FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1 - Checkpoint 1

[7506/9558] Organizacion de Datos
Grupo 21
Primer cuatrimestre de 2023

Alumno	Padron	Email
Martin Pata Fraile de Manterola	106226	mpata@fi.uba.ar
Tobias Emilio Serpa	108266	tserpa@fi.uba.ar
Juan Francisco Cuevas	107963	jcuevas@fi.uba.ar

1. Exploración Inicial

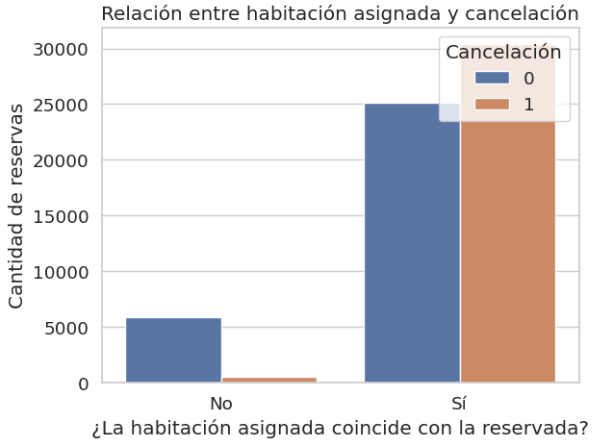
En primer lugar, agrupamos todas las variables por su naturaleza, en cuantitativas discretas/continuas o en categóricas. Hecho esto, nos dedicamos a analizar todas las variables cuantitativas, realizando los cálculos de sus cuantiles, mediana, moda, etc. Para cada variable realizamos los gráficos correspondientes (dependiendo de si son continuas o discretas). Luego, nos dedicamos a graficar las variables categóricas, para así poder ver que valores toman y con que frecuencia lo hacen. Cabe destacar que todos los gráficos mencionados anteriormente, se hicieron además indicando la tasa de cancelación de cada valor dentro de las variables, para de a poco ir buscando patrones para nuestro modelo.

A continuación, realizamos los pairplots para analizar la relación entre las variables, teniendo en cuenta nuestro target en todos los gráficos. Seguido a esto, realizamos una matriz de correlación (con el método de Pearson para las cuantitativas, y V de Cramer para las categóricas) para analizar estas relaciones.

Gracias a todos los gráficos individuales, los pairplots y la matriz de correlación, pudimos ver que había variables que de manera individual no presentan ningún patrón modelable para la deducción de si una reserva se cancelará o no. Además, tampoco poseen relaciones relevantes para el target a la hora de combinarse con otras variables.

2. Visualización de los datos

En esta sección, realizamos gráficos de relaciones que considerábamos que iban a mostrar un patrón interesante para la deducción de si una reserva se cancelará o no. Para dar un ejemplo de nuestro accionar, buscábamos situaciones que considerábamos que lógicamente mostrarían un patrón, como en el caso de que una reserva tenga un tipo de cuarto, pero que a la hora de asignación le den otro tipo. O que pasa con una persona que suele cancelar reservas.



3. Datos Faltantes

Para comenzar, calculamos el porcentaje de valores nulos (Nan) en cada columna, visualizando que la columna “Agente” posee una cantidad relevante, y “Compañía” posee casi su totalidad de valores en Nan. Luego, verificamos que no existan valores negativos, hallando solo uno en “ADR” considerandolo un error de tipeo y reemplazándolo por el mismo valor absoluto positivo. Luego entendimos que adr tampoco podía tener valor 0 (ya que sería no pagar la reserva) y se acomodó con un modelo definido por el grupo. Debido al alto porcentaje de valores nulo en “Company” decidimos eliminar esa columna. Además se chequeó que la cantidad de noches de reserva entre semana, tenga lógica con la cantidad de noches entre fin de semana.

4. Outliers

Por ultimo, en la sección de outliers en algunas parte fue mejor tomar la decisión de eliminar esos outliers que fuimos descubriendo con los métodos propuestos, y en otros casos simplemente no. Había casos en los que eliminar los outliers que los métodos nos proporcionaban por ahí no tenían tantos sentido, como por ejemplo, en el caso del "lead_time"(días de anticipo) había casos en los que los métodos como el z-score utilizando la regla de oro, te decía que una cantidad de casos eran "outliers" y eran casos en los que por ahí el huésped había reservado con 2 años de anticipo y si realmente te pones a pensar eso no seria para nada raro.

Por eso muchas veces aparte de usar los métodos (z-score, z-score modificado, boxplots, etc) hay que fijarse debidamente en cada caso analizando un poco mas allá de lo que nos dicen los datos para decidir lo que se va a terminar de hacer con esos posibles "outliers".
Claramente hay casos particulares en los que por ahí no podes llevarlo a la "realidad" y aplicando algún método podes llegar a una conclusión acerca de que datos son considerados outliers y así poder eliminarlos. Por ejemplo, en los "días de estadía" utilizando los métodos de z-score modificado y aparte el método de mahalanobis pudimos definir un umbral para el cual los valores que superen ese umbral serian considerados outliers y así eliminados del dataframe.