# Logistic Regression with titanic_train Dataset

Natthamon Ratthanasurakarn

### Load Library and Dataset

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## 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(titanic)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
data("titanic_train")
```

### Clean and Preview Data

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

**Step 1 : Split data (70% train : 30% test)**

```
titanic_train <- titanic_train %>%
  select(PassengerId,Name,Survived,Pclass,Sex,SibSp,Parch)

set.seed(555)
n <- nrow(titanic_train)
id <- sample(1:n, size = 0.7*n)
train_data <- titanic_train[id, ]
test_data <- titanic_train[-id, ]
paste("No. of train data :",nrow(train_data))
```

```
## [1] "No. of train data : 499"
```

```
paste("No. of test data :",nrow(test_data))
```

```
## [1] "No. of test data : 215"
```

**Step 2 : Train Model**

```
train_model <- glm(Survived ~ Pclass + Sex + SibSp + Parch,
                   data = train_data,
                   family = "binomial")
summary(train_model)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + SibSp + Parch, family = "binomial",
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2099  -0.7139  -0.4930   0.6469   2.2567
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.24375    0.39891   8.132 4.24e-16 ***
## Pclass      -0.89294    0.14105  -6.331 2.44e-10 ***
## Sexmale     -2.61096    0.24825 -10.518  < 2e-16 ***
## SibSp       -0.12541    0.14939  -0.839    0.401
## Parch        0.04151    0.14451   0.287    0.774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 674.33  on 498  degrees of freedom
## Residual deviance: 473.42  on 494  degrees of freedom
## AIC: 483.42
##
## Number of Fisher Scoring iterations: 4
```

```
prob_train <- predict(train_model,type = "response")
train_data$pred_Survived <- ifelse(prob_train >= 0.5,1,0)
head(train_data)
```

```
##     PassengerId                                Name Survived Pclass    Sex SibSp
## 628         628     Longley, Miss. Gretchen Fiske        1      1 female     0
## 430         430 Pickard, Mr. Berk (Berk Trembisky)        1      3   male     0
## 502         502               Canavan, Miss. Mary        0      3 female     0
## 436         436         Carter, Miss. Lucile Polk        1      1 female     1
## 461         461               Anderson, Mr. Harry        1      1   male     0
## 146         146      Nicholls, Mr. Joseph Charles        0      2   male     1
##     Parch pred_Survived
## 628     0             1
## 430     0             0
## 502     0             1
## 436     2             1
## 461     0             0
## 146     1             0
```

**Step 3 : Test Model**

```
prob_test <- predict(train_model, newdata = test_data,type = "response")
test_data$pred_Survived <- ifelse(prob_test >= 0.5,1,0)
head(test_data)
```

```
##    PassengerId                                                Name Survived
## 5            5                            Allen, Mr. William Henry        0
## 11          11                       Sandstrom, Miss. Marguerite Rut        1
## 13          13                        Saundercock, Mr. William Henry        0
## 16          16                      Hewlett, Mrs. (Mary D Kingcome)         1
## 19          19 Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)        0
## 21          21                                   Fynney, Mr. Joseph J        0
##    Pclass    Sex SibSp Parch pred_Survived
## 5       3   male     0     0             0
## 11      3 female     1     1             1
## 13      3   male     0     0             0
## 16      2 female     0     0             1
## 19      3 female     1     0             1
## 21      2   male     0     0             0
```

**Step 4 : Confusion matrix and Evaluate Model**

```
## confusion matrix
conM_train <- table(train_data$pred_Survived,
                          train_data$Survived,
                          dnn = c("Predicted","Actual"))
conM_test <- table(test_data$pred_Survived,
                         test_data$Survived,
                          dnn = c("Predicted","Actual"))
cat("Confusion matrix of train model :\n")
```

```
## Confusion matrix of train model :
```

```
print(conM_train)
```

```
##          Actual
## Predicted   0   1
##         0 254  65
##         1  42 138
```

```r
cat("Confusion matrix of test model :\n")
```

```
## Confusion matrix of test model :
```

```r
print(conM_test)
```

```
##           Actual
## Predicted   0   1
##         0 106  28
##         1  22  59
```

```r
## evaluate train model
acc_train <- (conM_train[1,1] + conM_train[2,2]) / sum(conM_train)
precision_train <- (conM_train[2,2]) / (conM_train[2,1] + conM_train[2,2])
recall_train <- (conM_train[2,2]) / (conM_train[1,2] + conM_train[2,2])
f1_train <- (2*(precision_train*recall_train)/(precision_train+recall_train))

## evaluate test model
acc_test <- (conM_test[1,1] + conM_test[2,2]) / sum(conM_test)
precision_test <- (conM_test[2,2]) / (conM_test[2,1] + conM_test[2,2])
recall_test <- (conM_test[2,2]) / (conM_test[1,2] + conM_test[2,2])
f1_test <- (2*(precision_test*recall_test)/(precision_test+recall_test))

## print evaluation model
df_accuracy <- data.frame(
  Model_name = c("Train model","Test model"),
  Accuracy = c(acc_train,acc_test),
  Precision = c(precision_train,precision_test),
  Recall = c(recall_train,recall_test),
  F1 = c(f1_train,f1_test)
)

print(df_accuracy)
```

```
##    Model_name  Accuracy Precision    Recall        F1
## 1 Train model 0.7855711 0.7666667 0.6798030 0.7206266
## 2  Test model 0.7674419 0.7283951 0.6781609 0.7023810
```