Project1LR.R

RaxyR

2021-03-07

```
#title: "CS 4375 Project 1 Logistic Regression"
#author: "Ramesh Kanakala"
#subtitle: "This is an R script with the purpose of running logistic regression
#on a titanic data set to observe run time and other metrics"

### Logistic Regression
#load the data
ttnc <- read.csv(file = 'titanic_project.csv')
#ttnc$pclass <- as.factor(ttnc$pclass)
ttnc$sex <- as.factor(ttnc$sex)
ttnc$survived <- as.factor(ttnc$survived)

#dividing into train/test, putting 75% in train
i <- 1:900
train <- ttnc[i,]
test <- ttnc[-i,]

start <- Sys.time()
#train logistic regression model
glm1 <- glm(survived~pclass, family = "binomial", data = train)
end <- Sys.time()

#print coefficients of model
glm1$coefficients[]
```

```
## (Intercept)      pclass
##    1.297166   -0.779929
```

```
#test on test data
probs <- predict(glm1, newdata=test, type="response")
pred <- ifelse(probs>0.5, 1, 0)

#print accuracy, sensitivity, and specificity #check spelling
library(caret)
```

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

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.4
```

```
confusionMatrix(as.factor(pred), as.factor(test$survived))$overall[1]
```

```
##  Accuracy
## 0.6712329
```

```
confusionMatrix(as.factor(pred), as.factor(test$survived))$byClass[1]
```

```
## Sensitivity
##   0.8481013
```

```
confusionMatrix(as.factor(pred), as.factor(test$survived))$byClass[2]
```

```
## Specificity
##   0.4626866
```

```
#time difference
end - start
```

```
## Time difference of 0.009507895 secs
```

```
#DATA EXPLORATION: FUNCTIONS 1-3
#data exploration 1
str(ttnc)
```

```
## 'data.frame':    1046 obs. of  5 variables:
##  $ X       : int  738 868 971 938 456 139 840 510 626 1099 ...
##  $ pclass  : int  3 3 3 3 2 1 3 2 3 3 ...
##  $ survived: Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 1 2 1 ...
##  $ sex     : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 2 1 1 ...
##  $ age     : num  19 22 20 1 63 38 19 39 17 3 ...
```

```
#data exploration 2
summary(ttnc)
```

```
##        X              pclass      survived sex          age
##  Min.   :   1.0   Min.   :1.000   0:619    0:388   Min.   : 0.1667
##  1st Qu.: 299.2   1st Qu.:1.000   1:427    1:658   1st Qu.:21.0000
##  Median : 575.5   Median :2.000                    Median :28.0000
##  Mean   : 600.2   Mean   :2.207                    Mean   :29.8811
##  3rd Qu.: 875.5   3rd Qu.:3.000                    3rd Qu.:39.0000
##  Max.   :1309.0   Max.   :3.000                    Max.   :80.0000
```

```
#data exploration 3
summary(glm1)
```

```
##
## Call:
## glm(formula = survived ~ pclass, family = "binomial", data = train)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4035  -0.7771  -0.7771   0.9671   1.6399
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.29717    0.19678   6.592 4.34e-11 ***
## pclass      -0.77993    0.08521  -9.153  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1211.4  on 899  degrees of freedom
## Residual deviance: 1122.1  on 898  degrees of freedom
## AIC: 1126.1
##
## Number of Fisher Scoring iterations: 4
```

```
confusionMatrix(as.factor(pred), as.factor(test$survived))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 67 36
##          1 12 31
##
##                Accuracy : 0.6712
##                  95% CI : (0.5887, 0.7467)
##     No Information Rate : 0.5411
##     P-Value [Acc > NIR] : 0.0009418
##
##                   Kappa : 0.3195
##
##  Mcnemar’s Test P-Value : 0.0009009
##
##             Sensitivity : 0.8481
##             Specificity : 0.4627
##          Pos Pred Value : 0.6505
##          Neg Pred Value : 0.7209
##              Prevalence : 0.5411
##          Detection Rate : 0.4589
##    Detection Prevalence : 0.7055
##       Balanced Accuracy : 0.6554
##
##        ’Positive’ Class : 0
##
```