R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
# Logistic Regression & Linear Discriminant Analysis
# Reading a CSV file
BC <- read.csv(file = "C:/Venu/UCI DataSets/Breast-cancer.csv", header = TRUE
,stringsAsFactors = TRUE)

# Data Cleansing
sum(is.na(BC))
```

```
## [1] 9
```

```
BCdata <- na.omit(BC)
levels(BCdata$Class)
```

```
## [1] "no-recurrence-events" "recurrence-events"
```

```
levels(BCdata$Class)[1]<-"0"
levels(BCdata$Class)[2]<-"1"
str(BCdata)
```

```
## 'data.frame':    277 obs. of  10 variables:
##  $ Class      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age        : Factor w/ 6 levels "20-29","30-39",..: 2 3 3 5 3 5 4 5 3 3
...
##  $ Menopause  : Factor w/ 3 levels "ge40","lt40",..: 3 3 3 1 3 1 3 1 3 3 .
..
##  $ Tumor.size : Factor w/ 11 levels "0-4","10-14",..: 6 4 4 3 1 3 5 4 11 4
...
##  $ Inv.nodes  : Factor w/ 7 levels "0-2","12-14",..: 1 1 1 1 1 1 1 1 1 1 .
..
##  $ Node.caps  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Deg.Malig  : int  3 2 2 2 2 2 2 1 2 2 ...
##  $ Breast     : Factor w/ 2 levels "left","right": 1 2 1 2 2 1 1 1 1 2 ...
##  $ Breast.quad: Factor w/ 5 levels "central","left_low",..: 2 5 2 3 4 2 2
2 2 3 ...
##  $ IR.Radiat  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int  146 164 165 184 185 207 234 264
265
##   ..- attr(*, "names")= chr  "146" "164" "165" "184" ...
```

```r
# Logistic Regression
BC.logit <- glm(Class~.,data = BCdata,family = binomial)
summary(BC.logit)
```

```
##
## Call:
## glm(formula = Class ~ ., family = binomial, data = BCdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6909  -0.7663  -0.4461   0.8386   2.3143
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -20.14038 3956.18073  -0.005  0.99594
## Age30-39               16.98184 3956.18052   0.004  0.99658
## Age40-49               16.43693 3956.18051   0.004  0.99669
## Age50-59               16.31422 3956.18052   0.004  0.99671
## Age60-69               16.99021 3956.18055   0.004  0.99657
## Age70-79                1.13566 4272.57421   0.000  0.99979
## Menopauselt40         -16.11590 1696.19453  -0.010  0.99242
## Menopausepremeno        0.36858    0.49565   0.744  0.45711
## Tumor.size10-14        -1.86583    1.67130  -1.116  0.26425
## Tumor.size15-19        -0.06938    1.35521  -0.051  0.95917
## Tumor.size20-24         0.11514    1.28639   0.090  0.92868
## Tumor.size25-29         0.30421    1.29502   0.235  0.81428
## Tumor.size30-34         0.45836    1.28732   0.356  0.72180
## Tumor.size35-39        -0.06334    1.39648  -0.045  0.96382
## Tumor.size40-44        -0.24392    1.38881  -0.176  0.86058
## Tumor.size45-49        -0.16288    1.79225  -0.091  0.92759
## Tumor.size5-9         -15.95139 1941.92629  -0.008  0.99345
## Tumor.size50-54         0.59842    1.49301   0.401  0.68856
## Inv.nodes12-14          0.98717    1.44260   0.684  0.49379
## Inv.nodes15-17          0.68968    0.97900   0.704  0.48113
## Inv.nodes24-26         17.15306 3956.18041   0.004  0.99654
## Inv.nodes3-5            0.70832    0.51421   1.377  0.16837
## Inv.nodes6-8            0.89930    0.68577   1.311  0.18974
## Inv.nodes9-11           1.50225    0.99683   1.507  0.13180
## Node.capsyes            0.15739    0.47939   0.328  0.74267
## Deg.Malig               0.81556    0.25272   3.227  0.00125 **
## Breastright            -0.30037    0.34017  -0.883  0.37724
## Breast.quadleft_low     0.44714    0.75495   0.592  0.55367
## Breast.quadleft_up      0.27630    0.76855   0.360  0.71921
## Breast.quadright_low    0.05287    0.94064   0.056  0.95518
## Breast.quadright_up     0.93832    0.83101   1.129  0.25884
## IR.Radiatyes            0.39683    0.36962   1.074  0.28300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 334.78  on 276  degrees of freedom
## Residual deviance: 261.60  on 245  degrees of freedom
## AIC: 325.6
##
## Number of Fisher Scoring iterations: 16
```

Note that only one of the predictor variables (Degree of Malignancy) is stati
stically significant at 1% level of significance

