HOMEWORK 3

Gurpreet Singh

Data 621

TABLE OF CONTENTS

DATA EXPLORATION:

The dataset contains information on crimes for various neighborhoods of a major city. Each record has a response variable indicating whether or not the crime rate is above the median crime rate (1) or not (0). The dataset will be utilized to build up a binary logistic regression model to detect the risk of high crime levels in the neighborhood.

zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv	target
0	19.58	0	0.605	7.929	96.2	2.0459	5	403	14.7	3.7	50	1
0	19.58	1	0.871	5.403	100	1.3216	5	403	14.7	26.82	13.4	1
0	18.1	0	0.74	6.485	100	1.9784	24	666	20.2	18.85	15.4	1
30	4.93	0	0.428	6.393	7.8	7.0355	6	300	16.6	5.19	23.7	0
0	2.46	0	0.488	7.155	92.2	2.7006	3	193	17.8	4.82	37.9	0
0	8.56	0	0.52	6.781	71.3	2.8561	5	384	20.9	7.67	26.5	0

Dimension and Structure:

The dataset has 466 records and 13 variables. The variables names and descriptions are described below:

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

The variable target is our response variable. The target variable is binary variable. Remaining 12 variables are independent variables for prediction of target variables.

Summary Statistic:

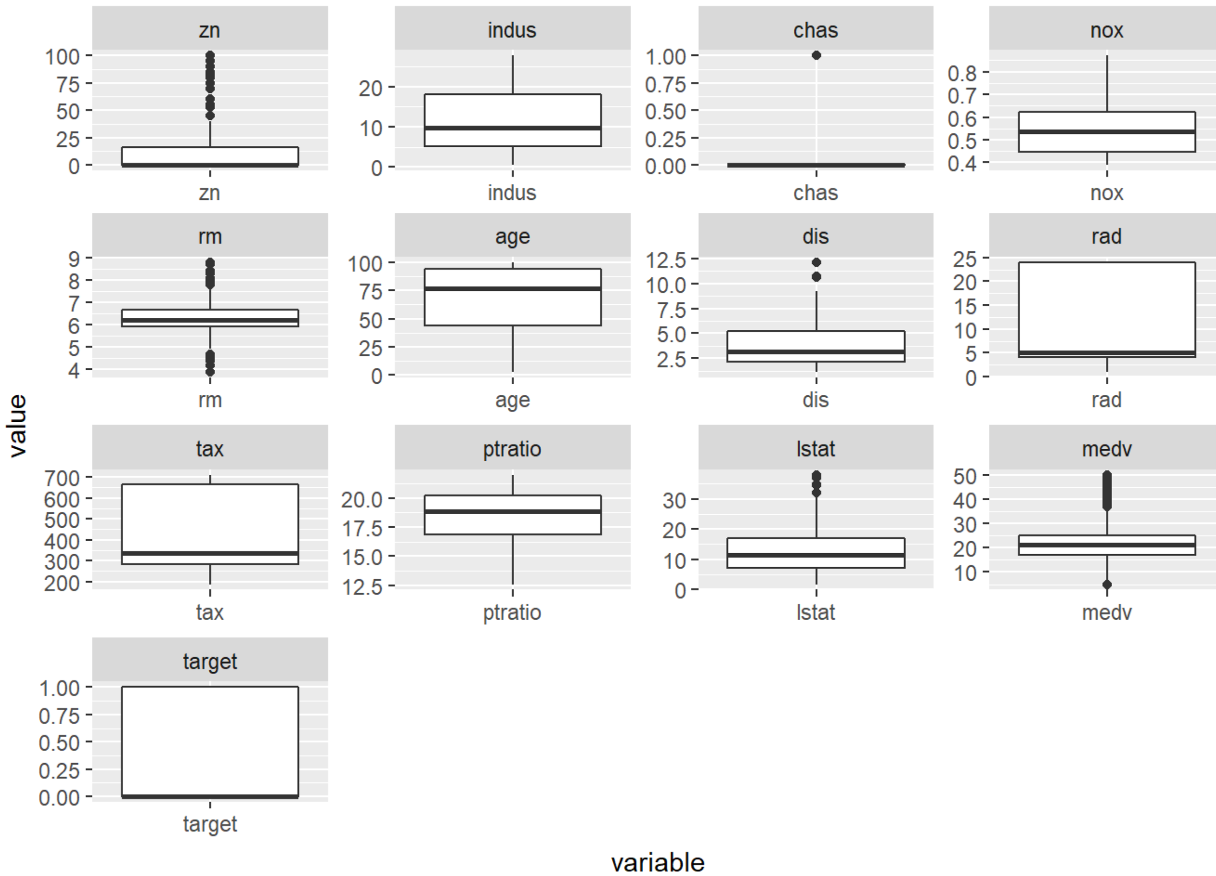
Variable	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
zn	11.58	23.36	0	5.35	0	0	100	100	2.18	3.81	1.08
indus	11.11	6.85	9.69	10.91	9.34	0.46	27.74	27.28	0.29	-1.24	0.32
chas	0.07	0.26	0	0	0	0	1	1	3.34	9.15	0.01
nox	0.55	0.12	0.54	0.54	0.13	0.39	0.87	0.48	0.75	-0.04	0.01
rm	6.29	0.7	6.21	6.26	0.52	3.86	8.78	4.92	0.48	1.54	0.03
age	68.37	28.32	77.15	70.96	30.02	2.9	100	97.1	-0.58	-1.01	1.31
dis	3.8	2.11	3.19	3.54	1.91	1.13	12.13	11	1	0.47	0.1
rad	9.53	8.69	5	8.7	1.48	1	24	23	1.01	-0.86	0.4
tax	409.5	167.9	334.5	401.51	104.52	187	711	524	0.66	-1.15	7.78
ptratio	18.4	2.2	18.9	18.6	1.93	12.6	22	9.4	-0.75	-0.4	0.1
lstat	12.63	7.1	11.35	11.88	7.07	1.73	37.97	36.24	0.91	0.5	0.33
medv	22.59	9.24	21.2	21.63	6	5	50	45	1.08	1.37	0.43
target	0.49	0.5	0	0.49	0	0	1	1	0.03	-2	0.02

