Homework Assignment - 03

Critical Thinking Group 5
Arindam Barman
Mohamed Elmoudni
Shazia Khan
Kishore Prasad

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)

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.

1.3 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.

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

Table 2: 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

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.

1.4 Outliers and Missing Values Identification

1.4.1 Missing Values

As per Table .3 below, we see that we have no missing values which is good thing as we don't have to carry out any imputation tasks.

Table 3: 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

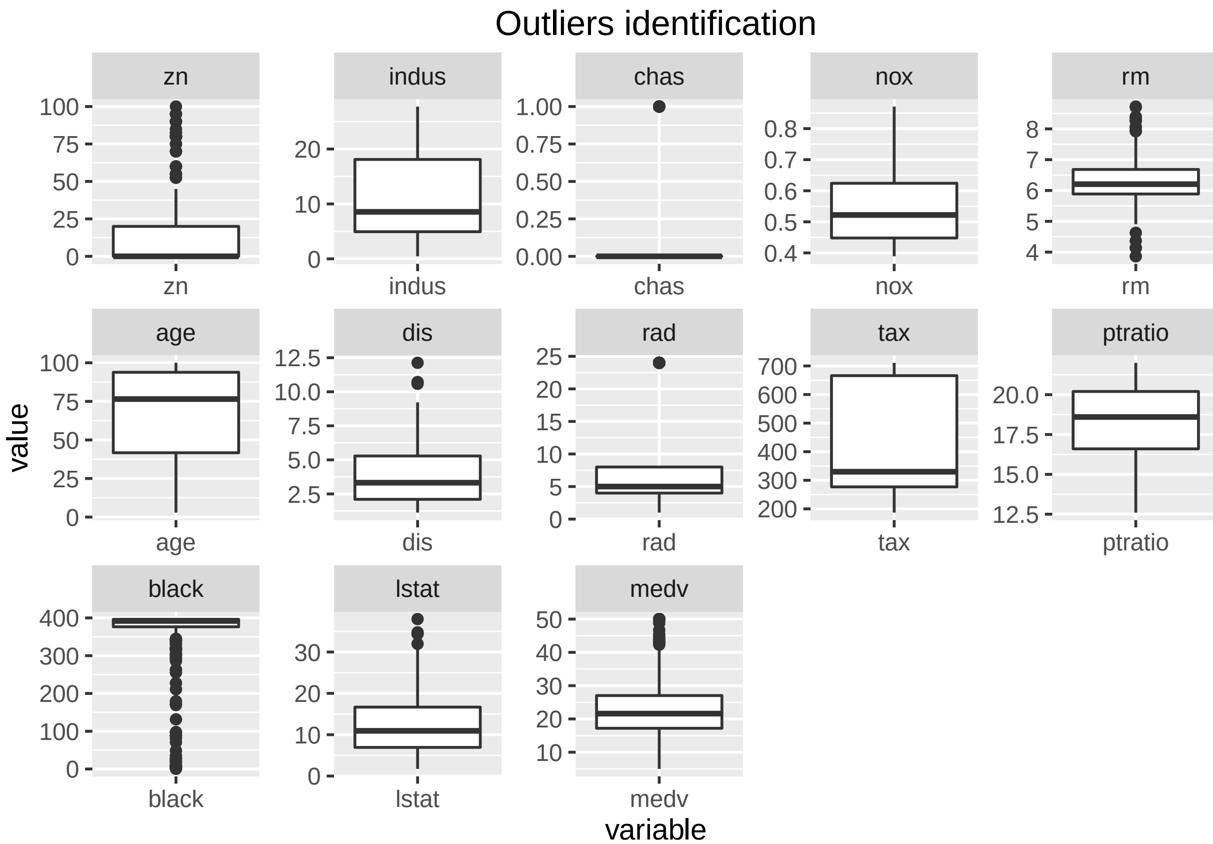
Also, as per Table .4 below, we can confirm that our target variable is binary as expected.

