DATA 621 HW3

Ahsanul Choudhury

April 15, 2018

Contents

Introduction

The purpose of this analysis is to build a logistic regression model that will predict whether a particular neighborhood of a major city will be at risk for high crime levels.

Data Exploration

Let's look at the data first; there are 466 observations and 13 variables, following table contains the names of the variable and a brief description of each variable:

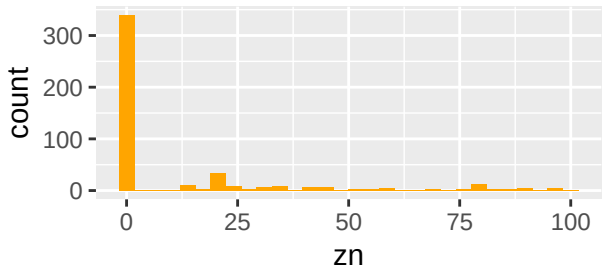
Variables	Description
zn	proportion of residential land zoned for large lots (over 25000 square feet)
indus	proportion of non-retail business acres per suburb
chas	a dummy var. for whether the suburb borders the Charles River (1) or not (0)
nox	nitrogen oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted mean of distances to five Boston employment centers
rad	index of accessibility to radial highways
tax	full-value property-tax rate per $10,000
ptratio	pupil-teacher ratio by town
lstat	lower status of the population (percent)
medv	median value of owner-occupied homes in $1000s
target	whether the crime rate is above the median crime rate (1) or not (0)

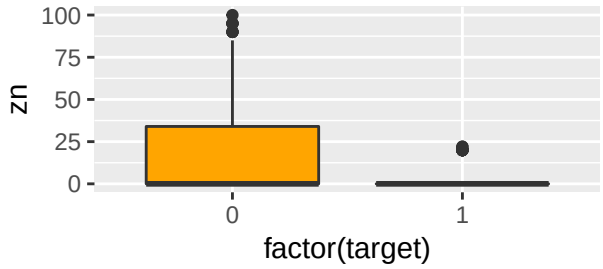
The first 12 variables are the predictor variables and the last variable, *target*, is the response variable. Also note the response variable is a binary variable.

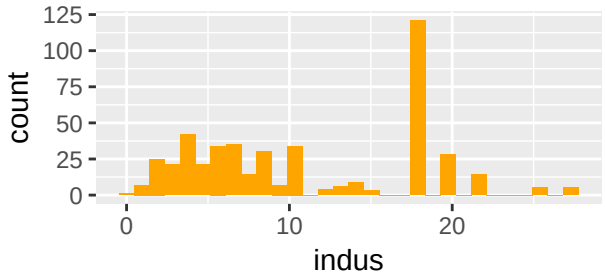
Now, lets checkout a brief descriptive summary of our variables:

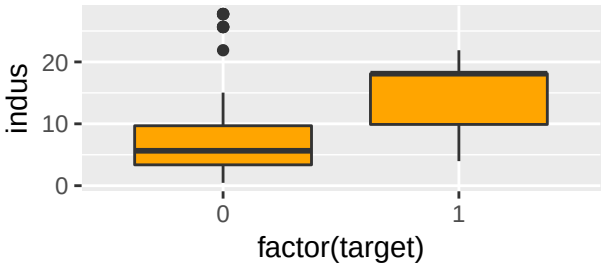
	mean	sd	median	min	max	skew	kurtosis	se
zn	11.5772532	23.3646511	0.00000	0.0000	100.0000	2.1768152	3.8135765	1.0823466
indus	11.1050215	6.8458549	9.69000	0.4600	27.7400	0.2885450	-1.2432132	0.3171281
chas	0.0708155	0.2567920	0.00000	0.0000	1.0000	3.3354899	9.1451313	0.0118957
nox	0.5543105	0.1166667	0.53800	0.3890	0.8710	0.7463281	-0.0357736	0.0054045
rm	6.2906738	0.7048513	6.21000	3.8630	8.7800	0.4793202	1.5424378	0.0326516
age	68.3675966	28.3213784	77.15000	2.9000	100.0000	-0.5777075	-1.0098814	1.3119625
dis	3.7956929	2.1069496	3.19095	1.1296	12.1265	0.9988926	0.4719679	0.0976026
rad	9.5300429	8.6859272	5.00000	1.0000	24.0000	1.0102788	-0.8619110	0.4023678
tax	409.5021459	167.9000887	334.50000	187.0000	711.0000	0.6593136	-1.1480456	7.7778214
ptratio	18.3984979	2.1968447	18.90000	12.6000	22.0000	-0.7542681	-0.4003627	0.1017669
lstat	12.6314592	7.1018907	11.35000	1.7300	37.9700	0.9055864	0.5033688	0.3289887
medv	22.5892704	9.2396814	21.20000	5.0000	50.0000	1.0766920	1.3737825	0.4280200

