FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1 - Checkpoint 2

Grupo 21 - 'df' is not defined

[7506/9558] Organizacion de Datos
Primer cuatrimestre de 2023

Alumno	Padron	Email
Martin Pata Fraile de Manterola	106226	mpata@fi.uba.ar
Tobias Emilio Serpa	108266	tserpa@fi.uba.ar
Juan Francisco Cuevas	107963	jcuevas@fi.uba.ar

Primero que nada, comenzamos la búsqueda del mejor modelo probando que *encoding* era el mejor para conseguir los scores, en este caso probamos *One Hot Encoding* y *Label Encoding*. Haciendo esta prueba nos dimos cuenta de que usando OHE pudimos obtener un mejor resultado, ambas pruebas hechas únicamente con una poda.

Luego hicimos un análisis para ver qué métrica nos daba el mejor modelo y llegamos a la conclusión de que *f1 score* era el mejor, ya que es el que más métricas utiliza (recall y accuracy). Esto garantiza que, al predecir, se obtiene un resultado más aproximado.

Una vez elegido el método de encoding y la métrica con la cual queríamos evaluar nuestros modelos, procedimos a hacer diferentes pruebas de modelos, utilizando algunos árboles con *Random Search* y otros con *Grid Search*. Observamos que los resultados obtenidos mediante Random Search eran muy similares a los obtenidos mediante Grid Search.

También se pudo observar que, al usar una cantidad exagerada de folds, en muchas ocasiones el modelo empeoraba, mientras que mantener un rango de folds entre 5 y 10 daba mejores resultados. Algo similar pasaba con las combinaciones posibles: para una cantidad determinada de variables y datos, a veces no tenía sentido probar un número alto de combinaciones, ya que esto no mejoraría necesariamente el modelo y podría, en cambio, empeorarlo o agregar tiempo de ejecución innecesario. Por lo tanto, los mejores modelos que obtuvimos fueron los que mantuvieron una coherencia entre los parámetros elegidos, la cantidad de combinaciones y folds.

Hay modelos donde el score es muy bueno, pero el árbol queda demasiado *overfitteado* y con una mala relación entre las variables. Por eso, en algunos casos, la poda era muy importante, pero no siempre, ya que si se tenía pocos niveles o se podaba demasiado, el resultado no era un buen modelo.

En conclusión, el mejor modelo que pudimos lograr fue mediante Grid Search, donde se puede ver un árbol bastante grande (en este caso, no *overfitteado*, ya que se podían observar buenas decisiones), con un score y una predicción bastante buena también. No consideramos que sea un mal modelo o que se haya hecho *overfitting*. Fue un modelo donde únicamente usamos la poda y una alta cantidad de combinaciones, pero que terminó dando un buen resultado. Además, elegimos un segundo árbol elaborado con Random Search, con una mejor poda, obteniendo una profundidad mucho menor y una buena performance (salvando las distancias con el primer árbol, el cual posee una diferencia notoria, pero es lógico ya que posee una profundidad notablemente mayor).