Informe del Trabajo Practico 1

Facultad de Ingeniería de la Universidad de Buenos Aires

Grupo 09

Castro Martinez, Jose Ignacio
Padrón: 106957
email: jacastrom@fi.uba.ar

Douce, German Alejandro
Padrón: 106001
email: gdouce@fi.uba.ar

Orsi, Tomas Fabrizio
Padrón: 109735
email: torsi@fi.uba.ar

Hotels

El data frame está compuesto por un conjunto de 61913 filas (registros) y 31 columnas en donde cada registro representa una reserva de hotel y cada columna es una variable distinta que brinda información acerca de dichas reservas. El objetivo de esta primera parte es realizar un análisis exploratorio del DataFrame e identificar Outliers para poder trabajar con uno más "limpio" a la hora de predecir nuestra variable target: "is_cancelled". Variables con pocos valores que tienen mucha desviación podrían generar "ruido" y falsas tendencias en las reservas a cancelar por sí o por no. También se puede hacer una predicción más "honesta" evitando sesgos sobre algunas variables. El notebook se divide de la siguiente manera:

- Preparación del ambiente de trabajo
- Análisis univariado
  - Cuantitativas
  - Cualitativas
- Análisis multivariado
- Relación contra el target: "is_canceled"

Preparación del ambiente de trabajo

En esta sección, realizamos los imports necesarios para trabajar con el Data Frame. Luego mostramos algunos registros y listamos las columnas con sus tipos de datos. Además, cambiamos los nombres de las variables por otros más declarativos.

Análisis univariado

Listamos las variables y las dividimos en Cualitativas y Cuantitativas. De las 31 variables, 16 son cuantitativas y 15 son cualitativas.

Cuantitativas

Para cada una de ellas:

- Vemos sus valores estadísticos más relevantes;
- Detectamos la cantidad de valores nulos;
- Construimos una gráfica de distribución de sus valores en algunos casos gráficas de tipo box-plot para identificar más fácilmente valores atípicos (Outliers) y en el caso de la variable "average_daily_rate" agregamos un análisis con Z_score;
- Realizamos un tratamiento sobre dichos valores eliminando sus registros o reemplazando los valores atípicos por otros.

Para estas variables concluimos lo siguiente:

- La única variable con datos nulos o faltantes es "babies_number";
- Decidimos eliminar la variable "previous booking not cancelled" debido a que, después de remover unos pocos valores outliers, todo el resto de los registros presentan el mismo valor. Esto no nos aporta ningún tipo de información

Cualitativas

Para estas, en primer lugar analizamos las variables con valores nulos. Estas fueron: "agent_id", "company_id" y "country" de las cuales se decidió eliminar company_id por contener un 92% de

valores faltantes.
Para cada una de ellas:

- Mostramos los valores que toman;
- Realizamos una gráfica de distribución de sus valores
- En el caso de “agent_id” y “country” se corrigieron sus valores valores faltantes

Análisis multivariado

En esta sesión hicimos multivariado entre las siguientes parejas de variables. En algunos de ellos encontramos algunos Outliers que luego eliminamos y en otros no. Además agregamos la columna “dias_totales” al Data Frame.

- “weekend nights” vs “weekend nights num” (de las suma de estas 2 nace la columna dias totales)
- adr vs customer Type
- room type vs adr
- Children Num vs Babies num vs Adult Num

Relación contra el target: “is_canceled”

Se realizaron gráficas de las siguientes variables comparándolas contra el target.

- Lead Time
- Previous_cancellations_num
- Average_daily_rate
- dias_totales
- reserved_room_type

De las 5 variables analizadas solo 2 parecen tener cierta influencia sobre el target: “Lead time” y ”Previous cancellations number”. En el caso de ”Lead time”, su gráfica podría sugerir que reservas realizadas con mayor anticipación tendrían más probabilidad de ser canceladas.
Para Previous cancellations number, hay un salto grande en el numero de cancelaciones cuando hay una cancelación previa
El resto de las gráficas tienen distribuciones muy homogéneas al hacer foco en el número de reservas canceladas. Mostramos abajo las distribuciones de ”lead time” y ”previous cancellations” con foco en ”is canceled”