Experimento 3

Zoe Borrone Luca Mazzarello Ignacio Pardo

2023-10-14

Introducción

Para desarrollar este experimento estudiaremos como, para cada uno de los tres datasets analizados, cómo impacta tratar las variables categóricas ordinales con OHE estándar, versus convertirlas en un número que refleje el orden de las categorías (algo que se logra con discretize igual a TRUE, n_bins igual a 10 y ord_to_numeric igual a TRUE en preprocess_control).

Resultados

A simple vista, los datasets de *Churn* y *Heart* no parecen verse afectados por la discretización de las variables categóricas, mientras que el dataset de *Student* parece verse afectado negativamente. Sin embargo, para poder analizarlo mejor, vamos a graficar la performance de los modelos para cada dataset en función de la profundidad máxima del árbol.

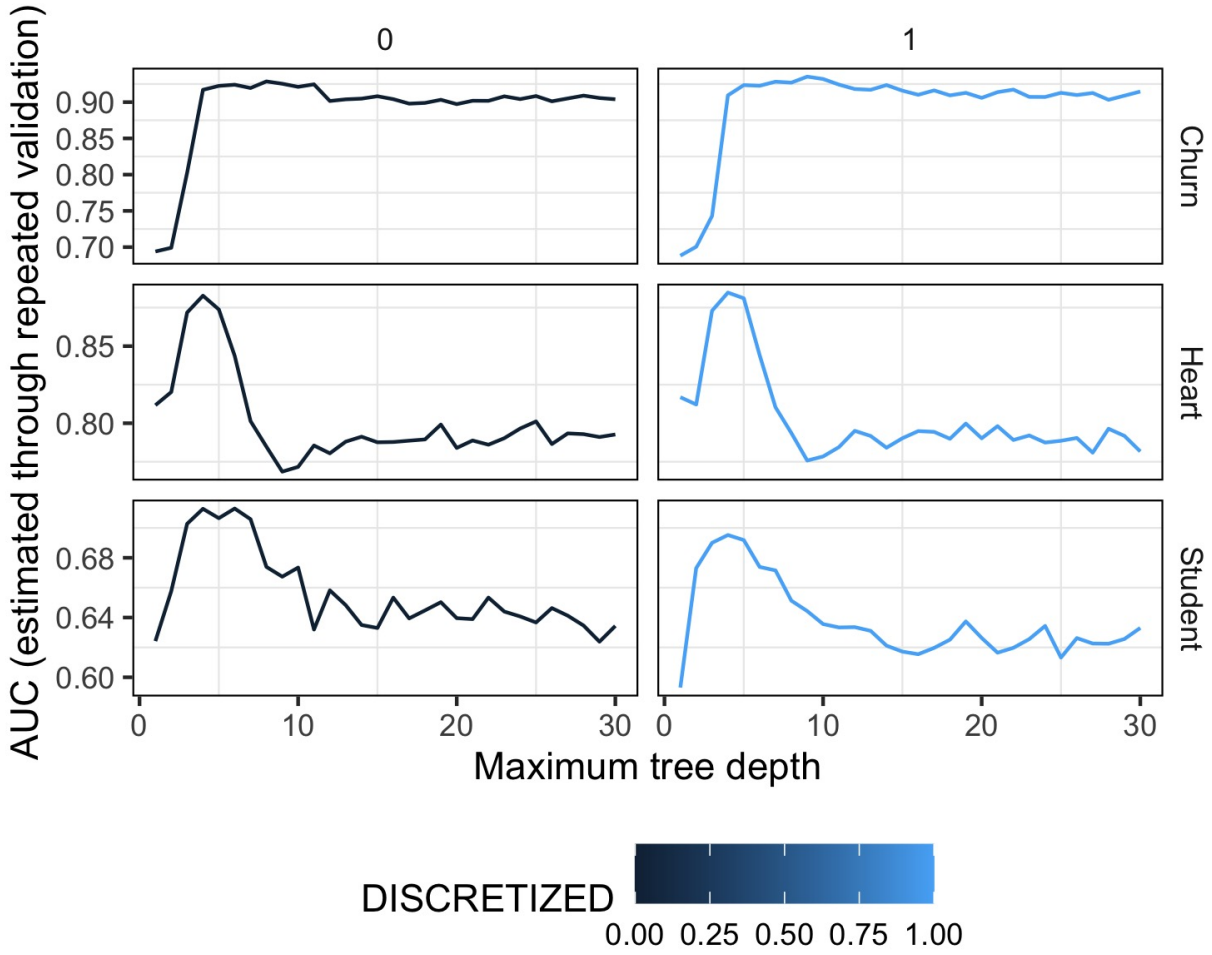


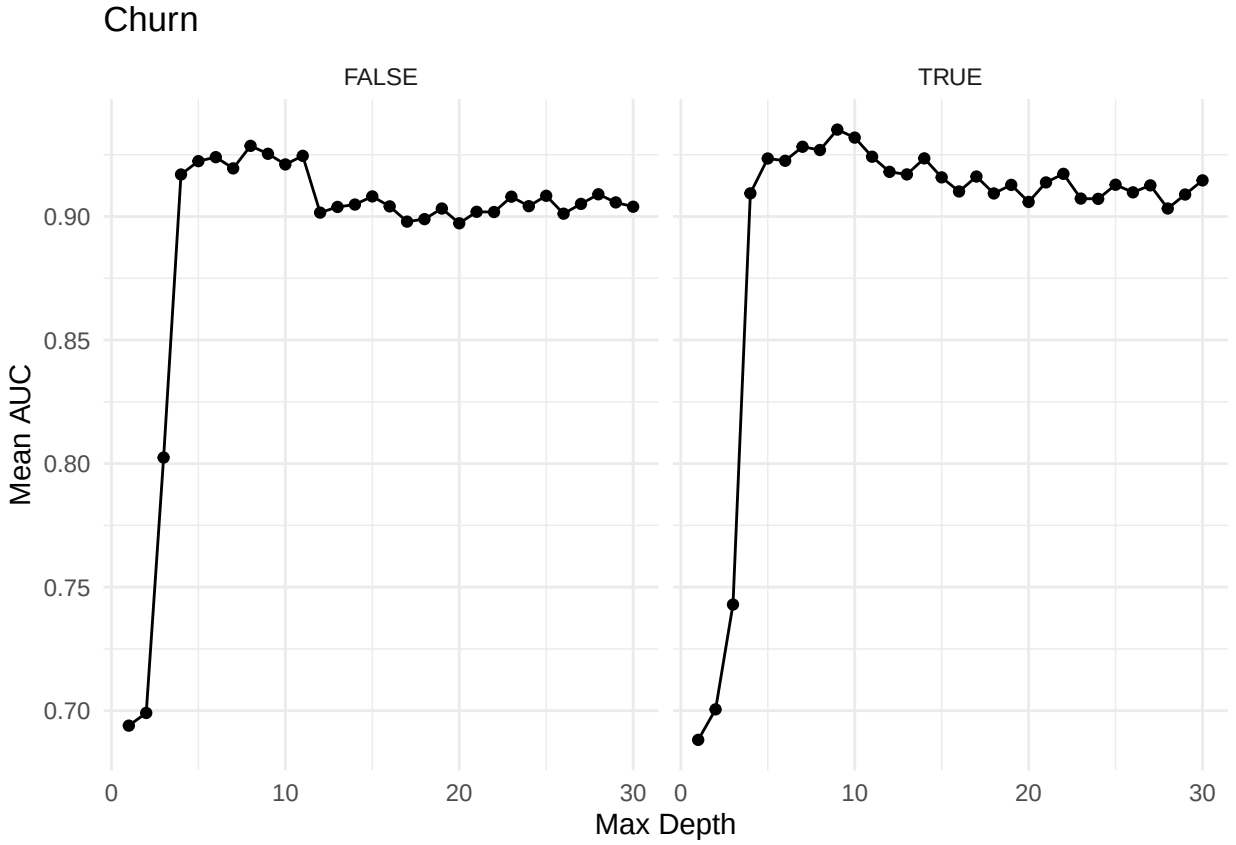
Figure 1: Resultados sin (0/izquierda) y con (1/derecha) Discretización y Numericas

Churn

Como se puede observar en la "*Tabla 1*", para el dataset de *Iranian Churn*, el mayor valor de ROC AUC se obtiene al procesar las variables categoricas convirtiendo sus valores en números que reflejen el orden de las categorías. Ademas, para el mayor valor de AUC alcanzado por el dataset sin procesar, se requirió de una mayor profundidad del árbol (`max_depth=9`) que para el dataset procesado (`max_depth=7`). Esto se puede observar en la "*Figura Churn*" donde se grafica la performance de los modelos en función de la profundidad máxima del árbol.

Table 1: Churn Max AUCs

maxdepth	mean_auc	DISCRETIZED
8	0.9285919	No
9	0.9351698	Yes



Por último, en la cola derecha de los gráficos, parece observarse que el modelo entrenado con el dataset procesado mantiene una performance más estable a medida que aumenta la profundidad del árbol.

