Trabajo Práctico 1 - CHP2
Reservas de Hotel

[75.06/95.58] Organización de Datos
Primer cuatrimestre de 2023

ALUMNO	PADRON	CORREO
BARTOCCI, Camila	105781	cbartocci@fi.uba.ar
HERBAS, Enzo	103747	jeherbas@fi.uba.ar
PATIÑO, Franco	105126	fpatino@fi.uba.ar

Sobre el modelo construido final y sus iteraciones previas

Nuestro modelo tuvo varias iteraciones, tanto respecto a los rangos de hiperparámetros utilizados para la búsqueda y las múltiples ejecuciones para optimizarlos, como respecto a la utilización de distintas técnicas y combinaciones de hiperparámetros. El modelo final consta de un árbol de decisión con parámetros optimizados mediante la técnica de Random Search Cross Validation, utilizando 7 folds. La elección del número de folds tiene relación con que consideramos que teníamos una buena cantidad de datos de prueba para subdividirla en 7 subconjuntos, obteniendo así resultados consistentes y reproducibles, y aumentando la confianza en los resultados. Además, buscamos mantener una evaluación precisa del rendimiento de nuestro modelo sin subestimar la complejidad del mismo, por lo que 7 nos pareció una cantidad adecuada para probar maximizar el uso de datos y reducir la variabilidad en la evaluación.
Decidimos utilizar una proporción del 70 % de datos de train y 30 % de test. Tras haber entrenado el árbol con distintos hiperparámetros, tanto utilizando Random Search como Grid Search para su optimización y variando la cantidad de folds, decidimos quedarnos con un estimador que utiliza Gini y tiene una profundidad de 24, y una cantidad mínima de muestras por hoja de 5, entre otros hiperparámetros y podando el resultado final para evitar el overfitting.
En cuanto a la importancia de características, nuestro modelo consideró como más importante la columna de tipo de depósito, seguida del tiempo de antelación con el cual se hace una variable y de si el tipo de cuarto reservado coincide con el asignado. Con estas tres, observamos que puede hacer primeras predicciones con bastante seguridad.
Las métricas obtenidas serán detalladas en el apartado siguiente.

Sobre los resultados obtenidos

Para nuestro modelo, optamos por optimizar la precisión manteniendo un recall considerablemente bueno. Buscamos lograr un equilibrio entre la minimización de los falsos positivos y los falsos negativos y obtener un modelo preciso y confiable para predecir las cancelaciones de las reservas del hotel.
Estratificamos nuestros conjuntos proporcionalmente, de tal manera que el conjunto de train quedó dividido en un 49 % de cancelaciones (21559 registros) y el resto de no cancelaciones (21777 registros). De todos estos registros, nuestro modelo predijo como verdaderos negativos 7744 registros y como verdaderos positivos 7028. A su vez, predijo 1451 falsos positivos y 2350 falsos negativos. Con este modelo, obtuvimos aproximadamente las siguientes métricas:

- 79.5 % accuracy
- 75 % recall
- 82.88 % precisión
- 78.7 % f1 score

A su vez, calculamos el área bajo la curva ROC, obteniendo un valor aproximado de 0.88.
Podemos concluir que nuestro modelo logra predecir correctamente el 82.8 % de reservas realmente no canceladas y el 74.9 % de las que se cancelan. Por otro lado, el accuracy nos dice que clasifica correctamente el 79.5 % de las reservas. El f1 score sugiere que hay un buen equilibrio entre precisión y recall. En cuanto al área bajo la curva ROC, el valor de 0.88 indica que el modelo tiene una buena capacidad para distinguir entre reservas canceladas y no canceladas.