

HOTEL BOOKING EDA

CAPSTONE PROJECT

Completed by:

Ankush Kumar

Abstract:

In our Hotel Booking EDA (Exploratory Data Analysis) we are performing various types of analytic strategies based on historical data to analyze whether a reservation is going to canceled or not. It is a machine learning classification problem. The data for this project is provided by program resources for analysis. This data was collected from 2015 to 2017 by extracting it from the hotel's asset management system.

We have performed different level of data analysis from handling the missing and soiled data to convert our data to a proper format.

Performed various analysis for different levels like univariate, Bivariate and Multivariate analysis to analyze the data with different possibilities to extract the proper reasons behind the occurrence and cancellation of hotel booking.

Analysed and finalized significant outcomes that provide us essential acumens about the hotel booking occurrences.

Problem Statement / Analysis Direction:

- Because of the rising tendency of cancellations from year to year, several hoteliers believe that high cancellation rates are the new industry standard, which is entirely incorrect.
- One out of every four hotel visitors cancels their reservations before their stay.
- This cancellation tendency has resulted in the hotel's revenue management system being unable to appropriately estimate occupancy.
- This cancellation tendency has resulted in a loss of potential cost for hotels (unsold room due to cancellation)

Introduction

The hotel industry is one of the fastest-growing sectors of the tourist industry, especially with the development of massive online travel agencies (OTAs) that make reserving a hotel as simple as it's ever been.

Hotel income increased by 18% to \$3.6 billion in 2017. The rise of the hotel business may also be observed in the fact that the overall number of hotel visitors in 2017.

According to the statistics, the rate of cancellations from July 2015 to August 2017 was 37.1 percent, which is somewhat lower than the industry standard of 41.3% at the time.

Steps For Exploratory Data Analysis(EDA)

1. Data information:

Collected various information about the data by reading the data from csv file name as “Hotel Booking”.

We get to know there are total 32 columns, in which 12 columns are numeric and remaining 10 are categorical columns. And our target/label class column is “is_canceled” which contains 0 , 1 binary values in which 1 -> Cancelled Booking and 0 -> Booking not Cancelled.

There are total 119390 entries for each columns but also it some columns have missing data i.e. Null/NaN values.

2. Data Cleaning and Removal of Duplicacy:

Detected various columns with missing values ['children', 'country', 'agent', 'company']

Out of Total 119390 rows , 'agent' and 'company' column contains very large amount of Null/NaN values in it i.e. 16340 and 112593 and this much of Null/NaN values will impact our analysis. We have Dropped these columns as we haven't used these columns for our analysis.

And 488 in 'country' and 4 in 'children' column contains Null/NaN values and which is very less so we have just deleted these missing value rows from our dataset.

After that we have checked for data duplicacy in which we have found that there are 31984 duplicate entries. and we have just deleted these duplicate entries.

3. Data Visualization

We have performed different types of visualization on the basis of Univariate, Bivariate and multivariate analysis.

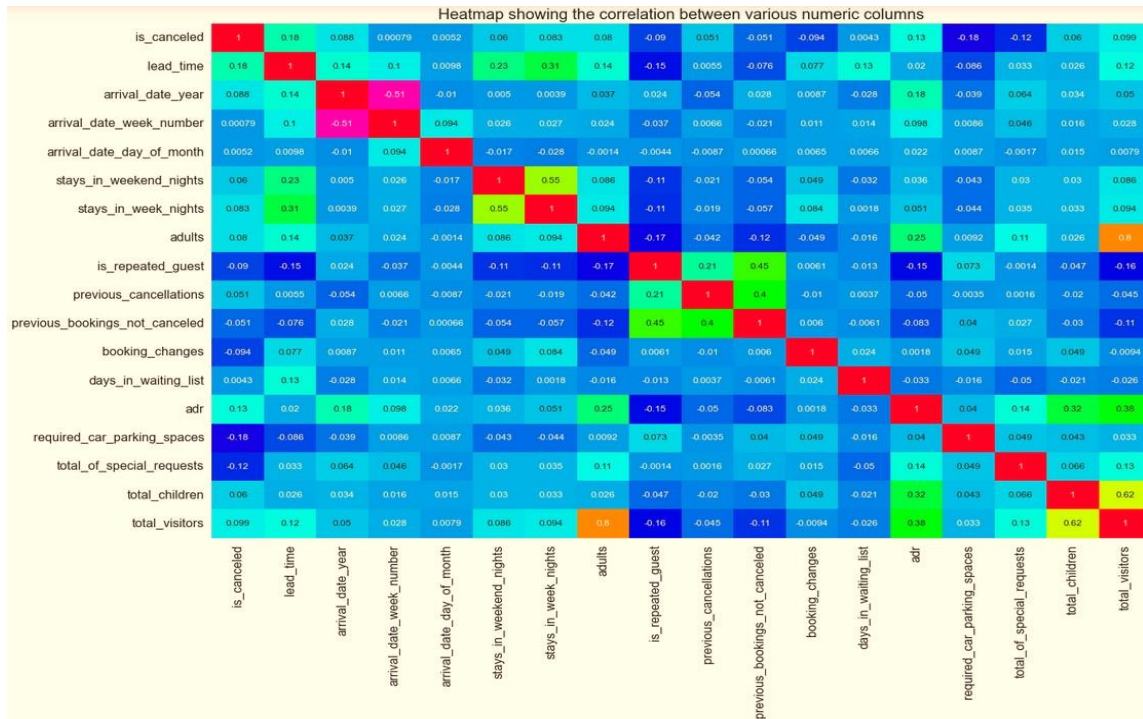
Firstly we have conducted Univariate analysis as we also need to understand various features/columns individually that what kind of importance and insights they brings for our analysis.

