621-Homework3-HoddeFarrisBurmoodLin

Rob Hodde, Matt Farris, JeffreyBurmood, Bin Lin

3/28/2017

Problem Description

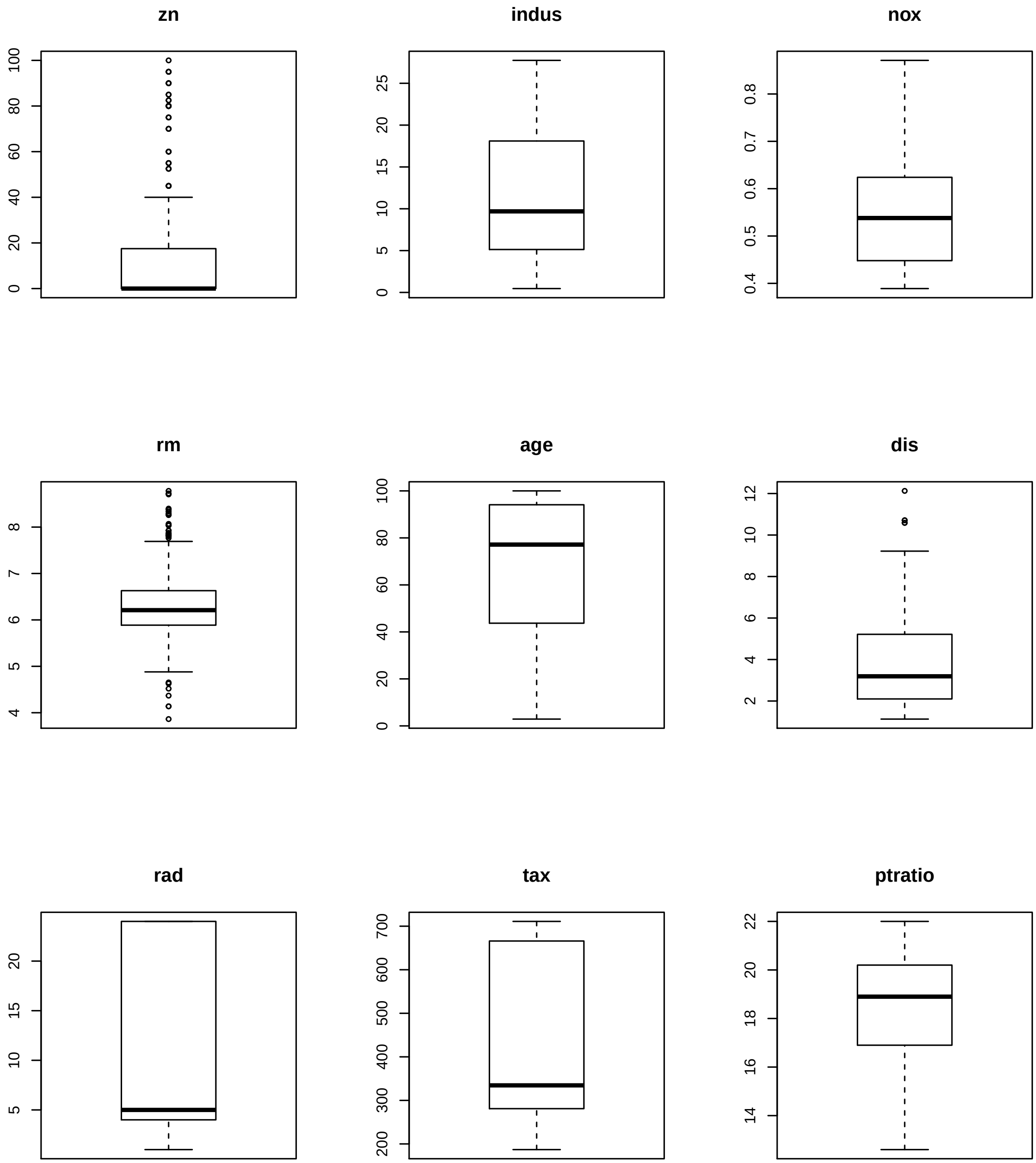
Explore, analyze and model a data set containing information on crime for various neighborhoods of a major city. Using the data set build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. Provide classifications and probabilities for the evaluation data set using the developed binary logistic regression model.

