Checkpoint 1 - Grupo 12

Análisis Exploratorio

Nos encontramos con un dataset sobre reservas de hotel de 61913 filas y 31 columnas. Entre esas columnas podemos hallar 15 columnas que contienen datos cuantitativos discretos, 1 columna con datos cuantitativos continuos, 14 cualitativas nominales y 1 cualitativa ordinal.

Preprocesamiento de Datos

1. Columnas eliminadas:

Decidimos eliminar las siguientes columnas:

-*"id"*: Consideramos que no era necesaria para el análisis de datos, ya que solo representa un único id para cada fila.

-*"company"*: Al encontrarnos con 94.90% de sus datos faltantes y no encontrar correlación alguna con el target decidimos eliminarla.

2. Correlaciones detectadas:

Las mayores correlaciones entre variables y el target son:

-**"***lead_time***"** : Se puede ver que las reservas que se realizan con mayor anticipación tienen una tendencia a ser menos propensas a la cancelación, y esto se expresa por la correlación positiva de 0.29 (coeficiente de punto-biserie).

-*"required_car_parking_spaces"*: A medida que la cantidad de cocheras pedidas aumenta, la probabilidad de que una reserva sea cancelada tiende a disminuir en cierta medida, y esto se expresa por la correlación negativa de -0.23 (coeficiente de punto-biserie).

-*"total_of_special_requests"*: A medida que la cantidad de pedidos especiales del cliente aumenta, entonces la probabilidad de que la reserva sea cancelada, tiende a disminuir y esto se expresa por la correlación negativa de -0.24 (coeficiente de punto-biserie).

"deposit_type" y "*agent"* : Utilizando el método de chi-cuadrado y el coeficiente de Cramer, pudimos ver que estas variables tienen p-values muy bajos, lo que indica una asociación significativa con "*is_canceled*". Además, los coeficientes de Cramer's V indican

una correlación moderada (aproximadamente 0.440 para "*deposit_type*" y 0.396 para "*agent*") con respecto a la variable "*is_canceled*".

Y respecto de correlaciones calculadas con el coeficiente de Pearson (a destacar dentro de todas las variables cuantitativas), podemos decir que las variables *children* y *adult* se correlacionan positivamente con la variable *adr* (0.35 y 0.29).

3. Columnas recodificadas:

Nos encontramos con las variables: *arrival_date_day_of_month*, *arrival_date_year* y arrival_date_month representando la fecha en que se realizó la reserva y decidimos unirlas en una.

4. Valores atípicos:

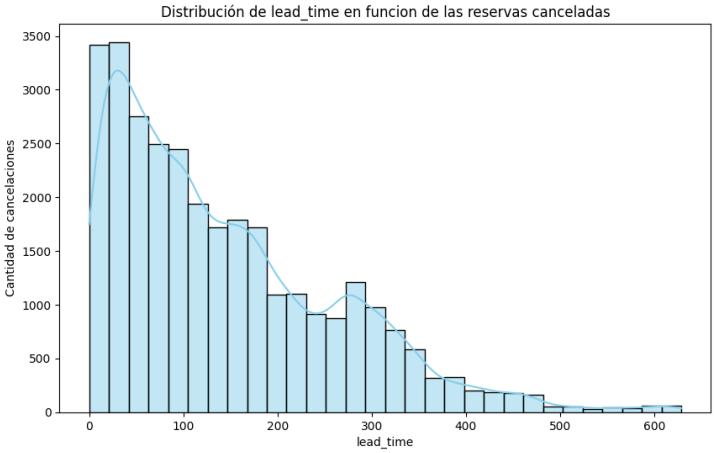
Realizando un análisis univariado a través de boxplots y calculando el rango intercuartílico detectamos como outliers:
-En "*lead_time*" arriba de los 400 días de espera.
-"*adults*" distintos de 1, "*children*" y "*babies*" distintos de 0.
-"*stays_in_week_nights*" superior a 6 y "*stays_in_weekend_nights*" superior a 5.
-"*total_of_special_requests*" mayor a 2.
-"*required_car_parking_spaces*" distinto de 0
-"*adr*" superior a 200
-"*days_in_waiting_list*" distinto de 0
También realizamos un análisis multivariado calculando la distancia de Mahalanobis sin embargo todos los outliers que detectamos caen fuera del rango de los outliers detectados por el análisis univariado.

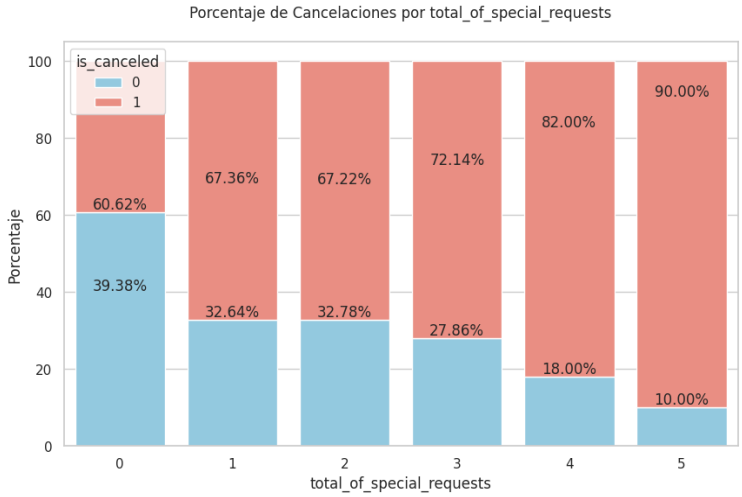
5. Valores faltantes:

-"*company*": Esta variable tiene 94,90% de los datos faltantes, luego de realizar una matriz de correlación con el target con un coeficiente muy bajo (-0,04) decidimos eliminar la columna.
-"*agent*": 12,74% de sus datos son nulos, decidimos rellenar los valores faltantes con la moda.
-"*meal*": 0,95% de sus datos están marcados como Undefined pero según el paper representa el mismo valor que 'SC' por lo tanto se cambió el valor de todos los Undefined por SC.
-"*country*": Con 0.35% de datos faltantes se los reemplazó con la moda.
-"*children*": 4 filas que contenían NaN en esta variable fueron eliminadas ya que representan menos del 0.01% del total del dataset.

-"*distribution_channel*" y "*market_segment*": Ambas columnas tienen Undefined en sus valores pero que representan menos del 0.01% de las filas y fueron eliminadas.

Visualizaciones





En el primer gráfico, se puede ver un histograma que muestra que a medida que aumenta la diferencia de días entre la reserva y el día que efectivamente el cliente llega al hotel, la probabilidad de que se cancele la reserva disminuye.

En el segundo gráfico observamos un gráfico de barras, que muestra que a medida de que aumenta la cantidad de pedidos especiales, la probabilidad de que se cancele la reserva disminuye.

Tareas Realizadas

Integrante	Tarea
Gaston Sabaj	Detección de correlaciones entre variables Análisis de datos faltantes y variables irrelevantes Análisis gráfico de distribución de variables
Juan Angel Gomez	Armado de Reporte Análisis de Variables Análisis de los Outliers
Juan Yago Pimenta	Análisis de Variables Imputación de datos Armado de Reporte