Checkpoint 2 - Grupo 29 - PyDataBros

Introduccion

Para iniciar con el análisis predictivo, realizamos unas nuevas modificaciones del dataset. Estas fueron necesarias para asegurar que los datos estuvieran en un formato adecuado para alimentar el modelo de árbol de decisión. En el proceso de construcción de estos árboles, y optimización de hiper parámetros, utilizamos k-fold Cross Validation con k=5. Evaluamos el modelo en los conjuntos de entrenamiento y validación usando métricas como F1-score, precisión y recall mostrando la matriz de confusión. Estas implementaciones se llevaron a cabo para obtener el mejor modelo posible y mejorar la precisión en Kaggle.

Construcción del modelo

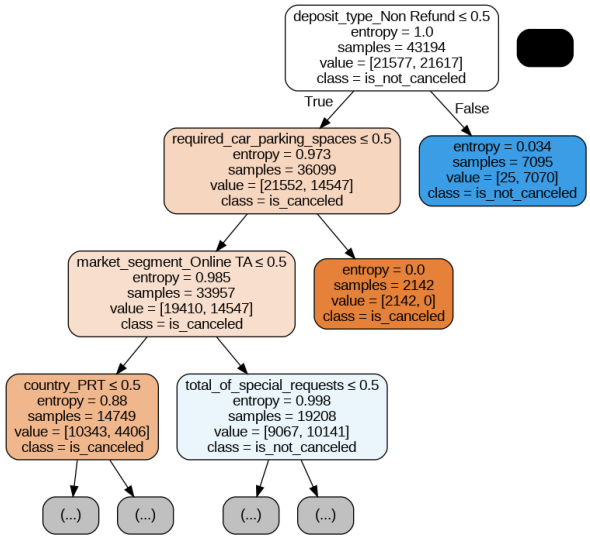
Al momento de realizar el análisis exploratorio, la creación de variables dummy fue esencial para tener un modelo preciso.

- Para la optimización de hiperparametros probamos F1-score. Usamos un gran conjunto de parámetros como: criterion, min_samples_leaf, min_samples_split, ccp_alpha y max_depth.

  Tomamos para el K Fold Cross Validation que la cantidad de splits sea 5.

- Decidimos que la métrica más adecuada para buscar hiperparametros es F1-score
- En el modelo inicial planteado sin procesamiento de hiperparametros, tenemos un resultado en kaggle de 0,6682. Al momento de realizar el análisis de hiperparametros, nuestro resultado en kaggle aumenta considerablemente hasta tener una puntuación de 0,827. Esto es una mejora de más del 20%.

Al tener una gran cantidad de valores para predecir, nuestro árbol de decisión quedaba con una altura demasiado grande, por eso decidimos mostrar esta versión acotada.



Cuadro de Resultados

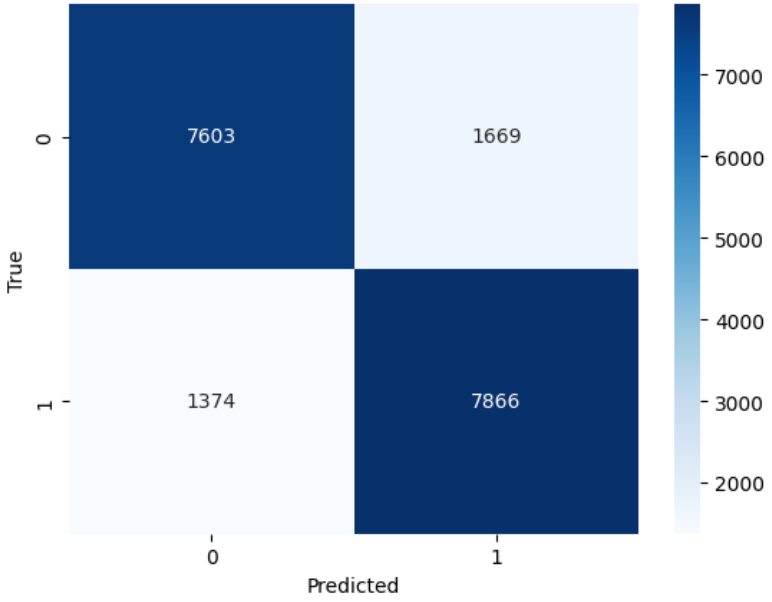
Modelo	F1-Test	Precision Test	Recall Test	Kaggle
modelo_1	0,829417	0.824057	0.83484	0,82585
modelo_2	0,829074	0.824328	0.833874	0.827
modelo_3	0.712146	0.836032	0.620238	0.79222

Matriz de Confusion

True Positives: 7866
False Positives: 1669
True Negatives : 7603
False Negatives : 1374

Vemos que tenemos una buena predicción para los falsos negativos y para los verdaderos positivos. Esto determina un buen modelo de predicción, más precisamente para la métrica Recall.

matriz de confusión del último modelo



Tareas Realizadas

Integrante	Tarea
Delfina Cano Ros	Desarrollo del entrenamiento y predicción
Isidro Borthaburu	Mejorar el desempeño de la predicción y armado del reporte
Martin Wainwright	Desarrollo del entrenamiento y predicción