DSC_520_week9_Assignment01

Reenie Christudass

2022-08-15

Load Libraries

```
if(!require('foreign')) {
  install.packages('foreign')
  library('foreign')
}
```

```
## Loading required package: foreign
```

```
if(!require('tidyr')) {
  install.packages('tidyr')
  library('tidyr')
}
```

```
## Loading required package: tidyr
```

```
## Warning: package 'tidyr' was built under R version 4.2.1
```

```
install.packages("MASS", repos="http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/chris/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'MASS' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\chris\AppData\Local\Temp\Rtmpa6Oqaa\downloaded_packages
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.2.1
```

```
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/chris/dsc520/data")
```

```r
## Load the `data/r4ds/heights.csv` to
newdata <- read.csv("C:/Users/chris/dsc520/data/binary-classifier-data.csv")
head(newdata)
```

```
##   label        x        y
## 1     0 70.88469 83.17702
## 2     0 74.97176 87.92922
## 3     0 73.78333 92.20325
## 4     0 66.40747 81.10617
## 5     0 69.07399 84.53739
## 6     0 72.23616 86.38403
```

```r
newdata2 <-newdata[,c("label","x","y")]

riskmodel<-glm(label~x+y,family=binomial,data=newdata2)
summary(riskmodel)
```

```
##
## Call:
## glm(formula = label ~ x + y, family = binomial, data = newdata2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3728  -1.1697  -0.9575   1.1646   1.3989
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.424809   0.117224   3.624  0.00029 ***
## x           -0.002571   0.001823  -1.411  0.15836
## y           -0.007956   0.001869  -4.257 2.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2075.8  on 1497  degrees of freedom
## Residual deviance: 2052.1  on 1495  degrees of freedom
## AIC: 2058.1
##
## Number of Fisher Scoring iterations: 4
```

```r
##VARIABLE SELECTION
riskmodel_new <- stepAIC(riskmodel)
```

```
## Start:  AIC=2058.07
## label ~ x + y
##
##        Df Deviance    AIC
## - x     1   2054.1 2058.1
## <none>      2052.1 2058.1
## - y     1   2070.4 2074.4
##
```

```
## Step:  AIC=2058.06
## label ~ y
##
##        Df Deviance    AIC
## <none>      2054.1 2058.1
## - y     1   2075.8 2077.8
```

```
summary(riskmodel_new)
```

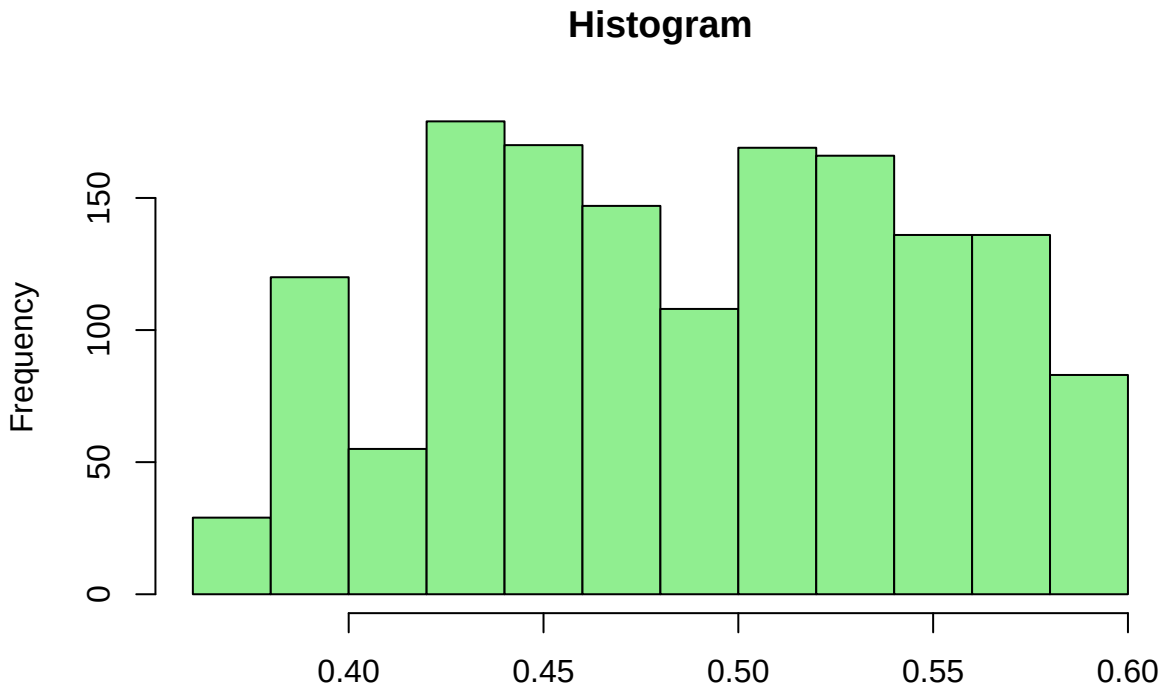
```
##
## Call:
## glm(formula = label ~ y, family = binomial, data = newdata2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3335  -1.1350  -0.9886   1.1771   1.4287
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.332800   0.097188   3.424 0.000616 ***
## y           -0.008480   0.001831  -4.630 3.65e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2075.8  on 1497  degrees of freedom
## Residual deviance: 2054.1  on 1496  degrees of freedom
## AIC: 2058.1
##
## Number of Fisher Scoring iterations: 4
```

```
##Analysis of the outcome

summary(newdata2$fitted.values)
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

```
hist(riskmodel_new$fitted.values,main = " Histogram ",xlab = "", col = 'light green')
```




```
newdata2$Predict <- ifelse(riskmodel_new$fitted.values >0.5,"0","1")
head(newdata2)
```

```
##   label        x        y Predict
## 1     0 70.88469 83.17702       1
## 2     0 74.97176 87.92922       1
## 3     0 73.78333 92.20325       1
## 4     0 66.40747 81.10617       1
## 5     0 69.07399 84.53739       1
## 6     0 72.23616 86.38403       1
```

```
##Model Performance Evaluation
riskmodel$aic
```

```
## [1] 2058.067
```

```
riskmodel_new$aic
```

```
## [1] 2058.06
```

```
## CONCLUSION : A model with minimum AIC value is preferred.The above shows the AIC of the original mod
```

```
##Confusion Matrix
mytable <- table(newdata2$label,newdata2$Predict)
mytable
```

```
##
##       0   1
##   0 333 434
##   1 357 374
```

```
efficiency <- sum(diag(mytable))/sum(mytable)
efficiency
```

```
## [1] 0.4719626
```

```
## CONCLUSION: The accuracy of our model is 47.1%
```