EDA on Hotel Booking Analysis

Kishor Shivaji Patil

Data science trainee, AlmaBetter , Bangalore.

1) **ABSTRACT :** This hotel booking dataset contains booking information for city and resort hotel. Both datasets share the same structure, with 32 variables describing the 40,060 observations of H1 and 79,330 observations of H2. Each observation represents a hotel booking. Both datasets comprehend bookings between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were cancelled. Since this is hotel real data, all data elements pertaining hotel or costumer identification were deleted. Due to the scarcity of real business data for scientific and educational purposes, these datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

2) **INTRODUCTION :** In tourism and travel related industries, most of the research on Revenue Management demand forecasting and prediction problems employ data from the aviation industry, in the format known as the Passenger Name Record (PNR). This is a format developed by the aviation industry. However, the remaining tourism and travel industries like hospitality, cruising, theme parks, etc., have different requirements and particularities that cannot be fully explored without industry's specific data. Hence, two hotel datasets with demand data are shared to help in overcoming this limitation. The datasets now made available were collected aiming at the development of prediction models to classify a hotel booking's likelihood to be cancelled. Nevertheless, due to the characteristics of the variables

included in these datasets, their use goes beyond this cancellation prediction problem. One of the most important properties in data.

3) PROBLEM STATEMENT :

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests? This hotel booking dataset can help you explore those questions ! This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has been removed from the data. Explore and analyze the data to discover important factors that govern the bookings.

4) FEATURE DESCRIPTION :

for prediction models is not to promote leakage of future information. In order to prevent this from happening.

1) **hotel** : Hotel(Resort Hotel or City Hotel)
2) **is_canceled** : Value indicating if the booking was canceled (1) or not ()
3) **lead_time** : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
4) **arrival_date_year** : Year of arrival date
5) **arrival_date_month** : Month of arrival date
6) **arrival_date_week_number** : Week number of year for arrival date
7) **arrival_date_day_of_month** : Day of arrival date
8) **stays_in_weekend_nights** : Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
9) **stays_in_week_nights** : Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel.
10) **adults** : Number of adults
11) **children** : Number of children
12) **babies** : Number of babies
13) **meal** : Type of meal booked. Categories are presented in

standard hospitality meal packages

14) **country** : Country of origin

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

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

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

18) **previous_cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking

19) **previous_bookings_not_canceled** : Number of previous bookings not cancelled by the customer prior to the current booking

20) **reserved_room_type** : Code of room type reserved. Code is presented instead of designation for anonymity reasons

21) **assigned_room_type** : Code for the type of room assigned to the booking

22) **booking_changes** : Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation

23) **deposit_type** : Indication on if the customer made a deposit to guarantee the booking

24) **agent** : ID of the travel agency that made the booking

25) **company** : ID of the company/entity that made the booking or responsible for paying the booking.

26) **days_in_waiting_list** : Number of days the booking was in the waiting list before it was confirmed to the customer

27) **customer_type** : Type of booking, assuming one of four categories

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

29) **required_car_parking_spaces** : Number of car parking spaces required by the customer

30) **total_of_special_requests** : Number of special requests made by the customer (e.g. twin bed or high floor)

31) **reservation_status** : Reservation last status, assuming one of three categories • Canceled – booking was canceled by the customer • Check-Out – customer has checked in but already departed • No-Show – customer did not check-in and did inform the hotel of the reason why

32) **reservation_status_date** : Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel.

5) EXPLORATORY DATA ANALYSIS :

- **DATA PREPARATION :** Firstly, we imported libraries and dataset, some of the libraries used are NumPy, pandas, matplotlib, seaborn, Once the data is collected, process of analysis begins. But data has to be translated in an appropriate form. This process is known as Data Preparation like Validate data, Clean the dataset, Checking and deleting the duplicate values, Statically adjust the data, Store the dataset for analysis & Analyze the data.