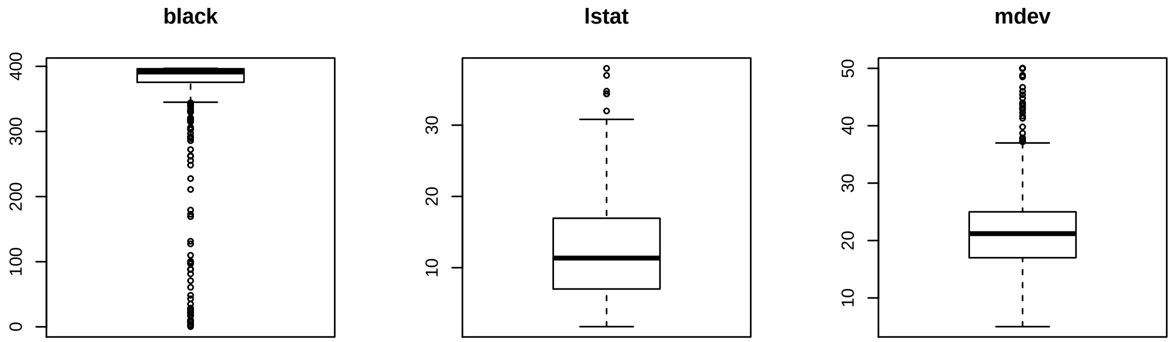
Data Exploration

VAR	TYPE
zn	double
indus	double
chas	integer
nox	double
rm	double
age	double
dis	double
rad	integer
tax	integer
ptratio	double
black	double
lstat	double
medv	double
target	integer

zn	indus	chas	nox	rm	age
Min. : 0.00	Min. : 0.460	Min. :0.00000	Min. :0.3890	Min. :3.863	Min. : 2.90
1st Qu.: 0.00	1st Qu.: 5.145	1st Qu.:0.00000	1st Qu.:0.4480	1st Qu.:5.887	1st Qu.: 43.88
Median : 0.00	Median : 9.690	Median :0.00000	Median :0.5380	Median :6.210	Median : 77.15
Mean : 11.58	Mean :11.105	Mean :0.07082	Mean :0.5543	Mean :6.291	Mean : 68.37
3rd Qu.: 16.25	3rd Qu.:18.100	3rd Qu.:0.00000	3rd Qu.:0.6240	3rd Qu.:6.630	3rd Qu.: 94.10
Max. :100.00	Max. :27.740	Max. :1.00000	Max. :0.8710	Max. :8.780	Max. :100.00

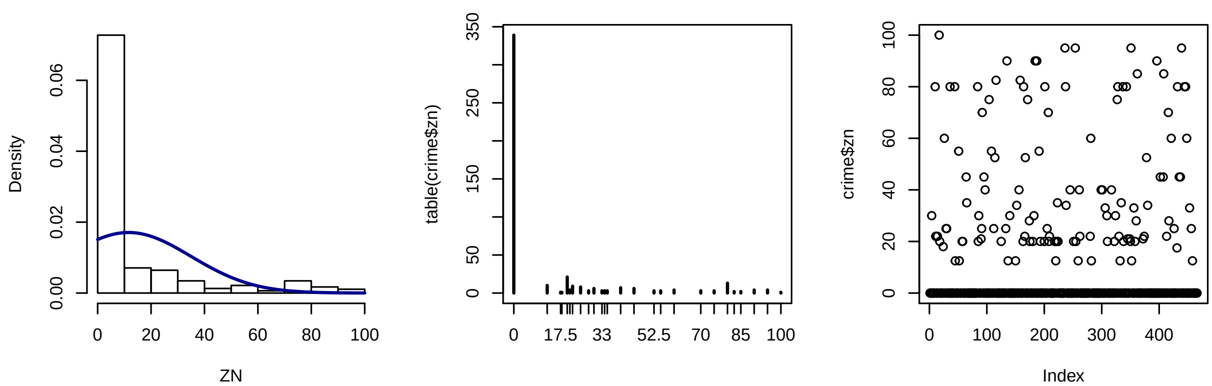
dis	rad	tax	ptratio	black	lstat
Min. : 1.130	Min. : 1.00	Min. :187.0	Min. :12.6	Min. : 0.32	Min. : 1.730
1st Qu.: 2.101	1st Qu.: 4.00	1st Qu.:281.0	1st Qu.:16.9	1st Qu.:375.61	1st Qu.: 7.043
Median : 3.191	Median : 5.00	Median :334.5	Median :18.9	Median :391.34	Median :11.350
Mean : 3.796	Mean : 9.53	Mean :409.5	Mean :18.4	Mean :357.12	Mean :12.631
3rd Qu.: 5.215	3rd Qu.:24.00	3rd Qu.:666.0	3rd Qu.:20.2	3rd Qu.:396.24	3rd Qu.:16.930
Max. :12.127	Max. :24.00	Max. :711.0	Max. :22.0	Max. :396.90	Max. :37.970

