EDA ON HOTEL BOOKINGS

Prajwal D U, Data Science Trainees, Almabetter, Bangalore.

ABSTRACT

Hotels use Database management systems to accept payments and keep track of sales. The hotel database management system is most commonly known as POS (Property Management System).

Our experiment is to analyze the data given by the hotelier. Also, make that data clean and ready to use to make conclusions from that data.

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. However, the remaining tourism and travel industries like hospitality, cruising, theme parks, etc., have different requirements and particularities that cannot be fully explored without industry's specific data. Hence, two hotel datasets with demand data 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. One of the most important properties in data for prediction models is not to promote leakage of future information. In order to prevent

this from happening, the timestamp of the target variable must occur after the input variables' timestamp. Thus, instead of directly extracting variables from the bookings database table, when available, the variables' values were extracted from the bookings change log, with a timestamp relative to the day prior to arrival date (for all the bookings created before their arrival date).

PROBLEM STATEMENT

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests? This hotel booking dataset can help you explore those questions!

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.

FEATURE DESCRIPTION

The data feature in this dataset respectively

- ADR (Numeric) Average Daily Rate as defined.
- Adults (Integer) Number of adults.
- Agent (Categorical) ID of the travel agency that made the booking.
- ArrivalDateDayOfMonth (Integer) Day of the month of the arrival date.
- ArrivalDateMonth (Categorical) Month of arrival date with 12 categories: "January" to "December".

- arrival_date_week_numbe r (Integer) Week number of year for arrival date.
- arrival_date_year (Integer) Year of arrival date.
- Babies (Integer)number of babies in count.
- Children (Integer) number of children.
- Company (Integer) ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons.
- Country(object) Country of origin. Categories are represented in the ISO 3155–3:2013 format.
- customer_type(categorical) Type of booking, assuming one of four categories: Contract - when the booking has an allotment; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract; Transient-party – when the booking is transient, but is associated to at least other transient booking.
- distribution_channel(categ orical) Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
- days_in_waiting_list (Integer) Number of days the booking was in the waiting list before it was confirmed to the customer.
- Hotel(categorical) Hotel (H1 = Resort Hotel or H2 = City Hotel).
- is_canceled (Integer) Value indicating if the booking was canceled (1) or not (0).
- is_repeated_guest (Integer) Value indicating if the booking name was from a repeated guest (1) or not (0)
- lead_time (Integer) Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.

- Meal(categorical) Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal.
- market_segment(categorical) Market segment designation. In categories, the term “TA”
- means “Travel Agents” and “TO” means “Tour Operators”
- previous_cancellations(categorical) Number of previous bookings that were canceled by the customer prior to the current booking.
- previous_bookings_not_canceled (Integer) Number of previous bookings not canceled by the customer prior to the current booking.
- reservation_status(categorical) Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out.
- reservation_status_date (Date) Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when the booking was canceled or when the customer checked-out of the hotel.
- stays_in_weekend_nights (Integer) Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
- stays_in_week_nights (Integer) Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel.
- total_of_special_requests (Integer) Number of special requests made by the customer (e.g., twin bed or high floor).

EXPLORATORY DATA ANALYSIS (EDA)

Firstly, we imported libraries and dataset, some of the libraries used are NumPy, pandas, matplotlib, seaborn, and warnings. Once the data is collected, the process of analysis begins. But data has to be

translated in an appropriate form. This process is known as Data Preparation.

- Validate data
- Clean the data set Checking and deleting the duplicate values Statically adjust the data
- Store the data set for analysis
- Analyze the data

DATA PREPROCESSING

A dataset may contain noise, missing values, and inconsistent data, thus, pre-processing of data is essential to improve the quality of data and time required in the data mining.

- CLEANING AND MANIPULATING THE DATASET

CLEANING

After completing the Data Sourcing, the next step in the process of data preprocessing is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system. Irregularities are of different types of data.

- Missing Values
- Incorrect Format
- Incorrect Headers
- Anomalies

We dropped features like the company where we had null values nearly 94 %.

After you find any missing values, you can categorize your values to help determine what statistical and visualization methods can work with your dataset like categorical,

DATA MANIPULATION

Manipulation of data is the process of manipulating or changing information to make it more organized and readable. Made some new features with the help of columns present in the datasets .

- The next step is to find the relationship between the features to do the analysis.
- Locating outliers in your dataset is another important step to conducting EDA. Outliers

are values in your dataset that are significantly different from the rest of the values. Outliers can be much higher or lower than the other values in a dataset.

- It's helpful to conduct exploratory data analysis to help you understand a dataset before you start to model it as we did by asking the following questions.

1. Which year did the hotels (both city and resort) have the highest footfall?
2. Which are the top 10 countries from which travelers and tourists visit hotels?
3. What type of Bookings is majorly seen in the hotels according to market segment?
4. How much do customers pay for accommodation per night?
5. What is the booking relation between Resort Hotel and City Hotel?
6. Find Top 20 agents with the highest bookings?
7. Which was the most booked accommodation type (Single, Couple, Family)?
8. Which is the busiest month for hotels?
9. How much Booking cancellation is per month
10. Finding the occupancy and level of profitability over months.
11. How long does the customer stay in the hotel?

By solving these questions we tried to do some exploratory data analysis and tried to find the relationship between features.

CHALLENGES

Following are we faced during EDA :

Data acquisition and importing necessary libraries.

Data wrangling and analyzing data.

Handling null values of variables like country, agent, etc.

Cleansing data and dropping columns which are not necessary like we do for a variable like company

While doing data wrangling we came to know that there are such rows in which adults, children, and babies are null at the same time that will be not possible in case of booking.

Performing EDA on the dataset to find relation and find some insights from hotel booking data.

CONCLUSION

Exploratory data analysis is one of the key competencies of a data scientist at a startup. You should be able to dig into a new data set and determine how to improve your product based on the results. EDA is a way of understanding the shape of a data set, exploring correlations within the data, and determining if there's a signal for modeling an outcome based on the different features. By performing EDA we came to know the footfall of the hotels per year. We also came to know from which countries most of the guests visit the hotel. We also analyzed that most of the booking that hotels receive is from an online travel agent. We also visualized how the average daily rate hotels charge according to the room type. The city hotel has bookings near to 60% while the resort hotel has nearly 40% booking out total booking. We also came to know the trend of booking according to the month with respect to hotels. We also analyzed the trend of the average daily rate changes according to month. By correlation matrix, we came to know that there is no such positive linear correlation because most of the values are near zero. Hence variables do not show any correlation.