EDA CAPSTONE PROJECT- HOTEL BOOKING

HOTEL BOOKING ANALYSIS

Abhishek Kumar,
Mohita Rathour,
Mukesh Sablani,
Sanjay Paul.

Index: -

I. Abstract

The success factoring a profitable hotel industry has been changing over time, driven by global competition and increasingly high customer expectations. Hotels focus on customer satisfaction and to exceed customer expectations.

We have a hotel booking dataset containing information for city and resort hotels. This dataset has 32 variables with around 1,19,000 entries. The study has data recorded between 2015 to 2017 which have bookings that shows effectively arrived and bookings that were canceled.

We have a hotel booking dataset. We are using our Python skills to perform EDA and gain informative insights about factors in hotel bookings and how they affect hotel bookings.

II. Introduction :-

In the hotel industry, focus is on revenue and research is on forecasting demand and predicting the needs of the customers.

This dataset has a collection of two types of hotels. The resort hotel and the city hotel. It is collected in order to predict hotel bookings and its probability of cancellation. Some attributes here are to understand what factors do bring in the revenue for the business. Some attributes show the customers preference for booking whereas some attributes show the factors leading to cancellations.

III. Project Goal :-

Purpose of our study is to find the best time of year to book a hotel room. The optimal length of stay in order to get the best daily rate. Study on special requests.

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.

IV. Attributes:-

We have a sample hotel bookings dataset. Here are its attributes:

❖ hotel: Indicates types of Hotels ● City hotel ● Resort type
❖ is_canceled: Indicates the cancellation of the hotel booking ● Cancellation = 1 ● No Cancellation = 0
❖ lead_time: Time (in days) between booking transaction and actual arrival.
❖ arrival_date_year: Year of arrival

- ❖ arrival_date_month: Month of arrival
- ❖ arrival_date_week_number: week number of arrival date.
- ❖ arrival_date_day_of_month: Day of month of arrival date
- ❖ stays_in_weekend_nights: No. of weekend nights stayed in a hotel
- ❖ stays_in_week_nights: No. of weeknights stayed in a hotel
- ❖ adults: No. of adults in a single booking record.
- ❖ children: No. of children in a single booking record.
- ❖ babies: No. of babies in single booking record.
- ❖ meal: Type of meal chosen
    - BB:- bed and breakfast
    - HB:-Half board (Breakfast and dinner)
    - FB:- Full Board (All meals included)
    - SC:- Self catering (No meals Included)
- ❖ country: Country of origin. Categories are represented in the ISO 3155–3:2013 format.
- ❖ market_segment: market segment for booking
    - Aviation
    - Complimentary
    - Corporate
    - Direct
    - Groups
    - Online (TA)
    - Offline (TA/TO)
- ❖ distribution_channel: Via which medium booking
    - Corporate
    - Direct
    - GDS: - Global Distribution System
    - TA/TO: - Travel Agent/Operator
- ❖ is_repeated_guest:
    - 0 for new customer

- 1 for repeated customer

❖ previous_cancellations: No. of previous canceled bookings.

❖ previous_bookings_not_canceled: No. of previous non-canceled bookings.

❖ reserved_room_type: Room type reserved by a customer.

❖ assigned_room_type: Room type assigned to the customer.

❖ booking_changes: No. of booking changes done by customers

❖ deposit_type: Type of deposit at the time of making a booking
- No deposit
- Refundable
- No refund

❖ agent: Id of agent for booking

❖ company: Id of the company making a booking

❖ days_in_waiting_list: No. of days in waiting to book

❖ customer_type: Type of customer
- Contract: - bookings done by the contract
- Group: - Group booking
- Transient: - Customer staying for shorter period
- Transient-Party: - Group of customers staying for a shorter period

❖ adr: Average Daily rate of hotels.

❖ required_car_parking_spaces: No. of car parking preferred by customers at the time of booking

❖ total_of_special_requests: total no. of special request.

❖ reservation_status:
- checked out
- canceled
- not showed

❖ reservation_status_date: Date of making reservation status.

IV. Exploratory Data Analysis:-

First step is to import libraries such as NumPy, pandas, matplotlib, seaborn.Then load the raw dataset. This data has many unprocessed values which cannot be considered for the study. Here is the workflow of correcting it for our analysis.

1. Data Cleaning

a. Handling Null Values

- **Company Id and Agent Id: -** These columns have null values of 93% and 15% respectively. Hence, these columns are dropped.
- **Country:** - This has null values less than 5% thus the null values are filled with the mode value.
- **Children and babies:** - There are only 4 null values so the null value is filled with mean

b. Handling Outliers

An outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data graph or dataset we work with.

