Home Work Assignment - 03

Critical Thinking Group 5

Contents

Overview

The data set contains approximately 466 records and 14 variables. Each record has information on crime 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 objective is to build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. In addition, we will provide classifications and probabilities for the evaluation data set using the binary logistic regression model.

1 Data Exploration Analysis

In section we will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:
-Variable identification
-Variable Relationships
-Data summary analysis
-Outliers and Missing Values Identification

1.1 Variable identification

First let's display and examine the data dictionary or the data columns as shown in table 1

Table 1: Variable Description

Variable	Description	Datatype	Role
zn	proportion of residential land zoned for large lots (over 25000 square feet)	numeric	predictor
indus	proportion of non-retail business acres per suburb	numeric	predictor
chas	a dummy var. for whether the suburb borders the Charles River (1) or not (0)	binary	predictor
nox	nitrogen oxides concentration (parts per 10 million)	numeric	predictor
rm	average number of rooms per dwelling	numeric	predictor
age	proportion of owner-occupied units built prior to 1940	numeric	predictor
dis	weighted mean of distances to five Boston employment centers	numeric	predictor
rad	index of accessibility to radial highways	integer	predictor
tax	full-value property-tax rate per $10,000	integer	predictor
ptratio	pupil-teacher ratio by town	numeric	predictor
black	1000(Bk - 0.63)2 where Bk is the proportion of blacks by town	numeric	predictor
lstat	lower status of the population (percent)	numeric	predictor
medv	median value of owner-occupied homes in $1000s	numeric	predictor
target	whether the crime rate is above the median crime rate (1) or not (0)	binary	response

We notice that all variables are numeric except for two variables: the response variable "target" which is binary and the predictor variable "chas" which is a dummy binary variable indicating whether the suburb borders the Charles River (1) or not (0).

Based on the original dataset, our predictor input is made of 13 variables. And our response variable is one variable called target.

1.2 Variable Relationships

The variables seem to not have any arithmetic relations. In other words, there are no symmetricity or transitivity relationships between any two variable in the independent variable set.
In addition, since this is Logistic Regression, we will be making the below assumptions on the variables:
-The dependent variable need not to be normally distributed
-Errors need to be independent but not normally distributed.
- We will be using GLM and GLM does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.
- Also does not use OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE)

1.3 Data summary analysis

```
summary(city_crime_train_full)
```

```
##        zn              indus             chas              nox        
##  Min.   :  0.00   Min.   : 0.460   Min.   :0.00000   Min.   :0.3890  
##  1st Qu.:  0.00   1st Qu.: 5.145   1st Qu.:0.00000   1st Qu.:0.4480  
##  Median :  0.00   Median : 9.690   Median :0.00000   Median :0.5380  
##  Mean   : 11.58   Mean   :11.105   Mean   :0.07082   Mean   :0.5543  
##  3rd Qu.: 16.25   3rd Qu.:18.100   3rd Qu.:0.00000   3rd Qu.:0.6240  
##  Max.   :100.00   Max.   :27.740   Max.   :1.00000   Max.   :0.8710  
##        rm              age              dis              rad        
##  Min.   :3.863   Min.   :  2.90   Min.   : 1.130   Min.   : 1.00  
##  1st Qu.:5.887   1st Qu.: 43.88   1st Qu.: 2.101   1st Qu.: 4.00  
##  Median :6.210   Median : 77.15   Median : 3.191   Median : 5.00  
##  Mean   :6.291   Mean   : 68.37   Mean   : 3.796   Mean   : 9.53  
##  3rd Qu.:6.630   3rd Qu.: 94.10   3rd Qu.: 5.215   3rd Qu.:24.00  
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.00  
##       tax             ptratio          black             lstat       
##  Min.   :187.0   Min.   :12.6   Min.   :  0.32   Min.   : 1.730  
##  1st Qu.:281.0   1st Qu.:16.9   1st Qu.:375.61   1st Qu.: 7.043  
##  Median :334.5   Median :18.9   Median :391.34   Median :11.350  
##  Mean   :409.5   Mean   :18.4   Mean   :357.12   Mean   :12.631  
##  3rd Qu.:666.0   3rd Qu.:20.2   3rd Qu.:396.24   3rd Qu.:16.930  
##  Max.   :711.0   Max.   :22.0   Max.   :396.90   Max.   :37.970  
##       medv            target      
##  Min.   : 5.00   Min.   :0.0000  
##  1st Qu.:17.02   1st Qu.:0.0000  
##  Median :21.20   Median :0.0000  
##  Mean   :22.59   Mean   :0.4914  
##  3rd Qu.:25.00   3rd Qu.:1.0000  
##  Max.   :50.00   Max.   :1.0000
```

Table 2: Missing Values

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

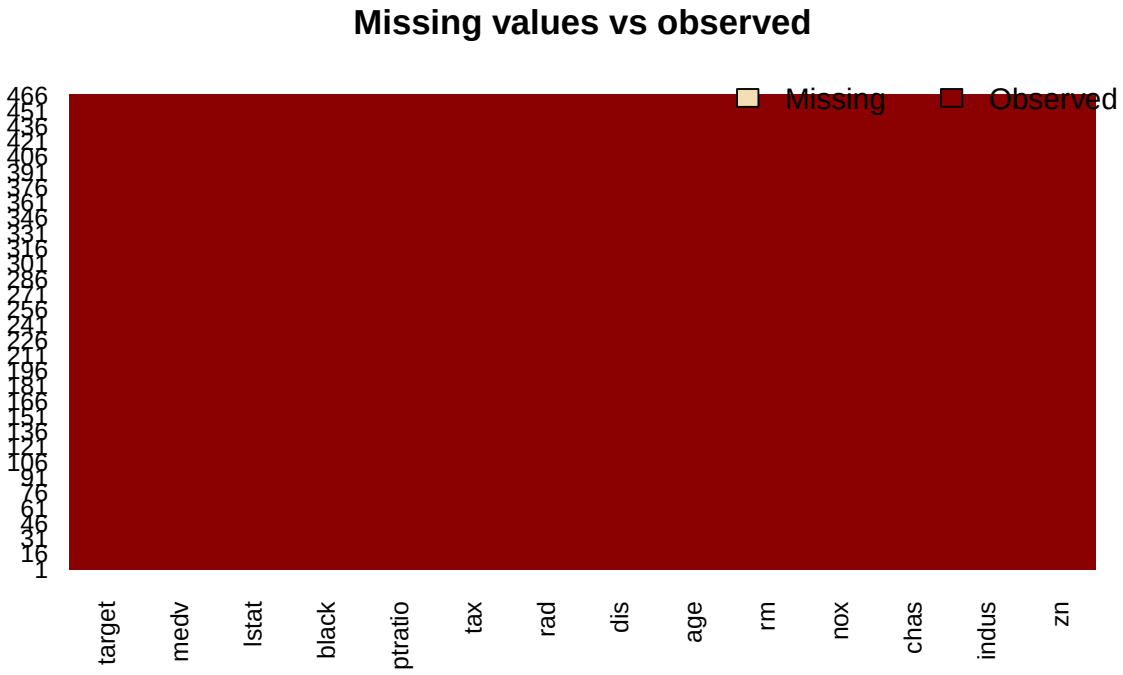


Table 3: Unique Values

zn	26
indus	73
chas	2
nox	79
rm	419
age	333
dis	380
rad	9
tax	63
ptratio	46

black	331
lstat	424
medv	218
target	2

```
##
##         0         1
## 0.5085837 0.4914163
```

Based on the analysis above it can be seen that there is no missing value in the data set. Also count of unique values for each variable is shown above. Also % split of target variable is given above table which shows data is almost evenly split between binary outcome 0 and 1.

Train data set will be Split into train data(80% of train set) and validation set (20% of train set)to evaluate the performance of the models on the validation set. Train subset will be used to build the models.

