Experimento Propio

Patricio Guledjuan Azul Noguera Rocio Gonzalez

2023-08-22

Experimento propio

Introducción

El objetivo de nuestro experimento consistió en explorar y comprender el impacto de las variables en el rendimiento del modelo. A través del análisis de las variables más influyentes, buscamos obtener una visión más profunda de cómo se toman decisiones en el contexto de cada conjunto de datos. Además, evaluamos si la selección de características relevantes podría potencialmente mejorar el desempeño del modelo.

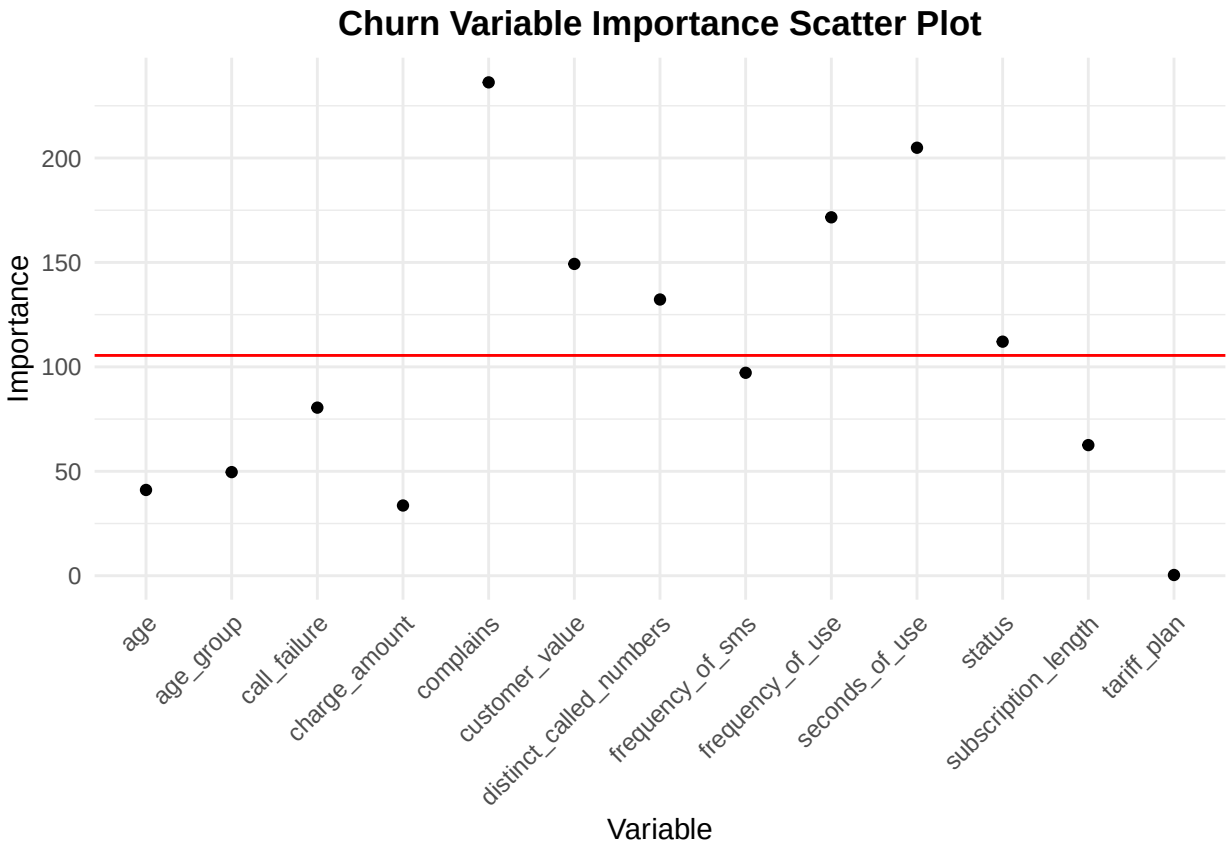
Nuestro enfoque implicó identificar las variables más significativas en función del comportamiento del modelo original y, posteriormente, elegir las variables más influyentes para construir un nuevo modelo. Esta estrategia nos permitió comparar el rendimiento del nuevo modelo con el del original y analizar cualquier diferencia resultante.

