Logistic Regression

Ben

28/09/2020

R Markdown

#Remove warnings

```
options(warn=-1)
```

#Reading libraries

```
library(caTools)
library(car)
```

```
## Loading required package: carData
```

```
library(DAAG)
```

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

```
##
## Attaching package: 'DAAG'
```

```
## The following object is masked from 'package:car':
##
##     vif
```

```
library(ROCR)
```

#Removing env variables

```
rm(list=ls(all=TRUE))
```

#Setting working directory

```
getwd()
```

```
## [1] "C:/Users/Ben Roshan/Documents"
```

```
setwd("C:/Users/Ben Roshan/Documents")
```

#Reading csv files

```
#flierresponse=read.csv(file='FlierResponse.csv',header=T)
framingham=read.csv(file='framingham.csv',header=T)
```

#Studying the data

```
#str(flierresponse)
#summary(flierresponse)
summary(framingham)
```

```
##       male             age          education     currentSmoker   
##  Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000  
##  Median :0.0000   Median :49.00   Median :2.000   Median :0.0000  
##  Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941  
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000  
##                                   NA's   :105                     
##    cigsPerDay         BPMeds        prevalentStroke     prevalentHyp   
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000  
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000  
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000  
##  Mean   : 9.006   Mean   :0.02962   Mean   :0.005896   Mean   :0.3106  
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000  
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000  
##  NA's   :29       NA's   :53                                           
##     diabetes          totChol          sysBP           diaBP      
##  Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.0  
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.0  
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.0  
##  Mean   :0.02571   Mean   :236.7   Mean   :132.4   Mean   : 82.9  
##  3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 90.0  
##  Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.5  
##                    NA's   :50                                     
##       BMI          heartRate         glucose         TenYearCHD    
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.0000  
##  1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.0000  
##  Median :25.40   Median : 75.00   Median : 78.00   Median :0.0000  
##  Mean   :25.80   Mean   : 75.88   Mean   : 81.96   Mean   :0.1519  
##  3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:0.0000  
##  Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.0000  
##  NA's   :19      NA's   :1        NA's   :388                      
```

#Removing NA values

```
#flierresponse$Response=as.factor(flierresponse$Response)
framingham <- na.omit(framingham)
```

#Random split the data into training and testing sets

```
set.seed(1000)
split=sample.split(framingham$TenYearCHD,SplitRatio = 0.70)
train=subset(framingham,split==TRUE)
test=subset(framingham,split==FALSE)
```

#Logistic regression model

```
framinghamlog=glm(TenYearCHD~.,data=train,family = binomial)
summary(framinghamlog)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9465  -0.6019  -0.4168  -0.2723   2.8342
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.147517   0.856122  -9.517  < 2e-16 ***
## male             0.562997   0.131368   4.286 1.82e-05 ***
## age              0.066380   0.007983   8.315  < 2e-16 ***
## education       -0.130789   0.060676  -2.156  0.03112 *
## currentSmoker    0.031966   0.188375   0.170  0.86525
## cigsPerDay       0.019760   0.007455   2.650  0.00804 **
## BPMeds           0.146584   0.283906   0.516  0.60564
## prevalentStroke  0.633471   0.527053   1.202  0.22940
## prevalentHyp     0.254990   0.166855   1.528  0.12646
## diabetes         0.138585   0.368311   0.376  0.70671
## totChol          0.003480   0.001325   2.626  0.00864 **
## sysBP            0.012884   0.004570   2.819  0.00482 **
## diaBP           -0.003368   0.007699  -0.437  0.66176
## BMI             -0.001536   0.015467  -0.099  0.92089
## heartRate       -0.003204   0.005094  -0.629  0.52945
## glucose          0.007366   0.002807   2.624  0.00868 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2185.3  on 2560  degrees of freedom
## Residual deviance: 1914.3  on 2545  degrees of freedom
## AIC: 1946.3
##
## Number of Fisher Scoring iterations: 5
```

#Checking for multicollinearity

```
car::vif(framinghamlog)
```

```
##            male             age       education   currentSmoker      cigsPerDay
##        1.249028        1.267015        1.064799        2.588899        2.744778
##          BPMeds prevalentStroke    prevalentHyp        diabetes         totChol
##        1.106263        1.030437        2.015416        1.722506        1.070313
##           sysBP           diaBP             BMI       heartRate         glucose
##        3.521935        2.809076        1.235812        1.096363        1.732848
```

#Accuracy on training set

```
predictTrain=predict(framinghamlog,type="response",newdata=train)
#predictTrain
```

