Checkpoint 2 - Grupo 24

Introduccion

Comenzamos haciendo ingeniería de características en el dataset de train como por ej: juntar los países más relevantes, juntar la cantidad de personas total por habitación, entre otras cosas. Creamos varios tipos de árboles aplicando distintas técnicas como K-Fold Cross validation, split tree y luego optimizamos los valores con Random search.

Construcción del modelo

- Optimizacion de hiperparámetros:
  Como optimización de hiperparametros utilizamos random search, como bien mencionamos anteriormente, sin embargo no nos fue de mucha ayuda, ya que o bien mantenía el f1 score bajaba apenas el score.

  Los parámetros optimizados fueron los siguientes:
  min_samples_split, min_samples_leag, max_depth, criterion y ccp_alpha.

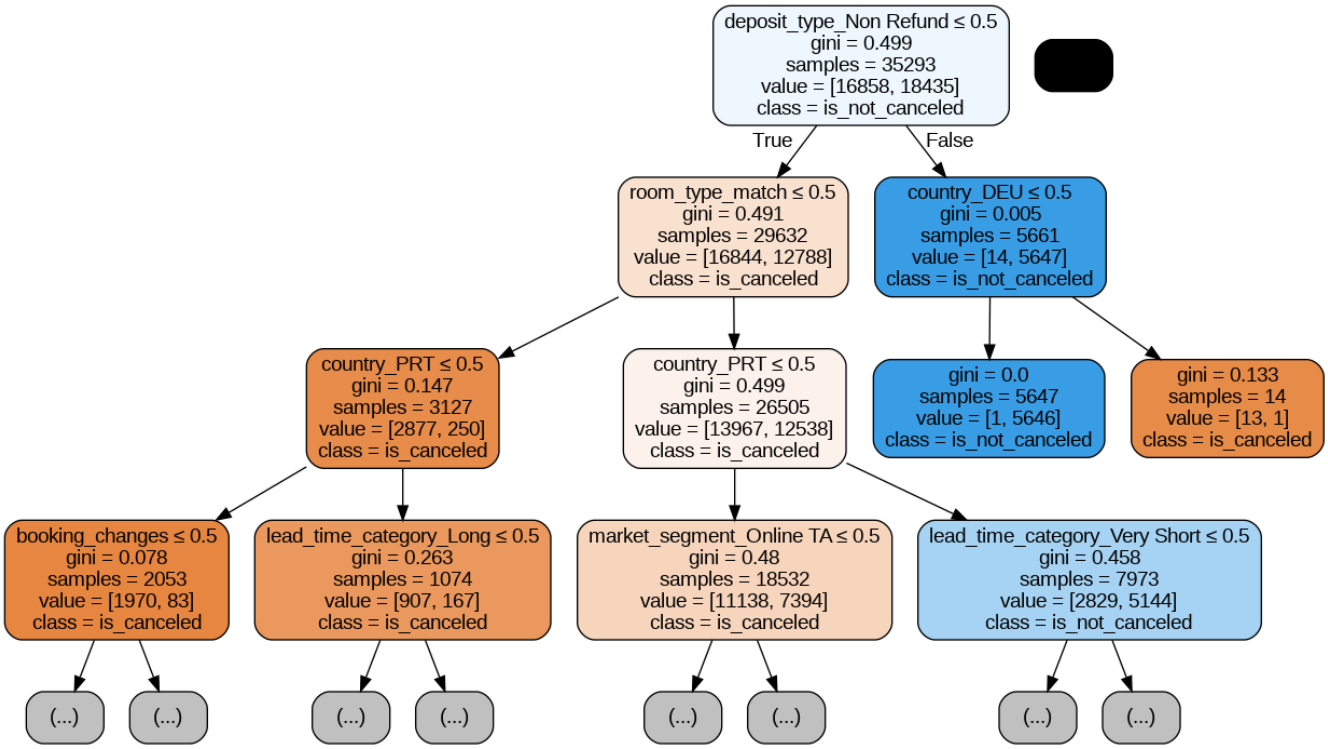
- Utilizamos K-fold Cross Validation y los fold utilizados fueron **15**.

