Project

Project name – Hotel Booking Analysis

Name : Ankita Mishra

Email :ankitaamishra09@gmail.com

Please write a short summary of your project and its components. Describe the problem statement, your approaches and your conclusion.

For this hotel booking analysis, the goal was to explore the customer data of a hotel and identify any potential trends or correlations. The purpose of this exploratory data analysis (EDA) was to explore the hotel booking data set and identify potential relationships between key variables.

The data set included customer booking information. As part of the analysis, descriptive statistics were calculated for each variable, and visualizations were created to explore the relationships between various variables. To get insight from the dataset, we built a variety of charts, including a count plot, bar plot, kdeplot, heatmap, pairplot, violin plot, and boxplot.

The data set was composed of over 119390 hotel bookings, each containing several variables such as 'hotel', 'is_canceled', 'lead_time', 'arrival_date_year', 'arrival_date_month', 'arrival_date_week_number', 'arrival_date_day_of_month', 'stays_in_weekend_nights', 'stays_in_week_nights', 'adults', 'children', 'babies', 'meal', 'country', 'market_segment', 'distribution_channel', 'is_repeated_guest', 'previous_cancellations', 'previous_bookings_not_canceled', 'reserved_room_type', 'assigned_room_type', 'booking_changes', 'deposit_type', 'agent', 'company', 'days_in_waiting_list', 'customer_type', 'adr', 'required_car_parking_spaces', 'total_of_special_requests', 'reservation_status', and 'reservation_status_date'.

Dataset variables are in int64, float64, and object datatypes. There are 32 variables: 12 variables are objects, 16 are int64, and 4 are float64. 31994 duplicate values were removed. The variables country had 452, children had 4, agent had 12193, and company had 82137 null values. We replaced the null value with the mode of each variable (country, children, agent) for these variables, but the variable "company" had more than 50% null value, so we removed it. Further, we removed outliers from lead_time and adr. The final dataset had 87396 observations.

We also changed the data types of variables children, agent, and reservation_status_date to int64, int64, and datetime64, respectively. We performed some feature engineering for more convenience and created new variables: total_stays, total_people, total_childrens, reserved_room_assigned, guest_category, and lead_time_category. Now total_people and total_childrens are in the floated 64 datatype, so we converted them to int64. Now We were removed from the observation because having total_people at 0 made no sense.

After data cleaning, exploratory data analysis revealed several interesting findings as following :

- The top country with the most number of bookings is PRT, and the number one agent with the most number of bookings is 9.
- Customers favored city hotels more than resort hotels by a margin of 61.07 percent.
- One of the four reservations is canceled.
- The most popular food is BB.
- The Online (internet) platform is used to make the majority of bookings.
- The majority of the bookings are made using TA/TO, the leading distribution channel.
- The vast majority of hotel bookings are made by new guests. Almost no consumers (3.86%) returned.
- The customer wants Room A to be reserved the most.
- Customers do not wish to make a bookings with a pre-deposit.
- Customers (80%) favored making a hotel reservation for a short visit.
- Only 10% of people require space to park their cars.
- Most visitors are couples.
- The inability to assign a reserved room to a customer is not grounds for cancellation.

- Booking cancellations are not caused by a longer Lead time.
- A city hotel is busier than a resort.
- The busiest months for hotels are October and September. There isn't a lengthy wait for reservations in July.
- Not assigning a reserved room does not affect ADR.

We had some difficulties with the data when we were cleaning and analyzing it. There were a lot of duplicate values in the dataset. Null values were present in the dataset. Choosing the most effective visualization method is difficult. Performing feature engineering was more challenging.