Homework 3

Group 1

Contents

Prepared for:

Dr. Nathan Bastian

City University of New York, School of Professional Studies - Data 621

Prepared by:

Group 1

Senthil Dhanapal

Yadu Chittampalli

Christophe Hunt

1 Introduction

Crime has a high cost to all parts of society and it can have severe long term impact on neighborhoods. If crime rises in the neighborhood or it is invaded by criminals, then families and those with the economic means to leave for more stable areas will do so[1]. Additionally, crime can even have a health cost to the community in that the perception of a dangerous neighborhood was associated with significantly lower odds of having high physical activity among both men and women[2]. It is important to understand the propensity for crime levels of a neighborhood before investing in that neighborhood.

2 Statement of the Problem

The purpose of this report is to develop a statistical model to determine the variables that are independently associated with neighborhoods with crime rates above or below the median. Note that neighborhoods with crime rates above or below the median have already been provided in our evaluation data set.

3 Data Exploration

3.1 Variables Explained

The variables provided in our evaluation data set our explained below:

Abbreviation	Definition
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
black	1000(Bk - 0.63)2 where Bk is the proportion of blacks by town
lstat	lower status of the population (percent)
medv	median value of owner-occupied homes in $1000s

[1]Effect of Crime on Real Estate Values. (1952). The Journal of Criminal Law, Criminology, and Police Science, 43(3), 357-357. Retrieved from http://www.jstor.org.remote.baruch.cuny.edu/stable/1139159

[2]Bennett GG, McNeill LH, Wolin KY, Duncan DT, Puleo E, Emmons KM (2007) Safe To Walk? Neighborhood Safety and Physical Activity Among Public Housing Residents. PLoS Med 4(10): e306. doi:10.1371/journal.pmed.0040306

3.2 Exploration of Variables

The skewness of each input variable is shown below. The two variables with the strongest skew are the proportion of residential land zoned for large lots and the proportion of blacks by town. Respectively the magnitudes of the skewness of these two variables are 2.18 and 2.92. This indicates that the distributions for these two variables are far from symmetrical. The skewness of the dummy variable (whether the suburb borders the river or not) can be neglected because it is a binary variable. All of the other variables skewnesses that are approximately of magnitude 1 or less. This indicates that the distributions for those variables can be considered symmetric even though for three of the variables (concentration of nitrogen oxides, index of accessibility to radial highways, and median value of owner-occupied homes) are multimodal.

variables	skew
zn	2.1768152
indus	0.2885450
chas	3.3354899
nox	0.7463281
rm	0.4793202
age	0.5777075
dis	0.9988926
rad	1.0102788
tax	0.6593136
ptratio	0.7542681
black	2.9163108
lstat	0.9055864
medv	1.0766920
target	0.0342293

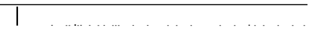
According to the standard deviations of each variable, the variable that has the highest difference from the mean is tax.

variables	sd
zn	23.3646511279634
indus	6.84585491881262
chas	0.256791996193711
nox	0.116666665669521
rm	0.704851288243787
age	28.3213784029166
dis	2.10694955535994
rad	8.68592724130043
tax	167.900088684704
ptratio	2.19684473073614
black	91.3211298387792
lstat	7.10189067779907
medv	9.23968141143397
target	0.500463581298941

Histograms of most of our variables have been plotted below so that distribution can be visualized. We have excluded `target` and `chas` due to being binary and not being well represented in the below visualization. We also excluded `rad` as it is an index variable and also is not best represented in the below visualization.

Table 1 : Descriptive Statistics
11 Variables 466 Observations

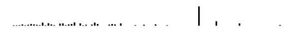
zn



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	26	0.61	12	0	0	0	0	16	45	80

`lowest :   0  12  18  18  20, highest:  82  85  90  95 100`

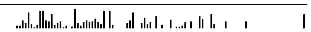
indus



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	73	0.98	11	2	3	5	10	18	20	21

`lowest :  0.5  0.7  1.2  1.2  1.2, highest: 18.1 19.6 21.9 25.6 27.7`

nox



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	79	1	0.6	0.4	0.4	0.4	0.5	0.6	0.7	0.8

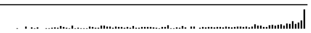
`lowest : 0.4 0.4 0.4 0.4 0.4, highest: 0.7 0.7 0.7 0.8 0.9`

rm

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	419	1	6	5	6	6	6	7	7	8

`lowest : 4 4 4 5 5, highest: 8 8 9 9 9`

age



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	333	1	68	18	26	44	77	94	99	100

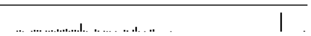
`lowest :   3   6   6   6   7, highest:  99  99  99  99 100`

dis

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	380	1	4	1	2	2	3	5	7	8

`lowest :  1  1  1  1  1, highest:  9  9 11 11 12`

tax



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	63	0.98	410	222	233	281	334	666	666	666

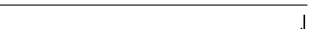
`lowest : 187 188 193 198 216, highest: 432 437 469 666 711`

ptratio

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	46	0.98	18	15	15	17	19	20	21	21

`lowest : 13 13 14 14 15, highest: 21 21 21 21 22`

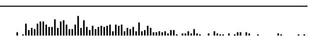
black



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	331	0.99	357	88	295	376	391	396	397	397

```
lowest :   0.3   2.5   2.6   3.5   3.6
highest: 396.3 396.3 396.3 396.4 396.9
```

lstat



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	424	1	13	4	5	7	11	17	23	27

`lowest :  2  2  2  2  3, highest: 34 34 35 37 38`

medv

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	218	1	23	10	13	17	21	25	35	43

`lowest :  5  6  6  7  7, highest: 46 47 48 49 50`

3.3 Correlation Matrix

We implement a correlation matrix to better understand the correlation between variables in the data set. The below matrix is the results and we noticed a few interesting correlations.

