script_baseline.R

sergio

2020-05-27

```
source('../utils/utils_oblig.R')
```

```
## Loading required package: caret
```

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

```
## Loading required package: ggplot2
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
set.seed(117)

script.name <- 'baseline'

script.date <- date()

script.start <- Sys.time()

print('Start')
```

```
## [1] "Start"
```

```
# leer el archivo dataset.csv de la carpeta

dataset <- read.csv('../data/dataset.csv')

print('** Tratamiento inicial de los datos')
```

```
## [1] "** Tratamiento inicial de los datos"
```

```
# ver la estructura del dataset

# str(dataset)

# asignar el nombre del jugador como nombre de la fila

rownames(dataset) <- dataset$CustomerID

df <- na.omit(dataset[,-1])
df$ServiceArea <- NULL

df.output.levels <- levels(df$Churn)

print('** Distribucion a-priori de la variable a predecir')
```

```
## [1] "** Distribucion a-priori de la variable a predecir"
```

```
df.apriori <- prop.table(table(df$Churn))

print(df.apriori)
```

```
##
##        No       Yes
## 0.7131871 0.2868129
```

```
df.part <- train_dev_partition(df, p = 0.9)

df.fn_summary <- fn_summaryUtility

df.metric <- 'utility'

df.form <- Churn ~ .

print('** Utilidad maxima en dev')
```

```
## [1] "** Utilidad maxima en dev"
```

```
df.max_utility <- fn_utility(df.part$dev$Churn, df.part$dev$Churn)

print(df.max_utility)
```

```
## [1] 7.32696
```

```
print('** Baseline -- todos Yes')
```

```
## [1] "** Baseline -- todos Yes"
```

```
df.baseline.pred <- factor(rep('Yes', nrow(df.part$dev)),
                           levels = df.output.levels)

print('Utilidad de la prediccion en dev')
```

```
## [1] "Utilidad de la prediccion en dev"
```

```
df.baseline.utility <- fn_utility(df.baseline.pred, df.part$dev$Churn)

print(df.baseline.utility)
```

```
## [1] 0.2002777
```

```
print('Matriz de confusion')
```

```
## [1] "Matriz de confusion"
```

```
df.baseline.cm <- conf_matrix(df.baseline.pred, df.part$dev$Churn)

print(df.baseline.cm)
```

```
##           Reference
## Prediction   No  Yes
##        No     0    0
##        Yes 3336 1345
```

```
print('** Generacion de la prediccion sobre test sample')
```

```
## [1] "** Generacion de la prediccion sobre test sample"
```

```
test_sample <- read.csv('../data/test_sample.csv')
rownames(test_sample) <- test_sample$CustomerID
test_sample$CustomerID <- NULL
test_sample$ServiceArea <- NULL

test_sample.pred <- factor(rep('Yes', nrow(test_sample)),
                           levels = df.output.levels)

file_id <- paste0(c(script.name, script.date), collapse = ' ')

gen_output(test_sample, test_sample.pred, file_id)

print('Done')
```

```
## [1] "Done"
```

```
script.done <- Sys.time()

print(script.done - script.start)
```

```
## Time difference of 2.26763 secs
```