Tipologia i cicle de vida de les dades

Gerard Cegarra Dueñas

December 27, 2021

1. Descripció del dataset

Per a aquest projecte s'ha escollit el dataset de la plataforma Kaggle de la competició "Titanic - Machine Learning from Disaster". Les dades estan separades en dos conjunts: entrenament i test. En el conjunt d'entrenament es troben 891 registres corresponents a passatgers que estaven al Titanic en el moment del seu naufragi. Una de les 13 variables indica si el passatger en qüestió va sobreviure a la tragedia. El conjunt de test conté 418 passatgers amb les mateixes variables però sense estar informada la variable que indica la supervivència del passatger. L'objectiu de la competició és predir amb exactitud quines de les 418 persones del conjunt de test van sobreviure i quines no. Les variables del conjunt de dades són les següents:

- `PassengerId`: identificador únic del passatger en el conjunt de dades.
- `Survival`: variable que indica la supervivència del passatger.
- `Pclass`: classe del ticket del passatger (primera, segona o tercera).
- `Sex`: sexe biològic del passatger.
- `Age`: edat del passatger.
- `SibSp`: nombre de parents del passatger a bord (del tipus conjuge/germà).
- `Parch`: nombre de parents del passatger a bord (del tipus pare/fill).
- `Ticket`: número de bitllet del passatger.
- `Fare`: preu que ha pagat el passatger per estar a bord del vaixell.
- `Cabin`: cabina assignada al passatger.
- `Embarked`: lloc d'embarcament del passatger.

2. Integració i selecció de les dades

Es mostren el resum de les dades i el tipus de cada variable. Les dades de tipus text no es transformen a factors de moment per a poder-les manipular amb més facilitat. El conjunt de dades d'entrenament i de test es combinen en un únic conjunt amb una columna addicional `Train` que indica la procedencia de cada registre.

```
train <- read.csv("data/train.csv", stringsAsFactors = FALSE)
test <- read.csv("data/test.csv", stringsAsFactors = FALSE)

train$Train <- TRUE
test$Train <- FALSE

data <- bind_rows(train, test)
summary(data)
```

```
##   PassengerId      Survived          Pclass          Name
##  Min.   :   1   Min.   :0.0000   Min.   :1.000   Length:1309
##  1st Qu.: 328   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median : 655   Median :0.0000   Median :3.000   Mode  :character
##  Mean   : 655   Mean   :0.3838   Mean   :2.295
```

```
##  3rd Qu.: 982   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :1309   Max.   :1.0000   Max.   :3.000
##                 NA's   :418
##      Sex                 Age            SibSp            Parch
##  Length:1309        Min.   : 0.17   Min.   :0.0000   Min.   :0.000
##  Class :character   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000
##  Mode  :character   Median :28.00   Median :0.0000   Median :0.000
##                     Mean   :29.88   Mean   :0.4989   Mean   :0.385
##                     3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000
##                     Max.   :80.00   Max.   :8.0000   Max.   :9.000
##                     NA's   :263
##     Ticket               Fare            Cabin
##  Length:1309        Min.   :  0.000   Length:1309
##  Class :character   1st Qu.:  7.896   Class :character
##  Mode  :character   Median : 14.454   Mode  :character
##                     Mean   : 33.295
##                     3rd Qu.: 31.275
##                     Max.   :512.329
##                     NA's   :1
##    Embarked           Train
##  Length:1309        Mode :logical
##  Class :character   FALSE:418
##  Mode  :character   TRUE :891
##
##
##
##
```

```
sapply(data, class)
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##   "integer"   "integer"   "integer" "character" "character"   "numeric"
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##   "integer"   "integer" "character"   "numeric" "character" "character"
##       Train
##   "logical"
```

Es pot observar com les variables `Cabin` i `Embarked` tenen valors buits representats per un string buit. De manera anàloga, la variable `Fare` conté valors `0` que també es poden entendre com a valors buits. Aquests valors se substitueixen per `NA`.

```
data$Cabin[data$Cabin == ""] <- NA
data$Embarked[data$Embarked == ""] <- NA
data$Fare[data$Fare == 0] <- NA
```

2.1. Creació de nous atributs

En aquest apartat es creen nous atributs a partir de la manipulació de variables ja existents en el conjunt de dades. De la variable `Name`, es pot extreure el títol de la persona (`Title`) i el seu cognom (`FamilyName`). A partir de les variables `Parch` i `SibSp`, es pot calcular el total del nombre de familiars a bord de cadascún dels passatgers (`FamilySize`). Agafant la primera lletra de la variable `Cabin`, es pot extreure la coberta a la qual estava ubicada la cabina del passatger. Finalment, a partir de la variable `Ticket` es poden extreure categories que informen sobre el venedor del tiquet del passatger.

```
get_family_name <- function(row){
  str_split(row["Name"], ", ")[[1]][1]
```

```
}
get_family_size <- function(row){
  as.numeric(row["Parch"]) + as.numeric(row["SibSp"]) + 1
}
get_title <- function(row){
  str_replace(str_split(row["Name"], "[,.]")[[1]][2], " ", "")
}
get_cabin_letter <- function(row){
  if (is.na(row["Cabin"])) return(NA)
  return(substr(row["Cabin"], 1, 1))
}
get_ticket_src <- function(row){
  ticket <- row["Ticket"]

  if (check.numeric(ticket)){
    return('N')
  }

  ticket <- trim(str_replace(ticket, "[./]", ""))

  return(substr(ticket, 1, 1))
}

data$FamilyName <- apply(data, FUN=get_family_name, MARGIN=1)
data$FamilySize <- apply(data, FUN=get_family_size, MARGIN=1)
data$Title <- apply(data, FUN=get_title, MARGIN=1)
data$CabinLetter <- apply(data, FUN=get_cabin_letter, MARGIN=1)
data$TicketSrc <- as.factor(apply(data, FUN=get_ticket_src, MARGIN=1))
```

