DATA621-Homework3-HoddeFarrisBurmoodLin

Rob Hodde, Matt Farris, Jeffrey Burmood, Bin Lin

4/05/2017

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

Data Exploration

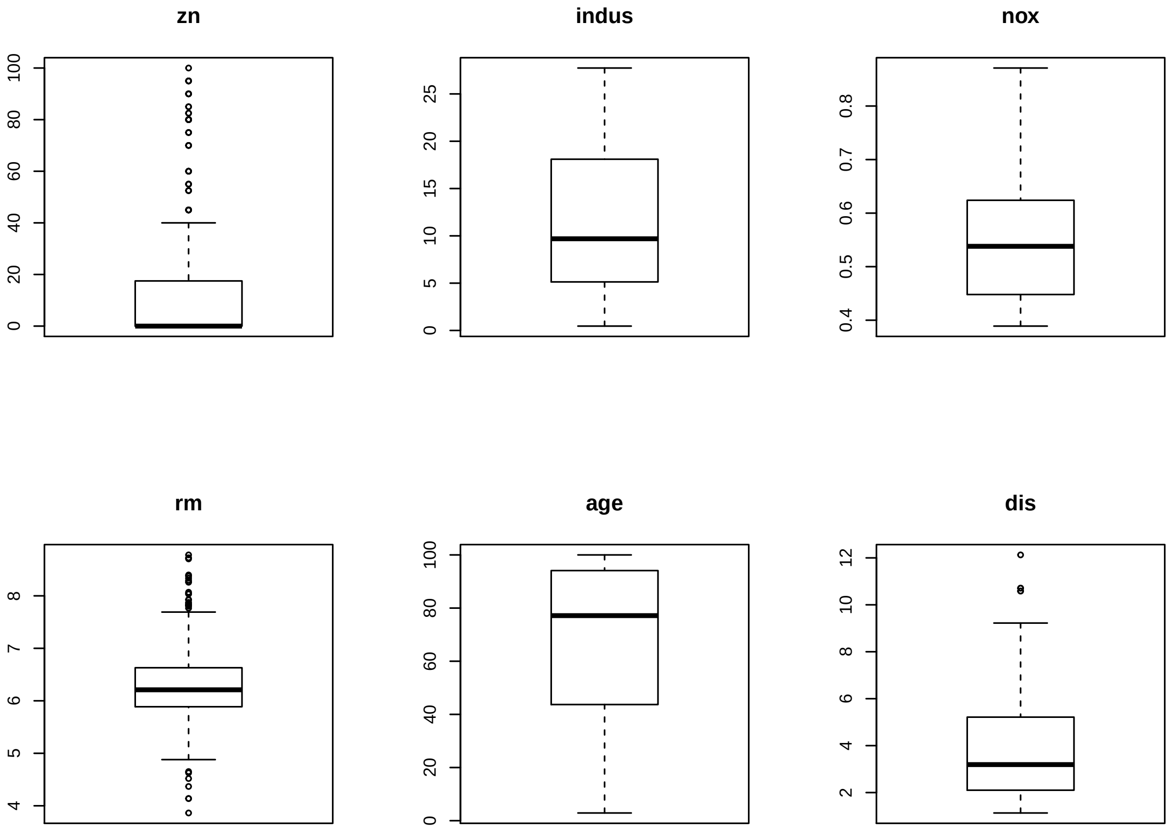
The first thirteen variables in the table below are potential predictor variables that could affect the response variable *target*.

