Experimento 3

Zoe Borrone Luca Mazzarello Ignacio Pardo

2023-10-12

Introducción

Para desarrollar este experimento se planteo testear el comportamiento del modelo al aplicar **One-Hot Encoding** (OHE) a las variables de los datasets. El objetivo es observar si el modelo mejora su performance al aplicarle OHE a las variables categóricas de los datasets. Además veremos si el modelo se ve afectado por la imputación de los valores faltantes.

Resultados

A simple vista parecería ser que los modelos entrenados con los datasets *Churn* y *Student* no se ven demasiado alterados al hacerles **OHE**, pero al modelo entrenado con el dataset de *Heart Disease* parece alcanzar mayor performance, especialmente al imputarle los valores faltantes, pero además luego de caer a un menor valor mínimo de AUC parece remontar un poco más su performance que aquel modelo al que no se lo preprocesó haciendo OHE.

Como en la figura no se logran observar diferencias muy significativas entre los resultados obtenidos con y sin *One-Hot Encoding* para los modelos de *Churn* y *Student* generamos 3 figuras nuevas una por cada dataset para poder observar mejor los resultados.

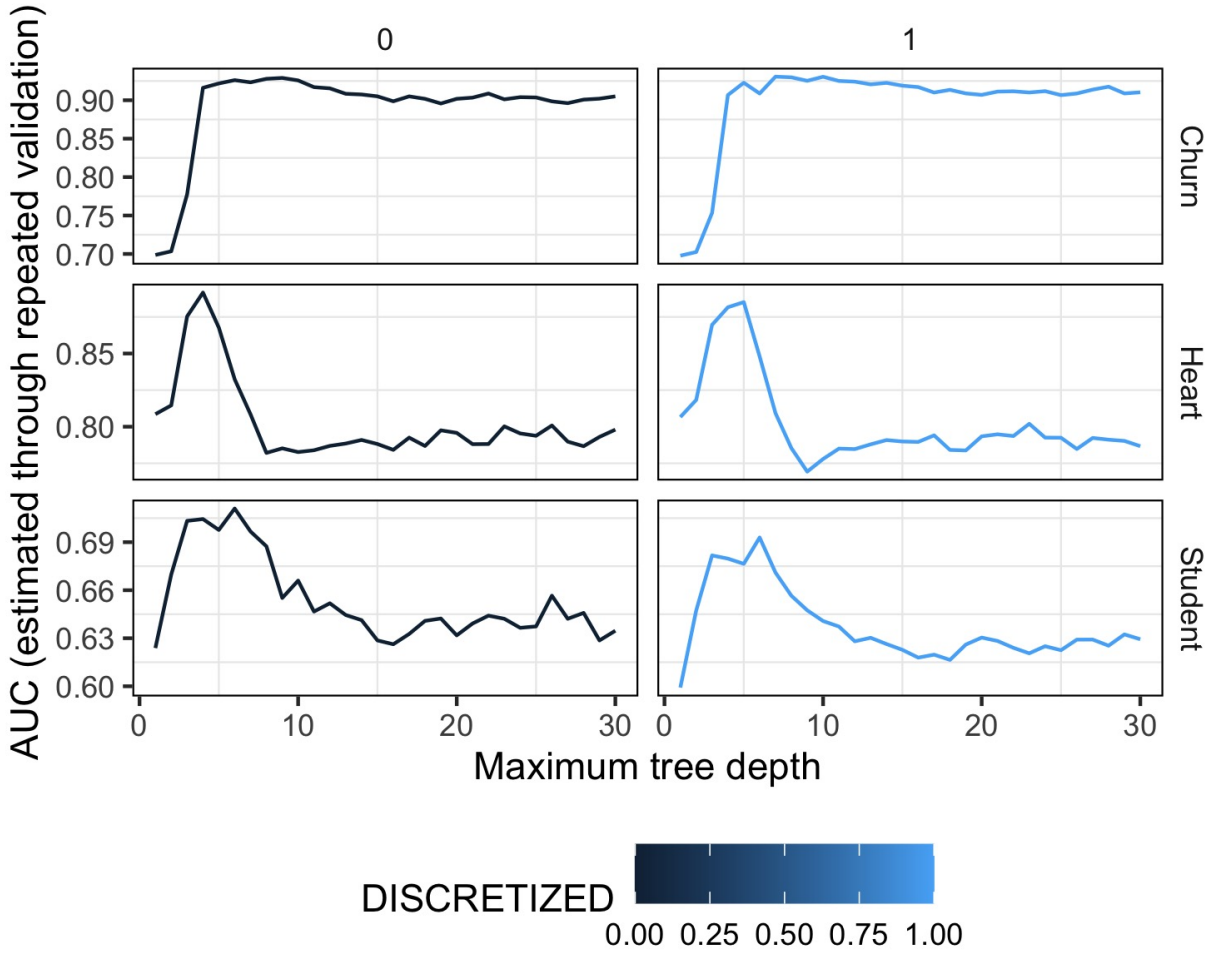


Figure 1: Resultados sin (0/izquierda) y con (1/derecha) One-Hot Encoding

Churn

Como se puede observar en la "*Tabla 1*", para el dataset de *Iranian Churn* el mayor valor de AUC alcanzado para el dataset sin *One-Hot Encoding* se obtiene con un árbol de `maxdepth=5` cuando no se imputan los valores faltantes y con un árbol de `maxdepth=7` cuando si se imputan los valores faltantes con la media.

