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)

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.

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.

1.3 Outliers and Missing Values Identification

1.3.1 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

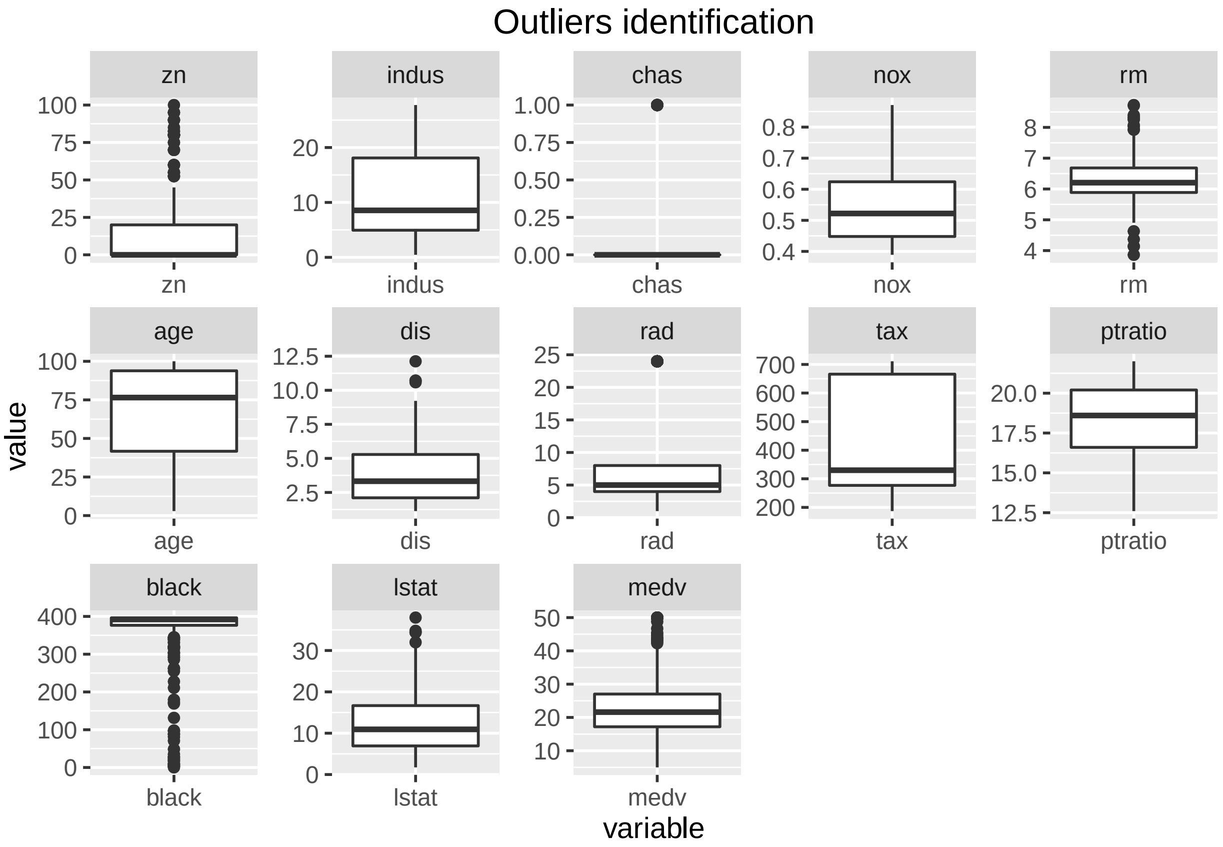
Also, as per Table .4 below, we can confirm that our target variable is binary as expected.

