Expedia Hotel Dataset Analysis

FEBRUARY 22 2021

IE School of Human Sciences and Technology
Authored by: Ayan Ghosh
MBD Section 1
September 2020 Intake

Expedia Dataset

Background/scenario description

Which hotel type will an Expedia customer book?

While a person plans a vacation or a weekend escape to a nearby place, it can turn out to be an exhausting affair. With having so many choices of hotels to choose from at every destination, it is difficult for Expedia to understand a visitor's needs and recommend a proper hotel based on their personal preference. Should the visitor be shown a new hotel with a jacuzzi, or would he enjoy a vacation staying in a room with great view from the window? Would he rather prefer a hotel with a proper family setup, or have some alone time and not be disturbed by the people around? These are few questions that Expedia needs to understand from this data to provide better recommendation with the aim of making it a customer centralized business.

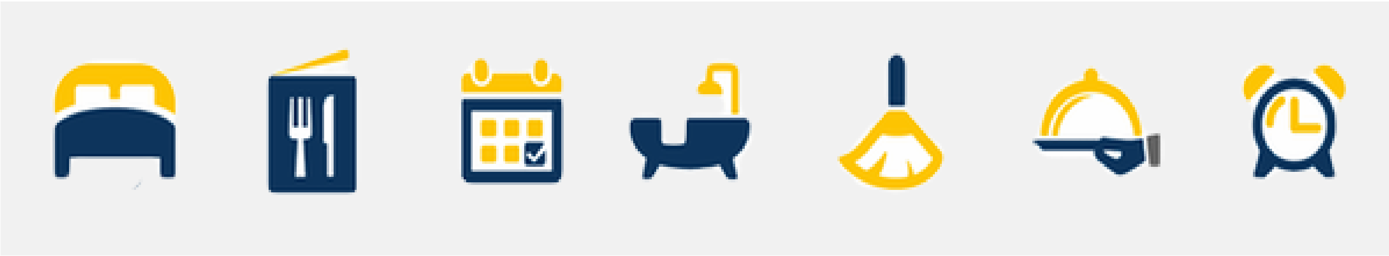
Expedia provided this dataset; in hope the analysts build a viable and sustainable model to predict and provide their millions of customer base with personalized recommendations to boost their market share in the travel industry segment.

Normally, during the time when this dataset was released, Expedia uses search parameters to adjust their hotel recommendations, but there are not enough customer specific data to personalize them for each user. In this competition, Expedia expected the analysts to contextualize customer data and predict the likelihood a customer chooses a personalized recommendation provided by Expedia, with 100 different hotel clusters to choose from.

The mentality of Expedia of the following exercise is as follows:

"Human-centered & tailored experiences do really make the difference and will boost company's profit"

One thing about the dataset, is that due to data privacy concerns, the data has been anonymized and we will be dealing with IDs for most of the attributes. This data in this competition is a random selection from Expedia and is not representative of the overall statistics.



Goal of the analysis

The following dataset consists of the customer's behaviour within various Expedia sites, including the search details, destination location details, whether the search resulted in a package or a single hotel, or after viewing all the recommendations and choices, the customer was convinced enough that he booked finally. It also includes the user location and logged timestamp data, along with their used device selection. The following table consists of all the details of the attributes and their datatypes have also been mentioned.

Expedia was interested in building a viable predicting model, which will be suggesting personalized hotel cluster a user would be interested in and will have a higher likelihood of a visit being converted to a booking. The already existing algorithm used by Expedia at that point, recommended hotel cluster mainly based on previously popular hotels at a particular destination. In this analysis, the various attributes like historical price, customer ratings, geographical locations relative to city center are grouped together and they came up with 100 different hotel clusters. The goal is to predict 5 hotel clusters where a user is more likely to stay, out of a total of 100 hotel clusters. These will be serving as good identifiers to which type of hotels a customer will be preferring based on their personal choices. The main aim of the analysis is to predict these hotel clusters to a customer based on their requirements and search pattern thus retaining and increasing market size in the travel sector.

In this report we will only be considering profiling analysis and finding out the trends between various attributes within Spark Environment, and we will not be focusing on building a model to predict hotel clusters – thus we would not take the test data into consideration. The Training data consists of customer preferences and search patterns, including the details whether they finally booked or not, the period was mainly from the year 2013 and 2014.

The training set consists of 37,670,293 entries. Apart from this, the dataset also provides some latent features for each of the destinations recorded in the train data sets. The data is anonymized and almost all the fields are in numeric format. Missing data, ranking requirement, and the curse of dimensionality are the main challenges posed by this dataset, the report focuses on finding trends between different attributes to find the driving factors within this dataset helping to find customer clusters and finally build a classification model (out of scope of this report).