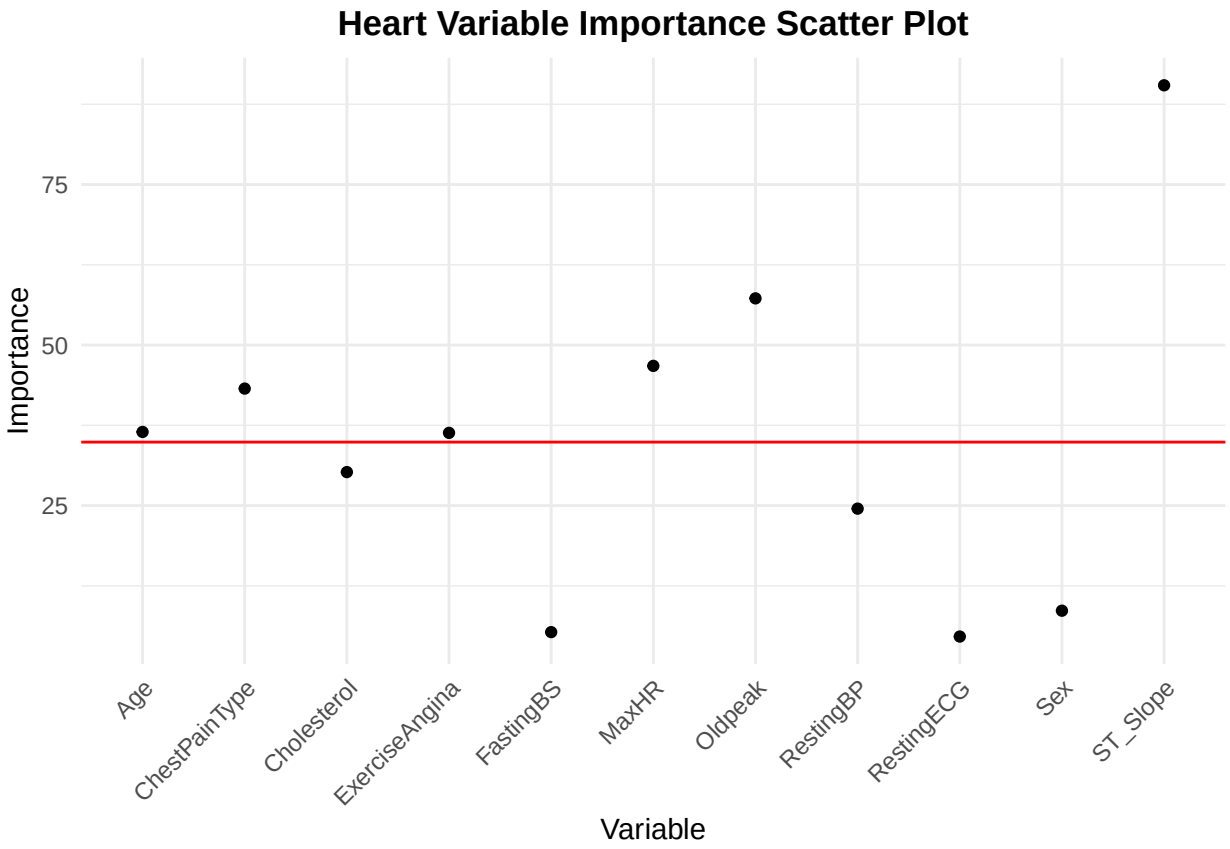
En una primera etapa, evaluamos la importancia de los atributos en el modelo original utilizando rpart. En este contexto, una alta puntuación de importancia sugiere que la variable ejerce una mayor influencia en las decisiones tomadas por el árbol, y por ende, en las predicciones resultantes.