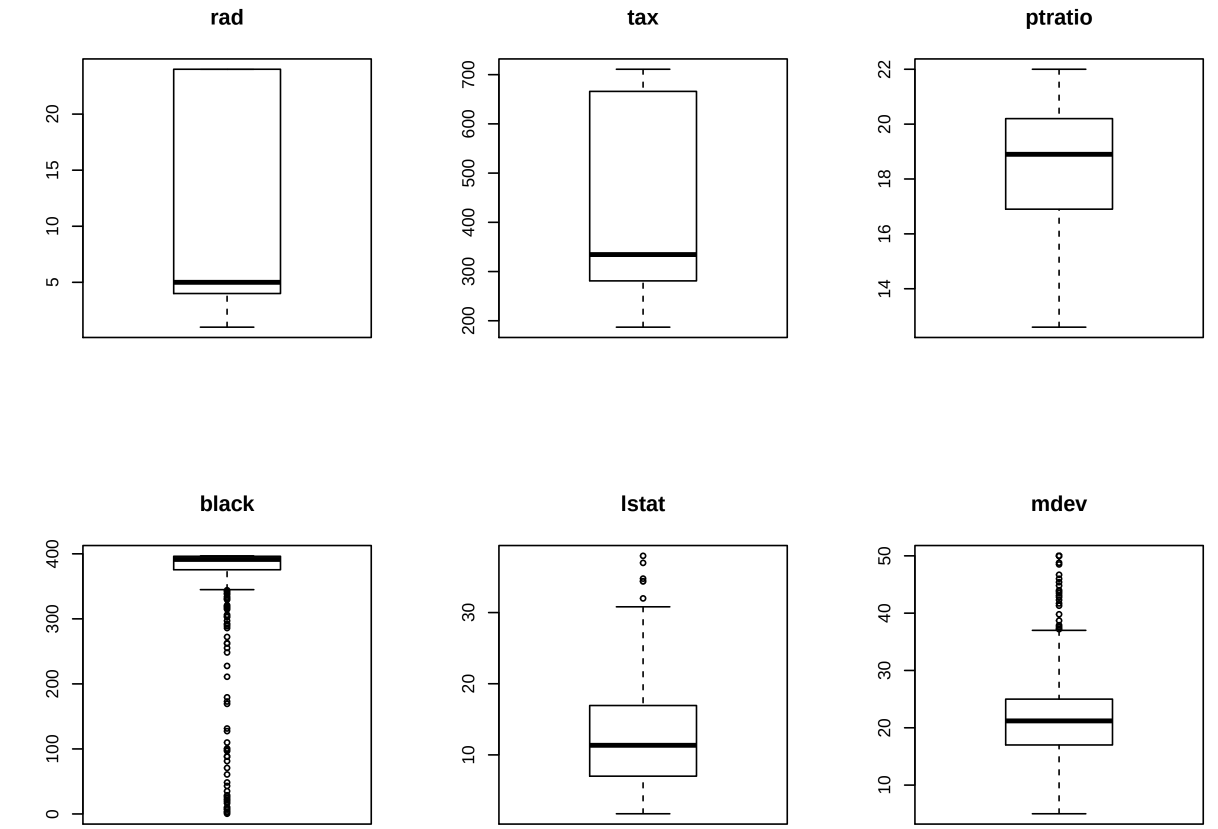
Variable	Type	Definition
zn	Double	proportion of residential land zoned for large lots
indus	Double	proportion of non-retail business acres per suburb
chas	Integer	suburb borders the Charles River (1 = Yes, 0 = No)
nox	Double	nitrogen oxides concentration (parts per 10 million)
rm	Double	average number of rooms per dwelling
age	Double	proportion of owner-occupied units built prior to 1940
dis	Double	mean of distances to five Boston employment centers
rad	Integer	index of accessibility to radial highways
tax	Integer	full-value property-tax rate per $10,000
ptratio	Double	pupil-teacher ratio by town
black	Double	1000(Bk - 0.63)2 where Bk = proportion of blacks by town
lstat	Double	lower status of the population (percent)
medv	Double	median value of owner-occupied homes in $1000s
target	Integer	crime rate is above the median crime rate (1 = Yes, 0 = No)

Below is a summary of each predictor variable's basic statistics, followed by boxplots which illustrate the spread and outliers for each variable.

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

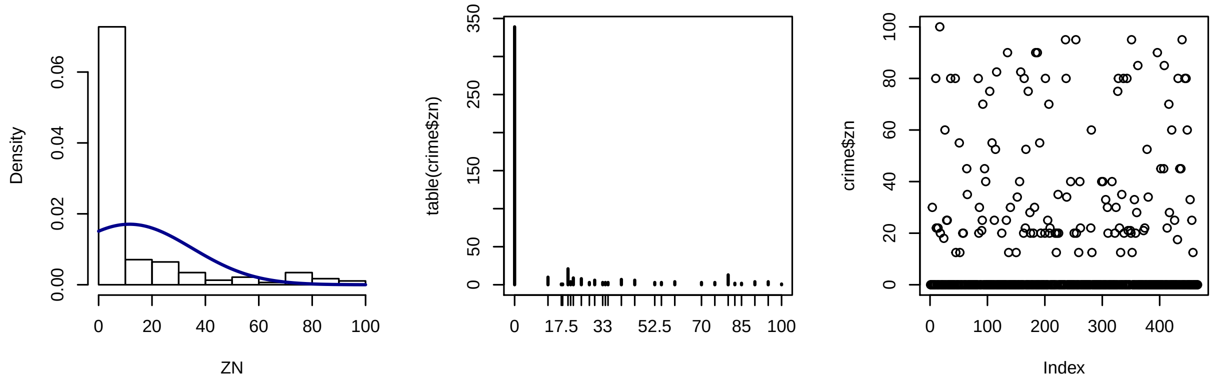
dis	rad	tax	ptratio	black	lstat
Min. : 1.130	Min. : 1.00	Min. :187.0	Min. :12.6	Min. : 0.32	Min. : 1.730
1st Qu.: 2.101	1st Qu.: 4.00	1st Qu.:281.0	1st Qu.:16.9	1st Qu.:375.61	1st Qu.: 7.043
Median : 3.191	Median : 5.00	Median :334.5	Median :18.9	Median :391.34	Median :11.350
Mean : 3.796	Mean : 9.53	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