```
# Calculating the Logistic Probabilities
BC.logit.probs <- predict(BC.logit,type = "response")
attach(BCdata)

# Predicting the Class
BC.logit.pred <- rep("0", 277)
BC.logit.pred[BC.logit.probs>0.5]="1"

# Constructing 2x2 table
table(BC.logit.pred,Class)
```

```
##              Class
## BC.logit.pred   0   1
##             0 180  37
##             1  16  44
```

```
(180+44)/277
```

```
## [1] 0.8086643
```

```
mean(BC.logit.pred==Class)
```

```
## [1] 0.8086643
```

Accuracy based on the overall data set is about 80%

```
# Creating training and testing data sets
train <- sample(1:nrow(BCdata), 200)
BCdata.test <- BCdata[-train,]
Class.test <- Class[-train]

# Logistic model for training set
BC.logit.train<- glm(Class~.,data = BCdata,family = binomial,subset = train)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(BC.logit.train)
```

```
##
## Call:
## glm(formula = Class ~ ., family = binomial, data = BCdata, subset = train)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5705  -0.7475  -0.2983   0.6217   2.3802
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -36.84050 7066.34051  -0.005   0.9958
## Age30-39               18.03568 6522.63858   0.003   0.9978
## Age40-49               18.34256 6522.63857   0.003   0.9978
## Age50-59               17.60359 6522.63857   0.003   0.9978
## Age60-69               18.88658 6522.63860   0.003   0.9977
## Age70-79                1.92702 7021.87829   0.000   0.9998
## Menopauselt40         -16.86408 3186.30564  -0.005   0.9958
## Menopausepremeno        0.44763    0.60866   0.735   0.4621
## Tumor.size10-14        14.33111 2718.15296   0.005   0.9958
## Tumor.size15-19        16.77689 2718.15277   0.006   0.9951
## Tumor.size20-24        16.28290 2718.15274   0.006   0.9952
## Tumor.size25-29        16.76501 2718.15274   0.006   0.9951
## Tumor.size30-34        16.68692 2718.15272   0.006   0.9951
## Tumor.size35-39        16.58949 2718.15279   0.006   0.9951
## Tumor.size40-44        16.07245 2718.15280   0.006   0.9953
## Tumor.size45-49        15.33707 2718.15306   0.006   0.9955
## Tumor.size5-9          -0.90966 4619.05607   0.000   0.9998
## Tumor.size50-54        -1.01834 3887.94243   0.000   0.9998
## Inv.nodes12-14        -18.01177 6522.63872  -0.003   0.9978
## Inv.nodes15-17          0.61702    1.08214   0.570   0.5686
## Inv.nodes24-26         18.06600 6522.63867   0.003   0.9978
## Inv.nodes3-5            0.98434    0.61949   1.589   0.1121
## Inv.nodes6-8            1.08706    0.93811   1.159   0.2465
## Inv.nodes9-11          19.31592 4599.38283   0.004   0.9966
## Node.capsyes            0.07717    0.61152   0.126   0.8996
## Deg.Malig               0.67853    0.30056   2.258   0.0240 *
## Breastright            -0.86289    0.41873  -2.061   0.0393 *
## Breast.quadleft_low    -0.21503    0.96895  -0.222   0.8244
## Breast.quadleft_up     -0.56334    0.97826  -0.576   0.5647
## Breast.quadright_low   -0.72693    1.24110  -0.586   0.5581
## Breast.quadright_up     0.74315    1.03048   0.721   0.4708
## IR.Radiatyes            0.27337    0.47912   0.571   0.5683
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 239.05  on 199  degrees of freedom
## Residual deviance: 169.52  on 168  degrees of freedom
## AIC: 233.52
##
## Number of Fisher Scoring iterations: 17
```

Observe that on the training data set 2 predictor variables i) Degree of Malignancy and ii) Breast Right are statistically significant at 5% level of significance.