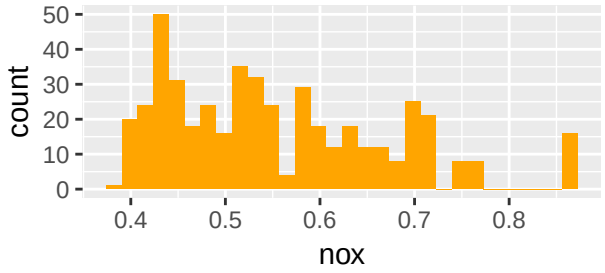
We will now plot a histogram and boxplot to see the distribution of the predictor variables and look for outliers:

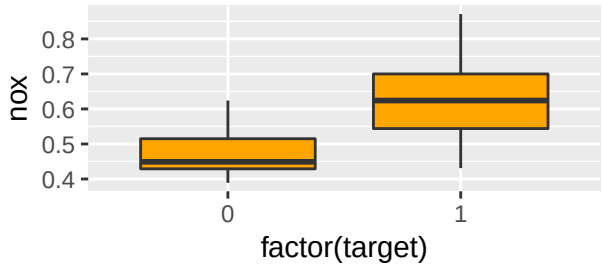

count
300
200
100
0
0
25
50
75
100
zn

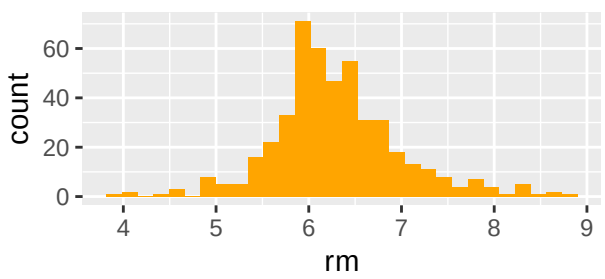

zn
100
75
50
25
0
0
1
factor(target)

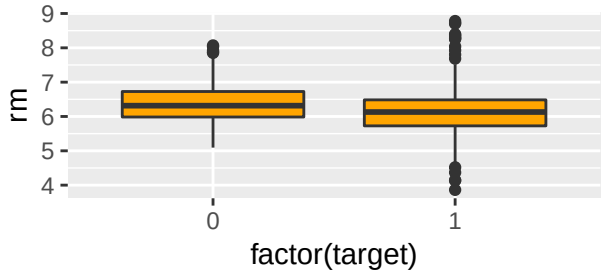

count
125
100
75
50
25
0
0
10
20
indus

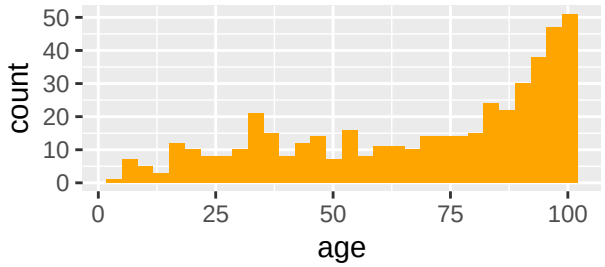

indus
20
10
0
0
1
factor(target)

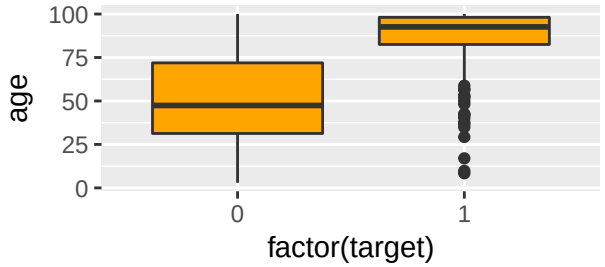

count
50
40
30
20
10
0
0.4
0.5
0.6
0.7
0.8
nox

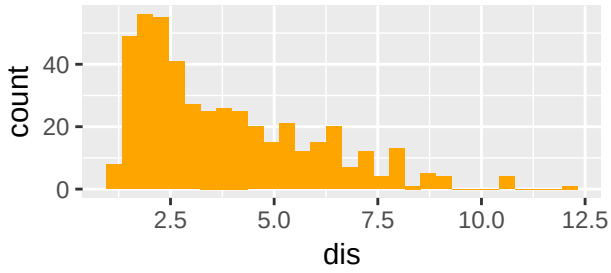

nox
0.8
0.7
0.6
0.5
0.4
0
1
factor(target)