Two data set has been created city_crime_train (80% of train data), and train_test (20% of train data).
In next step below relationship between the target variable and dependent variables is shown in three charts.

1.2 Data Summary Analysis

In this section, we will create summary data to better understand the initial relationship variables have with our dependent variable using correlation, central tendency, and dispersion As shown in table 2.

Table 4: Data Summary

	vars	n	mean	sd	median	trimmed	mad
zn	1	372	12.3615591	24.0566280	0.0000	6.0436242	0.0000000
indus	2	372	10.8997043	6.9049747	8.5600	10.6588591	7.9022580
chas	3	372	0.0645161	0.2460010	0.0000	0.0000000	0.0000000
nox	4	372	0.5512384	0.1192316	0.5220	0.5394315	0.1230558
rm	5	372	6.2950027	0.6994228	6.2055	6.2664161	0.5315121
age	6	372	67.4131720	28.6905417	76.5000	69.8332215	30.9122100
dis	7	372	3.8438124	2.1293108	3.3246	3.5968611	2.0548095
rad	8	372	9.2043011	8.5398184	5.0000	8.2818792	1.4826000
tax	9	372	403.6854839	167.0523120	330.0000	394.0033557	108.2298000
ptratio	10	372	18.2325269	2.2232031	18.6000	18.4144295	2.3721600
black	11	372	359.6269355	88.5960400	391.9550	384.7739262	7.3314570
lstat	12	372	12.3974731	7.0278483	10.9250	11.6246309	6.7680690
medv	13	372	22.8473118	9.0745857	21.6000	21.9842282	6.9682200
target	14	372	0.4731183	0.4999493	0.0000	0.4664430	0.0000000

Table 5: Data Summary (Cont)

	min	max	range	skew	kurtosis	se
zn	0.0000	100.0000	100.0000	2.0480221	3.1952491	1.2472781
indus	0.4600	27.7400	27.2800	0.3403513	-1.2138927	0.3580063
chas	0.0000	1.0000	1.0000	3.5309878	10.4961121	0.0127546
nox	0.3890	0.8710	0.4820	0.8365808	0.0921839	0.0061819
rm	3.8630	8.7250	4.8620	0.3906669	1.4757590	0.0362634
age	2.9000	100.0000	97.1000	-0.5301804	-1.0930455	1.4875353
dis	1.1296	12.1265	10.9969	0.9561753	0.3805285	0.1103996
rad	1.0000	24.0000	23.0000	1.0969598	-0.6717893	0.4427690
tax	187.0000	711.0000	524.0000	0.7166865	-1.0538103	8.6612589
ptratio	12.6000	22.0000	9.4000	-0.6734659	-0.5183076	0.1152677
black	0.3200	396.9000	396.5800	-3.0998189	8.5467305	4.5934907
lstat	1.7300	37.9700	36.2400	0.9457306	0.6003832	0.3643770
medv	5.0000	50.0000	45.0000	0.9704021	1.1065501	0.4704954
target	0.0000	1.0000	1.0000	0.1072487	-1.9938358	0.0259212

Now we will produce the correlation table between the independent variables and the dependent variable

```
Correlation <- sort(Correlation, decreasing = TRUE)
kable(Correlation, caption = "Variable Correlation")
```

Table 6: Variable Correlation

target	1.0000000
nox	0.7290920
rad	0.6307187
age	0.6275762
indus	0.6034795
tax	0.6021403
lstat	0.4808888
ptratio	0.2198922
chas	0.0579716
rm	-0.1605913
medv	-0.2724789
black	-0.3463425
zn	-0.4239382
dis	-0.6167264

*** Curious It is clear from the table that most of the variables are having strong correlation with the target variable.

Correlation analysis suggests that there are strong positive and negative between the independent variables and the dependent variable. For instance, we notice that there is a strong correlation of .73 between the concentration of nitrogen oxides and crime rate being above average. We will need to perform more investigations about this correlation as it is not obvious the concentration of nitrogen oxides would results in high crime rate; perhaps it impacts the crime rate indirectly by impacting other independent variables that we may or may not have in our data set.
In addition, we noticed that accessibility to radial highways also has a strong correlation with the crime rate being average average. Again we will investigate such correlation. We also noticed that unit or house age, property tax, and non-retail businesses having a positive impact on the crime rate being above average.

It is also worth noting that that distances to five Boston employment centers, large residential lots, the proportion of blacks by town, median value of owner-occupied homes, and the average number of rooms per dwelling, all have negative correlation to the crime rate being above crime rate average. In other words, the closer people are to the five Boston employment centers, the more likely the crime rate will be below the crime average.

1.3 Outliers and Missing Values Identification

In this section uni variate analysis is being carried out and boxplots diagrams are being used to determine the outliers in variables and decide on whether to act on the outliers. Along with boxplot, Histrogram, Sin, Log,Sqrt,nth transformation diagrams are used to evaluate best transformation to handle outliers.

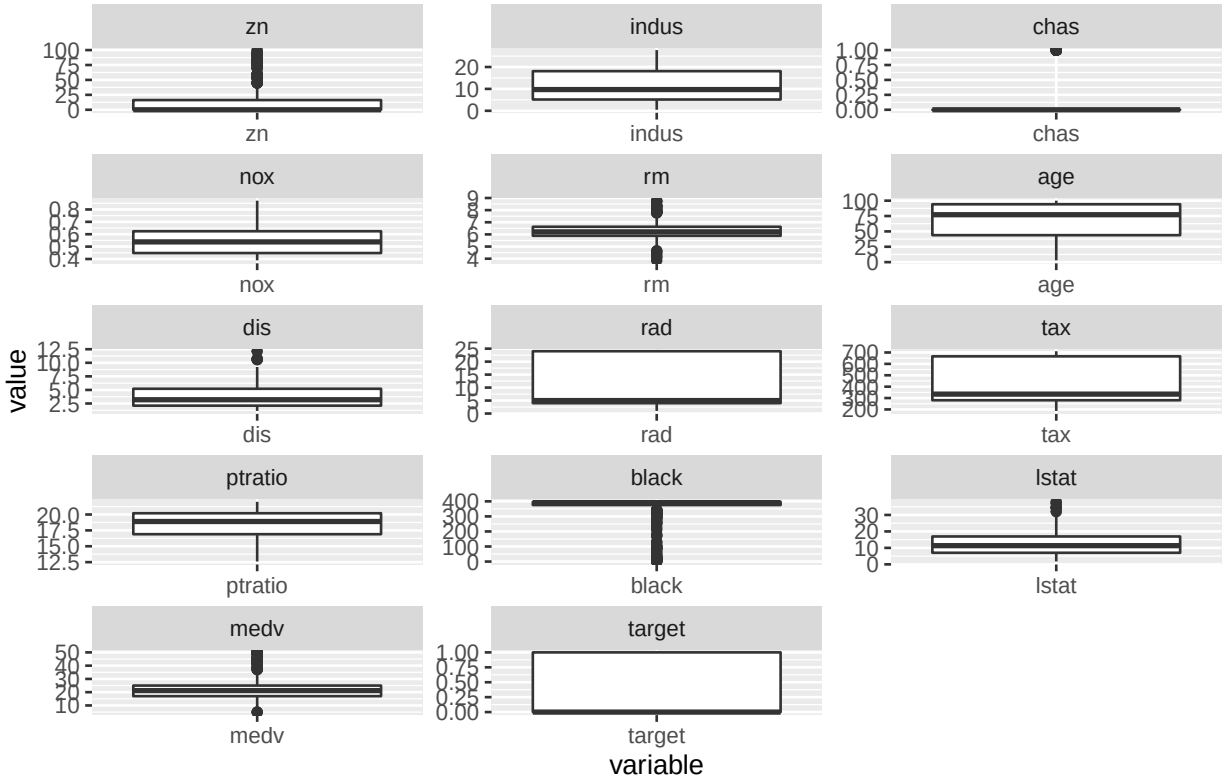
```
library(ggplot2)
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 3.2.3
```