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

1.3.2 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

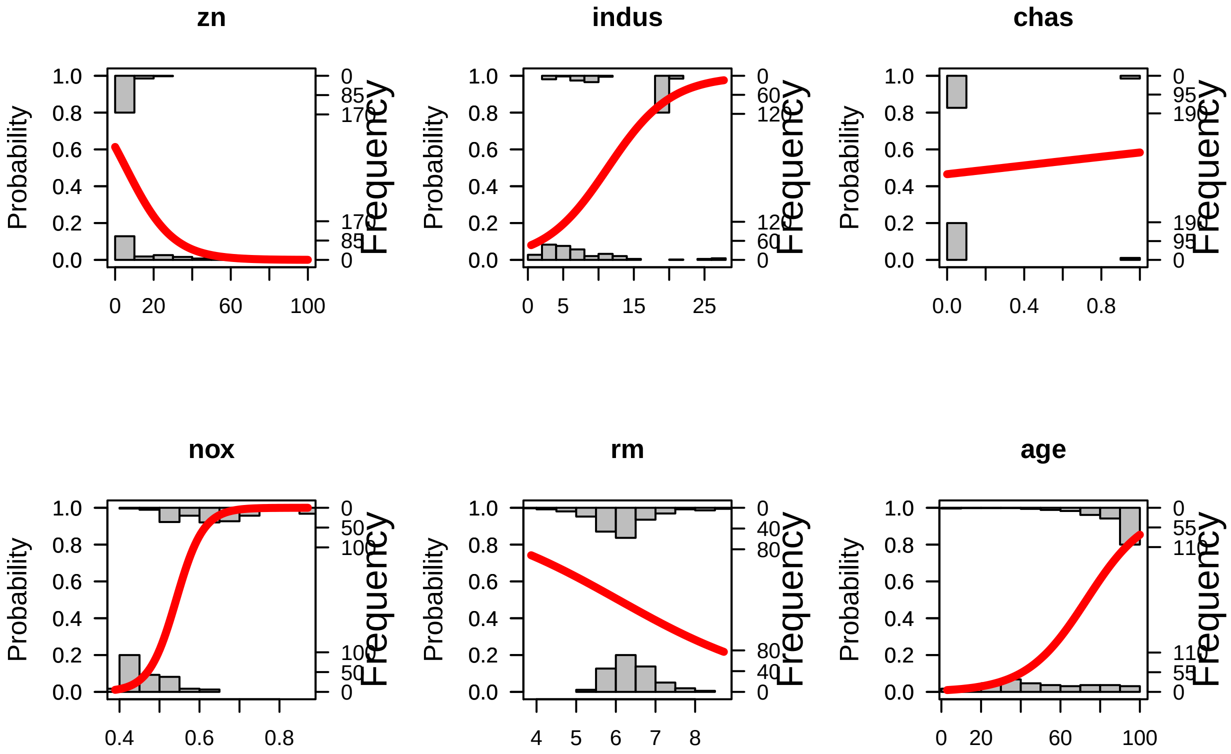
1.3.3 Analysis the link function

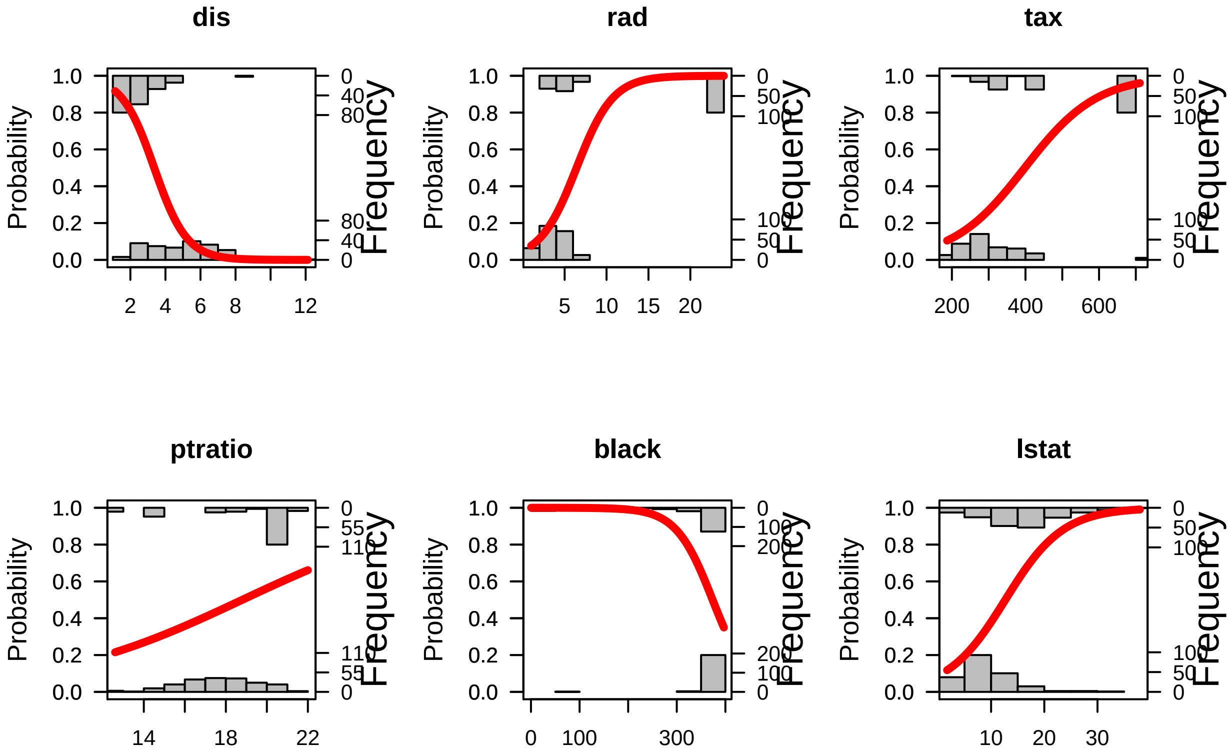
In this section, we will investigate how our initial data aligns with a typical logistic model plot.