Es mostren les primeres files del conjunt de dades per a il · lustrar els atributs creats.

```
head(data)
```

```
##   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
## 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 Train FamilyName
## 1     0        A/5 21171  7.2500  <NA>        S  TRUE     Braund
## 2     0         PC 17599 71.2833   C85        C  TRUE    Cumings
## 3     0 STON/O2. 3101282  7.9250  <NA>        S  TRUE  Heikkinen
## 4     0           113803 53.1000  C123        S  TRUE   Futrelle
## 5     0           373450  8.0500  <NA>        S  TRUE      Allen
## 6     0           330877  8.4583  <NA>        Q  TRUE      Moran
```

```
##   FamilySize Title CabinLetter TicketSrc
## 1          2    Mr        <NA>         A
## 2          2   Mrs           C         P
## 3          1  Miss        <NA>         S
## 4          2   Mrs           C         N
## 5          1    Mr        <NA>         N
## 6          1    Mr        <NA>         N
```

2.2. Discretització i normalització de les variables

A continuació s'afegeixen algunes variables noves més, fruit de la discretització dels valors de les variables `Age` i `FamilySize`. Aquestes variables seran creades com a factors per a poder usar-les en els mètodes d'anàlisi posteriors. En el cas de la variable `Pclass`, mantenir els seus valors numèrics és una opció correcte, ja que les diferents categories mantenen una relació ordenada entre elles.

```
data$AgeCat <- as.factor(cut(data$Age, c(0, 20, 60, 80), c("young", "adult", "senior")))
data$FamilySizeCat <- as.factor(cut(data$FamilySize, c(0, 1, 3, 11), c("single", "medium", "large")))
```

Quant a la variable `Title`, els títols menys representats s'afegiran a una nova categoria `Other`.

```
summary(factor(data$Title))
```

```
##         Capt          Col          Don         Dona           Dr
##            1            4            1            1            8
##     Jonkheer         Lady        Major       Master         Miss
##            1            1            2           61          260
##         Mlle          Mme           Mr          Mrs           Ms
##            2            1          757          197            2
##          Rev          Sir the Countess
##            8            1            1
```

```
data$TitleCat <- data$Title
data$TitleCat[!(data$TitleCat %in% c("Master", "Miss", "Mrs", "Mr"))] <- "Other"
data$TitleCat <- as.factor(data$TitleCat)
summary(data$TitleCat)
```

```
## Master   Miss     Mr    Mrs  Other
##     61    260    757    197     34
```

3.1. Valors buits

En primer lloc, els registres amb el valor `NA` a les variables `Cabin` i `CabinLetter` representen passatgers sense una cabina assignada. Es crea una categoria `N` per a les variables `Cabin` i `CabinLetter` que indiquen aquesta situació.

```
data$Cabin[is.na(data$Cabin)] <- 'N'
data$CabinLetter[is.na(data$CabinLetter)] <- 'N'
```

A continuació es mostra el nombre de valors buits per cada variable del dataset. Les variables `Age` i `AgeCat`, `Fare` i `Embarked` presenten valors buits, a part de la variable a predir `Survived` en el conjunt de test.

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

```
##   PassengerId      Survived        Pclass          Name           Sex
##             0           418             0             0             0
##           Age         SibSp         Parch        Ticket          Fare
##           263             0             0             0            18
##         Cabin      Embarked         Train    FamilyName    FamilySize
```

```
##             0             2             0             0             0
##         Title   CabinLetter     TicketSrc        AgeCat FamilySizeCat
##             0             0             0           263             0
##      TitleCat
##             0
```

Quant a l'edat dels passatgers, la imputació de valors es farà a la variable `AgeCat`, que és la variable que s'utilitzarà posteriorment per a l'anàlisi, i a més, a l'estar categoritzada per franges la imputació afegirà menys soroll.

Els passatgers amb l'edat no informada i que tenen el títol `Master` es poden imputar amb la categoria `young`, ja que aquest títol s'utilitza per als nens amb edat inferior o igual a 11 anys.

```
data$AgeCat[is.na(data$Age) & data$Title == "Master"] <- "young"
```