Data fields

Column name	Description	Data type
date_time	Timestamp	string
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)	int
posa_continent	ID of continent associated with site_name	int
user_location_country	The ID of the country the customer is located	int
user_location_region	The ID of the region the customer is located	int
user_location_city	The ID of the city the customer is located	int
orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated	double
user_id	ID of user	int
is_mobile	1 when a user connected from a mobile device, 0 otherwise	tinyint
is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise	int
channel	ID of a marketing channel	int
srch_ci	Check-in date	string
srch_co	Checkout date	string
srch_adults_cnt	The number of adults specified in the hotel room	int
srch_children_cnt	The number of (extra occupancy) children specified in the hotel room	int
srch_rm_cnt	The number of hotel rooms specified in the search	int
srch_destination_id	ID of the destination where the hotel search was performed	int
srch_destination_type_id	Type of destination	int
hotel_continent	Hotel continent	int
hotel_country	Hotel country	int
hotel_market	Hotel market	int
is_booking	1 if a booking, 0 if a click	tinyint
cnt	Number of similar events in the context of the same user session	bigint
hotel_cluster	ID of a hotel cluster	int

Analysis deep dive

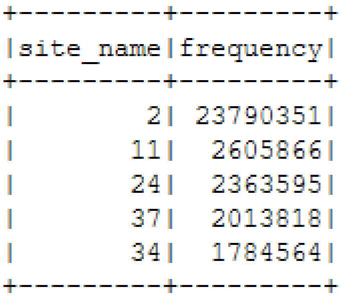
The report analysis will be focusing on profiling users, their search characteristics, user location analysis. It will help to understand the customers better, understand how they are distributed geographically, the time when customers are most active and destination profiling. The dataset is analyzed at the beginning, looking into the basic statistics, null values within columns and the datatypes for each column, those which will be dealt with as the analysis proceeds.

User Profiling:

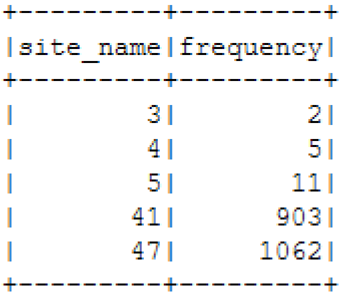
Querying using SQL API, the dataset consists of 1198786 unique users. The number of website visits for top 5 users, irrespective of booking a hotel or not, was having a range of 530-491 visits within a period of two years, whereas for the customers on the trailing end, who rarely visits the website had a range of 2 visits.

Site Profiling:

Similar to the process used in user profiling, Expedia has 45 sites with different domains. Site ID 2 generated the most traffic with respect to other websites, with 23790351 visits within these two years, whereas Site ID is the website which generated the least amount of traffic, only 2 visits in two years. We can delve deep into it if the site was down or had some issues explaining this numbers, but that analysis is out of the scope of this report. Below is the table, chalking out the most and least popular websites.



```
+---------+---------+
|site_name|frequency|
+---------+---------+
|        2| 23790351|
|       11|  2605866|
|       24|  2363595|
|       37|  2013818|
|       34|  1784564|
+---------+---------+
```



```
+---------+---------+
|site_name|frequency|
+---------+---------+
|        3|        2|
|        4|        5|
|        5|       11|
|       41|      903|
|       47|     1062|
+---------+---------+
```

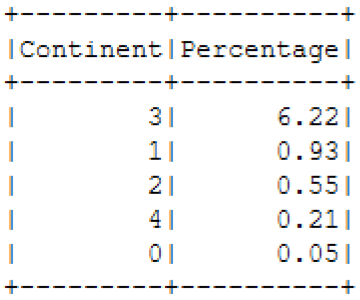
Booking:

Among all the visits, only 7.965% of the visits were converted to booking.

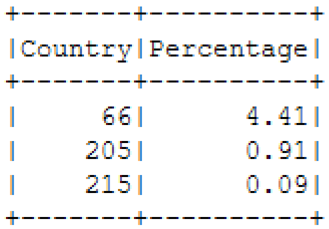
Continent, Country, region and City profiling:

The dataset contains the anonymized details about the user's geographical location. It should be understood what continent, country or the region the user is based from. This information is required to understand the geolocalized clusters – as customers from same regions will have a higher tendency to visit the same destinations or prefer to spend their vacations in a similar hotel cluster. After doing the profiling, we tried finding the largest continent, country, city and region; but this analysis may turn out to be misleading, in the sense, different continents have the same city name, same region name, and as we drill down it becomes evident. Another reason can be how the system logged the geographical details – for example Spain and Espana are provided with different IDs. But it provides with the geolocalized profile of the customer segment. The largest continent based on the number of countries was Continent 3, within Continent 3 the largest country based on number of regions was Country 69 and the largest region within them was 844 and 664, having 110 cities within them.