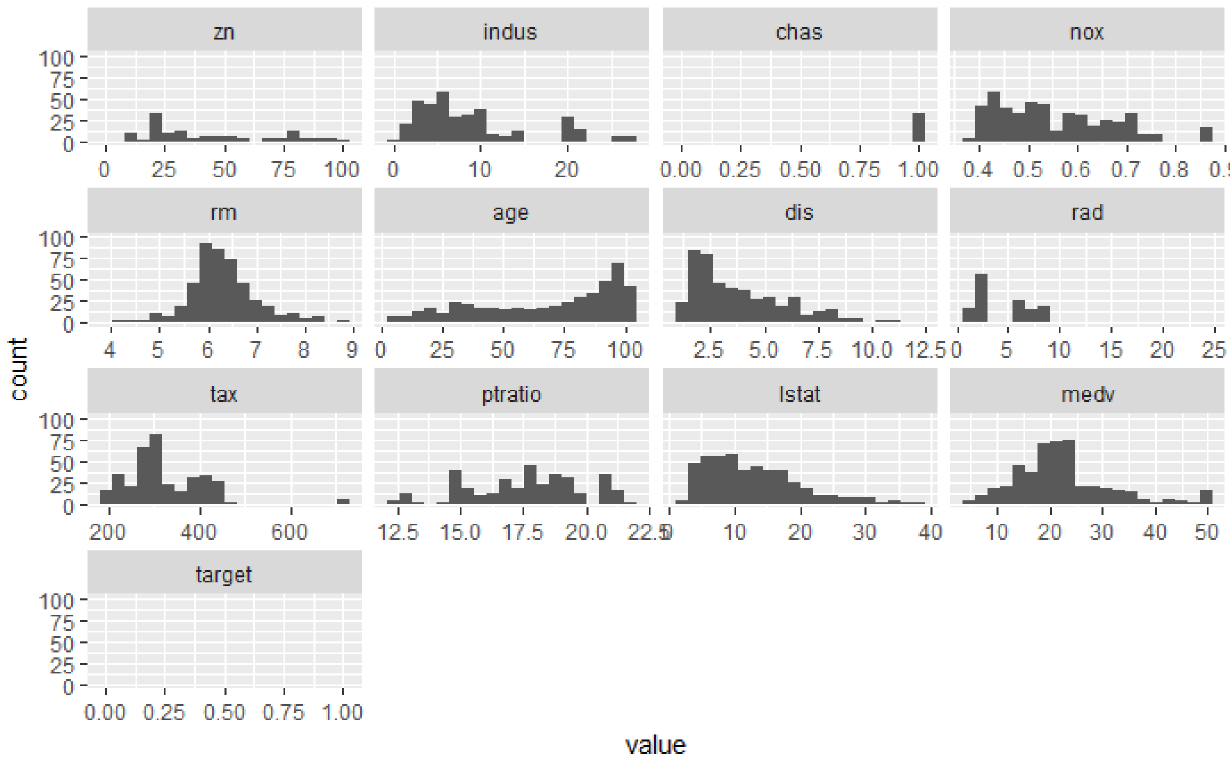
Missing Values:

There are no missing values in the dataset.

