Lab5.R

rstudio-user

2021-02-20

```
#1.Install the package "titanic"
install.packages("titanic")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
#2.Load Titanic library to get the dataset
train <- read.csv("~/R/x86_64-pc-linux-gnu-library/4.0/titanic/data-raw/train.csv")
#View(train)
test <- read.csv("~/R/x86_64-pc-linux-gnu-library/4.0/titanic/data-raw/test.csv")
#View(test)

#3. Set Survived column for test data to NA
test$Survived <- NA

#4. Combine the  Training and Testing dataset
dataset <- rbind(train, test)
#View(dataset)

#5.Get the data structure
class(dataset)
```

```
## [1] "data.frame"
```

```
dim(dataset)
```

```
## [1] 1309   12
```

```
#6. Check for any missing values in the data
print("Null values in each column:")
```

```
## [1] "Null values in each column:"
```

```
sapply(dataset, function(x) sum(is.na(x)))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         418           0           0           0         263
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           1           0           0
```

```
print("Summary of null values")
```

```
## [1] "Summary of null values"
```

```
table(is.na(dataset))
```

```
##
## FALSE  TRUE
```

```
## 15026   682
```

```
#7.Check for any empty values
sum(dataset=="")
```

```
## [1] NA
```

```
#8.Check number of unique values for each column to find out which column we
#can convert to factors
sapply(dataset, function(x) length(unique(x)))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##        1309           3           3        1307           2          99
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           7           8         929         282         187           4
```

```
#9.Remove Cabin as it has very high missing values, passengerId, Ticket and
#Name are not required
ds <- subset(dataset, select=-c(PassengerId, Cabin, Ticket, Name))

#10.Convert "Survived","Pclass","Sex","Embarked" to factors
ds$Survived <- as.factor(ds$Survived)
ds$Pclass <- as.factor(ds$Pclass)
ds$Sex <- as.factor(ds$Sex)
ds$Embarked <- as.factor(ds$Embarked)

#11.Splitting training and test data
set = sort(sample(nrow(ds), nrow(ds)*.7))
train = ds[set,]
test = ds[-set,]

#12.Create a model
lmodel <- glm(Survived ~ Age+Sex+Pclass+Fare+Embarked+SibSp+Parch, data=ds,
              family="binomial")

#13.Visualize the model summary
summary(lmodel)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + Fare + Embarked +
##     SibSp + Parch, family = "binomial", data = ds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7220  -0.6455  -0.3770   0.6293   2.4461
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  16.691979 607.920015   0.027 0.978095
## Age          -0.043308   0.008322  -5.204 1.95e-07 ***
## Sexmale      -2.637859   0.223006 -11.829  < 2e-16 ***
## Pclass2      -1.189637   0.329197  -3.614 0.000302 ***
## Pclass3      -2.395220   0.343356  -6.976 3.04e-12 ***
## Fare          0.001451   0.002595   0.559 0.576143
## EmbarkedC   -12.259048 607.919885  -0.020 0.983911
```

```
## EmbarkedQ   -13.082427 607.920088  -0.022 0.982831
## EmbarkedS   -12.661895 607.919868  -0.021 0.983383
## SibSp        -0.362925   0.129290  -2.807 0.005000 **
## Parch        -0.060365   0.123944  -0.487 0.626233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 632.34  on 703  degrees of freedom
##   (595 observations deleted due to missingness)
## AIC: 654.34
##
## Number of Fisher Scoring iterations: 13
```

```
#14.Analyse the test of deviance using anova()
anova(lmodel, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       713     964.52
## Age       1    4.288       712     960.23  0.038392 *
## Sex       1  210.271       711     749.96 < 2.2e-16 ***
## Pclass    2  102.674       709     647.28 < 2.2e-16 ***
## Fare      1    0.054       708     647.23  0.816314
## Embarked  3    4.556       705     642.67  0.207379
## SibSp     1   10.092       704     632.58  0.001489 **
## Parch     1    0.240       703     632.34  0.624303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```