Experimento opción 1

De Fino - Solari Barrios - Wurzel

2023-08-22

Hipótesis pre-experimentación

1. Por la gran cantidad de observaciones en nuestro conjunto de datos sospechamos que las predicciones serán mejores que en los conjuntos Churn y Heart.

```
#Cantidad de observaciones CO2
CO2 = read.csv("./data/CO2_Emissions_Transformado.csv")
nrow(CO2)
```

```
## [1] 7385
```

```
#Cantidad de observaciones Churn
Churn = read.csv("./data/customer_churn.csv")
nrow(Churn)
```

```
## [1] 3150
```

```
#Cantidad de observaciones Heart
Heart = read.csv("./data/heart.csv")
nrow(Heart)
```

```
## [1] 918
```

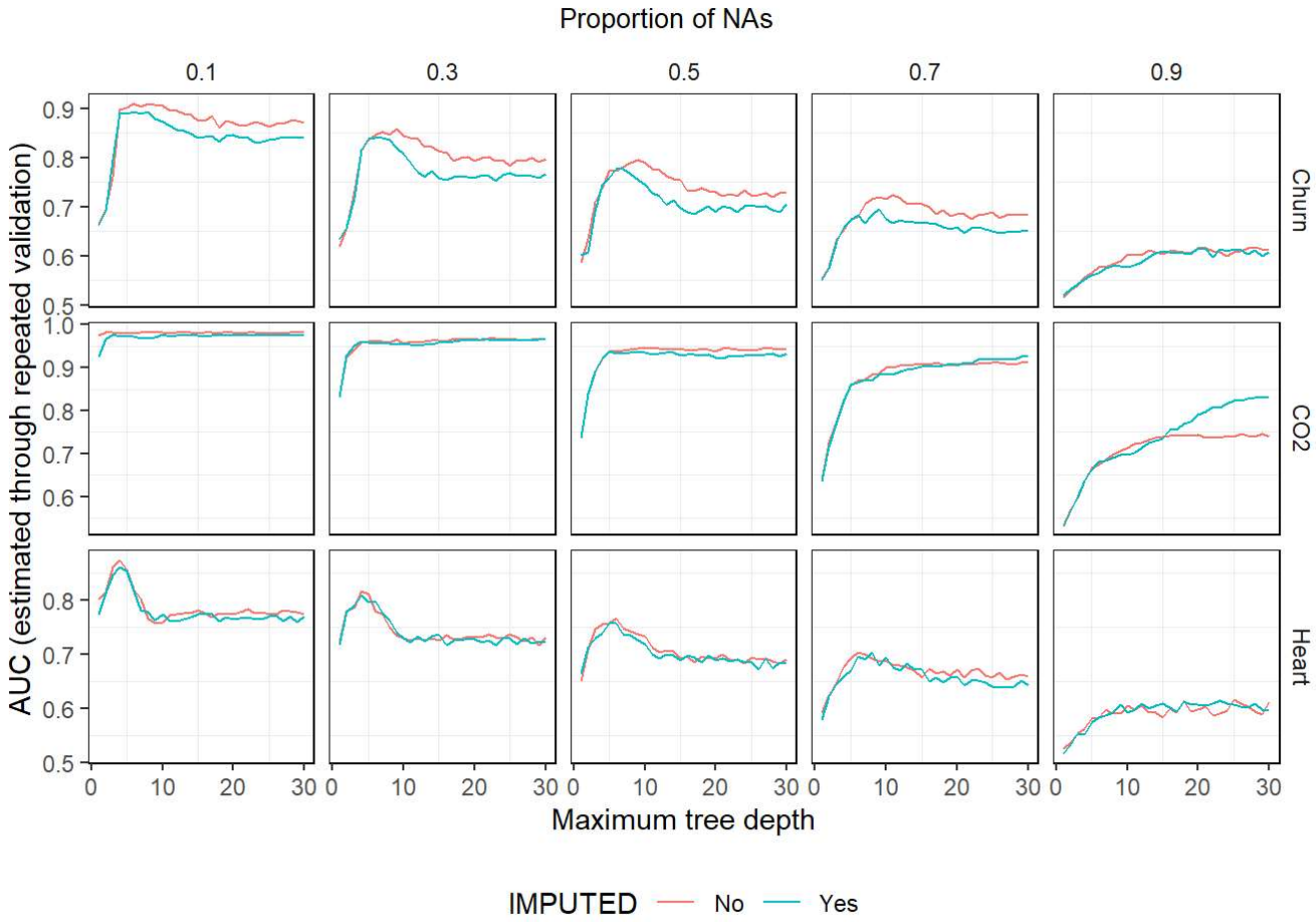
2. A mayor cantidad de NAs peores van a ser las predicciones para los tres conjuntos de datos tanto imputándolos como no.