Secondly we have performed Bivariate analysis so that we can analyse the impact of one column/feature to another feature and where these insights leads us.

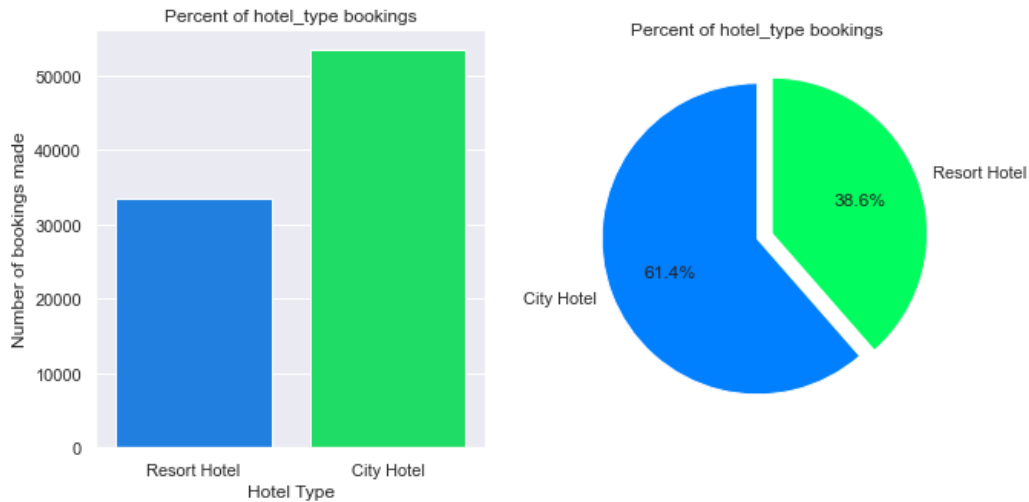
At last we have performed Multivariate analysis in which we came to know the impact of multiple features to the booking occurrence of hotels.

4. Correlation

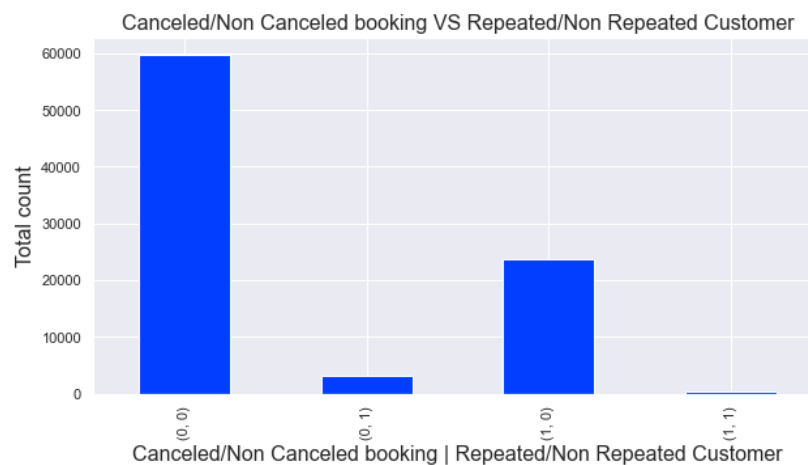
As we know that correlation is a very important observation task to get the proper idea that how our features are idependent from each other and if they are not independent and are related to each other then there will be lots of abiguity for our model creations that misleads our analysis.



