Organización de Datos (75.06)
Trabajo Práctico 1, checkpoint 1: Reservas de Hotel
Cuatrimestre, año y grupo: 2ºC 2023
Grupo: 11 - Los Pandas
Corrector: Mateo
Fecha de entrega: 22 de septiembre de 2023
Datos de los integrantes:

Labollita, Francisco	103456	flabollita@fi.uba.ar
Mundani Vegega, Ezequiel	102312	emundani@fi.uba.ar
Otegui, Matías Iñaki	97263	motegui@fi.uba.ar

Análisis exploratorio

El dataset provisto tiene 31 columnas y 61.913 filas. El nombre del *hotel* es una variable cualitativa de texto nominal, se la considera relevante porque hay una gran diferencia en el porcentaje de cancelaciones entre los dos hoteles dados. *lead_time* es una variable cuantitativa discreta que indica cuántos días pasan desde que se hizo la reserva hasta la fecha de esta, se observó un claro incremento en el porcentaje de relaciones a medida que pasa el tiempo. *country* es otra variable de cualitativa de texto nominal en la que se ven porcentajes de cancelaciones mayores para determinados países. *required_car_parking_spaces* es una variable cuantitativa discreta que se la considera importante porque todas las personas que pidieron estacionamiento no cancelaron, siendo este el 5 % de las personas. *total_special_requests* es cuantitativa discreta y se la considera importante porque hay una notable diferencia entre el porcentaje de gente que canceló dependiendo si tuvo pedidos especiales o no.

Se supuso de antemano que el *adr* de la reserva no afectaría a su cancelación.

Procesamiento de datos

Columnas creadas

Se creó la columna *room_type_match* que dice si la habitación asignada es la que fue pedida.

Columnas eliminadas

Una vez hecho el análisis de los diferentes atributos, se decidió eliminar todas las columnas consideradas irrelevantes. Estas son *arrival_date_year*, *arrival_date_day_of_month*, *stays_in_weeknd_nights*, *stays_in_week_days*, *children*, *babies* , *reserved_room_type*, *deposit_type*, *company*, *adr* e *id*.

Se consideraron irrelevantes a las columnas cuyos porcentajes de cancelación no parecían tener una clara relación con el atributo analizado.

adr se eliminó porque el código de la reserva no se supuse que influiría en su cancelación.

Correlaciones detectadas

Se detectó una correlación débil entre *arrival_date_month* y *arrival_date_week_number*. Su coeficiente de Correlación de Pearson es -0,54. Tiene sentido que estas variables estén correlacionadas, siendo que una se refiere al número de semana en el año de la reserva y la otra al mes de esta.

Columnas recodificadas

Los atributos *required_car_parking_spaces*, *days_in_waiting_list*, *babies*, *previous_cancelations*, *special_requests* *previous_bookings_not_canceled*, *booking_changes*, serán repensados como variables booleanas, siendo verdaderas si hay por lo menos uno o más de lo que indique el nombre del atributo y siendo falsa si no.

Valores atípicos

Para el atributo *adults* se encontraron valores atípicos de naturaleza univariada, siendo que hubo varios valores que ocurrían una única vez y estaban bastante espaciados de sus valores “vecinos”, como por ejemplo 55, 40, 27, 26, 20 y 10. Además hubo valores atípicos de naturaleza multivariada, que resultaron de calcular el porcentaje de cancelaciones en base a la cantidad de adultos, y se vio que para valores mayores a 4, por no haber poblaciones significativas daban porcentajes extremos. Por lo tanto se optó por eliminar las reservas con más de 4 adultos.

No tuvo sentido realizar el boxplot de *adults* dado a la mayoría de reservas tiene 2 adultos, por lo que el boxplot termina siendo una línea nada más.

Valores faltantes

Los atributos con porcentajes de valores faltantes son *company* (94,9 %), *agent* (12,7 %), *country* (0,38 %), *children* (0,006 %), *distribution_channel* (0,006 %) y *market_segment* (0,003 %).

Como *children* y *company* no son variables relevantes no tenía sentido analizar qué hacer con estos registros cuando sus columnas ya iban a ser eliminadas. Para el caso de *country*, *market_segment* y *distribution_channel* se opta por eliminar las entradas con estos datos faltantes dado a que representaban una cantidad insignificativa sobre el total del dataset. Para *agent* se dejarán las filas que no tienen este dato.

Visualizaciones

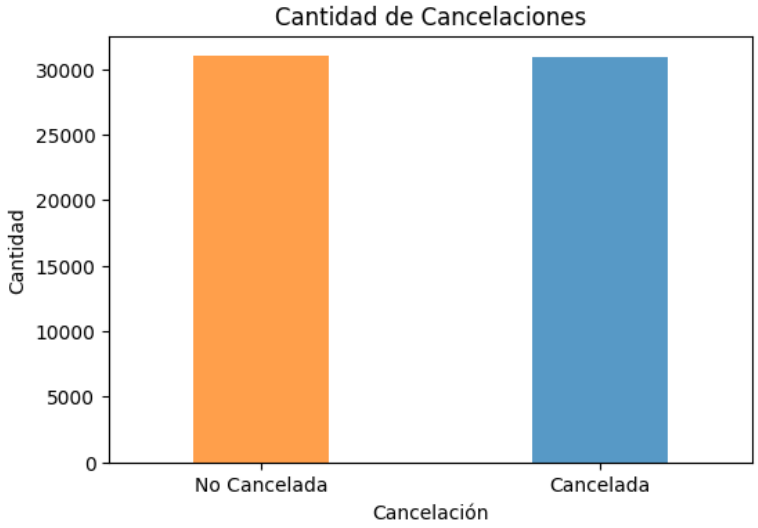

Figura 1: Cantidad de reservas canceladas y no canceladas.