3. Creemos que imputar los NAs con la media va a ser de gran ayuda dado que reemplazarlos por un valor representativo de los datos puede llevar a mejores predicciones. Por lo que encontramos de la documentación de rpart, a la hora de buscar los mejores splits los NAs se ignoran. Creemos que habría mejores predicciones si estas observaciones se tomaran en cuenta al no reducir tanto la cantidad de observaciones a ser consideradas.

Experimentación

Importamos el script y el gráfico que este realiza:

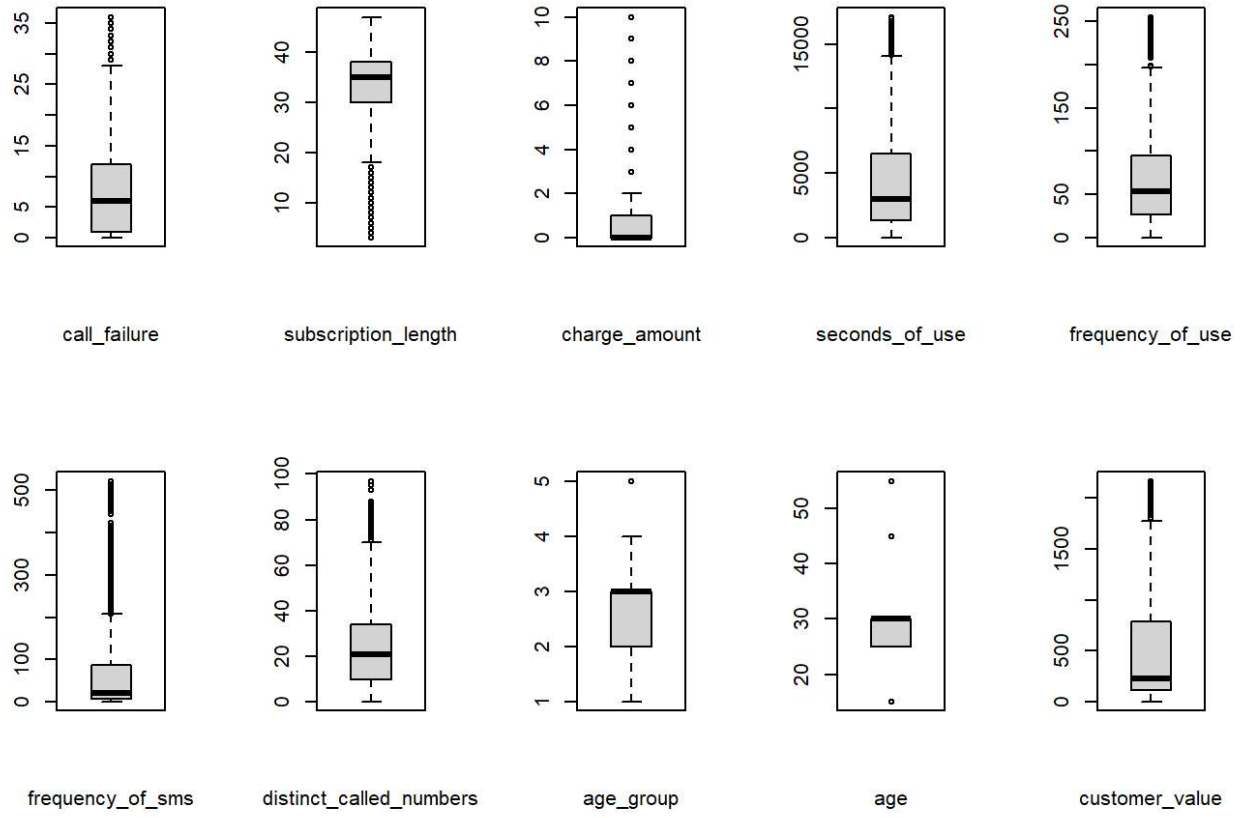
```
# Cargamos el script
# A la funcion plot_exp_results le agregamos un print del gráfico
# para facilitar el display acá
suppressWarnings(source("exp_1.R"))
```