Table 4: 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

1.4.2 Outliers identification

In this section univariate 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



From the “Outliers identification” plot above, we see that we have few outliers that we need to treat. We see that: zn (residential land zoned), rm (average number of rooms per dwelling), dis(weighted mean of distances to five Boston employment centers), black (the proportion of blacks by town),lstat (lower status of the population), and medv median value of owner-occupied homes in $1000s all need to be trated

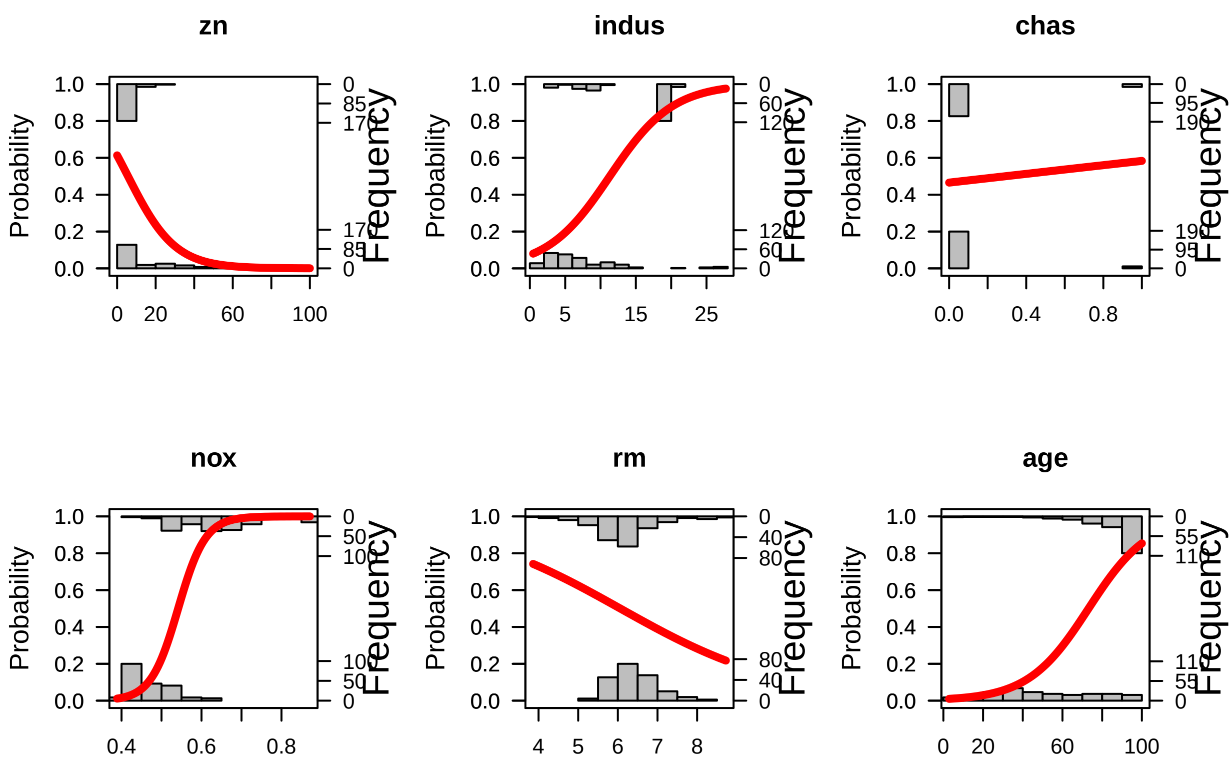
1.4.3 Analysis the link function

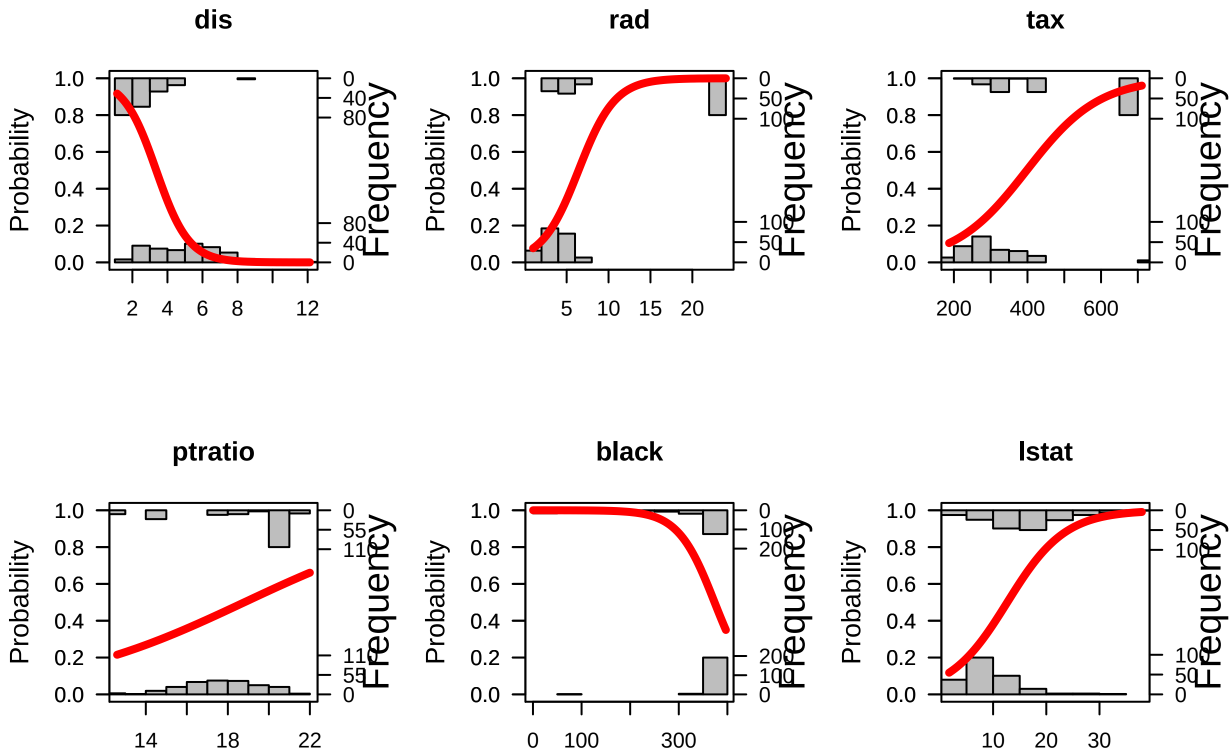
In this section, we will investigate how our initial data aligns with a typical logistic model plot.

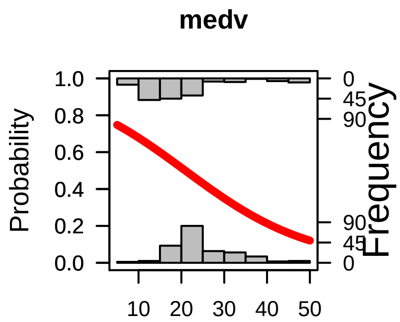
Recall the Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm). The fundamental equation of generalized linear model is: $g(E(y)) = a + Bx_1 + B_2x_2 + B_3x_3 + ...$ where, g() is the link function, E(y) is the expectation of target variable and $B_0 + B_1x_1 + B_2x_2 + B_3x_3$ is the linear predictor (B_0, B_1, B_2, B_3 to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

In logistic regression, we are only concerned about the probability of outcome dependent variable (success or failure). As described above, g() is the link function. This function is established using two things: Probability of Success (p) and Probability of Failure (1-p). p should meet following criteria: It must always be positive (since p >= 0) It must always be less than equals to 1 (since p <= 1).

Now let's investigate how our initial data model aligns with the above criteria. In other words, we will plot regression model plots for each variable and compare it to a typical logistic model plot:




dis
rad
tax
ptratio
black
lstat
Probability
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
12
5
10
15
20
200
400
600
14
18
22
0
100
300
10
20
30



- Interpretation

You can see clearly that the probability of crime being above average increases as we get closer to the "1" classification for the indus,nox,age,rad,tax,and lstat variables. In the middle, the probability changes at the highest rate, while it tails off at each end in order to bound it between 0 and 1.
You can see clearly that the probability of crime being above average decreases as we get closer to the "1" classification for the zn, dis,black, and mdev variables. In the middle, the probability changes at the lowest rate. However, it does not tails off at each end for all of the variables.