```
par(mfrow=c(2,2))
plot(BC.logit.train)
```

```
## Warning: not plotting observations with leverage one:
##   70, 128, 153
```

```
## Warning: not plotting observations with leverage one:
##   70, 128, 153
```

```
# Predicting the probabilities and Classes for the test data set
BC.logit.test.probs <- predict(BC.logit.train,BCdata.test,type = "response")
BC.logit.test.pred <- rep("0",77)
BC.logit.test.pred[BC.logit.test.probs>0.5] <- "1"
```

```
class(BC.logit.test.pred)
```

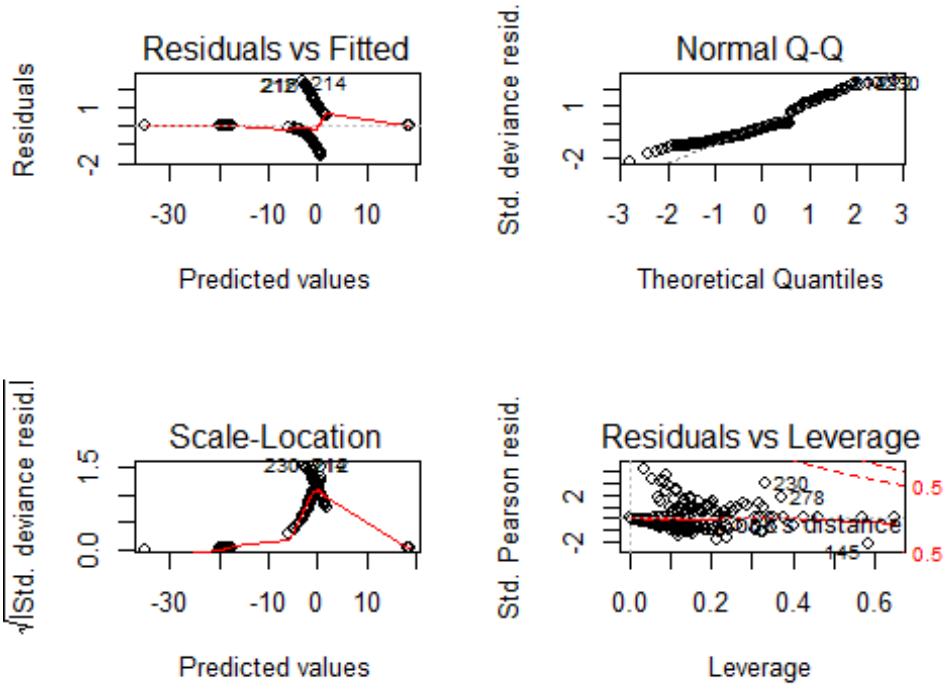
```
## [1] "character"
```

```
Pred.updated <- as.factor(BC.logit.test.pred)
Pred.updated1 <- as.numeric(BC.logit.test.pred)
```