- High nitrogen oxides concentration (parts per 10 million) ("nox") is positively correlated with higher than median crime rates. As defined by the EPA - "NOx pollution is emitted by automobiles, trucks and various non-road vehicles (e.g., construction equipment, boats, etc.) as well as industrial sources such as power plants, industrial boilers, cement kilns, and turbines"[3]. It is clear to see that nox is concentrated in areas of high road traffic and possible high industrial use which would be neighborhoods of low value and may attract crime.
- The weighted mean of distances to five Boston employment centers is negatively correlated with a city with higher than median crime rate. This is intuitive in that employment centers would be more closely located in cities of high crime due to high unemployment being positively correlated with higher crimes rates[4].
- The tax is positively correlated with higher than median crime rate which is counter intuitive because we would think as tax increases then crime would decrease (more valuable property = higher tax = less crime).
- We also see bk is negatively correlated with higher than median crime rates but it seems to be due to the transformation of 1000(Bk - 0.63)ˆ2. Further resources on why this type of transformation is being used were not available. It should be noted that this transformation causes a counter intuitive correlation.

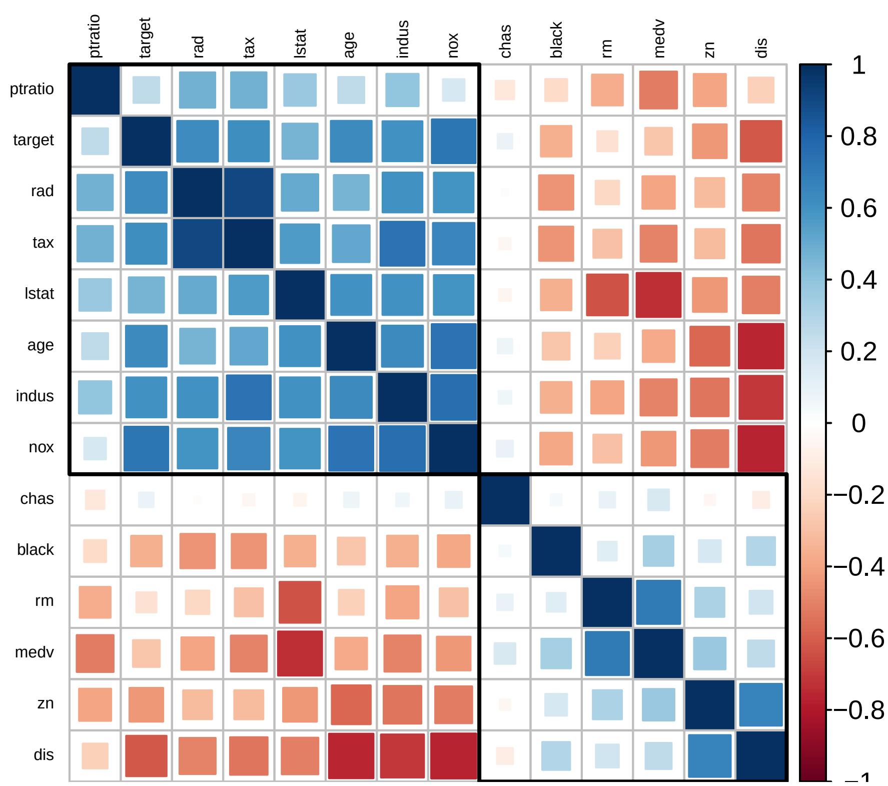


Figure 1: Correlation Plot of Training Data Set

[3] "Nitrogen Oxides Control Regulations | Ground-level Ozone | New England | US EPA." EPA. Environmental Protection Agency, n.d. Web. 22 Oct. 2016.
[4] Ajimotokin, S., Haskins, A., & Wade, Z. (2015). The Effects of Unemployment on Crime Rates in the US.

4 Data Transformation

4.1 Outliers Treatment

We chose winsorizing as the method to address outliers. Instead of trimming values, winsorizing uses the interquantile range to replace values that are above or below the interquantile range multiplied by a factor. Those values above or below the range multiplied by the factor are then replaced with max and min value of the interquantile range. Using the factor 2.2 for winsorizing outliers is a method developed my Hoaglin and Iglewicz and published Journal of American Statistical Association in 1987[5].

The below table is the summary results of the winsorizing of the data.

Table 4:

Statistic	N	Mean	St. Dev.	Min	Max
zn	466	8.739	15.567	0.000	45.000
indus	466	11.105	6.846	0.460	27.740
chas	466	0.071	0.257	0	1
nox	466	0.554	0.117	0.389	0.871
rm	466	6.289	0.686	4.368	8.259
age	466	68.368	28.321	2.900	100.000
dis	466	3.793	2.096	1.130	10.710
rad	466	9.530	8.686	1	24
tax	466	409.502	167.900	187	711
ptratio	466	18.398	2.197	12.600	22.000
black	466	380.268	22.690	331.290	396.900
lstat	466	12.631	7.102	1.730	37.970
medv	466	22.273	8.399	5.000	42.300
target	466	0.491	0.500	0	1

[5] Hoaglin, D. C., and Iglewicz, B. (1987), Fine tuning some resistant rules for outlier labeling, Journal of American Statistical Association, 82, 1147-1149.

4.2 BoxCox Transformations

Using the `BoxCox.lambda` function from the `forecast` package we are able to determine our necessary transformations to our independent variables.

λ	Variables
0.1396180	zn
-0.0877933	indus
0.4722021	chas
-0.9999242	nox
0.0389955	rm
1.9999242	age
-0.6099464	dis
-0.3353947	rad
-0.9999242	tax
1.9999242	ptratio
1.9999242	black
-0.1792021	lstat
0.1044075	medv

Utilizing the below table of common transformations based on the lambda value of the BoxCox we further transform our independent variables.