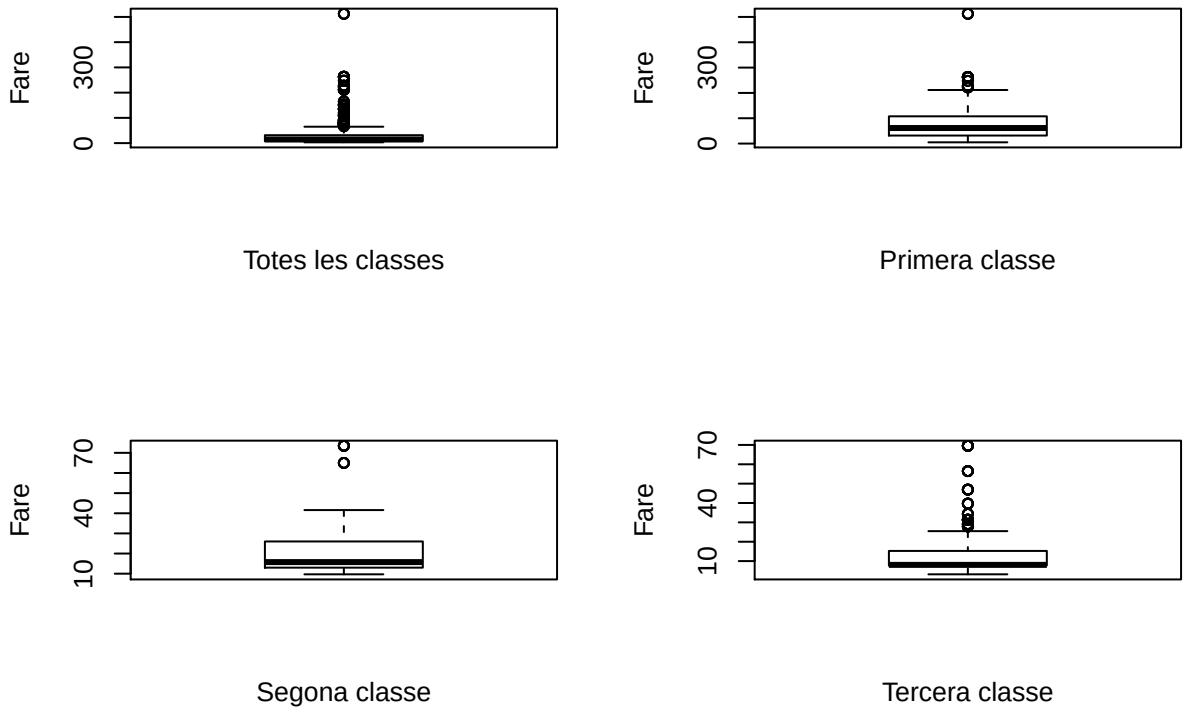
Per als valors de la variable `AgeCat` de la resta de registres, així com per a les variables `Fare` i `Embarked`, s'utilitza l'algorisme kNN com a mètode d'imputació. Les variables amb valors únics no seran utilitzades com a referència durant la imputació. Tampoc les versions originals d'aquelles variables que hagin sigut categoritzades. Es descarta també la variable `Survived`, ja que no és present a les dades de test. L'ordre d'imputació de les variables es defineix de forma ascendent pel nombre de valors buits de cada variable. Com que els valors imputats d'una variable s'utilitzen pel càlcul de la següent, seguint aquest ordre es minimitza el soroll que les variables amb més valors buits afegeixen a la resta.

```
data <- kNN(
  data,
  variable = c("Embarked", "Fare", "AgeCat"),
  dist_var = c("Pclass", "Sex", "SibSp", "Parch", "Ticket", "CabinLetter", "FamilySizeCat", "TitleCat",
  k = 10
)
```

3.2. Valors extrems

L'única variable que pot presentar valors extrems és la variable `Fare`, ja que els valors de la resta de variables numériques cauen dins un rang raonable segons el coneixement que tenim del domini. A continuació es mostra el diagrama de caixes de la variable `Fare` per a tots els registres, i també separats per classe.

```
par(mfrow=c(2, 2))
bpt <- boxplot(data$Fare, xlab="Totes les classes", ylab="Fare")
bp1 <- boxplot(data$Fare[data$Pclass == 1], xlab="Primera classe", ylab="Fare")
bp2 <- boxplot(data$Fare[data$Pclass == 2], xlab="Segona classe", ylab="Fare")
bp3 <- boxplot(data$Fare[data$Pclass == 3], xlab="Tercera classe", ylab="Fare")
```




```
length(bpt$out); length(bp1$out); length(bp2$out); length(bp3$out)
```

```
## [1] 174
```

```
## [1] 29
```

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

```
## [1] 65
```

Segons la regla intercuartílica, en prendre els valors de la variable `Fare` de tot el conjunt de dades, 174 d'aquests valors es podrien considerar atípics, ja que s'allunyen més de 3 desviacions estàndard de la mitjana. Tot i això, en prendre els valors separats per classe, el total de valors que es podrien considerar com a valors extrems es redueix a 102, distribuits en grups de 29, 12 i 65 passatgers entre la primera, segona i tercera classe respectivament. Aquests registres no s'exclouran del conjunt de dades per al posterior anàlisi perquè la desviació dels seus valors no semblen desorbitats donada la distribució de la variable i no es pot assegurar que siguin errors de mesura.

4. Anàlisi de les dades

L'objectiu d'aquest estudi és estimar la probabilitat de supervivència dels passatgers del conjunt de dades de test. També es duen a terme proves de contrast d'hipòtesis de les variables principals sobre la seva distribució al voltant de la variable `Survived`.

4.1. Selecció de les dades

De totes les variables del conjunt, s'estudiaran les següents:

- Numèriques: `Age`, `SibSp`, `Parch`, `Fare` i `FamilySize`.
- Categòriques: `Survived`, `Pclass`, `Sex`, `Embarked`, `CabinLetter`, `TicketSrc`, `AgeCat`, `FamilySizeCat` i `TitleCat`.

Els anàlisis que s'aplicaran són els següents:

- Numèriques: proves d'hipòtesis sobre la mitjana d'una variable en dues mostres.
- Categòriques: proves d'hipòtesis sobre la dependència de dues variables (`Survived` respecte les demés).

Els dos grups de variables s'utilitzaran també per a crear un model de regressió logística que predigui la variable `Survived`.