```
# Constrcuting the 2x2 cross table
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```




```
library(gmodels)
CrossTable(Class.test,Pred.updated)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  77
##
##
##                | Pred.updated
##   Class.test |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |        44 |         9 |        53 |
##              |     0.417 |     1.272 |           |
##              |     0.830 |     0.170 |     0.688 |
##              |     0.759 |     0.474 |           |
##              |     0.571 |     0.117 |           |
## -------------|-----------|-----------|-----------|
```

```
##            1 |        14 |        10 |         24 |
##              |     0.920 |     2.808 |            |
##              |     0.583 |     0.417 |     0.312 |
##              |     0.241 |     0.526 |            |
##              |     0.182 |     0.130 |            |
## -------------|-----------|-----------|-----------|
## Column Total |        58 |        19 |         77 |
##              |     0.753 |     0.247 |            |
## -------------|-----------|-----------|-----------|
##
##
```

```
confusionMatrix(Class.test,Pred.updated,positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 44  9
##          1 14 10
##
##                Accuracy : 0.7013
##                  95% CI : (0.5862, 0.8003)
##     No Information Rate : 0.7532
##     P-Value [Acc > NIR] : 0.8814
##
##                   Kappa : 0.2618
##
##  Mcnemar's Test P-Value : 0.4042
##
##             Sensitivity : 0.5263
##             Specificity : 0.7586
##          Pos Pred Value : 0.4167
##          Neg Pred Value : 0.8302
##              Prevalence : 0.2468
##          Detection Rate : 0.1299
##    Detection Prevalence : 0.3117
##       Balanced Accuracy : 0.6425
##
##        'Positive' Class : 1
##
```

As expected, on the test data set, the accuracy has decreased to 70% as compared to 80% on the overall data set. Note that the kappa value is about 0.26, implying that there is only fair agreement between model predictions and true values.

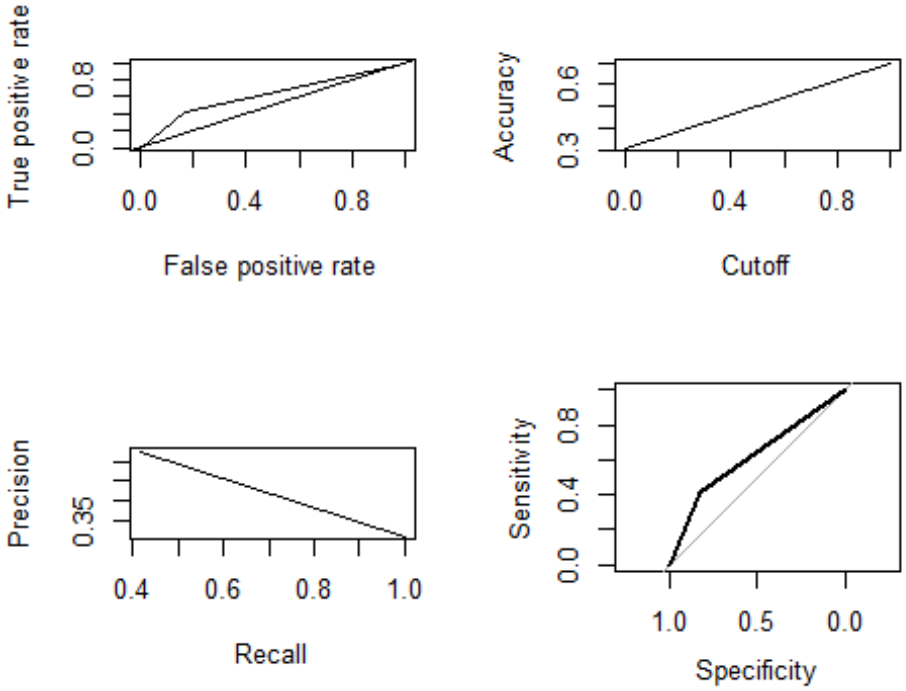
```
temp <- data.frame(Class.test,Pred.updated1)

# ROC plots and AUC using ROCR package
library(ROCR)
```

```
## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

data(temp)

## Warning in data(temp): data set 'temp' not found

rocr.pred <- prediction(temp$Pred.updated1,temp$Class.test)
rocr.perf <- performance(rocr.pred, "tpr","fpr")
plot(rocr.perf)
abline(a=0,b=1)

acc.perf = performance(rocr.pred, measure = "acc")
plot(acc.perf)

rocr.perf1 <- performance(rocr.pred,"prec","rec")
plot(rocr.perf1)


rocr.perf2 <- performance(rocr.pred, "auc")
rocr.perf2@y.values

## [[1]]
## [1] 0.6234277
```

