VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

ABHINAV VIJAYAKUMAR

19BCE1311

CSE3506 – ESSENTIALS OF DATA ANALYTICS
LAB-5

DR. LAKSHMI PATHI JAKKAMPUTI (L21 + L22)

Tasks for Week-5: Logistic Regression

Understand the following operations/functions on to perform logistic Regression and perform similar operations on 'Social_Network_Ads' dataset based on given instructions.

Aim: To perform logistic Regression and perform similar operations on the 'Social_Network_Ads' dataset.

Algorithm:

1. Import the dataset and load the caTools library.

2. Split the data using split function into test and train data in a ratio=0.8.

3. Convert the purchased and Gender variable to categorical variable using as.factor.

4. Apply the generalized linear model using glm command for the dependent and independent variables and print the summary.

5. Using the trained model, predict the output for the test data and observe the accuracy and plot the graphs.

6. Generate and Display the confusion matrix.

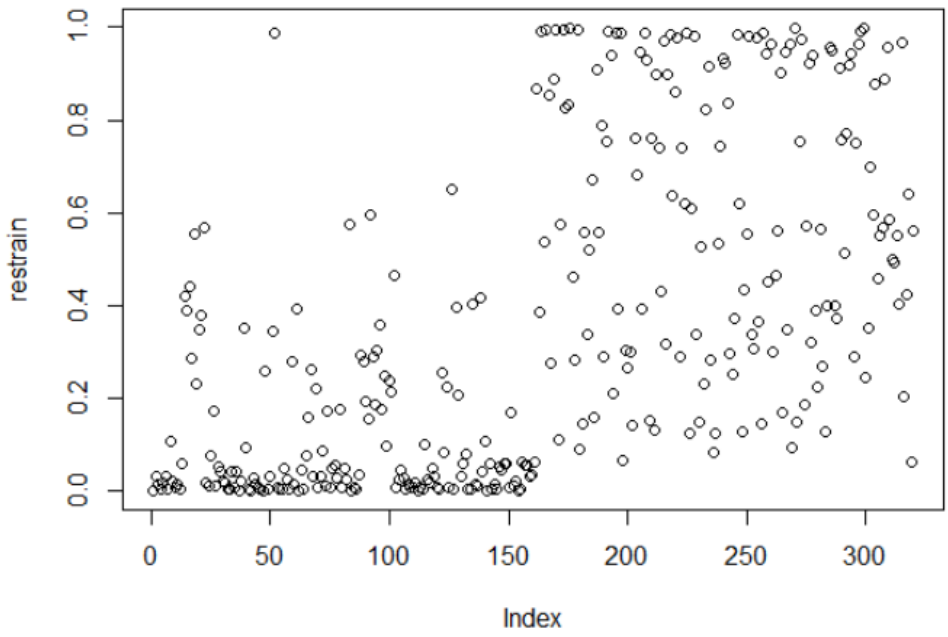
Statistics:

1. Summary of Applied Model:

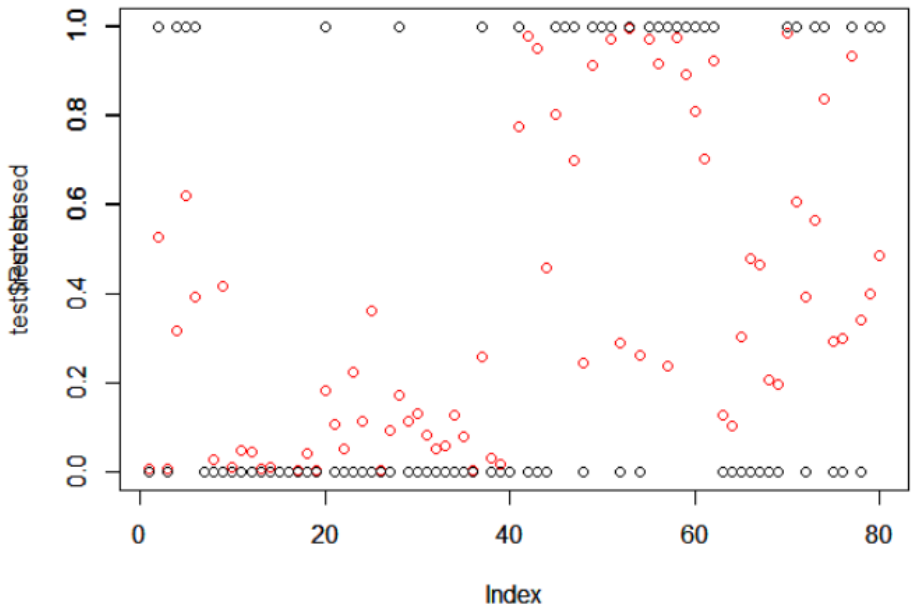
```
Call:
glm(formula = Purchased ~ Age + Gender + EstimatedSalary, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9347  -0.5373  -0.1411   0.3484   2.4629

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.283e+01  1.532e+00  -8.376  < 2e-16 ***
Age              2.415e-01  2.985e-02   8.091 5.92e-16 ***
GenderMale       2.981e-01  3.423e-01   0.871    0.384
EstimatedSalary  3.470e-05  6.134e-06   5.656 1.55e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 415.59  on 319  degrees of freedom
Residual deviance: 217.97  on 316  degrees of freedom
AIC: 225.97

Number of Fisher Scoring iterations: 6
```

