Checkpoint 1 - Grupo 17

Systeam

Integrantes

Del Rio, Juan Sebastián - 103337

Brizuela, Sebastián - 105288

Agha Zadeh Dehdeh, Lucía - 106905

Fecha de entrega: 21 de Septiembre de 2023

Análisis Exploratorio

El dataset en el cual se realizaron el análisis, es una colección de datos sobre reservas de Hoteles (City Hotel, Resort Hotel) que tiene 31 columnas y 61913 filas, donde cada fila corresponde a una reserva de hotel y cada columna es una característica de esa reserva.

Variables relevantes

hotel, lead_time, adults, country, is_repeated_guest, previous_cancellations, previous_bookings_not_canceled, reserved_room_type, assigned_room_type, booking_changes, deposit_type, agent, days_in_waiting_list, total_of_special_request.

Se consideran relevantes ya que al haber hecho el análisis con el target a través de gráficos, se observó que había cierta tendencia a que ocurra una cancelación, por ende, aportarían a la predicción.

Country (Cualitativa nominal) es la variable más destacada porque en el análisis se observó que dependiendo del país de orígen hay mayor y menor cantidad de cancelaciones.

Preprocesamiento de Datos

Columnas eliminadas

Se eliminó la columna “company” al detectar un porcentaje significativamente elevado (94,91%) de datos faltantes.

Correlaciones detectadas

Correlación Positiva (Cercanas a 1):

stays_in_week_nights - stays_in_weekend_nights (0.49), previous_booking_not_canceled - is_repeated_guest (0.41), company - agent (0.51), adr - children (0.35), agent - stays_in_week_nights (0.20), adr - arrival_date_year (0.22), adr - adults (0.22), is_canceled - lead_time (0.29), company - arrival_date_year (0.24), company - stay_in_week_nights (0.21)

Correlación Negativa (Cercanas a -1):

arrival_date_week_number - arrival_date_year (-0.54), is_canceled - required_car_parking_spaces (-0.23), is_canceled - total_of_special_requests (-0.24), company - previous_bookings_not_canceled(-0.20), company - is_repeated_guest (-0.23)

Columnas recodificadas

No se realizó recodificación de columnas.

Valores atípicos

Se eliminaron registros de reservas de clientes repetidos que no tenían cancelaciones previas ni reservas previas no canceladas. También se revisó si había días cero o negativos en la nueva variable creada (“dias_totales”) a partir de la suma de las variables "stay_in_week_nights" y "stay_in_weekend_nights" pero no se encontraron tales valores.

Univariado

La variable “adr” presenta valores atípicos como ceros y negativos, por ende se van a eliminar los registros con ese valor.

En la variable “adults”, se clasificaron como valores atípicos aquellos que eran cero o negativos.

Se identificaron valores atípicos en la variable "lead_time"(Visualización 2) utilizando el método del Rango Intercuartil (IQR). En este análisis, se encontraron 1368 casos moderados y 433 casos severos, los cuales se clasifican como valores atípicos de tipo "outlier colectivo".

Para la variable “days_in_waiting_list” se observó mediante un diagrama de caja, que a partir de 250 días de espera, hay datos dispersos dando a entender que podrían ser outliers y al hacer un análisis se ve que es una cantidad reducida de registros entonces se decidió eliminarlos.

Se analizó los outliers de la variable “total_of_special_requests” con el método IQR, se observó 1312 outliers moderados y 163 severos, sin embargo, se consideró que en este contexto no serían outliers.

Multivariado

Dentro del análisis de outliers multivariables se realizará un análisis de la cantidad total de personas en la reserva y el tipo de habitación asignada que se caractericen por la presencia de muchos registros de un tipo de habitación con un bajo número de personas y, al mismo tiempo, un registro con un gran número de personas para el mismo tipo de habitación.