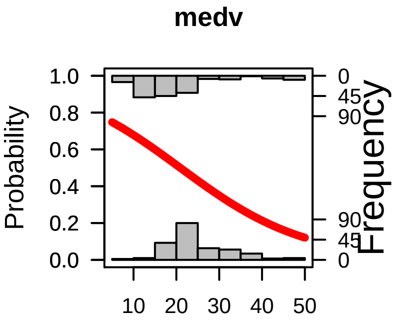
Recall the Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm). The fundamental equation of generalized linear model is: $g(E(y)) = a + Bx_1 + B_2x_2 + B_3x_3 + ...$ where, g() is the link function, E(y) is the expectation of target variable and $B_0 + B_1x_1 + B_2x_2 + B_3x_3$ 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
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
40
80
80
40
0
Frequency
2
4
6
8
12
rad
Probability
0
50
100
100
50
0
Frequency
5
10
15
20
tax
Probability
0
50
100
100
50
0
Frequency
200
400
600
ptratio
Probability
0
55
110
110
55
0
Frequency
14
18
22
black
Probability
0
100
200
200
100
0
Frequency
0
100
300
lstat
Probability
0
50
100
100
50
0
Frequency
10
20
30



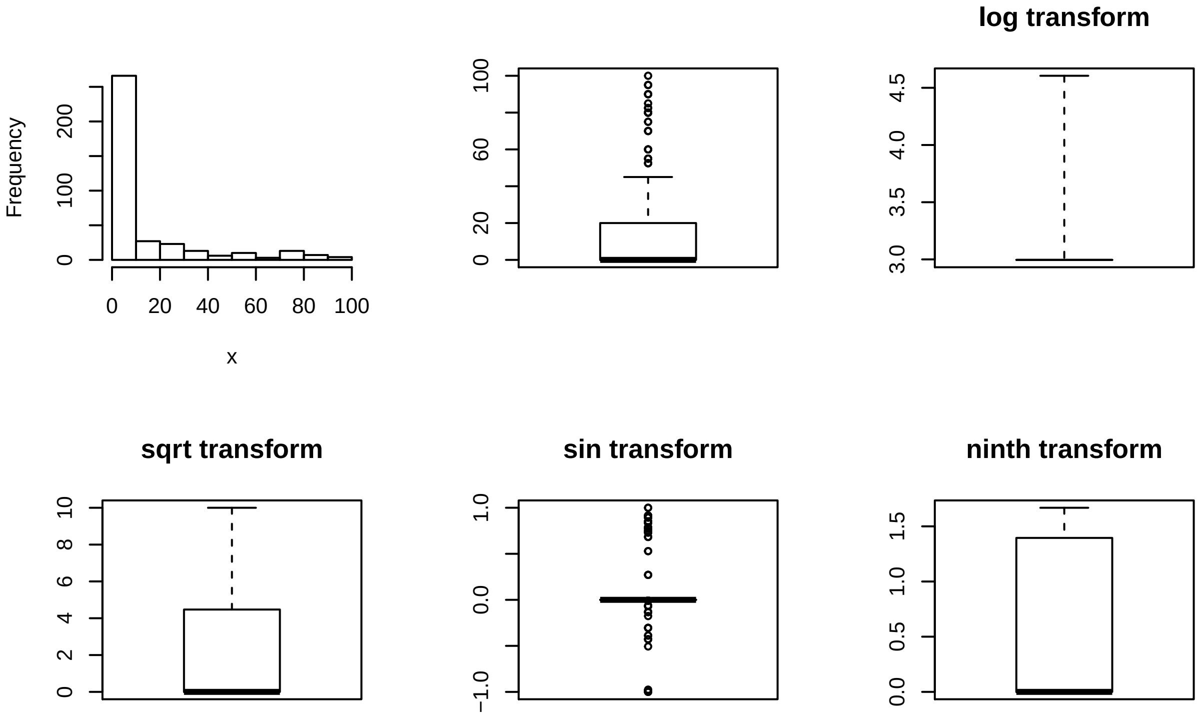
Interpretation

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.

****

In this section univariate analysis is being caarried 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.

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 puring the belwo steps as guidlines:

- Outliers treatment
- Missing values treatment
- Data transformation

2.1 Outliers treatment

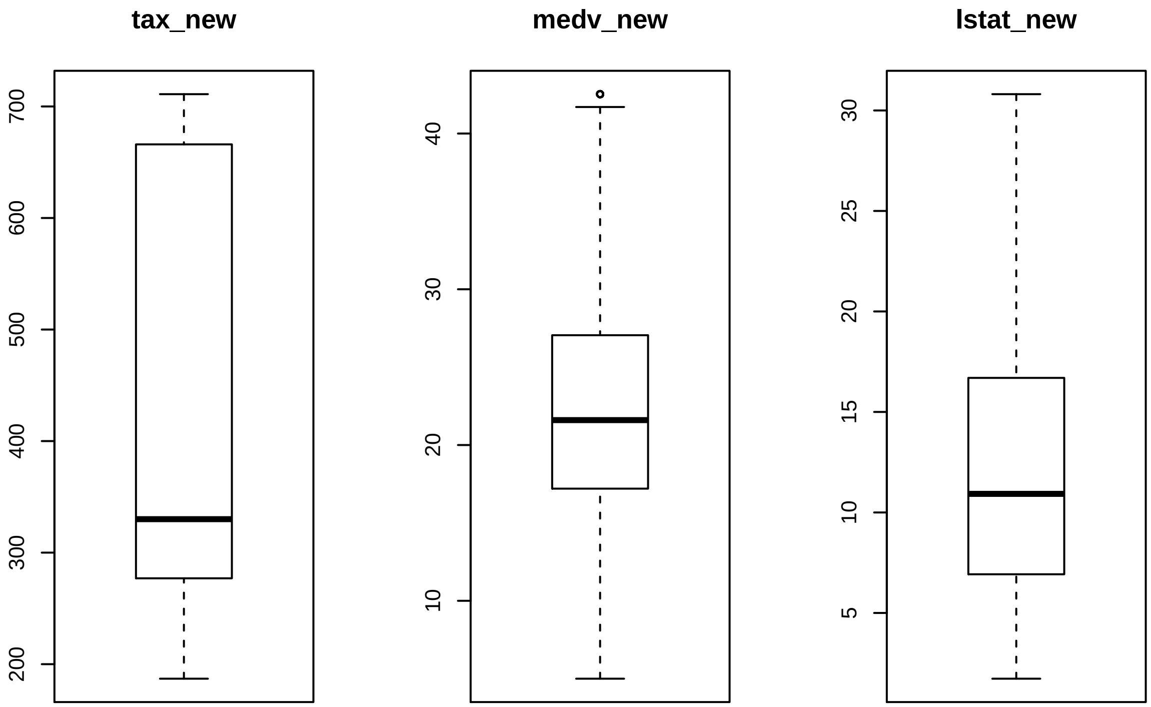
For outliers, we will create 2 sets of variables.