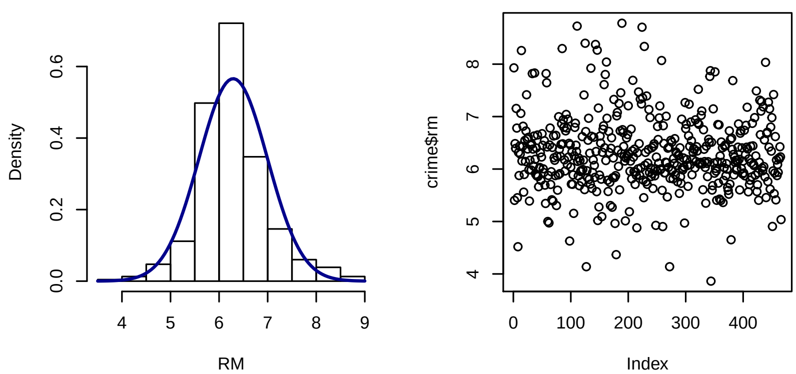

zn
0
20
40
60
80
100
indus
0
5
10
15
20
25
nox
0.4
0.5
0.6
0.7
0.8
rm
4
5
6
7
8
age
0
20
40
60
80
100
dis
2
4
6
8
10
12
rad
5
10
15
20
tax
200
300
400
500
600
700
ptratio
14
16
18
20
22