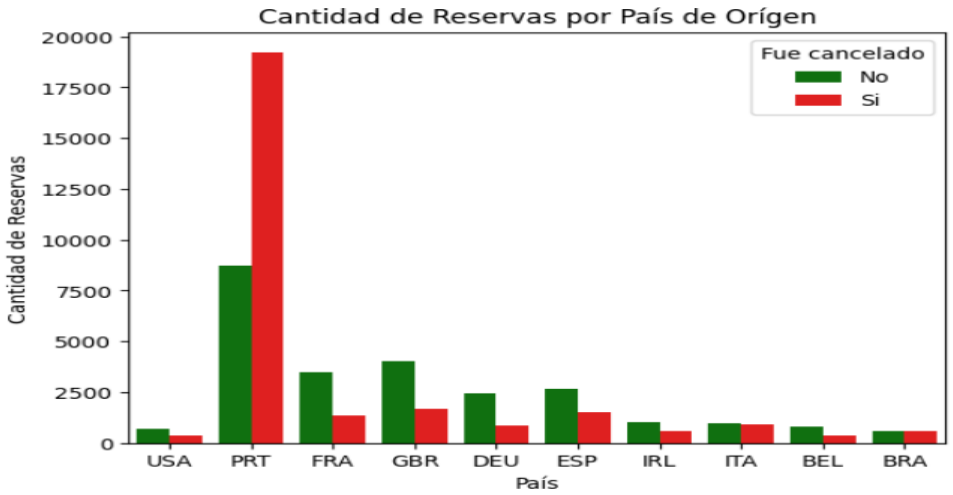
Mediante análisis gráfico se observó que por lo general en las habitaciones se asignan para menos de diez personas, por ende, las habitaciones que fueron asignadas para una cantidad mayor o igual se la considerara outlier.

Se realizó un análisis con previous_cancellations, “previous_bookings_not_canceled” y “booking_changes” con los métodos Z-Score, Z-Score Modificado e IQR, y se decidió quedarse con el resultado del Z-Score, ya que, al haber una gran cantidad de datos con valor cero, estaría afectando en los cálculos de los otros dos métodos, sin embargo, se decidió no eliminarlos debido a que no afecta al contexto del análisis.

Valores faltantes

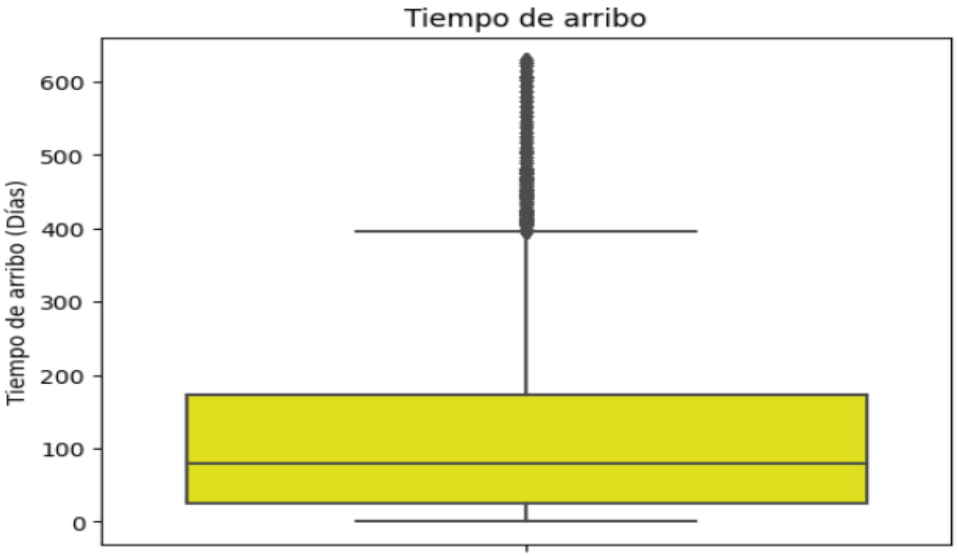
children: 0.0064% (4 datos faltantes) se imputan dichos datos utilizando la mediana.
agent: 12.74% (7890 datos faltantes) se creó una nueva categoría “agencia no existe” para la imputación.
company: 94.91% (58761 datos faltantes) se eliminó la variable.
country: 0,36% (221 datos faltantes) se utilizó la moda para imputarlos.

Visualizaciones



(Visualización 1)

Obs: Se eligió esta visualización ya que permite ver la relación entre la variable “country”, la cual se consideró más relevante respecto al resto, y el target.



(Visualización 2)

Obs: Se utilizó esta visualización para detectar valores atípicos, explicado anteriormente.
Se consideró como la segunda variable más importante para el análisis.

Tareas Realizadas

Se realizaron todas las tareas siempre en grupo por reuniones en meet por lo tanto no hubo una división clara de las mismas.