Statistic Project

Sarach S.

2023-03-08

Hello Reader!

This is markdown language. I've learned about a few topics in Statistic, such as:

- Linear Regression
- Logistic Regression
- Confusion Metric
- Model training

In order to put what I've learned today into practice, I have explore into titanic data set and create some modeling to determine the modeling concept.

Let's explore what I discovered in the Titanic database together!

The goal of this study is to identify which variable are significant to case a survived of people on titanic boat by using logistic regression.

Prepare a library

```
library(titanic)
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
```

Clean data

```
## convert Sex to factor
titanic_train$Sex <- factor(titanic_train$Sex,
                    levels = c("male", "female"),
                    labels = c(1,0))

## DROP NA (missing value)
titanic_train <- na.omit(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
## 7           7        0      1
##                                                  Name Sex Age SibSp Parch
## 1                             Braund, Mr. Owen Harris   1  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer)   0  38     1     0
## 3                              Heikkinen, Miss. Laina   0  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel)   0  35     1     0
## 5                            Allen, Mr. William Henry   1  35     0     0
## 7                             McCarthy, Mr. Timothy J   1  54     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
## 7            17463 51.8625   E46        S
```

Split Data into Train Model and Test Model

```
set.seed(11)
n <- nrow(titanic_train)
id <- sample(1:n, size = n*0.7) ## 70% train 30% set
train_data <- titanic_train[id, ]
test_data <- titanic_train[-id, ]
```

Prepare Model and Summary

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

summary(logis_model)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Fare, family = "binomial",
##     data = train_data)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.7931  -0.7095  -0.4129   0.6632   2.3916
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.287501   0.635501   3.600 0.000319 ***
## Pclass      -1.128555   0.190216  -5.933 2.97e-09 ***
## Sex0         2.527385   0.246401  10.257  < 2e-16 ***
## Age         -0.038156   0.009122  -4.183 2.88e-05 ***
## Fare         0.001784   0.002987   0.597 0.550348
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 681.62  on 498  degrees of freedom
## Residual deviance: 463.66  on 494  degrees of freedom
## AIC: 473.66
##
## Number of Fisher Scoring iterations: 4
```

In terms of the variable performance the summary includes the estimates, standard errors, z-values, and p-values for each variable. Here's a breakdown of the variables and their significance:

- (Intercept): The intercept term represents the baseline log-odds of survival when all other predictor variables are zero. In this case, the estimate is 2.287501 with a standard error of 0.635501. The z-value of 3.600 indicates that the intercept term is statistically significant (p-value = 0.000319). This suggests that the baseline log-odds of survival significantly differ from zero.
- Pclass: The variable "Pclass" represents passenger class. The estimate is -1.128555, implying that as the passenger class increases, the log-odds of survival decrease. The associated standard error is 0.190216, and the z-value of -5.933 indicates that the variable is highly significant (p-value = 2.97e-09). This suggests that passenger class has a significant impact on the likelihood of survival.
- Sex: The variable "Sex0" represents the gender of the passengers. The estimate is 2.527385, indicating that being female (coded as 0) is associated with higher log-odds of survival compared to males. The standard error is 0.246401, and the z-value of 10.257 indicates that the variable is highly significant (p-value < 2e-16). This suggests that gender is a significant predictor of survival, with females having a higher likelihood of survival compared to males.
- Age: The variable "Age" represents the age of the passengers. The estimate is -0.038156, implying that as age increases, the log-odds of survival decrease. The standard error is 0.009122, and the z-value of -4.183 indicates that age is statistically significant (p-value = 2.88e-05). This suggests that age has a significant impact on the likelihood of survival.
- Fare: The variable "Fare" represents the ticket fare paid by the passengers. The estimate is 0.001784, indicating a slight positive association between fare and log-odds of survival. However, the associated standard error is 0.002987, and the z-value of 0.597 suggests that fare is not statistically significant (p-value = 0.550348). This suggests that fare may not have a significant impact on survival, at least based on the available data and model.

In summary, based on the estimates, standard errors, z-values, and p-values, we can conclude that Pclass, Sex, and Age are significant variables in predicting survival on the Titanic. The variable Fare, however, does not appear to be statistically significant in this model.

Train model Performance

```
## Train Model
p_train <- predict(logis_model, type = "response") ## probability
train_data$pred <- ifelse(p_train >= 0.5, 1, 0)

## Accuracy
## Confusion Metric
conMT <- table(train_data$pred, train_data$Survived,
               dnn = c("Predicted", "Actual"))

## Train Model Evaluation
pre_train <- conMT[2,2]/(conMT[2,1] + conMT[2,2])
rec_train <- conMT[2,2]/(conMT[1,2] + conMT[2,2])
cat("Accuracy:", (conMT[1,1] + conMT[2,2])/sum(conMT),
    "Presition:", conMT[2,2]/(conMT[2,1] + conMT[2,2]),
    "Recall:", conMT[2,2]/(conMT[1,2] + conMT[2,2]),
    "F1-score:", 2*((pre_train*rec_train)/(pre_train+rec_train)))
```

```
## Accuracy: 0.8076152 Presition: 0.7920792 Recall: 0.7476636 F1-score: 0.7692308
```

Test model Performance

```
## Test Model
p_test <- predict(logis_model, newdata = test_data,type = "response") ## probability
test_data$pred <- ifelse(p_test >= 0.5, 1, 0)

## Accuracy
## Confusion Metric
conM <- table(test_data$pred, test_data$Survived,
              dnn = c("Predicted", "Actual"))

## Test Model Evaluation
pre_test <- conM[2,2]/(conM[2,1] + conM[2,2])
rec_test <- conM[2,2]/(conM[1,2] + conM[2,2])
cat("Accuracy:", (conM[1,1] + conM[2,2])/sum(conM),
    "Presition:", conM[2,2]/(conM[2,1] + conM[2,2]),
    "Recall:", conM[2,2]/(conM[1,2] + conM[2,2]),
    "F1-score:", 2*((pre_test*rec_test)/(pre_test+rec_test)))
```

```
## Accuracy: 0.7953488 Presition: 0.6904762 Recall: 0.7631579 F1-score: 0.725
```

In terms of the test model performance, the following metrics were calculated:

- Accuracy: The accuracy of the logistic regression model on the test data is 0.7953488. This indicates that approximately 79.53% of the predictions made by the model align with the actual survival outcomes in the test dataset.
- Precision: The precision of the model is 0.6904762, which means that around 69.05% of the positive predictions (predicted survivors) made by the model are correct.
- Recall: The recall, also known as sensitivity or true positive rate, is 0.7631579. This implies that approximately 76.32% of the actual survivors in the test dataset were correctly identified by the model.
- F1-score: The F1-score is a measure that combines precision and recall into a single metric. It is calculated as 0.725 in this case. A higher F1-score indicates a better balance between precision and recall.