

Tipología y ciclo de vida de los datos. Práctica 2

Baltasar Boix / Yago Ezcurra

13/5/2021

Contents

Titanic - Machine Learning from Disaster. Kaggle competition.	2
Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?	2
Lectura y análisis previo del dataset.	2
Integración y selección de los datos de interés a analizar.	4
Limpieza de los datos.	6
¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?	6
Identificación y tratamiento de valores extremos.	7
Análisis de los datos.	7
Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).	7
Comprobación de la normalidad y homogeneidad de la varianza.	7
Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.	
Aplicar al menos tres métodos de análisis diferentes.	7
Representación de los resultados a partir de tablas y gráficas.	8
Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones?	
¿Los resultados permiten responder al problema?	8

Titanic - Machine Learning from Disaster. Kaggle competition.

```
require(tidyverse)
require(laers)
require(GGally)
require(knitr)
require(kableExtra)
require(gridExtra)
require(DescTools)
```

Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Lectura y análisis previo del dataset.

```
df <- read_csv('../data/train.csv')

df$Survived <- factor(df$Survived)
df$Pclass <- factor(df$Pclass)
df$Sex <- factor(df$Sex)
df$SibSp <- factor(df$SibSp)
df$Parch <- factor(df$Parch)
df$Embarked <- factor(df$Embarked)

summary(df)
```

```
## PassengerId  Survived  Pclass    Name      Sex
## Min.   : 1.0      0:549    1:216  Length:891  female:314
## 1st Qu.:223.5    1:342    2:184  Class :character  male :577
## Median :446.0          3:491  Mode  :character
## Mean   :446.0
## 3rd Qu.:668.5
## Max.   :891.0
##
##      Age      SibSp  Parch      Ticket           Fare
## Min.   : 0.42    0:608    0:678  Length:891  Min.   : 0.00
## 1st Qu.:20.12    1:209    1:118  Class :character  1st Qu.: 7.91
## Median :28.00    2: 28    2: 80  Mode  :character  Median : 14.45
## Mean   :29.70    3: 16    3: 5           Mean   : 32.20
## 3rd Qu.:38.00    4: 18    4: 4           3rd Qu.: 31.00
## Max.   :80.00    5: 5     5: 5           Max.   :512.33
## NA's   :177      8: 7     6: 1
##      Cabin      Embarked
## Length:891      C   :168
## Class :character  Q   : 77
## Mode  :character  S   :644
##                  NA's: 2
##
```

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

pclass: A proxy for socio-economic status (SES)

1st = Upper

2nd = Middle

3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...

Sibling = brother, sister, stepbrother, stepsister

Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

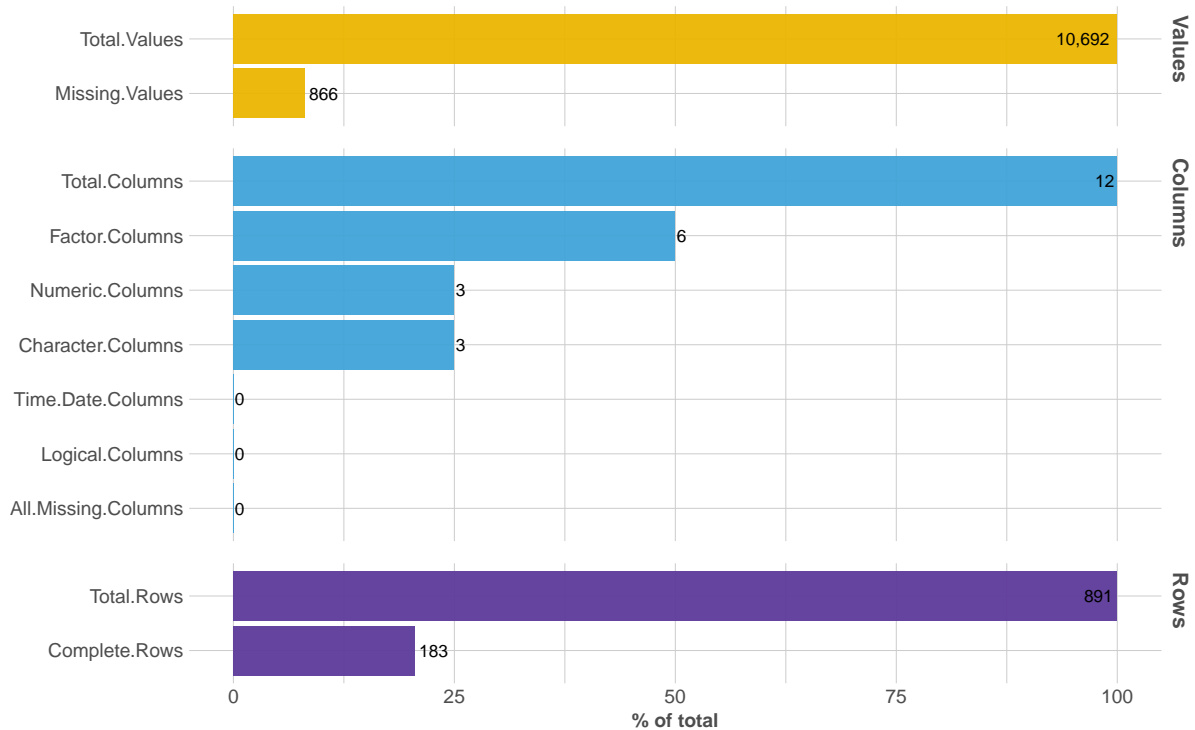
Some children travelled only with a nanny, therefore parch=0 for them.

