Business Problem

An Online travel booking company is suffering from loss in revenue because of the uncertain booking cancellation of its customers. The company wants to know which customer will cancel the booking.

Dataset Details

hotel (H1 = Resort Hotel or H2 = City Hotel)

is_canceled Value indicating if the booking was cancelled (1) or not (0)

ead_time Number of days that elapsed between the entering date of the booking into the PMS and the arrival date

arrival_date_year Year of arrival date

arrival_date_month Month of arrival date

arrival_date_week_number Week number of year for arrival date

arrival_date_day_of_month Day of arrival date

stays_in_weekend_nights Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

stays_in_week_nights Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel

adults Number of adults

children Number of children

babies Number of babies

meal Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal

country Country of origin. Categories are represented in the ISO 3155–3:2013 format

market_segment Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

distribution_channel Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”

is_repeated_guest Value indicating if the booking name was from a repeated guest (1) or not (0)

previous_cancellations Number of previous bookings that were cancelled by the customer prior to the current booking

previous_bookings_not_canceled Number of previous bookings not cancelled by the customer prior to the current booking

reserved_room_type Code of room type reserved. Code is presented instead of designation for anonymity reasons.

assigned_room_typeCode for the type of room assigned to the booking.Code is presented instead of designation for anonymity reasons.

booking_changes Number of changes made to the booking from the moment the booking was entered on the PMS until the moment of check-in or out

deposit_type Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No

agent ID of the travel agency that made the booking

company ID of the company that made the booking or responsible for paying the booking.

days_in_waiting_list Number of days the booking was in the waiting list before it was confirmed to the customer

customer_type Type of booking, assuming one of four categories:Transient - Transient-Party - Contract - Group

adr Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights

required_car_parking_spaces Number of car parking spaces required by the customer

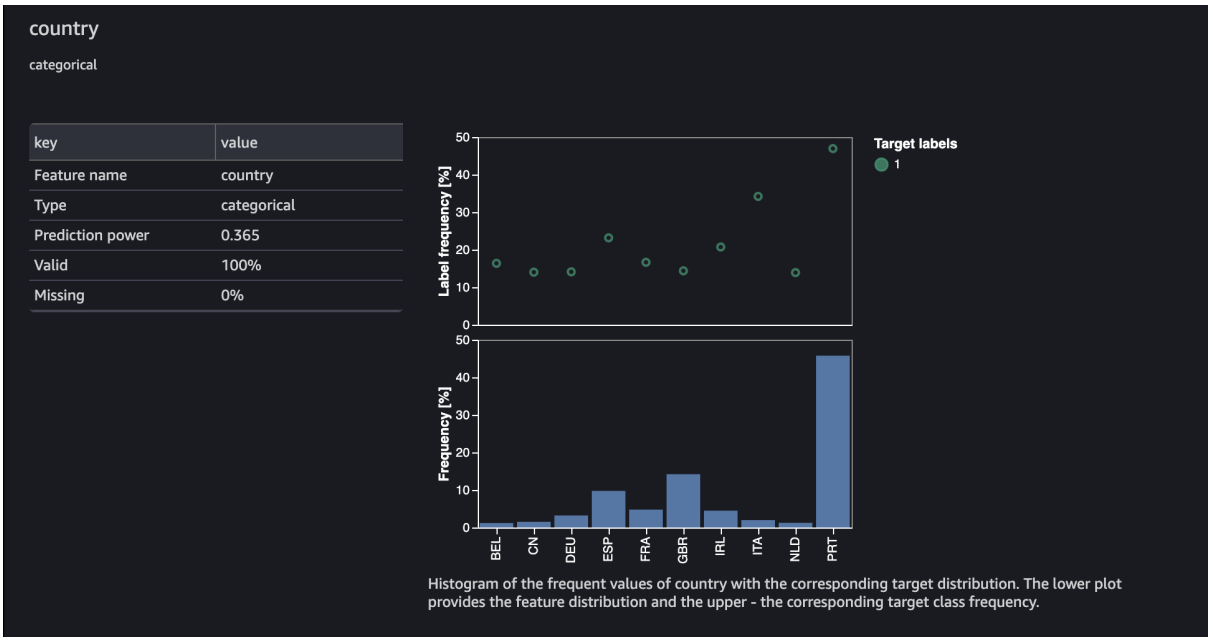
total_of_special_requestsNumber of special requests made by the customer (e.g. twin bed or high floor)

reservation_status Reservation last status, assuming one of three categories: Cancelled – booking was cancelled by the customer; Check-Out