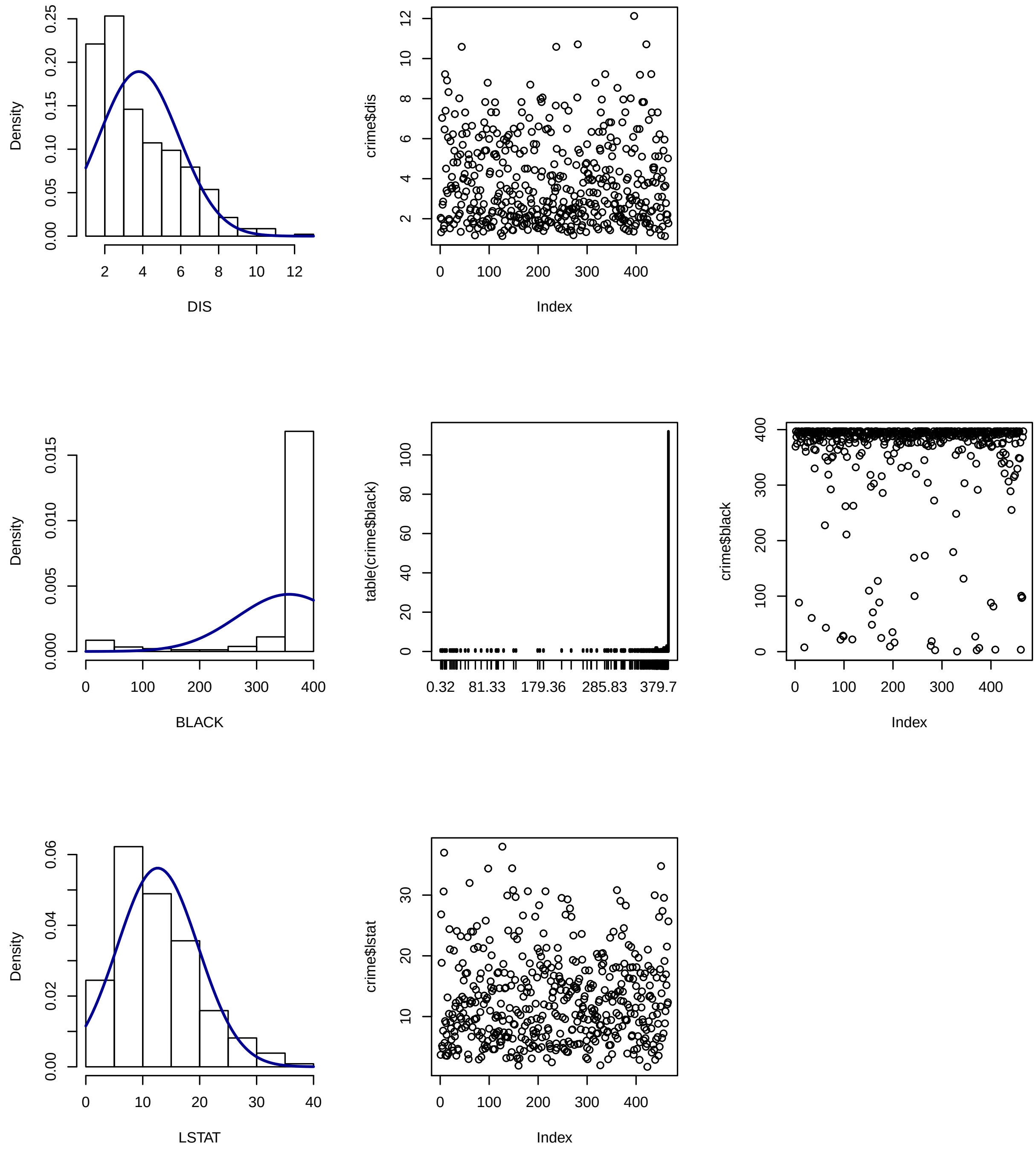


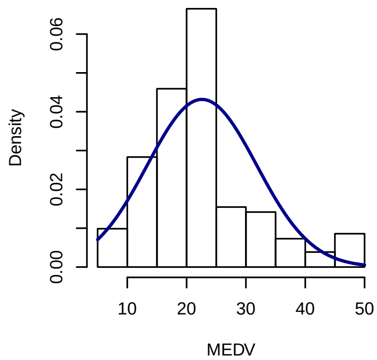
Based on an analysis of the box plots, the following variables have some outliers that may, or may not, exert influence on the regression results.
- zn, rm, dis, black, lstat, medv

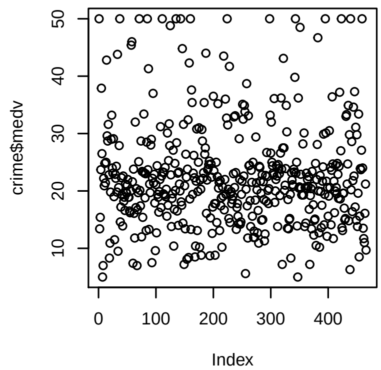
We'll next look at these variable mroe closely, starting with there histograms and frequency counts to better understand the nature of their distribution.






Density
DIS
crime$dis
Index
BLACK
table(crime$black)
0.32 81.33 179.36 285.83 379.7
crime$black
LSTAT
crime$lstat





According to the description, the variables zn, indus, and age are area, or land, proportions. According to the statistical summary, the values for these variables are all within the range [1,100] as you would expect.

Based on our detailed review of the variables that contained outliers, the following variables could be problematic:

The predictor variable zn is highly right skewed, we can confirm this by comparing the median and mean where the median is 0.0, but the median is 11.58. The frequency count plot shows how poor the distribution is due to clustering of the data at one extreme.

The predictor variable black is highly left skewed. We can confirm this by comparing the median and mean where the median is 391.34 and the mean is 357.12. The frequency count plot shows how poor the distribution is due to clustering of the data at one extreme.

The predictor variable dis is slightly right skewed. We can confirm this by comparing the median and mean where the median is 3.191 and the mean is 3.796.

