Titanic survives with logistic regression in R

Chayanis Tavichai

2023-12-31

Logistic Regression Model

In this project, the model was trained to be able to predict which passengers were likely to survive on Titanic.

Install packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(titanic)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(dplyr)
```

Check data

```
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
```

As this data has many variables, and somes have NA values so we need to clean data

Clean data (NA null)

```
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"~
```

Titanic_Train Dataset

```
print(tibble(titanic_train), n = 20)
```

```
## # A tibble: 714 x 12
##    PassengerId Survived Pclass Name   Sex     Age SibSp Parch Ticket  Fare Cabin
##          <int>    <int>  <int> <chr>  <chr> <dbl> <int> <int> <chr>  <dbl> <chr>
##  1           1        0      3 "Brau~ male     22     1     0 A/5 2~  7.25 ""
##  2           2        1      1 "Cumi~ fema~    38     1     0 PC 17~ 71.3  "C85"
##  3           3        1      3 "Heik~ fema~    26     0     0 STON/~  7.92 ""
##  4           4        1      1 "Futr~ fema~    35     1     0 113803 53.1  "C12~
##  5           5        0      3 "Alle~ male     35     0     0 373450  8.05 ""
##  6           7        0      1 "McCa~ male     54     0     0 17463  51.9  "E46"
##  7           8        0      3 "Pals~ male      2     3     1 349909 21.1  ""
##  8           9        1      3 "John~ fema~    27     0     2 347742 11.1  ""
##  9          10        1      2 "Nass~ fema~    14     1     0 237736 30.1  ""
```

```
## 10          11        1      3 "Sand~ fema~     4     1     1 PP 95~ 16.7  "G6"
## 11          12        1      1 "Bonn~ fema~    58     0     0 113783 26.6  "C10~
## 12          13        0      3 "Saun~ male     20     0     0 A/5. ~  8.05 ""
## 13          14        0      3 "Ande~ male     39     1     5 347082 31.3  ""
## 14          15        0      3 "Vest~ fema~    14     0     0 350406  7.85 ""
## 15          16        1      2 "Hewl~ fema~    55     0     0 248706 16    ""
## 16          17        0      3 "Rice~ male      2     4     1 382652 29.1  ""
## 17          19        0      3 "Vand~ fema~    31     1     0 345763 18    ""
## 18          21        0      2 "Fynn~ male     35     0     0 239865 26    ""
## 19          22        1      2 "Bees~ male     34     0     0 248698 13    "D56"
## 20          23        1      3 "McGo~ fema~    15     0     0 330923  8.03 ""
## # i 694 more rows
## # i 1 more variable: Embarked <chr>
```

Explore Titanic Train Dataset

```
# Convert survived to factor
titanic_train$Survived <- factor(titanic_train$Survived,
                                 levels = c(0,1))
# Convert sex to factor
titanic_train$Sex <- factor(titanic_train$Sex,
                            levels = c("male", "female"))
# Convert pclass to factor
titanic_train$Pclass <- factor(titanic_train$Pclass,
                               levels = c(1,2,3),labels = c("1","2","3"))
```

Split Data

Use sample for sampling titanic passenger

Use 75% for train model and 25% for test model

```
set.seed(123)
n <- nrow(titanic_train)
row <- sample(1:n, size = n * 0.75)
train_data <- titanic_train[row,]
test_data <- titanic_train[-row,]
```

Use logistic regression for train model

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

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Fare, family = "binomial",
##     data = train_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.0245514  0.4593403   2.230 0.025715 *
## Pclass2     -1.1877070  0.3579293  -3.318 0.000906 ***
## Pclass3     -2.3288785  0.3611389  -6.449 1.13e-10 ***
## Sexfemale    2.4266242  0.2345189  10.347  < 2e-16 ***
## Age         -0.0359752  0.0086592  -4.155 3.26e-05 ***
```

```
## Fare         0.0004319  0.0023284   0.185 0.852840
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 715.87  on 534  degrees of freedom
## Residual deviance: 500.29  on 529  degrees of freedom
## AIC: 512.29
##
## Number of Fisher Scoring iterations: 4
```

Testing Model and Evaluate using accuracy (of confusion matrix)

```
# Train data
train_data$prob_survived_train <- predict(model_mlm, type = "response")
train_data$pred_survived_train <- if_else(train_data$prob_survived_train >= 0.5, 1, 0)

# Test data
test_data$prob_survived_test <- predict(model_mlm, newdata = test_data, type = "response")
test_data$pred_survived_test <- ifelse(test_data$prob_survived_test >= 0.5, 1, 0)

df <- data.frame(train = mean(train_data$Survived == train_data$pred_survived_train),
                 test = mean(test_data$Survived == test_data$pred_survived_test))
df
```

```
##       train      test
## 1 0.7850467 0.8212291
```

We can observe that the logistic model has the ability to train on unseen data and does not suffer from overfitting.

Use confusion matrix for explain the model

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

```
##          Actual
## Predicted   0   1
##         0 275  64
##         1  51 145
```

Summary

```
Acc <- (conM[1,1] + conM[2,2]) / sum(conM)
Precision <- conM[2,2]/ sum(conM[2,])
Recall <- conM[2,2] / sum(conM[,2])
f1 <- (2*(Precision*Recall)/(Precision+Recall))
cat("Accuracy :", Acc,
    "\nPrecision :", Precision,
    "\nRecall :", Recall,
    "\nf1 :", f1)
```

```
## Accuracy : 0.7850467
## Precision : 0.7397959
```

```
## Recall : 0.6937799
## f1 : 0.7160494
```

This model, relying on factors like passenger class, gender, age, and ticket price, can estimate with 69%-78% accuracy whether someone aboard the Titanic would have survived, as shown in the confusion matrix.