Checkpoint 2 - Grupo 28

Introducción

Nuestro objetivo del checkpoint, es encontrar un árbol de decisión que mejore el score F1, esta es una métrica que pondera otras dos métricas que son precisión y recall. Debemos recordar que recall es una métrica que mide la capacidad de nuestro modelo de encontrar correctamente “true positive” (TP) y precisión es una métrica que mide los “true positive” con todos los casos positivos. Nosotros llegamos a la conclusión que f1 score es una buena métrica para determinar si es un árbol de decisión aceptable.

Antes de entrenar el modelo de árbol de decisión, nosotros tuvimos que realizar un preprocesamiento de los datos de los datasets.

Para realizar un análisis exhaustivo de distintos árboles y poder llegar a nuestra mejor opción, cada uno de los integrantes del grupo realizo una limpieza distinta de los datasets. Esta decisión se dio porque al cambiar la manera de limpieza y transformación del dataset, nos cambiaba bastante la situación del análisis.

Modelo 1

Archivo en el repositorio: 7506R_TP1_GRUPO28_CHP2_ENTREGA_N1.ipynb

La transformación del dataset tuvo unos pequeños cambios del checkpoint 1. Esos cambios se enumeran a continuación:

- Variable Agent: Se utiliza en el análisis porque termina siendo una variable importante en el análisis.
- Variable Country: Esta variable tiene muchos países diferentes (alrededor de 197) en sus reservas, entonces decidimos de agrupar en los top 5 países con más reservas. El resto de las reservas que no estén en ese top 5, se reemplaza con la palabra “otro” en el dataset. Entonces nos queda solo seis categorías para luego realizar el hot encoding.
- Variable previous_bookings_not_canceled: Se recupera esta variable porque tiene información importante para el entrenamiento y predicción.
- Variable assigned_room_type: Se elimina esta columna ya que proporciona la misma informcacion que reserved_room_type

Se realiza un entrenamiento sin optimizar los hiperparámetros, se utiliza un tamaño de test de 0.15, semilla de 28, criterio Gini y max_depth de 20. Con estos parámetros se llegó a una métrica f1-score de 0.852.

Luego se realiza un entrenamiento optimizando los hiperparámetros utilizando el modelo de Random Search Cross Validation, los parámetros que se utilizaron:

- Folds: 10
- Splitter: best
- Min_samples_split: 16
- Min_samples_leaf: 16
- Max_depth: 21
- Criterio: entropy
- Class_weight: balanced

F1-Score: 0.8535

Modelo 2

Archivo en el repositorio: 7506R_TP1_GRUPO28_CHP2_ENTREGA_N2.ipynb

La transformación del dataset tuvo unos pequeños cambios del checkpoint 1. Esos cambios se enumeran a continuación:

- Variable arrival_date_month: se transforman los meses en números para facilitar la cantidad de columnas.
- Variable meal: Se elimina esta columna porque se considera que no proporciona información valiosa para el árbol.
- Variable country: Se utiliza todos los valores en este modelo (a diferencia del modelo 1 que se agruparon valores).

Se realiza un entrenamiento sin optimizar los hiperparámetros, se utiliza un tamaño de test de 0.3, semilla de 2, criterio Gini y max_depth de 20. Con estos parámetros se llegó a una métrica f1-score de 0.842. Luego se realiza una poda y se utiliza un valor de ccp_alpha de 0.01 y tengo una métrica de f1 score de 0.7788

Modelo 3

Archivo en el repositorio :7506R_TP1_GRUPO28_CHP1_ENTREGA_N3.ipynb

La transformación del dataset tuvo unos pequeños cambios del checkpoint 1. Esos cambios se enumeran a continuación:

- Variable distribution_channel: Esta variable se elimina porque no encontramos relación para la predicción del árbol
- Variable assigned_room_type: Se elimina esta columna ya que proporciona la misma informcacion que reserved_room_type
- Variable country: Se utiliza todos los valores en este modelo (a diferencia del modelo 1 que se agruparon valores).

Se realiza un entrenamiento sin optimizar los hiperparámetros, se utiliza un tamaño de test de 0.15, semilla de 28, criterio Gini y max_depth de 20.