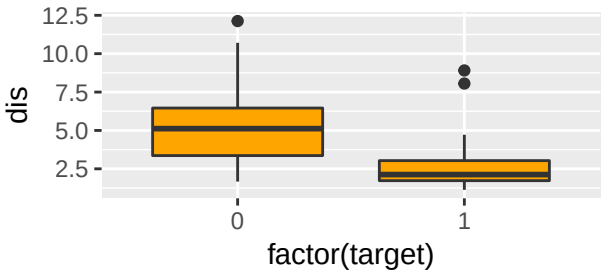

count
60
40
20
0
4
5
6
7
8
9
rm

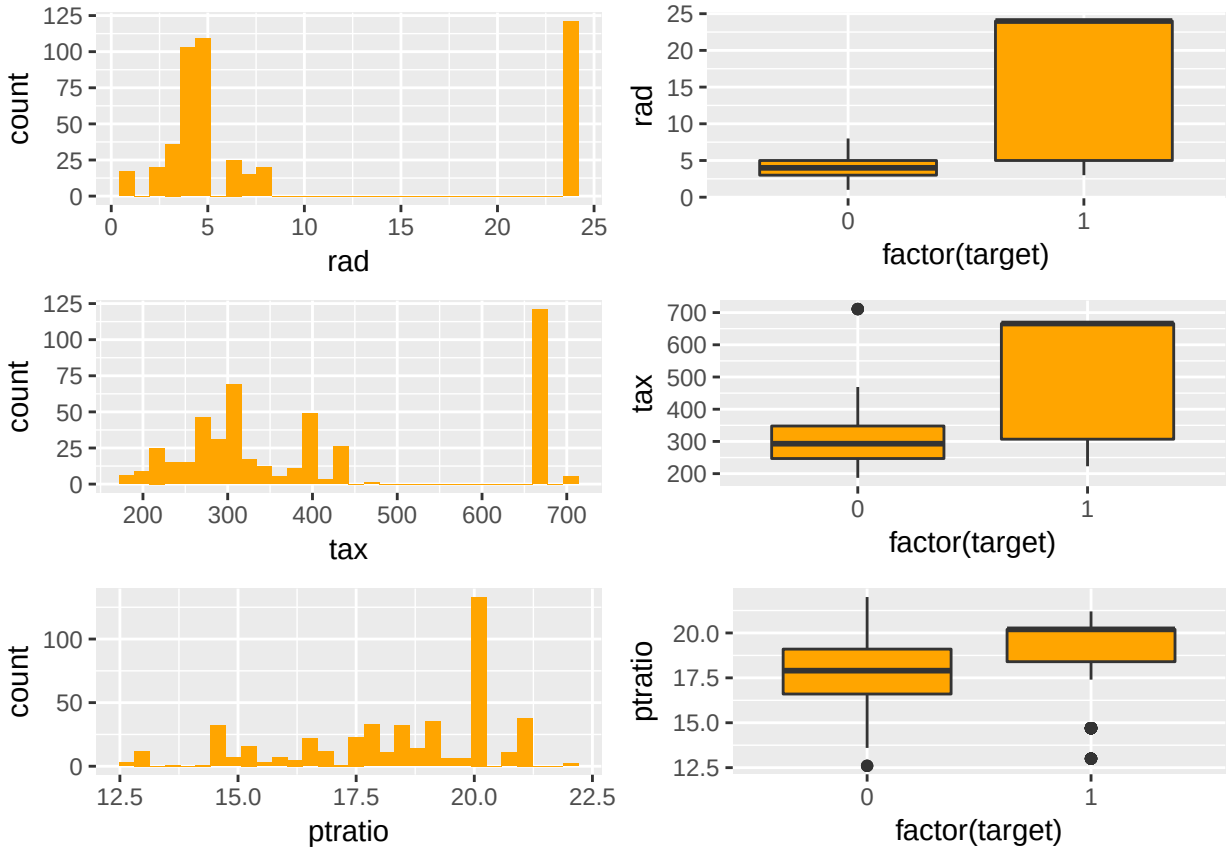

rm
9
8
7
6
5
4
0
1
factor(target)

