Hotel Booking Data Analysis

Samim Ali, Mohd. Izhar, Sarath Haridas
Data science trainees,
AlmaBetter, Bangalore

1. Problem Statement

For this project, we will be analyzing Hotel Booking data. This data set contains booking information for city hotels and resort hotels. It 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.

The hotel industry is a very volatile industry and the bookings depend on the above factors and many more. The main objective behind this project is to explore and analyze data

To discover important factors that govern the bookings and give insights to hotel management, which can perform various campaigns to boost the business and performance.

2. Introduction

This data article describes two datasets with hotel demand data. One of the hotels is a Resort hotel and the other is a City hotel. Each observation represents a hotel booking. 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.

3. Workflow

We will divide our workflow into 3 Following steps:

- Data Collection and Understanding.
- Data Cleaning and Manipulation.
- Exploratory Data Analysis (EDA)

EDA will be divided into the following 3 analyses**:**

- **Univariate analysis:** Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable.

- **Bivariate analysis:** Bivariate analysis is where you are comparing two variables to study their relationships.

- **Multivariate analysis:** Multivariate analysis is similar to Bivariate but you are comparing more than two variables.

4. Data collection and understanding

Data was obtained from Almabetter portal and stored as a CSV file in Google drive. This was later read in order to obtain the data.

5. Data Cleaning and Manipulation

- **Null values Treatment**

  Our dataset contains a large number of null values which might tend to disturb our Analysis, hence we dropped them at the beginning of our project in order to get a better result.

  There were 4 Columns that contain null values in our Data set: Country, agent, Company, and children
  Hence we drop all the null values from the columns which are mentioned above.

- **Handling Duplicates**

  Our Data had 31994 Duplicate values, so we dropped it from the Data.

- **Feature Engineering**

  We created 2 new Columns:
  1. 'Total_People' = From Children, Adults and babies.
  2. 'Total_Stay' = From Weekend nights and Weekdays nights

6. Exploratory Data analysis

Exploratory data analysis was then performed on the clean data set to obtain certain observations

- **Univariate analysis**

  1. Type of Hotels which the guests mostly prefer.
  2. Which agents made the highest booking
  3. Percentage of repeated guests in each hotel
  4. Percentage of cancellation
  5. Type of Food which the guests mostly prefer.
  6. Type of Rooms mostly preferred by the guests.
  7. From which country most of the guests are coming
  8. In which month most of the booking happened.
  9. Percentage of booking change made by the customers.
  10. Percentage distribution of customer types**.**

- **Multivariate Bivariate analysis**

1. Which hotel has the highest ADR
2. Which hotel type has the more Lead time
3. ADR across different Months
4. Which hotels has the longer waiting time
5. Which hotel has the most repeated guests
6. Bookings analysis according to Months and years
7. Which Distribution channel contributed more to ADR in order to increase the income
8. Which distribution channel has the highest cancellation rate
9. Which market segments has the highest cancellation rate
10. ADR Relationship with total number of people

11. Relationship between ADR and total stay

7. Correlation of all the Columns

- is_canceled and same_room_alloted_or_not are negatively correlated. That means a customer is unlikely to cancel his bookings if he doesn't get the same room as per reserved room. We have visualized it above.
- lead_time and total_stay is positively correlated.That means more is the stay of customers more will be the lead time.
- Adults, children and babies are correlated to each other. That means more people will be ADR.
- is_repeated guest and previous bookings not canceled has strong correlation. may be repeated guests are not more likely to cancel their bookings.

8. Conclusion

- City hotels are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
- 27.5 % bookings were cancelled out of all the bookings.
- Only 3.9 % of people were revisited the hotels. Rest 96.1 % were new guests. Thus retention rate is low.
- The percentage of 0 changes made in the booking was more than 82 %. Percentage of Single changes made was about 10%.
- BB (Bed & Breakfast) is the most preferred type of meal by the guests.
- Maximum number of guests were from Portugal, i.e. more than 25000 guests.
- Average ADR for city hotels is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
- Booking cancellation rate is high for City hotels which is almost 30 %.
- Average lead time for resort hotels is high.
- Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.
- Resort hotels have the most repeated guests.
- July and August had the most Bookings because of Summer vacations.
- '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.

References

- Stackoverflow.com
- Geeksforgeeks
- Kaggle.com

THANK YOU!