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 following the below steps as guidelines:
- Outliers treatment
- Missing values treatment
- Data transformation

2.1 Outliers treatment

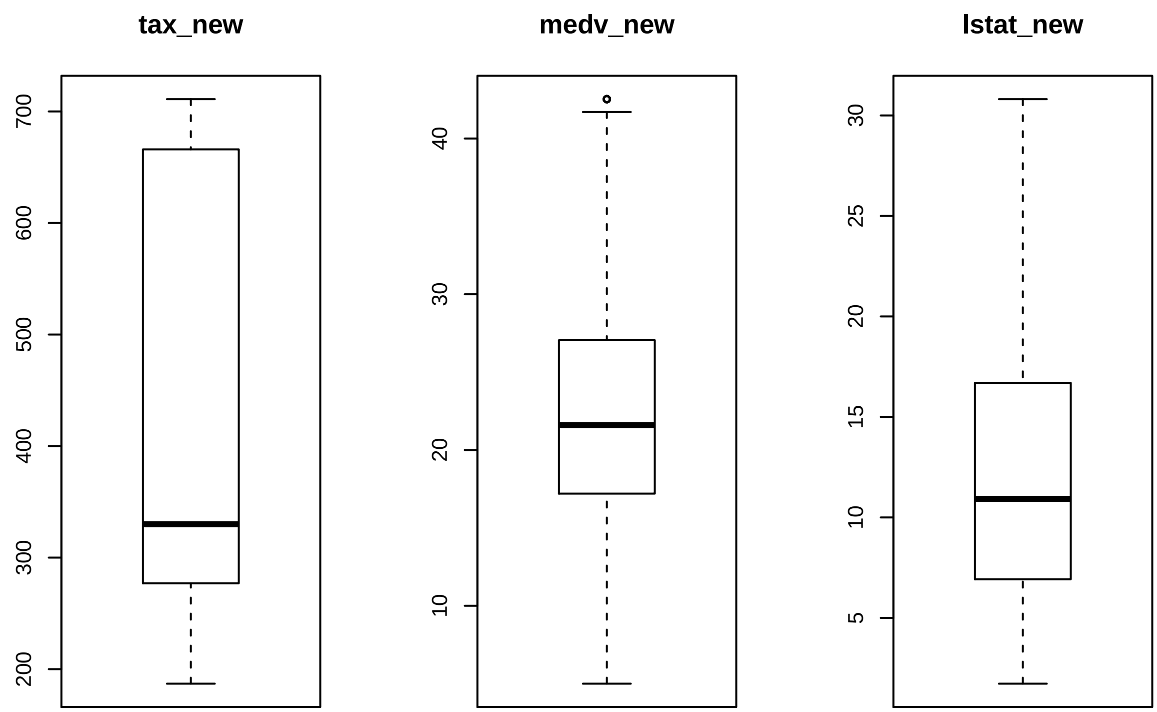
For outliers, we will create 2 sets of variables.

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_train$tax\ city_crime_train$medv
city_crime_train$lstat

Lets see how the new variables look in boxplots.



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

city_crime_train_mod$rm_new\ city_crime_train_mod$dis_new

2.3 Tranformation for Variables

In this section, we will analyze few transformation options using Sin, Log, Sqrt, nth transformations. Using histogram and boxplots to evaluate best transformation to handle outliers.
First we will start with variable, zn (proportion of residential land zoned for large lots) as per below

