HW-Sprint-06-Essential-Statistics.R

r1617576

2023-01-03

```
#Install Packages
install.packages("titanic")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
#Load Library
library(titanic)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.0      v purrr   1.0.0
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)

#Drop NA (Missing Value)
titanic_train <- na.omit(titanic_train)
nrow(titanic_train)
```

```
## [1] 714
```

```
#Glimpse Titanic
glimpse(titanic_train)
```

```
## Rows: 714
## Columns: 12
## $ PassengerId <int> 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19~
## $ Survived    <int> 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1~
## $ Pclass      <int> 3, 1, 3, 1, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 3, 2, 2, 3~
## $ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bradley (Fl~
```

```
## $ Sex         <chr> "male", "female", "female", "female", "male", "male", "mal~
## $ Age         <dbl> 22, 38, 26, 35, 35, 54, 2, 27, 14, 4, 58, 20, 39, 14, 55, ~
## $ SibSp       <int> 1, 1, 0, 1, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4, 1, 0, 0, 0~
## $ Parch       <int> 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1, 0, 0, 0, 0~
## $ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "113803", "37~
## $ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 51.8625, 21.0750~
## $ Cabin       <chr> "", "C85", "", "C123", "", "E46", "", "", "", "G6", "C103"~
## $ Embarked    <chr> "S", "C", "S", "S", "S", "S", "S", "S", "C", "S", "S", "S"~
```

```r
#1.Split Data
set.seed(24)
n <- nrow(titanic_train)
id <- sample(1:n,size=n*0.7) #70% train 30% test
train_data <- titanic_train[id, ]
test_data <- titanic_train[-id, ]

#2.Train Model
model_train <- glm(Survived ~ Pclass + Age + Sex,
                   data = train_data,
                   family ="binomial")
summary(model_train)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Age + Sex, family = "binomial",
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7144  -0.6682  -0.4020   0.6402   2.3580
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.010600   0.605971   8.269  < 2e-16 ***
## Pclass      -1.283121   0.166373  -7.712 1.24e-14 ***
## Age         -0.034415   0.009202  -3.740 0.000184 ***
## Sexmale     -2.534947   0.249507 -10.160  < 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: 673.56  on 498  degrees of freedom
## Residual deviance: 450.94  on 495  degrees of freedom
## AIC: 458.94
##
## Number of Fisher Scoring iterations: 5
```

```r
#3.Predict And Evaluate Model
train_data$prob_survived <- predict(model_train,type = "response")
train_data$pred_survived <- ifelse(train_data$prob_survived >= 0.5,1,0)

#4.Confusion matrix of Train Model
conM_titanic <- table(train_data$pred_survived,train_data$Survived,
                      dnn = c("Predicted","Actual"))
```

```r
acc_train <- (conM_titanic[1,1] + conM_titanic[2,2])/sum(conM_titanic)
prec_train <- conM_titanic[2,2] / (conM_titanic[2,1]+conM_titanic[2,2])
rec_train <- conM_titanic[2,2] / (conM_titanic[1,2]+conM_titanic[2,2])

f1_train <- 2*((prec_train*rec_train)/(prec_train+rec_train))

cat("Accuracy:", acc_train, "\nPrecision:", prec_train, "\nRecall:", rec_train,"\nF1:", f1_train)
```

```
## Accuracy: 0.7955912
## Precision: 0.7659574
## Recall: 0.7128713
## F1: 0.7384615
```

```r
#5.Test Model
model_test <- glm(Survived ~ Pclass + Age, data = test_data, family = "binomial")
summary(model_test)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Age, family = "binomial", data = test_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8168  -0.8308  -0.6149   0.9364   2.0955
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.42892    0.70859   4.839 1.30e-06 ***
## Pclass      -1.24956    0.21681  -5.763 8.25e-09 ***
## Age         -0.03906    0.01188  -3.287  0.00101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 290.94  on 214  degrees of freedom
## Residual deviance: 250.21  on 212  degrees of freedom
## AIC: 256.21
##
## Number of Fisher Scoring iterations: 4
```

```r
#6.Predict And Evaluate Model of Test Model
test_data$prob_survived <- predict(model_test,type = "response")
test_data$pred_survived <- ifelse(test_data$prob_survived >= 0.5,1,0)

#7.Confusion matrix of Test Model
conM_titanic2 <- table(test_data$pred_survived,test_data$Survived,
                       dnn = c("Predicted","Actual"))
acc_test <- (conM_titanic2[1,1] + conM_titanic2[2,2])/sum(conM_titanic2)
prec_test <- conM_titanic2[2,2] / (conM_titanic2[2,1]+conM_titanic2[2,2])
rec_test <- conM_titanic2[2,2] / (conM_titanic2[1,2]+conM_titanic2[2,2])

f1_test <- 2*((prec_test*rec_test)/(prec_test+rec_test))

cat("Accuracy:", acc_test, "\nPrecision:", prec_test, "\nRecall:", rec_test,"\nF1:", f1_test)
```

```
## Accuracy: 0.7069767
## Precision: 0.6712329
## Recall: 0.5568182
## F1: 0.6086957
```