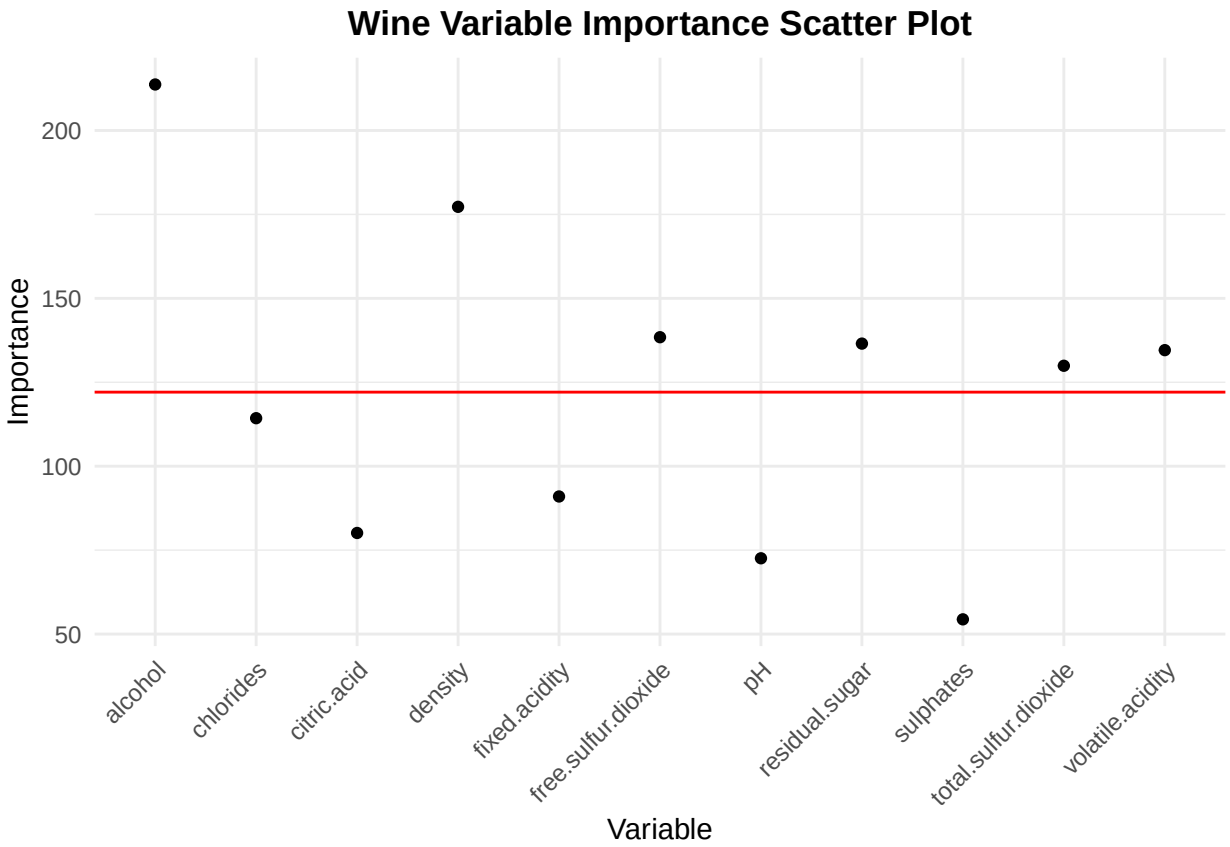
A partir de estos valores de importancia, establecimos una estrategia para determinar qué variables serían incluidas en el nuevo conjunto de datos. Para ello, calculamos el valor promedio de importancia de todas las variables presentes en cada uno de los conjunto de datos originales. Posteriormente, seleccionamos aquellas variables cuya importancia superara este promedio. Como resultado, cada uno de los conjuntos de datos originales se redujo a un conjunto de seis variables predictoras de mayor relevancia.

```
#Tabla con importancia promedio por dataset
data.frame(Dataset = c("ChurnRate", "HeartDisease", "WineQuality"),
           Importancia_Promedio = c(mean_variable_importance_churn, mean_variable_importance_heart, mean
           )
```

```
##        Dataset Importancia_Promedio
## 1    ChurnRate            105.47135
## 2 HeartDisease             34.88927
## 3  WineQuality            122.06333
```


