75.06 / 95.58

Organización de datos

Reporte TP N°1: Reservas de Hotel

Análisis Exploratorio y Preprocesamiento de Datos: Checkpoint N°1

Alen Davies Leccese: 107084

Luca Mauricio Lazcano: 107044

Abril 2023

1 Exploración inicial de los datos

Para familiarizarse con el dataset se utilizaron diversas funciones para ver su estructura, las columnas que posee, la cantidad de datos, tipos de datos de las columnas y determinamos el target (`is_canceled`).

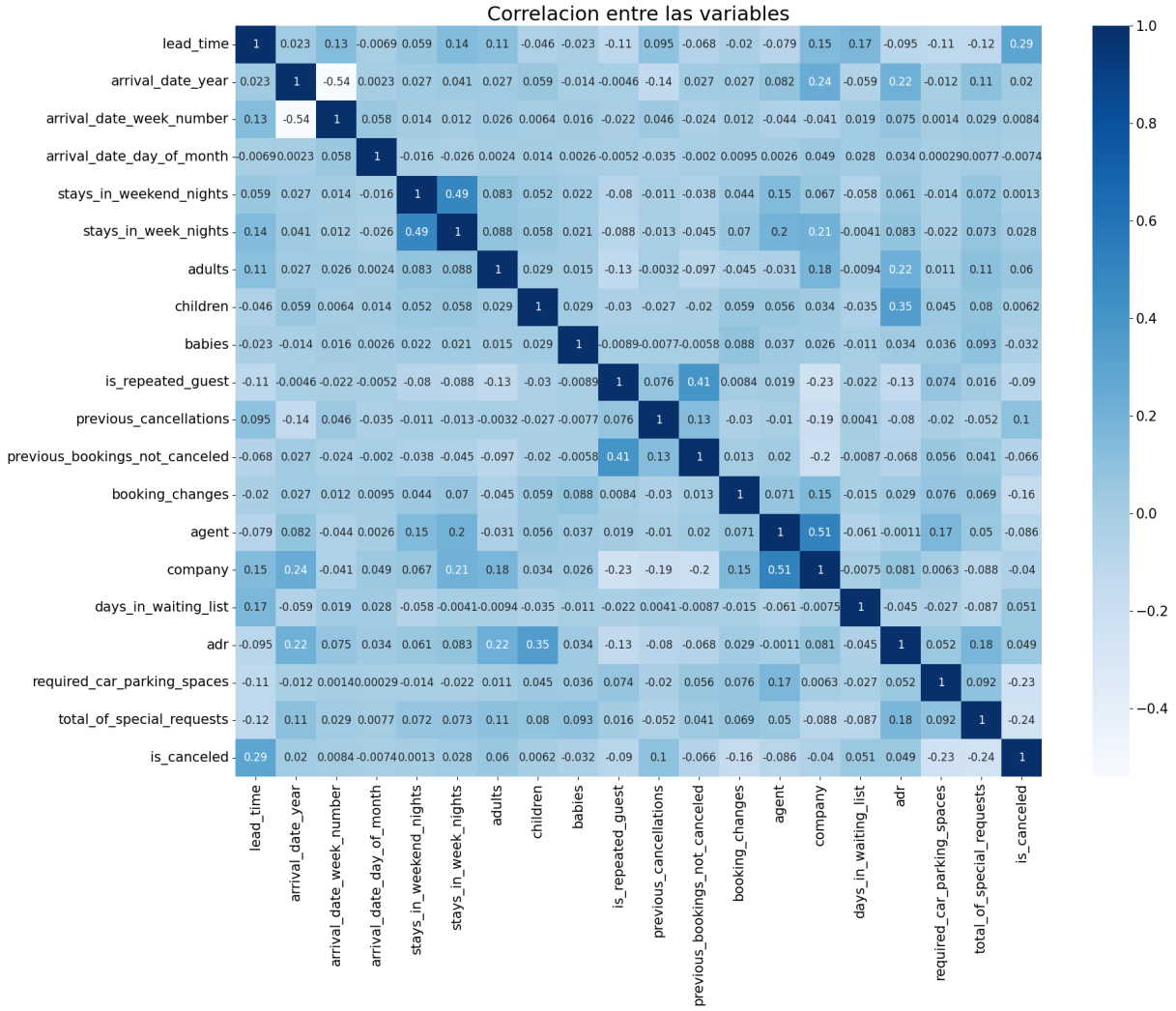
Se discriminaron las variables en cualitativas y cuantitativas, y se analizaron sus valores posibles, rangos, y resumen de estadísticas, según corresponda. Se analizaron gráficamente la distribución de todas las variables.