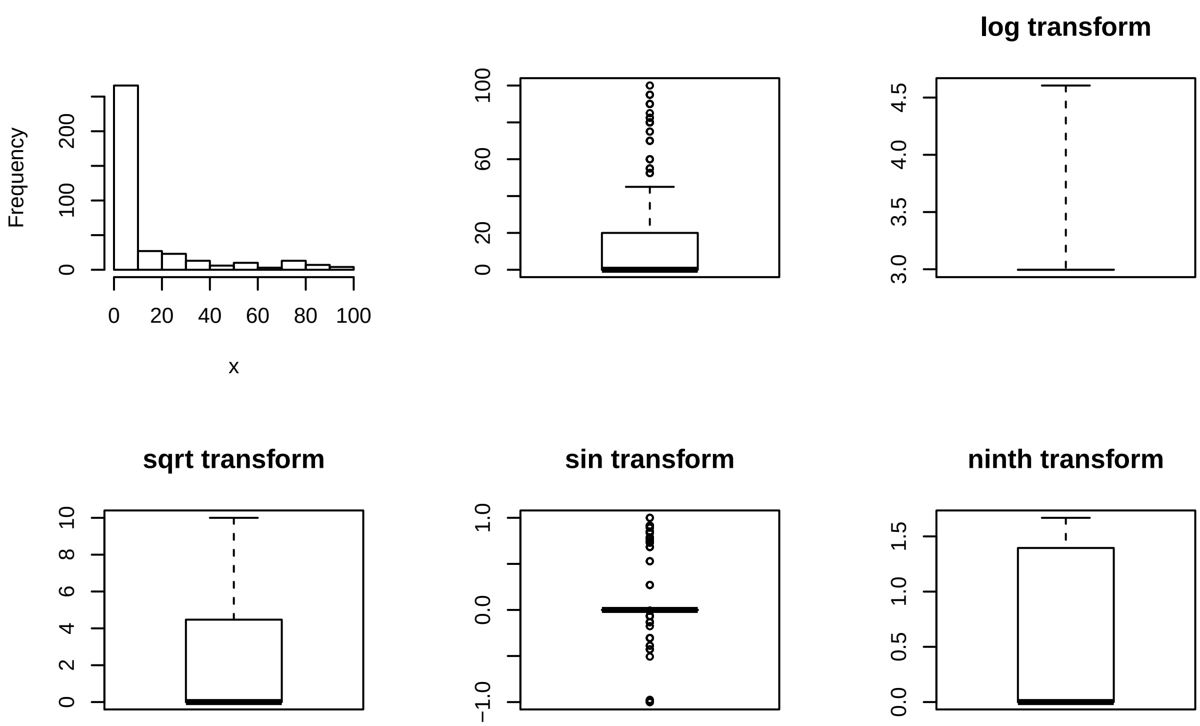


Figure 1: zn Transformation

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.

2.3.1 Variable Transformation Interpretation

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

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

\\ Following variables will need some transformation:

1. zn: proportion of residential land zoned for large lots (over 25000 square feet) (predictor variable)
2. chas: a dummy var. for whether the suburb borders the Charles River (1) or not (0) (predictor variable)
3. target: whether the crime rate is above the median crime rate (1) or not (0) (response variable)

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

3 Build Models

In this section, we will create six models. Aside from using original and transformed data, we will also using different methods and functions such as Linear Discriminant Analysis, step function, and logit function to enhance our models.

Below is our model definition:
-Model 1- This model will be created using the original variables in train data set with logit function GLM.
-Model 2- This model will be created using original variables; however using step function instead of GLM.
-Model 3- This model will be created using transformed variables using GLM function.
-Model 4- This model will be created using transformed variables using step function instead of GLM.
-Model 5: this model will be created using original variables using Linear Discriminant Analysis function lda in ISLR package.
-Model 6- This model will be created using transformed variables using Linear Discriminant Analysis

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

Table 5: 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

3.1.1 Model One

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 to the target variable. Other important variables are dis, rad, tax, ptratio, and medv. The 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 Two

This model, we will be using original variables; however using step function (backward process) instead of GLM.

```
##
## 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, and 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, and lstat. The 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

In this model, we will be using transformed variables 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
## chas          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 that the following variables are relevant for this model: nox, dis, rad, ptratio, tax_new, medv_new, and lstat_new.

(ii) The number of integration is 9 and AIC value =169.71.

(iii) nox and rad are the two most important variables. The 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 Four

In this model we will be using transformed variables using backward step function instead of GLM

```
##
## 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 that the following variables are relevant for this model: nox, dis, rad, ptratio , tax_new, medv_new,and lstat_new
(ii) The 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 Five

In this model we will be using original variables; however the Linear Discriminant Analysis function lda in ISLR package.

```
## 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 six

In this model we be using transformed variables using Linear Discriminant Analysis

```
## 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       chas       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
## chas      -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 also follow the below steps to select our final model:
-Model selection strategy
-Model1 Evaluation
-Final Model Selection
-Inference of Final Model

4.1 Model selection strategy:

The following model selection strategy has been used for this assignment:

(i) Compare accuracy & confusion matrix of the models.
(ii) Compare Precision, Sensitivity, Specificity, and 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 Model One Evaluation

Table 6: 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 Model Two Evaluation

Table 7: 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 Model Three Evaluation

Table 8: 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 Model Four Evaluation

Table 9: 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 Model Five Evaluation

Table 10: 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 Model Six Evaluation