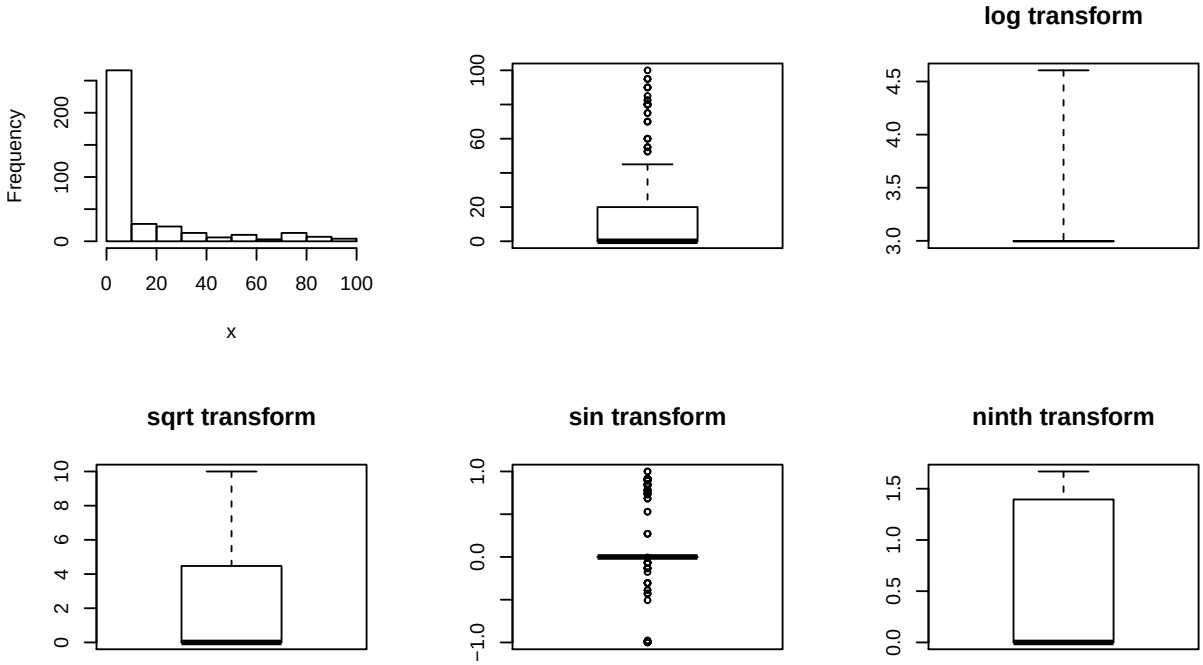
```
#create a new data frame with two columns only (variable, value) for all three predictors
mdata <- melt(city_crime_train_full)
```

```
## No id variables; using all as measure variables
```

```
# Output the boxplot
p <- ggplot(data = mdata, aes(x=variable, y=value)) +
  geom_boxplot()
p + facet_wrap( ~ variable, scales="free", ncol=3)
```



Analysis of variable zn:proportion of residential land zoned for large lots



For zn, we can see that there are large number of values with 0. ninth transformation seem better for this variable..(1)

*

**Please note that we have created similar figures to figure 1 above for each remaining variable. However, we hid the remaining figures for ease of streamlining the report as they have similar shapes. However, we have drawn the below observations from each remaining figure.

For indus, we can see that there is a spike toward right side of he distribution. Looking at the sqrt transformation it appears that distribution is close to normal and having two peaks after transformation.

For nox, there is a long right tail.

For rm, there are some outliers as we can see from box plot. This variable will need some transformation to handle the outliers.

age of the building variable is skewed heavily towards right side. We will need some transformation for this variable and looks sin transformation is best option for this case

For this variable dis, there are some outliers which needs transformation to handle those outliers. log transformation looks best suited for this scenario.

For rad variable distribution is not uniform as seen from the chart and will need transformation.

For tax variable is not uniformly distributed but there is no outlier for this variable.

For pratio has right aligned peak but no outliers are there in data set.

The variable lstat has long right tail and lef skewed

2. Data Preparation

Now that we have completed the preliminary analysis, we will be cleaning and consolidating data into one dataset for use in analysis and modeling. We will be pursing the below steps as guidelines:
- Outliers treatment
- Missing values treatment
- Data transformation

2.1 Outliers treatment and transformation

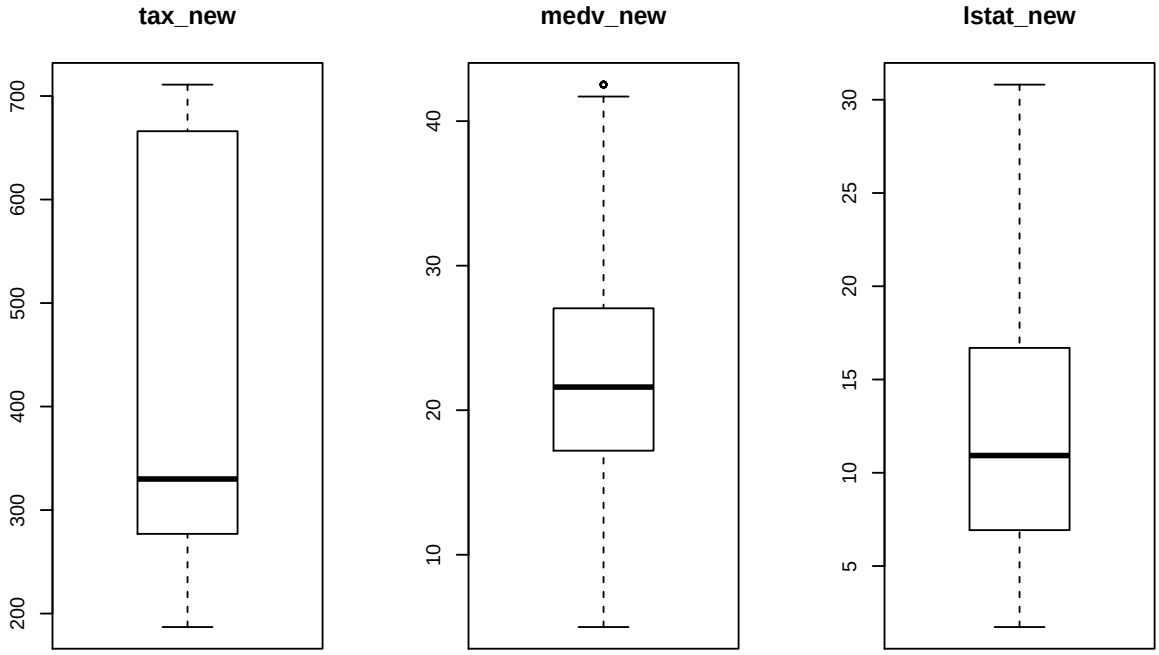
For outliers, we will create 2 sets of variables.

The first set uses the capping method. In this method, we will replace all outliers that lie outside the 1.5 times of IQR limits. We will cap it by replacing those observations less than the lower limit with the value of 5th %ile and those that lie above the upper limit with the value of 95th %ile.

