Experimento propio

Massri

2023-08-21

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

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

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

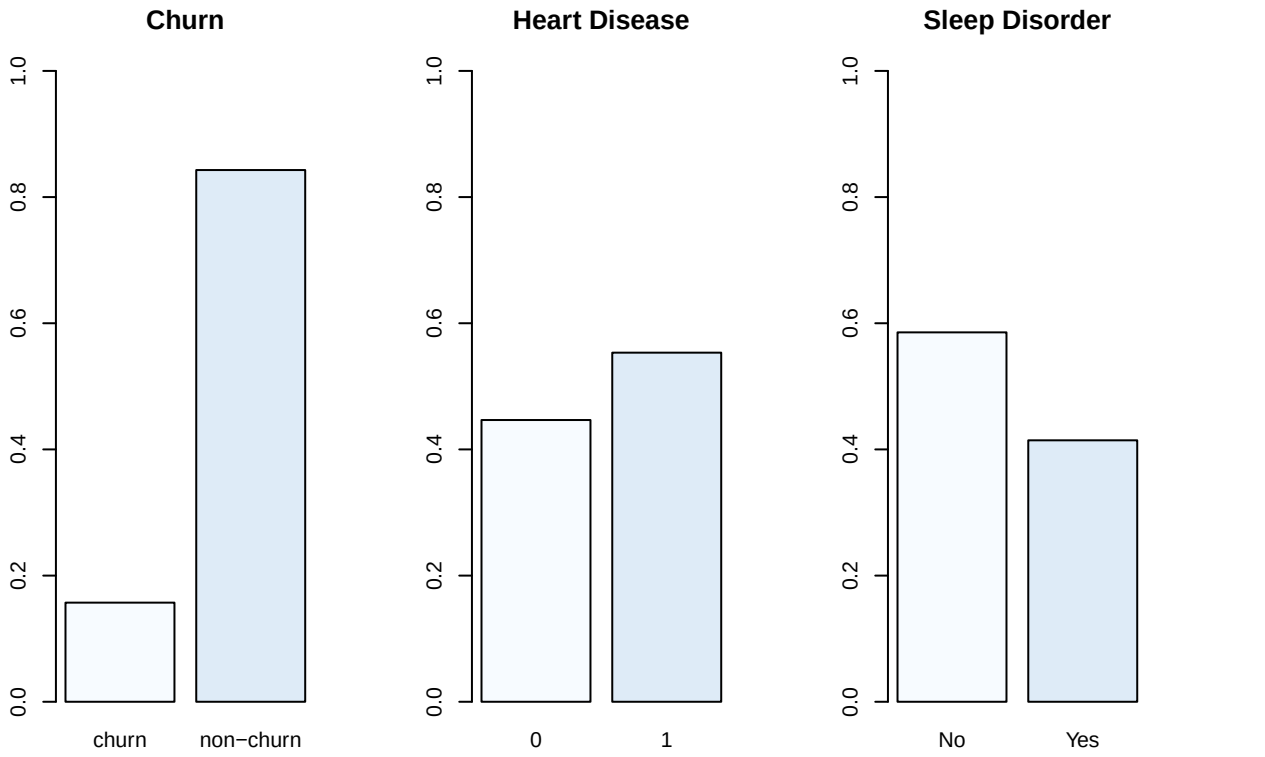
Decidimos hacer subsampling como nuestro experimento propio para estudiar si el balanceo/desbalanceo de clases en la variable a predciri tiene efecto en la performance del modelo. Para eso, eliminamos al azar algunas observaciones de la clase mayoritaria para lograr igual cantidad de observaciones en ambas clases. Luego entrenamos nuestro modelo y obtuvimos los resultados mostrados arriba. Comparando para cada dataset y 2 propNA, la performace del modelo que subsamplea y la que no lo hace.
Nustra hipotesis era que el modelo deberia predecir mejor en datasets balanceados que en los que no lo estan ya que podria haber un sesgo de predecir en mayor proporcion a clase mayoritaria.

```
data_heart = read.table("./data/heart.csv",header=TRUE, sep = ",")

data_churn = read.table("./data/customer_churn.csv", header=TRUE,sep = ",")

data_sleep = read.table("./data/sleep_health_proc.csv",header=TRUE, sep = ",")
```

```
par(mfrow=c(1, 3))
barplot(prop.table(table(data_churn$churn)), main = "Churn" , col = blues9, ylim = (c(0,1)))
barplot(prop.table(table(data_heart$HeartDisease)), main = "Heart Disease", col = blues9, ylim = (c(0,1
barplot(prop.table(table(data_sleep$Sleep.Disorder)), main = "Sleep Disorder", col = blues9, ylim = (c(
```




```
par(mfrow=c(1, 1))
```