#Confusion matrix with threshold of 0.5

```
table(train$TenYearCHD, predictTrain>0.5)
```

```
##
##      FALSE TRUE
##   0  2159   12
##   1   361   29
```

#Model metrics

```
accuracy=(3082+51)/(3082+506+19+51)
accuracy
```

```
## [1] 0.856479
```

```
precision=(2170)/(2170+357)
precision
```

```
## [1] 0.8587258
```

```
sensitivity_recall=(2170)/(2170+9)
sensitivity_recall
```

```
## [1] 0.9958697
```

```
specificity=(30)/(30+357)
specificity
```

```
## [1] 0.07751938
```

#Accuracy on test set

```
predictTest=predict(framinghamlog,type="response",newdata=test)
#predictTest
```

#Confusion matrix with threshold of 0.5

```
table(test$TenYearCHD, predictTest>0.5)
```

```
##
##      FALSE TRUE
##   0   926    4
##   1   151   16
```

```
table(test$TenYearCHD, predictTest>0.9)
```

```
##
##      FALSE TRUE
##   0   930    0
##   1   166    1
```

```
table(test$TenYearCHD, predictTest>0.7)
```

```
##
##      FALSE TRUE
##   0   930    0
##   1   164    3
```

```
table(test$TenYearCHD, predictTest>0.3)
```

```
##
##      FALSE TRUE
##   0   840   90
##   1   124   43
```

```
table(test$TenYearCHD, predictTest>0.1)
```

```
##
##      FALSE TRUE
##   0   430  500
##   1    26  141
```

#Model metrics

```
accuracy=(915+12)/(915+12+158+7)
accuracy
```

```
## [1] 0.8489011
```

```
precision=(915)/(915+158)
precision
```

```
## [1] 0.8527493
```

```
sensitivity_recall=(915)/(915+7)
sensitivity_recall
```

```
## [1] 0.9924078
```

```
specificity=(12)/(12+158)
specificity
```

```
## [1] 0.07058824
```

#Checking AIC AIC should be minimum

```
summary(framinghamlog)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9465  -0.6019  -0.4168  -0.2723   2.8342
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.147517   0.856122  -9.517  < 2e-16 ***
## male             0.562997   0.131368   4.286 1.82e-05 ***
## age              0.066380   0.007983   8.315  < 2e-16 ***
## education       -0.130789   0.060676  -2.156  0.03112 *
## currentSmoker    0.031966   0.188375   0.170  0.86525
## cigsPerDay       0.019760   0.007455   2.650  0.00804 **
## BPMeds           0.146584   0.283906   0.516  0.60564
## prevalentStroke  0.633471   0.527053   1.202  0.22940
## prevalentHyp     0.254990   0.166855   1.528  0.12646
## diabetes         0.138585   0.368311   0.376  0.70671
## totChol          0.003480   0.001325   2.626  0.00864 **
## sysBP            0.012884   0.004570   2.819  0.00482 **
## diaBP           -0.003368   0.007699  -0.437  0.66176
## BMI             -0.001536   0.015467  -0.099  0.92089
## heartRate       -0.003204   0.005094  -0.629  0.52945
## glucose          0.007366   0.002807   2.624  0.00868 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2185.3  on 2560  degrees of freedom
## Residual deviance: 1914.3  on 2545  degrees of freedom
## AIC: 1946.3
##
## Number of Fisher Scoring iterations: 5
```

