Logistic Regression

Titanic_train dataset 🚢

```
install.packages("titanic")
library(tidyverse)
library(titanic)
library(ggplot2)
library(dplyr)
glimpse(titanic_train)
```

```
Rows: 891
Columns: 12
$ PassengerId <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,
$ Survived    <int> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0,
$ Pclass      <int> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 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, NA, 54, 2, 27, 14, 4, 58, 20, 39, 14,
$ SibSp       <int> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4, 0, 1,
$ Parch       <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1, 0, 0,
$ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "113803", "3
$ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625,
$ Cabin       <chr> "", "C85", "", "C123", "", "", "E46", "", "", "", "G6", "
$ Embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S", "S", "C", "S", "S
```

```
Updating HTML index of packages in '.Library'
```

Clean Data -> Drop NA

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

Split Data to Train and Test (50:50)

```
#split data
set.seed(99)
n <- nrow(titanic_train)
id <- sample(1:n, size=n*0.5) #50% train 50% test
train_data <- titanic_train[id,]
test_data <- titanic_train[-id,]
```

Create Predicted and Evaluate Model

```
#Train model
train_model <- glm(Survived~Pclass + Age + Sex + SibSp,
                    data = train_data,
                    family = "binomial")
prob_train <- predict(train_model, type = "response")
train_data$pred_Survived <- ifelse(prob_train>=0.5,1,0)
##test model
prob_test <- predict(train_model, newdata = test_data, type = "response")
test_data$pred_Survived <- ifelse(prob_test >=0.5,1,0)

##confusion metrix
con_metrix_train <- table(train_data$pred_Survived,train_data$Survived,
                    dnn = c("Predicted","Actual"))
con_metrix_test <- table(test_data$pred_Survived,test_data$Survived,
                    dnn = c("Predicted","Actual"))

##Model Evaluation Train
acc_train <-(con_metrix_train[1,1]+con_metrix_train[2,2])/
              sum(con_metrix_train)
precision_train <- (con_metrix_train[2,2]/
              (con_metrix_train[2,1]+con_metrix_train[2,2]))
recall_train <- (con_metrix_train[2,2]/
              (con_metrix_train[1,2]+con_metrix_train[2,2]))
f1_train <- 2*((precision_train*recall_train)/(precision_train+recall_train))

##Model Evaluation Test
acc_test <- (con_metrix_test[1,1]+con_metrix_test[2,2])/sum(con_metrix_test)
precision_test <- (con_metrix_test[2,2]/(con_metrix_test[2,1]+con_metrix_test[2,2
recall_test <- (con_metrix_test[2,2]/(con_metrix_test[1,2]+con_metrix_test[2,2]))
f1_test <- 2*((precision_test*recall_test)/(precision_test+recall_test))
```

```
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)
)
cat("Hypothesis test of Multiple Regression:\n")
print(summary(train_model))
cat("------------------------------------------------------------------\n",
"Label of Values: 1 was Survived, 0 was Died\n","Confusion Matrix of Train:\n")
print(con_metrix_train)
cat("Confusion Matrix of Test:\n")
print(con_metrix_test)
cat("Accuracy:\n")
print(df_accuracy)
graph <- df_accuracy %>%
  gather(Accuracy:F1,
         key   = "type",
         value = "RMSE")
ggplot(graph, aes(x = type, y = RMSE, color = Model_name, group = Model_name)) +
  geom_line(size = 4) +
  coord_cartesian(ylim = c(0.73,0.83))+
  scale_color_manual(values = c("#F5C4C4","#ED9591")) +
  theme_minimal()+
  labs(title = "Model Evaluation",x = "Type of Accuracy",y = "Value")
```

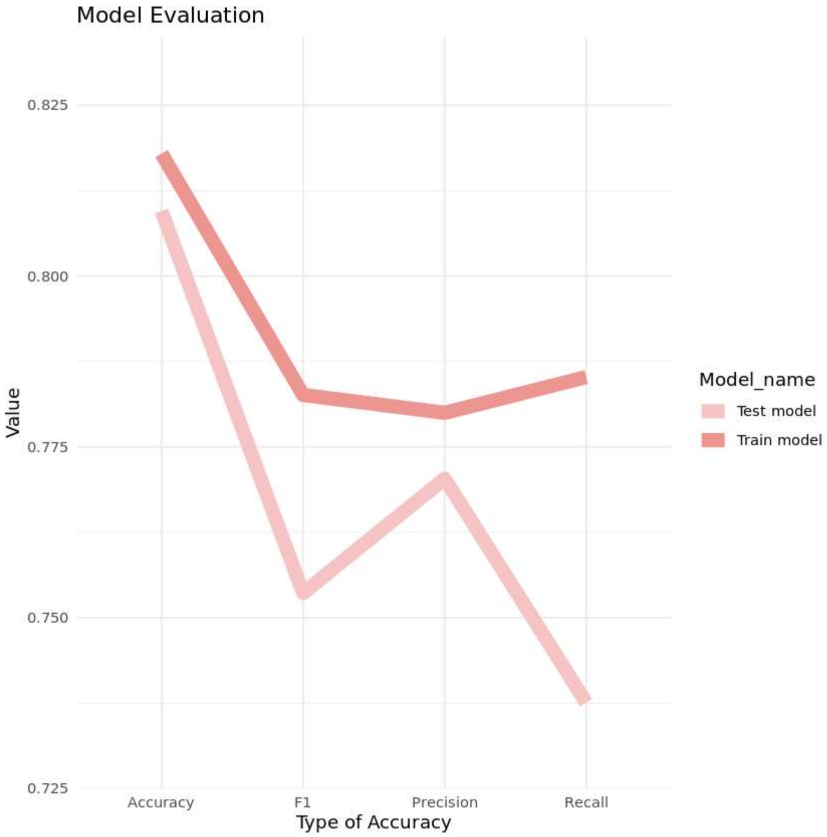
```
Hypothesis test of Multiple Regression:

Call:
glm(formula = Survived ~ Pclass + Age + Sex + SibSp, family = "binomial",
    data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1447  -0.6103  -0.3305   0.6229   2.6132

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.87162    0.87671   7.838 4.58e-15 ***
Pclass      -1.56949    0.22642  -6.932 4.16e-12 ***
Age         -0.06410    0.01214  -5.281 1.28e-07 ***
Sexmale     -2.72375    0.30931  -8.806  < 2e-16 ***
SibSp       -0.46045    0.18445  -2.496   0.0125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Download


Model Evaluation
0.825
0.800
0.775
0.750
0.725
Value
Accuracy
F1
Precision
Recall
Type of Accuracy
Model_name
Test model
Train model