Análisis de hipótesis

1. Dados los AUC reportados podemos confirmar que CO2 tiene mayor performance que Churn y Heart. Recordemos que una mayor cantidad de observaciones lleva a que el arbol pueda tomar decisiones más acertadas al tener más información para sacar conclusiones.
2. Se puede ver claramente en el gráfico que la performance para todos disminuye a medida que la proporción de NAs es mayor. Esto se encuentra un poco relacionado con el item anterior. Tener más información y de buena calidad (datos reales, no fabricados al imputar) lleva a una mayor performance.
3. Podemos observar varias cosas con respecto a imputar o no los datos al analizar el gráfico.
    - En primer lugar, vemos que con el dataset de Churn, parece ser mejor no imputarlos para todas las proporciones salvo para 0.9 que arroja AUCs más similares. Esto contradice nuestra hipótesis, veamos que está pasando con los datos y sus medias. Para esto decidimos realizar boxplots de las variables numéricas.

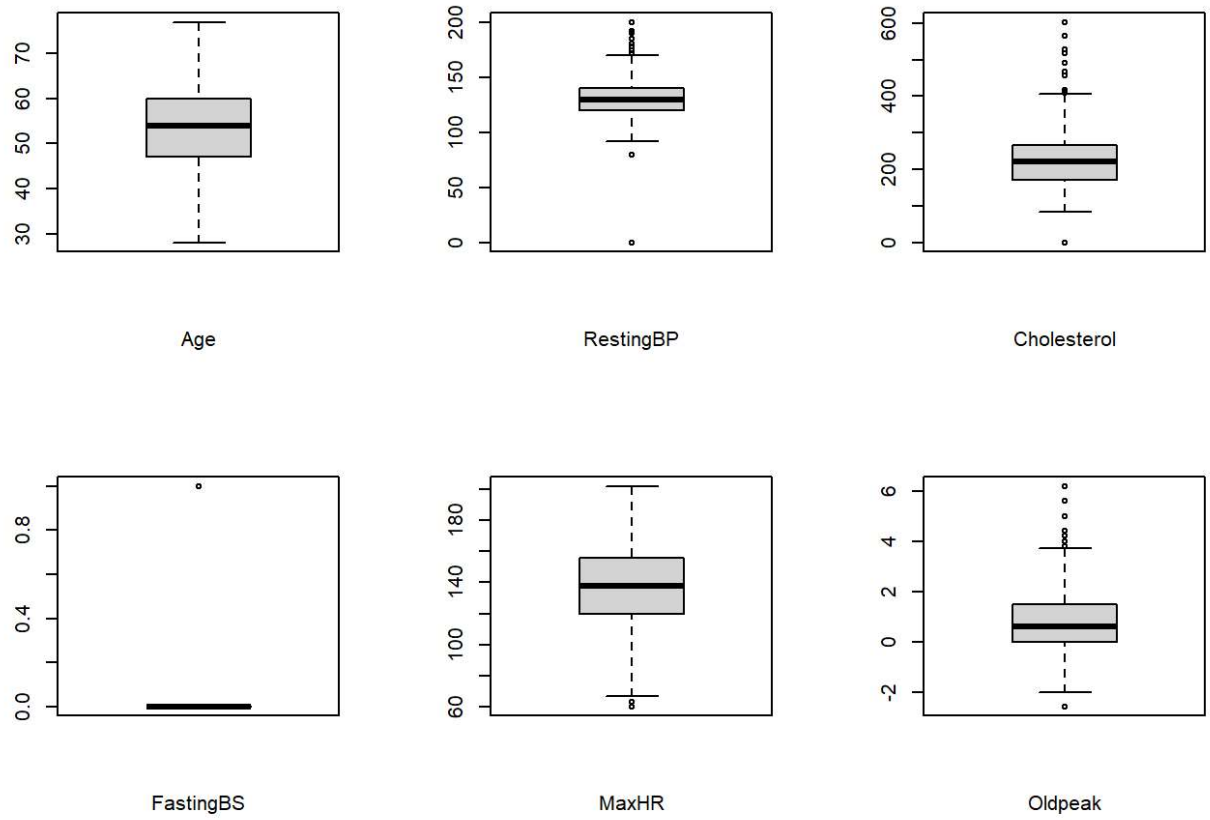
```
#Obtengo la lista de variables numericas para facilitar el ploteo
numerical_cols <- unlist(lapply(Churn, is.numeric))
ChurnNumerical <- Churn[ , numerical_cols]
par(mfrow=c(2,5))
boxplot(ChurnNumerical$call_failure, xlab="call_failure")
boxplot(ChurnNumerical$subscription_length, xlab="subscription_length")
boxplot(ChurnNumerical$charge_amount, xlab="charge_amount")
boxplot(ChurnNumerical$seconds_of_use, xlab="seconds_of_use")
boxplot(ChurnNumerical$frequency_of_use, xlab="frequency_of_use")
boxplot(ChurnNumerical$frequency_of_sms, xlab="frequency_of_sms")
boxplot(ChurnNumerical$distinct_called_numbers, xlab="distinct_called_numbers")
boxplot(ChurnNumerical$age_group, xlab="age_group")
boxplot(ChurnNumerical$age, xlab="age")
boxplot(ChurnNumerical$customer_value, xlab="customer_value")
```