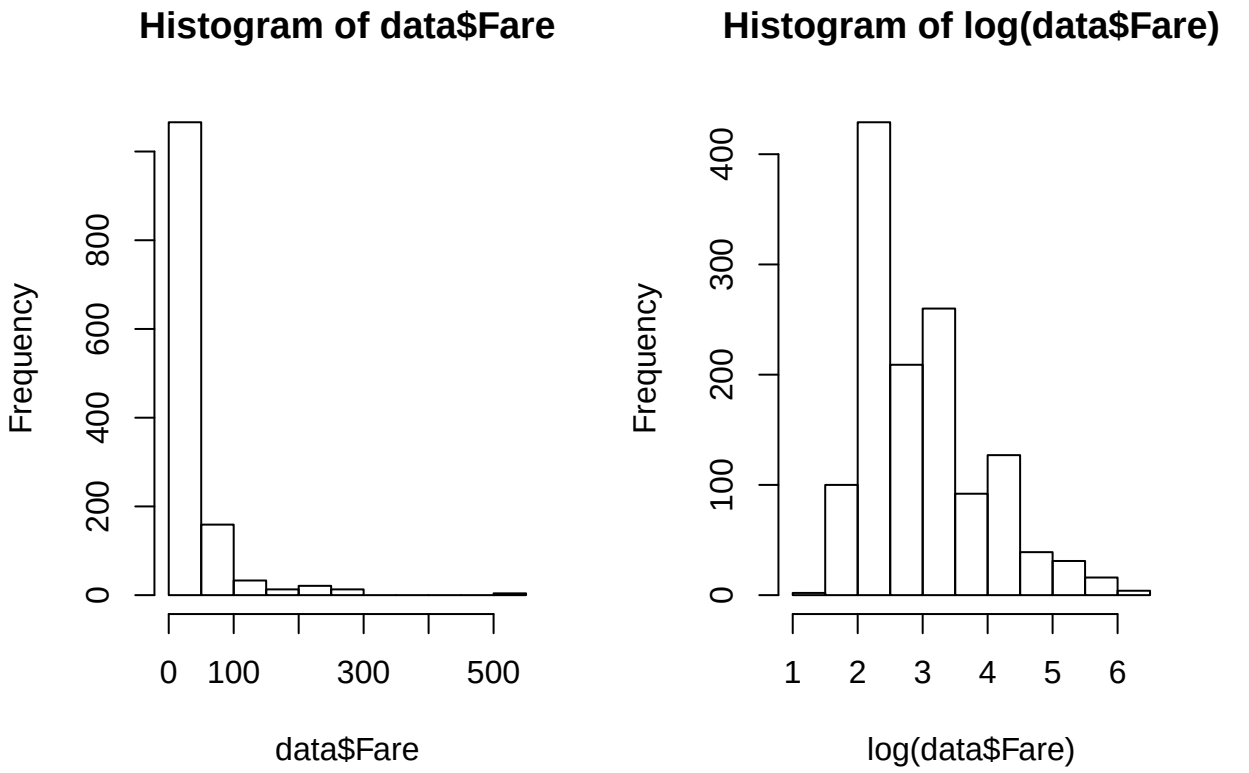
4.2. Comprovació de la normalitat i homogeneïtat de la variància

Per a poder aplicar les proves de contrast d'hipòtesis sobre les variables numèriques, cal assegurar que la mitjana mostral d'aquestes segueix una distribució normal. També és necessàri tenir en compte si al fer la separació de les dades sota els valors de la variable `Survived`, els dos grups presenten homoscedasticitat o no.

4.2.1. Comprovació de la normalitat

En aquest apartat, es comprova la normalitat de les variables numèriques. La variable `Fare` sembla presentar una distribució exponencial inversa, tal i com es pot veure a la següent visualització, i per tant, la comprovació de normalitat es durà a terme també sobre la transformació logarítmica d'aquesta variable.

```
par(mfrow=c(1,2))
hist(data$Fare)
hist(log(data$Fare))
```



El test de normalitat sobre la distribució de les dades es porta a terme mitjançant el test de Lilliefors. En tots els casos, el p-valor resultant del test és practicament 0, de manera que es pot rebutjar la hipòtesi nul · la de que els valors han sigut mesurats d'una distribució normal, per a totes les variables, amb un nivell de confiança del 95%.

Tot i que hem comprovat que no es pot assumir la distribució normal de la població de les variables numèriques, podem assumir pel Teorema Central del Límit que la mitjana mostral d'aquestes sí que tindrà una distribució normal, donat que la mida de la mostra és prou gran.

```
numVars <- c("Fare", "Age", "Parch", "SibSp", "FamilySize")

print("Lilliefors p-value:")
```

```
## [1] "Lilliefors p-value:"
```

```
cat("Log(Fare): ")
```

```
## Log(Fare):
```

```
cat(lillie.test(log(data$Fare))$p.value); cat("\n")
```

```
## 1.770086e-66
```

```
for (var in numVars){
  cat(var); cat(": ")
  cat(lillie.test(data[[var]])$p.value); cat("\n")
}
```

```
## Fare: 0
## Age: 6.670615e-17
## Parch: 0
## SibSp: 0
## FamilySize: 0
```

4.2.2. Comprovació de la homogeneïtat de la variància

Quant a la comprovació de la igualtat de variancies, es realitza el test sobre les dues mostres de cada variable al separar els seus registres pels valors de la variable `Survived`. Les mostres de les variables `Age` i `Parch` semblen presentar una variància igual amb un interval de confiança del 95%. En el cas de les variables `Fare`, `SibSp` i `FamilySize`, no hi ha homoscedasticitat. Amb aquesta informació, es realitzaran els test de contrast d'hipòtesis sobre la mitjana.

```
dataS0 <- data[data$Survived == 0, ]
dataS1 <- data[data$Survived == 1, ]

print("F test on variance p-value:")
```

```
## [1] "F test on variance p-value:"
```

```
for (var in numVars){
  cat(var); cat(": ")
  cat(var.test(dataS0[[var]], dataS1[[var]])$p.value); cat("\n")
}
```

```
## Fare: 9.931682e-55
## Age: 0.317044
## Parch: 0.1908444
## SibSp: 0
## FamilySize: 0
```