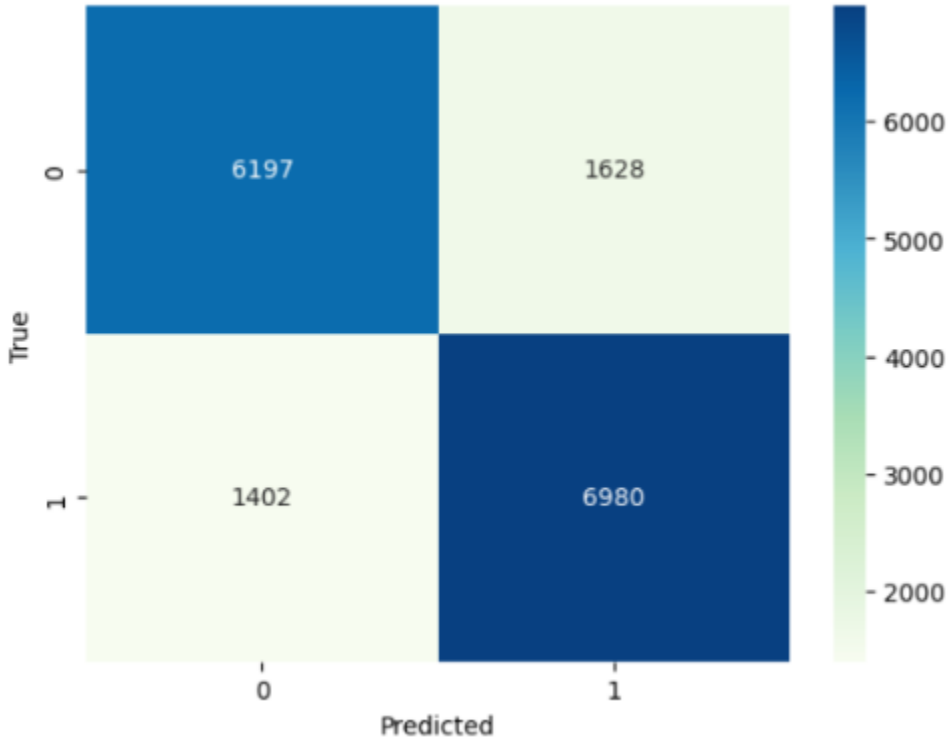
- Nosotros consideramos adecuado optimizar la métrica de **f1 score** para buscar los mejores hiperparametros, ya que es el que utiliza kaggle para evaluar la predicción.

- Realmente no hubo una mejora significativa desde la métrica inicial a la final y hasta en casos nos ha empeorado, sin embargo esto solo nos paso en nuestra notebook. En el Kaggle mejoraba respecto de nuestro colab pero no superaba nuestra mayor puntuación (realizada sin optimización de hiperparametros).

- Se utiliza un árbol de decisión para clasificar instancias. Comienza en el nodo raíz y se mueve hacia las hojas. La primera variable considerada es "Deposit_type" debido a su alta importancia. Dependiendo de su valor, la instancia se dirige a la izquierda o derecha. Luego se evalúa "Room_type_match" y “country” respectivamente. En el último nodo, se decide la clase “is_canceled”,basándose en las anteriores. Cuanto más fuerte sea el color del nodo, mayor es la confianza en la decisión.

  “Is_not_canceled’ hace referencia a un 0 en “is_canceled”, es decir no cancelo.





Matriz de Confusion

Esta es nuestra matriz de confusión de la mejor predicción presentada en Kaggle.

Se puede observar que tanto los falsos negativos como positivos tienen valores muy similares difiriendo en apenas un 13% y los verdaderos negativos y positivos difieren de un 12,56%, manteniendo así una relación entre ambas predicciones.

Tabla con las las 3 mejores predicciones:

Modelo	F1-Test	Precision Test	Recall Test	Accuracy	Kaggle
modelo_1	0.819739	0.826044	0.8135289	0.814956	0.81967
modelo_2	0.817819	0.828840	0.807086	0.81403	0.81851
modelo_3	0.816732	0.826181	0.799986	0.81201	0.81703

Tareas Realizadas

Integrantes	Tarea
Lautaro Torraca - 108813 Marco Tosi - 107237 Gianluca Negrotti - 108184	En este checkpoint la mayor parte del tiempo en pair-programming a excepción de los últimos 2 días, los cuales tratamos de forma individual aumentar el f1-score.