Tanto el dataset de sleep disorder como el de heart disease estan dentro de todo balanceados, diferente de el dataset de churn que la clase non-churn es muy mayoritaria. Creemos que balancear en los 2 datasets mas balanceados no tendra mucho efecto, sin embargo creemos que la perfomance sobre churn mejorara.

Luego de hacer el experimento, notamos que no hay tal mejoras en la performance en ninguno de los 3 datasets y para ningun propNA. Esperabamos que esto ocurra tanto para Heart como para Sleep pero no para Churn. Pensamos que puede ocurrir que cada clase este lo suficimientemente representada por lo que los resultados sin balancear son los maximos alcanzables.

Notamos que los modelos que subsamplean tienen menor cantidad de observaciones y que mantienen la performance. Por lo tanto, vamos a recrear el experiemento en donde los no subsampleados pasen a hacer un subsampleo random para poder comparar 2 modelos entrenados con la misma cantidad de datos.

```
source("exp_propio.R")
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following objects are masked from 'package:Metrics':
##
##     precision, recall

## Loading required package: iterators
```

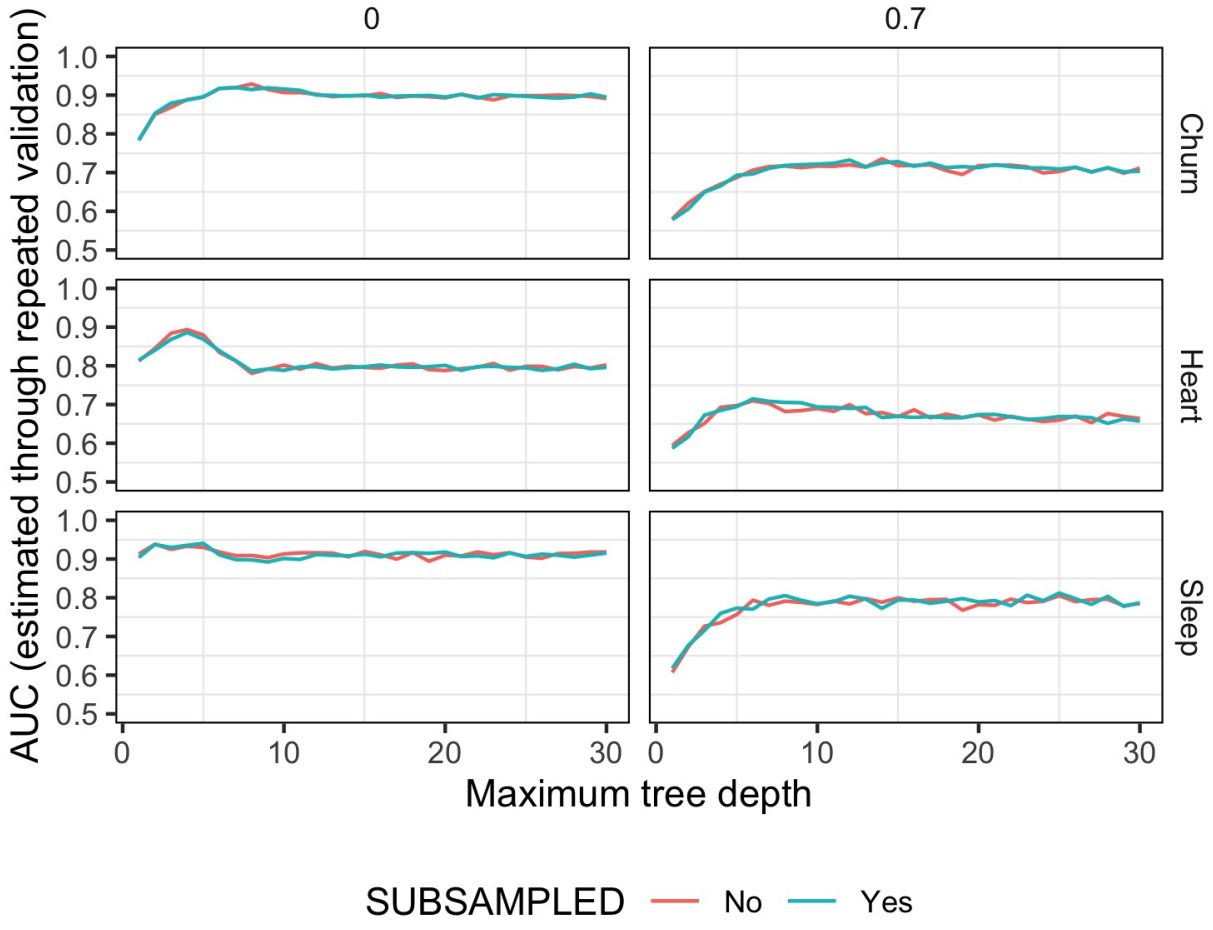
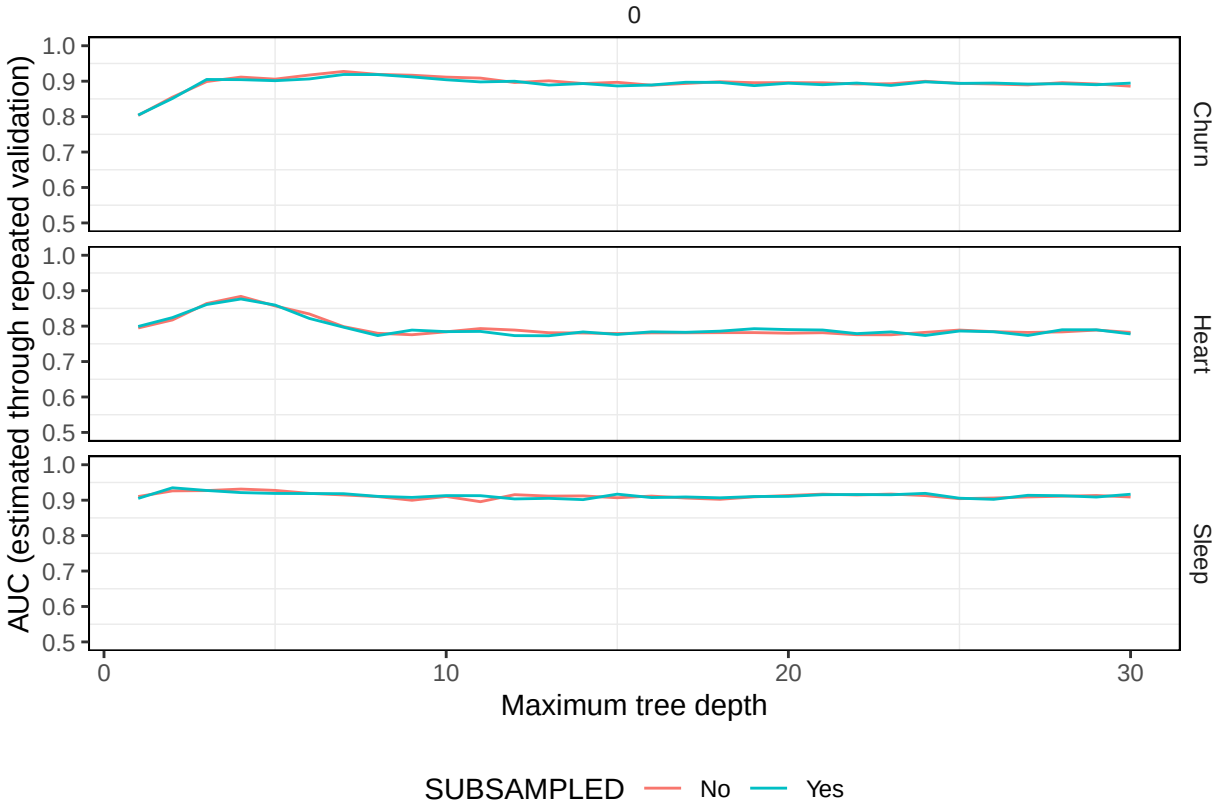


Figure 1: Graficos

```
## Loading required package: parallel
```



Luego de hacer el experimento con ambos modelos entrenados con la misma cantidad de observaciones, no notamos ninguna mejora en la performance del mismo.