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$tax\ city_{c}rime_{t}rain$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$rm_{n}ew\ city_{c}rime_{t}rain_{m}od$dis_new

2.3 Tranformation for Variables

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

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

VARIABLE_NAME	Comments	Theoretical.Effect	Model1	Model2	Model3	Model4
TEAM_BATTING_H	Given	Positive	Y			Y
TEAM_BATTING_2B	Given	Positive	Y			Y
TEAM_BATTING_3B	Given	Positive	Y			Y
TEAM_BATTING_HR	Given	Positive	Y			Y
TEAM_BATTING_BB	Given	Positive	Y			Y
TEAM_BATTING_HBP	Given	Positive				
TEAM_BATTING_SO	Given	Negative	Y			Y
TEAM_BASERUN_SB	Given	Positive	Y			Y
TEAM_BASERUN_CS	Given	Negative				
TEAM_FIELDING_E	Given	Negative	Y			Y
TEAM_FIELDING_DP	Given	Positive	Y			Y
TEAM_PITCHING_BB	Given	Negative	Y			Y
TEAM_PITCHING_H	Given	Negative	Y			Y
TEAM_PITCHING_HR	Given	Negative	Y			Y
TEAM_PITCHING_SO	Given	Positive	Y			Y
TEAM_BATTING_H_NEW	Derived	Positive		Y		Y
TEAM_BATTING_2B_NEW	Derived	Positive		Y		Y
TEAM_BATTING_3B_NEW	Derived	Positive		Y		Y
TEAM_BATTING_BB_NEW	Derived	Positive		Y		Y
TEAM_BASERUN_SB_NEW	Derived	Positive		Y		Y
TEAM_FIELDING_E_NEW	Derived	Negative		Y		Y
TEAM_FIELDING_DP_NEW	Derived	Positive		Y		Y
TEAM_PITCHING_BB_NEW	Derived	Negative		Y		Y
TEAM_PITCHING_H_NEW	Derived	Negative		Y		Y
TEAM_PITCHING_HR_NEW	Derived	Negative		Y		Y
TEAM_PITCHING_SO_NEW	Derived	Positive		Y		Y
TEAM_BATTING_H_SIN	Derived	Positive			Y	Y
TEAM_BATTING_2B_SIN	Derived	Positive			Y	Y
TEAM_BATTING_3B_SIN	Derived	Positive			Y	Y
TEAM_BATTING_BB_SIN	Derived	Positive			Y	Y
TEAM_BASERUN_SB_SIN	Derived	Positive			Y	Y
TEAM_FIELDING_E_SIN	Derived	Negative			Y	Y
TEAM_FIELDING_DP_SIN	Derived	Positive			Y	Y
TEAM_PITCHING_BB_SIN	Derived	Negative			Y	Y
TEAM_PITCHING_H_SIN	Derived	Negative			Y	Y
TEAM_PITCHING_HR_SIN	Derived	Negative			Y	Y
TEAM_PITCHING_SO_SIN	Derived	Positive			Y	Y
TEAM_BATTING_HBP_Missing	Derived				Y	Y
TEAM_BASERUN_CS_Missing	Derived				Y	Y
Hits_R	Derived				Y	Y
Walks_R	Derived				Y	Y
HomeRuns_R	Derived				Y	Y
Strikeout_R	Derived				Y	Y
TEAM_BATTING_EB	Derived				Y	Y
TEAM_BATTING_1B	Derived				Y	Y

3.1.1 Model One by using all given variable

In this model, we will be using the original variables. We will create model and we will highlight the variables that being recommended using the AIC value.
First we will produce the summary model as per below:

```
##
## 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
```

