**Hotel Booking Analysis**

**Rohith V, Abhishek V L**

**Nikita Hubballi, Aditi Sharma**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, seasonality, days of week and many more. This makes analyzing the patterns available in the past data more important to help the hotels plan well. Using the historical data, hotels can perform various campaigns to boost the business. We can use the patterns to predict the future bookings using time series or decision trees.

We will be using the data available to analyze the factors affecting the hotel bookings. These factors can be used for reporting the trends and predict the future bookings.

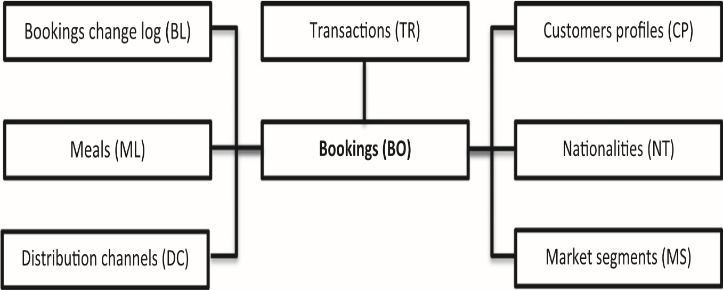
**1. Problem Statement**

The dataset represents the data of City hotels and Resort Hotels. The dataset have 32 variables with 119,390 observations. Each observation represents a hotel booking. The 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 canceled. Since this is hotel real data, all data elements pertaining hotel or customer identification were removed.

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.

Our main objective is perform EDA on the given dataset and draw useful conclusions about general trends in hotel bookings and how factors governing hotel bookings interact with each other.

1. In which of the hotel the highest booking was made
2. In which type of hotel people prefer stay in terms of days
3. Cancelation Done Based on months
4. High Cancelation rate in hotels
5. Arrival Seasonality based on years, months and days
6. Check whether the guest is repeated or not
7. Mode of bookings through market segment and distribution channel
8. When was the majority stays made?
9. What type of meal preferred by the customer in each hotel
10. In which country the highest booking was made
11. What kind of visitors travelled more?
12. Which was the most preferred deposit type made
13. Which kind of rooms were assigned for different adult size
    1. **Data Set**

* hotel: Name of hotel (City or Resort)
* is\_canceled: Whether the booking is canceled or not (0 for no canceled and 1 for canceled)
* 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 spent in a hotel
* stays\_in\_week\_nights: No. of weeknights spent in a hotel
* adults: No. of adults in single booking record.
* children: No. of children in single booking record.
* babies: No. of babies in single booking record.
* meal: Type of meal chosen
* country: Country of origin of customers (as mentioned by them)
* market\_segment: What segment via booking was made and for what purpose.
* distribution\_channel: Via which medium booking was made.
* is\_repeated\_guest: Whether the customer has made any booking before(0 for No and 1 for Yes)
* 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 on waiting list.
* customer\_type: Type of customer (Transient, Group, etc.)
* adr: Average Daily rate.
* required\_car\_parking\_spaces: No. of car parking asked in booking
* total\_of\_special\_requests: total no. of special request.
* reservation\_status: Whether a customer has checked out or canceled,or not showed
* reservation\_status\_date: Date of making reservation status.

**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. Hotel datasets 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 canceled. Nevertheless, due to the characteristics of the variables included in these datasets, their use goes beyond this cancellation prediction problem.

**3. Steps involved:**

* 1. **Null values Treatment**
* We created a copy of the given dataset, so that our original dataset remains unchanged
* Our dataset contains a large number of null values which might tend to disturb our accuracy hence we did some of the alterations so that it won’t affect for our analysis.
* We also found out the percentage of Null Values which we were dealing and found that Company column has got 94.3% of nulls, Agent column has got 13.7% and coming to children and country has got 0.003% and 0.4% which are less than 1%.
* Company column has got 112593 rows of nulls, so we removed the column.
* Agent column has got 16340 rows of nulls, so we replaced agent id with "Agent" and n8ull agent id with "No Agent".
* Then Country and Children column got 488 and 4 rows of null respectively, so we replaced the country column with ‘Unknown’ and Children column with “0”.
  1. **Finding Duplicates**
* From the given set of data we found there were 32014 rows of duplicate data, so we removed all the duplicates that were present.
  1. **Column Addition**
* From the data we got that stay in weekend nights and stay in week day nights, so we concatenated both of them to create total stay.
* We found adults, children, and babies these three were concatenated and created total people
  1. **Performing EDA**
     1. **Correlation Heatmap for Numerical Data:**

Using a describe function we got all the numerical data, from those we created a correlation heatmap.

