Boosting try

2025-12-02

```
load("~/GitHub/Mineria/DATA/dataaaaaaaaaaaaaa.RData")
```

```
train_df <- data_reducida[data_reducida$group == "train", ]
test_df  <- data_reducida[data_reducida$group == "test", ]
```

duplicados:

```
dup_completos_train <- train_df[
  duplicated(train_df) |
  duplicated(train_df, fromLast = TRUE),
]

nrow(dup_completos_train)
```

```
## [1] 1905
```

duplicados sin exited ni group:

```
##
## Adjuntando el paquete: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## [1] 2104
```

Hay filas idénticas con distinto Exited?

```
## [1] 176
```

Qué valores de Exited tienen los duplicados?

```
table(dup_completos_train$Exited)
```

```
##
##    0    1
## 1785  120
```

```r
prop.table(table(dup_completos_train$Exited))
```

```
##
##          0          1
## 0.93700787 0.06299213
```

Contra la proporcion total:

```r
table(train_df$Exited)
```

```
##
##    0    1
## 5550 1450
```

```r
prop.table(table(train_df$Exited))
```

```
##
##         0         1
## 0.7928571 0.2071429
```

Cuáles son los duplicados más frecuentes en train:

```
## # A tibble: 5,367 x 10
##    Geography Gender IsActiveMember NumOfProducts_grupo   Age Balance     n   n_0
##    <fct>     <fct>  <fct>          <fct>                <dbl>   <dbl> <int> <int>
##  1 France    Male   0              2                       37       0    22    18
##  2 France    Male   0              2                       38       0    21    19
##  3 France    Male   1              2                       32       0    18    18
##  4 France    Male   1              2                       33       0    18    18
##  5 France    Female 0              2                       35       0    17    15
##  6 France    Male   0              2                       30       0    16    16
##  7 France    Male   1              2                       37       0    16    16
##  8 France    Male   1              2                       38       0    16    16
##  9 France    Male   1              2                       40       0    16    16
## 10 France    Female 0              2                       34       0    14    13
## # i 5,357 more rows
## # i 2 more variables: n_1 <int>, prop_0 <dbl>
```