Kaggle Competition - Titanic: Machine Learning from Disaster

Paulo Cardoso

February 20, 2016

Overview

This competition aims to Predict survival on the Titanic, in this case using R. This challenge was proposed by Kaggle, an social web hub for Data Scientists. The tragedy that hapened with the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew.

The data used in this report were provided by the organizer of competition Kaggle, and are available on this link (https://www.kaggle.com/c/titanic/data). This data is composed of 4 files, 2 major (train.csv and test.CSV) that will be used in the prediction.

Set up

Loading the external libraries and the data.

```
# Loading libraries
# library(datasets)

# Loading the data
train <- read.csv("train.csv", header = TRUE, stringsAsFactors = FALSE)
test <- read.csv("test.csv", header = TRUE, stringsAsFactors = FALSE)

# Fixing valoues and transforming to factors
#   train$Survived <- factor(train$Survived, levels=c(1,0))
#   levels(train$Survived) <- c("Survived", "Died")
#   train$Pclass <- as.factor(train$Pclass)
#   levels(train$Pclass) <- c("1st Class", "2nd Class", "3rd Class")
#   train$Gender <- factor(train$Sex, levels=c("female", "male"))
#   levels(train$Gender) <- c("Female", "Male")
```

Exploratory Data Analyses

In order to understand the data set is necessary that a few metrics be presented.

I)Presenting the struture of the train dataset.

```
# Struture presentation
str(train)
```

```
## 'data.frame':    891 obs. of  13 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "Survived","Died": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Pclass     : Factor w/ 3 levels "1st Class","2nd Class",..: 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ 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" ...
##  $ Gender     : Factor w/ 2 levels "Female","Male": 2 1 1 1 2 2 2 2 1 1 ...
```

II)Display the amount of rows.

```
# Number of rows
nrow(train)
```

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

```
nrow(test)
```

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

III)Display the first six rows of train.

```
# Head of the dataset
head(train)
```

```
##   PassengerId Survived    Pclass
## 1           1     Died 3rd Class
## 2           2 Survived 1st Class
## 3           3 Survived 3rd Class
## 4           4 Survived 1st Class
## 5           5     Died 3rd Class
## 6           6     Died 3rd Class
##                                                   Name    Sex Age SibSp
## 1                              Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                               Heikkinen, Miss. Laina female  26     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                             Allen, Mr. William Henry   male  35     0
## 6                                     Moran, Mr. James   male  NA     0
##   Parch           Ticket    Fare Cabin Embarked Gender
## 1     0        A/5 21171  7.2500                S   Male
## 2     0         PC 17599 71.2833   C85        C Female
## 3     0 STON/O2. 3101282  7.9250                S Female
## 4     0           113803 53.1000  C123        S Female
## 5     0           373450  8.0500                S   Male
## 6     0           330877  8.4583                Q   Male
```

IV)Display the last six rows of test.

```r
# Tail of the dataset
tail(test)
```

```
##     PassengerId Pclass                           Name    Sex  Age SibSp
## 413        1304      3 Henriksson, Miss. Jenny Lovisa female 28.0     0
## 414        1305      3             Spector, Mr. Woolf   male   NA     0
## 415        1306      1   Oliva y Ocana, Dona. Fermina female 39.0     0
## 416        1307      3   Saether, Mr. Simon Sivertsen   male 38.5     0
## 417        1308      3            Ware, Mr. Frederick   male   NA     0
## 418        1309      3       Peter, Master. Michael J   male   NA     1
##     Parch             Ticket     Fare Cabin Embarked
## 413     0             347086   7.7750              S
## 414     0          A.5. 3236   8.0500              S
## 415     0           PC 17758 108.9000  C105        C
## 416     0 SOTON/O.Q. 3101262   7.2500              S
## 417     0             359309   8.0500              S
## 418     1               2668  22.3583              C
```

V) Presenting the summary of the train dataset.