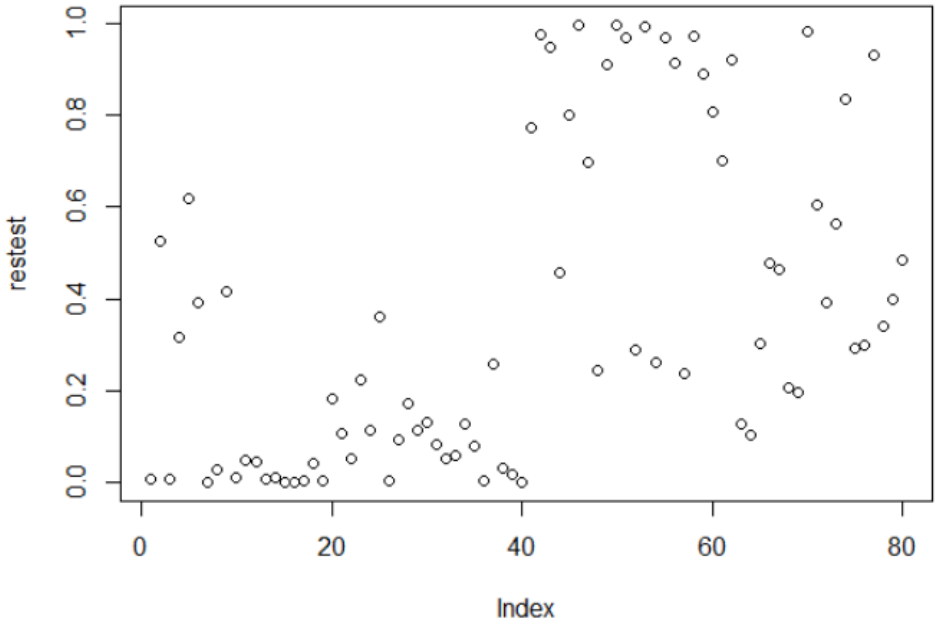
```
> restrain<-predict(mymodel,train,type='response')
> plot(restrain)
> restest<-predict(mymodel,test,type='response')
> plot(restest,col='red')
> par(new=TRUE)
> plot(test$Purchased)
> cfmatrix<-table(Act=test$Purchased, pred=restest>0.5)
> cfmatrix
   pred
Act FALSE TRUE
  0    48    2
  1     8   22
> Acc=(cfmatrix[[1,1]]+cfmatrix[[2,2]])/sum(cfmatrix)
> Acc
[1] 0.875
```

2. Graph of Predicted train data:



3. Graph of Predicted test





Inference:

1. As seen in the summary of glm model, the profit value (Pr(>|z|)) is less than 0.05 for all variables apart from gender male. Hence, all but one are accepted in the model**.**

2. The accuracy of the trained model is observed to be 0.875 i.e., 87.5%.

3. There are a total of 80 objectives as seen from the confusion matrix.

4. The graph shows the actual and predicted values of the trained model for the train and test data.

Program:

```
rm(list=ls())


setwd("C:/Users/Abhinav Vijayakumar/Desktop/VIT Academics/Sem 6/EDA/LAB/LAB 5")


mydata<-read.csv("Social_Network_Ads.csv")
library(caTools)
splitd<-sample.split(mydata,SplitRatio = 0.8)
train=subset(mydata,splitd=="TRUE")
test=subset(mydata,splitd=="FALSE")
train
mydata$Gender<-as.factor(mydata$Gender)
```

```
mydata$Purchased<-as.factor(mydata$Purchased)
mymodel <- glm(Purchased ~ Age+Gender+EstimatedSalary, data=train,
               family='binomial')
summary(mymodel)
restrain<-predict(mymodel,train,type='response')
plot(restrain)
restest<-predict(mymodel,test,type='response')
plot(restest,col='red')
par(new=TRUE)
plot(test$Purchased)
cfmatrix<-table(Act=test$Purchased, pred=restest>0.5)
cfmatrix
Acc=(cfmatrix[[1,1]]+cfmatrix[[2,2]])/sum(cfmatrix)
Acc
plot(restest)
```