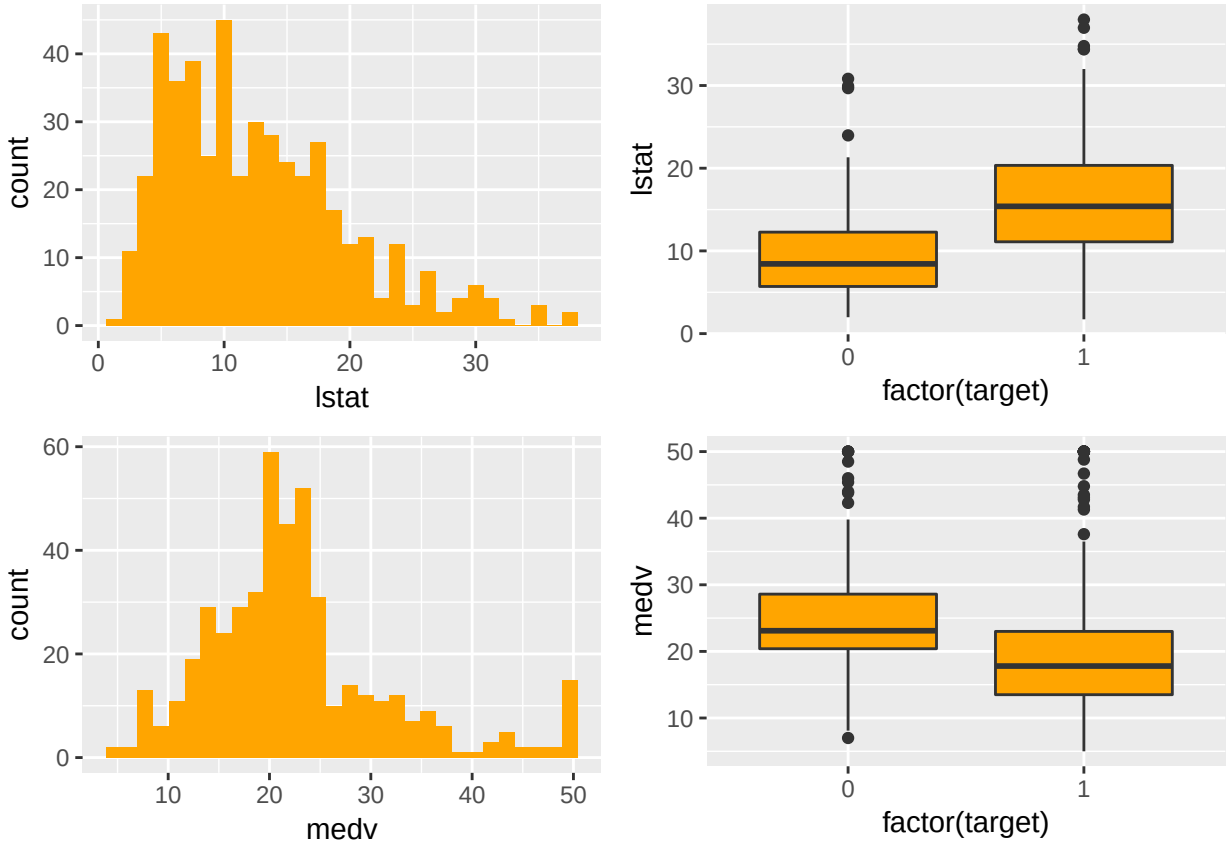

count
50
40
30
20
10
0
0
25
50
75
100
age


age
100
75
50
25
0
0
1
factor(target)


count
40
20
0
2.5
5.0
7.5
10.0
12.5
dis


dis
12.5
10.0
7.5
5.0
2.5
0
1
factor(target)

