Logistics Regression - Titanic Dataset

Olanrewaju Titilola

2024-02-18

Load Library packages

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(knitr)
library(ggplot2)
library(dplyr)
library(forcats)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.2
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

Set work directory

```
setwd("~/UPLIFT/Course 2/Course 2 - Assignments/Assignment 4 - Logistics Regression")
```

TRAIN DATASET

#Load datasets

```
train_data<-read.csv("train.csv", header = TRUE)
head(train_data, 5)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        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
##             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
```

Calculate and replace missing values in Age column with the mean of all ages

```
train_data$Age[is.na(train_data$Age)]=mean(train_data$Age, na.rm = TRUE)
```

```
view(train_data)
```

Load Test Dataset

```
test_data<-read.csv("test.csv", header=TRUE)
head(test_data, 5)
```

```
##   PassengerId Pclass                                          Name    Sex  Age
## 1         892      3                              Kelly, Mr. James   male 34.5
## 2         893      3              Wilkes, Mrs. James (Ellen Needs) female 47.0
## 3         894      2                     Myles, Mr. Thomas Francis   male 62.0
## 4         895      3                              Wirz, Mr. Albert   male 27.0
## 5         896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0
##   SibSp Parch  Ticket    Fare Cabin Embarked
## 1     0     0  330911  7.8292              Q
## 2     1     0  363272  7.0000              S
## 3     0     0  240276  9.6875              Q
## 4     0     0  315154  8.6625              S
## 5     1     1 3101298 12.2875              S
```

Calculate and replace missing values in Age column with the mean of all ages

```
test_data$Age[is.na(test_data$Age)]= mean(test_data$Age,na.rm = TRUE)
```

Check the types of dataset

```
str(test_data)
```

```
## 'data.frame':    418 obs. of  11 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles, Mr. Thomas
Francis" "Wirz, Mr. Albert" ...
##  $ Sex        : chr  "male" "female" "male" "male" ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : chr  "" "" "" "" ...
##  $ Embarked   : chr  "Q" "S" "Q" "S" ...
```

Create dataframe of dependent/independent variables

```
nonvars<-c("PassengerId","Name","Ticket","Cabin","Embarked")
train_data<-train_data[,!(names(train_data)%in%nonvars)]
str(train_data)
```

```
## 'data.frame':    891 obs. of  7 variables:
##  $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : chr  "male" "female" "female" "female" ...
##  $ Age     : num  22 38 26 35 35 ...
##  $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch   : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
```

Develop a model for training dataset

```
train_model<-glm(Survived~.,data= train_data,family=binomial)
summary(train_model)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial, data = train_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.960445   0.532937   9.308  < 2e-16 ***
## Pclass      -1.084297   0.139119  -7.794 6.49e-15 ***
## Sexmale     -2.762930   0.199011 -13.883  < 2e-16 ***
## Age         -0.039702   0.007797  -5.092 3.55e-07 ***
## SibSp       -0.350725   0.109552  -3.201  0.00137 **
## Parch       -0.111963   0.117400  -0.954  0.34024
## Fare         0.002852   0.002361   1.208  0.22718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  788.73  on 884  degrees of freedom
## AIC: 802.73
##
## Number of Fisher Scoring iterations: 5
```

Prediction of survival on Test dataset

```
test_data$predict <- predict(train_model, type = "response", newdata = test_data)
summary(test_data$predict)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0114  0.1125  0.2962  0.3988  0.6588  0.9663       1
```

No preference over error t = 0.5

```
test_data$survived <- as.numeric(test_data$predict>=0.5)
table(test_data$survived)
```

```
##
##   0   1
## 262 155
```

```
predictions= data.frame(test_data[c("PassengerId","survived")])
write.csv(file = "TitanicPred", x=predictions,)
head(predictions,5)
```

```
##   PassengerId survived
## 1         892        0
## 2         893        0
## 3         894        0
## 4         895        0
## 5         896        1
```

Interpretation of the Coefficients:

Pclass (Passenger Class) (-1.082896):

For each decrease in the passenger class, the estimated probability of survival decrease by about 1.08. Lower class is associated with lower chances of survival.

Sexmale (-2.763615):

Being a male reduces the estimated probability of survival by about 2.76.Being a male significantly decreases the chances of survival compared to females.

Age (-0.039746):

For each year increase in age, the estimated probability of survival decrease by about 0.04.Getting older is associated with a slight decrease in the chances of survival.

SibSp (Siblings/Spouses) (-0.351246):

For each additional sibling or spouse onboard, the estimated probability of survival decrease by about 0.35. Having more siblings or spouses onboard is associated with lower chances of survival.