Observamos una gran cantidad de outliers para varias de las variables demostrando que la media no es tan representativa de los datos. Por esto las predicciones al imputar están siendo peores que al no hacerlo.

- En segundo lugar, para el dataset CO2 parece ser indiferente imputar o no para proporciones menores de NAs. El cambio es realmente notorio cuando el 90% de los datos son NAs, imputarlos es drásticamente mejor que dejar que el arbol los maneje. Consideramos que esto puede estar pasando porque la media del 10% restante debe ser bastante representativa de los datos de test haciendo que imputar los NAs arroje buenos resultados.
- En tercer lugar vemos que en el dataset de Heart es casi completamente indiferente imputar o no los NAs para cualquier proporción. Repetimos el análisis que hicimos para Churn y contrario a este tenemos muy baja presencia de outliers haciendo que la media si sea significativa.

```
numerical_cols <- unlist(lapply(Heart, is.numeric))
HeartNumericals <- Heart[ , numerical_cols]
par(mfrow=c(2,3))
boxplot(HeartNumericals$Age, xlab="Age")
boxplot(HeartNumericals$RestingBP, xlab="RestingBP")
boxplot(HeartNumericals$Cholesterol, xlab="Cholesterol")
boxplot(HeartNumericals$FastingBS, xlab="FastingBS")
boxplot(HeartNumericals$MaxHR, xlab="MaxHR")
boxplot(HeartNumericals$Oldpeak, xlab="Oldpeak")
```