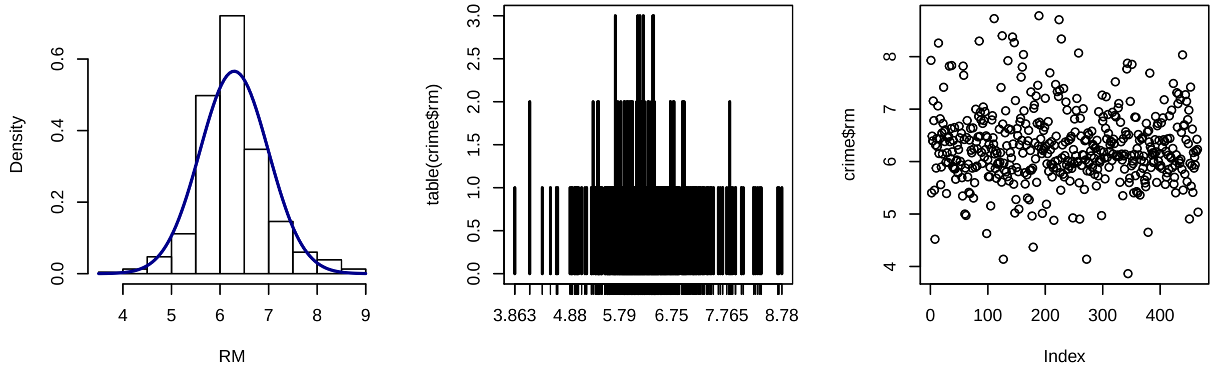


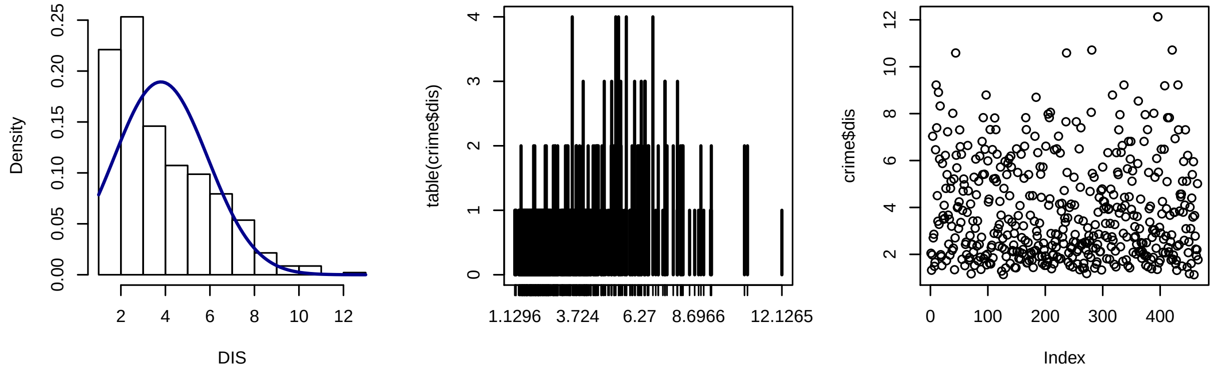


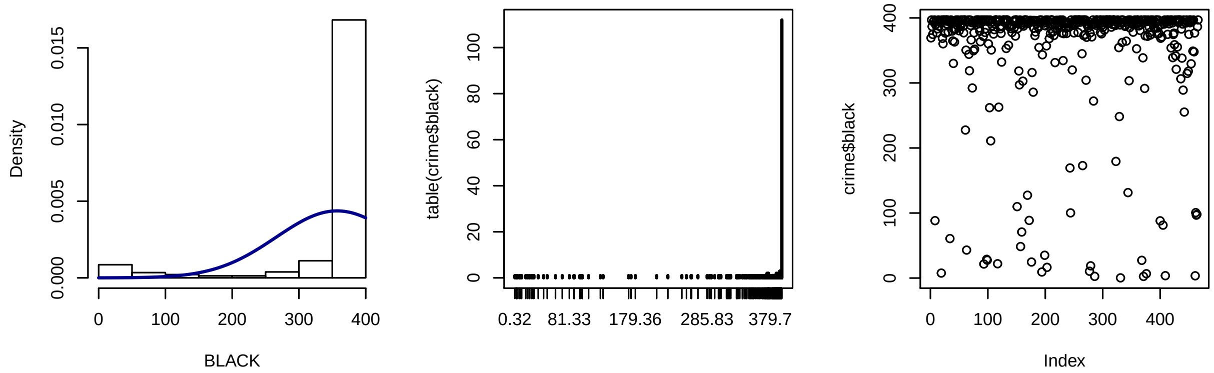
Based on an analysis of the box plots, the following variables have some outliers that may, or may not, exert influence on the regression results: - zn, rm, dis, black, lstat, medv

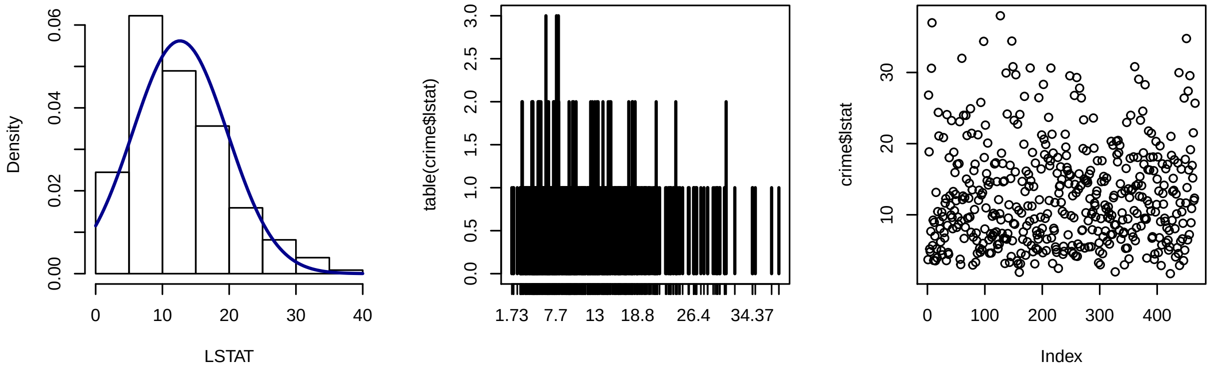