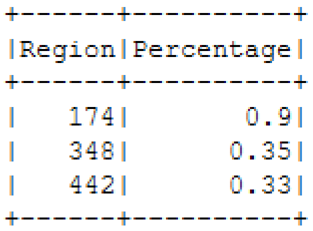
The analysis continues this profiling trying to find out which of the continents provided us with the maximum booking percentage and drilling down to the city and region level. The top continent from which the maximum bookings came was Continent 3 with 6.22% out of 7.965% of total booking, within that country 66, generated maximum bookings with 4.41% with Region 174 leading with respect to the others.



```
+---------+----------+
|Continent|Percentage|
+---------+----------+
|        3|      6.22|
|        1|      0.93|
|        2|      0.55|
|        4|      0.21|
|        0|      0.05|
+---------+----------+
```



```
+-------+----------+
|Country|Percentage|
+-------+----------+
|     66|      4.41|
|    205|      0.91|
|    215|      0.09|
+-------+----------+
```



```
+------+----------+
|Region|Percentage|
+------+----------+
|   174|       0.9|
|   348|      0.35|
|   442|      0.33|
+------+----------+
```

Log Time Analysis:

The following analysis is broken down into two parts: One which focused on the user data who completed all the steps and went forward with the bookings and the next part when the users just surfed around on the website without booking anything. This analysis will help to identify and segregate time periods when the users booked more than the other time. The analysis is done grouped by year, months, weekdays and even time of the day, to understand customer behaviour with more insights. The base attribute taken into account for this analysis is the logged time, date_time.

Booking:

In this part, the analysis concentrates on the users who completed all the necessary steps and completed a booking from any Expedia website. First, the analysis checks which year had more bookings, followed by which months and which weekdays generated more bookings than the rest, also querying the most and least active time periods within a day. The bookings nearly doubled in 2014, maximum bookings were done in the months of July, August and October, least were in January, February and March. Considering weekdays, most of the bookings were done mid-week (Tuesday- Wednesday-Thursday) and the least were on weekends and even on Monday. Looking into the active hours, customers booked mid-day(11am-1pm) the most and the least booking activity was recorded at late night (2am-4am).

Visits:

Similar concept as booking-time analysis, but the only difference is only the logs are considered which did not convert to booking. This will help to distinguish a time pattern behaviour to better understand the customers. The analysis finds visits in 2014 is nearly 2.5 times to that of 2013, this data when compared to the booking one infers that more people visited the websites, but a smaller section of people booked with respect to that of 2013. There are similar trends to that of booking month-wise, i.e., the months that generated more traffic are same as the ones which generated maximum booking, same with the months with least visits, and the same goes for weekdays trends. But, on analyzing the time periods, it revealed that customers tend to surf more without booking, during the evening (6pm-7pm), with respect to more bookings mid-day, least surfing time again remains the same.

Search Date Analysis:

This part of the analysis is done with respect to the search date by the users, keeping in mind that attributes srch-ci and srch_co have null values, but with respect to the size of the full dataset, it would be fine not to include those columns in this analysis. Also, another factor to be kept in mind, is that the search dates, specifically the srch_ci (Check-in Date) should be greater than or equal to the date_time attribute, also, the Check-in Date should be lesser than or equal to the Check-Out date.

When grouped by year, the average booking nights was maximum in 2023 with 7 frequency counts, there are other similar occurrences, but we can reject these because, the frequency counts are extremely low and if the days were checked, it mostly denotes a year, so probably the customer searched wrongly. The highest average booking nights was logged in 2015 with 4.15 nights, followed by 2016 and 2014. But, looking into the frequency counts, most of the customers looked for hotels in the year 2014. Month-wise, the highest average booking nights were searched for in the months of December, January and February and the least in October. But frequency-wise, most of the searched months were December, July and August and the least in February.

Destination Profiling:

To continue with this analysis, Dataframe API is used and also the null values of the attribute orig_destination_distance needs to be filled up with -1 as there is no clear data about these observations.

The analysis looks into the distinct values of the destination attributes namely, srch_destination_type_id, hotel_continent, hotel_country, hotel_market and hotel_cluster. Main points being the dataset has 9 search destination type in 7 continents, with 100 hotel clusters to choose from. In that, the most popular search destination type is 1 (23304952 counts) and the least is 2 (1 count), the most popular continent ID was 2 and least was 1. Among the hotel clusters, the customers mostly favoured cluster ID 91 and the least favourite was cluster ID 74.