Common Box-Cox Transformations[6]

λ	Y’
-2	$Y^{-2} = \frac{1}{Y^2}$
-1	$Y^{-1} = \frac{1}{Y^1}$
-0.5	$Y^{-0.5} = \frac{1}{\sqrt{(Y)}})$
0	$\log(Y)$
0.5	$Y^{0.5} = \sqrt{(Y)}$
1	$Y^1 = Y$
2	Y^2

Lambda values that did not fall in the proximity of common transformations were ignored. All other Lambda values were truncated to the nearest tenth that match a common transformation as per the below table.

variable	variable transformation
indus	$\log(indus)$
chas	$\sqrt{chas}$
nox	nox^{-1}
rm	$\log(rm)$
age	age^2
dis	$dis^{-.5}$
tax	tax^{-1}
ptratio	$ptratio^2$
black	$black^2$

[6] By Understanding Both the Concept of Transformation and the Box-Cox Method, Practitioners Will Be Better Prepared to Work with Non-normal Data. . “Making Data Normal Using Box-Cox Power Transformation.” ISixSigma. N.p., n.d. Web. 29 Oct. 2016.

5 Models Built

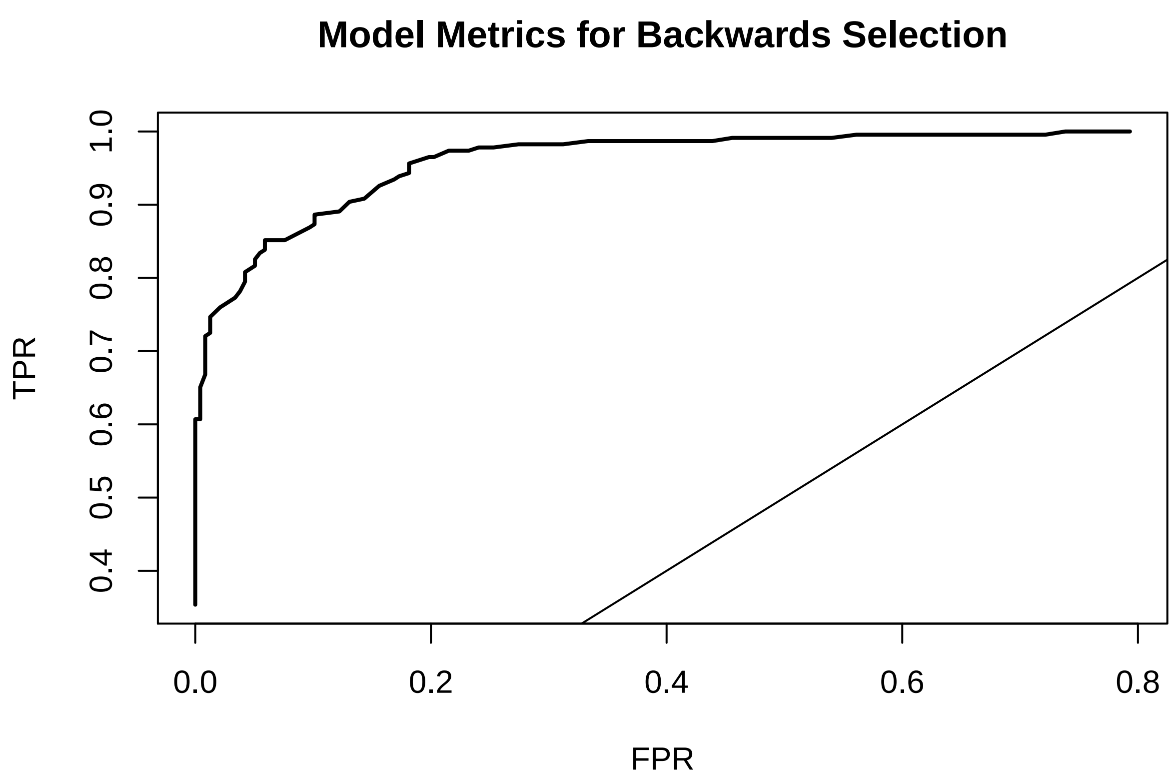
5.1 Model 1 - Backwards Selection Method

Table 6:

	Dependent variable:
	fullModel
I(nox^-1)	-0.830^{***}
	(0.090)
I(age^2)	0.00003^{***}
	(0.00001)
I(dis^-0.5)	-0.757^{***}
	(0.210)
rad	0.011^{***}
	(0.003)
I(tax^-1)	-66.824^{***}
	(23.469)
I(ptratio^2)	0.0004
	(0.0002)
I(black^2)	-0.00000^{***}
	(0.00000)
lstat	0.006^{*}
	(0.004)
medv	0.016^{***}
	(0.003)
Constant	2.288^{***}
	(0.292)
Observations	466
Log Likelihood	-92.296
Akaike Inf. Crit.	204.593
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

5.1.1 Model Metrics for Backwards Selection

We first use an established threshold of .50 to determine our best possible threshold. In this instances our best threshold is 0.500.



	Act-Pos	Act-Neg
Pred-Pos	195	34
Pred-Neg	14	223

	Model Metrics for Backwards Selection
accuracy	0.897
classif.error	0.103
precision	0.933
sensitivity	0.852
specificity	0.941
f1score	0.890
auc	0.759
best.threshold	0.500
aic	204.593

5.1.2 Multicollinearity for Backwards Selection

We will use a value of 5 as our threshold for multicollinearity of our variables[7]. Here in our backwards selection model we find that ourt transformed `nox` and `dis` exceeds our pre-established threshold.

variables	VIF
I(nox^-1)	5.755252
I(age^2)	3.580516
I(dis^-0.5)	5.308905
rad	3.233298
I(tax^-1)	3.145345
I(ptratio^2)	1.676025
I(black^2)	1.344496
lstat	3.680960
medv	3.532166

[7] "Variance Inflation Factor (VIF)." How2stats:. N.p., n.d. Web. 27 Oct. 2016.