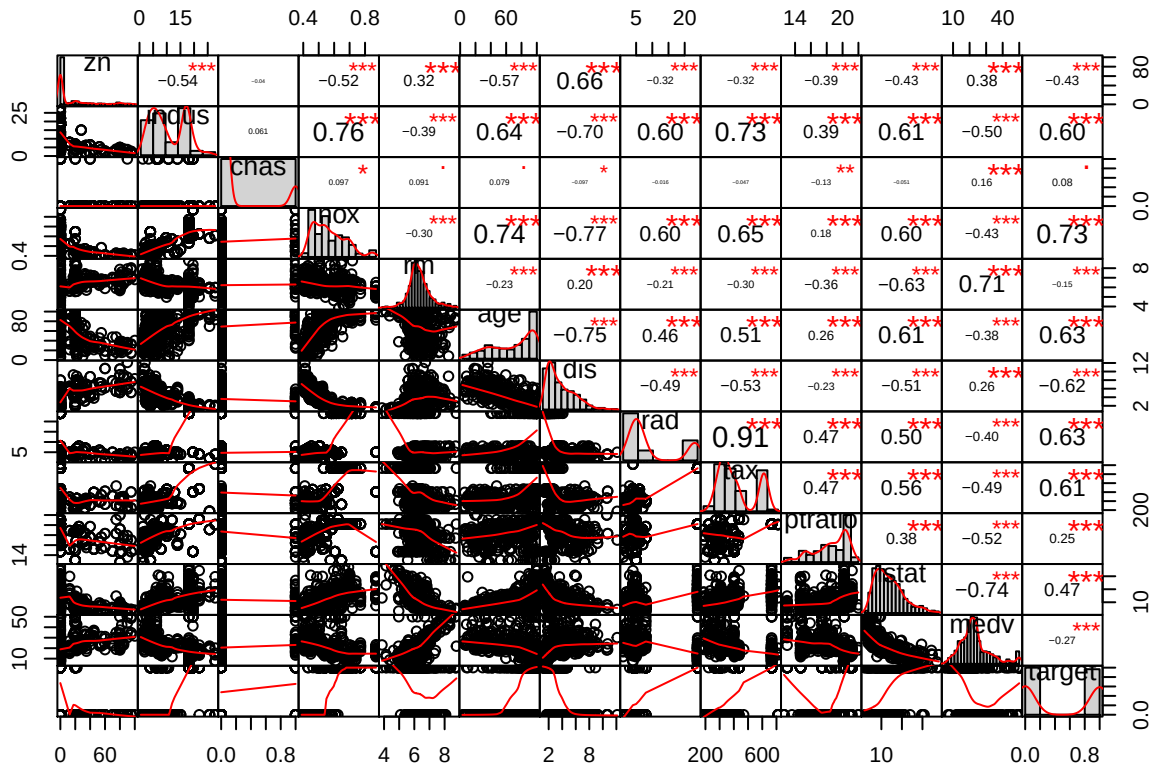

count
rad
rad
factor(target)
tax
tax
factor(target)
ptratio
ptratio
factor(target)



From the boxplots we can see variables *zn, rm, dis, lstat,* and *medv* contains outlier which can influence on our models.

```
##
##   0   1
## 237 229
```

We will now look at correlation matrix to get an understanding of relationships between variables:



We can see *tax* and *rad* has a very high correlation which raises the concern of multicollinearity concern in our dataset.

Data Preparation

To prepare our data, first, we will look for any missing value in our dataset and from the following table we can see there is no missing data in any of our variables.

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

As one of the condition of logistic regression is to have very little or no multicollinearity among the variables. As we have seen earlier *tax* and *rad* has a very high correlation, we will create a new variable putting *tax* in a bucket and dichotomize using median split.

Build Models

Model 1: Our first model uses the original *tax* and all the variables in the dataset.

	Estimate	Std. Error	z value	Pr(>|z|)
zn	-0.06595	0.03466	-1.903	0.05706
indus	-0.06461	0.04762	-1.357	0.1748
chas	0.9108	0.7555	1.205	0.228
nox	49.12	7.932	6.193	5.897e-10
rm	-0.5875	0.7228	-0.8127	0.4164
age	0.03419	0.01381	2.475	0.01333
dis	0.7387	0.2303	3.208	0.001338
rad	0.6664	0.1632	4.084	4.42e-05
tax	-0.006171	0.002955	-2.089	0.03674
ptratio	0.4026	0.1266	3.179	0.001477
lstat	0.04587	0.05405	0.8486	0.3961
medv	0.1808	0.06829	2.648	0.008103
(Intercept)	-40.82	6.633	-6.155	7.527e-10

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	645.9 on 465 degrees of freedom
Residual deviance:	192.0 on 453 degrees of freedom

