Hotel Booking Analysis

Prakhar Kumar

Data Science Trainee,

Almabetter

Abstract:

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 the 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 of data for prediction models is not to promote leakage of future information [3]. 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 before 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.

Introduction:

This data article describes two datasets with hotel demand data. One of the hotels is a resort hotel and the other is a city hotel.

Both datasets share the same structure, with 31 variables describing the 40060 observations of resort hotels and 79330 observations of city hotels.

Each observation represents a hotel booking. Both 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.

Variables:

- **hotel**: Two hotels are given: Resort Hotel City Hotel
- **is_canceled** : 1: Canceled 0: Not canceled
- **lead_time**: gap between booking and arrival
- **arrival_date_year** :arrival year
- **arrival_date_month** :arrival month
- **arrival_date_week_number** : arrival week
- **arrival_date_day_of_month** :arrival date
- **stays_in_weekend_nights**:count of nights the guests booked the hotel during Sat-Sun
- **stays_in_week_nights**:count of nights the guests booked the hotel during Mon-Fri
- **weekly_stays**:duration of stay including weekend nights and weeknights stay
- **adults**:count of adults
- **children**:count of children
- **babies**:count of babies
- **meal**: meal type (no meal package; BB; HB; FB)
- **country**: country of guests
- **market_segment** :Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

- **distribution_channel**: Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”

- **is_repeated_guest** :1: Yes 0: No
- **previous_cancellations**: count of previous bookings that were canceled by the customer before the final booking
- **previous_bookings_not_canceled**: count of no canceled bookings
- **reserved_room_type**: booked room category
- **assigned_room_type**: assigned room category
- **booking_changes**: count of changes made by the customer before final booking
- **deposit_type**: type of deposit made by the customer

- **agent**: travel agent id
- **company**: booking company id
- **days_in_waiting_list**: count of days the booking was on the waiting list before it was confirmed
- **customer_type** :Transient Contract Group Transient-party
- **adr**: average daily rate for the booking
- **price**:total price spent by a guest entity
- **required_car_parking_spaces**:count of car parking spaces allotted to the customer
- **total_of_special_requests**:count of special requests made by the customer
- **reservation_status**:status of reservation
- **reservation_status_date**:date corresponding to the status of reservation
- **total_guests**: sum of adults, children, babies

- **arrival_date** : in date format(yyyy-mm-dd)

Steps involved:

1) Collecting and understanding data
2) Cleaning of data i.e., removing or replacing null values, removing duplicate items
3) Exploratory data analysis (EDA): This process helped us figure out various aspects and relationships between the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

Statistics involved:

- **Categorical variable:** A **categorical variable** (also called a **qualitative variable**) is a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property. For e.g.- meal, children, hotel, etc
- **Continuous variable:** A continuous variable is defined as a variable that can take an uncountable set of values or an infinite set of values. For e.g. adr, weekend_stays, week_stays, etc.
- **Mean:** It is the total sum of observations divided by the no.of observation

- **Mode:** It is the highest occurring value in the dataset.
- **Univariate Analysis:** Analysis done on a single variable is called univariate analysis.
- **Bivariate Analysis:** Analysis done on the double variable is called univariate analysis.
- **Multivariate Analysis:** Analysis done on more than two variable is called univariate analysis.

Conclusion:

- Most bookings of City and Resort hotels are made in the year 2016.
- City hotels are most preferred by guests as it was booked by 66.4% of guests. Thus we can say that city hotels are the busiest hotel.
- About 37% of guest cancelled there reservation hotels in which City hotels have high cancelation rate than Resorts hotels.
- Majority of guest are from western Europe i.e Portugal,France,Great Britain
- Most guests are not preferred to made there changes in current booking.
- Only 3.2% people were revisited the hotels,rest 96.8% were new guests,thus retention rate is low.
- About 93.8% of guests does not required parking.
- BB(Bed and Breakfast) are the most preferred type of meal choose by guests.
- Booking made by TA/TO higher than the other distribution channels.
- On an avg guests stay for 2 to 3 nights in city hotels but as night of stays increases guests prefer resort hotels rather than city hotels.
- About 55.6% guests of city hotels and 32% guests of resort hotels choose no deposit as deposit type.
- Most of customers are Transient or Transient-party type of guests
- Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
- Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.
- Direct and TA/TO are generating more revenue than other distribution channels.

Reference:

Stackoverflow

GeeksforGeeks

Kaggle