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

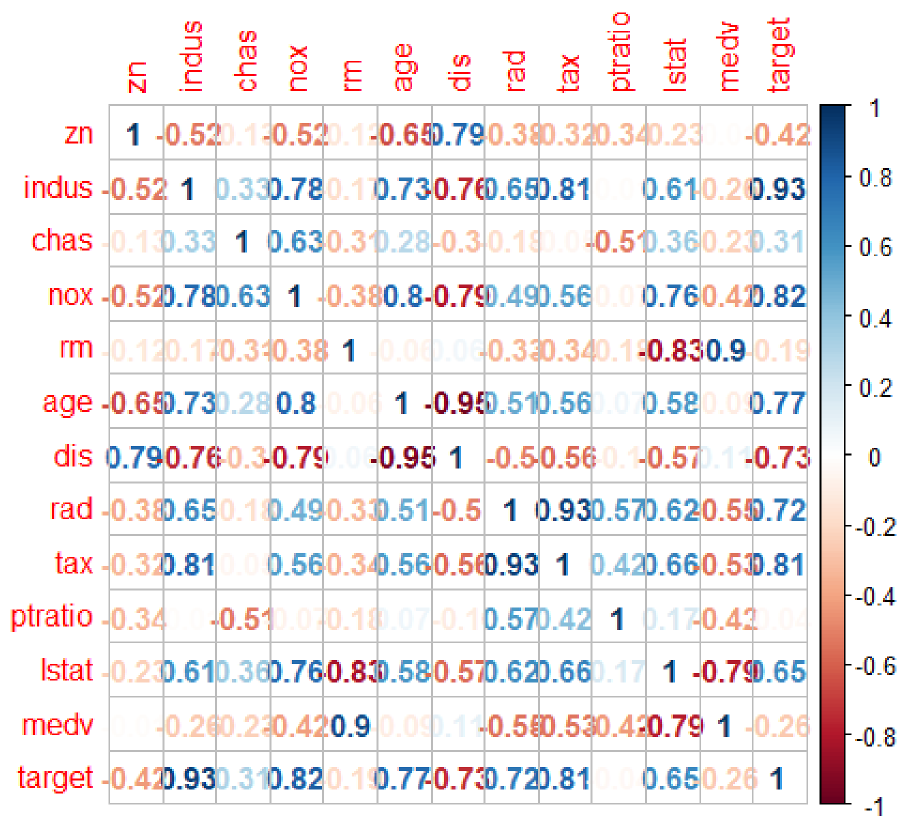
Outliers:

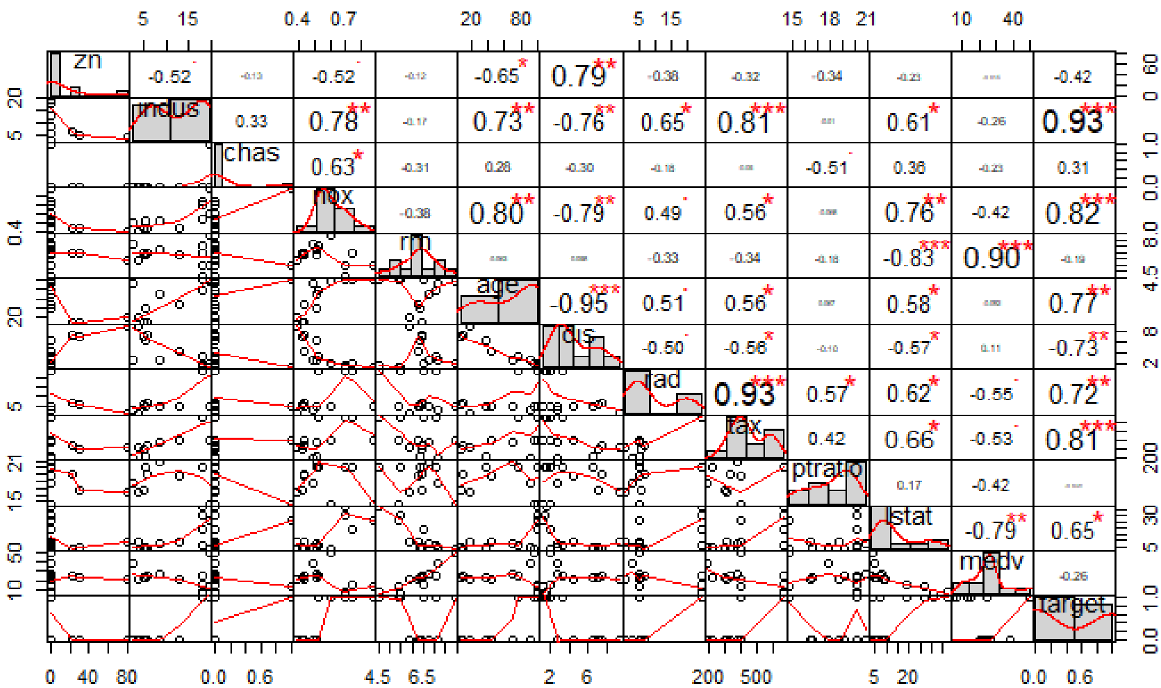

zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
lstat
medv
target
value
variable



Variables zn, rm, dis,lstat,medv contains outliers. We detected the outliers using boxplot. In addition we were created a list of the outliers in all the variables.