Heart

En este caso, en el gráfico inicial observamos que el dataset de *Heart Disease* parece obtener peores resultados al procesarlo. Para confirmar esto generamos la siguientes tablas (Table 2 y Table 3) donde podemos observar el valor máximo y mínimo de AUC para cada uno de los datasets.

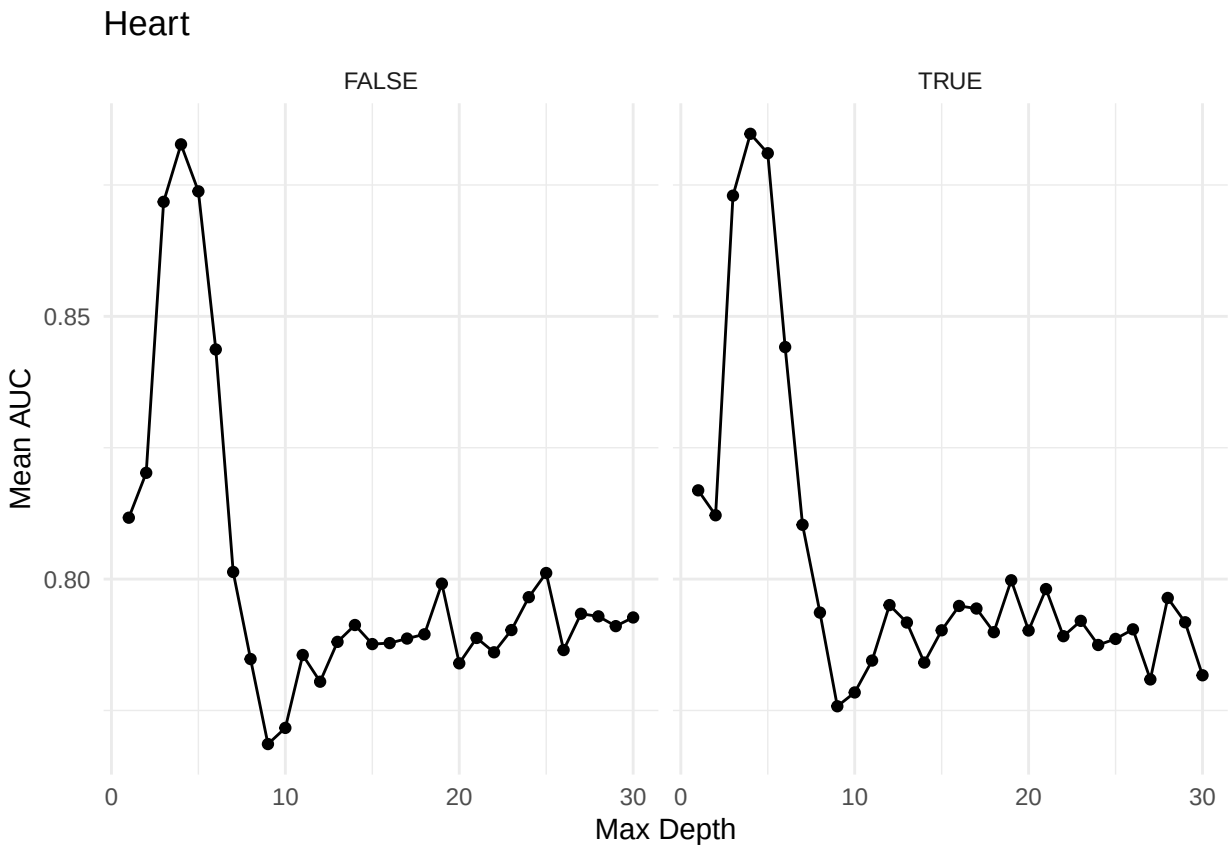
Tanto para el valor maximo alcanzado como para el valor minimo, el dataset sin procesar obtiene mejores resultados y con profundidades de arboles menores.

Table 2: Heart Max AUCs

maxdepth	mean_auc	DISCRETIZED
4	0.8827287	No
4	0.8847342	Yes

Table 3: Heart Min AUCs

maxdepth	mean_auc	DISCRETIZED
9	0.7686053	No
9	0.7757954	Yes


Heart
FALSE
TRUE
Mean AUC
0.85
0.80
0
10
20
30
Max Depth

Student

Al igual que en el caso de *Heart Disease*, el dataset de *Student Performance* parece obtener peores resultados al procesarlo. Nuevamente generamos las siguientes tablas donde podemos observar el valor máximo y mínimo de AUC para cada uno de los datasets.