5.2 Model 2 - Forwards Selection Method

The simplest data-driven model building approach is called forward selection. In this approach, one adds variables to the model one at a time. At each step, each variable that is not already in the model is tested for inclusion in the model.

Step function used in this assignment chooses a model by AIC in a Stepwise Algorithm. It continues including variables until the AIC value of variable is the least in the list of variables to choose.

```
## Start:  AIC=-644.15
## target ~ 1
##
##                   Df Sum of Sq     RSS      AIC
## + I(nox^-1)        1    66.116  50.349 -1032.94
## + I(age^2)         1    49.753  66.713  -901.81
## + I(dis^-0.5)      1    49.673  66.793  -901.25
## + rad              1    45.948  70.518  -875.96
## + I(log(indus))    1    42.398  74.068  -853.07
## + I(tax^-1)        1    41.918  74.547  -850.06
## + lstat            1    25.632  90.834  -757.98
## + zn               1    25.565  90.900  -757.64
## + I(black^2)       1    21.849  94.617  -738.97
## + medv             1    11.335 105.130  -689.87
## + I(ptratio^2)     1     8.986 107.479  -679.57
## + I(log(rm))       1     3.744 112.721  -657.38
## + I(sqrt(chas))    1     0.746 115.720  -645.15
## <none>                         116.466  -644.15
##
## Step:  AIC=-1032.94
## target ~ I(nox^-1)
##
##                   Df Sum of Sq    RSS     AIC
## + rad              1    5.0776 45.272 -1080.5
## + I(tax^-1)        1    3.3933 46.956 -1063.5
## + I(black^2)       1    1.6959 48.653 -1046.9
## + I(age^2)         1    0.6987 49.650 -1037.5
## + I(ptratio^2)     1    0.5710 49.778 -1036.3
## + I(log(rm))       1    0.4916 49.858 -1035.5
## + medv             1    0.3279 50.021 -1034.0
## + I(log(indus))    1    0.3061 50.043 -1033.8
## <none>                          50.349 -1032.9
## + zn               1    0.0500 50.299 -1031.4
## + I(sqrt(chas))    1    0.0440 50.305 -1031.3
## + lstat            1    0.0136 50.336 -1031.1
## + I(dis^-0.5)      1    0.0043 50.345 -1031.0
##
## Step:  AIC=-1080.48
## target ~ I(nox^-1) + rad
##
##                   Df Sum of Sq    RSS     AIC
## + medv             1   1.11058 44.161 -1090.1
## + I(age^2)         1   0.82921 44.442 -1087.1
## + I(black^2)       1   0.67914 44.592 -1085.5
## + I(log(rm))       1   0.59452 44.677 -1084.6
## + I(tax^-1)        1   0.25453 45.017 -1081.1
```

```
## <none>                      45.272 -1080.5
## + I(sqrt(chas))  1   0.15902 45.113 -1080.1
## + lstat          1   0.12216 45.149 -1079.7
## + I(dis^-0.5)    1   0.09113 45.180 -1079.4
## + I(ptratio^2)   1   0.03170 45.240 -1078.8
## + I(log(indus))  1   0.00470 45.267 -1078.5
## + zn             1   0.00091 45.271 -1078.5
##
## Step:  AIC=-1090.06
## target ~ I(nox^-1) + rad + medv
##
##                  Df Sum of Sq    RSS     AIC
## + I(age^2)       1   1.07406 43.087 -1099.5
## + I(black^2)     1   0.85978 43.301 -1097.2
## + I(tax^-1)      1   0.69865 43.462 -1095.5
## + lstat          1   0.22709 43.934 -1090.5
## + I(log(indus))  1   0.20749 43.954 -1090.2
## + I(dis^-0.5)    1   0.20159 43.959 -1090.2
## <none>                      44.161 -1090.1
## + I(ptratio^2)   1   0.09597 44.065 -1089.1
## + I(sqrt(chas))  1   0.03753 44.124 -1088.5
## + zn             1   0.02754 44.134 -1088.3
## + I(log(rm))     1   0.01119 44.150 -1088.2
##
## Step:  AIC=-1099.53
## target ~ I(nox^-1) + rad + medv + I(age^2)
##
##                  Df Sum of Sq    RSS     AIC
## + I(black^2)     1   0.88720 42.200 -1107.2
## + I(dis^-0.5)    1   0.79269 42.294 -1106.2
## + I(tax^-1)      1   0.64624 42.441 -1104.6
## <none>                      43.087 -1099.5
## + I(log(indus))  1   0.14658 42.940 -1099.1
## + I(ptratio^2)   1   0.04980 43.037 -1098.1
## + I(sqrt(chas))  1   0.02817 43.059 -1097.8
## + lstat          1   0.02274 43.064 -1097.8
## + zn             1   0.00171 43.085 -1097.5
## + I(log(rm))     1   0.00138 43.086 -1097.5
##
## Step:  AIC=-1107.22
## target ~ I(nox^-1) + rad + medv + I(age^2) + I(black^2)
##
##                  Df Sum of Sq    RSS     AIC
## + I(dis^-0.5)    1   0.81396 41.386 -1114.3
## + I(tax^-1)      1   0.48183 41.718 -1110.6
## <none>                      42.200 -1107.2
## + I(ptratio^2)   1   0.12372 42.076 -1106.6
## + I(log(indus))  1   0.12235 42.077 -1106.6
## + I(sqrt(chas))  1   0.02468 42.175 -1105.5
## + lstat          1   0.01587 42.184 -1105.4
## + zn             1   0.00785 42.192 -1105.3
## + I(log(rm))     1   0.00196 42.198 -1105.2
##
## Step:  AIC=-1114.3
```

```
## target ~ I(nox^-1) + rad + medv + I(age^2) + I(black^2) + I(dis^-0.5)
##
##                  Df Sum of Sq    RSS     AIC
## + I(tax^-1)      1   0.60854 40.777 -1119.2
## + I(log(indus))  1   0.28487 41.101 -1115.5
## <none>                       41.386 -1114.3
## + I(ptratio^2)   1   0.13167 41.254 -1113.8
## + lstat          1   0.12555 41.260 -1113.7
## + I(log(rm))     1   0.03131 41.355 -1112.7
## + I(sqrt(chas))  1   0.01758 41.368 -1112.5
## + zn             1   0.01480 41.371 -1112.5
##
## Step:  AIC=-1119.2
## target ~ I(nox^-1) + rad + medv + I(age^2) + I(black^2) + I(dis^-0.5) +
##     I(tax^-1)
##
##                  Df Sum of Sq    RSS     AIC
## + lstat          1  0.198103 40.579 -1119.5
## <none>                       40.777 -1119.2
## + I(log(indus))  1  0.162643 40.615 -1119.1
## + I(ptratio^2)   1  0.152425 40.625 -1119.0
## + zn             1  0.060083 40.717 -1117.9
## + I(log(rm))     1  0.038658 40.739 -1117.7
## + I(sqrt(chas))  1  0.027090 40.750 -1117.5
##
## Step:  AIC=-1119.47
## target ~ I(nox^-1) + rad + medv + I(age^2) + I(black^2) + I(dis^-0.5) +
##     I(tax^-1) + lstat
##
##                  Df Sum of Sq    RSS     AIC
## + I(ptratio^2)   1  0.207100 40.372 -1119.9
## <none>                       40.579 -1119.5
## + I(log(indus))  1  0.170875 40.408 -1119.4
## + zn             1  0.071496 40.508 -1118.3
## + I(sqrt(chas))  1  0.024015 40.555 -1117.8
## + I(log(rm))     1  0.003403 40.576 -1117.5
##
## Step:  AIC=-1119.86
## target ~ I(nox^-1) + rad + medv + I(age^2) + I(black^2) + I(dis^-0.5) +
##     I(tax^-1) + lstat + I(ptratio^2)
##
##                  Df Sum of Sq    RSS     AIC
## <none>                       40.372 -1119.9
## + I(log(indus))  1  0.103613 40.268 -1119.1
## + I(sqrt(chas))  1  0.033824 40.338 -1118.2
## + zn             1  0.012202 40.360 -1118.0
## + I(log(rm))     1  0.000701 40.371 -1117.9

##
## Call:
## lm(formula = target ~ I(nox^-1) + rad + medv + I(age^2) + I(black^2) +
##     I(dis^-0.5) + I(tax^-1) + lstat + I(ptratio^2), data = wCrimes)
##
## Coefficients:
```