Correlation:





Variables indus, nox, age,rad and tax have strong positive correlation, whereas variable dis has negative correlation with independent variable target.

DATA PREPARATION:

Outliers: The outliers were fixed by winsorization by replacing them with 5^{th} and 95^{th} percentile in lower and upper tail respectively in the variables zn, rm, dis, lstat, and medv.

Variable Transformation: We will create two new variables for ptratio and rm. We will use median split to categorize the variables into high and low values. The values above median will be flagged as 1 (high) and values below median will be flagged as 0 (low). The reasoning behind this is that due to low correlation of these variables with target, we think it is a better approach to include the important information only for these variables rather than the model testing significance of all the values. In addition it might be a better to remove the original variables with categorical to test another model. The dichotomizing approach sometimes can impact your results because losing data can lead to losing information. We are selecting variables with weak correlation to lower this impact.

New variables ptratio_bkt and rm_bkt were created by dichotomizing and median split.

Logistic regression requires little or no multicollinearity among the independent variables. Based on multicollinearity assumption, we selected variable tax strongly correlated with variables indus and rad. We will create a new variable by transforming using median split. New variable tax_bkt was created using dichotomization.

Normality assumption is not required for logistic models. There are no missing values in the dataset.

Outliers have been treated by using winsorization.

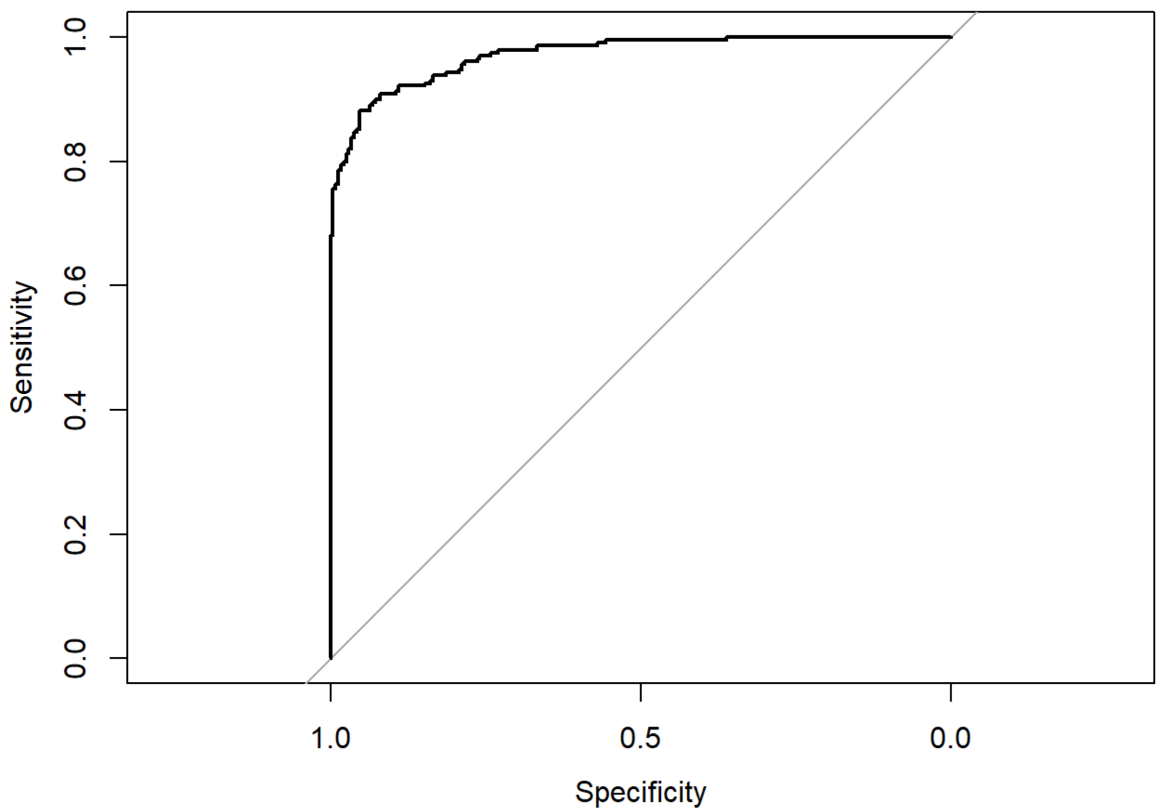
BUILD MODELS:

