Materia: 75.06/95.58 - Organización de Datos
Cuatrimestre: 1° cuatrimestre 2023
Grupo: 27

Padrón	Apellido y Nombre
94727	Jarmolinski, Arian
108405	Porro, Joaquín
97538	Bordón Villavicencio, Fernando Nahuel

Elegimos el modelo de skitlearn para predecir la variable “is_canceled”, para ello previamente pre procesamos los datos para que el modelo lo acepte transformando todas las variables categóricas con one hot encoding. Además, para las categorías con muchas opciones agrupamos en menos categorías para reducir las dimensiones del dataset final con el que entrenaremos el modelo, para reducir el tiempo de entrenamiento

Luego, buscamos los hiper parámetros del modelo, para lo cual utilizamos grid search cross validation para ir acotando la búsqueda y realizando cambios en la cantidad de folds.

La métrica utilizada es el impureza de Gini porque dio mejor resultados al maximizar la separación entre las clases en el árbol de decisión

La hiperparametros encontrados son:
métrica : gini, profundidad máxima = 19, muestra hoja mínima = 7, muestra mínima para split = 7 y cantidad de fold = 5

Para medir nuestro modelo utilizando 5 folds, es decir se particionó 80% train y test para calcular las siguientes métricas:

Accuracy: 0.9086136998691713
Recall: 0.9183284315309783
Precision: 0.9007449675067364
f1 score: 0.9094517171846493

Con el modelo entrenado se predijo el set de test y se subió a kaggle dando un f1score_test = 0.86482

Dado que los f1score utilizados para entrenar y el test dan parecidos se concluye que nuestro modelo ajusta bien a valores nuevos (no overfittea)

Observamos, que se puede mejorar al iterar el modelo mejorando el preprocesamiento.

Observamos, la importancia de realizar un buen preprocesamiento para ajustar mejor el modelo y que al entrenar un modelo con muchas dimensiones tiene un costo de tiempo mayor.