Preprocessing Titanic Dataset

Guillem Fernández Pallarès i Miquel Tomé Carreño

21 de maig, 2021

Índex

1 Descripció del dataset

El dataset escollit ha estat *Titanic: Machine Learning from Disaster* disponible a aquest enllaç. Amb aquest joc de dades es pretenen aplicar algorismes de Machine Learning posteriors al preprocessat de les dades per crear models predictius que permetin construir un model que prevegui, en funció d'unes variables determinades, un determinat passatger sobreviurà o no al conegut accident.

El dataset descarregat es composa de 3 fitxers, els quals es troben en format `csv`.

El primer fitxer, `gender_submission.csv`, és un exemple del fitxer resultant a presentar un cop realitzat l'exercici, i conté una relació dels passatgers que van sobreviure amb dues columnes: identificador del passatger i el sexe (0 = Dona i 1 = Home).

Els altres dos fitxers contenen el conjunt de dades que ens serviran per entrenar l'algorisme (conjunt `train.csv`) i les dades de test (conjunt `test.csv`) que ens serviran per calcular el nivell de precisió en les prediccions del nostre algorisme i que seran les dades que haurem d'entregar per tal que se'ns valori en la competició de Kaggle.

En una primera inspecció, veiem que el conjunt d'entrenament consta de 891 registres, mentre que el de test en té 418. Per saber com funciona la divisió de les dades entre els conjunts train i test és bo mirar aquest vídeo.

2 Integració i selecció de dades

Les dades utilitzades es troben dividides en dos datasets diferents: un d'ells conté el subconjunt de dades que serà utilitzat al set d'entrenament o *training set* i l'altre conté aquelles que seran utilitzades al test de prova o *testing set* per comprovar l'eficàcia del model construït. Els dos subconjunts s'integraran en un de sol per dur a terme el preprocessat de les dades.

Primerament, es llegiran ambdós fitxers i s'afegirà una columna a cadascun d'ells que indicarà si un determinat registre pertany al subconjunt d'entrenament o de prova. Addicionalment, cal esmentar que al subset de prova s'ha afegit la columna `Survived` amb valors perduts `NA`, que és la variable dicotòmica a predir i que està present a l'altre subset.

```
# Lectura del training set.
r_train <- read.csv("train.csv")
summary(r_train)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      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
##
```

```
# Lectura del testing set i addició de la variable a predir.
r_test <- read.csv("test.csv")
r_test$Survived <- NA
summary(r_test)
```

```
##   PassengerId         Pclass          Name               Sex
##  Min.   : 892.0   Min.   :1.000   Length:418         Length:418
##  1st Qu.: 996.2   1st Qu.:1.000   Class :character   Class :character
##  Median :1100.5   Median :3.000   Mode  :character   Mode  :character
##  Mean   :1100.5   Mean   :2.266
##  3rd Qu.:1204.8   3rd Qu.:3.000
##  Max.   :1309.0   Max.   :3.000
##
##       Age             SibSp           Parch           Ticket
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.0000   Length:418
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :27.00   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   :30.27   Mean   :0.4474   Mean   :0.3923
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000
```

```
##  Max.   :76.00   Max.   :8.0000   Max.   :9.0000
##  NA's   :86
##       Fare            Cabin             Embarked         Survived
##  Min.   :  0.000   Length:418         Length:418         Mode:logical
##  1st Qu.:  7.896   Class :character   Class :character   NA's:418
##  Median : 14.454   Mode  :character   Mode  :character
##  Mean   : 35.627
##  3rd Qu.: 31.500
##  Max.   :512.329
##  NA's   :1
```

```
# Addició de la columna amb la classificació train-test.
r_train$train_test <- "train"
r_test$train_test <- "test"

# Concatenació dels dos subsets.
data <- rbind(r_train,
              r_test)

# Comprovació que s'ha concatenat correctament.
nrow(r_train) + nrow(r_test) == nrow(data)
```

```
## [1] TRUE
```

Pel que fa a la selecció de les dades, es trindran en compte la totalitat de registres dels quals es disposa, els quals seran considerats durant la fase de preprocessat de les dades.

3 Neteja de les dades

El següent pas a dur a terme abans de l'anàlisi de les dades, és la neteja i preprocessat d'aquestes. En els dos següents apartats es realitzarà un breu estudi per determinar si existeixen valors perduts i/o valors extrems.