We will build three models for prediction. Model 1will be created using the full variables in the original dataset.

```
train1<- train[,-c(14:16)]
model1 <- glm(target ~.,family=binomial(link='logit'),data=train1)
summary(model1)
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = train1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1826  -0.1805  -0.0027   0.0037   3.3254
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -36.357916   7.372688  -4.931 8.16e-07 ***
## zn           -0.059247   0.032836  -1.804  0.07118 .
## indus        -0.059964   0.047829  -1.254  0.20995
## chas          0.916872   0.755100   1.214  0.22466
## nox          41.463864   7.182331   5.773 7.79e-09 ***
## rm            0.179293   1.144323   0.157  0.87550
```

```
## age           0.018491   0.011481   1.611  0.10727
## dis           0.400778   0.236485   1.695  0.09013 .
## rad           0.695189   0.161445   4.306 1.66e-05 ***
## tax          -0.008210   0.002824  -2.908  0.00364 **
## ptratio       0.314344   0.113269   2.775  0.00552 **
## lstat         0.070859   0.062441   1.135  0.25645
## medv          0.118212   0.083242   1.420  0.15558
## ---
## 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: 197.56  on 453  degrees of freedom
## AIC: 223.56
##
## Number of Fisher Scoring iterations: 9
```



Model 2: Backward Elimination using transformed variables

```
train2<- train[,-c(5,9:10)]
model2 <- glm(target ~.,family=binomial(link='logit'),data=train2)
summary(model2)
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = train2)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8353  -0.1582  -0.0005   0.0011   3.6099
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -31.57309    5.31509  -5.940 2.85e-09 ***
## zn           -0.07864    0.04235  -1.857   0.0633 .
## indus        -0.04698    0.05045  -0.931   0.3517
## chas          0.27028    0.85393   0.317   0.7516
## nox          43.63046    8.73783   4.993 5.94e-07 ***
## age           0.01797    0.01159   1.550   0.1211
## dis           0.13485    0.27041   0.499   0.6180
## rad           0.80242    0.19164   4.187 2.83e-05 ***
## lstat         0.05942    0.06287   0.945   0.3446
## medv          0.10285    0.04782   2.151   0.0315 *
## ptratio_bkt   0.61940    0.45080   1.374   0.1694
## rm_bkt        0.07883    0.48453   0.163   0.8708
## tax_bkt      -2.99033    0.69341  -4.313 1.61e-05 ***
## ---
## 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: 188.12  on 453  degrees of freedom
## AIC: 214.12
##
## Number of Fisher Scoring iterations: 9
```

```
model2 <- update(model2, .~. -chas,data=train2)
summary(model2)
```

```
##
## Call:
## glm(formula = target ~ zn + indus + nox + age + dis + rad + lstat +
##     medv + ptratio_bkt + rm_bkt + tax_bkt, family = binomial(link = "logit
"),
##     data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8677  -0.1595  -0.0005   0.0009   3.6176
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -31.53973    5.32281  -5.925 3.12e-09 ***
## zn           -0.08056    0.04200  -1.918   0.0551 .
## indus        -0.04306    0.04893  -0.880   0.3789
## nox          43.38409    8.71226   4.980 6.37e-07 ***
## age           0.01803    0.01156   1.560   0.1188
## dis           0.12736    0.26847   0.474   0.6352
## rad           0.81930    0.18582   4.409 1.04e-05 ***
## lstat         0.06175    0.06251   0.988   0.3232
## medv          0.10350    0.04771   2.169   0.0300 *
## ptratio_bkt   0.59351    0.44296   1.340   0.1803
## rm_bkt        0.09036    0.48238   0.187   0.8514
## tax_bkt      -3.03747    0.67885  -4.474 7.66e-06 ***
## ---
## 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: 188.22  on 454  degrees of freedom
## AIC: 212.22
##
```

```
## Number of Fisher Scoring iterations: 9
```

```
model2 <- update(model2, .~. -rm_bkt,data=train2)
summary(model2)
```

```
##
## Call:
## glm(formula = target ~ zn + indus + nox + age + dis + rad + lstat +
##     medv + ptratio_bkt + tax_bkt, family = binomial(link = "logit"),
##     data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8554  -0.1593  -0.0005   0.0009   3.6125
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -31.55636    5.33396  -5.916 3.30e-09 ***
## zn           -0.08041    0.04217  -1.907   0.0566 .
## indus        -0.04273    0.04891  -0.874   0.3823
## nox          43.32119    8.71455   4.971 6.66e-07 ***
## age           0.01854    0.01127   1.645   0.1000 .
## dis           0.13277    0.26758   0.496   0.6198
## rad           0.81916    0.18566   4.412 1.02e-05 ***
## lstat         0.05708    0.05732   0.996   0.3194
## medv          0.10703    0.04398   2.433   0.0150 *
## ptratio_bkt   0.59145    0.44290   1.335   0.1817
## tax_bkt      -3.02971    0.67721  -4.474 7.68e-06 ***
## ---
## 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: 188.26  on 455  degrees of freedom
## AIC: 210.26
##
## Number of Fisher Scoring iterations: 9
```

```
model2 <- update(model2, .~. -indus,data=train2)
summary(model2)
```

```
##
## Call:
## glm(formula = target ~ zn + nox + age + dis + rad + lstat + medv +
##     ptratio_bkt + tax_bkt, family = binomial(link = "logit"),
##     data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9392  -0.1639  -0.0005   0.0006   3.5964
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -29.90971    4.86255  -6.151 7.70e-10 ***
## zn           -0.08369    0.04073  -2.055   0.0399 *
## nox          39.48759    7.14935   5.523 3.33e-08 ***
## age           0.01778    0.01108   1.604   0.1088
## dis           0.11294    0.26542   0.426   0.6705
## rad           0.86757    0.18480   4.695 2.67e-06 ***
## lstat         0.05073    0.05627   0.902   0.3673
## medv          0.10779    0.04438   2.429   0.0152 *
## ptratio_bkt   0.57952    0.43896   1.320   0.1868
```

```
## tax_bkt      -3.16828    0.66064  -4.796 1.62e-06 ***
## ---
## 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: 189.04  on 456  degrees of freedom
## AIC: 209.04
##
## Number of Fisher Scoring iterations: 9
```

```
model2 <- update(model2, .~. -lstat,data=train2)
summary(model2)
```

