HW_State_Titanic

Annop Techa

#Install Packages

##Drop NA (Missing Values)

```
titanic_train <- na.omit(titanic_train)

nrow(titanic_train)
```

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

```
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"~
```

Convert Sex to factor

```
titanic_train$Sex = as.factor(titanic_train$Sex)
str(titanic_train)
```

```
## 'data.frame':    714 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 7 8 9 10 11 ...
##  $ Survived   : int  0 1 1 1 0 0 0 1 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 1 3 3 2 3 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 1 1 1 ...
##  $ Age        : num  22 38 26 35 35 54 2 27 14 4 ...
##  $ SibSp      : int  1 1 0 1 0 0 3 0 1 1 ...
##  $ Parch      : int  0 0 0 0 0 0 1 2 0 1 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
##  - attr(*, "na.action")= 'omit' Named int [1:177] 6 18 20 27 29 30 32 33 37 43 ...
```

```
##   ..- attr(*, "names")= chr [1:177] "6" "18" "20" "27" ...
```

Split Data

```
set.seed(42)
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, ]
```

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
## -2.8617  -0.6485  -0.3554   0.6129   2.3884
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.604600   0.637259   8.795  < 2e-16 ***
## Pclass      -1.443887   0.174955  -8.253  < 2e-16 ***
## Age         -0.041450   0.009522  -4.353 1.34e-05 ***
## Sexmale     -2.739281   0.262607 -10.431  < 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: 432.26  on 495  degrees of freedom
## AIC: 440.26
##
## Number of Fisher Scoring iterations: 5
```

##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)
```

Confusion Matrix

```
conM_train <- table(train_data$pred_survived,train_data$Survived,
                    dnn=c("Predicted","Actual"))
```

Model_train Evalution

```
Acc_train <- (conM_train[1,1] + conM_train[2,2]) /sum(conM_train)
Pre_train <- (conM_train[2,2])/(conM_train[2,1]+conM_train[2,2])
Re_train <- (conM_train[2,2])/(conM_train[1,2]+conM_train[2,2])

F1_train <- 2*(Pre_train*Re_train)/(Pre_train+Re_train)

cat("Accuracy:",Acc_train,"\nPrecision:",Pre_train,"\nRecall:",Re_train,
    "\nF1:",F1_train)
```

```
## Accuracy: 0.7975952
## Precision: 0.7671958
## Recall: 0.7178218
## F1: 0.741688
```

Test Model

```
  model_test <- glm(Survived ~ Pclass + Age + Sex,data = test_data ,
                    family="binomial")
  summary(model_test)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Age + Sex, family = "binomial",
##     data = test_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2203  -0.7320  -0.4827   0.7004   2.2363
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.12628    0.85093   4.849 1.24e-06 ***
## Pclass      -1.01374    0.24149  -4.198 2.69e-05 ***
## Age         -0.02946    0.01312  -2.244   0.0248 *
## Sexmale     -2.17447    0.35253  -6.168 6.91e-10 ***
## ---
## 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: 211.31  on 211  degrees of freedom
## AIC: 219.31
##
## Number of Fisher Scoring iterations: 4
```

Predict and Evaluate Model

```
test_data$prob_survived <- predict(model_test,type="response")
test_data$pred_survived <- ifelse(test_data$prob_survived >= 0.5,1,0)
```

Confusion matrix

```
conM_test <- table(test_data$pred_survived,test_data$Survived,
                   dnn=c("Predicted","Actual"))
```

Model_train Evaluation

```
Acc_test <- (conM_test[1,1] + conM_test[2,2]) /sum(conM_test)
Pre_test <- (conM_test[2,2])/(conM_test[2,1]+conM_test[2,2])
Re_test <- (conM_test[2,2])/(conM_test[1,2]+conM_test[2,2])

F1_test <- 2*(Pre_test*Re_test)/(Pre_test+Re_test)

cat("Accuracy:",Acc_test,"\nPrecision:",Pre_test,"\nRecall:",Re_test,
    "\nF1:",F1_test)
```

```
## Accuracy: 0.7767442
## Precision: 0.7380952
## Recall: 0.7045455
## F1: 0.7209302
```