Maestría en Ciencia de Datos

Balanceo de Datos

Paúl Arévalo

2024-07-04

Balanceo de Datos

Carga de datos y visualización de la estructura

```
library(readr)
datos <- read_delim("Datos.csv",
  delim = ";",
  escape_double = FALSE, trim_ws = TRUE
)
```

```
## Rows: 3333 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (2): Plan_Internacional, Desafiliado
## dbl (4): Min_En_Dia, Min_Internacionales, Reclamos, Llamadas_Internacionales
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(datos)
```

```
## # A tibble: 6 x 6
##   Plan_Internacional Min_En_Dia Min_Internacionales Reclamos
##   <chr>                   <dbl>               <dbl>    <dbl>
## 1 no                       265.                10          1
## 2 no                       162.                13.7        1
## 3 no                       243.                12.2        0
## 4 si                       299.                 6.6        2
## 5 si                       167.                10.1        3
## 6 si                       223.                 6.3        0
## # i 2 more variables: Llamadas_Internacionales <dbl>, Desafiliado <chr>
```

```
summary(datos)
```

```
##  Plan_Internacional   Min_En_Dia    Min_Internacionales    Reclamos
##  Length:3333        Min.   :  0.0   Min.   : 0.00       Min.   :0.000
##  Class :character   1st Qu.:143.7   1st Qu.: 8.50       1st Qu.:1.000
##  Mode  :character   Median :179.4   Median :10.30       Median :1.000
##                     Mean   :179.8   Mean   :10.24       Mean   :1.563
##                     3rd Qu.:216.4   3rd Qu.:12.10       3rd Qu.:2.000
##                     Max.   :350.8   Max.   :20.00       Max.   :9.000
```

```
##  Llamadas_Internacionales Desafiliado
##  Min.   : 0.000           Length:3333
##  1st Qu.: 3.000           Class :character
##  Median : 4.000           Mode  :character
##  Mean   : 4.479
##  3rd Qu.: 6.000
##  Max.   :20.000
```

```
str(datos)
```

```
## spc_tbl_ [3,333 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Plan_Internacional      : chr [1:3333] "no" "no" "no" "si" ...
##  $ Min_En_Dia              : num [1:3333] 265 162 243 299 167 ...
##  $ Min_Internacionales     : num [1:3333] 10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
##  $ Reclamos                : num [1:3333] 1 1 0 2 3 0 3 0 1 0 ...
##  $ Llamadas_Internacionales: num [1:3333] 3 3 5 7 3 6 7 6 4 5 ...
##  $ Desafiliado             : chr [1:3333] "no" "no" "no" "no" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Plan_Internacional = col_character(),
##   ..   Min_En_Dia = col_double(),
##   ..   Min_Internacionales = col_double(),
##   ..   Reclamos = col_double(),
##   ..   Llamadas_Internacionales = col_double(),
##   ..   Desafiliado = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Exploración de cada variable

```
skimr::skim(datos)
```

Table 1: Data summary

Name	datos
Number of rows	3333
Number of columns	6
Column type frequency:	
character	2
numeric	4
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Plan_Internacional	0	1	2	2	0	2	0
Desafiliado	0	1	2	2	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Min_En_Dia	0	1	179.78	54.47	0	143.7	179.4	216.4	350.8	
Min_Internacionales	0	1	10.24	2.79	0	8.5	10.3	12.1	20.0	
Reclamos	0	1	1.56	1.32	0	1.0	1.0	2.0	9.0	
Llamadas_Internacionales	0	1	4.48	2.46	0	3.0	4.0	6.0	20.0	

Verificar si existe un desbalance en los datos

