Checkpoint 2 - Grupo 32

Introduccion

Separamos este checkpoint, en dos notebooks diferentes, una para el tratamiento de datos, y la otra para la realización de los árboles.
Probamos distintas técnicas, las cuales no todas fueron utilizadas, debido a que algunas nos daban pérdida de rendimiento en nuestro modelo.

- Ingeniería de características: manejo de columnas, creación, empalme y eliminación de las mismas.
- Normalizacion y estandarizacion de variables numéricas
- Refactorización de variables y tratamiento de datos
- Validación cruzada y evaluación
- Codificación de Características Categóricas

Modificaciones del dataset:

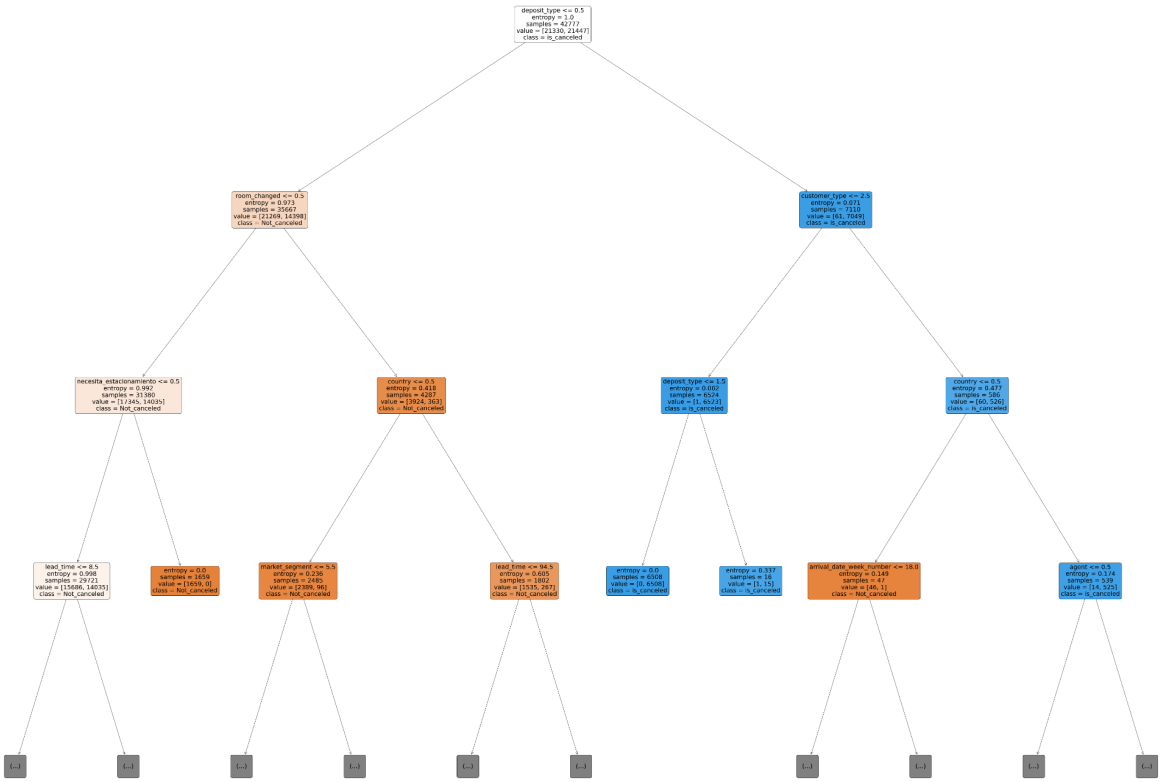
- Se cambió la variable country, a un valor booleano, la cual toma valor TRUE si el país era portugal, FALSE en caso contrario (tomando los valores nulos cómo FALSE), debido a que la mayoría de los datos de esta columna representa dicho país.
- Reconfiguramos variables categoricas, a valores booleanos, o numericos, para que tenga sentido buscar un mayor o un menos
- Agregamos nuevas variables, realizando operaciones aritméticas y comparativas sobre otras variables
- Replicamos las modificaciones en el archivo de test, para poder realizar todo el modelo

Construcción del modelo

Hiperparámetros optimizados:

- K: cantidad de folds
- N: cantidad de iteraciones
- Max_depth: mayor cantidad de ramificaciones que puede tener el arbol
- Min_sample_split: mínima cantidad de casos que le tienen que llegar a un nodo para que se divida
- Min_sample_leaf: mínima cantidad de casos que para poder ser un nodo hoja

Para optimizar los hiperparámetros del modelo, llevamos a cabo un proceso iterativo utilizando K-fold Cross Validation con 7 folds. En cada iteración, de las 20 realizadas, exploramos una nueva combinación de hiperparametros y evaluamos su rendimiento. Calculamos la métrica de evaluación F1-Score en cada fold y luego promediamos los resultados. La combinación de hiperparametros que produjo el mejor rendimiento se seleccionó cómo la configuración óptima. Cómo resultado de esta optimización, observamos una mejora aproximada del 10%, al realizar cambios en el dataset, basados en resultados de las iteraciones. Luego, presenciamos otra mejoría aproximada de un 4% modificando valores de hiperparametros, buscando así el resultado más óptimo.



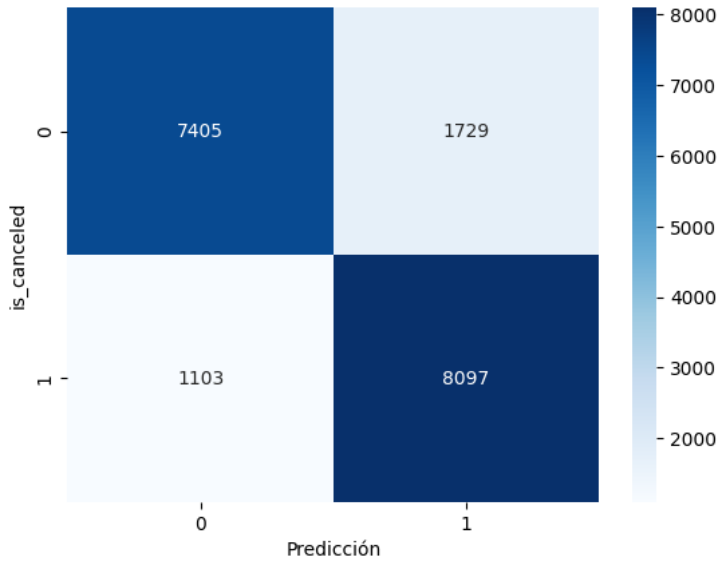
(Ver en el notebook para mayor profundidad)

Cuadro de Resultados

Modelo	F1-Test	Accuracy Test	Precision Test	Recall Test	Kaggle
Modelo simple	0,844	0.845	0.851	0.836	-------
Modelo sin poda	0.798	0.812	0.863	0.742	0,816
Modelo con poda	0.809	0.782	0.720	0.925	0,774
Mejor modelo	0.851	0.846	0.824	0.880	0,843

El mejor modelo, fue un DecissionTree sin poda, debido a que la misma, nos bajaba el valor de nuestra métrica F1-Score.

Matriz de Confusion



Tareas Realizadas

Integrante	Tarea
Daniel Agustin Marianetti	Tratamiento de datos Ajuste de hiperparámetros (iteraciones)
Ezequiel Lazarte	Armado de arboles Confección de la notebook
Franco Ezequiel Rodriguez	Ajuste de hiperparámetros (iteraciones) Armado de reporte