Assignment 8

Alka Santosh Naik

2023-10-24

#Alka Santosh Naik  
rm(list=ls())  
library("rio")  
emaildata=import("6304 Module 8 Assignment Data.xlsx")  
colnames(emaildata)=tolower(make.names(colnames(emaildata)))  
set.seed(16999752)  
emailsample=emaildata[sample(1:nrow(emaildata),750),]  
attach(emailsample)

#1  
str(emailsample)

## 'data.frame': 750 obs. of 12 variables:  
## $ row : num 3678 1555 236 434 1071 ...  
## $ spam : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ to\_multiple : num 0 0 0 0 0 0 1 0 0 0 ...  
## $ image : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ dollar : num 0 21 0 0 0 0 0 0 4 13 ...  
## $ winner : chr "no" "no" "no" "no" ...  
## $ inherit : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ password : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ format : num 1 1 1 1 1 1 0 1 1 1 ...  
## $ re\_subj : num 1 0 0 0 1 1 0 0 0 0 ...  
## $ urgent\_subj : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ exclaim\_subj: num 0 0 0 0 0 0 0 0 1 1 ...

Analysis1: structure of emailsample.

#2  
maineffect.out=glm(spam~to\_multiple+image+dollar+winner+inherit+password+format+re\_subj+urgent\_subj+exclaim\_subj,data=emailsample,  
 family="binomial")

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

summary(maineffect.out)

##   
## Call:  
## glm(formula = spam ~ to\_multiple + image + dollar + winner +   
## inherit + password + format + re\_subj + urgent\_subj + exclaim\_subj,   
## family = "binomial", data = emailsample)  
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.79529 0.20577 -3.865 0.000111 \*\*\*  
## to\_multiple -2.69375 0.75788 -3.554 0.000379 \*\*\*  
## image -0.10339 0.72209 -0.143 0.886143   
## dollar -0.06002 0.04447 -1.350 0.177153   
## winneryes 2.17863 0.69841 3.119 0.001812 \*\*   
## inherit -0.53492 0.98167 -0.545 0.585813   
## password -13.69142 643.74595 -0.021 0.983032   
## format -1.53189 0.28348 -5.404 6.52e-08 \*\*\*  
## re\_subj -2.90025 0.75587 -3.837 0.000125 \*\*\*  
## urgent\_subj NA NA NA NA   
## exclaim\_subj 0.59480 0.46453 1.280 0.200392   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 465.27 on 749 degrees of freedom  
## Residual deviance: 370.45 on 740 degrees of freedom  
## AIC: 390.45  
##   
## Number of Fisher Scoring iterations: 18

Analysis2: full regression.

Analysis3: Here the residual deviance is 332.22 which is lower than null deviance is 427.85 showing a better model fit for the full regression than the null regression. Lower the residual deviance value, better the model fits to the data so residual deviance can be lowered by identifying the significant independent variables.

Analysis4: Variable winner=yes will have the greatest influence in increasing the modeled probability that an email will be spam

Analysis5: Variable image will have the greatest influence in decreasing the modeled probability that an email will be spam.

#6  
unique(urgent\_subj)

## [1] 0

Analysis6: As we see urgent\_subj has only one value i.e zero , due to this R was not able to calculate the beta coefficient for URGENT\_SUBJ variable.

#7  
reduced.out=glm(spam~to\_multiple+dollar+inherit+password+format,data=emailsample,  
 family="binomial")

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

Analysis7: Reduced regression.

#8  
email.predictions=expand.grid(to\_multiple=unique(emailsample$to\_multiple),  
 image=quantile(emailsample$image,c(0,.33,.50,.67,1)),  
 dollar=quantile(emailsample$dollar,c(0,.33,.50,.67,1)),  
 winner=unique(emailsample$winner),  
 inherit=quantile(emailsample$inherit,c(0,.33,.50,.67,1)),  
 password=quantile(emailsample$password,c(0,.33,.50,.67,1)),  
 format=unique(emailsample$format),  
 re\_subj=unique(emailsample$re\_subj),  
 urgent\_subj=unique(emailsample$urgent\_subj),  
 exclaim\_subj=unique(emailsample$exclaim\_subj)  
 )  
  
