Essential Statistics : Titanic homework

Kaew Tibkham

2023-02-04

Load library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(titanic)
```

Load dataset

```
data(titanic_train)
head(titanic_train)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                            Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                             Heikkinen, Miss. Laina female  26     0     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                           Allen, Mr. William Henry   male  35     0     0
## 6                                   Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

Glimpse Dataset

```
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, 1~
## $ Pclass      <int> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 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, 0~
## $ Parch       <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1, 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, 8.4583, 51.8625,~
## $ Cabin       <chr> "", "C85", "", "C123", "", "", "E46", "", "", "", "G6", "C~
## $ Embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S", "S", "C", "S", "S"~
```

Preps data

Drop missing value

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

Transform data

male = 0 female = 1

```
titanic_train$Sex <- if_else(titanic_train$Sex == "male", 0, 1)
```

Split data

```
set.seed(33)
n <- nrow(titanic_train)
id <- sample(1:n, size = n*0.7)
train_data <- titanic_train[id,]
test_data <- titanic_train[-id,]
```

Glimpse Dataset

```
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         <dbl> 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1~
## $ 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"~
```

Train model

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

```
##
## Call:
## glm(formula = Survived ~ Pclass + Age + Sex, family = "binomial",
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7789  -0.6828  -0.4072   0.6306   2.4690
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.528527   0.536696   4.711 2.46e-06 ***
## Pclass      -1.243977   0.166863  -7.455 8.98e-14 ***
## Age         -0.040819   0.008917  -4.578 4.70e-06 ***
## Sex          2.637010   0.252330  10.451  < 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: 674.33  on 498  degrees of freedom
## Residual deviance: 450.46  on 495  degrees of freedom
## AIC: 458.46
##
## Number of Fisher Scoring iterations: 5
```

```
train_data$prob_Survived <- predict(model, type="response") ## probability
train_data$pred_Survived <- ifelse(train_data$prob_Survived >= 0.5, 1, 0)
```

Test

```
test_data$prob_Survived <- predict(model, newdata = test_data, type="response") ## probability
test_data$pred_Survived <- if_else(test_data$prob_Survived >=0.5, 1, 0)
```

Model evaluation

Confusion metric of train model

```
train_con_metrix <- table(train_data$pred_Survived, train_data$Survived, dnn=c("predicted", "actual"))

train_accuracy <- (train_con_metrix[1, 1] + train_con_metrix[2, 2]) / sum(train_con_metrix)
train_precision <- train_con_metrix[2, 2] / (train_con_metrix[2, 1] + train_con_metrix[2, 2])
train_recall <- train_con_metrix[2, 2] / (train_con_metrix[1, 2] + train_con_metrix[2, 2])
train_f1_score <- 2 * train_precision*train_recall / (train_precision + train_recall)
cat("train_accuracy:", train_accuracy)
```

```
## train_accuracy: 0.8016032
```

```
cat("\ntrain_precision:", train_precision)
```

```
##
## train_precision: 0.7765957
```

```
cat("\ntrain_recall:", train_recall)
```

```
##
## train_recall: 0.7192118
```

```
cat("\ntrain_f1_score:", train_f1_score)
```

```
##
## train_f1_score: 0.7468031
```

```
cat("\n")
```

Confusion metric of test model

```
test_con_metrix <- table(test_data$pred_Survived, test_data$Survived, dnn=c("predicted", "actual"))

test_accuracy <- (test_con_metrix[1, 1] + test_con_metrix[2, 2]) / sum(test_con_metrix)
test_precision <- test_con_metrix[2, 2] / (test_con_metrix[2, 1] + test_con_metrix[2, 2])
test_recall <- test_con_metrix[2, 2] / (test_con_metrix[1, 2] + test_con_metrix[2, 2])
test_f1_score <- 2 * test_precision*test_recall / (test_precision + test_recall)
cat("test_accuracy:", test_accuracy)
```

```
## test_accuracy: 0.7813953
```

```
cat("\ntest_precision:", test_precision)
```

```
##
## test_precision: 0.7222222
```

```
cat("\ntest_recall:", test_recall)
```

```
##
## test_recall: 0.7471264
```

```
cat("\ntest_f1_score:", test_f1_score)
```

```
##
## test_f1_score: 0.7344633
```

```
cat("\n")
```

Summary

Train and test have similar accuracy

```
summary_model <- data.frame(
  "Group" = c('Train', 'Test'),
  "Accuracy" = c(train_accuracy, test_accuracy),
  "Precision" = c(train_precision, test_precision),
  "Recall" = c(train_recall, test_recall),
  "F1 Score" = c(train_f1_score, test_f1_score)) %>%
```

```
  pivot_longer(-Group ,
    names_to = "Type",
    values_to = "Percent")

ggplot(summary_model, aes(Type, Percent, fill = Group)) +
  geom_bar(stat='identity', position = 'dodge') +
  coord_cartesian(ylim = c(0, 1)) +
  theme_minimal()
```