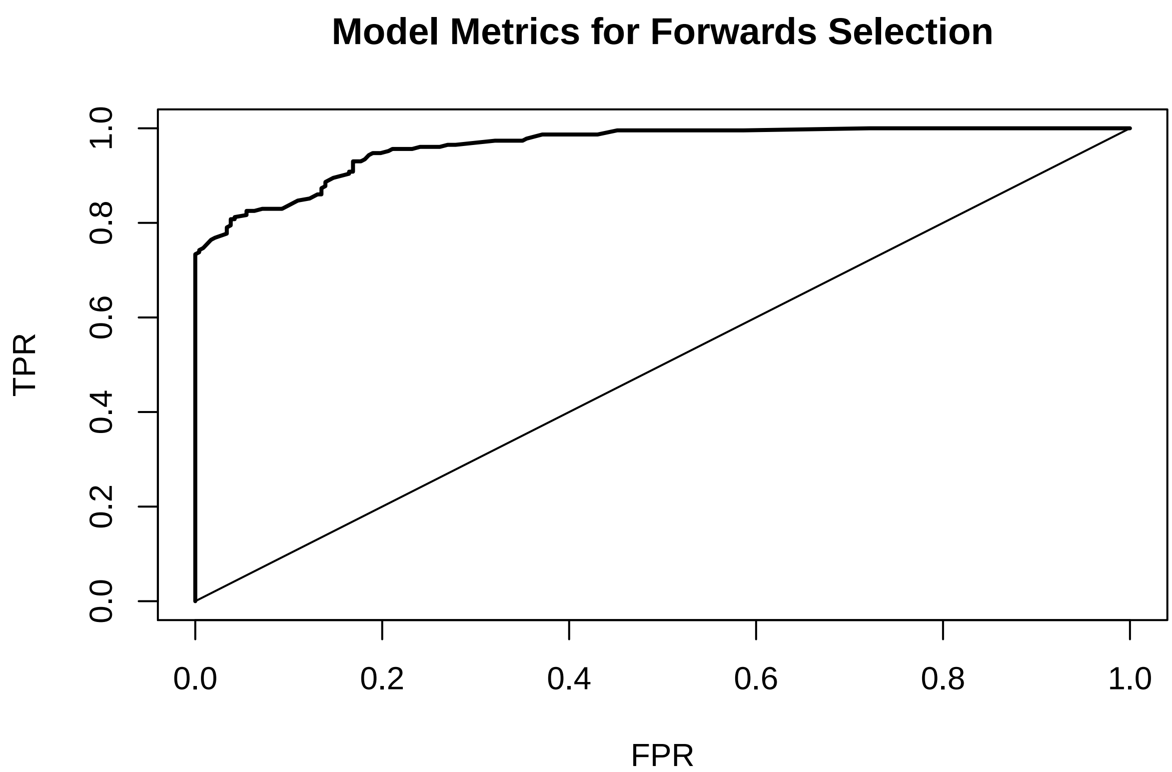
```
##   (Intercept)     I(nox^-1)           rad          medv      I(age^2)
##     2.288e+00    -8.299e-01     1.063e-02     1.553e-02     2.828e-05
##    I(black^2)   I(dis^-0.5)     I(tax^-1)         lstat  I(ptratio^2)
##    -2.871e-06    -7.566e-01    -6.682e+01     6.299e-03     3.532e-04
```