Figure 1: Descripción del dataset obtenido en Kaggle

```
##
```

```
df_str(df)
```

Dataset overall structure



Memory Usage: 0.18 Mb

Creamos la variable dicotómica `Child` para diferenciar los niños de los adultos (>12 años).

Creamos la variable `n_ticket` con el número de personas que viajan con el mismo ticket.

Separamos del `Name` el título (`title_name`) y el primer apellido (`first_name`).

Simplificamos `title_name` en cuatro niveles.

Integración y selección de los datos de interés a analizar.

```
df <- df %>%
  mutate(Child=factor(Age<=12))

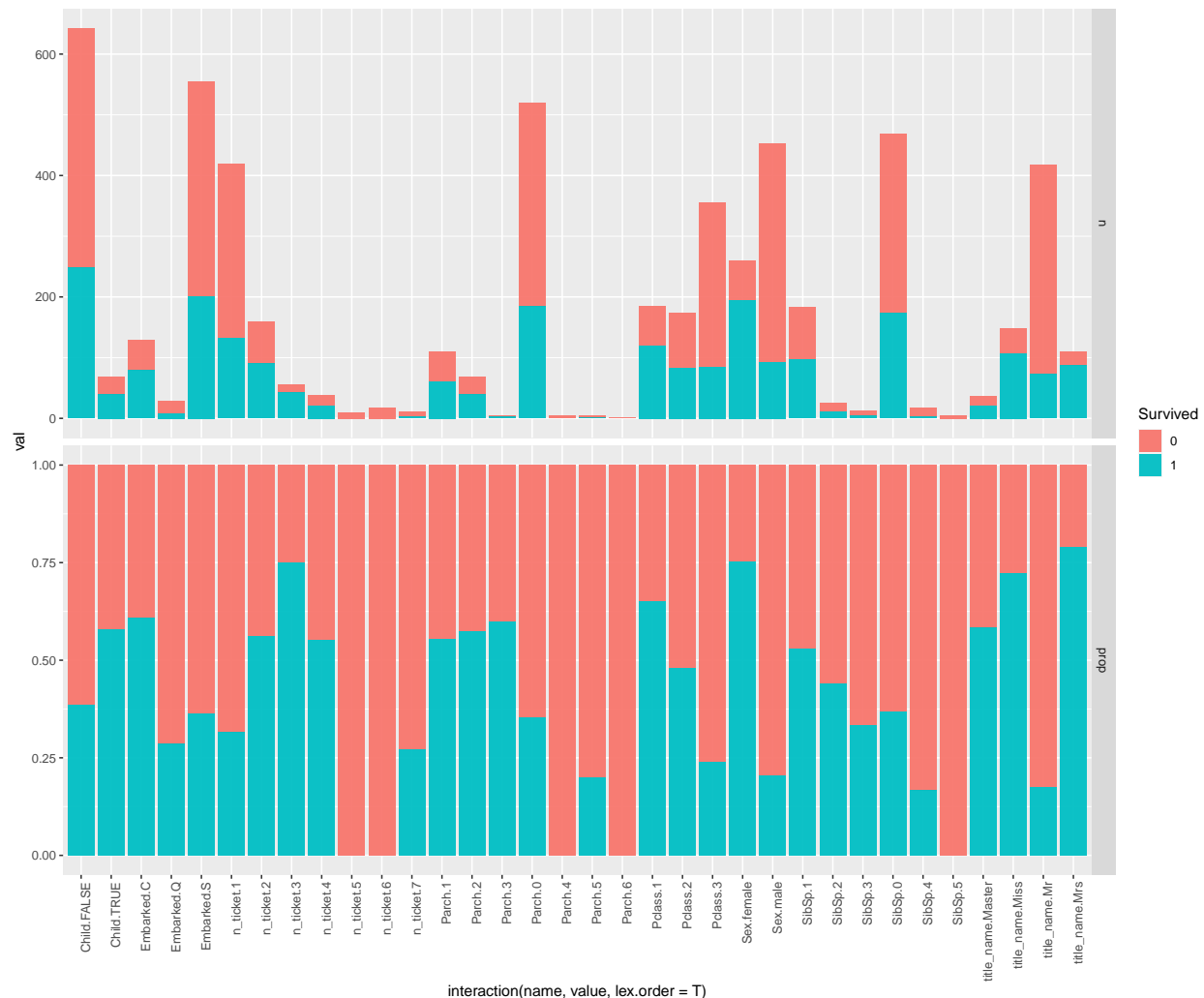
df <- left_join(df, df %>%
  group_by(Ticket) %>%
  summarize(n_ticket=n()) %>%
  mutate(n_ticket=factor(n_ticket))

df <- df %>%
  separate(Name, c('first_name', 'rest_name'), sep=', ', remove=F) %>%
  separate(rest_name, c('title_name', 'rest_name'), sep='\\.\\.') %>%
  select(-rest_name)

df$title_name[df$title_name %in% c('Capt', 'Col', 'Don', 'Dr', 'Jonkheer', 'Major', 'Rev', 'Sir')] <- 'I'
df$title_name[df$title_name %in% c('Lady', 'Mme', 'the Countess')] <- 'Mrs'
```

```
df$title_name[df$title_name %in% c('Mlle', 'Ms')] <- 'Miss'
df$title_name <- factor(df$title_name)

df %>%
  select(where(is.factor)) %>%
  na.omit() %>%
  pivot_longer(-Survived) %>%
  group_by(name, value, Survived) %>%
  summarize(n=n()) %>%
  mutate(prop=prop.table(n)) %>%
  pivot_longer(c(n,prop), names_to='tipo', values_to='val') %>%
  ggplot(aes(x=interaction(name,value, lex.order = T), y=val, fill=Survived)) +
  geom_bar(stat='identity', position='stack') +
  facet_grid(tipo ~ ., scale='free_y') +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1))
```



```
df %>%
  select(Survived, where(is.numeric), -PassengerId) %>%
  ggpairs(aes(color=Survived))
```