We'll next look at these variables more closely, starting with their histograms and frequency counts to better understand the nature of their distribution.

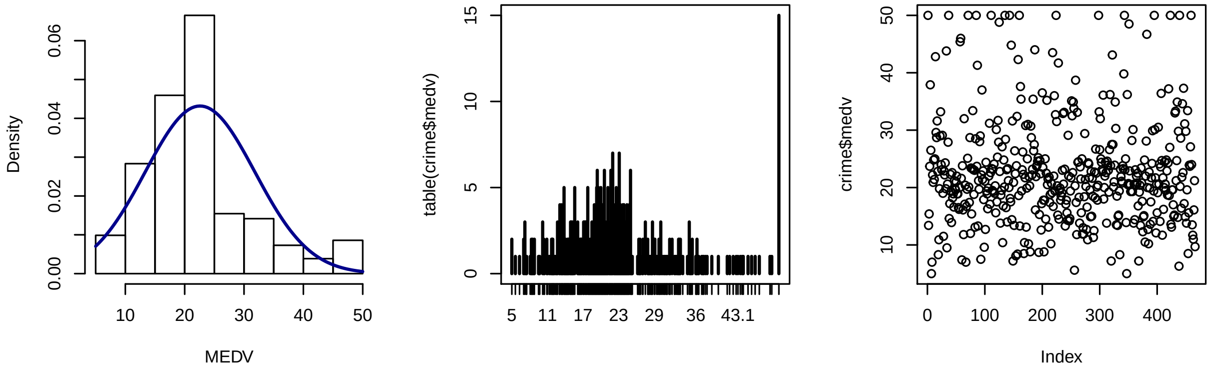



Density
0.0
0.2
0.4
0.6
4
5
6
7
8
9
RM
table(crime$rm)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.863
4.88
5.79
6.75
7.765
8.78
crime$rm
4
5
6
7
8
0
100
200
300
400
Index