4.3. Proves estadístiques

4.3.1. Proves de contrast d'hipòtesis

A continuació es comparen les distribucions de les variables del conjunt d'entrenament sota la separació en dos grups dels registres per la variable `Survived`. Per cadascuna de les variables numèriques, es duen a terme tests d'hipòtesis sobre la mitjana poblacional dels dos grups. Per a les variables categoriques, es comprova si hi ha una diferència estadisticament significativa entre la distribució poblacional de les categories respecte els dos grups.

```
homVars <- c(F, T, T, F, F)

print("Welch Two Sample t-test p-values:")
```

```
## [1] "Welch Two Sample t-test p-values:"
```

```
i=1
for (var in numVars){
  cat(var); cat(": ")
  t <- t.test(dataS0[[var]], dataS1[[var]], var.equal=homVars[i])
  cat("p-value: "); cat(t$p.value); cat("\n")
  cat("estimate: "); cat(t$estimate); cat("\n")
  i <- i + 1
}
```

```
## Fare: p-value: 1.062184e-10
## estimate: 22.98753 48.41892
## Age: p-value: 0.03912465
## estimate: 30.62618 28.34369
## Parch: p-value: 0.01479925
## estimate: 0.3296903 0.4649123
## SibSp: p-value: 0.2326626
## estimate: 0.5537341 0.4736842
## FamilySize: p-value: 0.5853351
## estimate: 1.883424 1.938596
```

Amb un nivell de confiança del 95%, no es pot descartar la hipòtesi nul · la de que les mitjanes poblacionals de les variables `SibSp` i `FamilySize` son iguals entre els grups de passatgers que sobreviuen i els que no. Les variables `Age` i `Parch` sí que semblen mostrar una diferència significativa tot i que no podriem assumir-la amb un interval de confiança del 99%. Les mitjanes de la mostra ens indiquen que els sobrevivents són una mica més joves en mitjana i tenen més parents a bord del tipus pare/fill. En el cas de la variable `Fare`, la diferència és molt significativa amb un p-valor pròxim a 0 i unes mitjanes que indiquen que els sobrevivents van pagar, en mitjana, un preu més de dues vegades superior.

```
catVars <- c("Pclass", "Sex", "Embarked", "CabinLetter", "TicketSrc", "AgeCat", "FamilySizeCat", "Title

print("Chi squred test p-value:")
```

```
## [1] "Chi squred test p-value:"
```

```
for (var in catVars){
  cat(var); cat(": ")
  t <- table(data$Survived, as.factor(data[[var]]))
  cat(chisq.test(t)$p.value); cat("\n")
}
```

```
## Pclass: 4.549252e-23
## Sex: 1.197357e-58
## Embarked: 2.300863e-06
```

```
## CabinLetter:
## Warning in chisq.test(t): Chi-squared approximation may be incorrect
## 6.32602e-18
## TicketSrc:
## Warning in chisq.test(t): Chi-squared approximation may be incorrect
## 5.323006e-06
## AgeCat: 0.003084686
## FamilySizeCat: 8.643996e-12
## TitleCat: 4.305036e-60
```

Quant a les variables categoriques, totes semblen presentar una diferència significativa en la seva distribució amb un interval de confiança del 95%. Aquelles variables amb un p-valor més baix (`Sex` i `TitleCat`) són les que probablement influeixen més a la probabilitat de supervivència d'un passatger.

4.3.2. Regressió logística

Finalment, es construeix un model de regressió logística per a estimar la probabilitat de supervivència a partir de les variables seleccionades. Els registres del conjunt d'entrenament se separen en primer lloc en dos subconjunts d'entrenament i test per a avaluar la bondat del model construit. A partir de la regressió construida amb les dades del subconjunt d'entrenament, s'avalua l'exactitud del model per a diferents thresholds de decisió sobre les dades del subconjunt test. Dels valors provats, un threshold de 0.5 és el que dona el millor resultat amb una exactitud de 0.8715084. Es pot comprovar a la corba ROC com aquest threshold és una de les millors opcions en quant al balanç entre els ràtios de positius vertaders i falsos.

També s'observa als coeficients del model com les variables categòriques amb un p-valor més baix en l'anàlisi de comparació de la distribució (`Sex` i `TitleCat`), són de les que tenen valors més significatius a l'hora de predir la probabilitat de supervivència. Es penalitza la categoria `male` de la variable `Sex` i totes les categories de la variable `TitleCat` tret de la categoria `Master`.