En cambio, para el dataset con *One-Hot Encoding* se obtiene el mayor valor de AUC con un árbol de `maxdepth=8` cuando no se imputan los valores faltantes y con un árbol de `maxdepth=7` cuando se imputan los valores faltantes con la media. Además parecería ser que luego de llegar a sus respectivos máximos para cada caso, los modelos con *One-Hot Encoding* se mantienen más cercanos al máximo inicial que aquellos sin *One-Hot Encoding*. Esto podría ser porque su pico no fue dado por volverse overfitteado.

Table 1: Churn Max AUCs

maxdepth	mean_auc	DISCRETIZED
9	0.9290292	No
7	0.9306897	Yes

Heart

Como planteamos anteriormente, al aplicar **OHE** al dataset de *Heart Disease* previo a entrenar los modelos, podemos observar en la Figura Heart que dichos modelos "exageran" sus métricas de AUC respecto a los modelos entrenados con el dataset sin *One-Hot Encoding*. Esto se debe a que al aplicar **OHE** a las variables categóricas, estas pasan a ser numéricas y por lo tanto el modelo puede entrenarse con ellas. En contraste con el dataset de *Churn* que contiene 3/13 variables categóricas, el dataset *Heart Disease* contiene 5/11 variables categóricas lo que parece mejorar la performance.

Para los modelos entrenados con el dataset de *Heart Disease* podemos observar en la Tabla 2 como para tanto imputar datos faltantes y como para hacer o no **OHE**, los arboles alcanzan todos sus máximos valores de AUC cuando su `maxdepth=4`. Sin embargo, el valor medio de sus AUC entre los repetidos entrenamientos de cada modelo con dichos hiper-parametros es mayor para los modelos con *One-Hot Encoding* (~0.894 sin imputar y ~0.881 imputando) que para los modelos sin *One-Hot Encoding* (~0.871 sin imputar y ~0.873 imputando).

Aunque a simple vista parecía ser que el modelo alcanza menores mínimos al hacer **OHE**, en la Tabla 3 podemos observar que la diferencia es minúscula.

