Checkpoint 2 - Grupo GPWIN 16

Introduccion

Breve comentario de técnicas exploradas, pruebas realizadas y si efectuaron nuevas modificaciones sobre el dataset. Cualquier implementación realizada por el equipo se debe detallar en esta sección.

Construcción del modelo

A fin de realizar las predicciones, se llevaron a cabo 3 modelos. Los primeros dos partiendo de los datos del checkpoint 1 en el cual se depuro el dataset eliminando la variable “Company”, realizando un proceso de imputación MICE en los valores nulos y eliminando la totalidad de los outliers.

- El primer modelo, consistió en realizar una segmentación del dataset Train, en 70/30 (entrenamiento/test). Luego se realizó una elección arbitraria de hiperparametros para la construcción de un árbol de decisión. Se procedió a calcular las métricas (accuracy/recall/precision/f1 score) así como también se generó la matriz de confusión.
  Si bien la performance fue muy alta (f1=0,81), notamos que el modelo presentaba overfitting para la selección de datos de entrenamiento, dado que su respuesta al realizar el commit en kaggle fue muy inferior f1= 0.583.
- El segundo modelo se realizó tomando el mismo set de datos depurados pero en este caso se buscó mediante la función de “parametersGrid”, el set de hiperparámetros que maximizan la eficiencia del árbol que se utilizará como predictor.

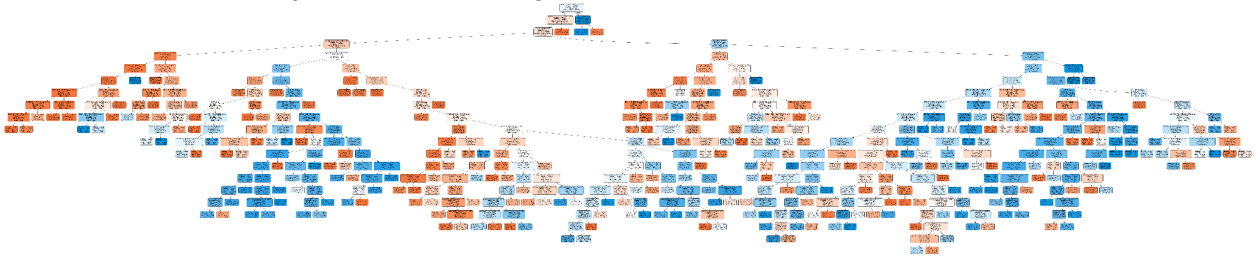
Se corrieron gran cantidad de configuraciones diferentes, utilizando k-fold Cross Validation con seteos entre k-Fold={1… 10} y permitiendo n combinaciones entre 1000 y 8000 dado que si superábamos ese límite de combinaciones nos produce un error de overflow.
Dentro de los hiperparametros habilitamos las combinaciones utilizando como criterio para medir la cantidad de información tanto “entropy” como “gini”.
La métrica para la búsqueda de hiperparametros fue f1 score. Siendo el mejor resultado en kaggle obtenido de 0.82906.

- El tercer modelo es idéntico al segundo pero se utilizó un dataset diferente. En este caso se se optó por utilizar el dataset “hotels_train” original, eliminando la variable “company” dado que al igual que antes esta no aporta información pero a su vez en lugar de realizar imputaciones también se eliminaron los registros/observaciones que

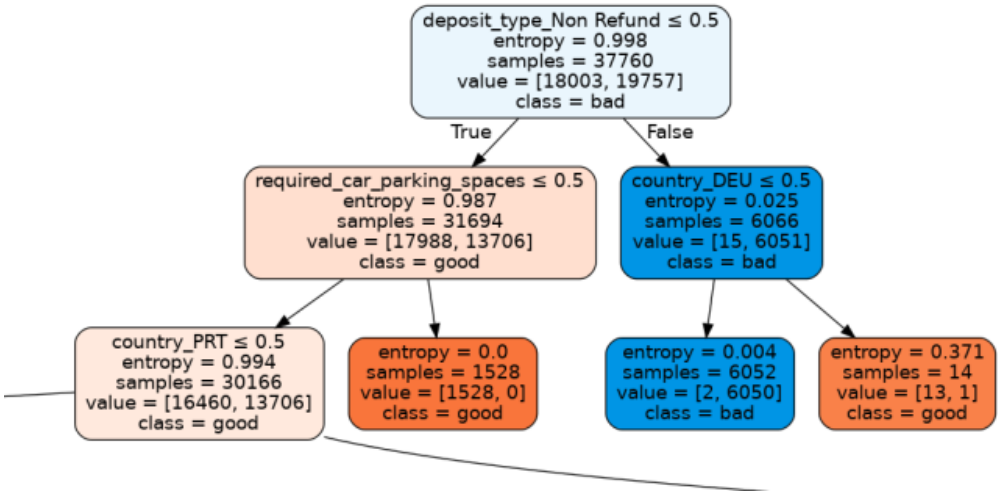
presentaban algún valor nulo mientras que se permitió la permanencia de los outliers.
Se realiza esto considerando que nuestras imputaciones podrían estar sesgando el modelo de predicción así como el dejar afuera los outlier evitar ganar un aprendizaje en la predicción de estos valores. Utilizando tambien f1 score, conseguimos el mejor resultado en nuestro modelos pasando a ser en kaggle ahora de 0,84482.