Model 2: Our 2nd model is a *AIC based Backward Stepwise Model* and uses transformed *tax*.

```
## Start:  AIC=200.46
## target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + lstat + medv
##
##           Df Deviance    AIC
## - chas     1   174.62 198.62
## - lstat    1   175.05 199.05
## - indus    1   175.22 199.22
## - rm       1   175.90 199.90
## <none>         174.46 200.46
## - zn       1   177.78 201.78
## - dis      1   178.25 202.25
## - age      1   181.84 205.84
## - medv     1   185.08 209.08
## - ptratio  1   188.82 212.82
## - tax      1   196.59 220.59
## - rad      1   247.27 271.27
## - nox      1   253.11 277.11
##
## Step:  AIC=198.62
## target ~ zn + indus + nox + rm + age + dis + rad + tax + ptratio +
```

```
##     lstat + medv
##
##           Df Deviance    AIC
## - indus    1   175.25 197.25
## - lstat    1   175.29 197.29
## - rm       1   176.14 198.14
## <none>         174.62 198.62
## - dis      1   178.33 200.33
## - zn       1   178.44 200.44
## - age      1   182.33 204.33
## - medv     1   185.76 207.76
## - ptratio  1   188.83 210.83
## - tax      1   199.62 221.62
## - rad      1   252.06 274.06
## - nox      1   253.30 275.30
##
## Step:  AIC=197.25
## target ~ zn + nox + rm + age + dis + rad + tax + ptratio + lstat +
##     medv
##
##           Df Deviance    AIC
## - lstat    1   175.82 195.82
## - rm       1   176.75 196.75
## <none>         175.25 197.25
## - dis      1   178.85 198.85
## - zn       1   179.76 199.76
## - age      1   182.78 202.78
## - medv     1   186.30 206.30
## - ptratio  1   189.42 209.42
## - tax      1   205.42 225.42
## - rad      1   253.76 273.76
## - nox      1   262.53 282.53
##
## Step:  AIC=195.82
## target ~ zn + nox + rm + age + dis + rad + tax + ptratio + medv
##
##           Df Deviance    AIC
## <none>         175.82 195.82
## - rm       1   178.59 196.59
## - dis      1   179.53 197.53
## - zn       1   179.92 197.92
## - medv     1   186.74 204.74
## - age      1   187.08 205.08
## - ptratio  1   190.84 208.84
## - tax      1   205.70 223.70
## - rad      1   256.95 274.95
## - nox      1   264.89 282.89
```

	Estimate	Std. Error	z value	Pr(>|z|)
zn	-0.06787	0.03655	-1.857	0.06331
nox	51.05	7.933	6.435	1.231e-10
rm	-1.164	0.7083	-1.643	0.1003
age	0.0408	0.0128	3.188	0.001431
dis	0.456	0.2453	1.859	0.06303

	Estimate	Std. Error	z value	Pr(>|z|)
rad	1.002	0.2034	4.924	8.465e-07
tax	-3.253	0.6945	-4.684	2.819e-06
ptratio	0.4801	0.1333	3.603	0.000315
medv	0.2177	0.07125	3.055	0.002251
(Intercept)	-42.74	6.747	-6.335	2.378e-10

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	645.9 on 465 degrees of freedom
Residual deviance:	175.8 on 456 degrees of freedom

Model 3: Our 3rd model is a *AIC based Forward Stepwise Model* and uses transformed *tax*

```
## Start:  AIC=200.46
## target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + lstat + medv
```

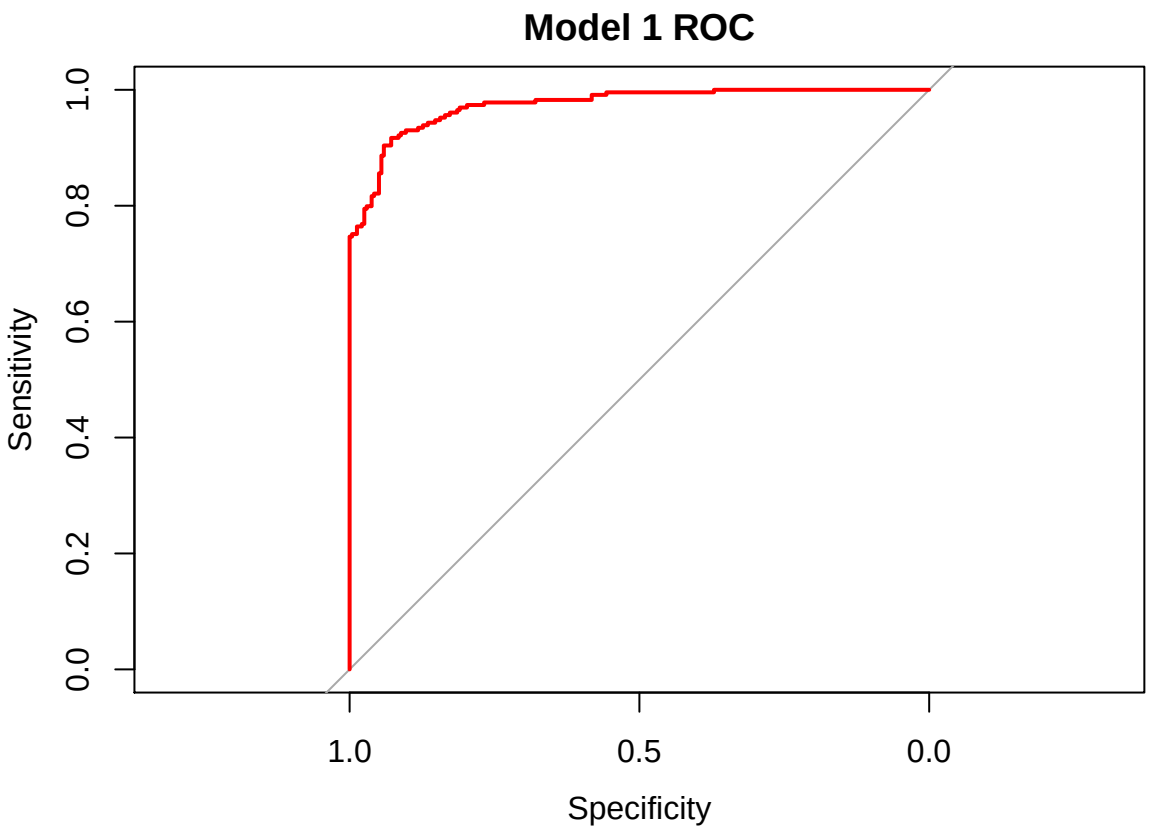
	Estimate	Std. Error	z value	Pr(>|z|)
zn	-0.06583	0.03931	-1.675	0.09395
indus	-0.04232	0.04915	-0.8611	0.3892
chas	0.3428	0.8474	0.4045	0.6859
nox	53.93	8.91	6.053	1.424e-09
rm	-0.9214	0.7729	-1.192	0.2332
age	0.03675	0.0142	2.588	0.009659
dis	0.4629	0.2463	1.88	0.06017
rad	0.9324	0.213	4.378	1.2e-05
tax	-3.058	0.7357	-4.156	3.232e-05
ptratio	0.4752	0.1344	3.537	0.0004054
lstat	0.04204	0.0547	0.7686	0.4421
medv	0.213	0.07086	3.006	0.00265
(Intercept)	-45.14	7.093	-6.364	1.964e-10