3.1 Valors perduts

Primerament, s'inspeccionaran les dades de les quals es disposa amb la finalitat de trobar els valors perduts. Es consideraran tant els que tenen associat el valor `NA` com els que tenen camps en blanc. Addicionalment, es comprovarà que els `NA` introduïts manualment a la variable `Survived` en apartats anteriors es corresponen a la totalitat de valors perduts de la columna, és a dir, es comprovarà que la variable que es vol predir no contingui valors perduts al subset d'entrenament de l'algorisme.

```
# Total de valors perduts al dataset.
colSums(is.na(data))
```

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

```r
colSums(data=="")
```

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

```r
# Comprovació que les dades d'entrenament no tenen valors perduts.
sum(is.na(data$Survived[data$train_test == "train"]))
```

```
## [1] 0
```

Com es pot observar, les columnes que contenen valors perduts són `Survived` (tot i que es corresponen amb els assignats manualment), `Age`, `Fare`, `Cabin` i `Embarked`. Es tractaran diferentment en funció del seu format.

1. Els valors perduts de la variable 'Survived' es deixaran sense tractar, ja que corresponen als que han estat inserits manualment a la columna per tal de poder ajuntar el subset d'entrenament i el de prova. Posteriorment, es descartarà aquesta variable quan es torni a separar els dos subconjunts de dades.
2. A les variables 'Age' i 'Fare', ambdues numèriques, es substituiran els valors perduts per la mediana de la variable. Es tria aquesta opció perquè la mediana és una mesura de tendència central menys sensible a valors extrems que la mitjana.
3. La columna 'Embarked', corresponent al port on va embarcar el passatger en qüestió, conté una variable categòrica que no pot ser ordenada en ordre creixent o decreixent, com podria fer-se amb una variable numèrica. Per tant, en aquest cas s'associarà als valors perduts d'aquesta columna un valor desconegut 'Unknown'.

```r
# Tractament variable Age.
data$Age <- ifelse(is.na(data$Age) | data$Age == "",
                   median(data$Age, na.rm = TRUE),
                   data$Age)

# Tractament variable Fare.
data$Fare <- ifelse(is.na(data$Fare) | data$Fare == "",
                    median(data$Fare, na.rm = TRUE),
                    data$Fare)

# Tractament variable Embarked.
data$Embarked <- ifelse(is.na(data$Embarked) | data$Embarked == "",
                        "Unknown",
                        data$Embarked)

# Comprovació del canvi.
colSums(is.na(data))
```

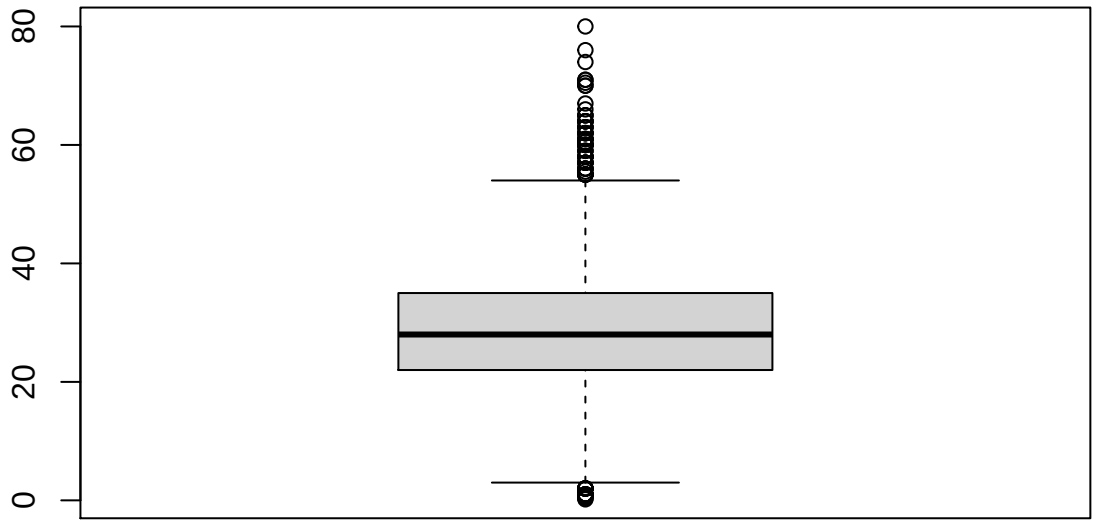
```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         418           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0         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