Table 10:

	Dependent variable:
	target
I(nox^-1)	-7.231^{***}
	(1.136)
rad	0.520^{***}
	(0.119)
I(age^2)	0.0002^{**}
	(0.0001)
medv	0.098^{***}
	(0.031)
I(ptratio^2)	0.008^{***}
	(0.003)
Constant	4.871^{**}
	(2.425)
Observations	466
Log Likelihood	-110.847
Akaike Inf. Crit.	233.694
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

5.2.1 Model Metrics for Forwards Selection

We first use an established threshold of .50 to determine our best possible threshold.



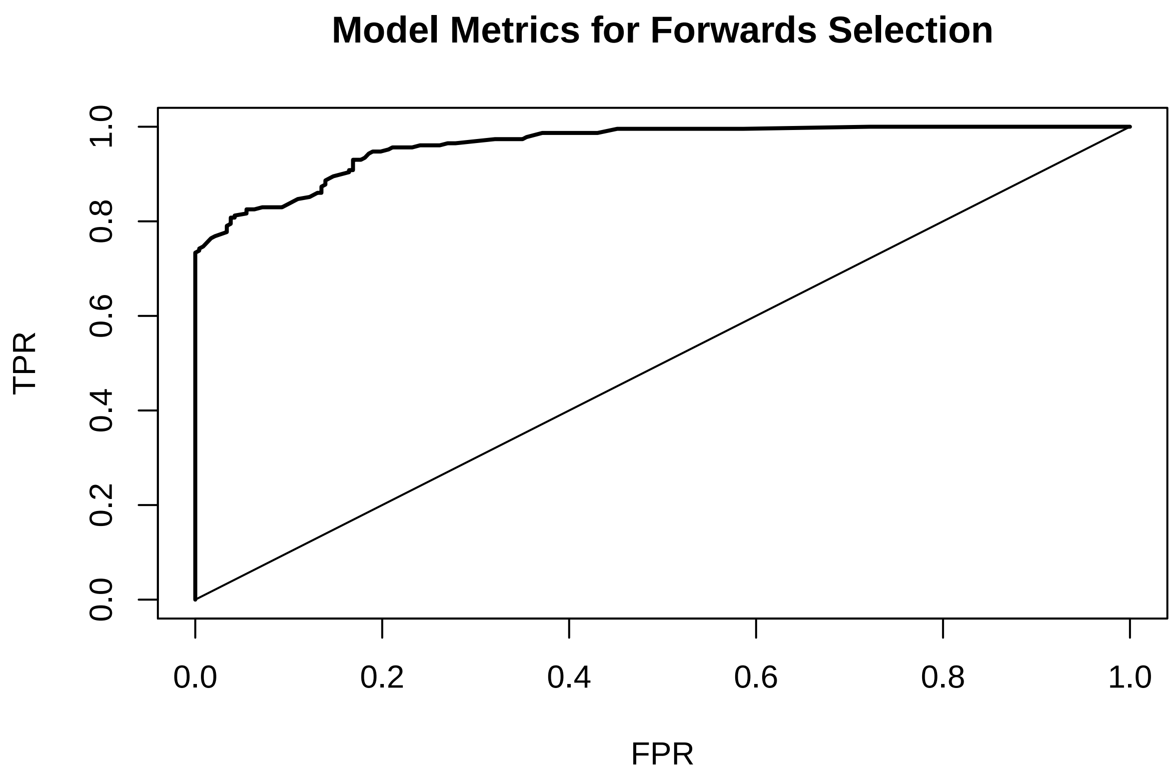
	Act-Pos	Act-Neg
Pred-Pos	195	34
Pred-Neg	29	208

	Model Metrics for Forwards Selection
accuracy	0.865
classif.error	0.135
precision	0.871
sensitivity	0.852
specificity	0.878
f1score	0.861
auc	0.962
best.threshold	0.610
aic	233.694

Our results indicate that .610 would be the best threshold for this model so we re-run our metrics using this threshold.

5.2.1.1 Model Metrics for Forwards Selection with best threshold

Model Metrics using best threshold of .610.



	Act-Pos	Act-Neg
Pred-Pos	189	40
Pred-Neg	13	224

	Model Metrics for Forwards Selection
accuracy	0.886
classif.error	0.114
precision	0.936
sensitivity	0.825
specificity	0.945
f1score	0.877
auc	0.962
best.threshold	0.610
aic	233.694

5.2.2 Multicollinearity for Forwards Selection

Here in our forward selection model we find that no variable exceeds our pre-established threshold of 5 for multicollinearity.

variables	VIF
I(nox^-1)	1.629476
rad	1.193812
I(age^2)	1.483391
medv	1.798920
I(ptratio^2)	1.645711

5.3 Model 3 - Subset Selection Method

Using the `leaps` package and the `regsubsets` function we are able to subset our independent variables by looking at the best model for each predictor. The variables as indicated in column `8` of the below table will be further implement into our subset selection model.

