Universidad de Buenos Aires
Facultad De Ingeniería
Año 2023 - 1er Cuatrimestre

Organización de Datos (75.06 /95.58)
Trabajo Practico 1: Reservas de Hotel

CHECKPOINT 3: Ensambles
FECHA: 12/05/23

INTEGRANTES:

Gonzalez,Tomás	#108193
<togonzalez@fi.uba.ar>	
Moreno del Ruvo, Valentina	#107948
<vmdr18@gmail.com>	
Pol, Juan Manuel	#108448
<jpol@fi.uba.ar>	

Procesamiento de datos

Para cada uno de los clasificadores entrenados realizamos el mismo procesamiento, utilizando “Ordinal encoding”para las variables categóricas dado que los algoritmos no pueden trabajar con variables no numéricas.

KNN

Inicialmente buscamos encontrar el mejor dataset teniendo en cuenta las 3 posibilidades con las variables “Country” y “Agent”. Dado los resultados de entrenar el modelo con cada uno notamos que nuevamente el dataset solo con “Country” fue el que mejor predijo.

SVM

Para emplear SVM decidimos hacer “Ordinal Encoding” en vez de “Hot Encoding” dada la cantidad de columnas innecesarias que se generaban, donde se tenía que procesar una columna completa para unos pocos casos. Mantuvimos este mismo procesamiento para todos los modelos que quedaban. Optamos por los kernels lineales y radiales para los modelos, y la busqueda de los hiperparámetros fue manual, debido al tiempo de espera que conllevaba entrenar las instancias.

RF y XGBoost

En ambos casos utilizamos Cross-Validation con GridSearch. Sin embargo no pudimos obtener diferencias significativas en la optimización de los hiperparámetros. Lo que notamos fue una gran diferencia entre los score obtenidos en el dataset de validación y el de testeo de la competencia. Para XGBoost también utilizamos la métrica ”Area under ROC curve”, con la cual obtuvimos un valor de 0.95 (Ver Figura 1)

Ensambles Híbridos

Para los ensambles hibridos creamos nuevas instancias de los modelos intentando utilizar los mismos parametros que anteriormente. Sin embargo, en muchos casos tuvimos que simplificar los modelos, es decir, volverlos menos precisos para poder ahorrar en el tiempo de ejecución, ya que notamos que para el entrenamiento podía tardar hasta multiples horas.