Checkpoint 2 - Grupo 17

Systeam

Integrantes

Del Rio, Juan Sebastián - 103337

Brizuela, Sebastián - 105288

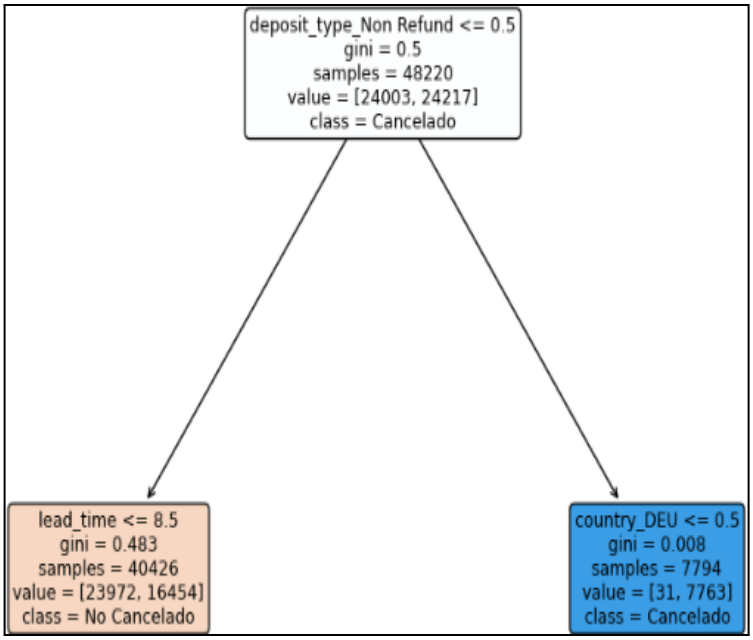
Agha Zadeh Dehdeh, Lucía - 106905

Fecha de entrega: 5 de Octubre de 2023

Introducción

Se optó por iniciar la construcción de árboles de decisión utilizando hiperparámetros arbitrarios como primera elección. Posteriormente, para mejorar su rendimiento y precisión, se implementó la técnica de validación cruzada k-fold con RandomizedSearchCV. Además, se incorporó la técnica de poda en el proceso.

Se determinó como nuevas variables relevantes “market_segment“, “distribution_channel“, “customer_type“, “required_car_parking_spaces“ al analizar nuevamente los gráficos, realizar pruebas con los modelos implementados y ver una mejoría en las métricas.



Construcción del modelo

Se optimizaron las siguiente métricas: “criterion”, "min_samples_leaf", "min_samples_split", "ccp_alpha", "max_depth".

En este proceso de búsqueda, se emplearon 8 folds para evaluar diferentes combinaciones de hiperparámetros de manera eficiente.

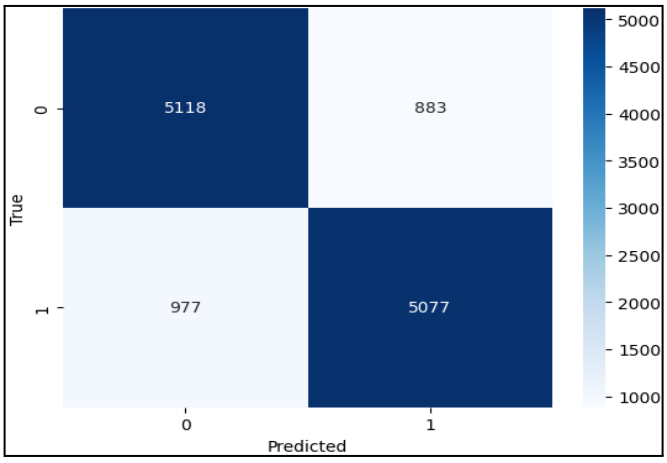
La métrica adecuada que se consideró fue f1-score ya que además de ser relevante para la competencia en Kaggle también combina dos medidas importantes a saber, la precisión y recall.

Desde la primera entrega en Kaggle, que tuvo un valor de 0.8167 y la última que tuvo 0.84082, se puede observar que el modelo mejoró aumentando un %0.03 el puntaje de f1_score.

Obs del gráfico: como regla principal clasifica si el tipo de depósito es NO reembolsable, en caso de serlo como segunda regla filtra según el país de origen en este caso Alemania. En caso de no serlo filtraría según el tiempo de espera.

Cuadro de resultados

Modelo	F1-Test	Precisión	Test Recall	Test Accuracy	Kaggle
modelo1	0.85184	0.84518	0.83861	0.84570	0.84082
modelo2	0.84910	0.84699	0.84489	0.84670	0.84045
modelo3	0.84464	0.84374	0.84555	0.84379	0.84007



Matriz de confusión

En la matriz se observa que el modelo fue capaz de predecir 5077 **Verdaderos Positivos** y 5118 **Verdaderos Negativos**, lo cual son cantidades bastantes buenas respecto de la cantidad total de datos. Por otro lado, el modelo tuvo 977 **Falsos Negativos** y 883 **Falsos Positivos**.

Tareas Realizadas: Se realizaron todas las tareas siempre en grupo por reuniones en meet por lo tanto no hubo una división clara de las mismas.