Organización de datos - TP 1

Grupo 10 - **Checkpoint 2:**

A lo largo de este checkpoint 2, generamos distintos modelos. El que finalmente entregamos es el que mejor Score F1 obtuvo.

Antes de entrenar el modelo de árbol de decisión, realizamos el siguiente preprocesamiento. Utilizamos la técnica de One Hot Encoding para transformar nuestras variables categóricas en variables numéricas. Eliminamos columnas que nos parecieron no relevantes como ‘id’ y ‘country’, entre otras. Y por último, transformamos las variables categóricas ‘country’, 'arrival_date_month', 'reserved_room_type' y 'assigned_room_type' en variables numéricas ordinales.

En cuanto a la construcción del modelo, utilizamos la técnica de K-Fold Cross Validation con 5 folds y la métrica de F1-score para evaluar el rendimiento del modelo.

Consideramos que F1 es adecuada para buscar los hiperparámetros ya que consigue un equilibrio entre Recall y Precision. De esta forma no estaríamos teniendo muchos falsos negativos por querer ser muy precisos, ni muchos falsos positivos por querer abarcar la mayor cantidad de positivos.

Probamos diferentes combinaciones de hiperparámetros del modelo. Finalmente, utilizamos la técnica de GridSearch para encontrar los mejores hiperparámetros y obtuvimos los mejores resultados con los parámetros ccp_alpha: 0.0, criterion: 'gini', max_depth: 16.

Atributos más importantes obtenidos:

- Deposit_type_non_Refund
- Lead_time
- Market_segment_online_ta
- country_numero