Trabajo Práctico 1

Reservas de Hotel

Integrantes

Apellido	Nombre	Grupo
Lopez	Francisco	17
Corn	Franco	17
Queirolo Dominguez	Cristian Daniel	17

Informe

Introducción

En esta primera parte del trabajo buscaremos realizar un análisis de los datos del dataset “hotels_train”. Analizaremos las variables y buscaremos dividirlas según su tipo, graficaremos su distribución, trataremos de determinar cuáles son útiles y cuales no apoyándonos en gráficos.

Por último, veremos si hay variables nulas y tomaremos decisión respecto a ello.

Objetivos

a) Exploración Inicial : analizar cada variable, considerando los siguientes aspectos

● Tipo de variable

● Variables Cuantitativas: calcular medidas de resumen: media, mediana,moda, etc

● Variables Cualitativas: reportar los posibles valores que toman y cuán frecuentemente lo hacen.

● Determinar si existen variables irrelevantes para el análisis

● Realizar un análisis gráfico de las distribuciones de las variables

● Analizar las correlaciones existentes entre las variables.

● Analizar la relación de las variables con el target

b) Visualización de los datos: en esta sección se espera que puedan realizar una primera aproximación a los datos apoyándose en visualizaciones, por ejemplo: gráficos de dispersión entre variables, histogramas, heatmaps, exploración de las columnas y cualquier otro gráfico adicional que se considere útil justificando su utilización.

c) Datos Faltantes : analizar la presencia de datos faltantes en el dataset

● Realizar análisis de datos faltantes a nivel de columna. Graficar para cada variable el porcentaje de datos faltantes con respecto al total del dataset

● Revisar los datos faltantes o mal ingresados y tomar una decisión sobre estos: reemplazo de valores, eliminación de registros incompletos, etc.

● En caso de realizar imputaciones comparar las distribuciones de cada atributo reparado con la distribución anterior a la imputación de los datos faltantes.

Desarrollo

Como primera aproximación a los datos mostramos el dataset y analizamos qué tipos de datos contiene, observando que muchas variables son del tipo object decidimos cambiarlas a un tipo adecuado para poder realizar los gráficos y otras acciones sobre ellas. Dividimos las variables en cuantitativas y cualitativas, una vez tengamos las variables divididas realizamos una lista de las cuantitativas y armamos dos dataframes (uno para cada tipo de variable). Esto lo hacemos con el objetivo de poder manipular las variables por separado si llega a ser necesario. Realizamos las acciones que nos pide el punto “a” sobre dichas variables y luego determinamos qué variables son irrelevantes para el análisis. En este caso decidimos que la variable “id” no es para nada útil para el análisis, ya que el id es un combinación única de caracteres que no se repite, por lo tanto no la tendremos en cuenta, luego reservation_status y reservation status_date tampoco las tendremos en cuenta, ya que nos dicen lo mismo que nuestra variable target y pueden llegar a confundir nuestro análisis.

Para la distribución de las variables vamos a mostrar la distribución de las que creamos más importantes (en este caso podríamos mostrar las 30 variables, pero no nos parece lo mejor), mostraremos primero algunas de las cualitativas y luego las cuantitativas (en este caso las mostraremos todas, ya que son más fáciles de graficar). Aclaramos que si bien “children” no es una variable cualitativa, vamos a tener que analizar en el punto c, por lo tanto la mostramos para que tengamos una mayor claridad de su distribución.

Analizamos la correlación entre todas nuestras variables y cuando analizamos nuestra variable target (“is_canceled”) mostraremos 2 heatmaps, para dejar más clara las correlaciones (dividiéndolos en los que tienen una relación negativa con nuestra variable target y una positiva).

Respecto al punto b) “visualización de datos” tomamos la decisión de analizar algunas de las variables cuantitativas y alguna cuantitativas, realizando gráficos de dispersión, para observar mejor cómo están dispersos nuestros datos. En este punto hacemos una primera aproximación a los datos viendo nuestros gráficos de dispersión (sin embargo, ya sabemos bastante de nuestros datos, ya que hemos graficado y tomado decisiones sobre ellos).

En el punto c) “datos faltantes” observamos los datos que nos faltan en el dataset (osea los que son nulos), graficamos dichos datos vemos que en el caso de children, agent y country no tenemos una cantidad de datos faltantes muy alta, por lo que decidimos tomar una moda del que más se repite y realizar un imput. company por otro lado, tiene una cantidad de datos faltantes altísima (94%) por lo que decidimos eliminarla (ya que casi no tenemos información sobre ella). Por último graficamos las columnas a las que le imputamos los

datos y comparamos nuestro primer gráfico de distribución con este último. En el cual solo notamos un cambio significativo en agent.