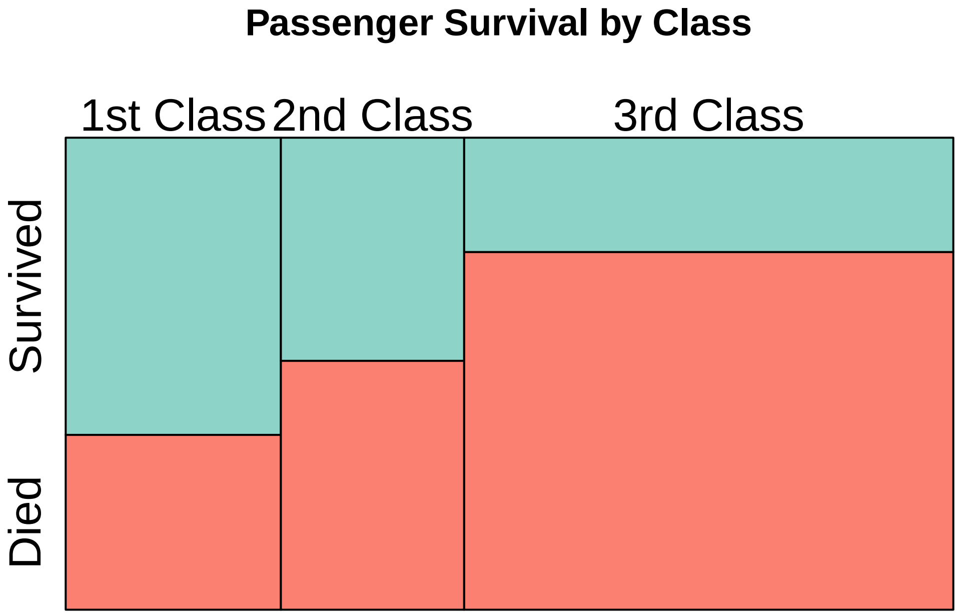
```r
# Summary
summary(train)
```

```
##   PassengerId        Survived         Pclass          Name
##  Min.   :  1.0   Survived:342   1st Class:216   Length:891
##  1st Qu.:223.5   Died    :549   2nd Class:184   Class :character
##  Median :446.0                  3rd Class:491   Mode  :character
##  Mean   :446.0
##  3rd Qu.:668.5
##  Max.   :891.0
##
##      Sex                 Age             SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
##     Gender
##  Female:314
##  Male  :577
##
##
```

```
##
##
##
```

VI) Mosaic plot presenting the distribution between Dead and Survival by Class and the probabity of survival for each person by Class.

```
# Mosaicplot
mosaicplot(train$Pclass ~ train$Survived, main="Passenger Survival by Class",
           color=c("#8dd3c7", "#fb8072"), shade=FALSE,  xlab="", ylab="",
           off=c(0), cex.axis=1.4)
```

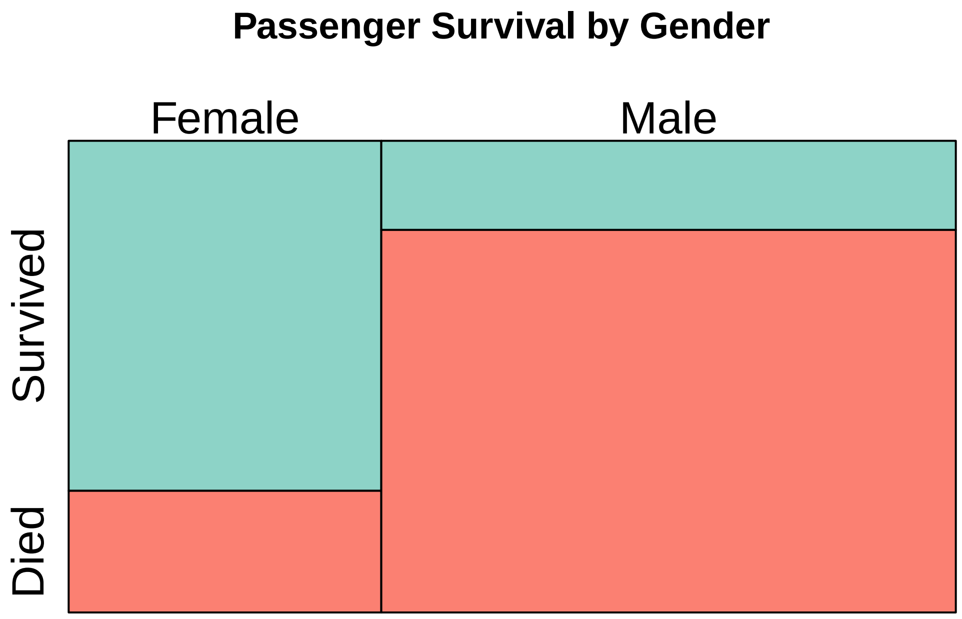


```
# Probability of sufvival by Class
prop.table(table(train$Pclass, train$Survived), 1)*100
```

```
##
##             Survived     Died
##   1st Class 62.96296 37.03704
##   2nd Class 47.28261 52.71739
##   3rd Class 24.23625 75.76375
```

VII) Mosaic plot presenting the distribution between Dead and Survival by Genderand the probabity of survival for each person by Gender.

```
# Mosaicplot
mosaicplot(train$Gender ~ train$Survived, main="Passenger Survival by Gender",
           color=c("#8dd3c7", "#fb8072"), shade=FALSE,  xlab="", ylab="",
           off=c(0), cex.axis=1.4)
```



```
# Probability of sufvival by Gender
prop.table(table(train$Sex, train$Survived), 1)*100
```

```
##
##          Survived     Died
##   female 74.20382 25.79618
##   male   18.89081 81.10919
```