- **Some glimpses we have gathered from our various visualtization analysis:**



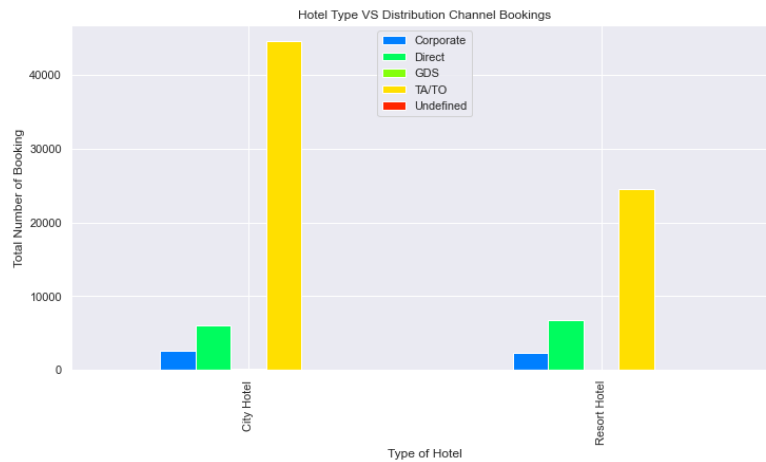
- ❖ City hotel has bookings above 50000 which is around 61.4% and Resort hotel has less than 35000 which is around 38.6% bookings.



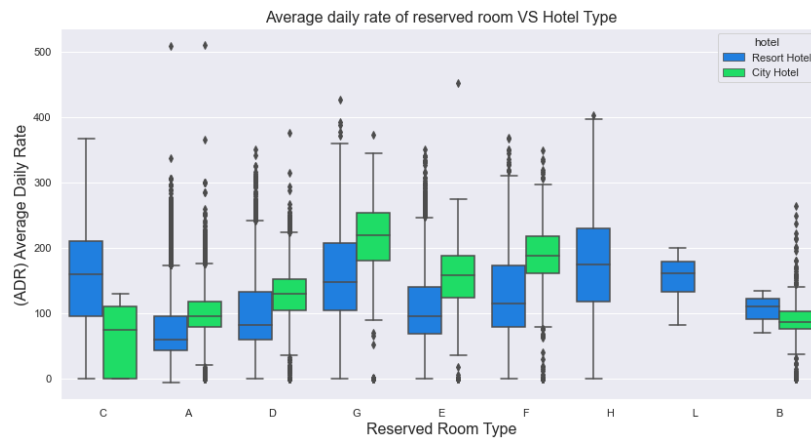
- ❖ The lowest cancellation occurs when the customer is repeated.
- ❖ That means the existing customers are very less likely to cancel the booking



- ❖ Maximum number of booking is done by the Online TA(Travel Agent) for the City Hotels as well as for the Resort Hotels.



- ❖ The most number of booking is done by the TA/TO Distribution Channels for both City and Resort Hotel.



- ❖ Room type 'G' has the highest and Room type 'C' has the lowest Average Daily Rates (ADR) in City Hotels.

- ❖ Room type 'C' and 'H' has the highest and Room type 'A' has the lowest Average Daily Rates (ADR) in Resort Hotels.

List of Question to help achieving the goal:

1. Do customers have any preference with or without children's
2. How daily average rate is impacting the reserved room type in hotels
3. Does total stay per month impacted by customers with and without children's
4. Do reservation status impacted by type of customers
5. Do customers cancel their booking if they are allotted with different room type.

We got very intuitive results by performing analysis on the above questions:

- The count for stays with and without children is pretty similar in both circumstances.
- Customers who stayed with and without children for 1-7 days have the same count.
- If we consider the usual stay, there will be three to four days during which people will have their children with them, otherwise only adults will be there.
- The majority of guests have no need for parking.
- The Online TA customer group has the highest number of hotel reservations in the city and at resort hotels, the number of customers who booked directly is the highest, however, it is clear that the majority of the appointments are made by consumers of the Online TA group.
- The impact of room type allocation on clients is minimal.

Conclusion and Summary

- The overwhelming majority of reservations are for hotels in cities. Resort hotels have fewer cancellations than city hotels. The aviation industry has the shortest wait time.
- The months of August, July, and should saw the foremost hotel bookings within the city. rock bottom cancellation rate occurs when a consumer is repeated.
- The time interval for July is sort of long, whereas the time interval for January and February is extremely short. In July, August, and September, the typical daily rate for Resort Hotels is higher.
- In June, July, and May, the typical daily rate for City Hotels is higher. Customers travelling with children have little preference for the sort of hotel they stay in.
- In both cases, the amount of stays with and without children is quite equal. Customers who stayed for 1-7 days with or without children had an equivalent count.
- If we take the typical stay, three to four days are going to be spent with youngsters, while the remainder of the time are going to be spent with adults solely.
- Online TA customer group has the very best number of hotel reservations within the city and at resort hotels.
- As a result, the impact of room type allocation on clients is minimal. the bulk of appointments are made by consumers of the web TA group who book directly through the web site .

Future work recommendation:

- For additional study, hotel parameters (hotel star rating, brand, and so on) might be added to the dataset in the hopes of improving the models and quantifying the value of these elements.
- Additional study in other locations and with different hotels might help to better understand why people cancel their reservations.

References:

- [Towardsdatascience](#)
- [Analyticsvidya](#)
- [Becominghumanai](#)
- [Siteminder.com](#)
- [Tmstudies.net](#)