Density
0.00
0.05
0.10
0.15
0.20
0.25
2
4
6
8
10
12
DIS
table(crime$dis)
0
1
2
3
4
1.1296
3.724
6.27
8.6966
12.1265
crime$dis
2
4
6
8
10
12
0
100
200
300
400
Index


Density
0.000
0.005
0.010
0.015
0
100
200
300
400
BLACK
table(crime$black)
0
20
40
60
80
100
0.32
81.33
179.36
285.83
379.7
crime$black
0
100
200
300
400
0
100
200
300
400
Index





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] that we 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 mean 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 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* variables that do not have the severe skew and outliers. This could allow us to remove the *zn* or *black* variables from the model.

Data Preparation

Data Preparation

The variable changes we identified so far include converting the predictor variable *chas* and the response variable *target* to factors. Next we will look at how each variable correlates to all the others:

	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
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

The correlation table above shows that the variable *zn* is moderately correlated to the variable *dis*. The plot of the *dis* data shows a much better distribution of values. Consequently, one possibility is to remove *zn* from the model and use *dis* instead. Before doing this, we should look at the real-world context of the two variables to determine if they are meaningfully related.

Build Models

Build Models

One method of developing multiple regression models is to take a stepwise approach. To accomplish this, we combine our knowledge from the data exploration above with logistic regression. Univariate Logistic

Regression is a useful method to understand how each predictor variable interacts individually with the target (response) variable. Looking at various statistics, we determine which variable may impact our 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

Here we see the selected output criteria for the linear models run with only a single predictor variable. We examine the p-value (significance), the AIC statistic (goodness-of-fit) and the AUC (Area Under Curve) to measure the potential predictive value of each variable, so we can decide whether or not to include it in our multiple regression model. We are looking for p-values below .05, AIC values as low as possible, and AUC values as high as possible.

From the above table, we can see that *chas* is the least likely to produce any meaningful inference because its p-value is well above .05 (not significant), it has the highest AIC (518, where 100 is considered excellent), and the lowest AUC (.54, where random chance would yield .50). Therefore, *chas* is the most likely candidate to be removed from our model.

Model 1

As a baseline, we start with a multiple logistic regression model that includes every predictor variable:

```
##
## 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
```

```
##           Reference
## Prediction  0  1
##          0 48  4
##          1  9 33
```

In this model-and in all the models- we set aside 20% of the training data and use 80% to train the model we then use the model to predict the outcome of the remaining 20% of the data. The model yields an Area Under Curve of .95, meaning it chose correctly 95% of the time.

Model 2

In this scenario we attempt to create the simplest model possible by using only one variable - the one that provides the highest overall AUC (performance) by itself. We calculate AUC for each variable separately and then select the highest result.

```
##
## Call:
## glm(formula = target ~ nox, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3967  -0.3300   0.0065   0.3147   2.6197
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -17.271      1.803  -9.577   <2e-16 ***
## nox           32.188      3.386   9.507   <2e-16 ***
## ---
## 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: 208.63  on 370  degrees of freedom
## AIC: 212.63
##
```

```
## Number of Fisher Scoring iterations: 6
```

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

The best predictor variable is *nox*, yielding an AUC of .87.

Next we combine *nox* with each of the remaining variables individually and select the highest AUC result.

```
##
## Call:
## glm(formula = target ~ nox + rad, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.02641  -0.31179   0.00106   0.01272   2.50496
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -17.6602     2.2085  -7.997 1.28e-15 ***
## nox          28.7157     3.7460   7.666 1.78e-14 ***
## rad           0.4144     0.1260   3.289  0.00101 **
## ---
## 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: 180.80  on 369  degrees of freedom
## AIC: 186.8
##
## Number of Fisher Scoring iterations: 8
```

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

We find that *nox* plus *rad* is the strongest combinaton of two variables, yielding an AUC of .93.