1. adr is slightly correlated with total\_people, which makes sense as more number of people means more revenue, therefore more adr.
2. total stay and lead time have slight correlation. This meeans that for longer hotel stays people generally plan before the actual arrival.
   * 1. **Checking for outliers:**

From the scatter plot we notice that there is an outlier present in adr column, so we filtered out the outlier which was greater than 5000, so we removed that outlier which was present in that column for the better scatter plot.

* + 1. **EDA based on Hotel:**

1. Here we compared both city hotel and resort hotel and analyzed the most preferred stay.
2. Then in both city and resort hotels we analyzed duration of stays in each of the hotel.
   * 1. **EDA based on Cancelation bookings:**
3. We analyzed the cancelation done based on the months.
4. We analyzed the cancelation rate on each hotel.
   * 1. **EDA Based on arrival Period:**

Here we analyzed the arrival period based on years, months and days in each hotels.

* + 1. **EDA based on Repeated Guests:**

Here we analyzed, whether the guests were repeated or not and if repeated, how many times they were repeated.

* + 1. **EDA based on Market segment and Distribution channel:**

From the given data we got types of market segment and we analyzed the majority mode of bookings done through Market segment and Distribution channel.

* + 1. **EDA based on duration of stays:**

We analyzed the duration of stays done in weekday nights and weekend nights in the hotels.

* + 1. **EDA on Meals:**

Coming to meal category we have BB (Bed and Breakfast), FB (Full board (breakfast, lunch and dinner), HB (Half board (breakfast and one other meal – usually dinner)), SC/Undefined (No meal package). Here we analyzed the meal category preferred.

**3.4.10 EDA based on Type of visitors:**

Here we analyzed that we had various types of visitors like Adults, Children and babies. So among these of the visitors which type of visitors travelled more?

* + 1. **EDA based on the Country:**

From this Data we analyzed in which country the maximum bookings were done for both City and Resort hotel

* + 1. **EDA based on Deposit Type:**

Here we analyzed the type of deposit done majorly for the hotels.

* + 1. **EDA based on Room Type:**

1. Firstly we analyzed the room type taken by a single adult.
2. Next we analyzed the room type taken by the adult size = 2, Lets imagine that the adult = 2 as Couples.
3. Lastly we analyzed for adults > 3, meaning the group/family Travelers room type.

So, these are all the EDA’s which we have analyzed and visualizations were also made upon all of these EDA’s respectively.

1. **Conclusion**

From the given dataset, after performing EDA based upon all of the above mentioned, we conclude the inferences which we found were:

* Majority of the hotels booked are city hotel. Definitely need to spend the most targeting fund on those hotel.
* People also prefer to stay for longer duration in Resort Hotels and prefer City Hotels for Shorter duration.
* We also realize that the high rate of cancellations can be due high no deposit policies.
* We should also target months between May to Aug. Those are peak months due to the summer period.
* We see there is less amount of repeated guest.
* Majority of the bookings are done through online travel agents.
* Most the peoples prefer to stay in week day nights compared to weekend’s night.
* Majority of the peoples prefer BB (Bed & Breakfast) category in the meal section.
* Generally couples travel most from the given data.
* From the given set of data Portugal is the country where majority of the bookings were done.
* Majority of the hotels does not require deposit type, so this may also be the reason for the high cancelation rates.
* Lastly coming to room type, solo travelers prefer type A rooms, couples prefer type A rooms and traveling in a group or families prefer G type of rooms.

Since we were working with limited amount of data there is a pretty good chance of model decay in the future. To counter this, periodic remodeling and configuring of the model would be required.

1. **References**
2. <https://matplotlib.org/>
3. https://seaborn.pydata.org/
4. GeeksforGeeks
5. Analytics Vidhya