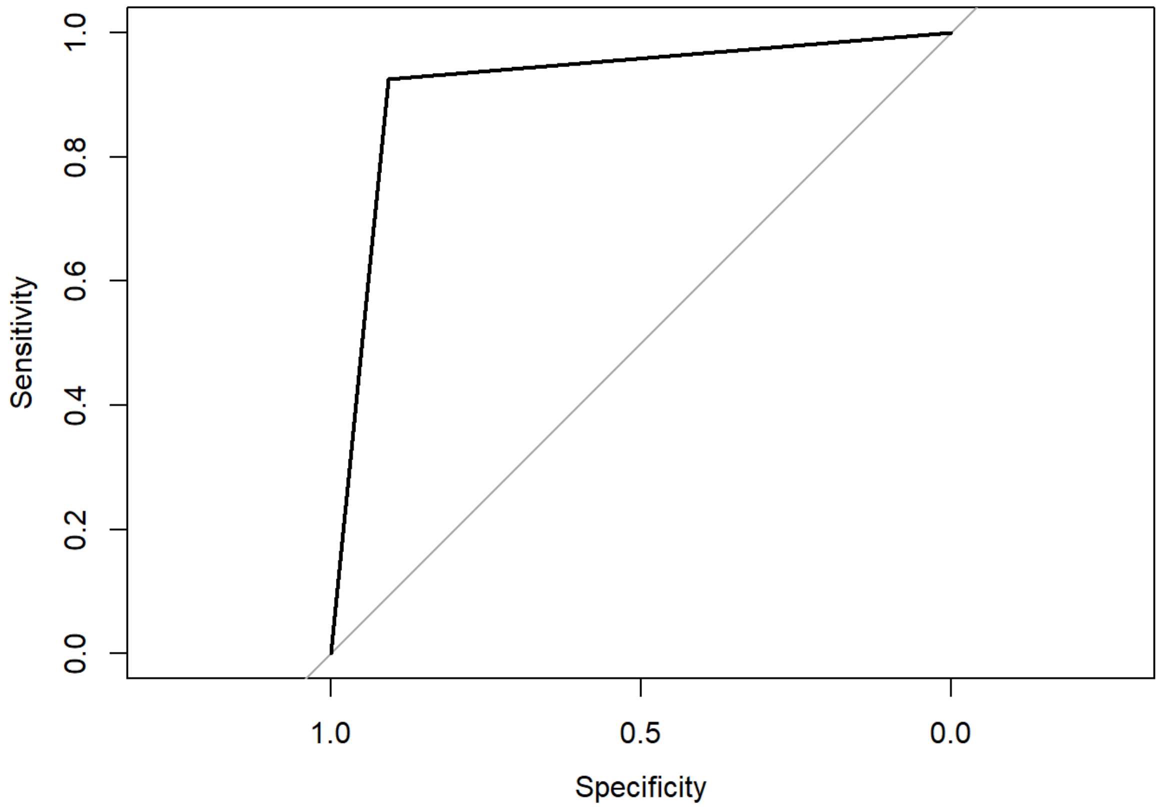
```
##
## Call:
## glm(formula = target ~ zn + nox + age + dis + rad + medv + ptratio_bkt +
##     tax_bkt, family = binomial(link = "logit"), data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9786  -0.1574  -0.0005   0.0006   3.6051
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -29.49503    4.84180  -6.092 1.12e-09 ***
## zn           -0.08149    0.04028  -2.023   0.0431 *
## nox          39.60330    7.14579   5.542 2.99e-08 ***
## age           0.02165    0.01038   2.087   0.0369 *
## dis           0.13992    0.26314   0.532   0.5949
## rad           0.85923    0.18313   4.692 2.71e-06 ***
## medv          0.10109    0.04404   2.295   0.0217 *
```

```
## ptratio_bkt   0.61839    0.43773   1.413   0.1577
## tax_bkt      -3.12848    0.65528  -4.774 1.80e-06 ***
## ---
## 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: 189.85  on 457  degrees of freedom
## AIC: 207.85
##
## Number of Fisher Scoring iterations: 9
```

```
model2 <- update(model2, .~. -ptratio_bkt,data=train2)
summary(model2)
```

```
##
## Call:
## glm(formula = target ~ zn + nox + age + dis + rad + medv + tax_bkt,
##     family = binomial(link = "logit"), data = train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0315  -0.1852  -0.0006   0.0011   3.4980
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -29.545791   4.807995  -6.145 7.99e-10 ***
## zn           -0.086158   0.041383  -2.082   0.0373 *
```

```
## nox          41.217873   7.087093   5.816 6.03e-09 ***
## age           0.018443   0.009957   1.852   0.0640 .
## dis           0.242477   0.256198   0.946   0.3439
## rad           0.797906   0.171174   4.661 3.14e-06 ***
## medv          0.084145   0.042110   1.998   0.0457 *
## tax_bkt      -2.996303   0.641854  -4.668 3.04e-06 ***
## ---
## 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: 191.87  on 458  degrees of freedom
## AIC: 207.87
##
## Number of Fisher Scoring iterations: 9
```