Limpieza de los datos.

¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Sustituimos los NA's de la variable Child con los siguientes criterios:

Asumimos que los viajeros con ticket unipersonal con SibSp==0 (sin hermanos o esposa a bordo) y Parch==0 (sin hermanos o padres a bordo) no son niños.

Asumimos que las personas con title_name=='Mrs' (mujeres casadas) no son niños.

Asumimos que las personas con title_name=='Master' son niños.

Asumimos que las personas con SibSp>0 y Parch>0 son niños.

```
df <- df %>%
  mutate(Child=if_else(SibSp==0 & Parch==0 & is.na(Child) & n_ticket==1, FALSE, as.logical(Child))) %>%
  mutate(Child=if_else(title_name=='Master' & is.na(Child), TRUE, as.logical(Child))) %>%
  mutate(Child=if_else(title_name=='Mrs' & is.na(Child), FALSE, as.logical(Child))) %>%
  mutate(Child=if_else(SibSp > '0' & Parch > '0' & is.na(Child), TRUE, as.logical(Child))) %>%
  mutate(Child=if_else(is.na(Child), FALSE, as.logical(Child)))
```

```
prop.table(table(df$Survived, df$Child, dnn=c('Survived', 'Child')), margin=2)
```

```
##           Child
## Survived  FALSE    TRUE
##           0 0.6332518 0.4246575
##           1 0.3667482 0.5753425
```

Identificación y tratamiento de valores extremos.

Análisis de los datos.

Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Comprobación de la normalidad y homogeneidad de la varianza.

Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

```
df %>%
  select(Survived, where(is.factor)) %>%
  pivot_longer(-Survived) %>%
  group_by(name) %>%
  summarize(phi=Phi(x=Survived, y=factor(value)),
            chisq.pvalue=chisq.test(Survived, factor(value))$p.value) %>%
  mutate(`signif 95%`=chisq.pvalue < 0.05) %>%
  arrange(chisq.pvalue) %>%
  kable(format='latex', digits=4, caption='Phi y chisq.test') %>%
  kable_styling(full_width = F, latex_options = "HOLD_position")
```

Table 1: Phi y chisq.test

name	phi	chisq.pvalue	signif 95%
title_name	0.5708	0e+00	TRUE
Sex	0.5434	0e+00	TRUE
Pclass	0.3398	0e+00	TRUE
n_ticket	0.3250	0e+00	TRUE
SibSp	0.2045	0e+00	TRUE
Embarked	0.1726	0e+00	TRUE
Parch	0.1770	1e-04	TRUE

```
df %>%
  select(Survived, where(is.numeric), -PassengerId) %>%
  pivot_longer(-Survived) %>%
  group_by(name, Survived) %>%
  summarize(value_list = list(value)) %>%
  pivot_wider(names_from=Survived, values_from=value_list) %>%
  mutate(vartest.pval=var.test(unlist(`0`), unlist(`1`))$p.val) %>%
```

```
mutate(ttest.pval=t.test(unlist(`0`), unlist(`1`), var.equal=var.test.pval>0.05)$p.val) %>%
mutate(`signif 95%`=ttest.pval < 0.05) %>%
select(`0`, -`1`) %>%
kable(format='latex', digits=4, caption='t.test') %>%
kable_styling(full_width = F, latex_options = "HOLD_position")
```

Table 2: t.test

name	var.test.pval	ttest.pval	signif 95%
Age	0.317	0.0391	TRUE
Fare	0.000	0.0000	TRUE

Representación de los resultados a partir de tablas y gráficas.

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?