FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1 — Checkpoint 1

[75.06/95.58] Organización de Datos
Segundo cuatrimestre de 2023

Grupo 15

Ayala, Tomás Gabriel - tayala@fi.uba.ar - 105336
Giacobbe, Juan Ignacio - jgiacobbe@fi.uba.ar - 109866
Olaran, Sebastian - solaran@fi.uba.ar - 109410

Docente corrector:
Pereira, Francisco

Índice

1. Análisis Exploratorio

Este primer checkpoint del trabajo práctico consistió en la familiarización con un dataset que describía reservas de hotel. En total, el dataset inicial contenía 61913 registros de reservas, con 31 columnas que nos daban información sobre las mismas(por ejemplo, una columna nos indicaba la cantidad de adultos que habían por reserva). Nuestro objetivo fue realizar un análisis exploratorio y preprocesar los datos, analizando cada variable por separado, observando correlaciones entre las mismas, etc.

Uno de los principales objetivos fue encontrar la relación entre cada variable con una variable target, la cual se llama is canceled(como su nombre indica, esta variable nos dice si la reserva se canceló o no). En las próximas entregas lo que se hará será crear modelos predictivos que determinen si una reserva se tiene que cancelar o no.

A la hora de analizar cada feature, se dió un enfoque distinto dependiendo el tipo de variable. Para las variables cuantitativas, en su mayoría variables numéricas, lo primero que se buscó hacer fue ver la distribución de sus estadísticas utilizando sus medidas de resumen, como por ejemplo su media. Usualmente, estas estadísticas son acompañadas por un box plot para poder observar gráficamente la distribución de dichos valores.

En cambio, para las variables del tipo cualitativas, por lo general del tipo object, hicimos un recuento de la cantidad total de valores y utilizamos gráficos de barras e histogramas para visualizar los posibles valores que tomaban y con qué frecuencia lo hacían.

2. Preprocesamiento de Datos

2.1. Valores faltantes

Durante el análisis general del dataset encontramos que hay exactamente 4 columnas con datos faltantes:

- Company: Es la que más valores faltantes tiene(58761 registros faltantes).
- Agent: Le faltaron 7890 registros.
- Country: Le faltaron 221 registros.
- Children: Le faltaron 4 registros.

2.2. Columnas recodificadas

Para las 4 variable mencionadas en la sección anterior, decidimos NO eliminar esas columnas para nuestro análisis, y lo que hicimos fue recodificar(o rellenar) aquellos valores faltantes. Para las columnas agent y company, al darnos IDs de agentes y compañías de bookings respectivamente, decidimos rellenar aquellos valores faltantes con un ID "falso.º "dummy", que toma el valor de -1. Para country hicimos algo parecido, pero la diferencia es que les pusimos a los datos faltantes el valor de ÜND"(undefined). Y por último, para los datos faltantes de children, tomamos como decisión cargarles el valor de 0(lo consideramos razonable para nuestro análisis, ya que pueden haber reservas sin niños).

2.3. Correlaciones destacadas

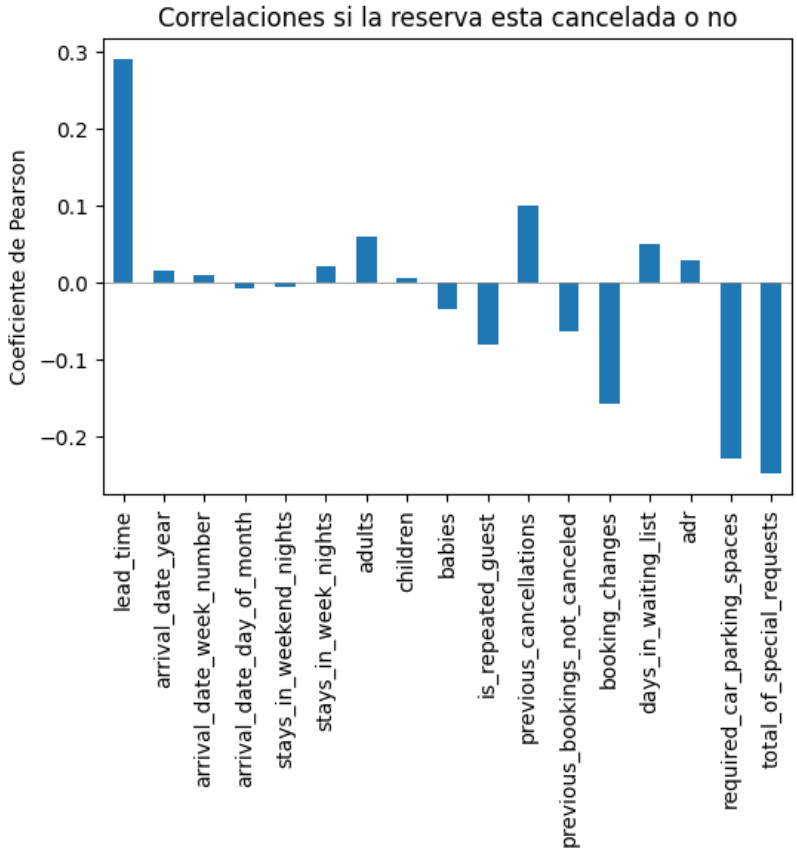

Figura 1: Correlaciones de Pearson entre algunas variables y la variable target.