Accordingly we create the following new variables while retaining the original variables.

city_crime_traintax $city_crime_train$medv
city_crime_train$lstat

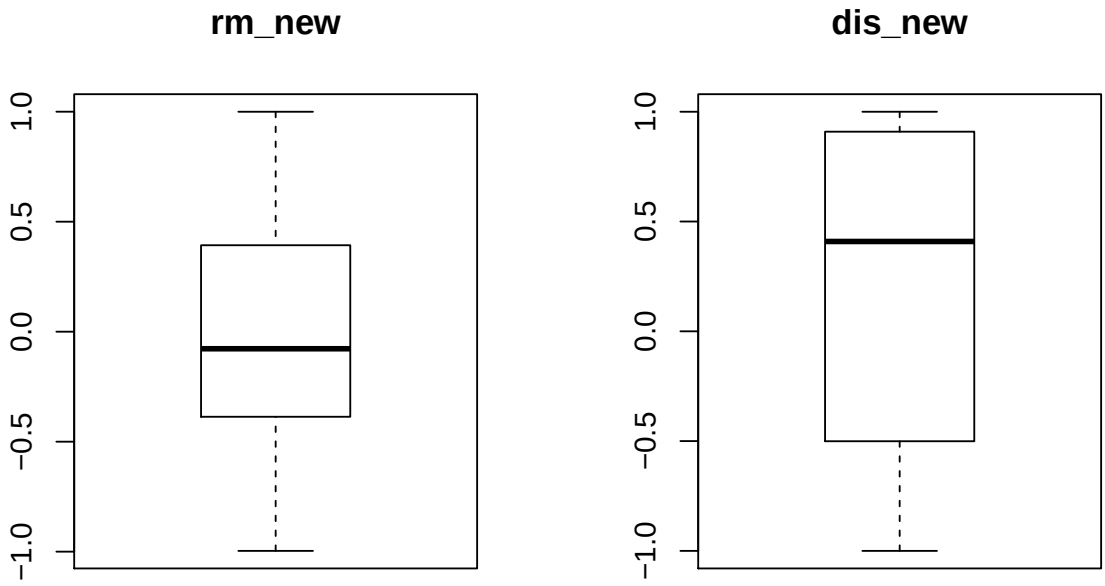
Below boxplots shows distribution of variables after outliers treatment.



In the second set, we will use the sin transformation and create the following variables:

city_crime_train_modrm_new $city_crime_train_mod$dis_new

Below is the boxplot after sin transformation of above variable.



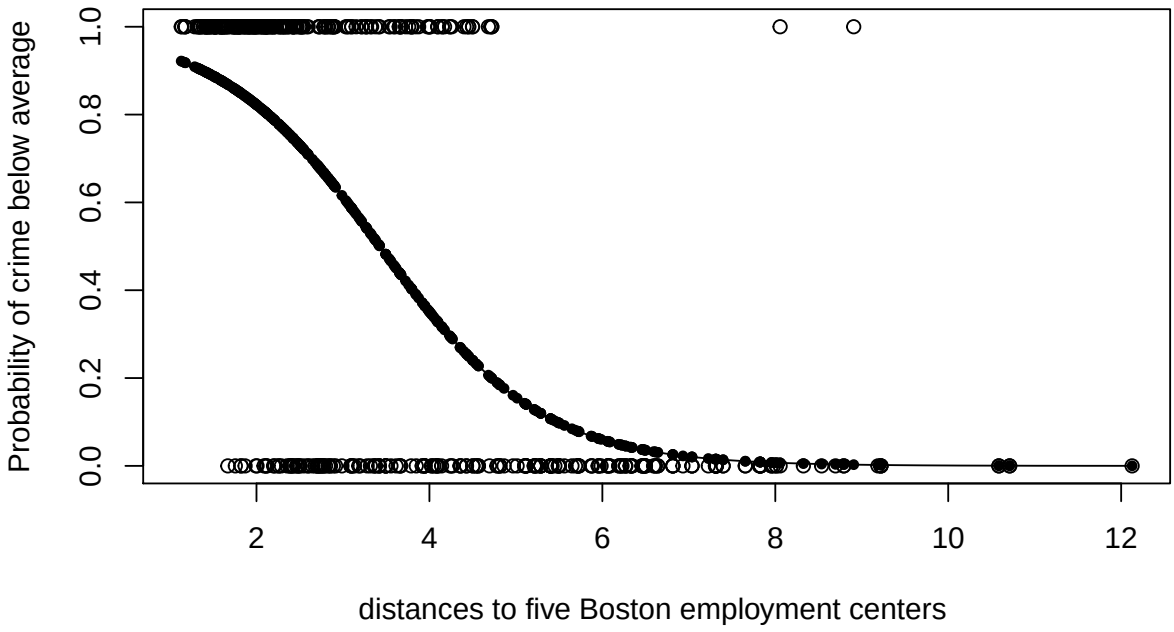
Additional transformation was performed on following variables
1. using bucket for zn, with set of values 0 and 1
2. Converting chas to a factor variable of 0 and 1
3. Converting target to a factor variable of 0 and 1

below we evaluate correlation of target with new variables

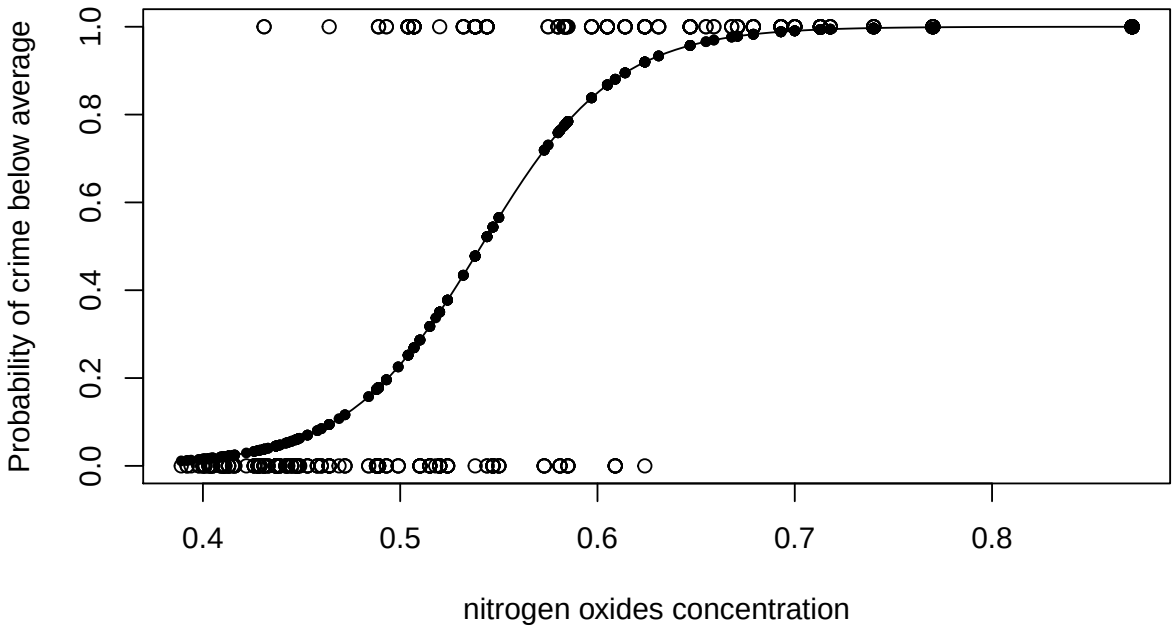
All new variables seem to have a positive correlation with target. However, some of them do not seem to have a strong correlation. Lets see how they perform while modeling.

