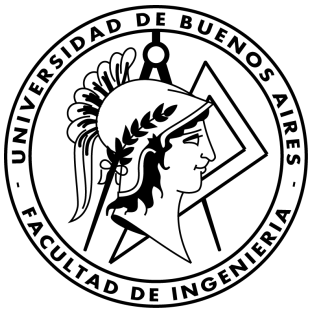

Trabajo Práctico 1
Reservas de hotel - Resumen checkpoint 1

[75.06] Organización de datos
Primer cuatrimestre 2023
Grupo 19: Sudanalytics

Integrantes

Nombre	Padrón	Mail
Re, Adrián Leandro	105025	are@fi.uba.ar
Lorenzo, Luciano Andrés	108951	llorenzo@fi.uba.ar
Toulouse, Alan	105343	atoulouse@fi.uba.ar

Fecha de entrega: 13/04/2023

Índice

1. Introducción

El objetivo del trabajo práctico es analizar reservas de dos hoteles para a través de ellas entrenar modelos que analicen los datos y sean capaces de poder predecir si una futura reserva será cancelada o no. Para esta primer parte, sólo nos limitamos a la ingeniería de características, es decir, analizar y visualizar cada columna y sus valores en relación a la variable target (la que nos indica si una reserva fue cancelada o no).

2. Dataset

Tenemos dos conjuntos de datos los cuales consisten de reservas en dos hoteles reales hechas desde el primero de julio de 2015 hasta el 31 de agosto de 2017, en donde uno corresponde a los 42.129 registros del primer hotel y otro contiene 19.784 del segundo. Cada registro representa una reserva. Ambos conjuntos cuentan con 31 variables en las cuales se describen las características de la reserva (al ser datos reales, se eliminó cualquier elemento de los conjuntos de datos que pueda exponer la identidad de los clientes).

Lo primero a tener en cuenta en el dataset de entrenamiento es que (casi) la mitad de los registros son de cancelaciones y la otra mitad no. Esto nos dice que al analizar independientemente una variable, esta no va a tener **individualmente** una relación con la cancelación de la reservación si tiene mitad de registros de cancelaciones y la otra mitad no cancelaciones. Esto no quiere decir necesariamente que la mitad de las reservaciones fueron canceladas, sino más bien que fueron seleccionadas para que haya la misma cantidad de registros de ambas para su análisis.

2.1. Variables descartadas

Como explicamos en el notebook, en principio solo descartamos la variable `arrival_date_year`, ya que no queremos entrenar a los modelos para encontrar un patrón con respecto a un año específico, sino mas bien generalizar para cualquier año. Ya que evidentemente en un futuro solo van a llegar reservas de años posteriores.

Aunque también encontramos otras variables como `meal`, que no parecen tener una fuerte relación con la cancelación de la reserva, pero no descartamos que puedan servir en conjunto con otras variables.

2.2. Variables mas útiles

Una de las variables que a priori parecen mas útiles es `lead_time`. observamos como las reservas que se hacen el mismo día de la llegada (lead time $== 0$) **no suelen ser canceladas**. Esto se puede deber a que llegan al hotel sin una reserva, y al hacerlo en el momento, evidentemente no la van a cancelar. En cambio si la reserva se hace con mucha antelación (más de 100 días), la **tendencia suele ser que la reserva sea cancelada**. Por estas razones, a priori `lead_time` parece tener una gran relación con `is_canceled`. Por otro lado, otra variable que resulta útil para el análisis de cancelaciones es la variable `required_car_parking_spaces` ya que notamos que ninguna de las reservas que fueron canceladas reservaron un lugar de estacionamiento.