Area Under Curve (AUC) is 0.6234

```
#ROC curve using pROC Package
library(pROC)

## Warning: package 'pROC' was built under R version 3.6.3

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

##
## Attaching package: 'pROC'

## The following object is masked from 'package:gmodels':
##
##     ci

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

roc.BC <- roc(temp$Class.test,temp$Pred.updated)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

roc.BC$auc

## Area under the curve: 0.6234

plot(roc.BC)
```

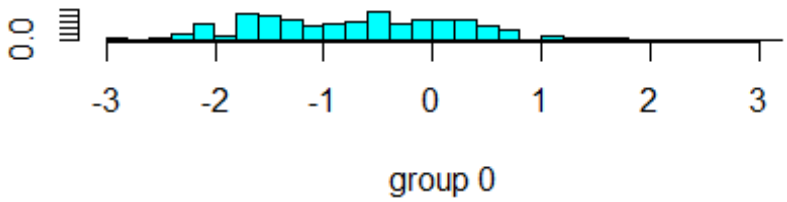



```
coords(roc.BC,x="best",input = "threshold", best.method = "youden",transpose
= TRUE)

##   threshold specificity sensitivity
##   0.5000000   0.8301887   0.4166667

#Linear Discriminant Analysis
library(MASS)
BC.lda <- lda(Class~.,data = BCdata,subset = train)
BC.lda

## Call:
## lda(Class ~ ., data = BCdata, subset = train)
##
## Prior probabilities of groups:
##     0     1
## 0.715 0.285
##
```