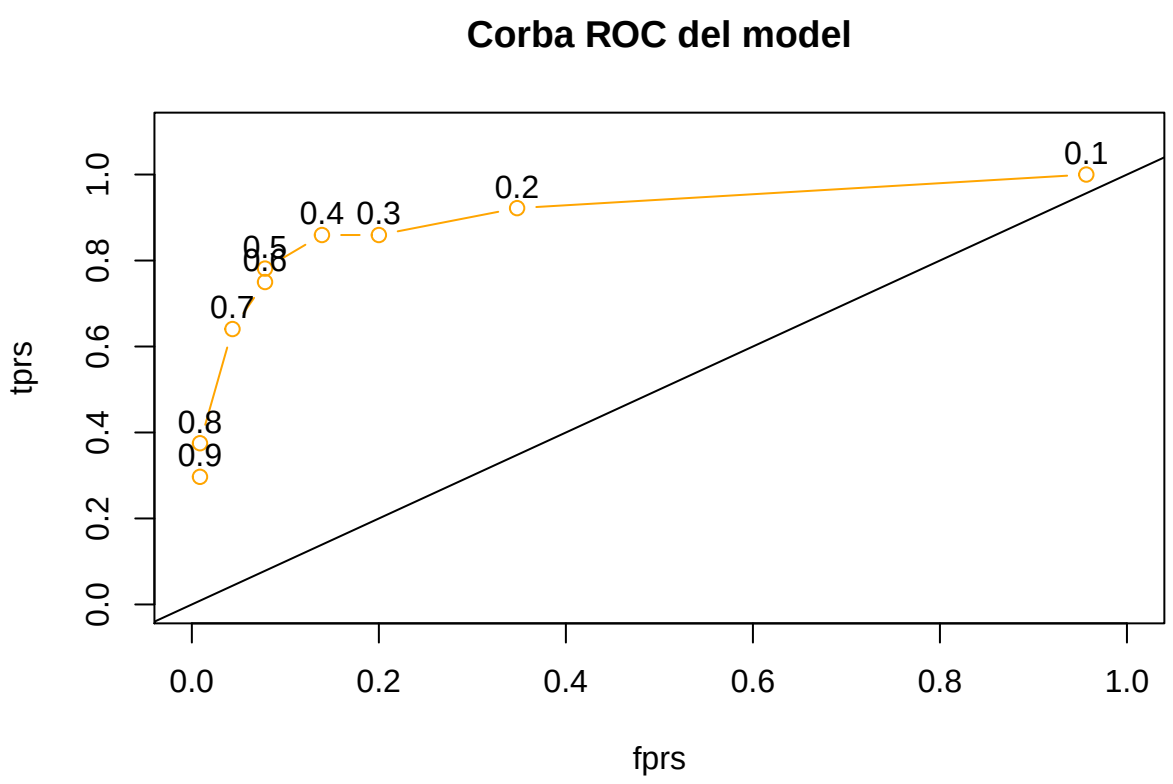
```
train <- data[data$Train == TRUE,]
test <- data[data$Train == FALSE,]

n <- nrow(train)
ntrain <- round(n * 0.8)
ntest <- n - ntrain

t_train <- train[1:ntrain,]
t_test <- train[ntrain:n,]

detach()
attach(t_train, warn.conflicts = FALSE)
model <- glm(Survived ~ Fare + Parch + SibSp + Sex + Embarked + CabinLetter + TicketSrc + AgeCat + Famil

summary(model)
prob <- predict(model, t_test, type="response")
ground <- t_test$Survived == 1

tprs <- c()
fprs <- c()
ts <- (1:9)/10
```

```
for (t in ts){
  pred <- prob > t
  acc <- sum(pred == ground) / length(pred)
```

```
  tpr <- sum(pred == TRUE & ground == TRUE) / sum(ground == TRUE)
  fpr <- sum(pred == TRUE & ground == FALSE) / sum(ground == FALSE)
  tprs <- append(tprs, tpr); fprs <- append(fprs, fpr)
  cat(t); cat(": "); cat(acc); cat("\n")
}
```

```
## 0.1: 0.3854749
## 0.2: 0.7486034
## 0.3: 0.8212291
## 0.4: 0.8603352
## 0.5: 0.8715084
## 0.6: 0.8603352
## 0.7: 0.8435754
## 0.8: 0.7709497
## 0.9: 0.7430168
```

```
plot(fprs, tprs, type="b", col="orange", ylim=c(0,1.1), xlim=c(0,1), main="Corba ROC del model")
abline(a=c(0,0), b=c(1,1))
text(fprs, tprs + 0.05, labels=round(ts, 2))
```




```
print(model$coefficients)
```

```
##          (Intercept)                 Fare                Parch
##         1.4598488442         0.0008211759        -0.0530306502
##                SibSp              Sexmale            EmbarkedQ
##        -0.0771888204        -0.5442022596        -0.0080737358
##            EmbarkedS         CabinLetterB         CabinLetterC
```

```
##       -0.0178491779       -0.0644736210       -0.1631754645
##        CabinLetterD        CabinLetterE        CabinLetterF
##        0.0013842384        0.0354004295       -0.0779481111
##        CabinLetterG        CabinLetterN        CabinLetterT
##       -0.4919719091       -0.2965495696       -0.4289055805
##          TicketSrcC          TicketSrcF          TicketSrcL
##        0.1977153808        0.1964877403        0.2133640027
##          TicketSrcN          TicketSrcP          TicketSrcS
##        0.0859940280        0.1116271853        0.1865276193
##          TicketSrcW         AgeCatadult         AgeCatsenior
##       -0.0996531666       -0.0512953276       -0.1635167413
## FamilySizeCatmedium  FamilySizeCatlarge        TitleCatMiss
##        0.0737293081       -0.0212938681       -0.5355621485
##          TitleCatMr         TitleCatMrs       TitleCatOther
##       -0.5327422694       -0.4708176109       -0.4274633412
```