  
  
email.predictions$pred\_prob=predict(reduced.out,  
 newdata=email.predictions,type="response")  
  
  
str(email.predictions)

## 'data.frame': 20000 obs. of 11 variables:  
## $ to\_multiple : num 0 1 0 1 0 1 0 1 0 1 ...  
## $ image : Named num 0 0 0 0 0 0 0 0 9 9 ...  
## ..- attr(\*, "names")= chr [1:20000] "0%" "0%" "33%" "33%" ...  
## $ dollar : Named num 0 0 0 0 0 0 0 0 0 0 ...  
## ..- attr(\*, "names")= chr [1:20000] "0%" "0%" "0%" "0%" ...  
## $ winner : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ inherit : Named num 0 0 0 0 0 0 0 0 0 0 ...  
## ..- attr(\*, "names")= chr [1:20000] "0%" "0%" "0%" "0%" ...  
## $ password : Named num 0 0 0 0 0 0 0 0 0 0 ...  
## ..- attr(\*, "names")= chr [1:20000] "0%" "0%" "0%" "0%" ...  
## $ format : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ re\_subj : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ urgent\_subj : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ exclaim\_subj: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pred\_prob : num 0.07016 0.00649 0.07016 0.00649 0.07016 ...  
## - attr(\*, "out.attrs")=List of 2  
## ..$ dim : Named int [1:10] 2 5 5 2 5 5 2 2 1 2  
## .. ..- attr(\*, "names")= chr [1:10] "to\_multiple" "image" "dollar" "winner" ...  
## ..$ dimnames:List of 10  
## .. ..$ to\_multiple : chr [1:2] "to\_multiple=0" "to\_multiple=1"  
## .. ..$ image : chr [1:5] "image=0" "image=0" "image=0" "image=0" ...  
## .. ..$ dollar : chr [1:5] "dollar= 0" "dollar= 0" "dollar= 0" "dollar= 0" ...  
## .. ..$ winner : chr [1:2] "winner=no" "winner=yes"  
## .. ..$ inherit : chr [1:5] "inherit=0" "inherit=0" "inherit=0" "inherit=0" ...  
## .. ..$ password : chr [1:5] "password= 0" "password= 0" "password= 0" "password= 0" ...  
## .. ..$ format : chr [1:2] "format=1" "format=0"  
## .. ..$ re\_subj : chr [1:2] "re\_subj=1" "re\_subj=0"  
## .. ..$ urgent\_subj : chr "urgent\_subj=0"  
## .. ..$ exclaim\_subj: chr [1:2] "exclaim\_subj=0" "exclaim\_subj=1"

Analysis8: Generated prediction object.

#9  
email.predictions=email.predictions[order(-email.predictions$pred\_prob),]  
#first 6 rows  
head(email.predictions,n=6)

## to\_multiple image dollar winner inherit password format re\_subj  
## 2501 0 0 0 no 0 0 0 1  
## 2503 0 0 0 no 0 0 0 1  
## 2505 0 0 0 no 0 0 0 1  
## 2507 0 0 0 no 0 0 0 1  
## 2509 0 9 0 no 0 0 0 1  
## 2511 0 0 0 no 0 0 0 1  
## urgent\_subj exclaim\_subj pred\_prob  
## 2501 0 0 0.2516821  
## 2503 0 0 0.2516821  
## 2505 0 0 0.2516821  
## 2507 0 0 0.2516821  
## 2509 0 0 0.2516821  
## 2511 0 0 0.2516821

#last 6 rows  
tail(email.predictions,n=6)

## to\_multiple image dollar winner inherit password format re\_subj  
## 19995 0 0 54 yes 6 28 0 0  
## 19996 1 0 54 yes 6 28 0 0  
## 19997 0 0 54 yes 6 28 0 0  
## 19998 1 0 54 yes 6 28 0 0  
## 19999 0 9 54 yes 6 28 0 0  
## 20000 1 9 54 yes 6 28 0 0  
## urgent\_subj exclaim\_subj pred\_prob  
## 19995 0 1 2.220446e-16  
## 19996 0 1 2.220446e-16  
## 19997 0 1 2.220446e-16  
## 19998 0 1 2.220446e-16  
## 19999 0 1 2.220446e-16  
## 20000 0 1 2.220446e-16