Accuracy=0.9042553

3.1.2 Model two- with backward step function with all given variables

```
stepmodel1<- step(model1, direction="backward")
```

```
## Start:  AIC=168.71
## target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + black + lstat + medv
##
```

```
##           Df Deviance    AIC
## - rm       1   141.22 167.22
## - chas     1   141.55 167.55
## - indus    1   141.93 167.93
## <none>         140.71 168.71
## - lstat    1   143.06 169.06
## - black    1   143.68 169.68
## - zn       1   143.99 169.99
## - age      1   144.45 170.45
## - tax      1   144.93 170.93
## - medv     1   148.67 174.67
## - ptratio  1   149.29 175.29
## - dis      1   150.97 176.97
## - rad      1   171.94 197.94
## - nox      1   195.65 221.65
##
## Step:  AIC=167.22
## target ~ zn + indus + chas + nox + age + dis + rad + tax + ptratio +
##     black + lstat + medv
##
##           Df Deviance    AIC
## - chas     1   142.10 166.10
## - indus    1   142.37 166.37
## <none>         141.22 167.22
## - black    1   144.02 168.02
## - age      1   144.48 168.48
## - zn       1   144.74 168.74
## - lstat    1   145.13 169.13
## - tax      1   145.97 169.97
## - ptratio  1   149.78 173.78
## - dis      1   150.97 174.97
## - medv     1   156.73 180.73
## - rad      1   172.26 196.26
## - nox      1   196.29 220.29
##
## Step:  AIC=166.1
## target ~ zn + indus + nox + age + dis + rad + tax + ptratio +
##     black + lstat + medv
##
##           Df Deviance    AIC
## - indus    1   142.85 164.85
## <none>         142.10 166.10
## - black    1   144.69 166.69
## - age      1   145.65 167.65
## - zn       1   146.09 168.09
## - lstat    1   146.43 168.43
## - tax      1   148.34 170.34
## - ptratio  1   149.90 171.90
## - dis      1   151.42 173.42
## - medv     1   157.16 179.16
## - rad      1   177.68 199.68
## - nox      1   196.44 218.44
##
## Step:  AIC=164.85
## target ~ zn + nox + age + dis + rad + tax + ptratio + black +
##     lstat + medv
##
##           Df Deviance    AIC
## <none>        142.85 164.85
## - black    1   145.21 165.21
## - age      1   146.69 166.69
## - lstat    1   146.75 166.75
## - zn       1   146.89 166.89
## - ptratio  1   150.46 170.46
## - dis      1   151.87 171.87
## - tax      1   154.08 174.08
## - medv     1   157.59 177.59
## - rad      1   184.71 204.71
## - nox      1   203.12 223.12
```

```
pre_train1_step<-predict(stepmodel1,type="response",newdata=train_test)

table(pre_train1_step>0.5,train_test$target)
```

```
##
##          0  1
##   FALSE 34  5
##   TRUE   7 48
```

Accuracy=0.8723404

3.1.3 Model three- model with transformed variables

In this model, we will be using the some transformed variables.

First we will produce the summary model as per below:

```
##
## Call:
## glm(formula = target ~ . - zn - rm - dis - tax - lstat - medv,
##     family = "binomial", data = city_crime_train_mod)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8490  -0.1466  -0.0024   0.0004   3.5826
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -45.541738   8.544894  -5.330 9.84e-08 ***
## indus         0.014531   0.064926   0.224 0.822909
## chas          0.108863   0.811295   0.134 0.893257
## nox          50.472586   9.083435   5.557 2.75e-08 ***
## age           0.036435   0.016117   2.261 0.023780 *
## rad           0.871309   0.241452   3.609 0.000308 ***
## ptratio       0.495086   0.172513   2.870 0.004107 **
## black        -0.010433   0.005881  -1.774 0.076036 .
## tax_new      -0.005498   0.003495  -1.573 0.115648
## medv_new      0.297542   0.090676   3.281 0.001033 **
## lstat_new     0.053168   0.069612   0.764 0.444998
## rm_new       -1.774497   1.144107  -1.551 0.120904
## dis_new      -2.191201   0.532281  -4.117 3.84e-05 ***
## zn_new        0.465684   0.892871   0.522 0.601978
## ---
## 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: 129.52  on 358  degrees of freedom
## AIC: 157.52
##
## Number of Fisher Scoring iterations: 9
```

```
##
##          0  1
##   FALSE 35  3
##   TRUE   6 50
```

