HOTEL BOOKING ANALYSIS

Shubham Chougule : sc9395288@gmail.com

Data Science Trainee, Almabetter

1. **ABSTRACT:** This hotel booking dataset contains booking information for city and resort hotel. Both datasets share the same structure, with 31 variables describing the 53,428 observations of City Hotel and 33,968 observations of Resort Hotel. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive 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, aswell as in other fields.

2. **INTRODUCTION:** In tourism and travel related industries, most of the research on Revenue Management demand forecasting and prediction problems employee 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 for prediction models is not to promote leakage of future information. In order to prevent this from happening, the timestamp of the target variable must occur after the input variables timestamp. Thus, instead of directly extracting variables from the bookings database table, when available, the variables values were extracted from the bookings change log, with a timestamp relative to the day prior.

PROBLEM STATEMENT: We are here to explore a hotel booking dataset to discover important factors that govern the bookings, which contain booking information for a city hotel and a resort hotel. Wewill analyze some important aspects of hotel bookings which will helps us identify major loopholes and give us insights which will be helpful to run profitable hotel business are as follows:

a. The time of year to book a hotel room?
b. Optimal length of stay to get the best daily rate?
c. To predict whether or not a hotel was likely to receive a disproportionately high number of special requests?

FEATURE DESCRIPTION: The data feature in this dataset respectively:

- ADR (Numeric) Average Daily Rate as defined.
  d. Adults (Integer) Number of adults.
  e. Agent (Categorical) ID of the travel agency that made the booking.

3. arrival_date_day_of_month (Integer) : Day of the month ofthe arrival date.
4. arrival_date_month (Categorical) : Month of arrival date with 12 categories:"January" to "December".
5. arrival_date_week_number (Integer) : Week number of year for arrival date.
6. arrival_date_year (Integer) : Year of arrival date.
7. Babies (Integer) : number of babies in count.
8. Children (Integer) : number of children.
9. Company (Integer) : ID of the company/entity that made the booking or responsible forpaying the booking. ID is presented instead of designation for anonymity reasons.
10. customer_type (categorical) : Type of booking, assuming one of four categories.
11. Transient : when the booking is not part of a group or contract.
12. Transient_party : when the booking is transient,but is associated to at least other transient booking.
13. distribution_channel (categorical) : Booking distribution channel. The term "TA" means"Travel Agents" and "TO" means "Tour Operators".
14.days_in_waiting_list (Integers) : Number of days the booking was in the waiting list before it was confirmed to the customer.
15. Hotel (categorical) : Hotel (Resort Hotel or CityHotel).

16. is_canceled (Integer) : Value indicating if the booking was canceled (1) or not (0).

17. is_repeated_guest (Integer) : Value indicating if the bookingname was from a repeated guest (1) or not (0).

18. lead_time (Integer) : Number of days that elapsed between the entering date of the booking into the PMS and thearrival date.

19. Meal (categorical) : Type of meal booked. Categories are presented in standardhospitality

20. Meal_packages : Undefined/SC – no meal.

21. previous_cancellations (categorical) : Number of previous bookings that were cancelled by the customer prior to the current booking.

22.previous_bookings_not_cancelled (Integer) : Number of previous bookings not cancelled by the customer prior to the current booking.

23.reservation_status(categorical) : Reservation last status,assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out.

24. reservation_status_date (Date) : Date at which the laststatus was set. This variable can be used in conjunction with the Reservation Status to under stand when was the booking canceled or when did the customer checked-out of the hotel.

25. stays_in_weekend_nights (Integer) : Number of weekendnights (Saturday or Sunday)the guest stayed or booked to stay at the hotel.

26. total_of_special_requests (Integer) : Number of specialrequests made by the customer (e.g., twin bed orhigh floor).

1. EXPLORATORY DATA ANALYSIS:

- **DATA PREPARATION:** Firstly, we imported libraries and dataset, some of the libraries used are NumPy, pandas, matplotlib, seaborn, warnings. 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.
- **Validate data.**
- **Clean the data set.**
- **Checking and deleting the duplicate values.**
- **Statically adjust the data.**
- **Store the data set for analysis.**

- **Analyze the date.**

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

- **CLEANING AND MANIPULATING THE DATASET:**

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. Irregularities are of different types of data.

- Missing Values.
- Incorrect Format.
- Incorrect Headers.

MANIPULATING:

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 .

CHALLENGES:

- Dealing with such big dataset is quite difficult some times , lots of missing values made things some more complicated , defining a function which is used to annotate the histogram percent according to their respective count taken a big notch of this obstacle part.
  Coming to the visualization part , more or less makes our challenges addresses to code in such a way to visualize the graphs as per rows and columns with fixed figure size to retain as per the sub plots.

CONCLUSION:

Our analysis, would be capable of helping prospective guests in choosing the right hotel, right stay duration and much more for their stay and moreover, would also be introspecting for hotel management in bringing out changes in theirservices for the guests.

- City hotels are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
- Most of the bookings for City hotels and Resort hotel were happened in 2016.
- Resort hotels have the most repeated guests.
- October month has most of the bookings.
- The most preferred Room type by customer is "A","D","E", etc.
- 'Online T/A' has the highest cancellation in both type of Hotels. In order to reduce the booking cancellations, hotels need to set the refundable/non-refundable and deposit policies.
- Maximum number of guests were from Portugal, i.e. more than 25000 guests.
- Transient customer type is more which is 82.4 %. percentage of Booking associated by the Group is vey low.
- Among all the bookings made by customer. 27.5 % of the bookings were cancelled.
- Booking cancellation rate is high for City hotels which almost 30 %. Average lead time for resort hotel is high.
- From the total bookings there are 98.7 % of the guests prefer "No deposit" type of deposit.
- Most of the customers (91.6%) do not require car parking spaces.
- BB(Bed & Breakfast) is the most preferred type of meal by the guests.
- Optimal stay in both the type hotel is less than 7 days. Usually people stay for a week.
- Almost 72.5 % people did not cancel their bookings even after not getting the same room which they reserved while booking hotel.

Average ADR for city hotel is high as compared to resort hotels. These City

hotels are generating more revenue than the resort hotels. As for the prediction of cancellations concerns. it is clear that better results can be achieved that includes more models into consideration. Besides, this data is some what limited (only two years). A wider time window and more features, which sure will be at the hands of every hotelier in the business, better results could be obtained.