- **MISSING VALUES AND OUTLIER TREATMENT :** There are different ways and methods of identifying outliers, but we are only going to use some of the most popular techniques:

Visualization: by boxplot or histogram plot

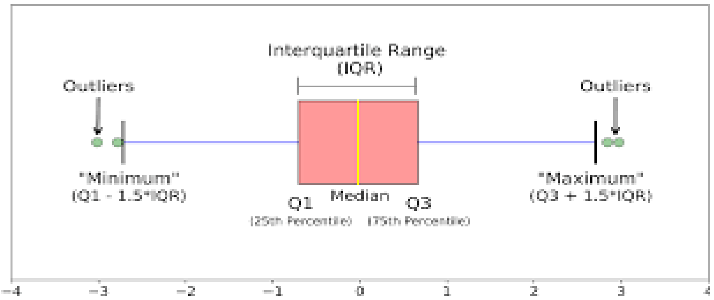
• **Skewness:** The skewness value should be within the range of -1 to 1 for a normal distribution, any major changes from this value may indicate the presence of outliers.

• **Interquartile Range:** IQR
• Standard Deviation: It shows the variability distribution of the data**.**

• **Flooring or capping**

• **Trimming**

Firstly, we demonstrate and remove the outlier based upon own understanding by setting up the threshold limit. And in terms of outlier, we used IQR Technique, In descriptive statistics, the interquartile range (IQR) is a measure of statistical dispersion, which is the spread of the data. The IQR may also be called the mid spread, middle 50%, or Hspread. It is defined as the difference between the 75th and 25th percentiles of the data. And lastly, we used quantile-based technique to treat the outlier, Capping is replacing all higher side values exceeding a certain theoretical maximum or upper control limit (UCL) by the UCL value



DATA PREPROCESSING: A dataset may contain noise, missing values, and inconsistent data, thus, preprocessing of data is essential to improve the quality of data and time required in the data mining.

CLEANING: After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

DATA MANIPULATION: Manipulation of data is the process of manipulating or changing information to make it more organized and readable. Made some new features with the help of column present in the datasets.

UNIVARIATE ANALYSIS: In Univariate Analysis, we choose a single feature from the data and try to determine what the output or the target value i.e.one feature/variable at a time.

- Understand the trends and patterns of data

- Analyze the frequency and other such characteristics of data

- Know the distribution of the variables in the data.

- Visualize the relationship that may exist between different variables.

BIVARIATE ANALYSIS: In a Bivariate Analysis, we try to analyze two features instead of one, and finally determine the classification of output we are looking for. It is a methodical statistical technique applied to a pair of variables (features/ attributes) of data to determine the empirical relationship between them. In order words, it is meant to determine any concurrent relations. There are three main types of bivariate analysis. They are as follows: • Scatter Plots - It makes use of dots to represent the values for two different numeric variables.

• **Regression Analysis-** This involves a wide range of tools that can be utilized to determine just how the data points might be related. It tends to provide us with an equation for the curve/line along with giving us the correlation coefficient.

• **Correlation Coefficients** - This shows how one particular variable moves about with relation to another.

MULTIVARIATE ANALYSIS:

• Multivariate analysis deals with such complex set of data with more than two feature and variables. There are two types of multivariate analysis techniques: Dependence techniques, which look at cause-and-effect relationships between variables, and interdependence techniques, which explore the structure of a dataset.

6) CHALLENGES :

● The name of the countries was not in the proper format,because of which we are not able to plot the geomap plot.

● Company and agent column has lots of duplicate value.

● There were many rows with almost similar data.

● Lots of null values in the dataset.

7) CONCLUSION :

● Month of August and july receives most no. of booking.

● Booking for city hotels is twice as for resort hotels.

● Repeated costumers cancel their hotel in very rare cases.

● Customers coming from aviation industry has very less time i.e. they book urgently.

● People with no kid prefer to choose city hotel over resort hotel.