*****For every variable the the model model

```
x = city_crime_train_full
plot(x$dis,x$target,xlab="distances to five Boston employment centers ",ylab="Probability of crime belo
g=glm(target~dis,family=binomial,x) # run a logistic regression model (in this case, generalized linear
curve(predict(g,data.frame(dis=x),type="resp"),add=TRUE) # draws a curve based on prediction from logis
points(x$dis,fitted(g),pch=20)
```




```
x = city_crime_train_full
plot(x$nox,x$target,xlab="nitrogen oxides concentration",ylab="Probability of crime below average") #
g=glm(target~nox,family=binomial,x) # run a logistic regression model (in this case, generalized linear
curve(predict(g,data.frame(nox=x),type="resp"),add=TRUE) # draws a curve based on prediction from logis
points(x$nox,fitted(g),pch=20)
```



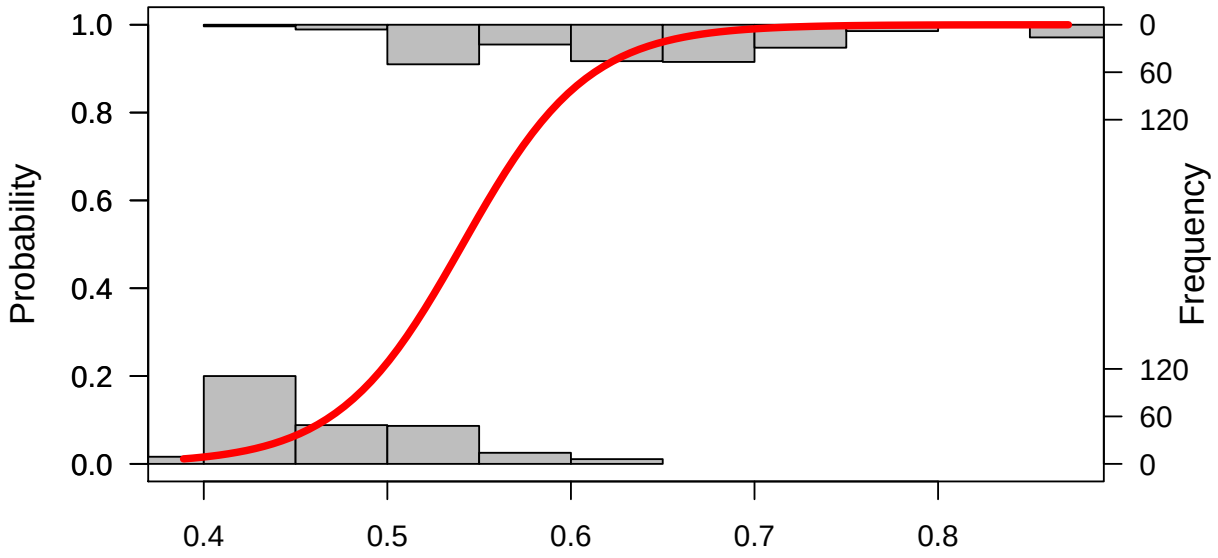

```
library(popbio)
```

```
## Warning: package 'popbio' was built under R version 3.2.4
```

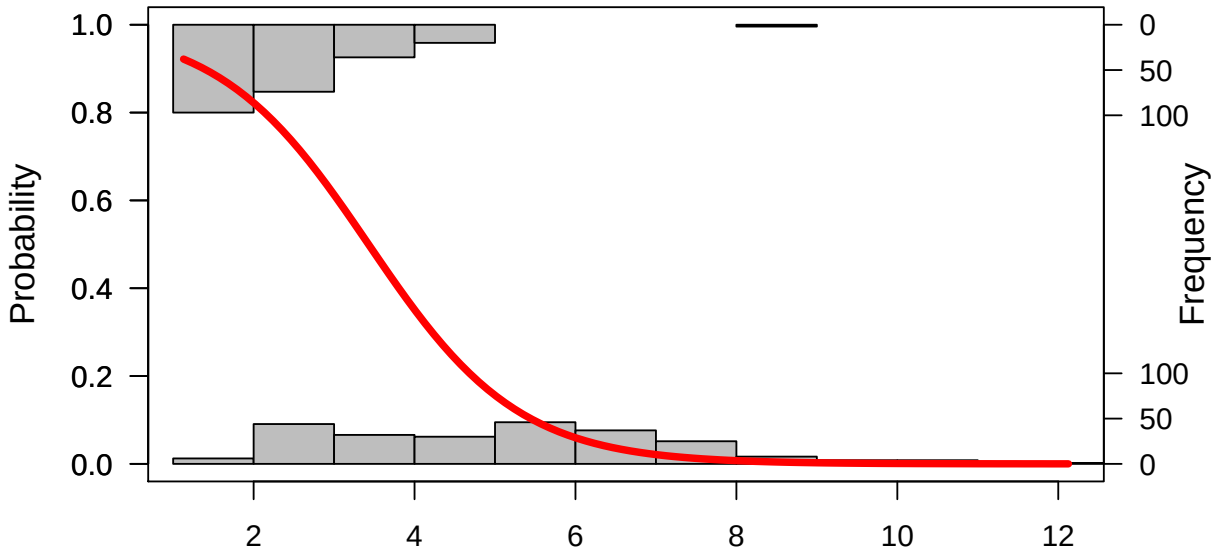
```
##
## Attaching package: 'popbio'
```

```
## The following object is masked from 'package:AUC':
##
##     sensitivity
```

```
logi.hist.plot(x$nox,x$target,boxp=FALSE,type="hist",col="gray")
```




```
library(popbio)
logi.hist.plot(x$dis, x$target,boxp=FALSE,type="hist",col="gray")
```


1.0
0.8
0.6
0.4
0.2
0.0
Probability
0
50
100
100
50
0
Frequency
2
4
6
8
10
12

3 Build Models

Below is a summary table showing models and their respective variables.

Table 7: Variables used in different models

Variables	Model.1	Model.2	Model.3	Model.4	Model.5	Model.6
zn	y	y			y	y
indus	y	y	y	y	y	y
chas	y	y	y	y	y	y
nox	y	y	y	y	y	y
rm	y	y	y	y	y	y
age	y	y	y	y	y	y
dis	y	y	y	y	y	y
rad	y	y	y	y	y	y
tax	y	y			y	y
ptratio	y	y	y	y	y	y
black	y	y	y	y	y	y
lstat	y	y			y	y
medv	y	y			y	y
tax_new			y	y		y
medv_new			y	y		y
lstat_new			y	y		y
zn_new			y	y		y

Following strategy has been adopted to build models for this scenario:

Model 1- This model has been created using the available variables in train data set with logit function GLM.
Model 2- In this model step function is being used to enhance model 1 using train data.
Model 3- Here , a new model has been created using GLM function and with transformed variables.
Model 4- In this model model 3 has been enhanced by using step function on the transformed data .
Model 5 Linear Discriminant Analysis function lda in ISLR package has been used to create model 5 with given variables.
Model 6- Here Linear Discriminant Analysis model is used on transformed variables.

3.1.1 Model One by using all given variables

In this model, we will be using all the given variables in train data set. We will create model using logit function and we will highlight the summary of the model.

```
##
## Call:
## glm(formula = target ~ ., family = "binomial", data = city_crime_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8791  -0.1299  -0.0025   0.0011   3.4785
##
## Coefficients:
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -41.462153   8.250799  -5.025 5.03e-07 ***
## zn           -0.060580   0.039153  -1.547 0.121799
## indus        -0.063885   0.059335  -1.077 0.281618
## chas          0.789391   0.865818   0.912 0.361912
## nox          53.413503  10.013666   5.334 9.60e-08 ***
## rm           -0.647942   0.904430  -0.716 0.473739
## age           0.028835   0.015680   1.839 0.065915 .
## dis           0.800917   0.268877   2.979 0.002894 **
## rad           0.721751   0.195662   3.689 0.000225 ***
## tax          -0.007065   0.003490  -2.024 0.042948 *
## ptratio       0.440768   0.159366   2.766 0.005679 **
## black        -0.009591   0.006025  -1.592 0.111412
## lstat         0.096941   0.062429   1.553 0.120469
## medv          0.236940   0.091276   2.596 0.009436 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 514.63  on 371  degrees of freedom
## Residual deviance: 140.71  on 358  degrees of freedom
## AIC: 168.71
##
## Number of Fisher Scoring iterations: 9
```

Interpretation for model 1

(i) Based on the outcome it can be seen that indus,chas,rm,age,black and lstat are not statistically significant.

(ii)As for the statistically significant variables, nox has the lowest p-value suggesting a strong association of the nox of the target variable. other important variables are dis,rad,tax,ptratio,medv. AIC value for the model1 =168.71.