Finally, we search for a third critical predictor by combining *nos* plus *rad* with the remaining variables, individually.

```
##
## Call:
## glm(formula = target ~ nox + rad + zn, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.98598  -0.29067   0.00113   0.01409   2.64355
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -16.45904    2.27151  -7.246 4.30e-13 ***
## nox          26.55221    3.85722   6.884 5.83e-12 ***
## rad           0.42817    0.13029   3.286  0.00102 **
## zn           -0.03045    0.02176  -1.400  0.16163
## ---
## 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: 178.43  on 368  degrees of freedom
## AIC: 186.43
##
## Number of Fisher Scoring iterations: 8
```

```
## [1] 0.9419156
```

```
##           Reference
## Prediction  0  1
##          0 51  7
##          1  6 30
```

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 94% 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 (95%), we can be confident in using these three variables for our decision support process, and disregarding the others.

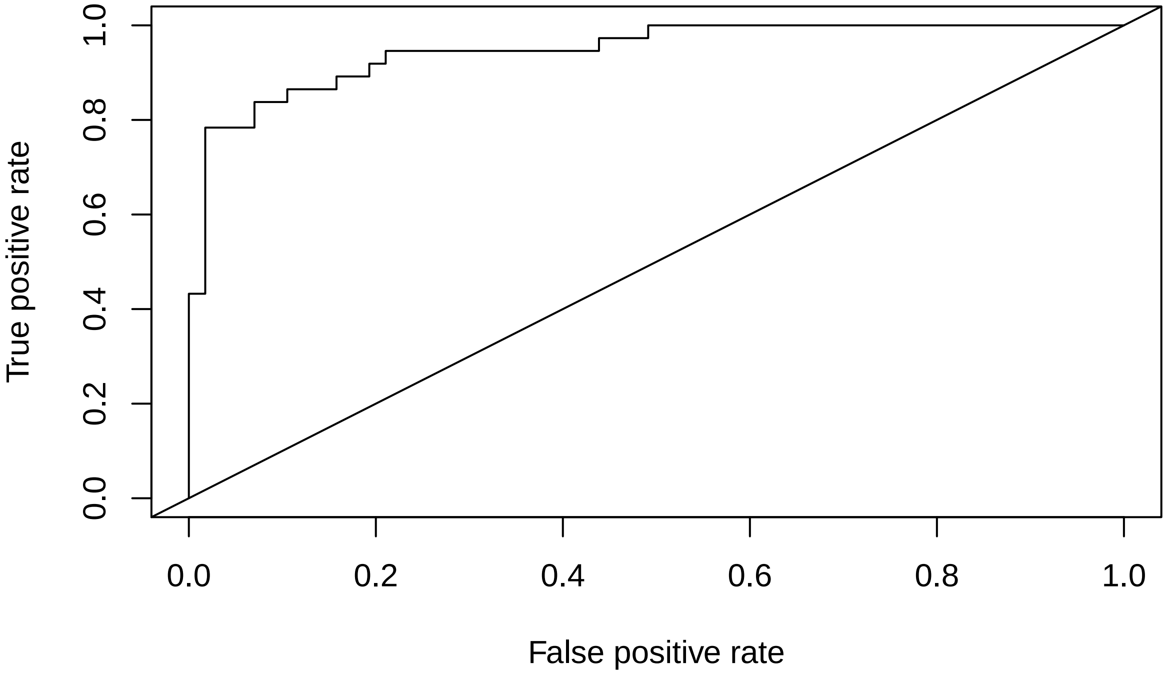
Model 3

The GLM Model summary in Model 1 illustrates the outsize impact of the predictor variable *nox* compared to all the others. It carries an Estimate of 53.3 where the next closest in magnitude is only 1.2. We thought it would be interesting to remove *nox* from the model just to see how the other variables perform without it. First we will perform a simple backward variable selection optimization process including it.