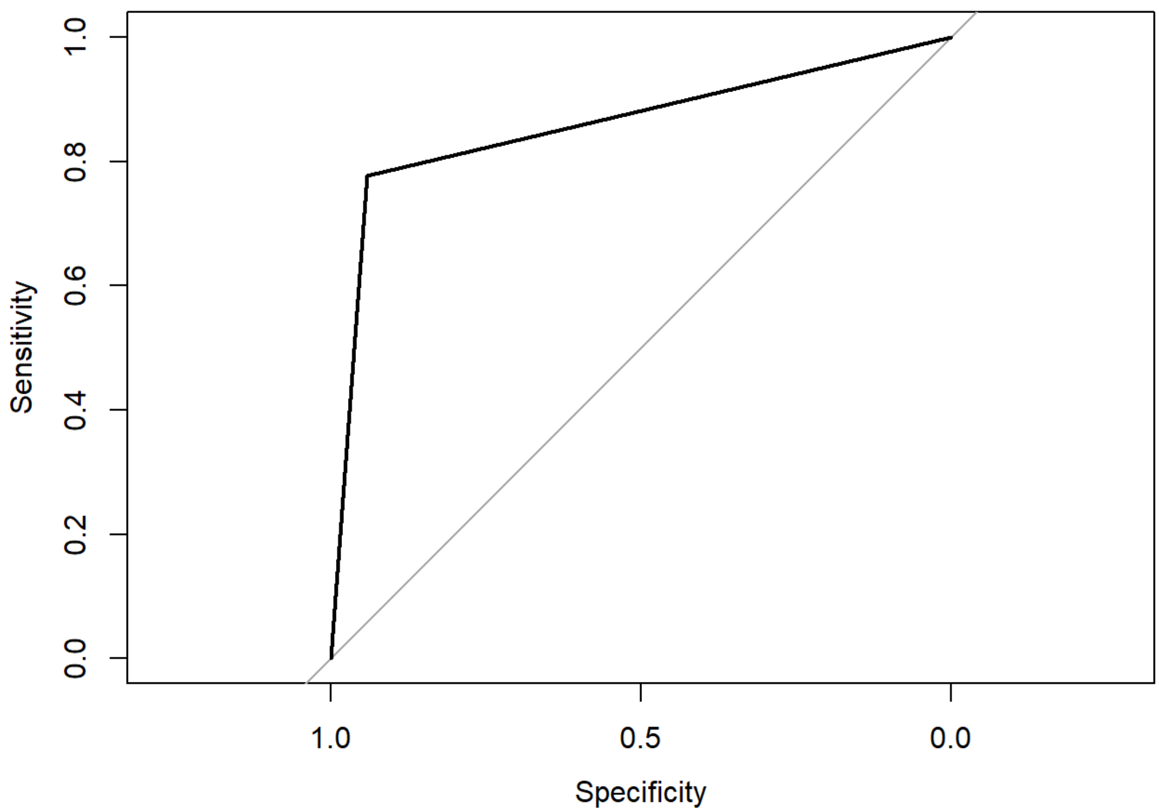
Model 3: Model 3 will be created using transformed variables and forward elimination.

```
train3<- train[,-c(5,9:10)]

model3 <- step(glm(target~ 1, data=train3), direction='forward', scope=~ zn+i
ndus+chas+nox+age+dis+age+dis+rad+lstat+medv+
                 ptratio_bkt+rm_bkt+tax_bkt)
```

```
##
## Call:
## glm(formula = target ~ nox + rad + age + medv + tax_bkt + indus,
##     data = train3)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.59008  -0.19860  -0.05885   0.14116   0.88736
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.172872   0.111328 -10.535  < 2e-16 ***
## nox          1.851349   0.227555   8.136 3.87e-15 ***
## rad          0.019303   0.002252   8.571  < 2e-16 ***
## age          0.003306   0.000764   4.327 1.86e-05 ***
## medv         0.009019   0.002517   3.583 0.000377 ***
## tax_bkt     -0.115219   0.040370  -2.854 0.004512 **
## indus        0.005491   0.003716   1.477 0.140246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09585466)
##     Null deviance: 116.466  on 465  degrees of freedom
## Residual deviance:  43.997  on 459  degrees of freedom
## AIC: 238.66
## Number of Fisher Scoring iterations: 2
```