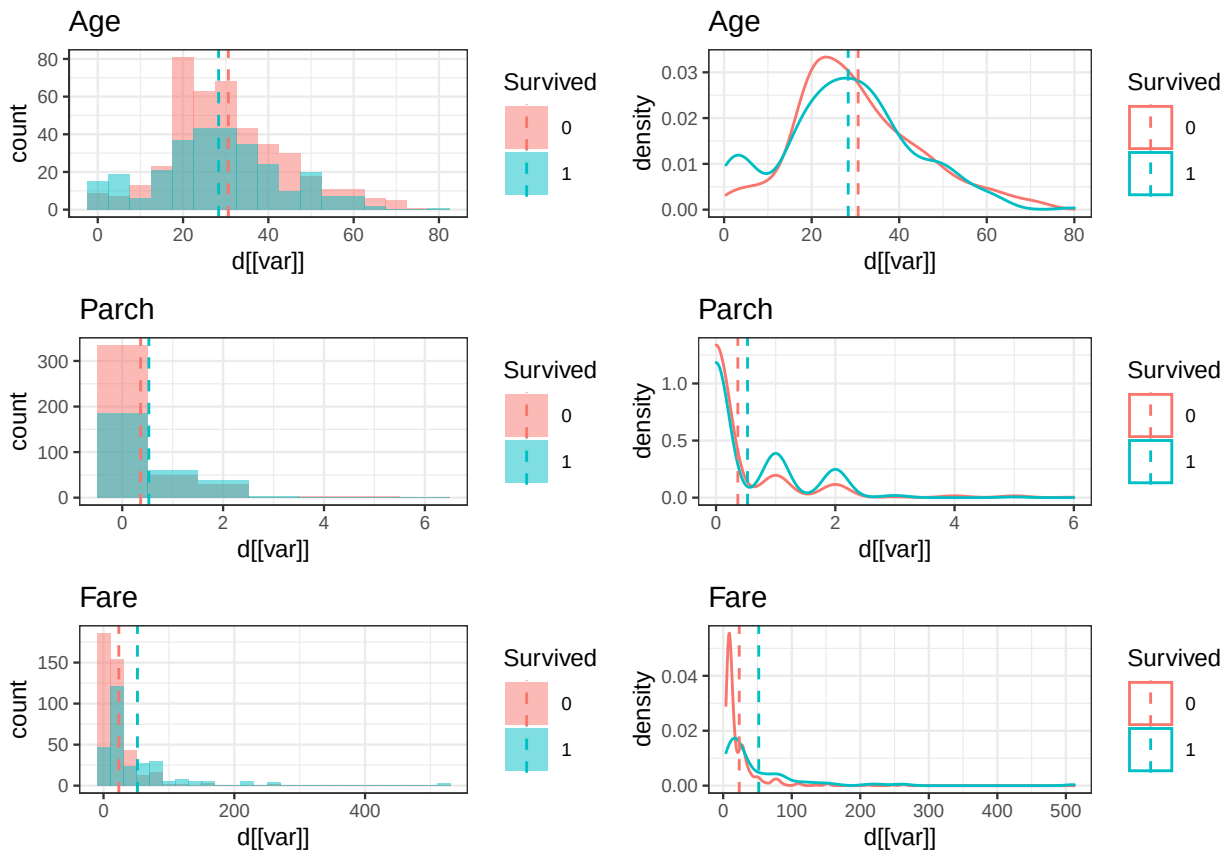
Per acabar d'avaluar la bondat de l'ajustament de la regressió, un nou model de regressió logística és entrenat sobre tot el conjunt de dades d'entrenament i es prediu el valor de la variable `Survived` del conjunt de test (del qual desconeixem els valors reals). Aquest resultat obté una exactitud de 0.77751 a la plataforma Kaggle.

```
apply_threshold <- function(row){
  if (as.numeric(row["Prob"]) > 0.5) return(1)
  return(0)
}

detach()
attach(train, warn.conflicts = FALSE)

model <- glm(Survived ~ Fare + Parch + SibSp + Sex + Embarked + CabinLetter + TicketSrc + AgeCat + Famil

test$Prob <- predict(model, test, type="response")
test$Survived <- apply(test, FUN=apply_threshold, MARGIN=1)

write.csv(test[, c("PassengerId", "Survived")], "data/result.csv", row.names = FALSE, quote = FALSE)
```