De esta forma iniciamos con una métrica de 0.583 y terminamos con 0,84482, logrando asi una mejora en la predicción de 45%.

El árbol de nuestro mejor modelo es el siguiente:



Se puede apreciar como notable que el mayor elemento de decisión es la variable deposite_type lo cual tiene correlato con lo visto en el análisis de datos preliminar dada la gran correlación que observamos entre esa variable y el target.



Cuadro de Resultados

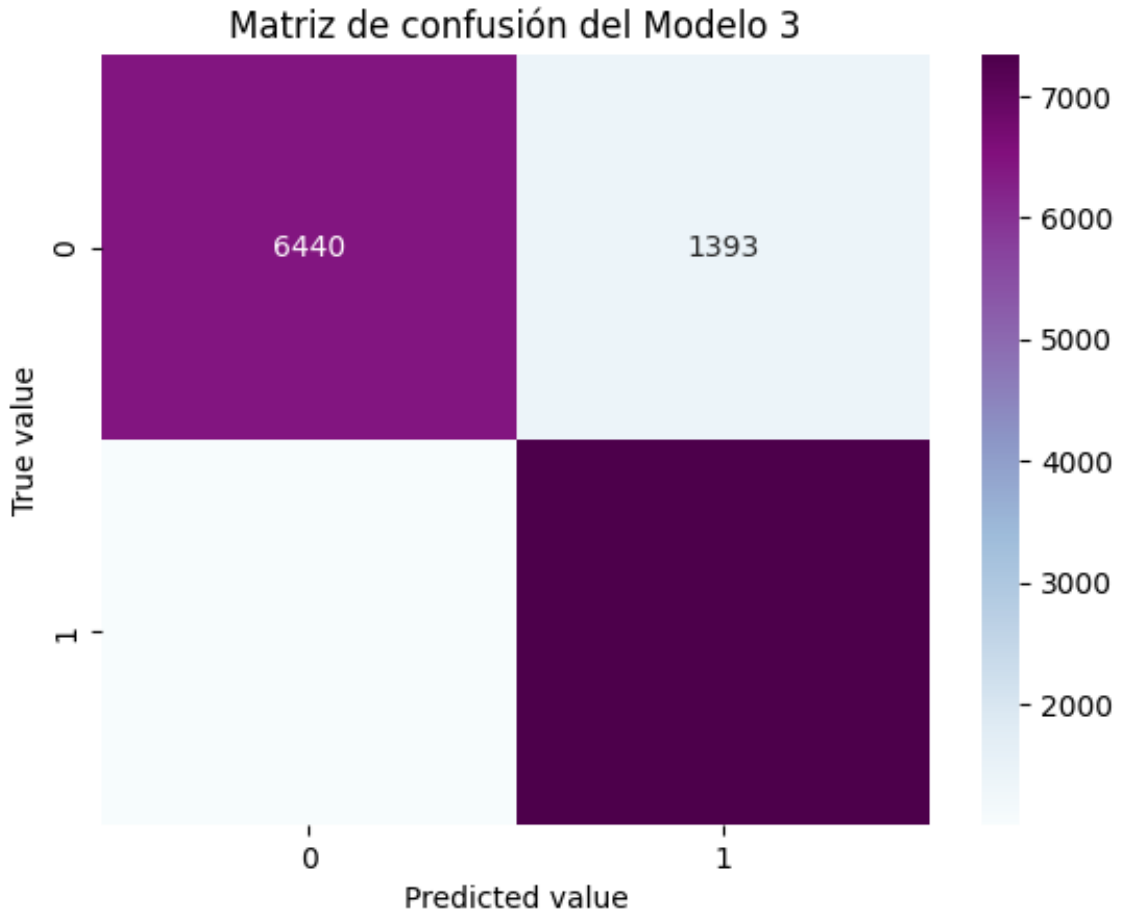
Realizar un cuadro de resultados comparando 3 modelos de los que entrenaron (entre ellos debe figurar el que seleccionaron como mejor predictor). Confeccionar el siguiente cuadro con esta información:

Medidas de rendimiento en el conjunto de TEST:

- F1
- Precision
- Recall
- Métrica XXX (Cualquier otra que considere relevante)
- Resultado obtenido en Kaggle.

Modelo	F1-Test	Presicion Test	Recall Test	Kaggle
modelo_1	0.8267	0.8114	0.8426	0.583
modelo_2	0.8382	0.8120	0.8662	0.82906
modelo_3	0.83826	0.81206	0.86620	0,84482

Matriz de Confusion



Tareas Realizadas

Indicar brevemente en qué tarea trabajo cada integrante del equipo, si trabajaron en las mismas tareas lo detallan en cada caso (como en el ejemplo el armado de reporte).

Integrante	Tarea
DIEM, Walter Gabriel	Desarrollo Modelo 1 y 2
MAIOLO, Alejandro	Desarrollo Modelo 1 y 3
RUIZ, Karen Belén	Desarrollo Modelo 2