Table 2: Heart Max AUCs

maxdepth	mean_auc	DISCRETIZED
4	0.8916253	No
5	0.8850997	Yes

Table 3: Heart Min AUCs

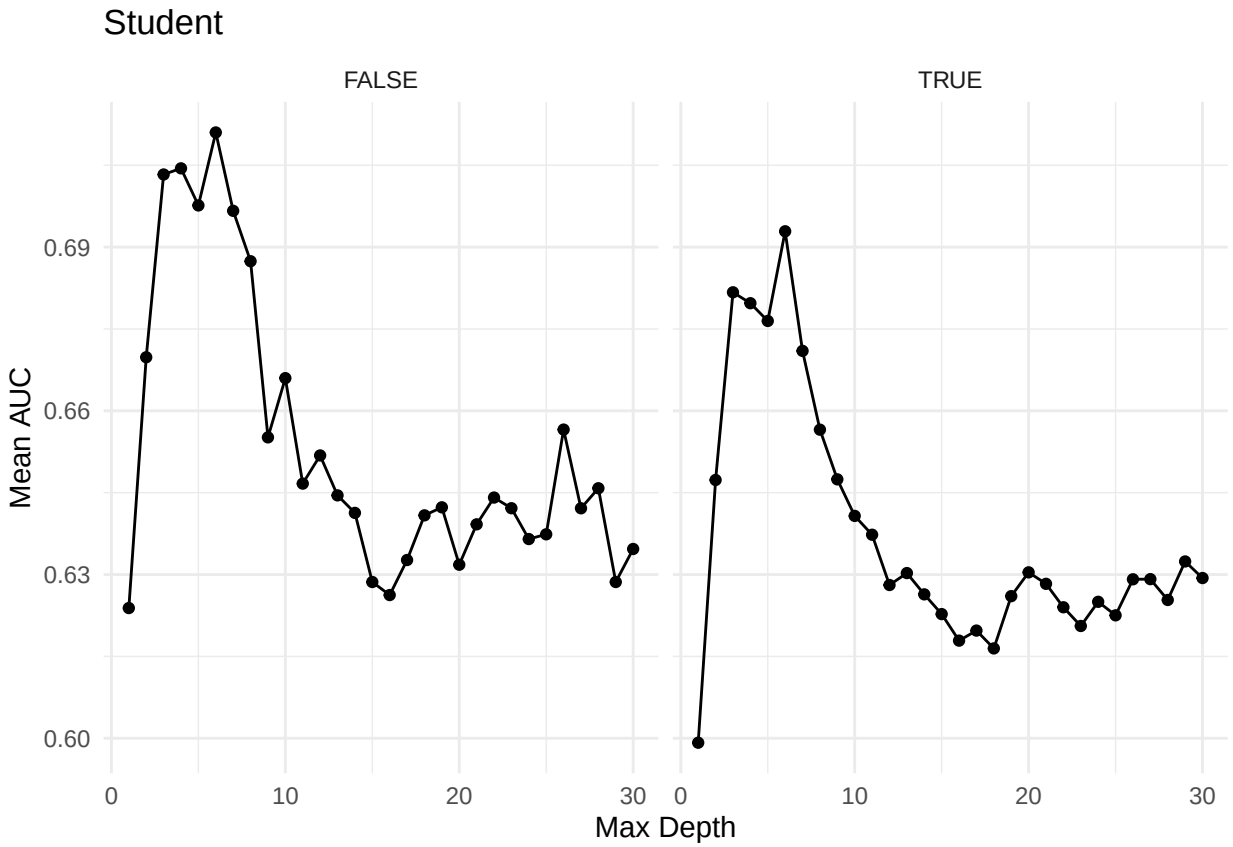
maxdepth	mean_auc	DISCRETIZED
8	0.7821061	No
9	0.7693144	Yes

De todas formas, las "colas" para los modelos de mayor profundidad, parecen tener mayores valores medios de AUC que para los modelos sin *One-Hot Encoding*, casi como si el modelo fuese a repuntar. Esto podría ser porque el modelo no se volvió overfitteado y por lo tanto puede seguir aprendiendo de los datos, aunque al ser mínimo no nos pareció suficiente como para avanzar.


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

Student

El último caso no respeta lo planteado para los datasets anteriores, ya que al aplicarle **OHE** al dataset de Student Performance previo a entrenar los Arboles de Decisión, podemos observar que dichos modelos no alcanzan los mismos valores medios máximos de AUC que aquellos a los que no se le aplico **OHE**. Esto creemos que se debe a que el dataset aunque no contenía muchas variables numéricas (13/30) con respecto al total, el resto de las variables categóricas se veían distribuidas entre binarias (13/30) y nominales (4/30), lo cual parecería ser empeora la performance del modelo al aplicarle **OHE**. Probablemente ocurra por convertir las variables binarias más que por las nominales.



Conclusiones

Aunque no parecería haber un patrón discernible entre los comportamientos de los modelos con y sin **OHE** entre los distintos datasets, una buena interpretación es que sus curvas de performance se ven "escaladas", ya sea para mejor como en "Heart Disease" como para peor en "Iranian Churn" y "Student Performance". Nuestra interpretación esta basada en que este "escalado" esta dado 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 beneficiado por el **OHE** y en aquellos cuya proporción era menor (Churn 3/13), el modelo se veía perjudicado, aunque mínimo, por el **OHE**.