```r
colSums(data=="")
```

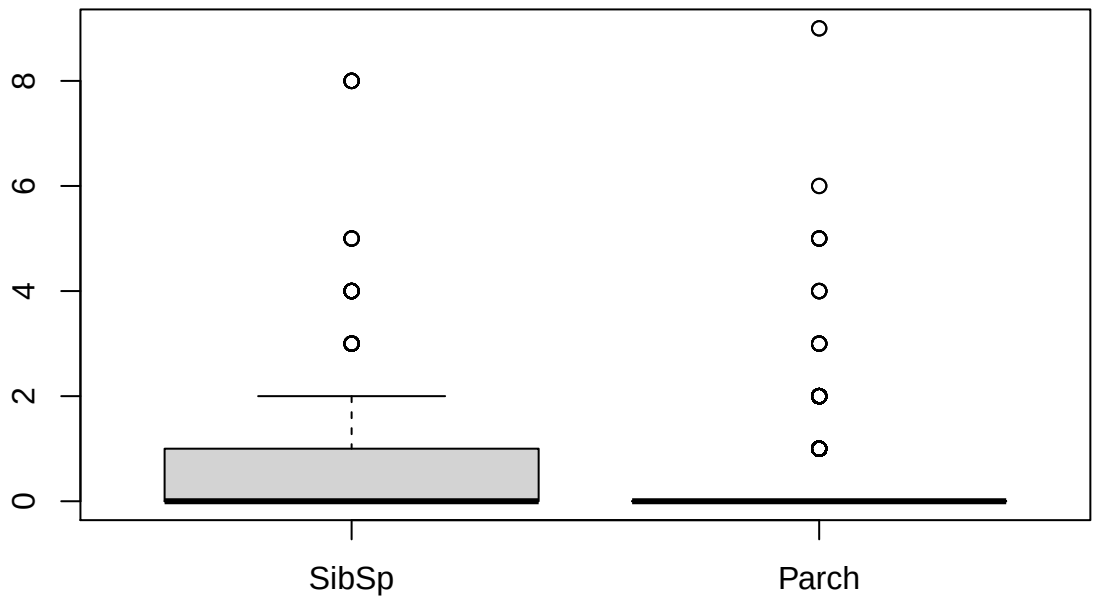
```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0          NA           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0        1014         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

3.2 Valors extrems o *outliers*

```r
# Boxplot per la variable Age.
boxplot(data$Age)
```



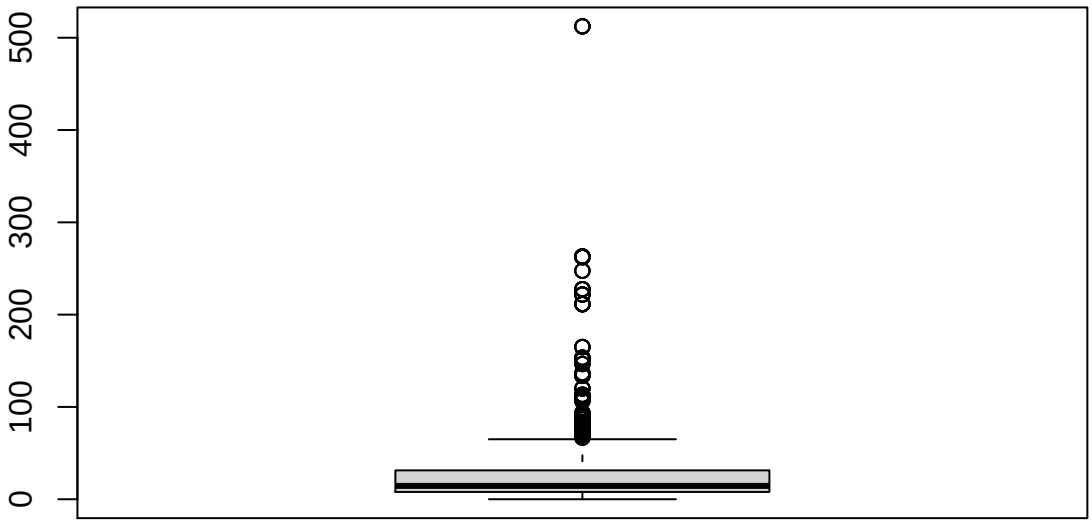

```r
# Boxplot per SibSp i Parch.
boxplot(data[c("SibSp", "Parch")])
```




```
# Taula de freqüències absolutes de Parch.
table(data$Parch)
```

```
##
##    0    1    2    3    4    5    6    9
## 1002  170  113    8    6    6    2    2
```

```
# Boxplot per Fare.
boxplot(data$Fare)
```




```
outliers_SibSp <- boxplot.stats(data$SibSp)$out
outliers_SibSp
```

```
##  [1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3 5 4 3
## [39] 4 8 4 3 4 8 4 8 3 4 5 3 4 8 4 8 4 3 3
```