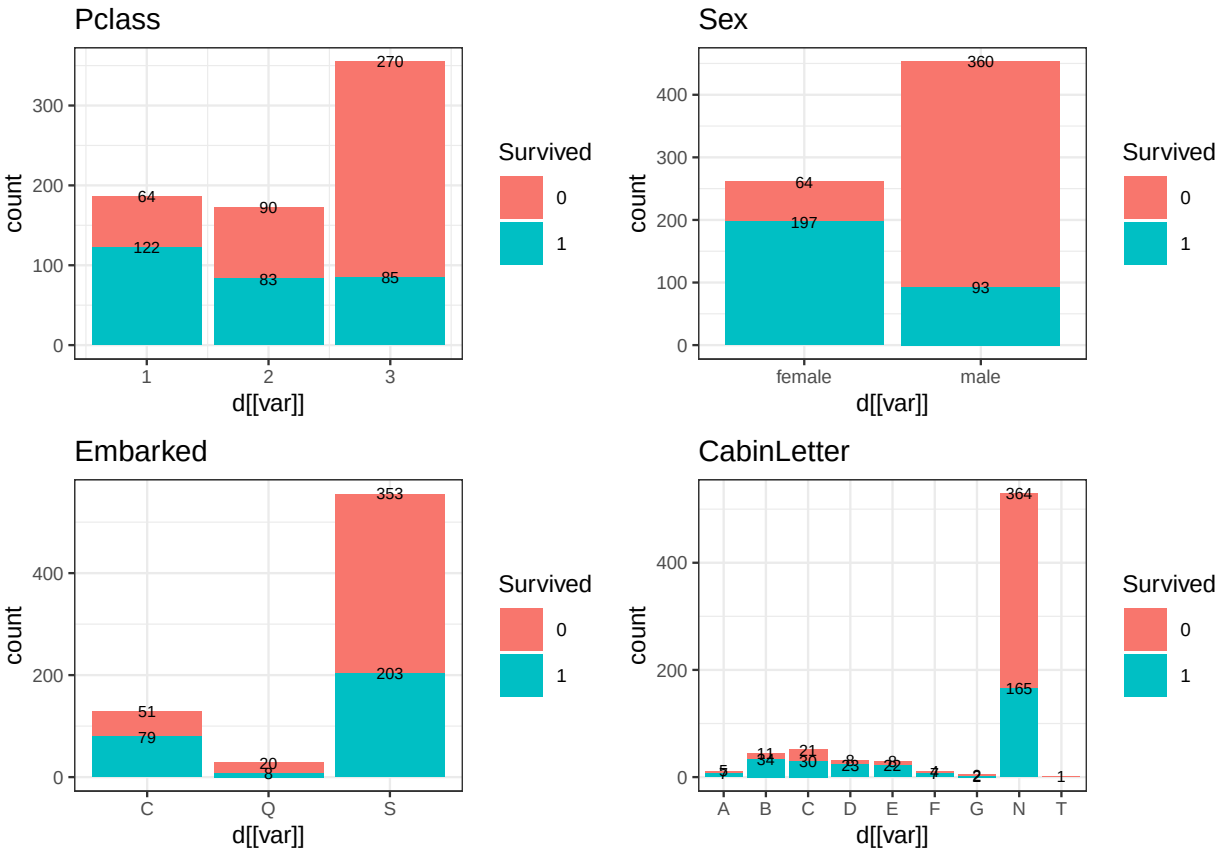
5. Representació dels resultats

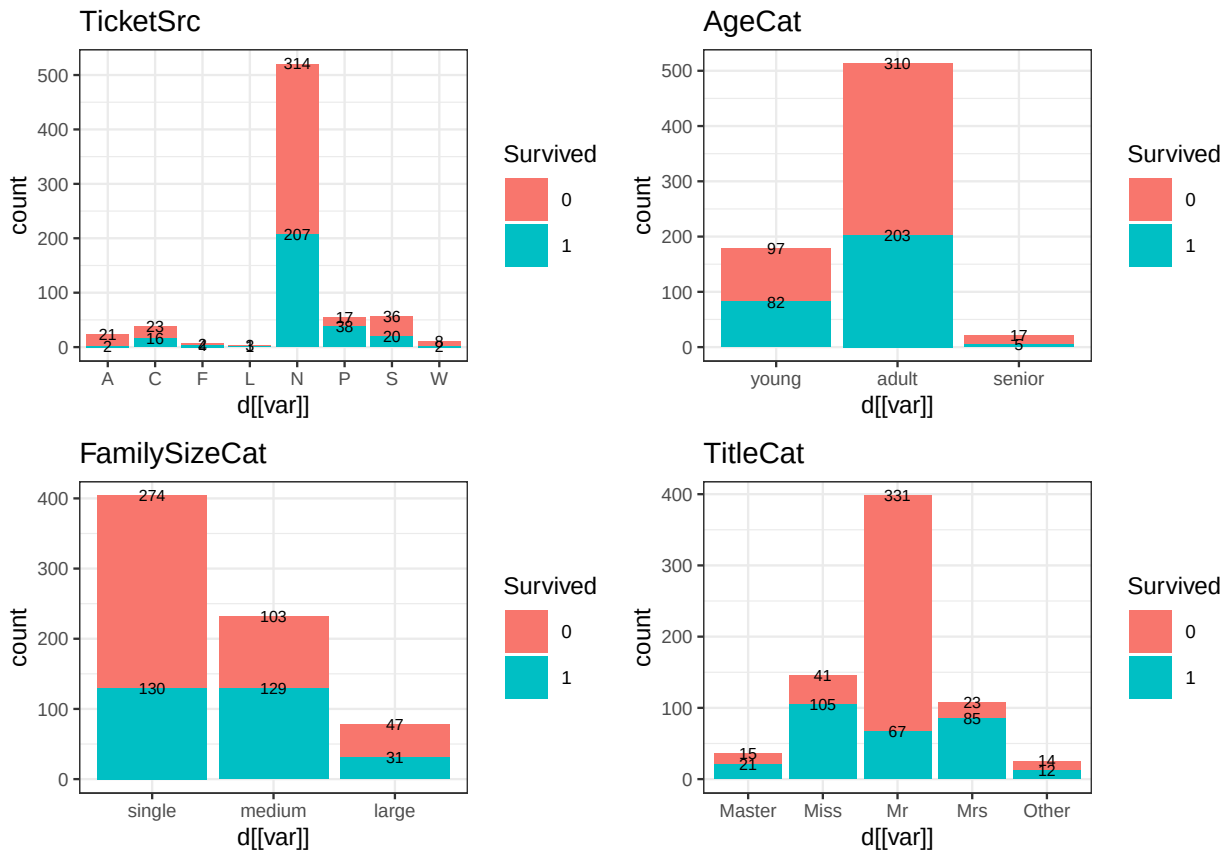
En aquest apartat es visualitzen els resultats obtinguts en les anàlisis de les dades de l'apartat anterior.

Sobre les variables numèriques, s'ha pogut comprobar com per algunes d'aquestes (`Age`, `Parch` i `Fare`) la seva mitjana és diferent segons si es mesura sobre la mostra de passatgers sobrevivents o no sobrevivents. A continuació es visualitza la distribució d'aquestes variables en ambdues mostres.



Quant a les variables categoriques, totes elles presenten una distribució diferent segons si s'observen per a la mostra de sobrevivents o no sobrevivents. A les seguents visualitzacions es mostren les distribucions d'aquestes variables per les dues mostres.


Pclass
count
d[[var]]
1
2
3
64
122
90
83
270
85
Survived
0
1
Sex
female
male
64
197
360
93
Embarked
C
Q
S
51
79
20
8
353
203
CabinLetter
A
B
C
D
E
F
G
N
T
5
7
11
34
21
30
8
23
9
22
4
2
364
165
1



6. Conclusions

De les proves de contrast i les visualitzacions sobre les variables es poden extreure diverses conclusions:

- `AgeCat`: els passatgers entre 0 i 20 anys tenen més probabilitat de sobreviure mentre que els majors de 60 són els que ho tenen més difícil.
- `Parch` i `FamilySizeCat`: les persones que viatjen soles tenen menys probabilitat de supervivència. Les families d'entre 2 i 4 persones tenen més probabilitat de supervivència, així com les persones que viatjen amb entre 1 i 3 familiars del tipus pare/fill.
- `Fare`, `Pclass` i `CabinLetter`: els passatgers de primera classe o que han pagat més pel seu bitllet tenen més probabilitat de spuervivència. A més, les persones que viatjen en una cabina sobreviuen més que la resta.
- `Sex` i `TitleCat`: els homes adults tenen menys probabilitat de supervivència que les dones. Les dones casades tenen una mica més de probabilitat de sobreviure que les que no ho estan.

D'altra banda, el model de regressió logística construit sobre les dades preprocessades aconsegueix predir correctament el 77.75% del conjunt de test. Es pot considerar un resultat prou satisfactori donat que les dades contenen un soroll inherent al problema que no és possible predir amb les variables de les que es disposen. L'exactitud del model demostra com els factors estudiats tenen una influència real en les probabilitats de supervivència dels passatgers del Titànic, reflexant els comportaments i valors estructurals de la societat de l'època.