Reservas de Hotel

Introducción:

El presente informe busca analizar el propósito y procedimiento del checkpoint 1 del TP1: Reservas de hotel. El objetivo para este checkpoint era realizar un análisis exploratorio y procesar datos. Para esto se nos brindó un dataset, el cual tuvimos que estudiar y modificar para obtener la información necesaria para poder trabajar con el modelo, el cual nos debe permitir poder predecir la variable **`is_canceled`** de manera correcta.

Checkpoint 1: Análisis Exploratorio y Preprocesamiento de Datos:

En esta sección fuimos introducidos al problema y comenzamos a investigar el dataset de entrenamiento.

Comenzamos analizando las variables que teníamos a nuestra disposición, se observó que las 31 categorías podían ser de 3 tipos, object, float64 e int64. Estas fueron divididas en variables cuantitativas y cualitativas, para las primeras se observó su media, moda, mediana, mínimo y máximo. Y para las siguientes, se observó cuántos valores posibles podía tomar y cuales eran los más comunes.

Se tomó la decisión de despreciar las variables, **`company`** y **`id`**, ya que, la variable **`company`** era un valor nulo para el 95% del dataset. Y, **`id`** fue despreciada por sus características, por ser una variable de identificación, la cual, al final, no mostraba ningún tipo de variación en las predicciones, aun así esta variable no fue eliminada, ya que era de utilidad para la creación de nuevos datasets.

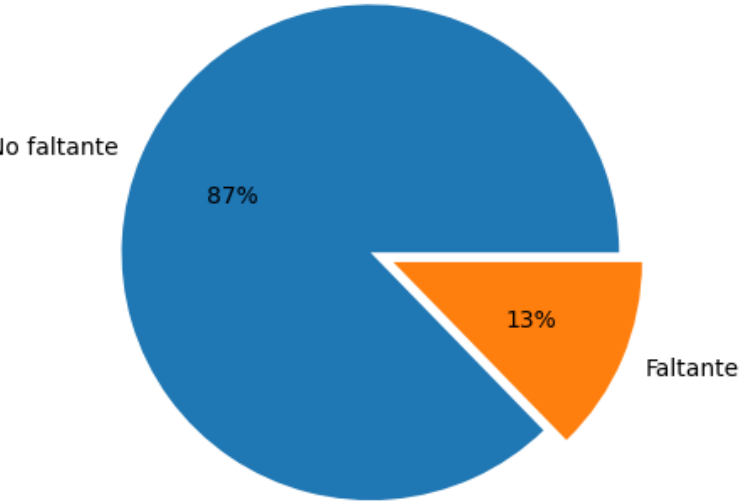
Una vez fuera las variables despreciables pudimos ver de mejor manera la distribución de datos y cómo se correlacionaban con el target. Para visualizar las variables utilizamos librerías como pandas, seaborn y matplotlib para graficar las variables, se utilizaron histogramas para la mayoría de las variables excepto en algunos casos particulares donde se utilizó boxplot. Para ver la correlación de las variables cuantitativas entre sí con el target utilizamos el coeficiente de correlación de Pearson y luego graficamos un heatmap para obtener una mejor visión de las correlaciones. Para las variables cualitativas hicimos un procedimiento similar pero utilizamos el coeficiente de la V de Cramer.

Con las correlaciones pudimos comenzar a encarar los valores nulos o mal cargados, así como los valores atípicos. Se decidió imputar la variable **`agent`**, ya que los valores nulos no estaban mal cargados, si no, que se indicaba de esta manera cuando un huésped no utilizaba un agente de viajes, en estos casos se cambió el nulo por el 0 (cero), y las variables **`children`** y **`country`** no se imputaron, la primera por ser pocos valores y la segunda se decidió esto ya que la variable se encuentra moderadamente relacionada con el target, y así, evitamos contaminar los datos. Luego para 13 variables se decidió eliminar los

valores atípicos donde 4 los hicimos mediante el método de IQR y los 9 restantes fueron analizados individualmente ya que muchos de los valores atípicos eran coherentes y no resultaba lógico eliminarlos.

Una vez estos procesos terminados, volvimos a graficar histogramas para comparar cómo se modificaron las variables.
Durante el proceso se utilizaron 2 data frames, los cuales se mergearon al final para proporcionar un dataframe limpio.

En el gráfico de arriba podemos ver el heatmap de la matriz de correlación de las variables cuantitativas.



En este gráfico podemos ver la cantidad de datos nulos por agente de viajes.