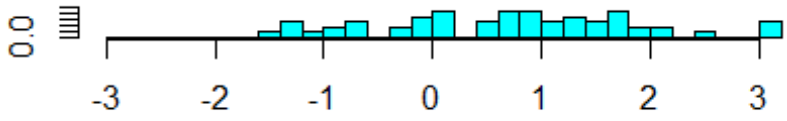
```
## Group means:
##    Age30-39  Age40-49  Age50-59  Age60-69   Age70-79 Menopauselt40
## 0 0.1188811 0.2867133 0.3776224 0.1748252 0.03496503    0.02797203
## 1 0.1578947 0.3684211 0.2456140 0.2280702 0.00000000    0.00000000
##   Menopausepremeno Tumor.size10-14 Tumor.size15-19 Tumor.size20-24
## 0        0.4965035      0.17482517      0.09090909       0.1608392
## 1        0.5964912      0.01754386      0.08771930       0.1754386
##   Tumor.size25-29 Tumor.size30-34 Tumor.size35-39 Tumor.size40-44
## 0       0.1748252       0.1398601      0.06993007      0.08391608
## 1       0.2807018       0.2631579      0.08771930      0.07017544
##   Tumor.size45-49 Tumor.size5-9 Tumor.size50-54 Inv.nodes12-14 Inv.nodes15
-17
## 0      0.01398601    0.02097902      0.03496503    0.006993007     0.02097
902
## 1      0.01754386    0.00000000      0.00000000    0.000000000     0.05263
158
##   Inv.nodes24-26 Inv.nodes3-5 Inv.nodes6-8 Inv.nodes9-11 Node.capsyes Deg.
Malig
## 0     0.00000000   0.08391608   0.02797203    0.00000000    0.1188811  1.9
09091
## 1     0.01754386   0.22807018   0.10526316    0.03508772    0.3684211  2.4
21053
##   Breastright Breast.quadleft_low Breast.quadleft_up Breast.quadright_low
## 0   0.5384615           0.3706294          0.3636364           0.09090909
## 1   0.3859649           0.4385965          0.2456140           0.07017544
##   Breast.quadright_up IR.Radiatyes
## 0           0.1048951    0.1818182
## 1           0.1929825    0.3684211
##
## Coefficients of linear discriminants:
##                              LD1
## Age30-39              1.9840937
## Age40-49              2.1913515
## Age50-59              1.6782875
## Age60-69              2.5295532
## Age70-79              1.5894554
## Menopauselt40        -1.1627443
## Menopausepremeno      0.2825485
## Tumor.size10-14      -0.2059686
## Tumor.size15-19       0.8761902
## Tumor.size20-24       0.4288697
## Tumor.size25-29       0.8994243
## Tumor.size30-34       0.9032952
## Tumor.size35-39       0.7302907
## Tumor.size40-44       0.2563836
## Tumor.size45-49      -0.2633864
## Tumor.size5-9        -0.7284808
## Tumor.size50-54      -0.6543492
## Inv.nodes12-14       -1.8511564
## Inv.nodes15-17        0.5850138
```

```
## Inv.nodes24-26         2.3049189
## Inv.nodes3-5           0.9568149
## Inv.nodes6-8           0.9916368
## Inv.nodes9-11          3.0113830
## Node.capsyes           0.1801135
## Deg.Malig              0.4936606
## Breastright           -0.5908274
## Breast.quadleft_low   -0.1146165
## Breast.quadleft_up    -0.4064085
## Breast.quadright_low -0.4371246
## Breast.quadright_up    0.7884922
## IR.Radiatyes           0.2458616
```

```
plot(BC.lda)
```



group 0



group 1

```
BC.lda.pred = predict(BC.lda,BCdata.test)
names(BC.lda.pred)
```

```
## [1] "class"     "posterior" "x"
```

```
BC.lda.class <- BC.lda.pred$class
CrossTable(BC.lda.class,Class.test)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
```

```
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  77
##
##
##              | Class.test
## BC.lda.class |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |        42 |        14 |        56 |
##              |     0.310 |     0.684 |           |
##              |     0.750 |     0.250 |     0.727 |
##              |     0.792 |     0.583 |           |
##              |     0.545 |     0.182 |           |
## -------------|-----------|-----------|-----------|
##            1 |        11 |        10 |        21 |
##              |     0.826 |     1.823 |           |
##              |     0.524 |     0.476 |     0.273 |
##              |     0.208 |     0.417 |           |
##              |     0.143 |     0.130 |           |
## -------------|-----------|-----------|-----------|
## Column Total |        53 |        24 |        77 |
##              |     0.688 |     0.312 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
confusionMatrix(Class.test,Pred.updated,positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 44  9
##          1 14 10
##
##                Accuracy : 0.7013
##                  95% CI : (0.5862, 0.8003)
##     No Information Rate : 0.7532
##     P-Value [Acc > NIR] : 0.8814
##
##                   Kappa : 0.2618
##
##  Mcnemar's Test P-Value : 0.4042
##
##             Sensitivity : 0.5263
```

```
##             Specificity : 0.7586
##          Pos Pred Value : 0.4167
##          Neg Pred Value : 0.8302
##              Prevalence : 0.2468
##          Detection Rate : 0.1299
##    Detection Prevalence : 0.3117
##       Balanced Accuracy : 0.6425
##
##        'Positive' Class : 1
##
```

Using Linear Discriminant Analysis, the accuracy obtained is 70% which same as that of Logistic Regression. Even the kappa statistic is 0.2618 which means only a fair agreement between model predicted values and actual values. Hence, no improvement in the model accuracy using Linear Discriminant Analysis over Logistic Regression.