En este gráfico podemos observar las correlaciones de pearson todas las variables con respecto con la variable is canceled . Para este caso destacamos la variable lead time.

La variable lead time hace referencia a la cantidad de dias de espera que hay entre la fecha en que se hizo la reserva y para cuando se hace el check-in de la reserva al hotel. Cuando miramos mas detalladamente y dividimos la distribución de los días entre el lead time de las reservas canceladas y las que no encontramos lo siguiente

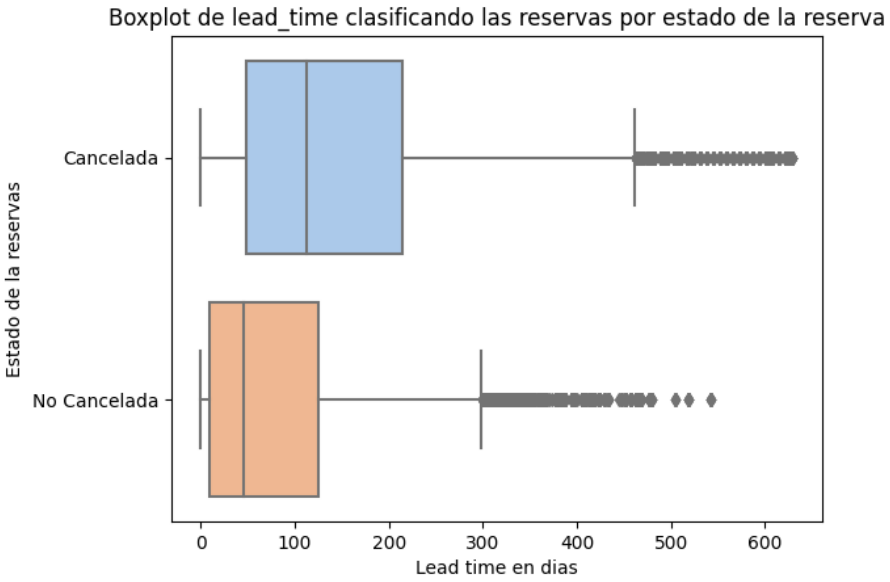

Figura 2: Boxplot de la variable lead time haciendo distinción entre reservas canceladas y no canceladas.

Algo que podemos apreciar a simple vista es que las reservas hechas mas cercanas a la fecha de ingreso, tienden a cancelarse menos. Y por el contrario, a medida que el lead time es mayor, aparecen más reservas canceladas. Esta es una de las conclusiones que tendremos en cuenta a la hora de armar nuestros modelos predictivos

2.4. Columnas candidatas a ser eliminadas

Las columnas agent y company, son candidatas a ser eliminadas debido a que son las columnas con más cantidad de registros faltantes. Sin embargo, en el afán de no perder valores a la hora de crear modelos predictivos, decidimos llenar esta columnas con valores falsos de -1. Este valor falso permitirá su filtración a la hora de armar modelos predictivos. Pero en caso de ser necesario es una columna que puede ser eliminada. Para las demás variables, no encontramos necesidad en eliminar ninguna columna, ya que consideramos que todas las columnas nos brindarán un buen valor a la hora de armar nuestros modelos predictivos.

2.5. Valores atípicos

Durante la limpieza de datos se encontraron una serie de valores atípicos que parecieran estar mal cargados o fuera de rango. Esto se pudo ver con las columnas de ADR (promedio del precio por noche) cuyo valores llegaban a 0 y valores negativos, lo cual implicaría que una reserva seria gratis o que le pagaran a alguien para ir a un hotel. Este es un caso de valor atípico univariado. Un comportamiento similar se puede observar en con las columnas adults, children y babies. Ya que hay reservas que poseen 0 adultos y con valores de children o babies mayores a 0. Esto se consideran valores atípicos multivariados ya que no puede haber una reserva sin adultos. Otro caso de valores atípicos que se pueden observar son los casos en los que ambas columnas stays in week nights y stays in weekend nights ambas poseen valores de 0. Si individualmente alguna posee un valor de 0, no hay ningún problema, pero para el dominio del problema, no tiene sentido que ambas valgan 0 porque no se puede hacer una reserva para un hotel y que no se quede ninguna noche. Para todos estos casos se eliminaron los registros que presentarn estos valores atipicos ya que si se buscan reemplazar los valores podrian alterar los modelos predictivos que se querían hacer con data inválida.

3. Tareas realizadas

Integrante	Tarea
Tomas Gabriel Ayala	Limpieza de Datos, Armado de Reporte
Sebastian Olaran	Análisis de Correlaciones, Detección de Outliers
Juan Ignacio Giacobbe	Análisis de Valores Faltantes, Imputación de Datos, Análisis Exploratorio