

Trabajo Práctico 1 — Reservas de Hotel

Organizacion de Datos
Curso Rodriguez
Primer cuatrimestre de 2023

Alumno	Padrón	gitHub
Camila Gonzalez	105661	c-gonzalez-a
Eduardo Martín Bocanegra	106028	martinboca
Mateo Cabrera	108118	m-cabrerar

1. Introducción

El objetivo de esta segunda parte, será buscar el mejor árbol de decisiones para predecir el estado de una reserva. Para ello usamos el Decision Tree Classifier de SkLearn, y utilizamos un método de GridSearch para encontrar los mejores hiperparámetros para el modelo.

2. Preparación del set de datos

El árbol de decisiones de SkLearn solo puede trabajar con parámetros numéricos, así que lo primero que hicimos fue decidir que columnas usar y como asegurarnos que sean numéricas. Buscamos las variables categóricas del set de datos y decidimos que hacer con cada una de ellas:

- 'hotel': la dejamos fuera del modelo.
- 'meal': la convertimos en números enteros, en orden ascendente de acuerdo a la cantidad de comidas.
- 'arrival_date_month': números enteros correspondientes al mes.
- 'reserved_room_type': número enteros correspondientes a cada letra.
- 'assigned_room_type': idéa al punto anterior.
- 'deposit_type': números enteros de 0 a 2, ascendente en orden del "riesgo" de cancelar. 0 para reservas sin pago por adelantado, 1 para las reservas con pagos por adelantado pero reembolsable, 2 para pagos no reembolsables.
- 'agent': 1 para reservas con agente, 0 para reservas sin agente.
- 'market_segment', 'distribution_channel' y 'customer_type'

3. Búsqueda del mejor árbol

Ahora es hora de buscar el árbol de decisiones que mejor se ajuste a nuestro modelo y set de datos. Para optimizar los hiperparámetros a usar, vamos a utilizar k-fold Cross Validation. Tendremos en cuenta 5 folds. Con una búsqueda randomizada con cross validation, el algoritmo probará distintas combinaciones de los hiperparámetros dados y calculará su score. Luego obtendremos la combinación de hiperparámetros que resultó en el mejor score.

4. Ponemos a prueba nuestro árbol

Ahora que tenemos nuestra mejor versión del árbol de decisiones, lo vamos a probar con nuestro conjunto de pruebas y graficamos la matriz de confusión.

