**Dataset**

Para la realización de este análisis contamos con un dataset con 119390 observaciones con las características asociadas a la reserva. El dataset tiene como target la variable “is\_canceled” que toma el valor 1 en el 37% de los casos.

Contamos inicialmente con 32 atributos con las siguientes característica:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  | |  |  | |
| Atributo | **Descripción** | **Tipo** | **Ppales características** | | |
| hotel |  | C | 2 categorías | | |
| lead\_time | Tiempo total del proceso de reserva | N |  | | |
| arrival\_date\_year/ arrival\_date\_month/  arrival\_date\_week\_number/  arrival\_date\_day\_of\_month | fecha de llegada: año, mes, semana del año, día del mes | N | Fechas comprendidas entre 2015 y 2017 | | |
| stays\_in\_weekend\_nights | Cantidad de noches de estadía según dia de semana /fin de semana | N | Entre 0 y 19 con media 0,9 | | |
| stays\_in\_week\_nights | Entre 0 y 50 con media 2,5 | | |
| adults | Composición del grupo que realiza la reserva | N | Valores comprendidos entre 0 y 55 – media 1,86 | | |
| children | N | Valores comprendidos entre 0 y 10 – media 0,1 | | |
| babies | N | Valores comprendidos entre 0 y 10  Alta asimetría (99% de los valores en 0) | | |
| meal | Tipo de comida | C | 5 posibles valores  77% de los casos en BB | | |
| country | Pais de origen | C | 177 posibles valores  73% concentrados en los primeros 5 | | |
| market\_segment |  | C | 8 posibles valores Online TA 47% | | |
| distribution\_channel |  | C | 5 posibles valores  TA/TO 92% | | |
| is\_repeated\_guest |  | C | Valores 0 y 1  97% de los valores en 0 | | |
| previous\_cancellations | Cantidad de cancelaciones previas | N | Valores entre 0 y 26 con alta asimetría (95% de los valores en 0) | | |
| previous\_bookings\_not\_canceled | Reservas previas no canceladas | N | Valores entre 0 y 26 con alta asimetría (97% de los valores en 0) | | |
| reserved\_room\_type | Tipo de habitación reservada | C | 10 posibles valores  Tipo A frec 72% | | |
| assigned\_room\_type | Tipo de habitación asignada | C | 12 posibles valores  Tipo A 72% | | |
| booking\_changes | Cambios en la reserva | N | Valores entre 0 y 21 con alta asimetría (85% de los valores en 0 | | |
| deposit\_type | Tipo de depósito | C | 3 posibles valores  No deposit 88% | | |
| agent | Agencia | N | 333 posibles valores  Agente 9 y 240 38% | | |
| company | Compañía | N | 352 valores posibles  94,3% missing | | |
| days\_in\_waiting\_list | Dias en lista de espera | N | Valores entre 0 y 391 con alta asimetría (97% de los valores en 0) | | |
| customer\_type | Tipo de cliente | C | 4 posibles valores  Transient 75% | | |
| adr | ¿? | N | Valores entre -6 y 5400 | | |
| required\_car\_parking\_spaces | Espacios de estacionamiento solicitados | N | Valores entre 0 y 8 con alta asimetría (94% de los valores en 0) | | |
| total\_of\_special\_requests | Pedidos especiales | N | Valores entre 0 y 5 con media 0,57 | | |
| reservation\_status | Estado de la Reserva | C | 3 posibles valores | | |
| reservation\_status\_date | Fecha estado reserva | N | 926 posibles valores | | |

**Selección de variables predictoras**

Teniendo como objetivo del modelo la predicción de la cancelación de la reserva se decidió descartar los siguientes atributos:

* Aquellos que se definen posteriormente ya que no estarán disponibles al momento de la utilización del modelo. Estos atributos son: **reservation status, reservation\_status\_date y required\_car\_parking\_spaces** (este última se deduce ya que en el 100% de los casos cancelados se encuentra en 0)
* **arrival\_date\_year** ya que, de tenerlo en cuenta, el modelo podría clasificar mal solicitudes realizadas en fechas futuras.

Por otro lado se generaron los siguientes atributos:

Estadía Total: variable numérica que resulta de la suma de stay

Tipo de familia: variable C que resume la composición de la/s persona/s que realizan la reserva

Diferencia\_room type: C que toma el valor 1 si existen cambios entre el tipo de habitación reservada y asignada

**Correlación entre atributos y variable target**

Por tener tanto variables Cs como continuas en el dataset siendo la variable target C se decidió realizar el análisis utilizando el coeficiente de correlación ϕK,

<https://arxiv.org/abs/1811.11440>

**Aplicación de árboles de decisión**

Dado que se decidió abordar este problema utilizando árboles de decisión para predecir la cancelación y el algortimo a utilizar es DecisionTree se deben imputar valores donde haya valores faltantes.

----------------------------------------------------------------------------------------

Decidir entre variables correlacionadas