```
table(datos$Desafiliado)
```

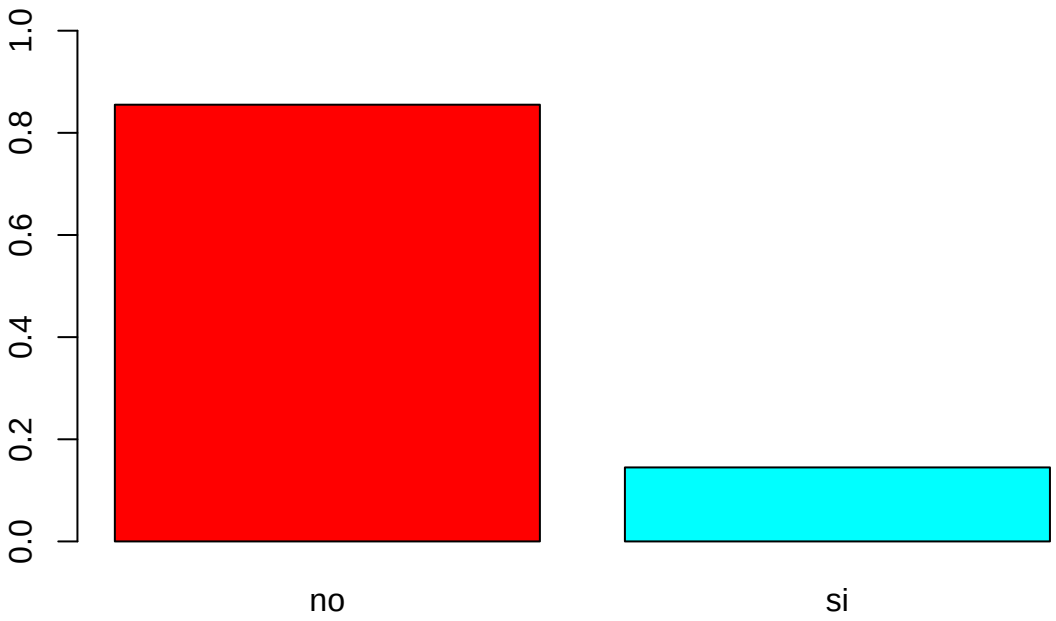
```
##
##   no   si
## 2850  483
```

```
prop.table(table(datos$Desafiliado))
```

```
##
##        no        si
## 0.8550855 0.1449145
```

Se puede ver que existe un desbalanceo de datos con respecto a la variable Desafiliado

```
barplot(prop.table(table(datos$Desafiliado)),
  col = rainbow(2),
  ylim = c(0, 1),
  main = "Distribución de Clases"
)
```



Balanceo - Oversampling

```
prop.table(table(datos$Desafiliado))
```

```
##
##        no        si
## 0.8550855 0.1449145
```

Cantidad de observaciones en desafiliado que son Desafiliado = no

```
table(datos$Desafiliado)[1]
```

```
##   no
## 2850
```

```
library(ROSE)
```

```
## Loaded ROSE 0.0-4
```

```
desafilado_bal_over <- ovun.sample(Desafiliado ~ .,
  data = datos, method = "over",
  N = table(datos$Desafiliado)[1] * 2
)$data
table(desafilado_bal_over$Desafiliado)
```

```
##
##   no   si
## 2850 2850
```

```
head(desafilado_bal_over)
```

```
##   Plan_Internacional Min_En_Dia Min_Internacionales Reclamos
## 1                 no      265.1                10.0        1
## 2                 no      161.6                13.7        1
## 3                 no      243.4                12.2        0
## 4                 si      299.4                 6.6        2
## 5                 si      166.7                10.1        3
## 6                 si      223.4                 6.3        0
##   Llamadas_Internacionales Desafiliado
## 1                        3          no
## 2                        3          no
## 3                        5          no
## 4                        7          no
## 5                        3          no
## 6                        6          no
```

Balanceo - Undersampling

Cantidad de observaciones en desafilado_train que son Desafiliado = yes

```
table(datos$Desafiliado)[2]
```

```
##  si
## 483
```

```
desafilado_bal_under <- ovun.sample(Desafiliado ~ .,
  data = datos, method = "under",
```

```
  N = table(datos$Desafiliado)[2] * 2
)$data
table(desafilado_bal_under$Desafiliado)
```

```
##
##  no  si
## 483 483
```

Undersampling y Oversampling

Cantidad de observaciones en el conjunto de datos

```
dim(datos)[1]
```

```
## [1] 3333
```

```
desafilado_bal_ambos <- ovun.sample(Desafiliado ~ .,
  data = datos,
  method = "both",
  p = 0.5,
  N = dim(datos)[1],
  seed = 1
)$data
# p es la prob. de la clase positiva en la nueva muestra generada
table(desafilado_bal_ambos$Desafiliado)
```

```
##
##   no   si
## 1721 1612
```

```
prop.table(table(desafilado_bal_ambos$Desafiliado))
```

```
##
##        no        si
## 0.5163516 0.4836484
```

SMOTE

```
library(DMwR)
```

```
## Loading required package: lattice
```

```
## Loading required package: grid
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
set.seed(2019) # Para tener resultados reproducibles
datos <- as.data.frame(datos)
datos$Plan_Internacional <- as.factor(datos$Plan_Internacional)
datos$Desafiliado <- as.factor(datos$Desafiliado)
desafilado_bal_smote <- SMOTE(Desafiliado ~ ., datos,
  perc.over = 200, k = 5,
  perc.under = 100
```

```
)
table(desafilado_bal_smote$Desafiliado)
```

```
##
##   no   si
##  966 1449
```