Churn Variable Importance Scatter Plot
Importance
200
150
100
50
0
age
age_group
call_failure
charge_amount
complains
customer_value
distinct_called_numbers
frequency_of_sms
frequency_of_use
seconds_of_use
status
subscription_length
tariff_plan
Variable


Heart Variable Importance Scatter Plot
Importance
75
50
25
Age
ChestPainType
Cholesterol
ExerciseAngina
FastingBS
MaxHR
Oldpeak
RestingBP
RestingECG
Sex
ST_Slope
Variable



A partir de esto, obtuvimos los nuevos datasets que contienen las variables que cumplían con la condición planteada anteriormente.

```
## Warning in write.csv(winequality_top, file = ruta_archivo_nuevo, sep = ",", :
## attempt to set 'sep' ignored
```

```
## Warning in write.csv(customer_churn_top, file = ruta_archivo_nuevo, sep = ",", :
## attempt to set 'sep' ignored
```

```
## Warning in write.csv(heart_top, file = ruta_archivo_nuevo, sep = ",", row.names
## = FALSE): attempt to set 'sep' ignored
```

```
##           Churn_variables Heart_variables    Quality_variables
## 1               complains        ST_Slope              alcohol
## 2           seconds_of_use         Oldpeak              density
## 3         frequency_of_use           MaxHR  free.sulfur.dioxide
## 4           customer_value   ChestPainType       residual.sugar
## 5 distinct_called_numbers             Age     volatile.acidity
## 6                   status  ExerciseAngina total.sulfur.dioxide
```

Resultados

Llevamos a cabo este experimento con la finalidad de indagar si incrementar el número de variables predictoras inevitablemente resulta en una mejora en el rendimiento del modelo, o si este resultado está condicionado por la importancia intrínseca de cada variable y las características particulares de los conjuntos de