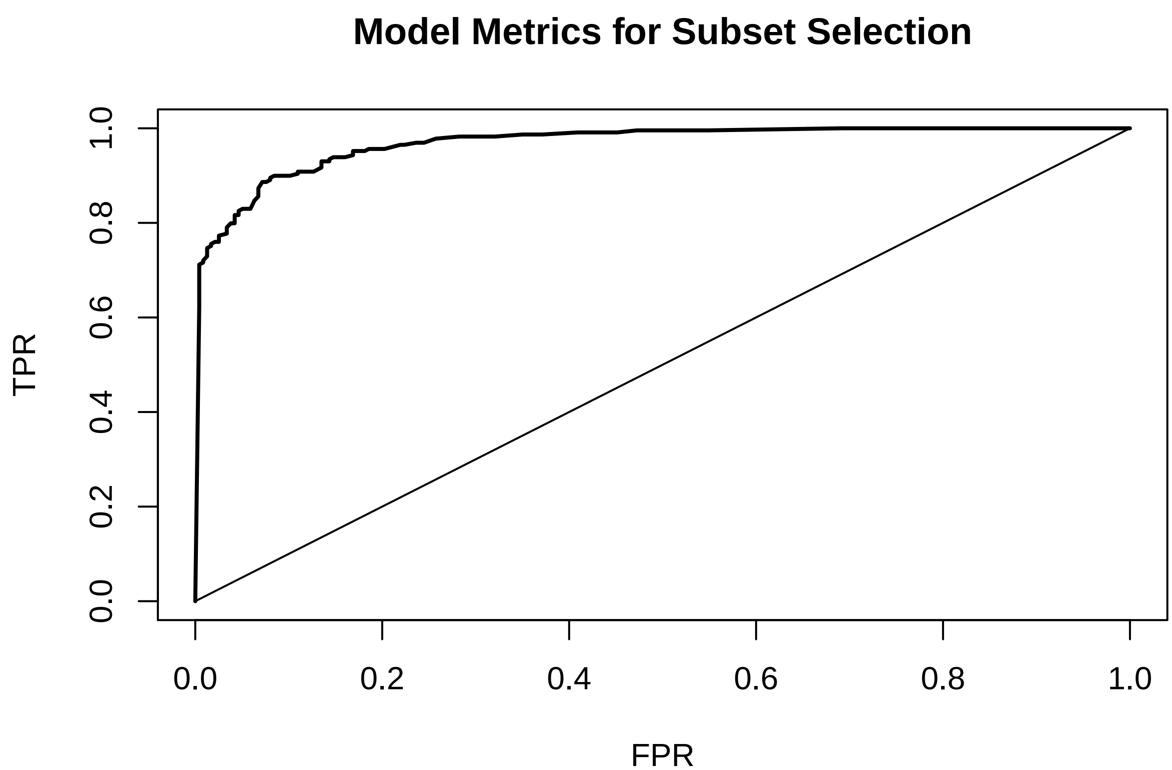
	1 (1)	2 (1)	3 (1)	4 (1)	5 (1)	6 (1)	7 (1)	8 (1)
zn								
I(log(indus))								
I(sqrt(chas))								
I(nox^-1)	*	*	*	*	*	*	*	*
I(log(rm))								
I(age^2)				*	*	*	*	*
I(dis^-0.5)						*	*	*
rad		*	*	*	*	*	*	*
I(tax^-1)							*	*
I(ptratio^2)								
I(black^2)					*	*	*	*
lstat								*
medv			*	*	*	*	*	*

Table 17:

	Dependent variable:
	target
I(nox^-1)	-11.595^{***}
	(1.813)
I(age^2)	0.0002^{***}
	(0.0001)
I(dis^-0.5)	-12.109^{***}
	(2.964)
rad	0.414^{***}
	(0.133)
I(tax^-1)	-4.690
	(276.467)
I(black^2)	-0.00004^{***}
	(0.00002)
lstat	0.033
	(0.048)
medv	0.126^{***}
	(0.044)
Constant	28.324^{***}
	(5.237)
Observations	466
Log Likelihood	-100.067
Akaike Inf. Crit.	218.135
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

5.3.1 Model Metrics for Subset Selection

We first use an established threshold of .50 to determine our best possible threshold.



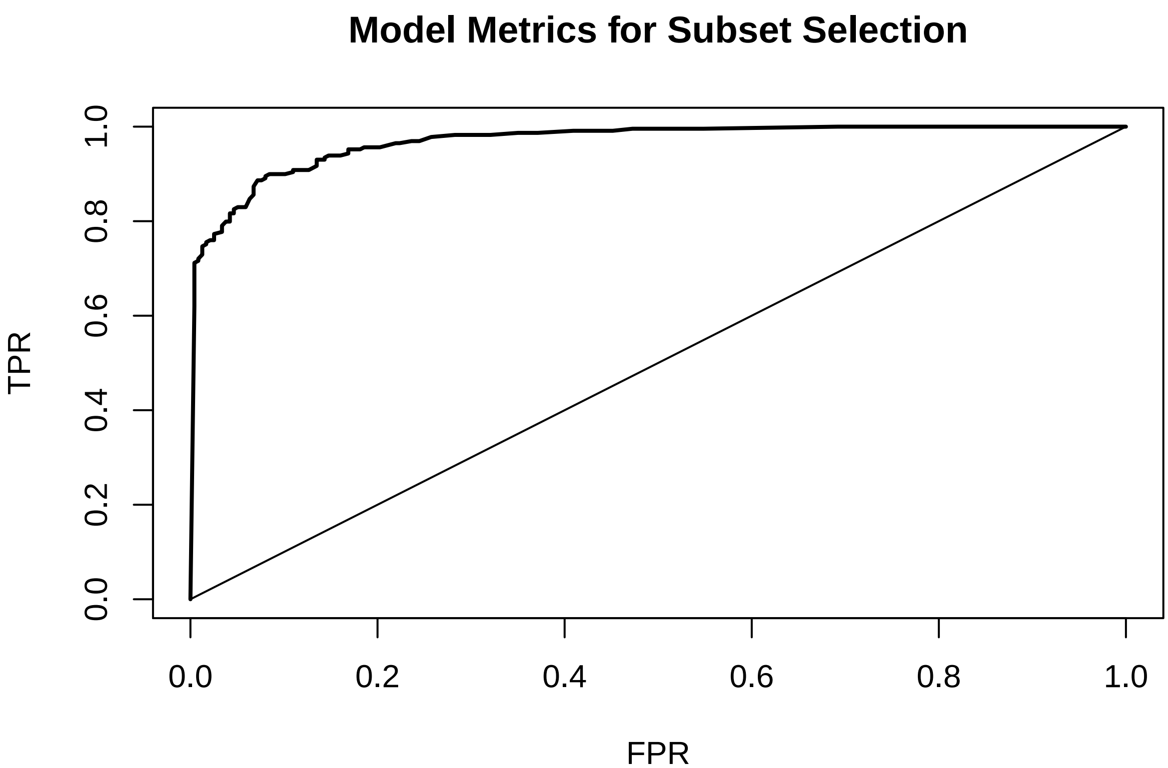
	Act-Pos	Act-Neg
Pred-Pos	205	24
Pred-Neg	19	218