Este gráfico es interesante porque muestra que el conjunto de datos para entrenar el modelo es bastante homogéneo en cuanto a si la reserva fue cancelada o no, y la posibilidad parece del 50 % de que sea un caso o el otro. Se va a querer que nuestro modela pueda predecir mucho mejor el tipo de reserva mejor que una moneda arrojada al aire.

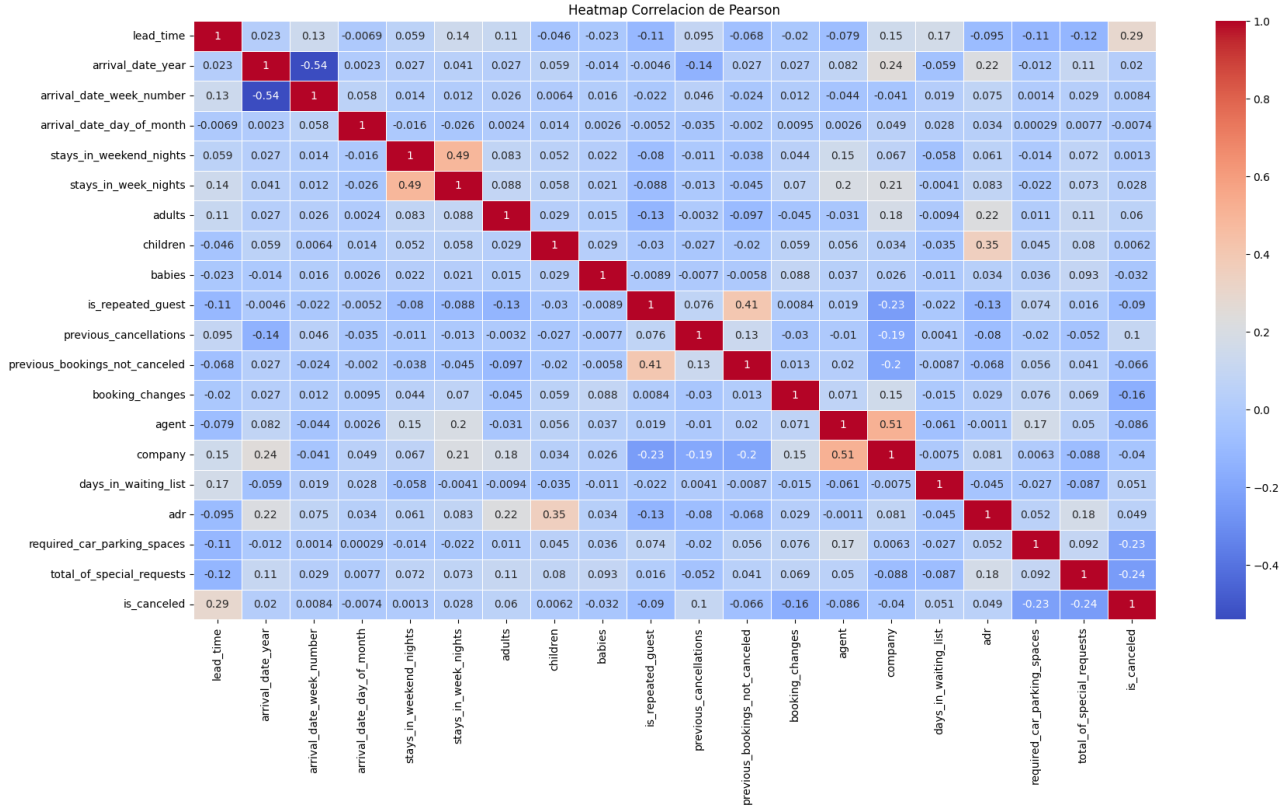

Figura 2: Heatmap de Correlación de Pearson.

El heatmap de la correlación de Pearson indica que no hay ninguna variable fuertemente relacionada de manera lineal con el target y que tampoco las variables se relacionan linealmente mucho entre sí. Excepto por los pares *arrival_date_week_number* y *arrival_date_month*, *stays_in_weekend_nights* y *stays_in_week_nigths*, *is_repeated_guest* y *previous_cancellations* que tienen una correlación lineal considerable.

Tareas Realizadas

Labollita, Francisco:

- Analizar y clasificar las variables
- Realización de gráficos de las variables
- Analizar la relación de las variables con target
- Analizar las correlaciones entre las variables
- Identificación de valores atípicos

Mundani Vegega, Ezequiel:

- Analizar las variables
- Realización de gráficos de las variables
- Analizar la relación de las variables con target
- Identificación de datos faltantes
- Identificación de valores atípicos
- Revisión final
- Informe

Otegui, Matías Iñaki:

- Analizar y clasificar las variables
- Realización de gráficos de las variables
- Analizar la relación de las variables con target
- Identificación de datos faltantes
- Identificación de valores atípicos
- Revisión final