Fortunately, no missing data, or NAs, were found.

The following data corrections were identified in this section:

(1) The predictor variable “chas” and the response variable “target” are supposed to be categorical (binary), so we need to convert them to factors.

(2) Need to determine if there are other variables highly coorelated with the zn or black variable that don’t have the severe skew and outliers. This would allow us to remove the zn or black variable from the model.

Data Preparation

The variable changes we identified so far include converting the predictor variable “chas” and the response variable “target” to factors.

	zn	indus	nox	rm	age	dis
zn	1.0000000	-0.5382664	-0.5170452	0.3198141	-0.5725805	0.6601243
indus	-0.5382664	1.0000000	0.7596301	-0.3927118	0.6395818	-0.7036189
nox	-0.5170452	0.7596301	1.0000000	-0.2954897	0.7351278	-0.7688840
rm	0.3198141	-0.3927118	-0.2954897	1.0000000	-0.2328125	0.1990158
age	-0.5725805	0.6395818	0.7351278	-0.2328125	1.0000000	-0.7508976
dis	0.6601243	-0.7036189	-0.7688840	0.1990158	-0.7508976	1.0000000
rad	-0.3154812	0.6006284	0.5958298	-0.2084457	0.4603143	-0.4949919
tax	-0.3192841	0.7322292	0.6538780	-0.2969343	0.5121245	-0.5342546
ptratio	-0.3910357	0.3946898	0.1762687	-0.3603471	0.2554479	-0.2333394

	zn	indus	nox	rm	age	dis
black	0.1794150	-0.3581356	-0.3801549	0.1326676	-0.2734677	0.2938441
lstat	-0.4329925	0.6071102	0.5962426	-0.6320245	0.6056200	-0.5075280
medv	0.3767171	-0.4961743	-0.4301227	0.7053368	-0.3781560	0.2566948

