Logistic-Regression-for-binary-classification.R

patriciamaya

2020-12-05

```
#LOGISTIC REGRESSION for binary classification

library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.0.2
```

```
data("Carseats")
attach(Carseats)
str(Carseats)
```

```
## 'data.frame':    400 obs. of  11 variables:
##  $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice  : num  138 111 113 117 141 124 115 136 132 132 ...
##  $ Income     : num  73 48 35 100 64 113 105 81 110 113 ...
##  $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
##  $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
##  $ Price      : num  120 83 80 97 128 72 108 120 124 124 ...
##  $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
##  $ Age        : num  42 65 59 55 38 78 71 67 76 76 ...
##  $ Education  : num  17 10 12 14 13 16 15 10 10 17 ...
##  $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
##  $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

```
set.seed(256)
#create new categorial variable
High <- as.factor(ifelse(Sales >= 8, "YES", "NO")) #categorical variable w/ 2 levels
Data <- data.frame(Carseats, High) #new df with High variable included
Data <- Data[ ,-1] #removes 1st column "Sales"
colnames(Data)[11] <- "Target" #change name to last (11th) column to Target
head(Data)
```

```
##   CompPrice Income Advertising Population Price ShelveLoc Age Education Urban
## 1       138     73          11        276   120       Bad  42        17   Yes
## 2       111     48          16        260    83      Good  65        10   Yes
## 3       113     35          10        269    80    Medium  59        12   Yes
## 4       117    100           4        466    97    Medium  55        14   Yes
## 5       141     64           3        340   128       Bad  38        13   Yes
## 6       124    113          13        501    72       Bad  78        16    No
##    US Target
## 1 Yes    YES
## 2 Yes    YES
## 3 Yes    YES
## 4 Yes     NO
## 5  No     NO
```

```
## 6 Yes    YES
```

```
indx <- sample(2,nrow(Data), replace=T, prob = c(0.8, 0.2))
train <- Data[indx ==1, ]
test <- Data[indx ==2, ]

#glm - generalized linear model (~)
#glm(categorical target ~ inputs, data= train, family= "binomial")
logitModel <- glm(Target ~ . , data = train, family = "binomial")
summary(logitModel)
```

```
##
## Call:
## glm(formula = Target ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.54086  -0.29416  -0.05406   0.16124   3.00877
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -5.042977   2.647769  -1.905   0.0568 .
## CompPrice        0.165231   0.025361   6.515 7.26e-11 ***
## Income           0.035677   0.009113   3.915 9.04e-05 ***
## Advertising      0.320174   0.059138   5.414 6.16e-08 ***
## Population      -0.002236   0.001649  -1.356   0.1751
## Price           -0.161611   0.021128  -7.649 2.02e-14 ***
## ShelveLocGood    7.827315   1.073606   7.291 3.08e-13 ***
## ShelveLocMedium  2.932128   0.693583   4.228 2.36e-05 ***
## Age             -0.076811   0.015677  -4.900 9.60e-07 ***
## Education       -0.015547   0.082481  -0.188   0.8505
## UrbanYes        -0.396721   0.481712  -0.824   0.4102
## USYes           -0.666006   0.652072  -1.021   0.3071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 453.77  on 336  degrees of freedom
## Residual deviance: 149.73  on 325  degrees of freedom
## AIC: 173.73
##
## Number of Fisher Scoring iterations: 7
```

```
#Deviance: measure of goodness of fit of a glm : -2 log (likelihood)
  #higher number - worse fit
#Null deviance: deviance of model with NO input variables, only intercept
#Residual deviance: deviance of full model.

predictions <- predict(logitModel, newdata = test)
#predicted log of odds

predictions <- predict(logitModel, newdata = test, type="response")
#***** probability of being in class YES
```

```
Class <- ifelse(predictions >= 0.5, "YES", "NO")
Class
```

```
##     6    19    20    31    33    41    42    43    45    49    54    61    63
## "YES" "YES" "YES" "YES"  "NO"  "NO"  "NO" "YES"  "NO"  "NO"  "NO" "YES"  "NO"
##    65    69    86    98   100   108   118   131   138   150   156   160   167
## "YES" "YES"  "NO" "YES"  "NO" "YES"  "NO" "YES"  "NO" "YES"  "NO" "YES"  "NO"
##   168   172   187   189   191   199   204   211   223   238   245   247   250
##  "NO" "YES" "YES"  "NO" "YES"  "NO"  "NO"  "NO"  "NO"  "NO" "YES" "YES"  "NO"
##   256   262   266   267   268   274   277   280   296   304   314   316   318
##  "NO"  "NO"  "NO" "YES"  "NO" "YES"  "NO"  "NO"  "NO" "YES" "YES"  "NO"  "NO"
##   333   339   351   352   356   371   375   383   384   385   388
##  "NO"  "NO" "YES" "YES"  "NO"  "NO"  "NO"  "NO" "YES" "YES" "YES"
```

```
test$Target == Class
```

```
##     6    19    20    31    33    41    42    43    45    49    54    61    63
##  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##    65    69    86    98   100   108   118   131   138   150   156   160   167
## FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##   168   172   187   189   191   199   204   211   223   238   245   247   250
##  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
##   256   262   266   267   268   274   277   280   296   304   314   316   318
##  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##   333   339   351   352   356   371   375   383   384   385   388
##  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
```

```
#confusion matrix
table(test$Target, Class)
```

```
##      Class
##       NO YES
##   NO  31   3
##   YES  5  24
```