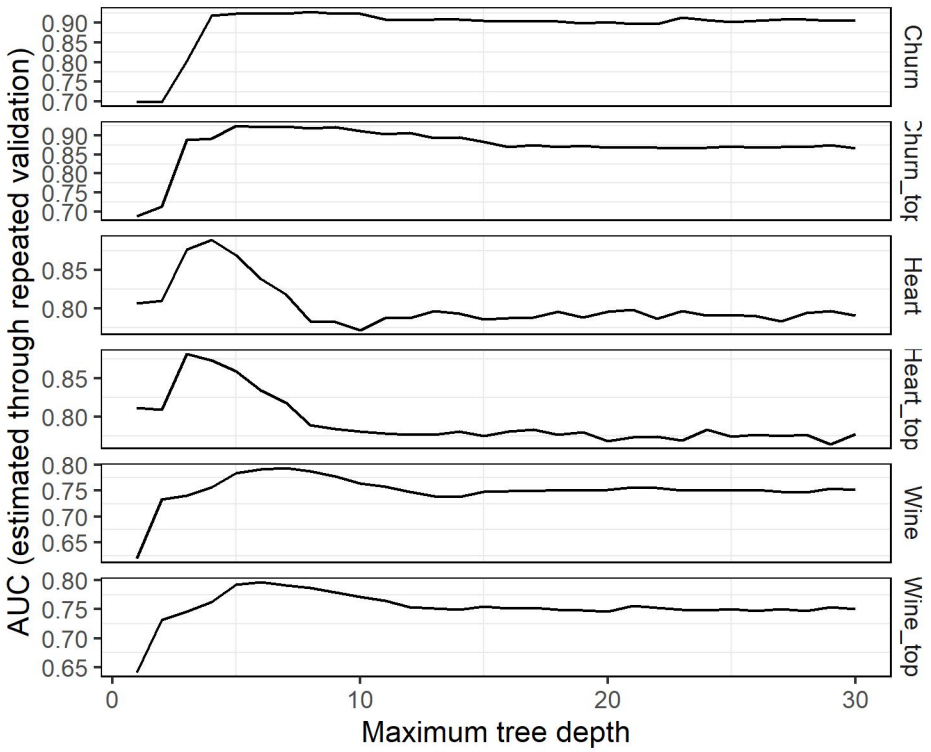


Figure 1: Resultados

datos. Para efectuar este análisis, evaluamos el desempeño de los tres conjuntos de datos originales, junto con el de tres conjuntos de datos que fueron reducidos únicamente a las variables más significativas.

A grandes rasgos, podemos observar que las diferencias de rendimiento entre los conjuntos de datos originales y los conjuntos de datos reducidos son en su mayoría mínimas, siendo las performances prácticamente idénticas. Si profundizamos en los detalles, a medida que aumentamos la profundidad del árbol, se observa una ligera tendencia en la cual los conjuntos de datos originales tienen un rendimiento infimamente superior en comparación con los conjuntos de datos reducidos.