	rad	tax	ptratio	black	lstat	medv
zn	-0.3154812	-0.3192841	-0.3910357	0.1794150	-0.4329925	0.3767171
indus	0.6006284	0.7322292	0.3946898	-0.3581356	0.6071102	-0.4961743
nox	0.5958298	0.6538780	0.1762687	-0.3801549	0.5962426	-0.4301227
rm	-0.2084457	-0.2969343	-0.3603471	0.1326676	-0.6320245	0.7053368
age	0.4603143	0.5121245	0.2554479	-0.2734677	0.6056200	-0.3781560
dis	-0.4949919	-0.5342546	-0.2333394	0.2938441	-0.5075280	0.2566948
rad	1.0000000	0.9064632	0.4714516	-0.4463750	0.5031013	-0.3976683
tax	0.9064632	1.0000000	0.4744223	-0.4425059	0.5641886	-0.4900329
ptratio	0.4714516	0.4744223	1.0000000	-0.1816395	0.3773560	-0.5159153
black	-0.4463750	-0.4425059	-0.1816395	1.0000000	-0.3533659	0.3300286
lstat	0.5031013	0.5641886	0.3773560	-0.3533659	1.0000000	-0.7358008
medv	-0.3976683	-0.4900329	-0.5159153	0.3300286	-0.7358008	1.0000000

Based on the correlation table, the variable zn has a moderate correlation with the variable dis. The plot of the dis data shows a much better distribution of values. Consequently, one possibility is to remove the zn variable from the data set for modeling.

Build Models

One analysis of multiple regression models is to take a stepwise approach, and to begin this step, we first take our knowledge from the data exploration, and combine it with a logistic regression. The Univariate Logistic Regression is a useful tool to understand how each variable plays against our target variable. Looking at various statistics, we can which variable impacts are target the most.

var	p_val	aic	auc
zn	0.0000000	413.2878	0.7076814
indus	0.0000000	345.8163	0.8091513
chas1	0.3188437	518.3011	0.5452821
nox	0.0000000	212.6269	0.8710289
rm	0.0010624	507.8644	0.5737316
age	0.0000000	317.3847	0.7937411
dis	0.0000000	307.0926	0.7970602
rad	0.0000015	330.3616	0.8440019
tax	0.0000000	353.7222	0.8319109
ptratio	0.0000011	493.3566	0.6600284
black	0.0000018	435.2948	0.7484590
lstat	0.0000000	416.8908	0.7015173

We took the p-value, the AIC statistic and then a measure of the Area under the curve to measure the variables potential in a multiple regression model. From the above table, we can see that the chas variable is least likely to be included in our model, as it isn't statistically signficant. From the above table, we can see 1 variable that has no significance and under a univariate regression model, and have high relative AIC, and accuracy that is barely higher than a random variable. The Chas variable is a viable candidate to remove from our modelling.

Model 1

A quick looke at the total model:

```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7132  -0.0934   0.0000   0.0016   3.4718
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -24.342449   9.762679  -2.493 0.012652 *
## zn           -0.038247   0.038733  -0.987 0.323420
## indus        -0.082035   0.066940  -1.225 0.220391
## chas1         1.189371   0.904623   1.315 0.188587
## nox          53.285171  10.168667   5.240  1.6e-07 ***
## rm           -1.183564   0.917904  -1.289 0.197252
## age           0.054774   0.016677   3.284 0.001022 **
## dis           0.710750   0.286890   2.477 0.013233 *
## rad           0.703069   0.203161   3.461 0.000539 ***
## tax          -0.010313   0.004648  -2.219 0.026491 *
## ptratio       0.560259   0.180922   3.097 0.001957 **
## black        -0.044213   0.018559  -2.382 0.017206 *
## lstat        -0.046652   0.067660  -0.690 0.490500
## medv          0.187979   0.084565   2.223 0.026223 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.31  on 371  degrees of freedom
## Residual deviance: 130.18  on 358  degrees of freedom
## AIC: 158.18
##
## Number of Fisher Scoring iterations: 9
```

```
## [1] 0.9506875
```

Model 2

We will attempt to create the simplest model possible by using only one variable - the one that provides us the highest overall AUC (performance) all by itself. We can plug in each variable separately and then select the highest result. The best variable is nox - the presence of nitrogen oxides (an industrial pollutant) on the property.

```
## [1] 0.8710289
```

By combining nos with all the remaining variables and selecting the highest resulting AUC result, we conclude that nox plus rad (access to radial highways) is the strongest combinaton of two variables.

```
## [1] 0.9338549
```

```
## [1] 0.9279279
```

By combining three variables - nox, rad and zn - that is, the concentration of nitrogen oxides, access to radial highways and the proportion of land zoned for large lots, we can predict with 95.8% accuracy whether the crime rate at this property is above or below average. Since this is very close to the performance of the model using all variables (96%), we can be confident in using these three variables for our decision support process, and disregarding the others.