(iii) The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variables.

a. For every one unit change in nox, the log odds of crime rate above median value increases by 53.41.

b. For a one unit increase in dis, the log odds of crime rate above median value increases by 0.80.

c. For a one unit increase in rad, the log odds of crime rate above median value increases by 0.72.

d. For a one unit increase in tax, the log odds of crime rate above median value increases by -0.007. Tax has a negative impact on crime rate.

e. For a one unit increase in ptratio, the log odds of crime rate above median value increases by 0.44.

f. For a one unit increase in medv , the log odds of crime rate above median value increases by 0.23.

(iv) No. of iterations are 9 before lowest value of AIC was derived for this model.

3.1.2 Model 2 with step function (backward process) with all given variables

In this model, model 1 will be enhanced with by using step function on the same train data set.

```
##
## Call:
## glm(formula = target ~ zn + nox + age + dis + rad + tax + ptratio +
##     black + lstat + medv, family = "binomial", data = city_crime_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9258  -0.1459  -0.0024   0.0013   3.3934
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -39.282116   7.705519  -5.098 3.43e-07 ***
## zn           -0.064656   0.037414  -1.728 0.083964 .
## nox          46.617168   8.074920   5.773 7.78e-09 ***
## age           0.025273   0.013545   1.866 0.062065 .
## dis           0.710480   0.249767   2.845 0.004447 **
## rad           0.775881   0.182072   4.261 2.03e-05 ***
## tax          -0.009144   0.003082  -2.967 0.003011 **
## ptratio       0.359297   0.135081   2.660 0.007817 **
## black        -0.008384   0.005737  -1.462 0.143871
## lstat         0.110624   0.055650   1.988 0.046829 *
## medv          0.181460   0.053572   3.387 0.000706 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 514.63  on 371  degrees of freedom
## Residual deviance: 142.85  on 361  degrees of freedom
## AIC: 164.85
##
## Number of Fisher Scoring iterations: 9
```

Interpretation for model 2

(i)It can be seen that zn,age,black are not statistically significant.

(ii)As for the statistically significant variables, nox has the lowest p-value suggesting a strong association of the nox of the target variable. other important variables are dis,rad,tax,ptratio,medv,lstat. AIC value for the model1 =164.85

(iii) The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variables.

a. For every one unit change in nox, the log odds of crime rate above median value increases by 46.61.

b. For a one unit increase in dis, the log odds of crime rate above median value increases by 0.71.

c. For a one unit increase in rad, the log odds of crime rate above median value increases by 0.77.

d. For a one unit increase in tax, the log odds of crime rate above median value increases by -0.009.

e. For a one unit increase in ptratio, the log odds of crime rate above median value increases by 0.35.

f. For a one unit increase in medv , the log odds of crime rate above median value increases by 0.18

(iv) there were 9 iterations in backward steps before final model was selected

3.1.3 Model three with transformed variables

In this model, transformed variables are being used with the logit function GLM.

```
##
## Call:
## glm(formula = target ~ . - zn - tax - lstat - medv, family = "binomial",
##     data = city_crime_train_mod)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7883  -0.1410  -0.0026   0.0005   3.3645
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -68.319369  16.418997  -4.161 3.17e-05 ***
## indus        -0.001867   0.067017  -0.028 0.977778
## chas1         0.366993   0.849076   0.432 0.665577
## nox          56.080643  10.147964   5.526 3.27e-08 ***
## rm            2.995884   2.385419   1.256 0.209147
## age           0.043435   0.018166   2.391 0.016805 *
## dis           0.472036   0.331312   1.425 0.154231
## rad           0.838409   0.237364   3.532 0.000412 ***
## ptratio       0.468316   0.176293   2.656 0.007896 **
## black        -0.010739   0.005922  -1.813 0.069782 .
## tax_new      -0.005285   0.003663  -1.443 0.149151
## medv_new      0.283102   0.106228   2.665 0.007698 **
## lstat_new     0.050027   0.074958   0.667 0.504515
## rm_new       -5.052053   2.830695  -1.785 0.074304 .
## dis_new      -1.886385   0.552223  -3.416 0.000636 ***
## zn_new       -0.363834   1.036508  -0.351 0.725574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 514.63  on 371  degrees of freedom
## Residual deviance: 124.11  on 356  degrees of freedom
## AIC: 156.11
##
## Number of Fisher Scoring iterations: 9
```

Interpretation for model 3

(i)From this model it can be seen following variables are relevant for this model-nox, dis, rad, ptratio , tax_new, medv_new,lstat_new.
(ii) number of integration is 9 and AIC value =169.71.
(iii) nox and rad are the two most important variables. New variables tax_new,medv_new,lstat_new are having minor impact on the model.

(iv) The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variables.

a. For every one unit change in nox, the log odds of crime rate above median value increases by 56.02.

b. For a one unit increase in rad, the log odds of crime rate above median value increases by 0.72.

c. For a one unit increase in dis, the log odds of crime rate above median value increases by 0.82.

3.1.4 Model with transformed variable and with with backward step function

In this model, transformed variables are being used with the step function and backward process.

```
##
## Call:
## glm(formula = target ~ nox + age + dis + rad + ptratio + black +
##     tax_new + medv_new + rm_new + dis_new, family = "binomial",
##     data = city_crime_train_mod)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0158  -0.1472  -0.0031   0.0005   3.1030
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -52.779764   9.739144  -5.419 5.98e-08 ***
## nox          56.509319   9.188179   6.150 7.74e-10 ***
## age           0.051467   0.016215   3.174 0.001503 **
## dis           0.564992   0.255943   2.207 0.027280 *
## rad           0.849127   0.212643   3.993 6.52e-05 ***
## ptratio       0.533319   0.159365   3.347 0.000818 ***
## black        -0.010960   0.005943  -1.844 0.065147 .
## tax_new      -0.004534   0.003144  -1.442 0.149355
## medv_new      0.342778   0.095427   3.592 0.000328 ***
## rm_new       -2.358513   1.028472  -2.293 0.021835 *
## dis_new      -1.865533   0.488896  -3.816 0.000136 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 514.63  on 371  degrees of freedom
## Residual deviance: 126.80  on 361  degrees of freedom
```

```
## AIC: 148.8
##
## Number of Fisher Scoring iterations: 9
```

Interpretation for model 4

(i)From this model it can be seen following variables are relevant for this model-nox, dis, rad, ptratio , tax_new, medv_new,lstat_new
(ii) number of integration is 9 and AIC value =165.8.

(iii) The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variables.

a. For every one unit change in nox, the log odds of crime rate above median value increases by 48.61.

b. For a one unit increase in rad, the log odds of crime rate above median value increases by 0.79.

c. For a one unit increase in dis, the log odds of crime rate above median value increases by 0.74.

(iv) same variables as model3 are being marked as relevant for model 4 after backward elimination process.

3.1,5 Model three with Linear Discriminant Analysis

In this model Linear Discriminant Analysis function has been used with given set of variables in training data.

```
## Call:
## lda(target ~ ., data = city_crime_train)
##
## Prior probabilities of groups:
##         0         1
## 0.5268817 0.4731183
##
## Group means:
##          zn     indus       chas       nox       rm      age      dis
## 0 22.012755  6.956327 0.05102041 0.4689730 6.401296 50.37398 5.086538
## 1  1.613636 15.291193 0.07954545 0.6428523 6.176631 86.38864 2.459868
##         rad      tax  ptratio    black     lstat     medv
## 0  4.107143 308.4949 17.76990 388.6647  9.199235 25.18724
## 1 14.880682 509.6932 18.74773 327.2894 15.959148 20.24148
##
## Coefficients of linear discriminants:
##                   LD1
## zn      -0.0047914631
## indus    0.0281044279
## chas    -0.0556293189
## nox      7.9109306913
## rm       0.1658180998
## age      0.0131973114
## dis      0.0840623852
```

```
## rad      0.1027832012
## tax     -0.0019152605
## ptratio  0.0090391049
## black   -0.0009160458
## lstat    0.0248449648
## medv     0.0425514709
```

Interpretation for model 5

(i) summary provides prior probability of outcome before star of model

(ii) Group means provides mean values for variables with respect to target variable values 0 and 1 here

(iii) One point to note here this model performs less accurately compared to earlier logistics models.LDA models assumes normality of its variable and hence the outliers than we have seen in actual model is impacting the result out of this model.