Luego se realiza un entrenamiento optimizando los hiperparámetros utilizando el modelo de Random Search Cross Validation, los parámetros que se utilizaron:

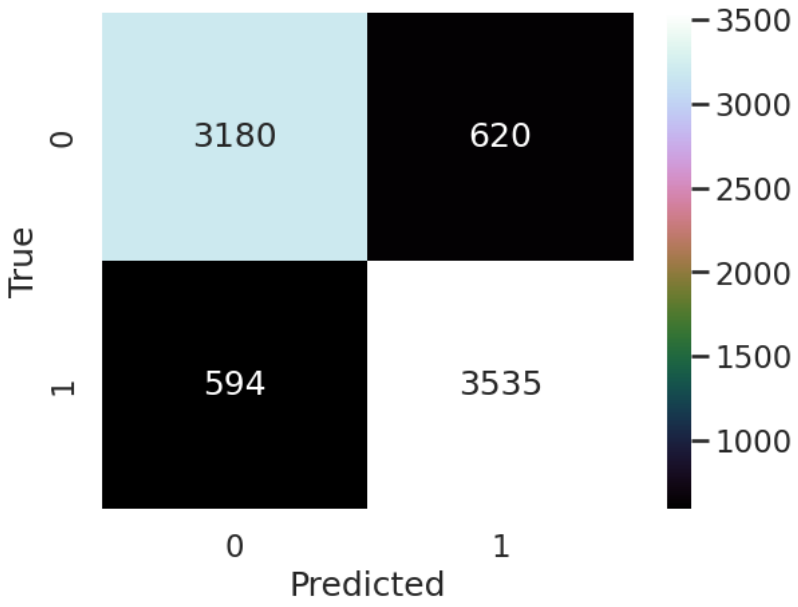
- Folds: 5
- Min_samples_split: 35
- Min_samples_leaf: 9
- Max_depth: 24
- Criterio: gini
- Ccp_alpha: 0.0001

F1-Score: 0.84531

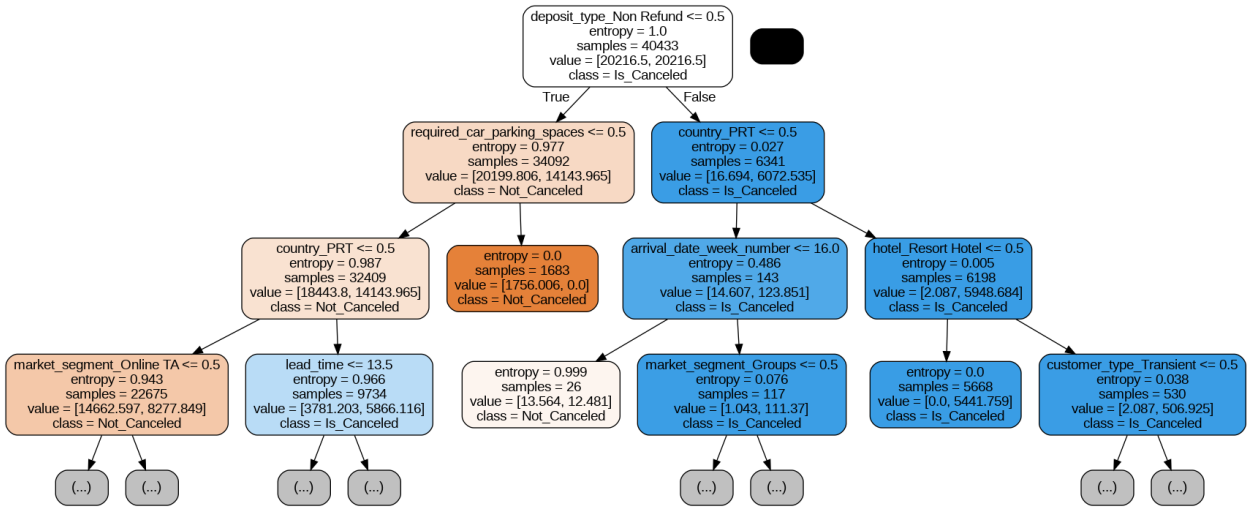
Cuadro de Resultados

Modelo	F1-Test	Presicion Test	Recall Test	Accuracy	Kaggle
modelo_1	0.8535	0.8508	0.8561	0.8469	0.8445
modelo_2	0.7788	-	0.7823	0.7747	0.6705
modelo_3	0.84531	0.8398	0.8714	0.85377	0.8434

Matriz de Confusión del Modelo 1



Árbol de Decisión del modelo 1



Tareas Realizadas

Integrante	Tarea
Jurgens, Cecilia Ines	Modelo 1

Schipani, Martin	Modelo 2
Soto, Marilyn Nicole	Modelo 3