Ahora, profundicemos en cada caso particular para determinar si este comportamiento tiene sentido en cada contexto.

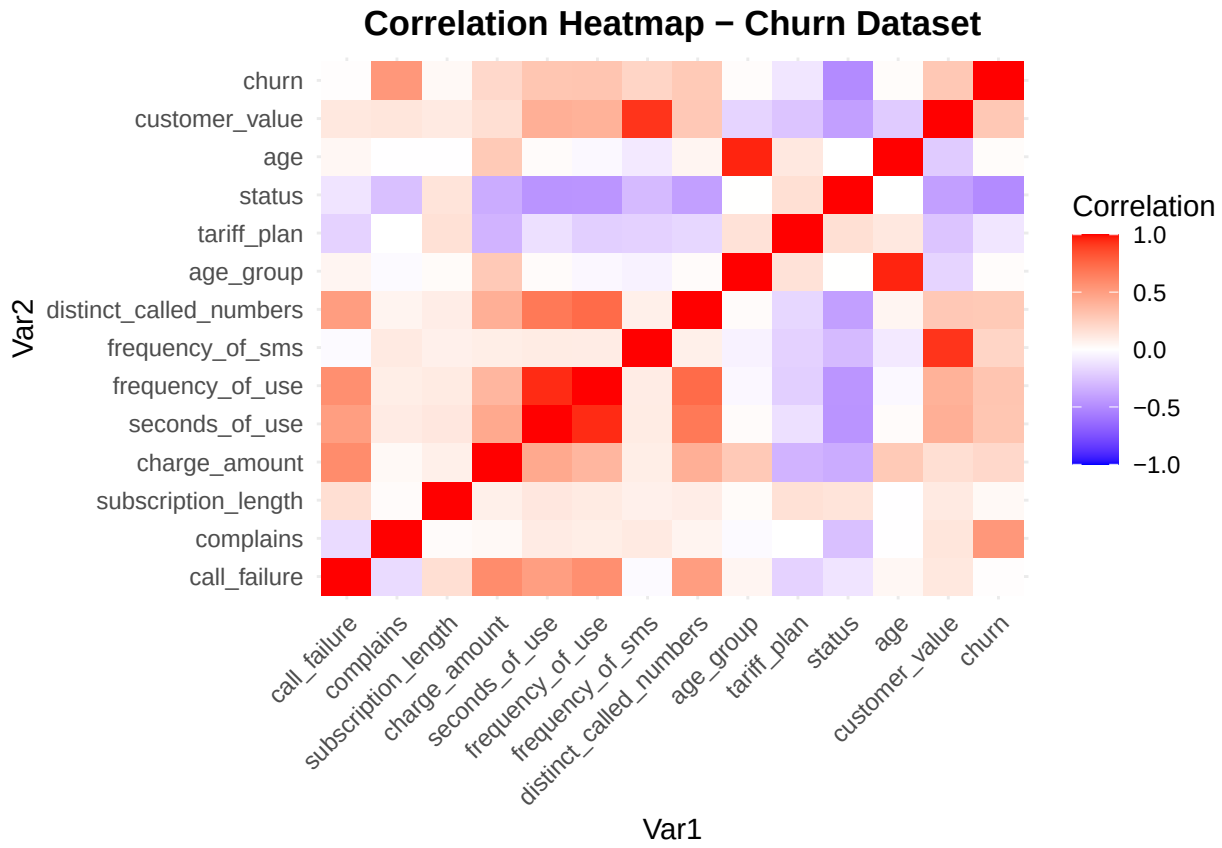
Churn: Tanto el conjunto de datos original de churn como el conjunto reducido exhiben un comportamiento sorprendentemente parecido. Las discrepancias en el rendimiento a lo largo de las profundidades del árbol son mínimas, y se observa que los picos de rendimiento se alcanzan prácticamente al mismo tiempo. Además, los valores máximos de rendimiento se sitúan en torno a 0.9, señalando un nivel de predicción bastante sólido en ambos conjuntos de datos. Tambien, es interesante notar que el conjunto de datos original mantiene una ligera ventaja en términos de rendimiento en todas las iteraciones de entrenamiento de los árboles.

