Data Science Bootcamp - Logistic Regression

Arraya

2023-10-20

Intruction

Create model by using **'Titanic'** dataset

- Split data
- Train model
- Test model
- Evaluate model and find accuracy

Before using **Titanic** dataset, Titanic packaged must be installed.

```
# installed package titanic
library(titanic)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## 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
```

```
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 data

- Drop missing values

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

```
## [1] 714
```

Split data

- Split data into 2 groups: 70% train data and 30% test data

```r
set.seed(42)
n <- nrow(titanic_train)
id <- sample(1:n, size=n*0.7)
train_data <- titanic_train[id, ]
test_data <- titanic_train[-id, ]
```

Train model

- Set threshold: 60%.

```r
model1 <- glm(Survived ~ Pclass, data=train_data, family="binomial")
train_data$prob_survived <- predict(model1, type="response")
train_data$pred_survived <- ifelse(train_data$prob_survived >= 0.6, 1, 0)
```

Test Model

- Use same threshold (60%) to find prediction from test data

```r
test_data$prob_survived <- predict(model1, newdata=test_data, type="response")
test_data$pred_survived <- ifelse(test_data$prob_survived >= 0.6, 1, 0)
```

Evaluate Model

- Find accuracy by creating Confusion Matrix

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

```
##           Actual
## Predicted   0   1
##         0 253 114
##         1  44  88
```

```r
train_acc <- (train_conM[1, 1] + train_conM[2, 2]) / sum(train_conM)
train_pre <- (train_conM[2,2]/ (train_conM[2,1] + train_conM[2,2]))
train_rec <- (train_conM[2,2]/ (train_conM[1,2] + train_conM[2,2]))
train_F1 <- 2*((train_pre*train_rec)/(train_pre+train_rec))
```

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

```
##           Actual
## Predicted   0   1
##         0 107  54
##         1  20  34
```

```
test_acc <- (test_conM[1, 1] + test_conM[2, 2]) / sum(test_conM)
test_pre <- (test_conM[2,2]/ (test_conM[2,1] + test_conM[2,2]))
test_rec <- (test_conM[2,2]/ (test_conM[1,2] + test_conM[2,2]))
test_F1 <- 2*((test_pre*test_rec)/(test_pre+test_rec))
```

Conclusion

- Accuracy from train model is 0.6833667.
- Accuracy from train model is 0.655814.

```
cat("Train Accuracy:", train_acc,
    "\nTest Accuracy", test_acc)
```

```
## Train Accuracy: 0.6833667
## Test Accuracy 0.655814
```