Accuracy=0.9042553

3.1.4 Model with transformed variable and with with backward step function

```
stepmodel2<- step(model2, direction="backward")
```

```
## Start:  AIC=157.52
## target ~ (zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + black + lstat + medv + tax_new + medv_new + lstat_new +
##     rm_new + dis_new + zn_new) - zn - rm - dis - tax - lstat -
##     medv
##
##             Df Deviance    AIC
## - chas       1   129.54 155.54
## - indus      1   129.57 155.57
## - zn_new     1   129.79 155.79
## - lstat_new  1   130.08 156.08
## <none>           129.52 157.52
## - tax_new    1   131.92 157.92
## - rm_new     1   131.97 157.97
## - black      1   132.86 158.86
## - age        1   135.31 161.31
## - ptratio    1   138.64 164.64
## - medv_new   1   142.81 168.81
## - dis_new    1   151.54 177.54
## - rad        1   155.24 181.24
## - nox        1   197.04 223.04
##
## Step:  AIC=155.54
## target ~ indus + nox + age + rad + ptratio + black + tax_new +
##     medv_new + lstat_new + rm_new + dis_new + zn_new
##
##             Df Deviance    AIC
## - indus      1   129.61 153.61
## - zn_new     1   129.79 153.79
## - lstat_new  1   130.13 154.13
## <none>           129.54 155.54
## - rm_new     1   131.99 155.99
## - tax_new    1   132.13 156.13
## - black      1   132.86 156.86
## - age        1   135.51 159.51
## - ptratio    1   138.79 162.79
## - medv_new   1   142.84 166.84
## - dis_new    1   152.03 176.03
## - rad        1   156.60 180.60
## - nox        1   197.61 221.61
##
## Step:  AIC=153.61
## target ~ nox + age + rad + ptratio + black + tax_new + medv_new +
##     lstat_new + rm_new + dis_new + zn_new
##
##             Df Deviance    AIC
## - zn_new     1   129.82 151.82
## - lstat_new  1   130.28 152.28
## <none>           129.61 153.61
## - rm_new     1   132.04 154.04
## - tax_new    1   132.51 154.51
## - black      1   132.99 154.99
## - age        1   135.51 157.51
## - ptratio    1   138.80 160.80
## - medv_new   1   143.10 165.10
## - dis_new    1   152.60 174.60
## - rad        1   161.77 183.77
## - nox        1   209.86 231.86
##
## Step:  AIC=151.82
## target ~ nox + age + rad + ptratio + black + tax_new + medv_new +
##     lstat_new + rm_new + dis_new
##
##             Df Deviance    AIC
## - lstat_new  1   130.87 150.87
## <none>           129.82 151.82
## - rm_new     1   132.04 152.04
## - tax_new    1   132.69 152.69
## - black      1   133.06 153.06
## - age        1   135.52 155.52
## - ptratio    1   139.74 159.74
## - medv_new   1   143.10 163.10
## - dis_new    1   152.65 172.65
## - rad        1   162.06 182.06
## - nox        1   212.46 232.46
##
## Step:  AIC=150.86
## target ~ nox + age + rad + ptratio + black + tax_new + medv_new +
##     rm_new + dis_new
##
##             Df Deviance    AIC
## <none>          130.87 150.87
## - tax_new   1   133.34 151.34
## - black     1   133.89 151.89
## - rm_new    1   135.44 153.44
## - age       1   139.74 157.74
## - ptratio   1   141.03 159.03
## - medv_new  1   143.94 161.94
## - dis_new   1   154.34 172.34
## - rad       1   163.53 181.53
## - nox       1   213.91 231.91
```

```
pre_train2_step<-predict(stepmodel2,type="response",newdata=train_test_mod)

table(pre_train2_step>0.5,train_test_mod$target)
```

```
##
##          0  1
##   FALSE 35  4
##   TRUE   6 49
```

Accuracy= 0.893617

3.1,5 Model three with Linear discrement analysis

```
##    class  posterior.0 posterior.1        LD1
## 3      1 0.0005609314  0.99943907  2.9179352
## 6      0 0.8593842086  0.14061579 -0.5664873
## 7      1 0.0040700562  0.99592994  2.1737359
## 8      1 0.0014576826  0.99854232  2.5596162
## 23     0 0.9672384727  0.03276153 -1.1568765
```

```
##
##      0  1
##   0 39 14
##   1  2 39
```