La aparente falta de distincion en el rendimiento podría sugerir que las variables que se eliminaron no desempeñaban un papel significativo en la capacidad del modelo para anticipar las tasas de abandono. Esto podría indicar que, dentro de este contexto particular, las variables que aún permanecen son capaces de capturar de manera adecuada la información relevante para realizar predicciones sólidas acerca del churn. En el marco de la tasa de abandono (churn rate), resulta lógico que las variables que estén directamente relacionadas con los comportamientos y características de los clientes tengan un impacto sustancial en las predicciones. Las variables que mantuvieron una alta importancia se centran en aspectos como los patrones de llamadas, la frecuencia de uso y el valor del cliente. Estos factores pueden representar señales fuertes acerca de si un cliente es más propenso a abandonar el servicio o a mantenerlo.

Una observación importante es que las variables eliminadas podrían haber carecido de aporte informativo único o estar altamente correlacionadas con las variables restantes. Esta correlación podría haber disminuido su valor predictivo cuando ya se consideran las otras características relevantes.

Veamos la corrrelación...

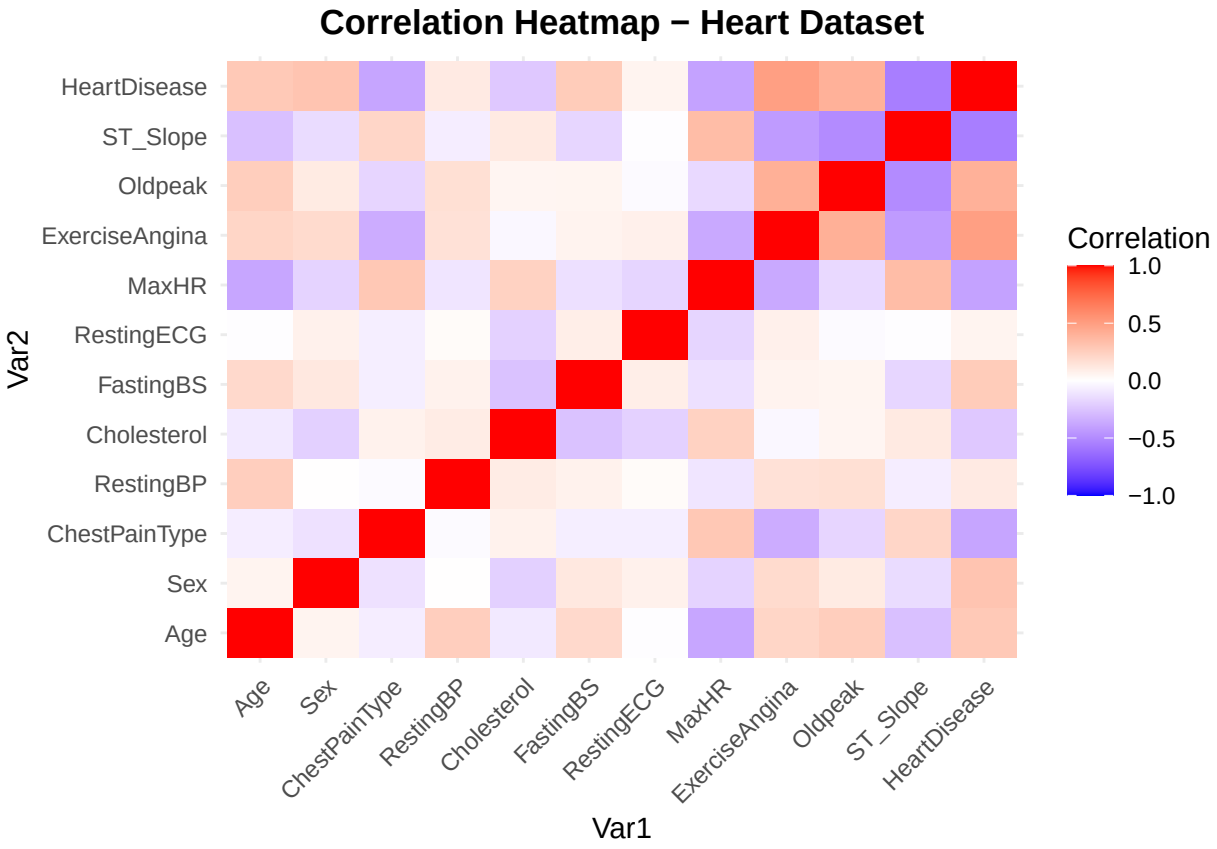
```
##
## Attaching package: 'dplyr'

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

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



Al analizar los valores de correlación entre las variables eliminadas y las restantes, notamos que en su mayoría son bajos. Esto sugiere que no existía una correlación fuerte entre las variables descartadas y las que se mantuvieron. Como resultado, es plausible pensar que las variables eliminadas posiblemente no añadían información distintiva que permitiera diferenciar entre las clases de predicción.

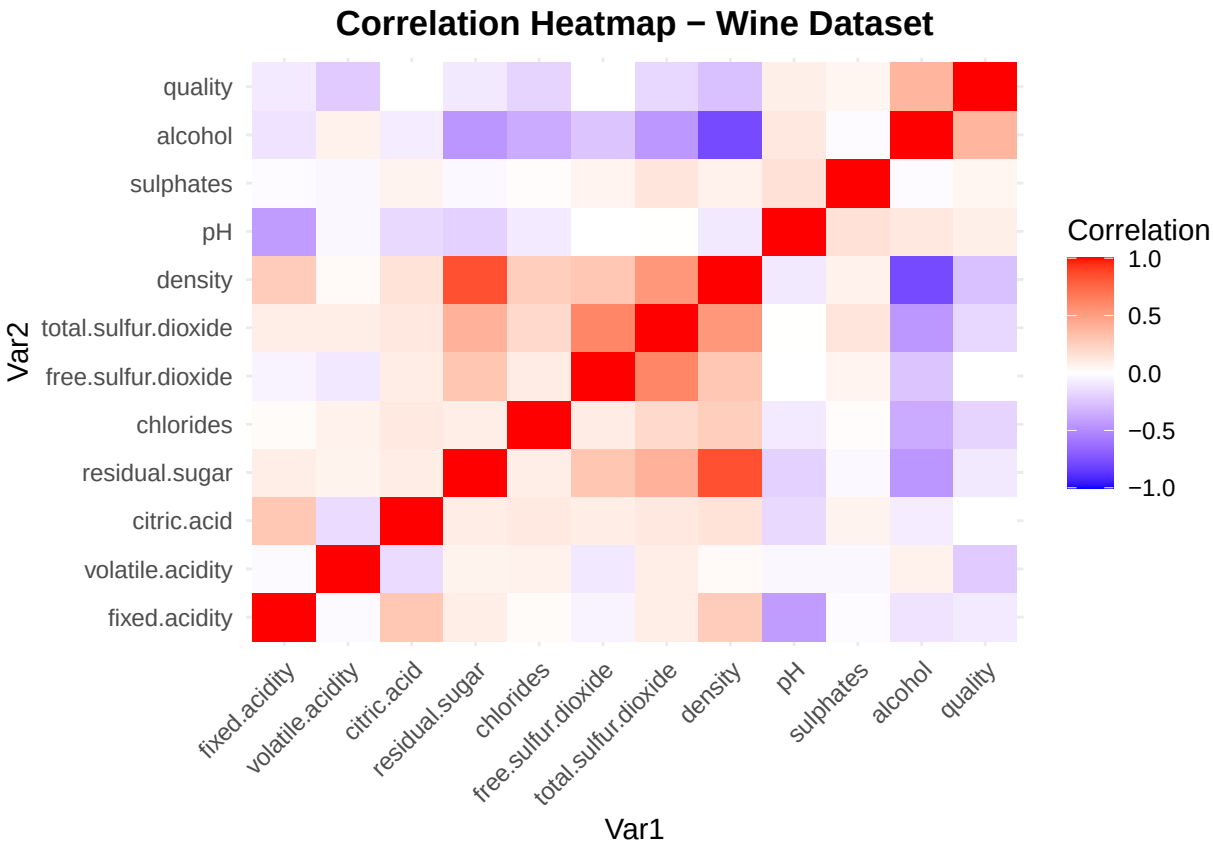