Para encontrar correlaciones entre variables, se realizó un heatmap estilo una matriz, con las correlaciones entre todos los pares de variables. Las correlaciones encontradas están debidamente explicadas junto a sus gráficos.

Luego se analizó la relación de cada variable con el target. Se detallan los hallazgos y se tantean hipótesis que puedan explicar la relación ciertas variables con el target.

Las principales correlaciones que encontramos son para:

- `lead_time`: Tiene sentido que a mayor tiempo hasta la fecha reservada, los clientes tienen más tiempo para cambiar de planes, encontrar otra opción o sufrir un imprevisto.

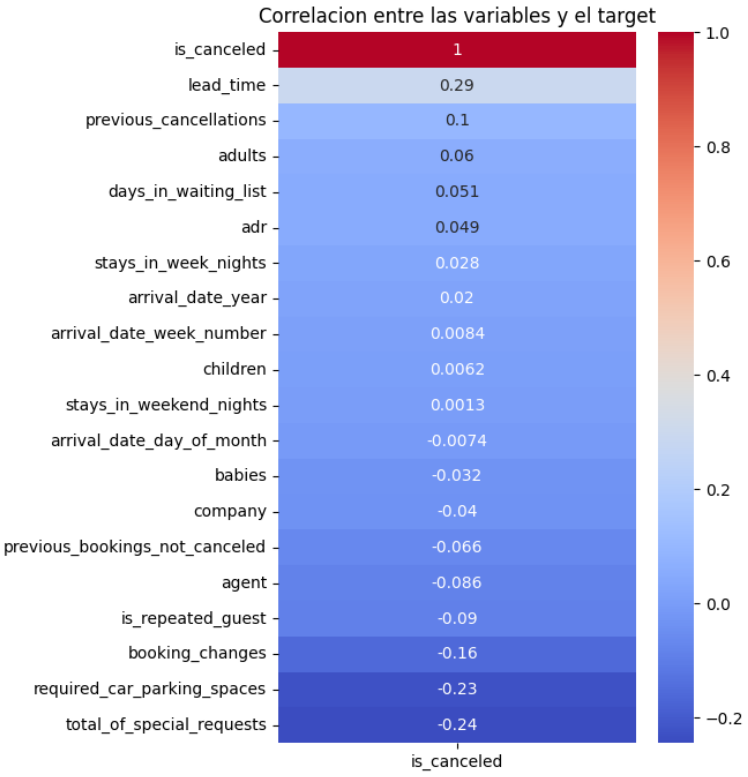


- `previous_cancelations`: Clientes con un historial previo de cancelaciones son más propensos a volver a cancelar.

Y entre las correlaciones negativas:

- `total_of_special_requests`: Esto podría pensarse como que el hotel tiene en cuenta las necesidades de los clientes y esto podría hacer que el cliente sea menos propenso a cancelar su reserva.

- `required_car_parking_spaces`: Siendo un subset de pedidos especiales, tiene sentido que la correlación sea parecida.



2 Datos faltantes

Es crucial identificar los datos faltantes y tratarlos de alguna manera. Los modelos no pueden ser entrenados con datos faltantes.