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity

MODEL SELECTION:

```
##                         parameters_model1 parameters_model2 parameters_model3
## Sensitivity                  8.864629e-01      9.257642e-01      7.772926e-01
## Specificity                  9.367089e-01      9.071730e-01      9.409283e-01
## Pos Pred Value               9.311927e-01      9.059829e-01      9.270833e-01
## Neg Pred Value               8.951613e-01      9.267241e-01      8.138686e-01
## Precision                    9.311927e-01      9.059829e-01      9.270833e-01
## Recall                       8.864629e-01      9.257642e-01      7.772926e-01
## F1                           9.082774e-01      9.157667e-01      8.456057e-01
## Prevalence                   4.914163e-01      4.914163e-01      4.914163e-01
## Detection Rate               4.356223e-01      4.549356e-01      3.819742e-01
## Detection Prevalence         4.678112e-01      5.021459e-01      4.120172e-01
## Balanced Accuracy            9.115859e-01      9.164686e-01      8.591104e-01
## Accuracy                     9.120172e-01      9.163090e-01      8.605150e-01
## Kappa                        8.238396e-01      8.326304e-01      7.201848e-01
## AccuracyLower                8.825347e-01      8.873668e-01      8.256964e-01
## AccuracyUpper                9.361226e-01      9.398120e-01      8.906714e-01
## AccuracyNull                 5.085837e-01      5.085837e-01      5.085837e-01
## AccuracyPValue               4.908194e-79      4.712099e-81      5.872987e-58
## McnemarPValue                1.183498e-01      5.218394e-01      7.997514e-06
```

Model1

```
## Area under the curve: 0.9715
```

Model2

```
## Area under the curve: 0.9165
```

Model3

```
Area under the curve: 0.8591
```

AIC.model1.	AIC.model2.	AIC.model3.
223.5611	207.8714	238.6639

BIC.model1.	BIC.model2.	BIC.model3.
277.4355	241.0249	271.8174

For our model selection, we will utilize all the key parameters for three models. Accuracy for model one and model two is comparatively similar, and beats model three. F1 score follows similar pattern for both models. AUC, AIC and BIC values for model 1 and model 2 is better than model 3. Based on these factors we can eliminate model3.

Model 1 and model 2 will be compared for key parameters for best prediction model.

Model 2has the best parameters for AIC, BIC, Accuracy, Sensitivity than model 1. AUC for model 1 is better than model2. Model 2 had seven significant variables as compared to full 12 variables in model1. Model 2 is providing best results based on the key parameters and number of significant variables. We Will select model 2 (backward elimination) model as the best model with zn, nox, age, dis, rad, medv and tax_bkt as the significant variables.

MODEL TEST:

We will test model 2 by predicting target variable using evaluation dataset. Evaluation data was Processed in a similar way to test with model 2. Based on model 2, we predicted 22 observations with Zero value and 18 observations with value one.

```
table(pred_df)
pred_df
 0   1
22  18
```

REFRENCES:

http://www.statisticssolutions.com/assumptions-of-logistic-regression/

https://frnsys.com/ai_notes/machine_learning/model_selection.html

https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/

http://ethen8181.github.io/machine-learning/unbalanced/unbalanced.html

APPENDIX:

R code :

https://github.com/gpsingh12/Data-621/blob/master/Hw3/Singh_Hw_3.Rmd