Universidad de
Buenos Aires

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

75.06/95.58 – Organización de Datos

Curso 02
1C 2023

Trabajo Práctico 1: Reservas de Hotel
Checkpoint 2

Grupo 14

Integrantes:

- *BOTOSHANSKY, ALAN* *108933*
- *BOTOSHANSKY, IVAN* *108897*
- *NIEVA, ULISES TADEO* *107227*

Corrector: Nacho

Fecha de entrega: *27* de *abril* del 2023

Universidad de
Buenos Aires

Resumen

En este reporte nos enfocaremos en detallar el proceso que realizamos para hallar el árbol de decisión con mejor performance y los resultados que obtuvimos a partir de este.

Desarrollo

Como primera medida importamos el dataset de entrenamiento con el preprocesamiento realizado en el Checkpoint 1, y luego aplicamos el mismo tratamiento al dataframe de prueba para que ambos estén en los mismos términos.

Luego, aplicamos One Hot Encoding a ambos datasets. Dado que algunas variables del dataframe de entrenamiento poseían valores que luego no se encontraban en el de prueba, y viceversa, optamos por quedarnos únicamente con las variables dummies que pertenecían a ambos.

Proseguimos con la búsqueda de los mejores hiperparámetros para el modelo, mediante K-Fold Cross Validation, siendo estos parámetros la máxima profundidad del árbol, la estrategia de poda y el criterio de elección de atributos de cada nodo. La cantidad de folds elegidos fueron 5. En cuanto a la métrica seleccionada, determinamos que sea F1 Score, debido a que el contexto del problema requería tener una cantidad de falsos positivos y falsos negativos balanceada, y lo más reducida posible para ambos casos.

Tras llevar a cabo un Grid Search, la combinación de hiperparámetros arrojada fue una máxima profundidad de 27, entropía como criterio de elección y un costo de complejidad de poda de aproximadamente 0.0001724.
Al entrenar un árbol con estos parámetros, usando un 80% del dataframe de entrenamiento, y a continuación realizar una predicción en el 20% restante, el F1 Score resultante fue 0.8545.

Observando tanto el gráfico del árbol resultante como un listado de los atributos de clasificación más importantes, apreciamos que el tipo de depósito es la variable más preponderante, siendo que lo primero que el modelo pregunta es si se hizo un depósito por el costo total de la reserva o no. La importancia de esto se traduce en que casi el total (8101 de 8134) de las reservas cuyo depósito fue por el costo total de la reserva fueron canceladas.

Respecto al resto de las métricas, conseguimos una precisión de 0.85447809 y un recall de 0.85447814, un equilibrio pretendido y esperable, dada la selección de F1 Score como optimizadora de la búsqueda de los hiperparámetros.

Conclusión

Para poder conseguir el modelo previamente detallado requerimos, además de los pasos mencionados, probar con modelos basados en versiones más simples del dataset para encontrar un piso de F1 Score a superar, realizar distintas búsquedas utilizando tanto Grid Search como Random Search y llevar a cabo podas mayores con el objetivo de analizar la diferencia de resultados con árboles menos complejos. Las distintas combinaciones y valores de hiperparámetros utilizados mostraron valores de F1 Score bastante similares, siendo el aquí expuesto el mejor conseguido.