Model 3

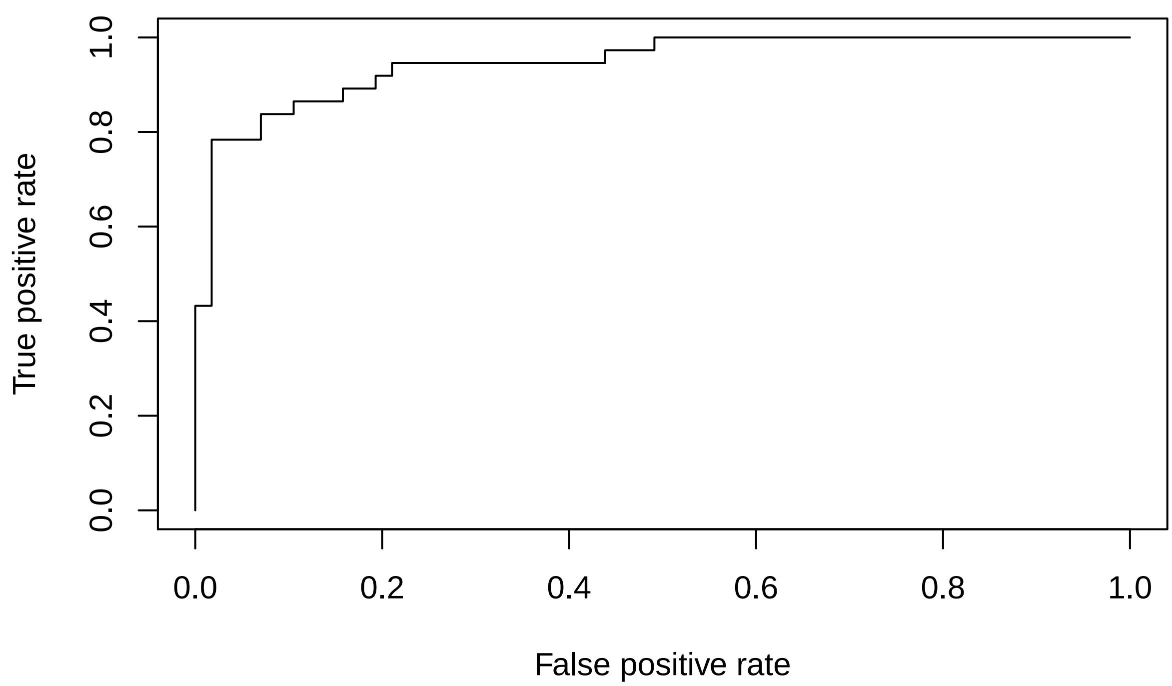
MODEL 3 WITH NOX VARIABLE

```
## Start:  AIC=158.18
## target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + black + lstat + medv
##
##           Df Deviance    AIC
## - lstat    1   130.66 156.66
## - zn       1   131.32 157.32
## - indus    1   131.71 157.71
## - rm       1   131.88 157.88
## - chas     1   131.90 157.90
## <none>         130.18 158.18
## - medv     1   135.74 161.74
## - tax      1   135.83 161.83
## - dis      1   137.13 163.13
## - black    1   141.32 167.32
## - ptratio  1   141.36 167.36
## - age      1   142.62 168.62
## - rad      1   160.19 186.19
## - nox      1   179.04 205.04
##
## Step:  AIC=156.66
## target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + black + medv
##
##           Df Deviance    AIC
## - rm       1   131.88 155.88
## - zn       1   131.97 155.97
## - chas     1   132.07 156.07
## - indus    1   132.13 156.13
## <none>         130.66 156.66
## - medv     1   135.85 159.85
## - tax      1   137.03 161.03
## - dis      1   137.28 161.28
## - ptratio  1   141.42 165.42
## - black    1   141.78 165.78
## - age      1   143.67 167.67
## - rad      1   161.04 185.04
## - nox      1   179.22 203.22
##
## Step:  AIC=155.88
## target ~ zn + indus + chas + nox + age + dis + rad + tax + ptratio +
```

```
##     black + medv
##
##           Df Deviance    AIC
## - indus    1   133.18 155.18
## - chas     1   133.31 155.31
## - zn       1   133.35 155.35
## <none>         131.88 155.88
## - dis      1   137.63 159.63
## - medv     1   138.52 160.52
## - tax      1   138.80 160.80
## - ptratio  1   141.43 163.43
## - black    1   143.36 165.36
## - age      1   143.79 165.79
## - rad      1   162.26 184.26
## - nox      1   179.24 201.24
##
## Step:  AIC=155.18
## target ~ zn + chas + nox + age + dis + rad + tax + ptratio +
##     black + medv
##
##           Df Deviance    AIC
## - chas     1   133.91 153.91
## - zn       1   134.74 154.74
## <none>         133.18 155.18
## - dis      1   138.30 158.30
## - medv     1   139.50 159.50
## - ptratio  1   141.70 161.70
## - black    1   144.01 164.01
## - age      1   144.79 164.79
## - tax      1   147.18 167.18
## - rad      1   169.58 189.58
## - nox      1   185.71 205.71
##
## Step:  AIC=153.91
## target ~ zn + nox + age + dis + rad + tax + ptratio + black +
##     medv
##
##           Df Deviance    AIC
## <none>         133.91 153.91
## - zn       1   135.92 153.92
## - dis      1   138.75 156.75
## - medv     1   139.87 157.87
## - ptratio  1   141.76 159.76
## - black    1   144.48 162.48
## - age      1   146.79 164.79
## - tax      1   149.18 167.18
## - rad      1   174.36 192.36
## - nox      1   185.81 203.81

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

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8622  -0.1135   0.0000   0.0018   3.3120
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -21.935844   9.197361  -2.385 0.017078 *
## zn           -0.048462   0.037132  -1.305 0.191848
## nox          44.193814   7.922146   5.579 2.43e-08 ***
## age           0.043782   0.013013   3.364 0.000767 ***
## dis           0.551173   0.260500   2.116 0.034359 *
## rad           0.764131   0.188980   4.043 5.27e-05 ***
## tax          -0.013328   0.004197  -3.176 0.001496 **
## ptratio       0.396941   0.145949   2.720 0.006534 **
## black        -0.041476   0.017228  -2.408 0.016062 *
## medv          0.097180   0.042373   2.293 0.021823 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.31  on 371  degrees of freedom
## Residual deviance: 133.91  on 362  degrees of freedom
## AIC: 153.91
##
## Number of Fisher Scoring iterations: 9
```



```
## [1] 0.9468943
```