Table 11: 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

Model_No	Accuracy	Error_Rate	AUC	Precision	sensitivity	specificity	F1_Score
1	0.9042553	0.0957447	0.9549011	0.9245283	0.9074074	0.9000000	0.9283174
2	0.8723404	0.1276596	0.9553613	0.9056604	0.8727273	0.8717949	0.9061444
3	0.8936170	0.1063830	0.9677865	0.9245283	0.8909091	0.8974359	0.9211541
4	0.8829787	0.1170213	0.9687069	0.9056604	0.8888889	0.8750000	0.9127916
5	0.8297872	0.1702128	0.9263691	0.7358491	0.9512195	0.7358491	0.8297872
6	0.8297872	0.1702128	0.9406351	0.7358491	0.9512195	0.7358491	0.8297872

From the comparison table it can be seen model 1 and mode 3 are very close. Model 1 has slightly better accuracy rate 90.42% compare to 89.36%.But Model 3 is the best in terms of AUC value which is .9677 but close to model 1 value. AUC provides the score best on probability correctly identifying the patterns at various cut off values and Accuracy on the other hand calculated as specific cut off value. For this assignment we will go with cut off value of 0.5 and choose the Model 1 based on Accuracy value for further prediction on evaluation data set.

4.2.1 Inference for Final Model

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

4.2.2 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 of the Model 1 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

