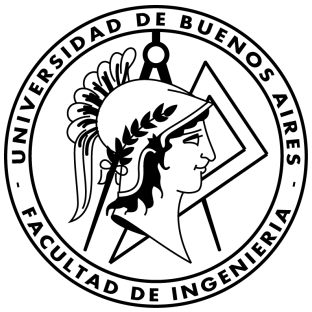

Trabajo Práctico 1
Reservas de hotel - Resumen checkpoint 2

[75.06] Organización de datos
Primer cuatrimestre 2023
Grupo 19: Sudanalytics

Integrantes

Nombre	Padrón	Mail
Re, Adrián Leandro	105025	are@fi.uba.ar
Lorenzo, Luciano Andrés	108951	llorenzo@fi.uba.ar
Toulouse, Alan	105343	atoulouse@fi.uba.ar
Tonizzo, Nicolas	107820	ntonizzo@fi.uba.ar

Fecha de entrega: 27/04/2023

Índice

1. Introducción

Ya tenemos procesado nuestro dataset teniendo en cuenta las conclusiones que sacamos en la instancia anterior, ahora vamos a buscar un árbol de decisión que mejor logre predecir los datos del dataset de prueba según la métrica "F1 score". El modelo generado se guardará en un archivo.

2. Entrenamiento de árboles

En primera instancia, separamos el dataset de entrenamiento en una proporción de 0.8/0.2, para reducir el posible overfitting que el modelo pueda tener. De esta manera nos podemos garantizar que el modelo esta generalizando correctamente, ya que podemos evaluar el modelo en datos que nunca ha visto.

Para el próximo paso, definimos los parámetros que vamos a usar en nuestro árbol de decisión, en este caso son la profundidad máxima, el mínimo de muestras para hacer una pregunta, el mínimo valor de un nodo para convertirse en hoja y la intensidad de la poda del árbol. El modelo a entrenar será una RandomizedSearchCV.

Por ultimo evaluamos el modelo en el test de validación y calculamos métricas de precisión y recall, creando la matriz de confusión. Mas que nada para ver como se comporta el modelo, ya que al solo optimizarlo para f1-score, podría pasar que tenga un muy buen recall pero una pésima precisión (o al revés), y no es lo que buscamos. Esta es la matriz de confusión del modelo entrenado (para el dataset de validación):

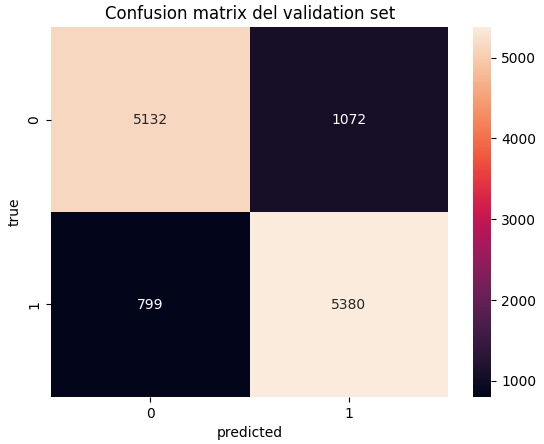

Figura 1: Matriz de confusión

3. Conclusiones

Llegamos a la conclusión de que los arboles de decisión son muy poderosos para su simplicidad, ya que el score que obtuvimos fue muy elevado (mas de lo que pensábamos en un principio). Pero lo que mas interesante que consideramos del decision tree, es que es muy fácil averiguar las decisiones que toma el modelo (plotteando el arbol), y averiguar cuales son las features que mas ayudan a predecir el target.

4. Referencias

- Paper "Hotel booking demand datasets"proporcionado por la cátedra.