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
```

```
##           Reference
## Prediction  0  1
##          0 48  5
##          1  9 32
```

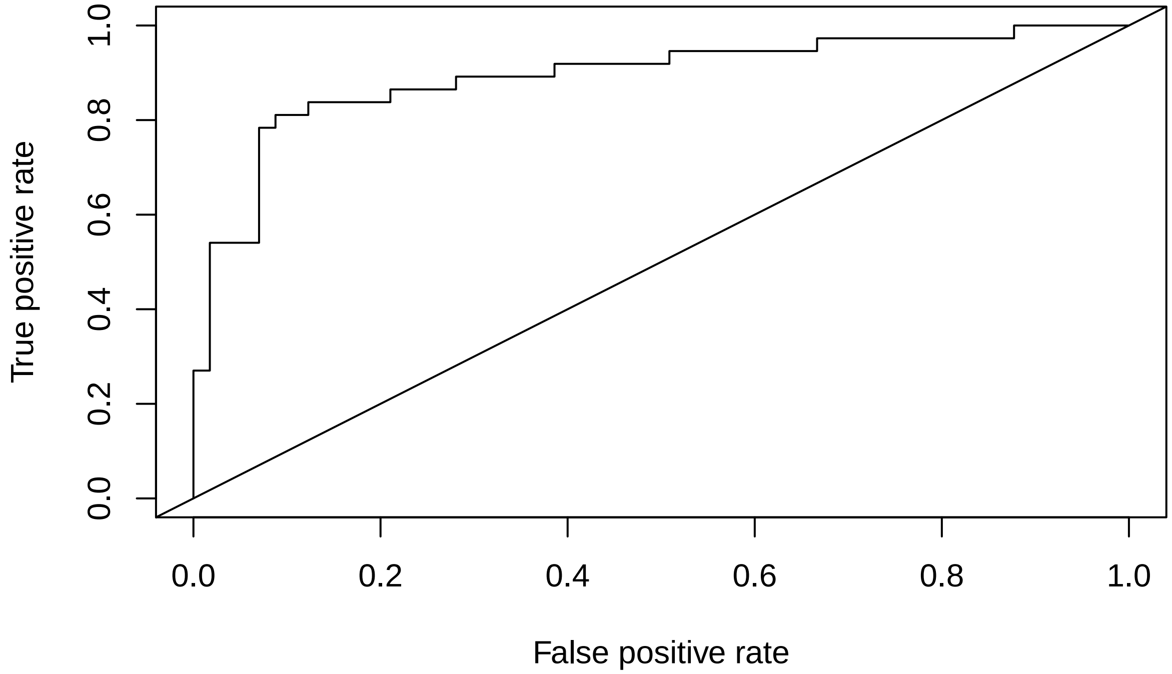
The model reduces to nine variables and yields a nice low residual deviance of 133.9, compared to a null deviance of 515.3. This roughly means that the model eliminates about 80% of the error compared to choosing at random. The AUC is .947 which is roughly the same as the full model using all variables.

Let's look at what happens when we remove the *nox* variable:

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
```

We still have a good model - the Residual Deviance increased to 182, but that is still much better than predicting with no model at all. The AUC is now .89 - again, very good. But the AUC with only one variable *nox* was .87. And in certain trials the AUC with *nox* exceeded .95 (due to randomly selected evaluation samples).

Why is the *nox* variable so powerful? We can look back at the Correlation table for clues. More variables are significantly correlated to *nox* than any other. It is like a super-variable, somehow encapsulating the properties of the variables around it. Is it because *nox* is an indicator of so many problems, like pollution, industrial decay, lax building codes? The *nox* variable is a stellar example of a finding that opens up many paths for further research.

Below is table illustrating the various fitness parameters that describe the effectiveness of the models. All the models are good - from a practical perspective, there is no difference between them.

Parameters	Model1	Model2	Model3
Accuracy	0.8617021	0.8617021	0.8510638
Classification Error Rate	0.1382979	0.1382979	0.1489362
Precision	NA	NA	NA
Sensitivity	0.8421053	0.8947368	0.8421053
Specificity	0.8918919	0.8108108	0.8648649
F1 Score	NA	NA	NA

Choose Model

Choose Model

We like *Model 3 With Nox* the best because it eliminates some of the questionable variables - the ones with high skew and many outliers, also it eliminates the chas variable, which was shown earlier as being insignificant. Ridding the model of these variables helps provide insurance against poor decisions that could arise, *even if they do not show up in the model.*

MODEL 3 WITH NOX VARIABLE USING FULL DATASETS

```
##
## Call:
## glm(formula = target ~ zn + nox + age + dis + rad + tax + ptratio +
##     black + medv, family = binomial(link = "logit"), data = crime)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2719  -0.1695  -0.0022   0.0022   3.4083
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -33.252218   6.510006  -5.108 3.26e-07 ***
## zn           -0.065747   0.031905  -2.061  0.03933 *
## nox          42.893366   6.744624   6.360 2.02e-10 ***
## age           0.031946   0.010928   2.923  0.00346 **
## dis           0.661897   0.216100   3.063  0.00219 **
## rad           0.724580   0.150914   4.801 1.58e-06 ***
## tax          -0.008216   0.002731  -3.009  0.00262 **
## ptratio       0.339874   0.114950   2.957  0.00311 **
## black        -0.011726   0.006535  -1.794  0.07276 .
## medv          0.117392   0.036009   3.260  0.00111 **
## ---
## 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: 192.57  on 456  degrees of freedom
## AIC: 212.57
##
## Number of Fisher Scoring iterations: 9
```