3.1.6 Model with Linear Discriminant Analysis with transformed data

In this model Linear Discriminant Analysis function has been used with transformed set of variables in training data.

```
## Call:
## lda(target ~ . - zn - rm - dis - tax - lstat - medv, data = city_crime_train_mod)
##
## Prior probabilities of groups:
##         0         1
## 0.5268817 0.4731183
##
## Group means:
##       indus      chas1       nox      age       rad  ptratio    black
## 0  6.956327 0.05102041 0.4689730 50.37398  4.107143 17.76990 388.6647
## 1 15.291193 0.07954545 0.6428523 86.38864 14.880682 18.74773 327.2894
##    tax_new medv_new lstat_new      rm_new    dis_new     zn_new
## 0 308.4949 25.04528  9.199235  0.08333182 -0.0504096 0.46938776
## 1 509.6932 19.86151 15.724247 -0.11166891  0.5106930 0.07954545
##
## Coefficients of linear discriminants:
##                    LD1
## indus      0.022452946
## chas1     -0.186416323
## nox        7.970446650
## age        0.015169354
## rad        0.100159450
## ptratio   -0.014404341
## black     -0.001159202
## tax_new   -0.001196341
## medv_new   0.047596449
## lstat_new  0.016840318
## rm_new    -0.008946209
## dis_new   -0.340985994
## zn_new    -0.001832533
```

Interpretation for model 6

(i) summary provides prior probability of outcome before star of model

(ii) Group means provides mean values for variables with respect to target variable values 0 and 1 here

(iii) One point to note here this model performs better than the previous one as outliers were taken care of in transformed set bringing more normality to the model. But overall this model also perform less than logistics model.

4 Model Selection

In section we will further examine all six models. We will apply a model selection strategy defined below to compare the models.

4.1 Model selection strategy:

Following model selection strategy has been used for this assignment:

(i) Compare accuracy of the models & confusion matrix
(ii) Compare Precision,Sensitivity,Specificity,F1 score
(iii) Compare AUC curve for the models

Following function Eval() will be used to calculate various metrics related to the model like Accuracy, Sensitivity, Precision , Specificity and F1 score

4.1.1 Model1 Evaluation

```
## Warning: package 'pROC' was built under R version 3.2.5
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:AUC':
##
##     auc, roc
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

Table 8: Model 1 evaluation KPIs

Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
0.9042553	0.0957447	0.9245283	0.9074074	0.9	0.9283174	0.9549011

Looking at the key metrics this can be concluded this model has high accuracy 0.9042553 rate. AUC for this model is 0.9549 which is very good. Always the optimal value for AUC is (0,1) and closer it goes to 1 values better the model outcome is.

4.1.2 Model2 Evaluation

Table 9: Model 2 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
2	0.8723404	0.1276596	0.9056604	0.8727273	0.8717949	0.9061444	0.9553613

Looking at the key metrics this can be concluded this model has high accuracy 0.8723404 and low error rate 0.12765957.AUC curve for this model is 0.9553 which is very good.

4.1.3 Model3 Evaluation

Table 10: Model 3 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
3	0.893617	0.106383	0.9245283	0.8909091	0.8974359	0.9211541	0.9677865

Looking at the key metrics this can be concluded this model has high accuracy 0.8936170and low error rate 0.10638298.AUC curve for this model is 0.9558 which is very good.

4.1.4 Model4 Evaluation

Table 11: Model 4 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
4	0.8829787	0.1170213	0.9056604	0.8888889	0.875	0.9127916	0.9687069

Looking at the key metrics this can be concluded this model has high accuracy 0.8829787 and low error rate 0.11702128.AUC curve for this model is 0.9549 which is very good.

4.1.5 Model5 Evaluation

Table 12: Model 5 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
5	0.8297872	0.1702128	0.7358491	0.9512195	0.7358491	0.8297872	0.9263691

Looking at the key metrics this can be concluded this model has relatively low accuracy0.8297872 and higher error rate 0.1702127 compared to other models.AUC curve for this model is 0.9263.

4.1.6 Model6 Evaluation

Table 13: Model 6 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
6	0.8297872	0.1702128	0.7358491	0.9512195	0.7358491	0.8297872	0.9406351

Looking at the key metrics this can be concluded this model has relatively low accuracy 0.8297872 and higher error rate 0.17021277 compared to other models.AUC curve for this model is 0.930.

4.2 Final Model Seletion

Following is the comparison of various metrics for above 6 models

Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
0.9042553	0.0957447	0.9245283	0.9074074	0.9000000	0.9283174	0.9549011
0.8723404	0.1276596	0.9056604	0.8727273	0.8717949	0.9061444	0.9553613
0.8936170	0.1063830	0.9245283	0.8909091	0.8974359	0.9211541	0.9677865
0.8829787	0.1170213	0.9056604	0.8888889	0.8750000	0.9127916	0.9687069
0.8297872	0.1702128	0.7358491	0.9512195	0.7358491	0.8297872	0.9263691
0.8297872	0.1702128	0.7358491	0.9512195	0.7358491	0.8297872	0.9406351

From the comparison table it can be seen model 1 is the best model with very high accuracy rate of 91.48%.But Model 2 is the best in terms of AUC value which is .9567 and Accuracy 90.42%. Model 1 also has close value of AUC score of 95.44. But Model2 has lower AIC value of 164.85 where as the other one has AIC value 170.93. Both model has same no of coefficient. Based on the above data points model 1 is selected for slightly better curacy. For final model following analysis has been carried out
(i) Relevant variables in the model (ii) Estimate confidence interval for coefficient (iii) odds ratios and 95% CI (iv) AUC curve (v) Distribution of prediction

Most important variables in the model

```
##
## Call:
## glm(formula = target ~ ., family = "binomial", data = city_crime_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8791  -0.1299  -0.0025   0.0011   3.4785
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -41.462153   8.250799  -5.025 5.03e-07 ***
## zn           -0.060580   0.039153  -1.547 0.121799
```