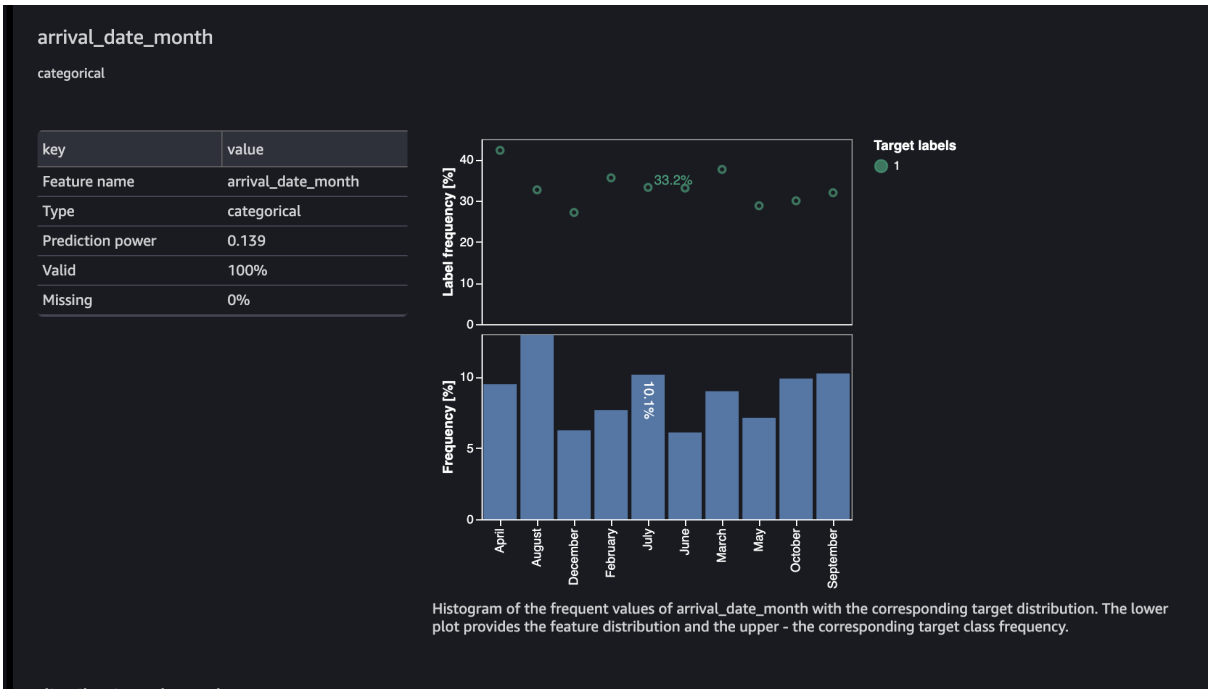
reservation_status_date Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to.

Visualisation and analysis

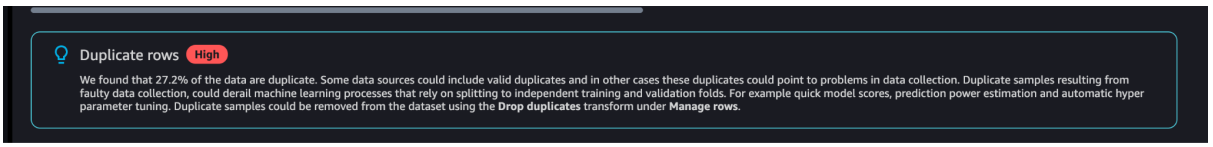
There are a lot of interesting visualisations about this dataset, notably are the following



The country with the most bookings is from Portugal.



Bookings soar in the summer season, which is expected.



There are a lot of duplicate rows.

Transformations

- Fill Missing
  - ☐ Fill all missing countries with PRT, because that's the mode.
  - ☐ Fill the rest(agent and company) with 0, because that's the mode.
- Remove bookings with no human beings
  The goal is to study the behaviour of bookings being cancelled, if a booking has no adults, children or babies, then it means no human beings are involved, making the booking invalid. **This was done with codes**.

```
filter = (df.children > 0) | (df.adults > 0) | (df.babies > 0)
df = df[filter]
```

- Remove Duplicate row
  A ton of the rows were duplicates which was highlighted in the visualisation, so they had to be removed. **This was done with codes.**

```
df = df.drop_duplicates()
```

- Remove Useless columns
  Some of the columns had no use, so they had to be removed. **This was done with codes.**

```
label =
['company','agent','total_of_special_requests','required_car_parkin
```

```
g_spaces','booking_changes',

'is_repeated_guest','reservation_status_date','stays_in_weekend_nig
hts','stays_in_week_nights',

'reserved_room_type','assigned_room_type','adults','children','babi
es']
df.drop(labels=label,axis=1,inplace=True)
```