	Model Metrics for Subset Selection
accuracy	0.9077
classif.error	0.0923
precision	0.9152
sensitivity	0.8952
specificity	0.9198
f1score	0.9051
auc	0.9681
best.threshold	0.4900
aic	218.1349

Our results indicate that .490 would be the best threshold for this model so we re-run our metrics using this threshold.

5.3.1.1 Model Metrics for Subset Selection with best threshold

Model Metrics using best threshold of .490.



	Act-Pos	Act-Neg
Pred-Pos	206	23
Pred-Neg	20	217

	Model Metrics for Subset Selection
accuracy	0.9077
classif.error	0.0923
precision	0.9115
sensitivity	0.8996
specificity	0.9156
f1score	0.9055
auc	0.9681
best.threshold	0.4900
aic	218.1349

5.3.2 Multicollinearity for Subset Selection

Here in our subset selection model we find that no variable exceeds our pre-established threshold of 5 for multicollinearity.

variables	VIF
I(nox^-1)	3.465847
I(age^2)	1.820623
I(dis^-0.5)	3.387638
rad	1.434276
I(tax^-1)	1.684888
I(black^2)	1.048190
lstat	2.408696
medv	2.906005

6 Selected Model

7 Appendix A

7.1 Session Info

- R version 3.3.1 (2016-06-21), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=English_United States.1252`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, utils
- Other packages: abc 2.1, abc.data 1.0, bibtex 0.4.0, car 2.1-3, corrplot 0.77, data.table 1.9.6, doParallel 1.0.10, dplyr 0.5.0, e1071 1.6-7, foreach 1.4.3, forecast 7.3, Formula 1.2-1, ggplot2 2.1.0, glmulti 1.0.7, highlight 0.4.7, Hmisc 3.17-4, iterators 1.0.8, itertools 0.1-3, knitcitations 1.0.7, knitr 1.14, lattice 0.20-34, leaps 2.9, locfit 1.5-9.1, magrittr 1.5, MASS 7.3-45, matrixStats 0.51.0, missForest 1.4, nnet 7.3-12, pacman 0.4.1, purrr 0.2.2, quantreg 5.29, randomForest 4.6-12, readr 1.0.0, rJava 0.9-8, scales 0.4.0, SparseM 1.72, stargazer 5.2, stringr 1.1.0, survival 2.39-5, tibble 1.2, tidyr 0.6.0, tidyverse 1.0.0, timeDate 3012.100, xlsx 0.5.7, xlsxjars 0.6.1, xtable 1.8-2, zoo 1.7-13
- Loaded via a namespace (and not attached): acepack 1.4.1, assertthat 0.1, bitops 1.0-6, chron 2.3-47, class 7.3-14, cluster 2.0.5, codetools 0.2-15, colorspace 1.2-7, DBI 0.5-1, digest 0.6.10, evaluate 0.10, foreign 0.8-67, formatR 1.4, fracdiff 1.4-2, grid 3.3.1, gridExtra 2.2.1, gtable 0.2.0, highr 0.6, htmltools 0.3.5, httr 1.2.1, latticeExtra 0.6-28, lazyeval 0.2.0, lme4 1.1-12, lubridate 1.6.0, Matrix 1.2-7.1, MatrixModels 0.4-1, mgcv 1.8-15, minqa 1.2.4, munsell 0.4.3, nlme 3.1-128, nloptr 1.0.4, pbkrtest 0.4-6, plyr 1.8.4, quadprog 1.5-5, R6 2.2.0, RColorBrewer 1.1-2, Rcpp 0.12.7, RCurl 1.95-4.8, RefManageR 0.11.0, RJSONIO 1.3-0, rmarkdown 1.1, rpart 4.1-10, splines 3.3.1, stringi 1.1.2, tools 3.3.1, tseries 0.10-35, XML 3.98-1.4, yaml 2.1.13

7.2 Data Dictionary

Abbreviation	Definition
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
black	1000(Bk - 0.63)2 where Bk is the proportion of blacks by town
lstat	lower status of the population (percent)
medv	median value of owner-occupied homes in $1000s

7.3 R source code

Please see Homework 3.rmd on GitHub for source code.

https://github.com/ChristopheHunt/DATA-621-Group-1/blob/master/Homework%203/Homework%203.Rmd