Hotel Booking Analysis

Midhun R
Data Science Trainee,
AlmaBetter, Bangalore

Abstract:

A customer's expectation of a hotel experience has grown over the years and hotel managements need to keep up with it. With AI and data science, hotels have advanced tools to evaluate and improve performance.

In this project, I have attempted to analyze a hotel booking dataset and come up with some relevant conclusions about the factors that contribute to count of bookings. No personal information of customer is provided in this dataset.

1. Problem Statement

A dataset containing 119390 records across 32 features has been given with information regarding bookings of two hotels from July 2015 to August 2017. These two hotels are City Hotel and Resort Hotel.

The main objective is to explore the given dataset and discover the factors which govern the bookings. The dataset will be analyzed and from the conclusions drawn from it will be used to recognize the missteps taken by the manager. With this information, hotels will be equipped to improve their performance.

Data analysis is performed to answer the following questions:

- Which hotel is more preferred among travelers?
- Which hotel retains more customers?
- Which is the busiest month?
- Which is the most popular room type?
- From which country the greatest number of bookings were made?
- How Long People Stay in the hotel?
- How many bookings were cancelled?

1. Introduction

Travelling is innate to humans. People wish to connect, explore and travel more than ever before. In this day and age, when data science has transformed into an invisible force that influences the decisions of the modern world, hospitality industry has also become beneficiary of data science.

Nowadays, having a proper understanding of data and connecting all data sources effectively is paramount in generating competitive advantage, providing superior customer value, and ultimately orientating the future of any business. Over the last few years, hospitality companies have begun to deploy predictive analytics to better anticipate and meet customer needs and preferences.

The goal is to inspect, clean and implement exploratory data analysis on the given dataset to derive some conclusions which is later used to find out which factors affect the bookings in hotels and how they will affect it.

2. Steps Involved:

- **Importing Libraries**
  Relevant libraries like NumPy for numerical operations, Pandas for data manipulation, matplotlib and seaborn for data visualization were loaded. In addition to these libraries, pycountry library was installed and loaded, which was used for converting country codes to names.

- **Reading Data**
  After drive was mounted, data from csv file was read and store in a pandas dataframe.

- **Data Inspection**
  After loading the dataset and importing relevant libraries, the dataset has been explored by thoroughly by looking into its head, tail, brief summary, number of records and features, etc. The dataframe contains 119390 rows of data, out of which 31,994 rows are duplicate rows, which must be removed later. The dataframe contains 32 columns. Four columns have missing values. Some columns require conversion of datatypes. Additionally, new columns need to be added from existing ones to make analysis easier. Then it was checked for any duplicated rows or null values. Unique values in each feature were also obtained. The features in the dataset were identified as:

1. hotel: Name of the hotel (Resort Hotel or City Hotel).
2. is_canceled: If the booking was canceled (1) or not (0).
3. lead_time: Number of days before the actual arrival of the guests.
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 month arrival date.
8. stays_in_weekend_nights: Number of weekend nights (Saturday or Sunday) spent at the hotel by the guests.
9. stays_in_week_nights: Number of weeknights (Monday to Friday) spent at the hotel by the guests.
10. adults: Number of adults among guests.
11. children: Number of children among guests.
12. babies: Number of babies among guests.
13. meal: Type of meal booked.
14. country: Country of guests.
15. market_segment: Designation of market segment.
16. distribution_channel: Name of booking distribution channel.
17. is_repeated_guest: If the booking 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.
21. assigned_room_type: Code of room type assigned.
22. booking_changes: Number of changes/amendments made to the booking.
23. deposit_type: Type of the deposit made by the guest.
24. agent: ID of travel agent who made the booking.
25. company: ID of the company that made the booking.
26. days_in_waiting_list: Number of days the booking was in the waiting list.
27. customer_type: Type of customer, 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.
31. reservation_status: Reservation status (Canceled, Check-Out or No-Show).
32. reservation_status_date: Date at which the last reservation status was updated.

● Data Cleaning

Data cleaning is done to ensure that the dataset is correct, consistent, and usable. It improves the efficiency and quality of analysis. Data cleaning was done in 4 steps. The first step was removing 31994 duplicate rows.

In the next step, missing values were handled. Four out of 32 columns have missing values in them. The column 'company' is dropped altogether since the number of missing values is extremely high compared to the number of rows. The number of missing values in column 'agent' is low compared to the number of rows. So, the missing values are filled with mode of the column 'agent'. The case is same with country. Missing values are filled with a constant string 'Others'. The number of missing values in column ‘children’ is negligible compared to the number of rows, so it doesn't affect the result of the analysis in a big way. The missing values with an integer constant 0.