zn	nox	age	dis	rad	tax	ptratio	black	medv	predict.prob	predict.result
0	0.469	61.1	4.9671	2	242	17.8	392.83	34.7	0.9500179	1
0	0.538	84.5	4.4619	4	307	21.0	380.02	18.2	0.5159788	1
0	0.538	94.4	4.4547	4	307	21.0	387.94	18.4	0.4898765	0
0	0.538	82.0	3.9900	4	307	21.0	232.60	13.2	0.4469112	0
0	0.499	41.5	3.9342	5	279	19.2	396.90	21.0	0.9091632	1
25	0.453	66.2	7.2254	8	284	19.7	395.11	18.7	0.7407794	1
25	0.453	93.4	6.8185	8	284	19.7	378.08	16.0	0.6637665	1
0	0.449	56.1	4.4377	3	247	18.5	392.30	26.6	0.9868215	1
0	0.449	56.8	3.7476	3	247	18.5	395.15	22.2	0.9950137	1
0	0.445	69.6	3.4952	2	276	18.0	391.83	21.4	0.9983607	1
0	0.581	97.0	1.9444	2	188	19.1	370.31	17.3	0.5858120	1
0	0.581	95.6	1.7572	2	188	19.1	359.29	15.7	0.6200110	1
0	0.624	94.7	1.9799	4	437	21.2	396.90	14.3	0.4492920	0
0	0.605	93.0	2.2834	5	403	14.7	240.16	25.0	0.3940307	0
0	0.605	97.3	2.3887	5	403	14.7	348.13	19.1	0.4957219	0
0	0.489	92.1	3.8771	4	277	18.6	393.25	21.7	0.8768745	1
0	0.504	21.4	3.3751	8	307	17.4	380.34	31.5	0.6518693	1
0	0.507	70.4	3.6519	8	307	17.4	378.95	48.3	0.3780390	0
22	0.431	6.8	8.9067	7	330	19.1	386.09	29.6	0.9128547	1
90	0.400	20.8	7.3073	1	285	15.3	394.72	32.2	0.9999990	1
80	0.385	31.5	9.0892	1	241	18.2	341.60	20.1	0.9999916	1
33	0.472	58.1	3.3700	7	222	18.4	393.36	28.4	0.9475884	1
0	0.544	52.8	2.6403	4	304	18.4	396.90	22.1	0.8769392	1
0	0.493	40.1	4.7211	5	287	19.6	396.90	25.0	0.8323536	1
0	0.493	28.9	5.4159	5	287	19.6	396.90	23.0	0.8483649	1
0	0.515	59.6	5.6150	5	224	20.2	394.81	18.5	0.5185803	1
80	0.435	29.7	8.3440	4	280	17.0	390.94	24.5	0.9998358	1
0	0.718	87.9	1.6132	24	666	20.2	354.70	27.5	0.3678794	0
0	0.631	97.5	1.2024	24	666	20.2	392.05	50.0	0.3678794	0
0	0.584	86.1	2.0527	24	666	20.2	83.45	14.5	0.3678795	0
0	0.740	87.9	1.8206	24	666	20.2	68.95	8.4	0.3678794	0
0	0.740	93.9	1.8172	24	666	20.2	396.90	12.8	0.3678794	0
0	0.740	92.4	1.8662	24	666	20.2	391.45	10.5	0.3678794	0
0	0.740	100.0	2.0048	24	666	20.2	395.69	18.4	0.3678794	0
0	0.740	96.6	1.8956	24	666	20.2	240.52	10.8	0.3678794	0
0	0.713	86.5	2.4358	24	666	20.2	50.92	14.1	0.3678794	0
0	0.713	88.4	2.5671	24	666	20.2	391.43	17.7	0.3678794	0
0	0.655	65.4	2.9634	24	666	20.2	396.90	21.4	0.3678795	0
0	0.585	70.6	2.8927	6	391	19.2	396.90	18.3	0.4666070	0
0	0.573	91.0	2.1675	1	273	21.0	396.90	23.9	0.6955442	1

The Smooth Operators of R Fusion Have Struck Again.