We have used the Interquartile range method to handle outliers. To find the interquartile range (IQR), we first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

2. Data Manipulation: - Creating new columns

- **Kids=** Children +babies
- **Total stay=** stays_in_weekend_nights+ stays_in_week_nights
- **Guest=** Adults+kids
- **Revenue=** stay of non-cancelled guests * ADR

➔ Data study

i) UNIVARIATE ANALYSIS:

a) Univariate analysis is the simplest form of analyzing data i.e study of one variable. Its major purpose is to describe; distribution of single data, and find patterns in the data.

ii) BIVARIATE ANALYSIS:

b) Bivariate analysis between two variables. One of the variables will be dependent and the other is independent. The study is analyzed

between the two variables to understand to what extent the change has occurred.

iii) MULTIVARIATE ANALYSIS

c) Multivariate data analysis is the study of relationships among the attributes, classify the collected samples into homogeneous groups, and make inferences about the underlying populations from the sample.

V. Data Visualization :-

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make it easier to understand and gain insights from them.
The graphs used here for study are: -

- Box Plot.
- Histogram.
- Pie Chart.
- Bar Plot.
- Line Plot.
- Scatter Plot.
- Geo Mapping.

VI. Conclusion :-

Now after our in-depth analysis of the data we have come up with some insights which help us to understand this industry better and also to focus on factors which help to improve the KPI.

- City Hotels are the most preferred hotel by guests.
- Even though the booking made in City hotel are greater than the resort hotel, almost double though the revenue by city hotel is less,.
- This shows that Resort hotels are bit expensive and receive less cancellation than City hotels
- Out of total no of reservations 63% actually showed up, 36% got canceled, and only 1% reservation got No-show.
- It seems that 2016 is the year where the hotel bookings are highest. So the cancellation % is also the highest this year. Every year there is 25-30% cancellation received for resort hotels and 40-45% cancellation received for city hotels

- From the above graph we can see in city hotels there is a peak from April to July and the booking is high in August. And in resort hotels we can see two peaks, first in June and second in September and booking is high in July,August and October, so people usually book hotels 30-60 days in advance.
- For Resort hotels-- ADR is increasing between May to September and then starts falling down, so the best time to book a resort hotel is from October to April as we are getting lower ADR.
- For City hotels--City hotels have nearly constant ADR from April to October and after that ADR starts decreasing, so the best time to book a City hotel is from November to March.

- Resort hotels and City hotels both are getting higher revenue between June to September. This is also because at the same time ADR is also high for both types of hotel as shown in the previous slide. Hence this period is best for hotels to generate more revenue.
- For Resort hotels-- ADR is increasing between April to September and then starts falling down, so the best time to book a resort hotel is from October to March as we are getting lower ADR.
- For City hotels--City hotels have nearly constant ADR from April to October and after that ADR starts decreasing, so the best time to book a City hotel is from November to March.

- Resort hotels and City hotels both are getting higher revenue between June to September. This is also because at the same the ADR is also high for both types of hotel as shown in the previous slide. Hence this period is best for hotels to generate more revenue.
- Here we can see that as lead time increases the ADR decreases. This means if a customer book a hotel in advance, he can get a better deal.
- Most bookings are done by transient customer types.
- Majority of people prefer room type-A. It seems to be more economical for booking as it has the least ADR.
- Most bookings are done by Transient customer type.
- Majority of the bookings and cancellations are made through Travel agencies (Online/Offline) and Tour Operators.
- Cancellation is more in City hotels as compared to Resort hotels.
- Chances of cancellation is high when there are no deposits taken by hotels. So minimum deposits should be taken by hotels to decrease the rate of cancellation.
- As length of total stay increases, adr decreases. This means for longer stay, the better deal for customers can be finalized.
- 77% of the people prefer the BB (bed & breakfast) meal type in both the hotel types.
- Maximum bookings and revenue are generated from Portugal.

- About 94% of people don't require the car parking spaces while booking hotel

- Mostly the guests are new customers and very small share for repeated customers

VII. Challenges: -

a. The amount of data collected: - There are some unnecessary raw data collected which do not contribute much to the study. Thus, to identify them and eliminate them will be a challenge
b. Handling Null Values and outliers: - Identifying Null values and outliers and handling them. Handling them is different in different cases. So we need to analyze it and to process.
c. Processing raw data into meaningful data: - Sometimes the columns cant be used as to understand how apply functions on these columns to get more relevant results
d. Visual representation: - One size doesn't fit all. Thus, finding exact graphs to represent the data is challenging.

THANK YOU!!!