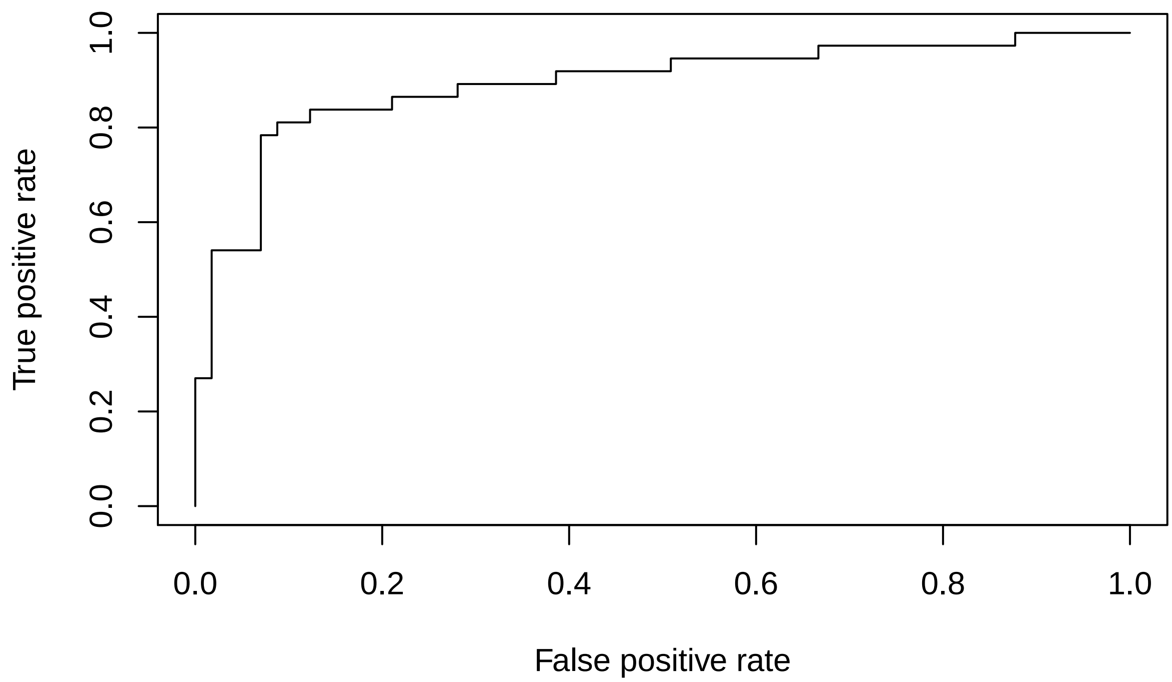
MODEL 3 WITHOUT NOX VARIABLE

```
## Start:  AIC=205.04
## target ~ (zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + black + lstat + medv) - nox
##
##           Df Deviance    AIC
## - rm       1   179.04 203.04
## - lstat    1   179.22 203.22
## - medv     1   179.49 203.49
## - ptratio  1   179.50 203.50
## - zn       1   179.72 203.72
## - chas     1   180.12 204.12
## <none>         179.04 205.04
## - dis      1   184.69 208.69
## - indus    1   185.26 209.26
## - tax      1   189.25 213.25
## - age      1   194.42 218.42
## - black    1   194.48 218.48
## - rad      1   224.28 248.28
##
## Step:  AIC=203.04
## target ~ zn + indus + chas + age + dis + rad + tax + ptratio +
##     black + lstat + medv
##
##           Df Deviance    AIC
## - lstat    1   179.24 201.24
## - ptratio  1   179.55 201.55
## - zn       1   179.74 201.74
## - chas     1   180.14 202.14
## - medv     1   180.33 202.33
## <none>         179.04 203.04
## - dis      1   184.75 206.75
## - indus    1   185.26 207.26
## - tax      1   189.48 211.48
## - black    1   194.49 216.49
## - age      1   199.65 221.65
## - rad      1   224.29 246.29
##
## Step:  AIC=201.23
## target ~ zn + indus + chas + age + dis + rad + tax + ptratio +
##     black + medv
##
##           Df Deviance    AIC
## - ptratio  1   179.66 199.66
## - zn       1   179.94 199.94
## - chas     1   180.24 200.24
## - medv     1   180.39 200.39
## <none>         179.24 201.24
## - dis      1   184.87 204.87
## - indus    1   185.71 205.71
```

```
## - tax      1   189.73 209.73
## - black    1   194.65 214.65
## - age      1   203.76 223.76
## - rad      1   224.87 244.87
##
## Step:  AIC=199.66
## target ~ zn + indus + chas + age + dis + rad + tax + black +
##     medv
##
##         Df Deviance    AIC
## - medv   1   180.44 198.44
## - chas   1   180.84 198.84
## - zn     1   180.88 198.88
## <none>       179.66 199.66
## - dis    1   184.91 202.91
## - indus  1   186.30 204.30
## - tax    1   189.81 207.81
## - black  1   194.65 212.65
## - age    1   203.82 221.82
## - rad    1   224.87 242.87
##
## Step:  AIC=198.44
## target ~ zn + indus + chas + age + dis + rad + tax + black
##
##         Df Deviance    AIC
## - zn     1   181.18 197.18
## - chas   1   181.53 197.53
## <none>       180.44 198.44
## - indus  1   186.61 202.61
## - dis    1   190.36 206.36
## - tax    1   193.03 209.03
## - black  1   195.03 211.03
## - age    1   203.82 219.82
## - rad    1   229.86 245.86
##
## Step:  AIC=197.18
## target ~ indus + chas + age + dis + rad + tax + black
##
##         Df Deviance    AIC
## - chas   1   182.11 196.11
## <none>       181.18 197.18
## - indus  1   187.57 201.57
## - dis    1   192.85 206.85
## - tax    1   193.44 207.44
## - black  1   196.02 210.02
## - age    1   206.37 220.37
## - rad    1   230.37 244.37
##
## Step:  AIC=196.11
## target ~ indus + age + dis + rad + tax + black
##
##         Df Deviance    AIC
## <none>       182.11 196.11
## - indus  1   187.66 199.66
```

```
## - tax    1   193.45 205.45
## - dis    1   193.52 205.52
## - black  1   196.74 208.74
## - age    1   206.70 218.70
## - rad    1   231.11 243.11
```

```
##
## Call:
## glm(formula = target ~ indus + age + dis + rad + tax + black,
##     family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3129  -0.3415   0.0000   0.0139   2.6900
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 12.339156   5.329705   2.315  0.02060 *
## indus        0.112098   0.050765   2.208  0.02723 *
## age          0.047489   0.010497   4.524 6.06e-06 ***
## dis         -0.475263   0.148846  -3.193  0.00141 **
## rad          0.647494   0.160850   4.025 5.69e-05 ***
## tax         -0.011195   0.003676  -3.045  0.00232 **
## black       -0.039095   0.013024  -3.002  0.00268 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.31  on 371  degrees of freedom
## Residual deviance: 182.11  on 365  degrees of freedom
## AIC: 196.11
##
## Number of Fisher Scoring iterations: 9
```




```
## [1] 0.8933144
```

All Done!