Hotel Bookings Exploratory Data Analysis

Business understanding

Analytics in the hotelier world today is important, and nowadays this business cannot be run with some sensible and smart use of data.

Here I demonstrate how to use data to analyse three business important concepts in the fields of revenue management and marketing.

The analysis tries to answer three questions

1. How strong is the seasonality in these hotels?
2. Up to what point ADR, length of stay, and lead time for groups reservation differ from individual/transient ones? are rooms harder to sell in nearby dates of a group's stay?
3. Can we predict a cancellation, just with the information available at the moment this reservation has been made?

The first question relates to detecting seasonality patterns. A basic principle to establish a pricing strategy. The second addresses the importance of having a specific pricing strategy for groups. Specific group's behaviours can end up with lower total revenue, because they may have been booked with too much time in advance, and in odd length of stays. Moreover, this awkward behaviour can make revenues having a hard time trying to sell rooms in the nearby dates. Finally, be able to detect whether or not a reservation will end up being cancelled, is a powerful weapon for the marketing department, with important potential gains

Data understanding

The original data has been extracted from the booking changelog, one day prior to arrival date, to avoid leakages. There are some variables that had to be extracted fro, other DB systems. The data presented consists of booking records from two hotels in Portugal: one resort hotel and one city hotel.

The direct source for this analysis is Kaggle and the original source is ScienceDirect.

In total there are 119,390 records and 32 features, with all of these features presenting almost (or none) null values, except for the variable "company" (94 % of records are missing).

Objective

We are provided with a hotel bookings dataset.

Out 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.

Data preparation

For the two first questions, the analysis has been performed only including confirmed bookings, given that some of the data can change when the client checks-in and cancelled bookings can have a different distribution for some variables. Also, this is an a-posteriori analysis, that is, only clients that have stayed in the hotel will be taken into account.

Also, a major transformation is carried out: from booking records we go into daily stay data: the data is aggregated by days and months, and not by check-in date, but by stay date. This requires some complex operations that are performed using helper functions built specifically for this analysis.

Several date variables are extracted from the original data, since some of the analysis deals with aggregation for different time windows.

For prediction of cancelled bookings, the analysis is performed at the booking level with the entire data set, and new features are created in the process. The major part, involves grouping qualitative data with too many categories into ones with a limited number of them.

Dataset

We are given a hotel bookings dataset. This dataset contains booking information for a city hotel and a resort hotel. It contains the following features

```
- 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.
.
```

- Total number of rows in data: 119390
- Total number of columns: 32

Data Cleaning and Feature Engineering

(1) Removing Duplicate rows

All duplicate rows were dropped.

(2) Handling null values

- Null values in columns `company` and `agent` were replaced by 0.
- Null values in column `country` were replaced by 'others'.
- Null values in column `children` were replaced by the mean of the column.

(3) Converting columns to appropriate data types

- Changed data type of `children`, `company`, `agent` to int type.
- Changed data type of `reservation_status_date` to date type.

(4) Removing outliers

- One outlier was found in the `adr` column. Simply dropped it.

(5) Creating new columns

- Created new column `total_stay` by adding `stays_in_weekend_nights`+`stays_in_week_nights`.
- Created new column `total_people` by adding `adults`+`children`+`babies`.

Exploratory Data Analysis

Performed EDA and tried answering the following questions:

```
 Q1) Which agent makes the most no. of bookings?
 Q2) Which room type is in most demand and which room type generatesthe  highest
adr?
 Q3) Which meal type isthe  most preffered meal of customers?
 Q4) What isthe  percentage of bookings in each hotel?
 Q5) Which is the most common channel for booking hotels?
 Q6) Which are the most busy months?
```

```
 Q7) From which country most of the guests are comin ?
 Q8) How long do people stay at the hotels?
 Q9)  Which hotel seems to make more revenue?
 Q10)  Which hotel hasa  higher lead time?
 Q11)  What is preferred stay length in each hotel?
 Q12)  Which hotel has higher bookings cancellation rate.
 Q13)  Which hotel hasa  high chance that its customer will return for another
stay?
 Q14)  Which channel is mostly used forthe  early booking of hotels?
 Q15)  Which channel hasa  longer average waiting time?
 Q16)  Which distribution channel brings betterrevenue-generatingg deals for
hotels?
 Q17)  Which significant distribution channel has the highest cancellation
percentage?
 Q18) Doesa  longer waiting period or longer lead time causes the cancellation of
bookings?
 Q19) Whether not getting allotted the same room type as demand is the main cause
of cancellation for bookings?
 Q20) Does not alloting the  same room as demanded affect adr?
 Q21) What is the trend of bookings within a month?
 Q22) Which types of customers mostly make bookings?
```

