Titanic Analysis

Taras the Analyst

2023-04-09

Introduction

This is the report produced from the Kaggle notebook 'Titanic Analysis' by Taras K. from 03/18/2023.

The original inspirational source is by Hilla Behar

In this analysis the following questions were asked:

1. What is the relationship the features and a passenger's chance of survival.
2. Prediction of survival for the entire ship.

Last update: 09/04/2023 (see the list of updates at the end of this work)

Setting the environment

Packages

```
# The following packages are to be used for the current analysis
library(dplyr)          # for data manipulation
library(tidyverse)      # for working operations
library(ggplot2)        # for data visualization
library(GGally)         # Extension to 'ggplot2'
library(rpart)          # decision tree model package
library(rpart.plot)     # decision tree visualization package
library(ggcorrplot)     # to understand the correlation matrix
library(randomForest)   # planting the trees needs some methodology...:)
library(pander)         # to create pretty tables
library(knitr)          # to create pretty tables
library(tinytex)        # to use the features for file rendering to .pdf
```

Loading the data sources

```
test <- read.csv('./test.csv',stringsAsFactors = FALSE)
train <- read.csv('./train.csv', stringsAsFactors = FALSE)
dim(test)   # check the test data frame dimensions
```

```
## [1] 418  11
```

```
dim(train)  # check the train data frame dimensions
```

```
## [1] 891  12
```

Data elaboration

Merging both datasets into a consolidated one*

bind_rows() is to be used, as **rbind()** doesn't work here due to different number of columns in *train* and *test*

```
full <- bind_rows(train,test)
dim(full)  # check the resulted data frame dimensions
```

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

```
str(full)  # check the resulted data frame structure
```

```
## 'data.frame':    1309 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heik
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

The data is to be checked for missing values

```
## [1] "Here is missing value check:"
```

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

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0          NA           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0          NA        1014           2
```

So, the ouput is: N/As - left table, NULLs - right table

```
knitr::kable(list(k1, k2))
```

```
# cross-checking the empty records for Embarked
filter(full, full$Embarked == "")
```

```
##   PassengerId Survived Pclass                                        Name    Sex
## 1          62        1      1                         Icard, Miss. Amelie female
## 2         830        1      1 Stone, Mrs. George Nelson (Martha Evelyn) female
##   Age SibSp Parch Ticket Fare Cabin Embarked
## 1  38     0     0 113572   80   B28
## 2  62     0     0 113572   80   B28
```