#ROC Curve

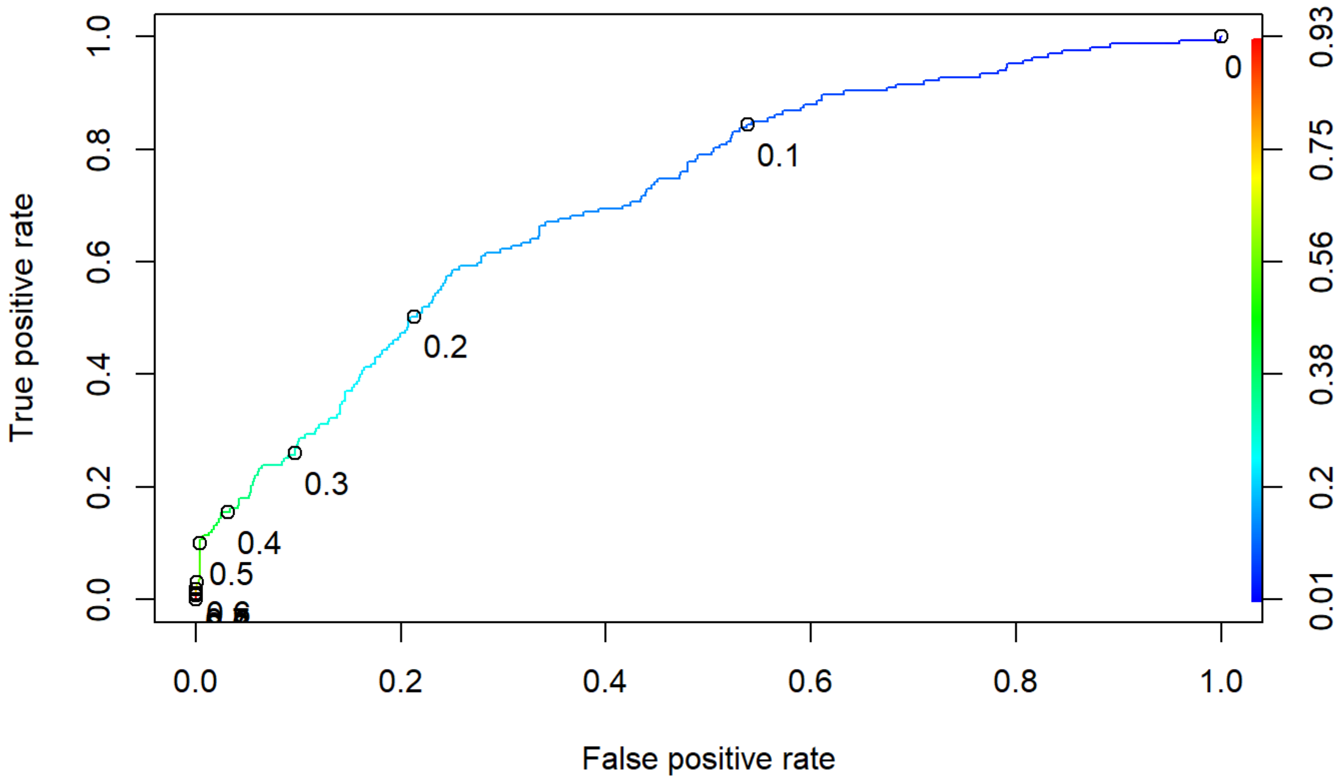
```
summary(test)
```

```
##       male             age         education    currentSmoker
##  Min.   :0.0000   Min.   :33.0   Min.   :1.00   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:42.0   1st Qu.:1.00   1st Qu.:0.000
##  Median :0.0000   Median :49.0   Median :2.00   Median :0.000
##  Mean   :0.4284   Mean   :49.7   Mean   :1.96   Mean   :0.474
##  3rd Qu.:1.0000   3rd Qu.:56.0   3rd Qu.:3.00   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :69.0   Max.   :4.00   Max.   :1.000
##    cigsPerDay         BPMeds        prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 8.658   Mean   :0.03191   Mean   :0.002735   Mean   :0.3054
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
##     diabetes          totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :113.0   Min.   : 85.5   Min.   : 51.00
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:118.0   1st Qu.: 76.00
##  Median :0.00000   Median :234.0   Median :129.0   Median : 82.00
##  Mean   :0.02735   Mean   :237.2   Mean   :133.2   Mean   : 83.23
##  3rd Qu.:0.00000   3rd Qu.:265.0   3rd Qu.:144.0   3rd Qu.: 89.00
##  Max.   :1.00000   Max.   :410.0   Max.   :248.0   Max.   :142.50
##       BMI          heartRate         glucose        TenYearCHD
##  Min.   :16.69   Min.   : 45.00   Min.   : 45.0   Min.   :0.0000
##  1st Qu.:23.10   1st Qu.: 67.00   1st Qu.: 71.0   1st Qu.:0.0000
##  Median :25.48   Median : 75.00   Median : 78.0   Median :0.0000
##  Mean   :25.81   Mean   : 75.15   Mean   : 82.3   Mean   :0.1522
##  3rd Qu.:28.09   3rd Qu.: 82.00   3rd Qu.: 87.0   3rd Qu.:0.0000
##  Max.   :44.55   Max.   :143.00   Max.   :394.0   Max.   :1.0000
```

```
ROCRpred = prediction(predictTest, test$TenYearCHD)
as.numeric(performance(ROCRpred, "auc")@y.values)
```

```
## [1] 0.7162514
```

```
ROCRperf <- performance(ROCRpred, "tpr", "fpr")
par(mfrow=c(1,1))
plot(ROCRperf, colorize = TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))
```


True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
0.93
0.75
0.56
0.38
0.2
0.01
0
0.1
0.2
0.3
0.4
0.5