The distance from the origin to destination being a continuous variable, to have a comprehensive analysis, it has been grouped into three categories:

- "Near” - orig_destination_distance = [0,1000] kms
- "Medium” - orig_destination_distance = (1000,6000] kms
- "Far" - orig_destination_distance> 6000 kms

Most of the customer preferred to travel to a nearby locations or locations with medium distance rather than travelling long distances (4.02%). The analysis does not take Not Applicable (observations with null values) group in the following queries.

Pivoting with the continents, it was revealed that Continent 3 and 5 attracted visitors from “Far” category the most, Continent 6 and 4 attracted customers from “Medium” travel distance category and Continent 2 had customers from nearby locations. One outlier is for continent 1 there is 80% from medium category, but looking into the total bookings, continent 1 has only 5 valid observations, thus can be treated as an outlier. Similar Analysis was done with respect to destination country. The analysis also looks into the hot hotel clusters and it sheds light on predicting the customer behaviour to choose a cluster based on the distance of the destination from the origin.

Family Profiling:

This part of the analysis looks into the distinct values of the customer's family attributes namely, srch_destination_type_id, orig_destination_distance, srch_adults_cnt, srch_children_cnt, is_mobile, channel and hotel_cluster. The total number of family members (adding srch_adults_cnt and srch_children_cnt) being a continuous variable, to have a comprehensive analysis, it has been grouped into three categories:

- "Alone" – total family member = 1
- "Couple" - total family member = 2
- "Fam&Frnds" - total family member > 2

After categorizing the majority of the customers in the dataset searched for a single person, followed by couples and family and friends. Not Applicable category being a very small percentage, it is not taken into account in the further analysis.

Channel Usage:

Channel 4 draws the maximum traffic with device not being mobile, whereas channel 6 and 7 draws the least. When the mobile device is considered, Channel 0 and 4 draws the highest traffic and Channel 5 and 6 being the lowest.

While comparing family profile to the distance profile, it was revealed all the three categories of family profile searched for a nearby vacation, the "Far" option was the least favourite. But among those numbers, couples favoured "Far" category with respect to the other two., i.e., out of 1000 couples 62 couples preferred vacation at a "Far" category, whereas Fam&Frnds preferred the "Near" one (56 out of 100 Fam&Frnds). When pivoted against the search destination type, Destination type 1 and 6 were clear winner among all the family profiles, least popular were destination type 9 and 7. One stand out thing in this analysis being, the couples preferred destination type 5 more than the customers falling in "Alone" and "Fam&Frnds".

Finally, the analysis looks into the most searched origin and destination combinations. Customers from countries 66 and 205 visit country 50 and 198 a lot, these types of insights are helpful while devising a recommendation algorithm predicting the destination country, as customers from same country seem to be exhibiting similar behaviour. The countries which have the maximum number of bookings per Expedia are 50, 8 and 198, whereas customers travelling outside their countries were mainly from countries 66 and 205. The analysis also looks into customers who tend to visit their own countries a lot, countries within this category are 68 and 77.

Conclusion

Everyone likes their products to be personalized and behave the way they want them to. Given a user, recommender systems aim to model and predict the preference of a product. Expedia wants to take the proverbial rabbit hole out of hotel search by providing personalized hotel recommendations to their users. During that time, Expedia uses search parameters to adjust their hotel recommendations, but there were not enough customer specific data to personalize them for each user.

The Expedia dataset was analyzed in Spark Ecosystem using SQL and Dataframe API to gather insights which will be helpful to built the recommendation model personalizing the prediction for each customer. The dataset can be basically divided into few categories, namely, customer behavior on the site based on search timeframe, their geolocation which is an important characteristic to form clusters, mode of device and channels used, and, destination and family profiling. The target variable, hotel_cluster needs to be determined based on these analyses. The analysis was successfully implemented using Expedia's dataset even though most of the data was anonymized which restricted the amount of feature engineering we could do. The most important and challenging part of the report was to create and extract meaningful features out of the 38 million data points provided. The exploration of data took a long time given the size of data and it helped to extract features that would seem to have high impact on predicting the hotel clusters.

To extend this analysis report, the dataset needs to be modelled as a multiclass classification problem and build variations of classic support vector machines (SVMs) and decision tree classifiers to predict the 5 most likely hotel groupings from which a user will book a hotel. To use feature selection techniques to select optimal feature subsets, the above analysis will play a main part, then build a unique combined ensemble model achieving a higher precision and recall than either individual model alone.

To improve the recommendation, few other features could also be included within the dataset, such as the demographics of the customer (age, gender), previous destinations, price range criteria, which will help the model to narrow down the predictions, providing a more personalized approach.

Dataset:
https://www.kaggle.com/c/expedia-hotel-recommendations/data (Train Dataset)

Video:
https://drive.google.com/file/d/1L42wepjZfFKKEbS9pVgUViS6wGSZp9hL/view?usp=sharing