Table 4: Student Max AUCs

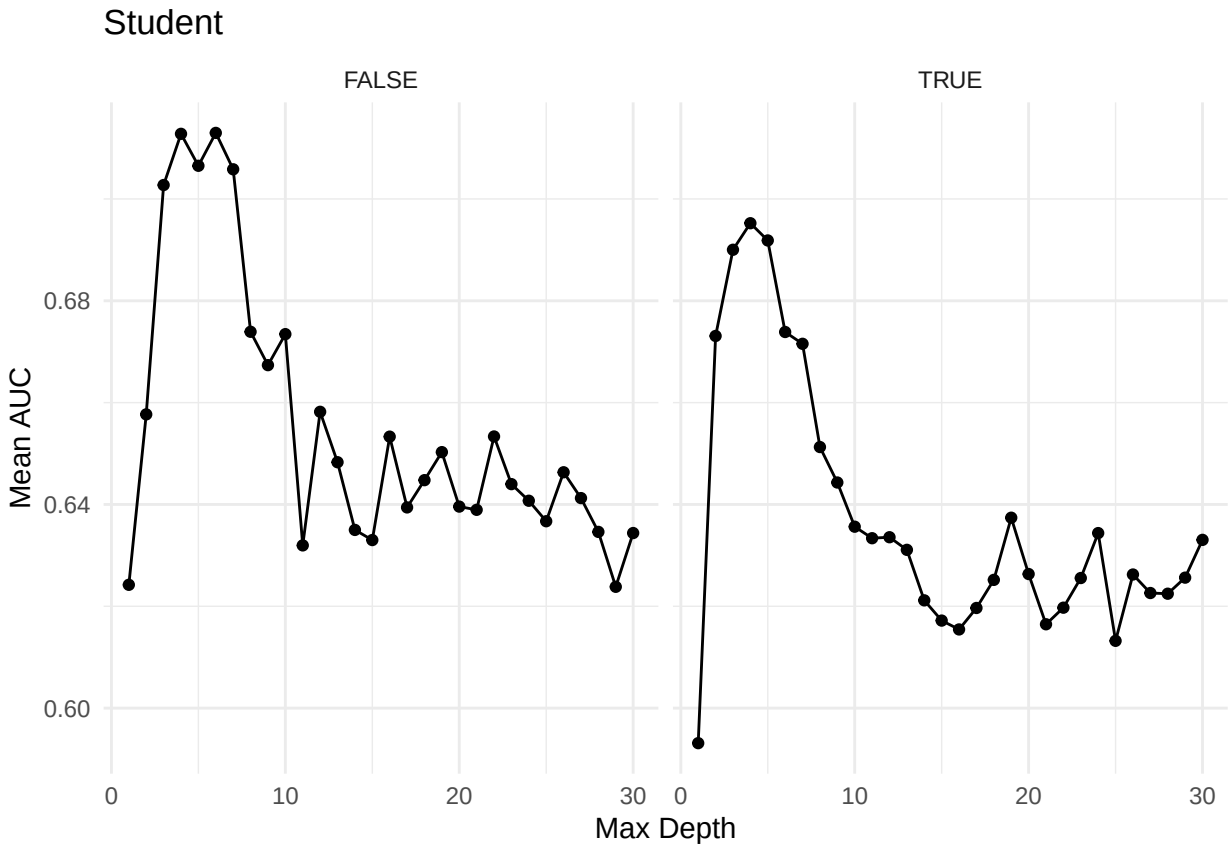
maxdepth	mean_auc	DISCRETIZED
6	0.7129718	No
4	0.6952341	Yes

Table 5: Student Min AUCs

maxdepth	mean_auc	DISCRETIZED
29	0.6238290	No
1	0.5931196	Yes

A diferencia del dataset *Heart Disease*, tanto a los valores máximos como minimos con el dataset de *Student Performance* con y sin procesar, se alcanzan con las mismas profundidades de arboles (`max_depth=6` a los máximos y `max_depth=1` a los mínimos).

Al igual que en el caso de *Heart Disease*, el dataset sin procesar alcanza un valor máximo mayor.



Conclusiones

Aunque no parecería haber un patrón discernible entre los comportamientos de los modelos con y sin la discretización y la conversión de variables ordinales a numericas entre los distintos datasets, los resultados que obtuvimos fueron mayormente negativos.

Nuestra interpretación esta basada en que estos resultados se encuentran dados por la proporción de variables categóricas contra las variables numéricas de cada dataset. En aquel dataset cuya proporción era necesariamente mayor (Heart 5/11), el modelo se veía perjudicado por el procesamiento y en aquellos cuya proporción era menor (Churn 3/13), el modelo se veía beneficiado, aunque mínimo, por el mismo. Para el caso de Student que contiene 13 variables numericas de 30 totales, el procesamiento afecta tambien negativamente al modelo.