Statistics: Logistics Regression Bootcamp Homework

Sakon Sukkhapanya Sangiamsri

2023-08-17

The assignment for this class is "to create Logistics Regression by use 'Titanic' data set"

Logistic Regression Creating Process

1. Activated library(titanic) to allow us use this data set to create Logistic Regression
2. Discover and clean the data
3. Split data by Random Sampling method from a whole data set
4. Create train and test model
5. Create Confuse model of train and test for model evaluation
6. Evaluate train and test model to verify the prediction accuracy of this model ## Activated Library

```
library(titanic)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     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
```

Discover and Clean The Data

```
# Discover the data
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"~
```

```
# Clean the data
titanic_train <- na.omit(titanic_train)
```

Split Data

```
# Data random sampling from 'titanic_train', split by 80/20
set.seed(30)
n <- nrow(titanic_train)
id <- sample(1:n, n*0.8)
```

```
# Assign splited data into train and test data
train_data <- titanic_train[id, ]
test_data <- titanic_train[-id, ]
```

```
# Verify the significant of data variable
model_titanic <- glm(Survived ~ Pclass + Sex + Age, data = train_data, family = "binomial")
summary(model_titanic)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age, family = "binomial",
##     data = train_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.747974   0.543134   8.742  < 2e-16 ***
## Pclass      -1.154199   0.150026  -7.693 1.43e-14 ***
## Sexmale     -2.490601   0.227339 -10.955  < 2e-16 ***
## Age         -0.035114   0.008201  -4.282 1.86e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 773.61  on 570  degrees of freedom
## Residual deviance: 535.02  on 567  degrees of freedom
## AIC: 543.02
##
## Number of Fisher Scoring iterations: 4
```

Create Train and Test Model

```
# Train model (Cut off at 0.6 of probability)
train_data$prob_survived <- predict(model_titanic, type = "response")
train_data$pred_survived <- ifelse(train_data$prob_survived >= 0.6, 1, 0)
```

```
# Test model (Cut off at 0.6 of probability)
test_data$prob_survived <- predict(model_titanic, newdata = test_data, type = "response")
test_data$pred_survived <- ifelse(test_data$prob_survived >= 0.6, 1, 0)
```

Create Confuse Matrix

```
# Train model
train_conM <- table(train_data$pred_survived,
      train_data$Survived,
      dnn = c("Predicted", "Actual"))
train_conM
```

```
##          Actual
## Predicted   0   1
##         0 302  88
##         1  34 147
```

```
# Test model
test_conM <- table(test_data$pred_survived,
      test_data$Survived,
      dnn = c("Predicted", "Actual"))
test_conM
```

```
##          Actual
## Predicted  0  1
##         0 78 15
##         1 10 40
```

```
trainAcc <- (train_conM[1,1] + train_conM[2,2])/sum(train_conM)
trainPre <- train_conM[2,2]/(train_conM[2,1] + train_conM[2,2])
trainRec <- train_conM[2,2]/(train_conM[1,2] + train_conM[2,2])
trainF1 <- 2*((trainPre*trainRec)/(trainPre+trainRec))

cat("Train Model Evaluation",
  "\nAccuracy", trainAcc,
  "\nPrecision", trainPre,
  "\nRecall", trainRec,
  "\nF1", trainF1)
```

```
## Train Model Evaluation
## Accuracy 0.7863398
## Precision 0.8121547
## Recall 0.6255319
## F1 0.7067308
```

```r
testAcc <- (test_conM[1,1] + test_conM[2,2])/sum(test_conM)
testPre <- test_conM[2,2]/(test_conM[2,1] + test_conM[2,2])
testRec <- test_conM[2,2]/(test_conM[1,2] + test_conM[2,2])
testF1 <- 2*((testPre*testRec)/(testPre+testRec))

cat("Test Model Evaluation",
    "\nAccuracy", testAcc,
    "\nPrecision", testPre,
    "\nRecall", testRec,
    "\nF1", testF1)
```

```
## Test Model Evaluation
## Accuracy 0.8251748
## Precision 0.8
## Recall 0.7272727
## F1 0.7619048
```

```r
# Verify the accuracy of train and test model prediction
# 1. Train model accuracy
traincheck <- mean(train_data$Survived == train_data$pred_survived)
cat("The percentage of the train model prediction accuracy is", traincheck)
```

```
## The percentage of the train model prediction accuracy is 0.7863398
```

```r
# 2. Test model accuracy
testcheck <- mean(test_data$Survived == test_data$pred_survived)
cat("The percentage of the test model prediction accuracy is", testcheck)
```

```
## The percentage of the test model prediction accuracy is 0.8251748
```