**Checkpoint 1 - Grupo 07**

Análisis exploratorio

El data frame “*Hotels*” está formado por 63193 registros y 31 columnas, en donde cada registro representa una reservación de hotel y cada columna aporta algún tipo de información sobre esta.

A continuación, se listan las variables de las columnas del data frame clasificadas según su tipo de dato:

1. **Cualitativas:**

* **Nominal:**
* Hotel
* Arrival date month
* Country
* Market Segment
* Distribution Channel
* Is Repeated Guest
* Reserved Room Type
* Assigned room type
* Deposit type
* Company
* ID
* Customer type
* Is canceled
* Agent
* **Ordinal:**
* Meal

1. **Cuantitativas:**

* **Discretas:**
* Lead time
* Arrival date year
* Arrival date week number
* Arrival date day of month
* Stays in weekend nights
* Stays in week nights
* Adults
* Children
* Babies
* Previous Cancellations
* Previous Bookings NotCanceled
* Booking changes
* Days in waiting list
* Required car parking spaces
* Total of special requests
* **Continuas:**
* ADR

Hipótesis

Basándonos en la lógica, el análisis realizado sobre los datos recibidos y los resultados obtenidos, el equipo ha sido capaz de plantear algunas hipótesis:

1. **Relación entre *Lead time* y target:** Los datos parecen indicar que un mayor lead time en una reserva, suele tener como resultado una mayor probabilidad de cancelamiento.
2. **Relación entre *Total of special requests* y target:** También parece ser que, a mayor cantidad de pedidos especiales hechos, el cliente es menos propenso a cancelar la reserva hecha.
3. **Relación entre *Required car parking spaces* y target:** Por último, creemos que, a mayor cantidad de espacios de estacionamiento pedidos, el cliente tiene nuevamente menor tendencia a cancelar su reserva.
4. **Relación entre *Previous Bookings NotCanceled* y target:** Creemos que un cliente usual con amplia cantidad de reservas previas no canceladas tendrá menor probabilidad de cancelar una reserva futura.

Preprocesamiento de Datos

**1. Variables eliminadas**

**1. Company:** Fue eliminada debido a presentar una gran cantidad de valores faltantes (> 90%).

**2. ID:** Se comprobó que no existen 2 reservas de igual ID, por lo que se puede despreciar esta variable ya que el data set corresponde a todas muestras de reservas diferentes.

**3. Agent:** Se imputaron los datos faltantes, ya que era < 13%, se rellenó con el valor más frecuente, ya que ‘agent’ es una variable categórica

**4. Arrival date week number:**Como vimos en las visualizaciones de variables numericas podemos observar, por ejemplo, que la variable’ arrival\_date\_week\_number’ representa la misma informacion que ‘arrival\_date\_month’ solo que en otra unidad de medida (una está representada en meses y la otra en semanas) por ende decidimos quedarnos solo con ‘arrival\_date\_month’ que nos pareció más representativa ya que podemos crear por ejemplo nuevos campos y ademas es mas abarcativa

**2. Correlaciones detectadas**

1.*Agent*y *Company* – Indice de correlación: 0.51

2. *Stays\_in\_week\_nights* y *Stays\_in\_weekend\_nights* - Indice de correlación: 0.49

3. *Arrival\_date\_year* y *Arrival\_date\_week\_number* - Indice de correlación: -0.54

4. *is\_repeated\_guest* y *previous\_bookings\_not\_canceled* - Indice de correlación: 0.41

5. *children* y *adr* - Indice de correlación: 0.35

6. *lead\_time* y *is\_canceled* - Indice de correlación: 0.29

7. *total\_of\_special\_requests* y *is\_canceled* - Indice de correlación: -0.24

8. *required\_car\_parking\_spaces* y *is\_canceled* - Indice de correlación: -0.23

**3. Columnas recodificadas**

1.Arrival\_date\_month – los meses pasaron de representarse como strings a números

2. Aquellas variables que eran floats64 como ‘children’ pero que en realidad eran enteros, se recodificaron a enteros.

**4. Valores atípicos**

**- Análisis univariado:** Se utilizó el rango intercuartílico (IQR) y el gráfico de caja (boxplot) para determinar la presencia de valores atípicos en las columnas. Las columnas identificadas mediante este método son las siguientes:

*-lead\_time***:** Se encontraron valores atípicos para > 600 días, los cuales se conservan.

*-stays\_in\_weekend\_nights* y *stays\_in\_week\_nights*: Se identificaron valores atípicos para > 9 y > 8, respectivamente, y se conservan.

*-adults*: Se consideraron atípicos aquellos valores que estaban fuera del rango < 2 y > 2, por lo que se eliminaron los valores mayores a 2.

*-children, babies, previous\_cancellations, days\_in\_waiting\_list,*

*previous\_booking\_not\_cancelled, booking\_changes* y *car\_parking\_spaces*: En todas estas variables se encontraron valores atípicos para > 0, y se optó por conservar todos ellos.

*-total\_of\_special\_requests*: Se identificaron valores atípicos para > 2 y se conservan.

*-adr*: Se encontró un valor atípico > 600, el cual se conserva.

Diagrama

Descripción generada automáticamente

**-Análisis multivariado:** Se utilizaron las técnicas de Mahalanobis e Isolation Forest para identificar valores atípicos en las variables. Las variables que se compararon incluyen:

*-adult*, *children*, *babies*: Se consideraron como valores atípicos aquellos registros en los que *'adult'* era igual a 0 y *'children'* era mayor que 0, y se eliminaron estos registros.

-*stays\_in\_weekend\_nights* y *stays\_in\_week\_nights*: Se identificaron valores atípicos para > 10 y > 30, respectivamente, y se eliminaron.

-*adr* y *stays\_in\_week\_nights*: Se eliminaron los registros en los que *'stays\_in\_week\_nights'* era mayor que 21 y *'adr'* era menor que 150.

-*days\_in\_waiting\_list* y *booking\_changes*: Se aplicaron los siguientes criterios de rango:

1. Para *'days\_in\_waiting\_list'*, se eliminaron los registros en los que *'days\_in\_waiting\_list'* era mayor que 300 o igual a 0.

2. Para '*booking\_changes'*, se eliminaron los registros en los que *'days\_in\_waiting\_list'* era igual a 0 y *'booking\_changes'* era mayor que 10, así como aquellos en los que *'days\_in\_waiting\_list'* era menor que 5.

**5. Valores faltantes**

1. La variable *Company* fue eliminada ya que presenta un porcentaje > 90% de valores faltantes.

2. La columna *Agent* presenta un porcentaje de valores faltantes levemente < 15%. Es por eso que se decidió ¿??.

3. El resto de las columnas presentan porcentajes de valores faltantes < 3% o directamente inexistentes, por lo que se decide eliminar las muestras incompletas.

4. Se corrobora que no haya muestras repetidas, ya que esto correspondería a un error. No podría haber dos muestras de mismo *ID*.

Visualizaciones

Imagen que contiene Gráfico

Descripción generada automáticamente

*Visualización de la matriz de correlación*

Gráfico, Gráfico de barras

Descripción generada automáticamente

*Visualización de cantidad de reservas canceladas y no canceladas.*

Tareas realizadas

|  |  |
| --- | --- |
| **Integrante** | **Tareas realizadas** |
| Guido Menendez | Exploración inicial, visualización de datos, análisis de valores faltantes, confección de reporte. |
| Rafael Wu | Detección de outliers, confección de reporte. |
| Mateo Riat Sapulia | Análisis de valores faltantes, visualización de datos, confección de reporte. |