```
## indus        -0.063885   0.059335  -1.077 0.281618
## chas          0.789391   0.865818   0.912 0.361912
## nox          53.413503  10.013666   5.334 9.60e-08 ***
## rm           -0.647942   0.904430  -0.716 0.473739
## age           0.028835   0.015680   1.839 0.065915 .
## dis           0.800917   0.268877   2.979 0.002894 **
## rad           0.721751   0.195662   3.689 0.000225 ***
## tax          -0.007065   0.003490  -2.024 0.042948 *
## ptratio       0.440768   0.159366   2.766 0.005679 **
## black        -0.009591   0.006025  -1.592 0.111412
## lstat         0.096941   0.062429   1.553 0.120469
## medv          0.236940   0.091276   2.596 0.009436 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 514.63  on 371  degrees of freedom
## Residual deviance: 140.71  on 358  degrees of freedom
## AIC: 168.71
##
## Number of Fisher Scoring iterations: 9
```

Following are the most relevant variables for the model- indus,nox,dis,rad,ptratio,medv
we can write the equation as:
log(y)=-41.426+53.41 x nox + 0.80 x dis + 0.721 x rad -0.007 x tax + 0.44 x Ptratio + 0.23 x medv

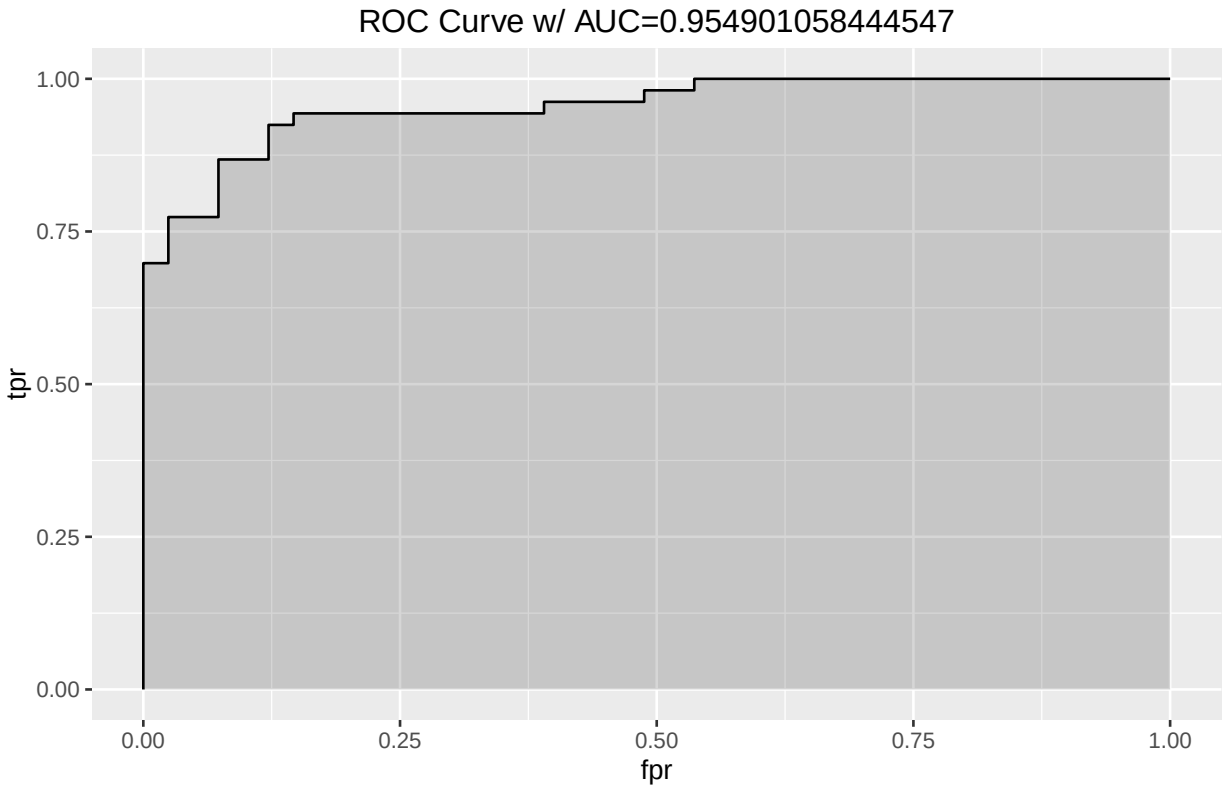
Analysis of odds ratios of variables 95% CI

```
##                       OR        2.5 %       97.5 %
## (Intercept) 9.844998e-19 9.335179e-26 1.038266e-11
## zn          9.412183e-01 8.716922e-01 1.016290e+00
## indus       9.381125e-01 8.351201e-01 1.053807e+00
## chas        2.202054e+00 4.034991e-01 1.201748e+01
## nox         1.574670e+23 4.715650e+14 5.258208e+31
## rm          5.231212e-01 8.886894e-02 3.079319e+00
## age         1.029255e+00 9.981051e-01 1.061378e+00
## dis         2.227583e+00 1.315121e+00 3.773133e+00
## rad         2.058033e+00 1.402507e+00 3.019950e+00
## tax         9.929600e-01 9.861908e-01 9.997758e-01
## ptratio     1.553900e+00 1.137027e+00 2.123615e+00
## black       9.904547e-01 9.788273e-01 1.002220e+00
## lstat       1.101795e+00 9.749020e-01 1.245205e+00
## medv        1.267365e+00 1.059759e+00 1.515641e+00
```

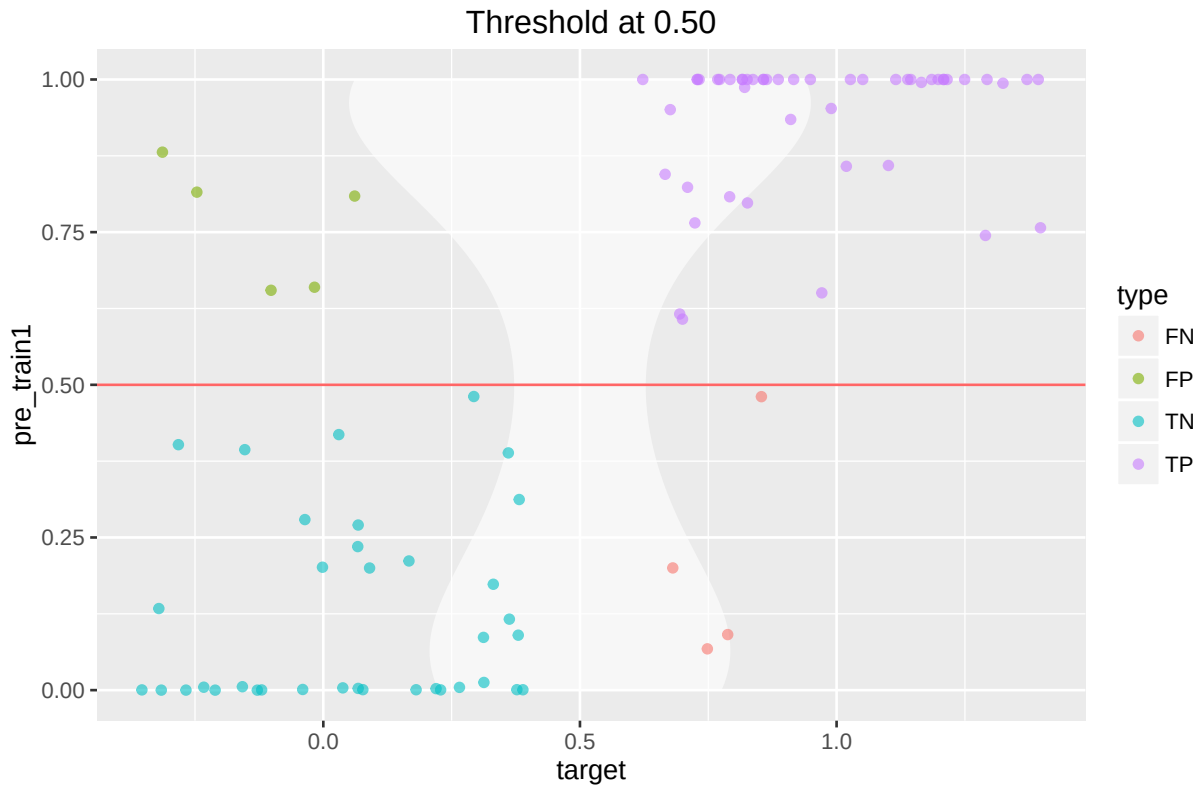
Following points can be interpreted for the above mentioned variables-

(i) nox variable has the most impact in odd ratio, keeping all other variables constant, odds of increase in crime rate above median increases 6.177550e+23 times with per unit change in nox variable. (ii)Keeping all other variables same odds of having crime rate above median value increases following way-0.875 for per unit change in indus,2.50 per unit change in dis,1.74 for per unit change in rad,1.51 for per unit change in ptratio and 1.30 for per unit change in medv. Any value which is less than 1 means less chance of an event with the per unit increase of the variable.

AUC curve for the selected model



Distribution of the Predictions



Considering the target has value 1 (crime above median) and 0 when crime is below median, then the above plot illustrates the trade off that to be made upon choosing a reasonable threshold. If threshold is increased the the number of false positive (FP) results is lowered, while the number of false negative (FN) results increases.

5 Prediction on test data

Table 15: Outcome on evaluation data set

Var1	Freq
FALSE	19
TRUE	21