En conjunto, estos resultados sugieren que las variables más influyentes para predecir el churn están siendo capturadas por las variables que se mantienen, y que la inclusión o exclusión de las variables menos relevantes no afecta de manera significativa la capacidad del modelo para realizar predicciones precisas en este contexto particular.

Heart: Dentro del ámbito de las enfermedades cardíacas, las variables que se conservan abordan elementos cruciales como la reacción del corazón al ejercicio, la manifestación de angina inducida por el ejercicio, y características relacionadas con la edad y la sensación de dolor en el pecho. Estas variables actúan como indicadores fundamentales para evaluar la salud cardíaca y podrían desempeñar un papel esencial en la predicción de enfermedades cardíacas. Es lógico suponer que ciertas variables, como el género, posiblemente carezcan de información única suficiente para diferenciar de manera relevante entre las distintas clases de predicción.



Wine: Este dataset es el que parece verse menos afectado por la eliminacion de variables, si comparamos el rendimiento del dataset original con el dataset reducido no parece existir ninguna diferencia de comportamiento.

En el contexto de la calidad del vino, las variables restantes incluyen características químicas y físicas del vino. Estas variables pueden influir en la calidad percibida y real del vino. Dado que la calidad del vino se basa en componentes químicos y sensoriales, las variables seleccionadas son fundamentales para determinar cómo se percibirá el vino en términos de calidad.


Correlation Heatmap − Wine Dataset
quality
alcohol
sulphates
pH
density
total.sulfur.dioxide
free.sulfur.dioxide
chlorides
residual.sugar
citric.acid
volatile.acidity
fixed.acidity
Var2
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
quality
Var1
Correlation
1.0
0.5
0.0
−0.5
−1.0