In the third step, some columns were converted to appropriate datatypes to make analysis easier and more accurate. These columns were ‘children’, ‘agent’ (int64) and ‘reservation_status_date’ (date). In the fourth step, some extra columns were added which will be useful during analysis. These are ‘total_stays_in_nights’, ‘total_guests’ and ‘is_reserved_room_type_assigned’.

Boolean data of ‘is_canceled’ and ‘is_repeated_guest’ is also converted to string for easy representation.

● Exploratory Data Analysis

Dataset, after cleaning, is subjected to exploratory data analysis, which will visualize the data and identify trends and patterns that can be later used to increase revenue.

EDA was carried out in 3 steps:

1. **Univariate Analysis**: Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately.

   Univariate analyses were done on:

   - Percentage of bookings in each hotel.
   - Percentage of repeated and non-repeated guests.
   - Percentage of bookings that got cancelled.
   - Number of bookings made for each room type.
   - Number of bookings made from each country.
   - For how long guests commonly stay in the hotel.
   - Number of bookings made in each month.

   Charts used for univariate analyses are pie chart, horizontal bar chart and vertical bar chart.

2. **Bivariate Analysis**: Bi means two and variate means variable, so here there are two variables. The analysis is related to cause and the relationship between the two variables.

   Bivariate analyses were done on:

   - Revenue generated by each hotel.
   - Percentage of repeated guests in each hotel.
   - Percentage of repeated guests in each distribution channel.
   - Percentage of cancelled and non-cancelled bookings in each hotel.
   - Number of cancelled and non-cancelled bookings among repeated and non-repeated guests.
   - Kernel density estimate of number of days in waiting list for cancelled and non-cancelled bookings.
   - Kernel density estimate of lead time for cancelled and non-cancelled bookings.
   - Number of bookings cancelled when reserved room type is the same and different as the assigned room type.
   - Percentage of cancelled and non-cancelled bookings in each distribution channel.
   - Change in the length of stay with the change in ADR.

   Charts used for univariate analyses are scatter plot, density plot, stacked bar chart and vertical bar chart.

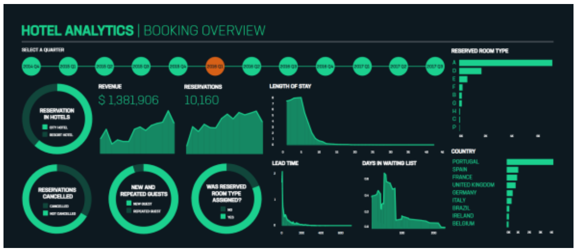
3. **Correlation Analysis**: It is used to measure the strength of the linear relationship between two

variables and compute their association. Correlation analysis calculates the level of change in one variable due to the change in the other.

Correlation analysis of the dataset was carried out using a correlation heatmap with the features, 'lead_time', 'adr', 'total_guests', 'total_stays_in_nights', 'previous_cancellations', 'booking_changes', 'days_in_waiting_list', 'required_car_parking_spaces', 'total_of_special_requests' and 'previous_bookings_not_canceled'.

4. Data Visualization

An interactive dashboard was also created with Tableau to display charts associated with the analysis.



Click here to interact with the data visualization.

5. Challenges:

Handling 119390 rows and 32 columns of data was a bit difficult as a beginner. The inspection and cleaning of dataset was a time-consuming process. Visualization of data was properly carried out after providing a great amount of attention to each and every details. In one case outliers had to be removed to get a proper visualization. Accurate visualization was achieved only after a lot of trial and errors.

If code blocks written to generate visualizations weren't modularized, it would have taken more effort.

6. Conclusion:

The following conclusions were drawn from analysis:

- City Hotel seems to be more preferred among travelers and it also generates more revenue.
- Most number of bookings are made in July and August.
- Room Type A is the most preferred room type among travelers.
- Most number of bookings are made from Portugal.
- Most of the guest stays for 1-4 days in the hotels.
- Resort Hotel retains a greater percentage of guests.
- Around one-fourth of the total bookings gets cancelled. More cancellations are from City Hotel.
- New guests tend to cancel bookings more than repeated customers.
- Lead time, number of days in waiting list or assignation of reserved room to customer does not affect cancellation of bookings.
- Corporate has the most percentage of repeated guests while TA/TO has the least whereas in the case of cancelled bookings TA/TO has the most percentage while Corporate has the least.
- The length of the stay decreases as ADR increases probably to reduce the cost.

- Visits planned with longer stays are booked earlier than those planned with shorter stay.

References-

1. Kaggle
2. GeeksforGeeks
3. Analytics Vidhya