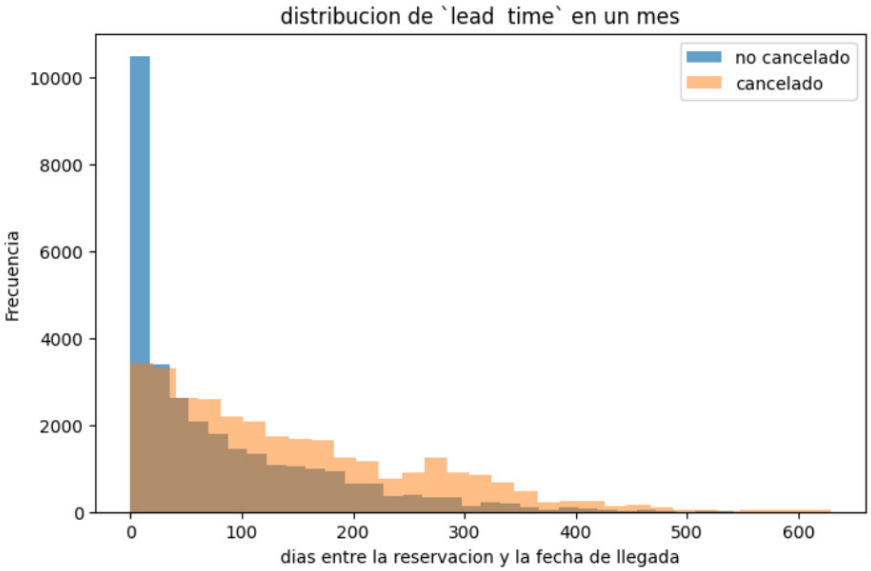

Figura 1: Gráfico del lead time vs. frecuencia

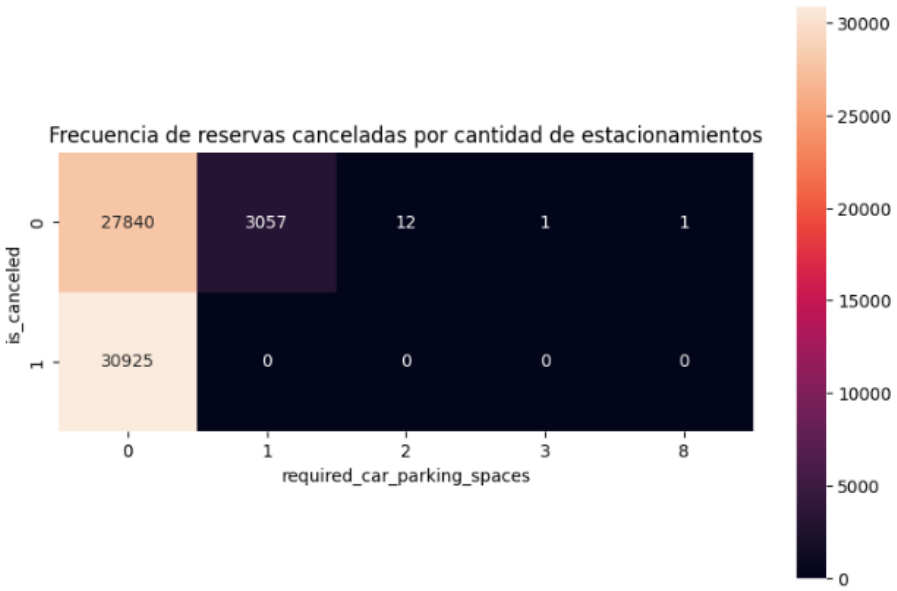

Figura 2: Gráfico de reservas canceladas por cantidad de estacionamientos

2.3. Nuevas variables

Creamos nuevas variables, que son derivadas de otras variables, y que nos podrían ayudar en la etapa de entrenamiento para crear modelos que generalicen mejor. Un caso de esta es `es_extranjero`, la cual creamos para poder seguir teniendo la noción de la procedencia del huésped, pero sin hacer que el modelo se 'memorice' los resultados de cada país. Otro caso parecido es el `tiene_hijos`. Estas variables vamos a probarlas en el momento del entrenamiento, y veremos si ayudan al modelo a generalizar mejor y mejorar los resultados.

3. Referencias

- Paper "Hotel booking demand datasets"proporcionado por la cátedra.