4.2.3 Analysis of odds ratios of variables 95% CI

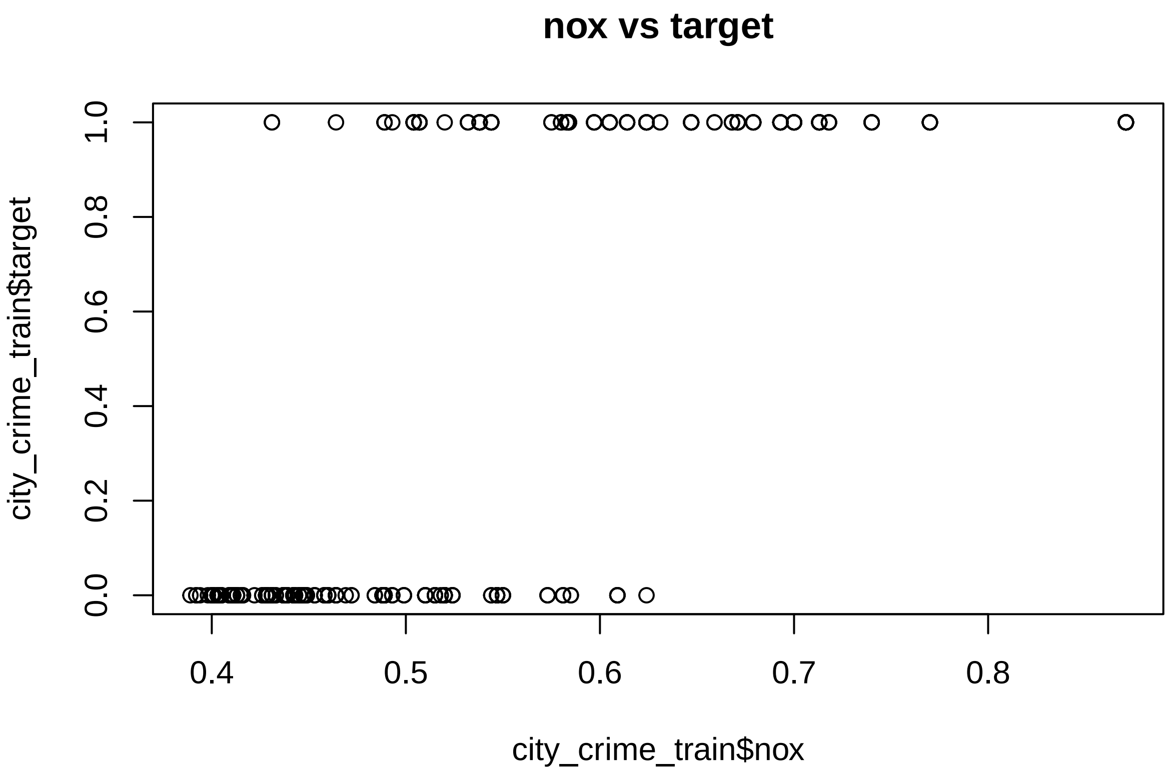
```
##                       OR        2.5 %       97.5 %
## (Intercept) 9.844998e-19 9.335179e-26 1.038266e-11
## zn          9.412183e-01 8.716922e-01 1.016290e+00
## indus       9.381125e-01 8.351201e-01 1.053807e+00
## chas        2.202054e+00 4.034991e-01 1.201748e+01
## nox         1.574670e+23 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 important variables in the model-

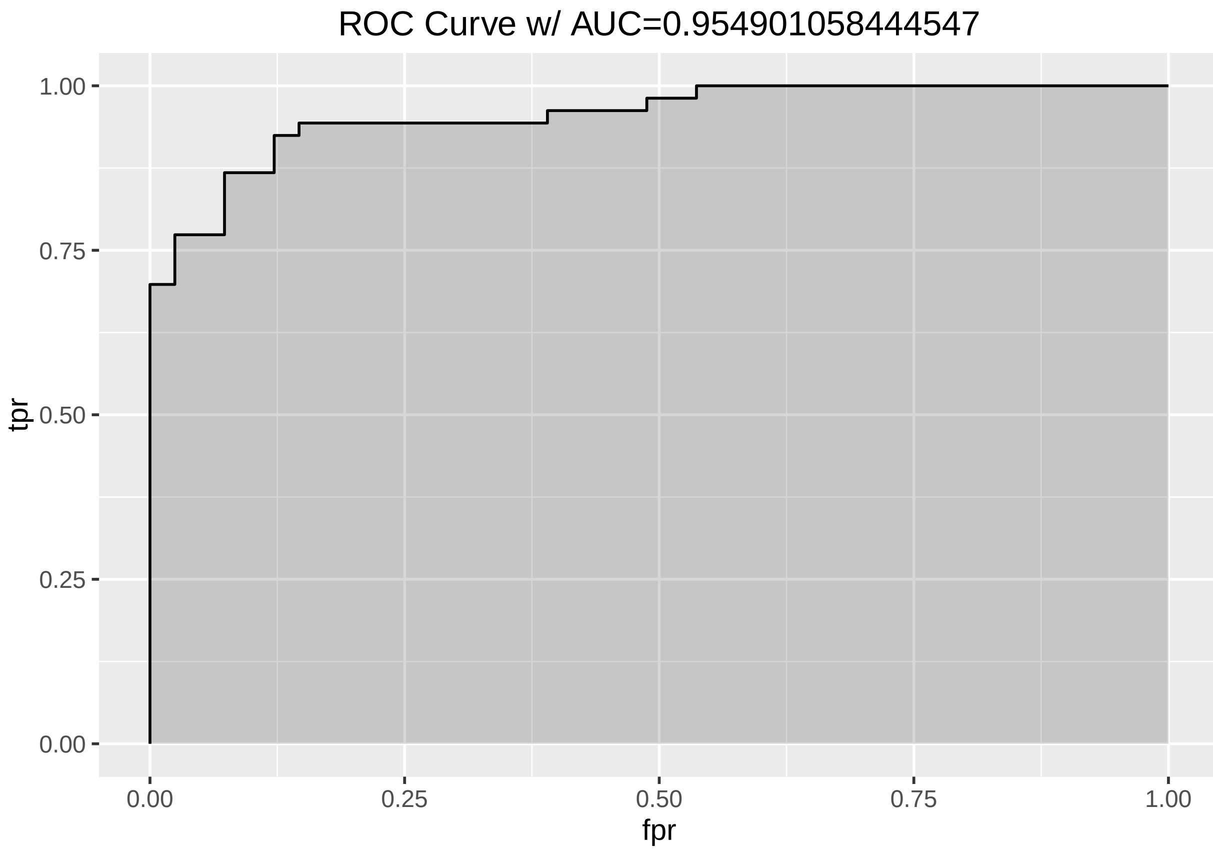
(i)Keeping all other variables same odds of having crime rate above median value increases following way -0.875 for per unit change in indus,2.50 per unit change in dis,1.74 for per unit change in rad,1.51 for per unit change in ptratio and 1.30 for per unit change in medv. Any value which is less than 1 means less chance of an event with the per unit increase of the variable.

(ii) nox variable has very high oddd ratio and reason is, in given data set there are records where for range of nox values there is only one outcome of target (either 0 or 1) as shown in the chrt below. That makes easy for prediction