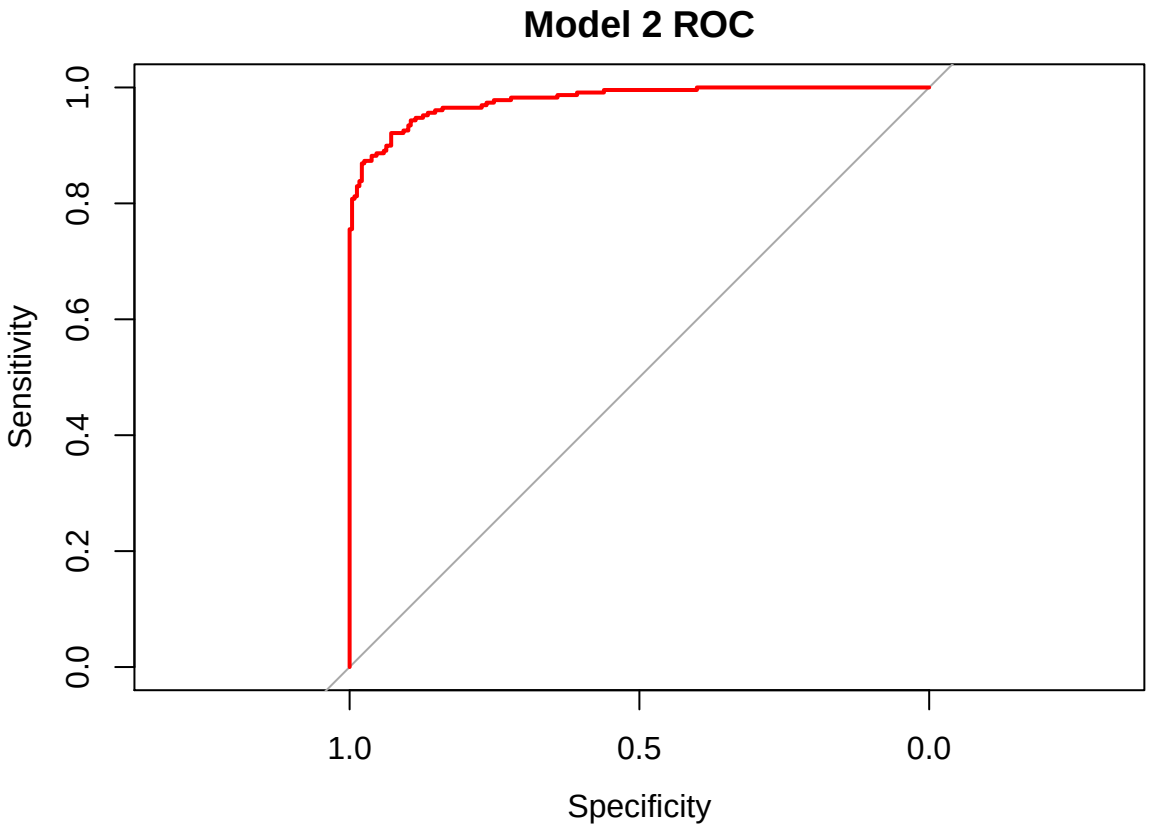
(Dispersion parameter for binomial family taken to be 1)