Analizamos cuáles son los datos faltantes, su cantidad y porcentaje. Los resultados son:

Columna	Cant. nulos	% nulos
`company`	58761	94.909
`agent`	7890	12.744
`country`	221	0.357
`children`	4	0.006

Table 1: Cantidad y porcentaje de nulos en el dataset.

Se eliminan los 4 registros con con `null` en `children`. Se asume que es un error.

El paper nos da información respecto a los `country` faltantes. La nacionalidad a veces se registra en el check-in, y no durante la reserva. Entonces si el cliente cancela, no quedará registrada su nacionalidad. Se imputan con el valor "Desconocido".

Para `agent` y `company` se desarrolla una explicación más exhaustiva en la sección correspondiente del notebook, pero básicamente se interpretan los valores `null` como un "no aplica".

Por último, tenemos el caso de los registros con 0 adultos. Esto es llamativo, ya que no imaginamos el escenario de niños o bebés reservando y viajando por hoteles sin el acompañamiento de adultos. La cantidad de reservas en esta condición es ínfima, así que simplementa las eliminamos.

La imputación y eliminación de los registros en esta sección tiene un impacto ínfimo en la distribución de los datos. Únicamente toman gran relevancia las categorías de "No aplica" para agentes y compañías, pero no afecta la distribución previa, simplemento los hace visibles.

3 Data cleaning

Aquí corregimos el tipo de dato inconsistentes de la columna children (decimal), a entero. Consideramos poco probable que se cuenten fracciones de niños...

También realizamos algo de "ingeniería de features", que consiste en seleccionar y transformar variables relevantes en un conjunto de datos. Por ejemplo, se agrupa la información relativa a las fechas en una sola columna de tipo `datetime`.

El dataset resultante fue guardado como `'hotels_cleaned.csv'`.

4 Outliers (valores atípicos)

4.1 Univariados

En esta sección exploramos los outliers, primero de forma univariada, y luego de forma multivariada.

Para encontrar outliers de forma univariada se utilizaron las técnicas del box-plot y z-score.

En el análisis del `adr` se eliminan los outliers superiores más severos, y también los registros con tarifas negativas o cero, ya que consideramos que estas no tienen sentido.

En el caso de `children` se detectan como outliers varios valores que son totalmente plausibles. Decidimos mantenerlos. El único registro eliminado es aquel que cuenta con 10 niños. Ídem con `babies`, donde se elimina el registro con 9 bebés.

Para los `adults`, consideramos outliers aquellas reservas con más de 4, las eliminamos.

Para `previous_cancellations` ponemos el límite en 6. Queda descartado un porcentaje ínfimo que tiene una cantidad absurda de cancelaciones previas.

En el caso de `days_in_waiting_list` se capturan bastantes outliers por el hecho de que la enorme mayoría de las reservas no pasan por un tiempo de espera. Se pone el corte en 200 días.

Para `lead_time` se detectan como outliers severos aquellos de más de 615 días. Se eliminan.

El dataset resultante queda con 59816 registros.

4.2 Multivariados

Para detectar los outliers en más dimensiones, que no se pueden detectar de forma sencilla analizando la distribución de forma multivariada, se utilizaron las técnicas de "isolation forest" y "local outlier factor". Primero, se cuantificaron los valores que eran de tipo objeto, convirtiendo por ejemplo los meses del año, los tipos de comida o el tipo de habitación a índices. De otra forma, no podrían ser analizados por los métodos de análisis multivariado.

Luego se aplicó el análisis a cada uno, detallado en el notebook.

5 Conclusiones

Esto da por concluida esta etapa del trabajo. Luego de un extenso análisis exploratorio, limpieza del dataset, detección y procesamiento de outliers, se obtuvo un dataset limpio.

A nuestro parecer, las decisiones tomadas a lo largo de este trabajo nos parecen las correctas para poder entrenar un modelo de ML que pueda predecir con bastante precisión la cancelación de una reserva de hotel.