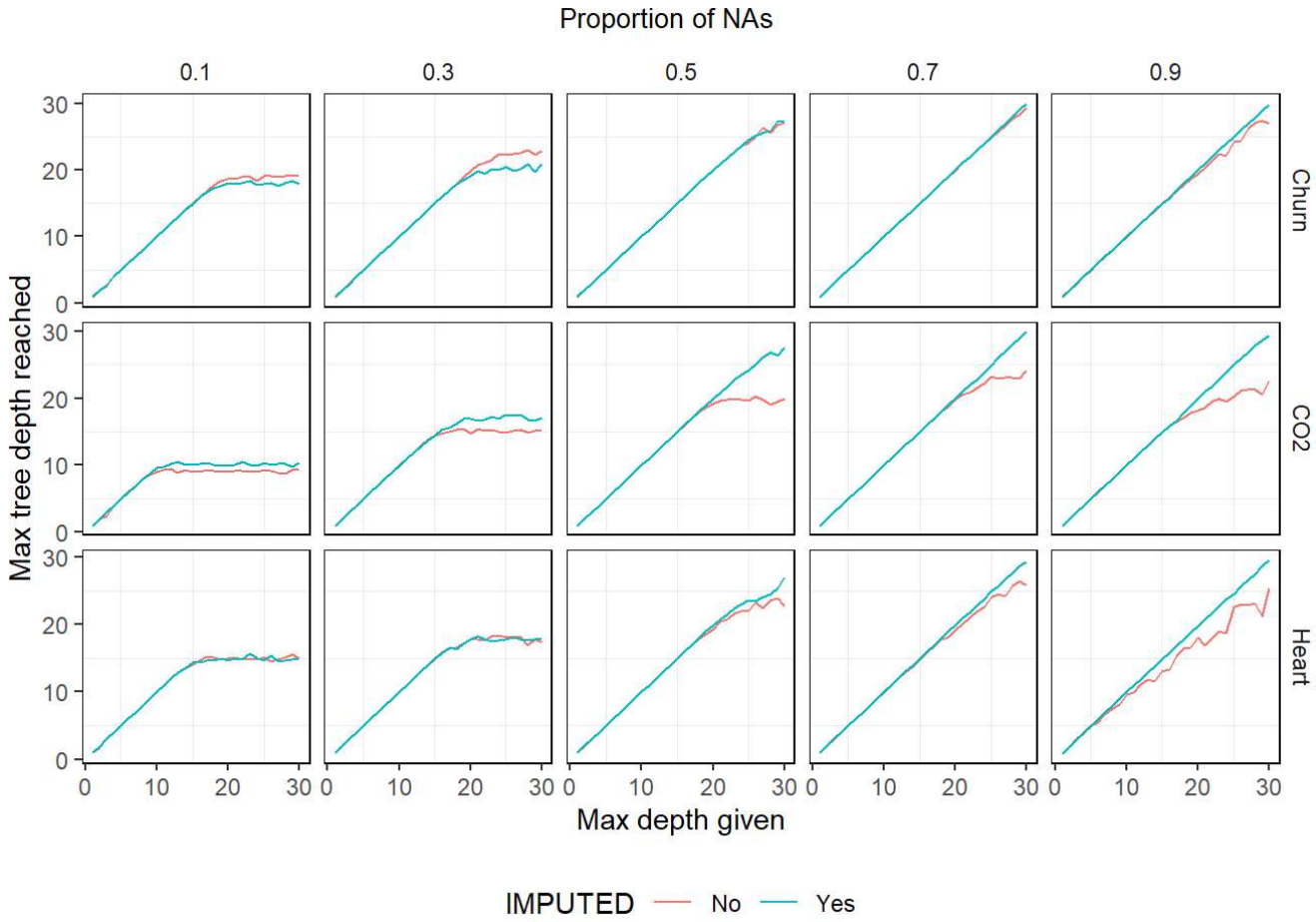


Por lo tanto rechazamos esta hipótesis dado que solo se cumple notoriamente para CO2 con *prop_NAs* = 0.9 .

Una observación extra

Quizás esto no tiene que ver tanto con la proporción de NAs elegida, si se imputan o no, pero notamos que los AUC se estancan a partir de ciertas alturas de los arboles. Entendemos que tiene que ver con la construcción de estos y que según los datos el algoritmo alcanza un tree_depth que le queda cómodo. Igualmente, nos pareció importante construir el siguiente gráfico y ver si se puede extraer alguna conclusión más con respecto a los NAs.

```r
#Modificamos la funcion provista para adecuarla a nuestro gráfico
plot_exp_results <- function(filename_exp_results, filename_plot, width, height) {
  # Load experiment results
  exp_results <- read.table(filename_exp_results, header=TRUE, sep="\t")

  # Calculamos la media de los tree_depth para cada grupo
  data_for_plot <- exp_results %>%
    group_by(dataset_name, prop_NAs, IMPUTED, maxdepth) %>%
    summarize(mean_tree_depth_alcanzado=mean(tree_depth), .groups='drop')

  g <- ggplot(data_for_plot, aes(x=maxdepth, y=mean_tree_depth_alcanzado, color=IMPUTED)) +
    geom_line() +
    theme_bw() +
    ggtitle("Proportion of NAs")+
    xlab("Max depth given") +
    ylab("Max tree depth reached") +
    facet_grid(dataset_name ~ prop_NAs, scales="free_y") +
    theme(legend.position="bottom",
          panel.grid.major=element_blank(),
          strip.background=element_blank(),
          panel.border=element_rect(colour="black", fill=NA),
          plot.title.position = 'plot',
          plot.title = element_text(hjust=0.5, size=10))

  ggsave(filename_plot, g, width=width, height=height)

  # Printeamos para mostrarlo por pantalla
  print(g)
}

# Generamos el nuevo plot llamado exp_1_alt.jpg
plot_exp_results( "./outputs/tables/exp_1.txt", "./outputs/plots/exp_1_alt.jpg", width=15, he
ight=8)
```



Para empezar, una mayor proporción de NAs requiere de una mayor altura. En las proporciones de 0.1 y 0.3 no se alcanza ni el máximo que estamos testeando (30). Notamos que en la mayoría de los casos se alcanza una altura menor al no imputar los datos. Para nuestro conjunto es especialmente notorio: cuando el maxdepth supera los 20 el tree_depth crece mucho más lento cuando no imputamos. Esto nos indica que dejar que rpart maneje los NAs como quiere lleva a requerir menos altura.

Conclusión

Habiendo descartado nuestra hipótesis número tres (imputar mejor que no imputar) y habiendo observado que no imputar lleva a un árbol con menor altura, podemos concluir que es mejor dejar que el árbol trabaje los NAs como le parezca.