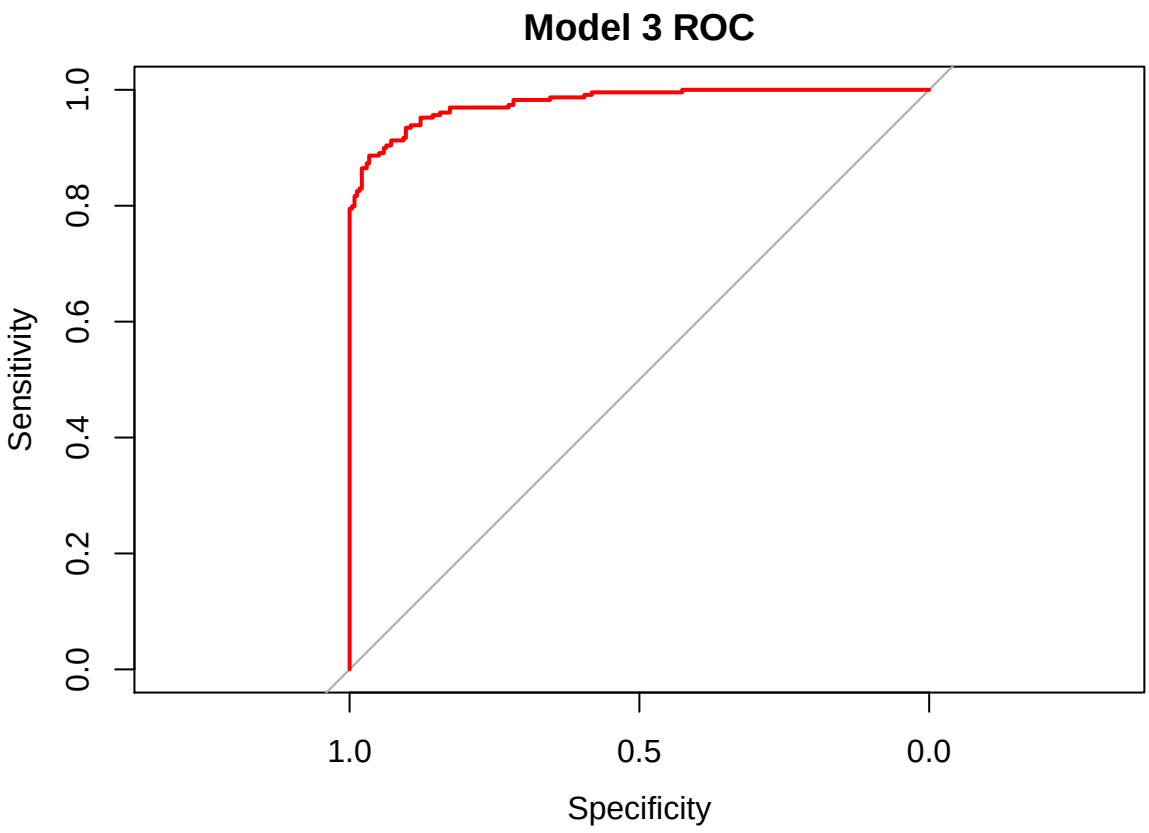
Null deviance:	645.9 on 465 degrees of freedom
Residual deviance:	174.5 on 453 degrees of freedom

Model Selection

ROC Curve


Model 1 ROC
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity


Model 2 ROC
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity



Area Under Curve

Model	Area.Under.Curve
Model 1	0.9737623
Model 2	0.9783686
Model 3	0.9783133

Confusion Matrix

```
## [1] "Model1"
```

	0	1
0	220	17
1	22	207

```
## [1] "Model2"
```

	0	1
0	220	17
1	19	210

```
## [1] "Model3"
```

	0	1
0	222	15
1	22	207

Based on our ROC curve, Area under curve, confusion matrix and AIC number we have selected *Model 2*.

Predictions

```
##
##  0  1
## 20 20
```

Appendix

Full R code can be found at the following URL:

https://github.com/choudhury1023/DATA-621/blob/master/HW%203/Ahsanul_Choudhury_HW3.Rmd

Reference:

http://rstudio-pubs-static.s3.amazonaws.com/2899_a9129debf6bd47d2a0501de9c0dc583d.html