Checkpoint 2 - Grupo 23

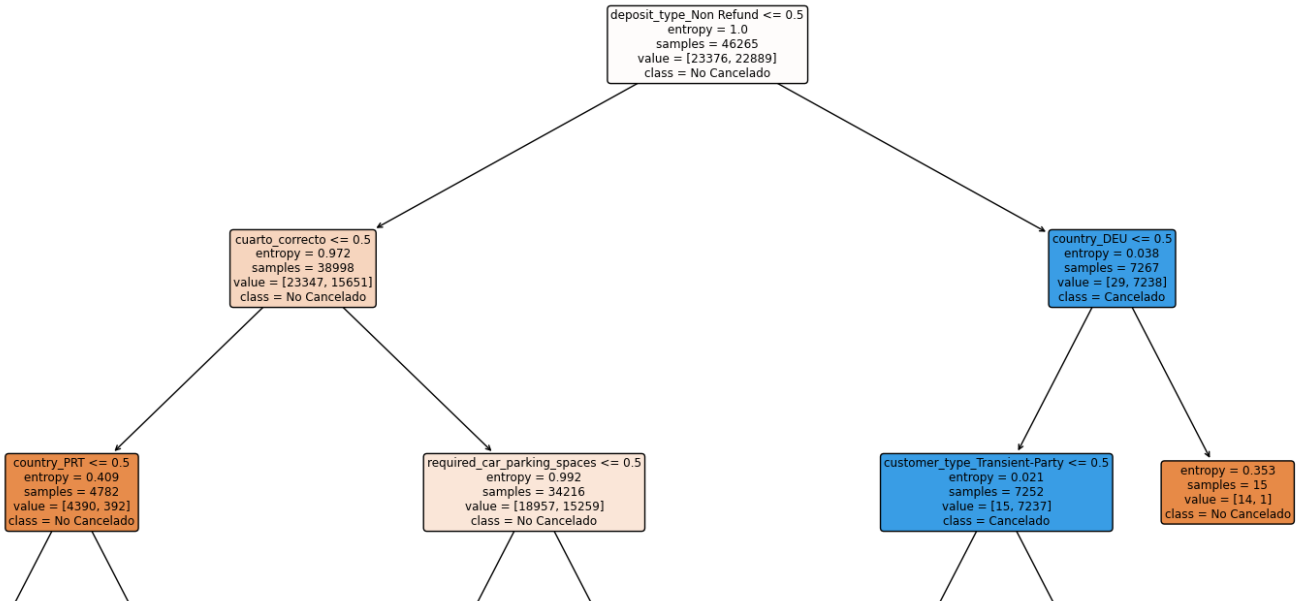
Introduccion

La herramienta utilizada en esta sección del trabajo práctico fueron los árboles de decisión. La principal técnica utilizada para este modelo fue la de One Hot Encoding, esto permite al árbol utilizar variables categóricas a la hora de predecir. Cabe recalcar que también tuvimos que agrupar variables como country o *agent_id* de cierta manera debido a que existían muestras de estas variables que estaban en el dataset de train que no estaban en el de test y viceversa, lo cual corrompía el modelo.
Primero, construimos árboles simples con pocas variables para darnos una idea de cómo iba variando su performance. Luego, en el último árbol, fue donde íbamos testeando variable por variable para ver si mejoraba el modelo, utilizando las relaciones que habíamos notado en el análisis exploratorio. Por último, buscamos los mejores hiperparametros para nuestro modelo por medio de RandomCvSearch.
Además, encontramos que era conveniente para el modelo eliminar los registros (de un df auxiliar) que contienen un *agent_id* que aparece menos de 5 veces en todo el dataset.

Construcción del modelo

- Hiperparametros: Decidimos hacer una corrida larga con un *n* elevado para poder notar diferencias ya que con pocos no notamos diferencias. Los parámetros optimizados fueron *max_depth, ccp_alpha , min_samples_leaf*.
- K-fold: Para ver cuantos folds eran los óptimos, hicimos un bucle de 3 a 7 folds para ver cuantos folds daban el score más alto, el resultado fueron 6 folds, así que usamos ese número.
- Métricas: La métrica que utilizamos para buscar los hiperparametros fue **f1**. Utilizamos esta métrica ya que tiene en cuenta tanto los falsos positivos como los falsos negativos y los aciertos.
- Progresión del modelo: Los modelos que creamos mejoraron sustancialmente, pasando de un f1 score de 0,715 a 0,8674. Esto fue optimizando hiperparámetros y variables utilizadas
- Arbol de decision: Los primeros nodos eran predecibles ya que con el análisis previamente hecho, sabíamos que las variables con mucho poder de decisión

eran *deposit type, required car_parking_spaces, cuarto_correcto* y las reservas provenientes de Portugal también



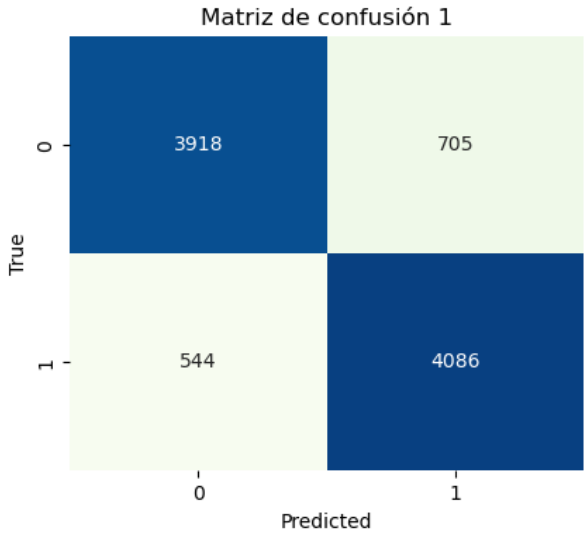
Cuadro de Resultados

Métrica X: Decidimos testear el f1 score del conjunto de train para evaluar overfitting

Modelo	F1-Test	Precision Test	Recall Test	Metrica X	Kaggle
modelo_1	0,79255	0,8065	0,7844	0,81649	0,78247
modelo_2	0,8257	0,83744	0,8143	0,8317	0,82512
modelo_3	0,8674	0,8528	0,8825	0,8768	0,85665

Matriz de Confusion

Podemos ver que a pesar que es un buen modelo, el mismo tiende a dar falsos positivos.



Tareas Realizadas

Dado que somos solo dos integrantes, decidimos ir tarea a tarea, tomando decisiones en conjunto y dividiéndonos la carga de cada parte a la mitad.

Integrante	Tarea
Mateo Vroonland	Todo
Juan Pablo Carosi Warburg	Todo