Data Cleaning

On this step of the project the data used for creation of the model need to be cleaned, for example missing value variables. And the variables that will not be used are removed from the "train" dataset. The variables removed are: PassengerID, Ticket, Fare, Cabin, and Embarked.

```
# Removing variables
ctrain = train[-c(1,9:12)]
```

Replace the content of the variable Gender(Male/Female) for (0/1) in order to fit to our model.

```r
# Replacing variable Gender Value
ctrain$Sex = gsub("female", 1, train$Sex)
ctrain$Sex = gsub("^male", 0, train$Sex)
```

Then in order to fix the missing values on the Age variable we try inference this missing values assuming that Mrs.X will older than Ms.X. Moreover, we're (naively) assuming that people with the same titles are closer together in age.

```r
master_vector = grep("Master.",ctrain$Name, fixed=TRUE)
miss_vector = grep("Miss.", ctrain$Name, fixed=TRUE)
mrs_vector = grep("Mrs.", ctrain$Name, fixed=TRUE)
mr_vector = grep("Mr.", ctrain$Name, fixed=TRUE)
dr_vector = grep("Dr.", ctrain$Name, fixed=TRUE)

for(i in master_vector) {
  ctrain$Name[i] = "Master"
}
for(i in miss_vector) {
  ctrain$Name[i] = "Miss"
}
for(i in mrs_vector) {
  ctrain$Name[i] = "Mrs"
}
for(i in mr_vector) {
  ctrain$Name[i] = "Mr"
}
for(i in dr_vector) {
  ctrain$Name[i] = "Dr"
}
```

Another step in order to normalize the Age variable is replace the missing values with the average age for all passangers with the same group title.

```r
master_age = round(mean(ctrain$Age[ctrain$Name == "Master"], na.rm = TRUE), digits = 2)
miss_age = round(mean(ctrain$Age[ctrain$Name == "Miss"], na.rm = TRUE), digits =2)
mrs_age = round(mean(ctrain$Age[ctrain$Name == "Mrs"], na.rm = TRUE), digits = 2)
mr_age = round(mean(ctrain$Age[ctrain$Name == "Mr"], na.rm = TRUE), digits = 2)
dr_age = round(mean(ctrain$Age[ctrain$Name == "Dr"], na.rm = TRUE), digits = 2)

for (i in 1:nrow(ctrain)) {
  if (is.na(ctrain[i,5])) {
    if (ctrain$Name[i] == "Master") {
      ctrain$Age[i] = master_age
    } else if (ctrain$Name[i] == "Miss") {
      ctrain$Age[i] = miss_age
    } else if (ctrain$Name[i] == "Mrs") {
      ctrain$Age[i] = mrs_age
    } else if (ctrain$Name[i] == "Mr") {
      ctrain$Age[i] = mr_age
    } else if (ctrain$Name[i] == "Dr") {
      ctrain$Age[i] = dr_age
    } else {
      print("Uncaught Title")
    }
  }
}
```

Strengthening the model by creating new variables we may be able to predict the survival of the passengers even more closely. We start by creating a child variable. This is done by appending an empty column to the dataset, titled “Child”.We then populate the column with value “1”, if the passenger is under the age of 12, and “2” otherwise.

```
#ctrain["Child"]
for (i in 1:nrow(ctrain)) {
  if (ctrain$Age[i] <= 12) {
    ctrain$Child[i] = 1
  } else {
    ctrain$Child[i] = 2
  }
}
```

With the intention of determining the size of the family of each passenger by adding the number of Siblings/Spouses and Parents/Children (we add 1 so minimum becomes 1). And thereby creating a variable Familia he ought to contain the amount of families each passenger, this variable will be used in the comparison of propabilidade of survival based on the size of the family.

```
ctrain["Family"] = NA

for(i in 1:nrow(ctrain)) {
  x = ctrain$SibSp[i]
  y = ctrain$Parch[i]
  ctrain$Family[i] = x + y + 1
}
```

Another variable added to the dataset in order to enrich the quality of the information present is the variable Mother. Which is the variable that will signal whether the passenger is a mother or not, the values 1 and 2.

```
#ctrain["Mother"]
for(i in 1:nrow(ctrain)) {
  if(ctrain$Name[i] == "Mrs" & ctrain$Parch[i] > 0) {
    ctrain$Mother[i] = 1
  } else {
    ctrain$Mother[i] = 2
  }
}
```