Mainly performed using Matplotlib and Seaborn library and the following graph and plots had been used:

- Bar Plot.
- Histogram.
- Scatter Plot.
- Pie Chart.
- Line Plot.
- Heatmap.
- Box Plot

Univariate Analysis:

Performed univariate analysis and made following conclusions:

```
1.) Agent no. 9 has made most no. of bookings.
 2.) Most demanded room type is A, but better adr generating rooms H, G and C.
Hotels should increase the no. of room types A and H to maximise revenue.
 3.) Most popular meal type is BB(Bed and Breakfast).
 4.) Around 60% bookings are for City hotel and 40% bookings are for Resort hotel,
therefore City Hotel is busier than Resort hotel.
 5.) Guests use different channels for making bookings out of which most preferred
way is TA/TO.
 6.) July- August are the most busier and profitable months for both of hotels.
 7.) Most of the guests came from european countries, with highest number of
guests from Portugal.
```

```
 8.) Most common stay length is less than 4 days and generally people prefer City
hotel for short stay, but for long stays, Resort Hotel is preferred.
```

Bivariate Analysis :

We tried to answer following questions

```
1.) Overall adr of City hotel is slightly higher than Resort hotel and no. of
bookings of City hotel is also higher than Resort hotel. Hence, City hotel is
makes more revenue.
 2.) City hotel has slightly higher median lead time. Also median lead time is
significantly higher for both hotels, this means customers generally plan their
hotel   visits way early.
 3.) Almost 30 % of City Hotel bookings got canceled.
 4.) Both hotels have very small percentage that customer will repeat, but Resort
hotel has slightly higher repeat % than City Hotel.
 5.) TA/TO is mostly used for planning Hotel visits well ahead of time.
 6.) While booking via TA/TO one may have to wait a little longer to confirm
booking of rooms.
 7.) GDS channel brings higher revenue generating deals for City hotel, in
contrast to that most bookings come via TA/TO. City Hotel can work to increase
outreach on GDS channels to get more higher revenue generating deals.
 8.) TA/TO has highest booking cancellation %. Therefore, a booking via TA/TO is
30% likely to get cancelled.
 9.) Longer lead time has no affect on cancellation of bookings.
 10.) Not getting same room as demanded is not the case of cancellation of rooms.
A significant percentage of bookings are not cancelled even after getting
different room as demanded.
 11.) Not getting same room do affects the adr, people who didn't got same room
have paid a little lower adr.
 12.) Arrivals in hotels increases at weekends and also the avg adr tends to go up
as month ends.
 13.)Moslty bookings are done by couples(bookings have two adults.)
```

Conclusion

```
(1) Around 60% bookings are for City hotel and 40% bookings are for Resort hotel,
therefore City Hotel is busier than Resort hotel. Also the overall adr of City
hotel is slightly higher than Resort hotel.
(2) Mostly guests stay for less than 5 days in hotel and for longer stays Resort
hotel is preferred.
(3) Both hotels have significantly higher booking cancellation rates and very few
guests less than 3 % return for another booking in City hotel. 5% guests return
for stay in Resort hotel.
(4) Most of the guests came from european countries, with most of guests coming
from Portugal.
(5) Guests use different channels for making bookings out of which most preferred
way is TA/TO.
(6) For hotels higher adr deals come via GDS channel, so hotels should increase
their popularity on this channel.
(7) Almost 30% of bookings via TA/TO are cancelled.
```

(8) Not getting same room as reserved, longer lead time and waiting time do not affect cancellation of bookings. Although different room allotment do lowers the adr.
(9) July- August are the most busier and profitable months for both of hotels.
(10) Within a month, adr gradually increases as month ends, with small sudden rise on weekends.
(11) Couples are the most common guests for hotels, hence hotels can plan services according to couples needs to increase revenue.
(12) More number of people in guests results in more number of special requests.
(13) Bookings made via complementary market segment and adults have on average high no. of special request.
(14) For customers, generally the longer stays (more than 15 days) can result in better deals in terms of low adr.

And many more conclusions.

Challenges

(1) There was a lot of duplicate data.
(2) Data was present in wrong datatype format.
(3) Choosing appropriate visualization techniques to use was difficult.
(4) A lot of null values were there in the dataset.