Accuracy=0.8297872

3.1.6 Model with Linear discrement analysis with transformed data

```
##
##      0  1
##   0 39 14
##   1  2 39
```

Accuracy=0.7978723

4 Model Selection

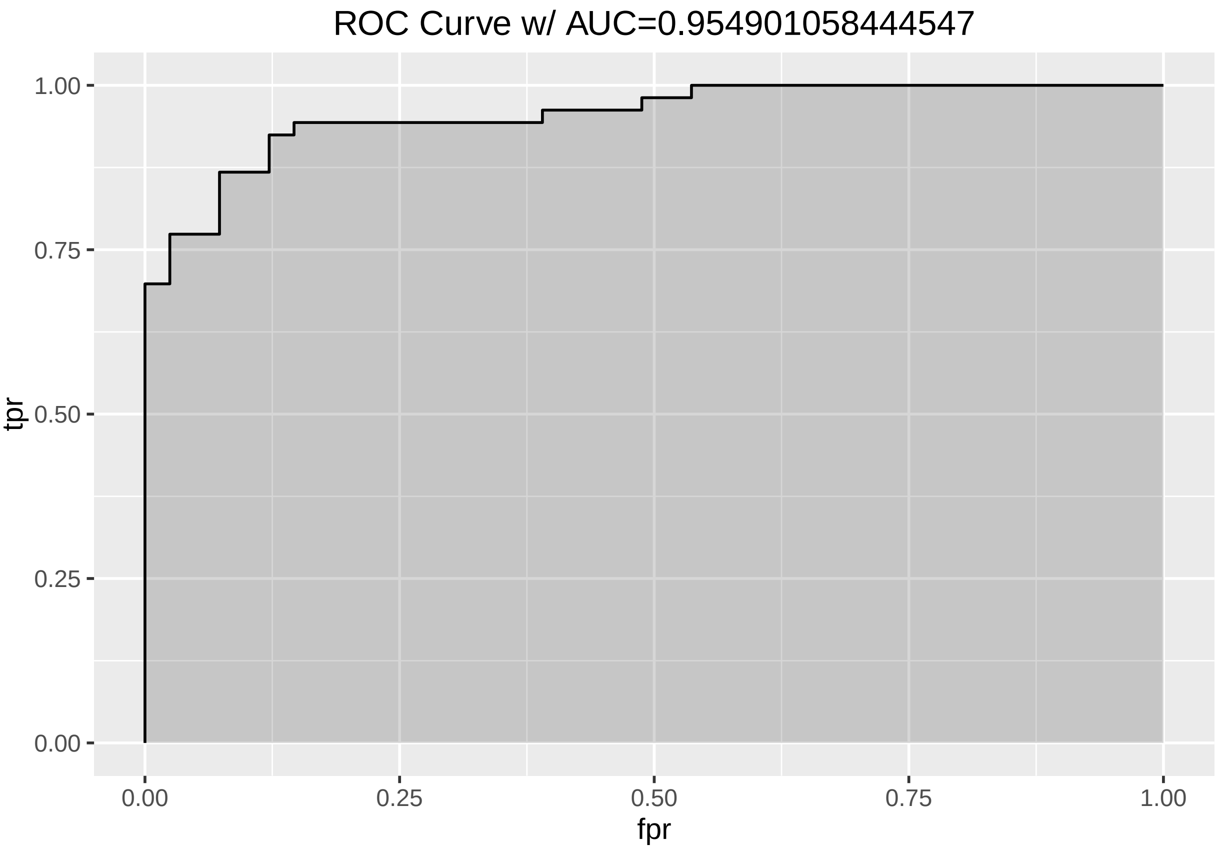
In section we will further examine all six models. We will apply a model selection strategy defined below to compare the models.

4.1 Model selection strategy:

Following model selection strategy has been used for this assignment:

(1) Compare accuracy of the models & confusion matrix
(2) Compare Precision,Sensitivity,Specificity,F1 score
(3) Compare AUC curve for the models

```
##
##          0  1
##   FALSE 36  4
##   TRUE   5 49
```


ROC Curve w/ AUC=0.954901058444547
tpr
fpr