Trabajo Práctico 1: Reservas de Hotel

Checkpoint 1: Análisis Exploratorio y Preprocesamiento de Datos

75.06 / 95.58 Organización de Datos - FIUBA
Ing. Rodríguez - 1°C 2023

Grupo 31 - "Datazo":

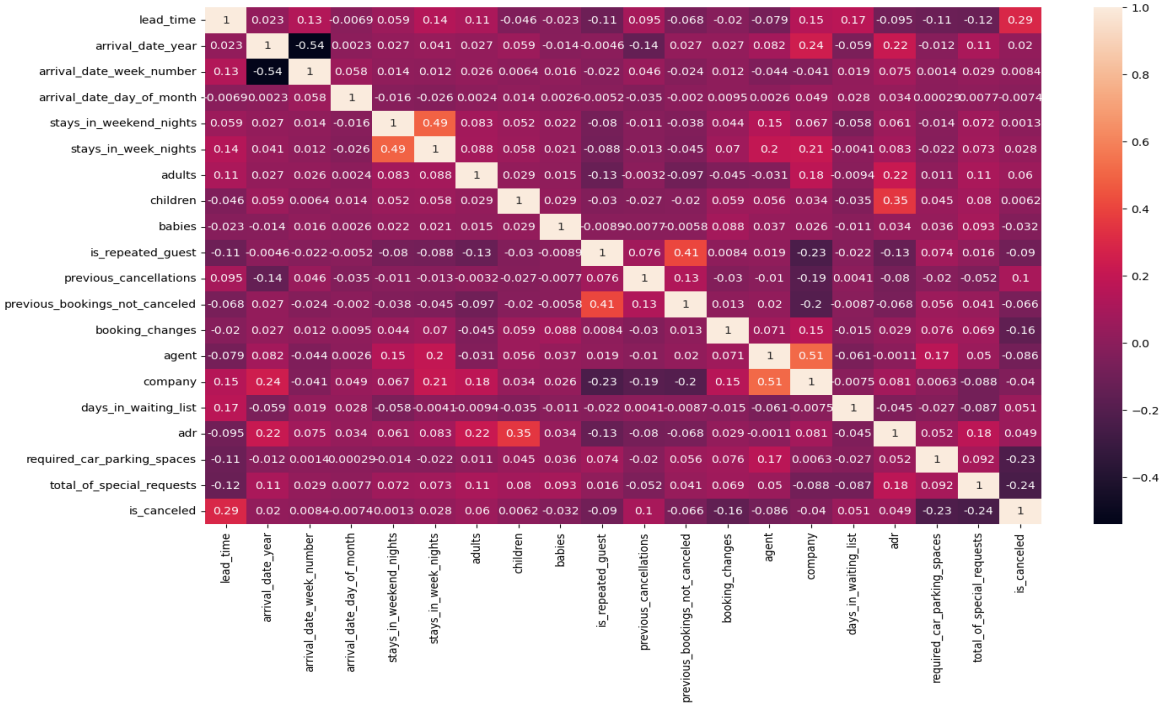
- 106203 - Kisinovsky, Diego Andrés
- 102685 - Mena Giraldo, Michael Gustavo
- 104256 - Brocca, Pablo Martín

Introducción

Esta primera parte del Trabajo Práctico consiste en el análisis exploratorio de las reservas de hotel contenidas en el dataset "hotels_train.csv", para poder estudiar las distribuciones de las variables y las relaciones entre ellas a modo de aproximación inicial. Para ello utilizamos herramientas de visualización y aplicamos lo aprendido sobre Ingeniería de Características para realizar un preprocesamiento de los datos que incluye limpieza e imputación de datos faltantes, selección de variables irrelevantes y generación de variables nuevas. Nuestro target es la variable *is_canceled*, que indica si la reserva fue cancelada o no.

Exploración inicial

Comenzamos por graficar la matriz de correlación para las variables numéricas del dataset:



Con ayuda de los colores del heatmap podemos observar a simple vista que la variable que produce el mayor valor absoluto de coeficiente de correlación con *is_canceled* es *lead_time*, con correlación positiva. El segundo y tercer valor absoluto mayor lo dan las variables *total_of_special_requests* y *required_car_parking_spaces* respectivamente, con correlación negativa. Se puede inferir prematuramente que hay una tendencia a que la reserva sea cancelada con menor probabilidad mientras menos tiempo transcurra entre esta y la llegada de los huéspedes al hotel, y mientras más peticiones especiales y cocheras se hayan solicitado.

Preprocesamiento

Determinamos que las siguientes variables no son relevantes para nuestro análisis:

- *arrival_date_week_number:* contando con la fecha de llegada en forma de mes y día, el número de semana se vuelve redundante y no proporciona información que no se pueda obtener de las otras variables de forma más clara.
- *company*: es la variable con mayor porcentaje de valores nulos, con casi 95%.
- *id*: se trata de un código único que sólo funciona como clave y no aporta información.
- *reservation_status*: es muy similar al target, puede generar confusiones o data leakage
- *reservation_status_date*: no tiene relevancia sin la variable *reservation_status*.

De estas columnas, eliminamos sólo *company*, *reservation_status* y *reservation_status_date*, las otras las ignoramos de momento.

A su vez, generamos las siguientes variables que unifican otras para simplificar su estudio:

- *stays_in_nights*: calculada como la suma entre *stays_in_week_nights* y *stays_in_weekend_nights*.
- *total_of_guests*: calculada como la suma entre *adults*, *children* y *babies*.

Por otro lado, realizamos imputación para datos faltantes en las siguientes columnas:

- *country*: se reemplazan los valores nulos por la moda, en este caso 'PRT'
- *agent*: si bien el paper indica que los valores nulos significan que no se utilizó un agente, se reemplazan por 0 para homogeneizar los tipos de dato de la variable, manteniendo el significado.
- *children*: se reemplazan los valores nulos por la moda, en este caso 0.

Visualizaciones destacadas