Trabajo Práctico 1 - Reservas de Hotel

Integrantes:
Ignacio Latorre - Padrón: 101305
Nicolas Ronchese - Padrón: 108169
Gastón Avila - Padrón: 104482

Introducción

El trabajo práctico consiste en realizar un análisis exploratorio de un conjunto de datos de reservas de hotel. Este análisis consiste en una exploración inicial describiendo cada variable, determinando su tipo y distribución. Luego graficar relaciones entre estas variables para contraer conclusiones e identificar valores atípicos y datos faltantes.
El objetivo de este análisis es obtener una base sólida de conocimiento para entrenar un modelo que pueda predecir el valor de la variable target **is_canceled** que indica si una reserva se encuentra cancelada o no.

Desarrollo

Exploración Inicial:

Analizamos cada variable, determinamos su tipo según la teoría (esto se puede ver en el documento **análisis_variables.txt**) y las graficamos para conocer cómo se distribuyen, es decir, cuales son los valores más comunes. Luego, generamos una matriz con las correlaciones y la graficamos en un heatmap para identificar relaciones fuertes, ya sean positivas o negativas entre las variables, y usamos esos datos para determinar cuales eran variables irrelevantes. Entre ellas están:

- **meal**: consideramos que meal no es una variable que valga la pena tener en cuenta para sacar conclusiones sobre la cancelación de la reserva
- **arrival_date_year**, **arrival_date_week_number**, **arrival_date_day_of_month:** son variables de tiempo que preferimos dejar de lado para para quedarnos solo con **arrival_date_month** como medida temporal para analizar.
- **id:** es un identificador de la reserva, lo cual no tiene ningún valor de análisis.
- **reservation_status_date:** preferimos hacer foco en el estado en sí, y no cuando cambio este último.

aclaración: elegimos utilizar el documento analisis_variables.txt ya que ocupaban mucho espacio en la notebook

Visualización de Datos:

Para analizar la relación entre las variables y el target optamos, principalmente, por graficar cada variable (acorde a su tipo) separando con diversos colores los casos de reserva canceladas o no canceladas. A partir de esto se pudieron extraer las siguientes conclusiones:

- Si el mismo cliente canceló reservas en ocasiones, es muy probable que vuelva a cancelar en las siguientes reservas.
- Los clientes que reservaron anteriormente y no cancelaron en esas ocasiones, suelen cumplir con la reserva y no cancelan
- Sorprendentemente, a la gente que se le asigna una habitación distinta a la que pidieron en la reserva no suele cancelarla.
- Los clientes que hicieron cambios en la reserva son menos propensos a cancelar
- Las reservas donde se abonó la totalidad de la estadía previamente y sin devolución fueron, en su amplia mayoría, canceladas.
- Aquellas reservas que se hicieron por intermedio de una compañía no suelen ser canceladas.
- Mientras más días transcurran en lista de espera, es más probable que las reservas sean canceladas.
- Los adultos que concurren con niños, gastan más durante la estadía.
- Las reservas hechas por grupos o que están asociadas a otras reservas tienen menos probabilidad de ser canceladas.
- Las reservas donde se pidieron lugares de estacionamiento no fueron canceladas.
- Aquellas reservas donde hubo al menos un pedido especial, son menos propensas a ser canceladas.
-

Correlación de Datos con el target:

Variables cuantitativas:

- Entre las correlaciones de **is_canceled** con las demás variables se destaca **lead_time**. Es una correlación positiva, lo cual indicaría que las reservas que se hacen con mayor anticipación, tienden a cancelarse. Si bien es solo de **0,29**, es la correlación más fuerte que tiene la variable **is_canceled**.
- La segunda más fuerte, se da con la variable **total_of_special-requests**. Es una correlación negativa de **-0,24** lo que indicaría que a mayor cantidad de requerimientos, menos cancelaciones. Se podría pensar que el hotel cumple con los requerimientos especiales, por lo que el cliente se siente satisfecho.

- Por último, bastante similar a la anterior es la correlación de **-0,23** con **RequiredCardParkingSpaces**. También se podría pensar que cuando el hotel cumple los requerimientos, el cliente no cancela.

Variables cualitativas:

- La mayoría de los portugueses cancela la reserva.
- La categoría "Groups" de **market_segment** se destaca debido a que la mayoría cancela la reserva.
- En la categoría “Transient”, de la variable **customer_type**, son más los que cancelan que los que no.
- Las reservas de habitaciones tipo A suelen ser canceladas.
- Se puede ver que las reservas donde se abono el total del costo en anticipación fueron canceladas en su gran mayoría. Mientras que aquellas donde hubo una seña parcial o no la hubo predominan las reservas no canceladas.

Datos Faltantes:

A la hora de analizar los datos faltantes el primer paso fue ver cuáles de las variables restantes contenían columnas con datos faltantes. En el archivo logramos identificar solo 4 columnas que contenían casilleros vacíos ‘children’, ‘country’, ‘‘agent’ y ‘company’.

En la columna de ‘children’ únicamente 4 filas se encontraban sin datos, un 0,0065% del total, por los que la decisión fue en lugar de eliminar las filas que el contenido de esa casillas sea el mismo que el de la mediana que en este caso era 0.

Similarmente, en la columna ‘country’ la cantidad de filas sin datos no representa una gran cantidad en comparación con el total, en este caso 221 o un 0.357 %, por lo cual decidimos tomar un camino similar al que utilizamos con ‘children’. Sin embargo, ‘country’ es una variable cualitativa por lo que en vez de utilizar la mediana utilizamos la variable de mayor probabilidad que era ‘PTR’.

Mientras que las primeras 2 contaban con una baja cantidad de datos faltantes (4 para ‘children’ y 221 para ‘country’) lo contrario pasaba para ‘agent’ y ‘company’ los cuales tenían casi miles de datos faltantes en ambos, pero significa que el cliente/s no había contratado dicho servicio por lo que no hace falta modificar

Valores Atípicos:

Para el análisis de esta sección comenzamos creando un box plot para cada una de las variables cualitativas con la intención de ver si dentro de los outliers había valores atípicos. Una vez solucionado esto decidimos pasar a investigar los outliers multivariados por lo que agrupamos las columnas que a nuestro entender tiene sentido agrupar, por ejemplo ‘adults’-’children’-’babies’.

Algunas valores atípicos son :

- que no haya adultos,
- que haya niños o bebes y no adultos
- que el adr sea 0 o negativo,
- que haya más autos que adultos,
- entre otros,

aclaración:la parte final de esta sección no la logramos realizar a tiempo por lo que solo se tendrá en cuenta la detección de los valores atípicos
aclaración 2: consideramos a children como menores de 13