

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

En este trabajo práctico se buscó analizar y tratar un problema real de ciencia de datos, trabajando y aplicando, en cada una de las etapas del proceso, los contenidos vistos en la materia. Utilizamos un conjunto de datos de reservas de hotel provisto por la cátedra. El objetivo principal del trabajo fue aplicar técnicas de análisis exploratorio, preprocesamiento de datos y entrenar modelos de clasificación para predecir si una reserva va a ser cancelada.

2. Resumen

Esta primera entrega se enfocó en el análisis de nuestro set de datos inicial. Para la exploración de los datos, extrajimos información como la cantidad de columnas, datos faltantes o filas que podrían haber estado mal cargadas e hicimos graficos para mejor visualización de la información obtenida.

3. Variables

En el dataset con el que trabajamos se tienen en cuenta 33 variables, una de ellas nuestra variable target. Podemos segregar estas variable en 2 principales categorías, variables categóricas y variables numéricas. Encontramos así 16 variables de cada grupo:

- **Variables categóricas:** Hotel, Mes de llegada, Comida, País, Segmento del mercado, Canal de distribución, Cliente fijo, Tipo de habitación, Forma de pago, Tipo de cliente, agente, compañía, estatus de reserva(variable objetivo).
- **Variables Numéricas:** Tiempo de anticipo, Año de llegada, Día de llegada, Semana de llegada, Número de noches de estadía durante semana, Número de noches de estadía durante fin de semana, Número de adultos, bebés y niños en la reserva, cancelaciones previas, reservas previas no canceladas, número de cambios en la reserva, días en lista de espera, etc.

4. Correlación entre variables

Para estudiar la correlación entre variables, utilizamos el coeficiente de Pearson, calculado para todas las variables numéricas. Esto lo graficamos en un Heatmap para poder visualizar los resultados.

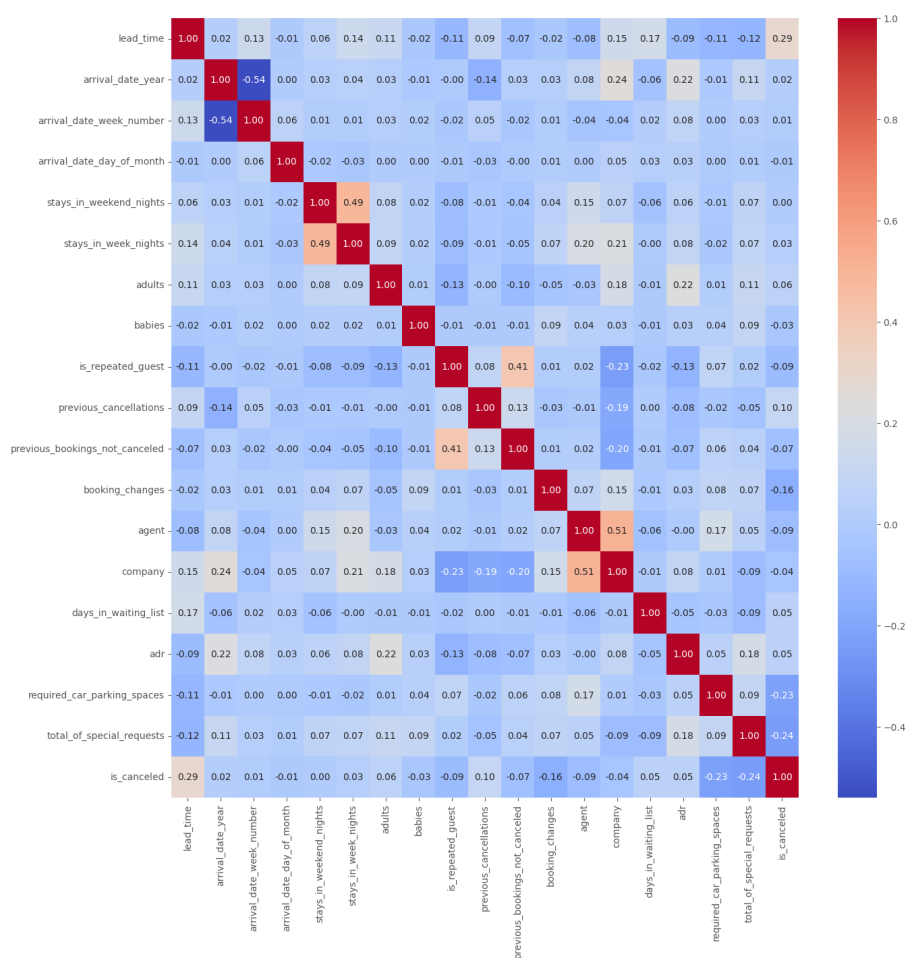
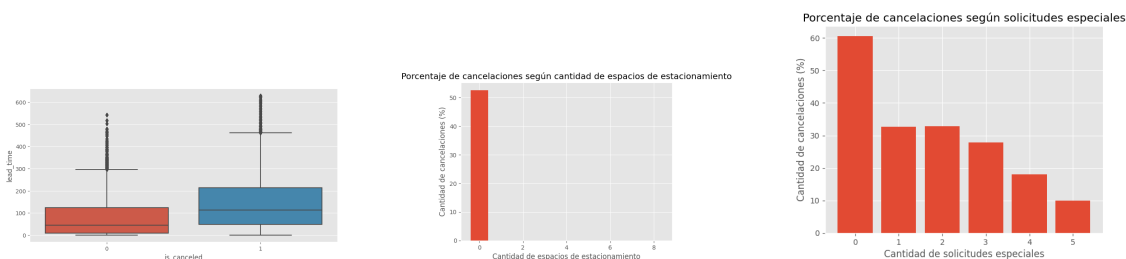


Figura 1: Gráfico de Correlaciones

Acá además podemos observar la relación entre las variables y nuestra variable objetivo. Vemos que las más correlacionadas son tiempo de anticipación de la reserva, cantidad de pedidos especiales y cantidad de espacios requeridos en el estacionamiento. Para estudiar estas variables un poco más en profundidad, graficamos como se relacionan con el target.



5. Datos Faltantes

Hicimos gráficos para analizar la cantidad de datos faltantes en cada columna. La mayor cantidad de datos faltantes se encontró en las columnas 'company', con 95 % de datos faltantes, y 'agent' con 13 % de datos faltantes

Nos dimos cuenta que esto no era información que faltaba, sino que significaba que la reserva no se había hecho por medio de un `.agent`". Lo mismo para la columna `'company'`, la falta de datos, significa que la reserva no está asociada a ninguna compañía.