of target variable and increases the accuracy of the model. Outlier treatment for this variable will result in 50% reduction on train data set. For this assignment we have kept this variable.

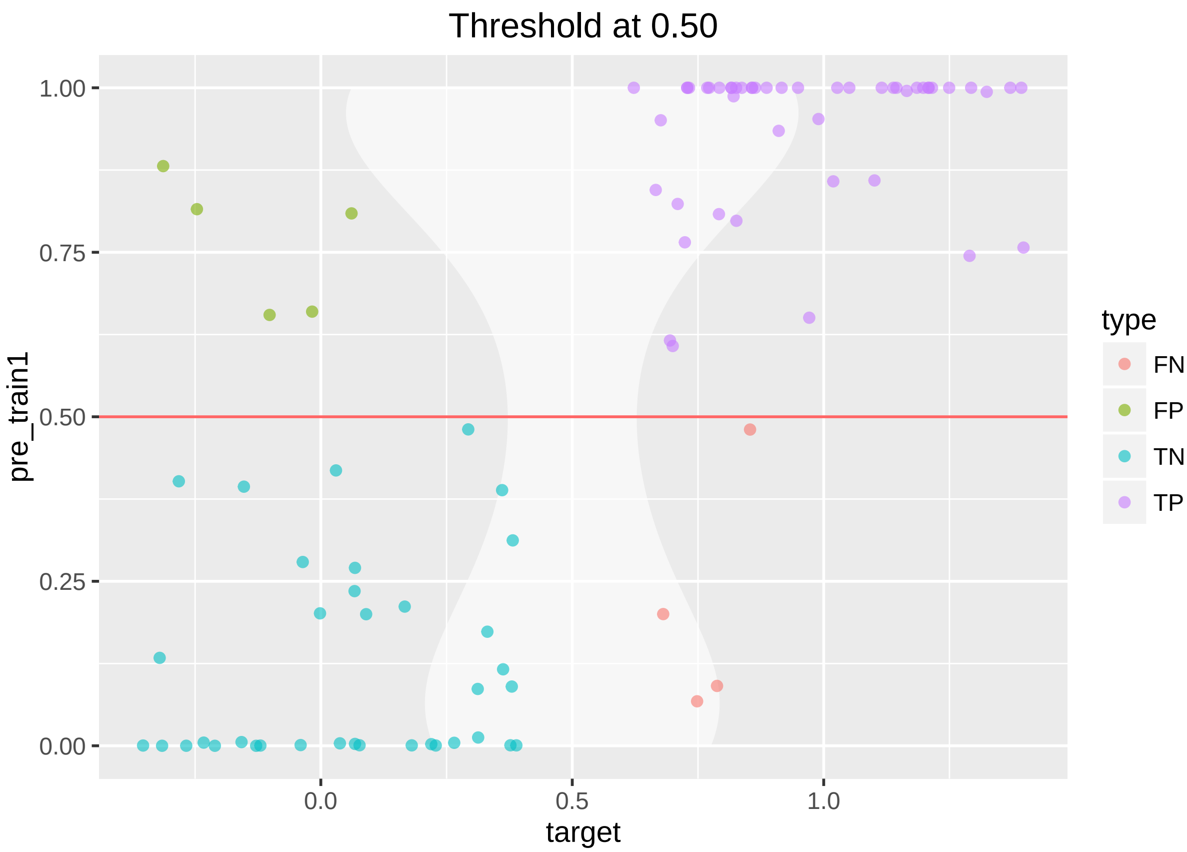
Note: There are approaches to handle such scenario like form of penalized regression. For this assignment purpose we have not perused further into any other model.



4.2.4 AUC curve for the selected model



4.2.5 Distribution of the Predictions



Considering the target has value 1 (crime above median) and 0 when crime rate 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 using final model on evaluation data set

In this section Model 1 has been used predict the outcome on evaluation dataset.

Table 13: Outcome on evaluation data set

Var1	Freq
FALSE	19
TRUE	21

Based on the outcome we can conclude in evaluation data set of 40, around 21 records are there where crime rate is above median and 19 records where crime rate is below median.

Appendix A: DATA621 Homework 03 R Code

```
#code=readLines(knitr::purl('https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW1/HW3_Final
```