**THE EXPEDIA DATASET**

Expedia provided data-set that captured the logs of user behavior. These provides various details as to where the customer was located when he made the search, what was the channel through which the search was made, if it was actually a search or a booking etc. The goal was to predict which “hotel cluster” the user is likely to book, given his search details. These “clusters” have been created by Expedia based on some undisclosed in-house algorithms. But the intuition is that hotels belonging to a cluster are similar for a particular search - based on historical price, customer ratings, geographical locations relative to city center, etc. These hotel clusters serve as good identifiers to which types of hotels people are going to book, while avoiding outliers such as new hotels that don't have historical data.

The training dataset consists of the following columns that includes both, the click events and the booking events. The data is huge in terms of size. The overall size of data is 3.93 GB. The total entries was approx. 376 million and 24 fields. The details of fields are as below:

**Data Dictionary:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Col.No** | **Column name** | **Description** | **Data type** |
| 1 | date\_time | Timestamp | Object |
| 2 | site\_name | ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...) | Int |
| 3 | posa\_continent | ID of continent associated with site\_name | Int |
| 4 | user\_location\_country | The ID of the country the customer is located | Int |
| 5 | user\_location\_region | The ID of the region the customer is located | Int |
| 6 | user\_location\_city | The ID of the city the customer is located | Int |
| 7 | 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 | Float |
| 8 | user\_id | ID of user | Int |
| 9 | is\_mobile | 1 when a user connected from a mobile device, 0 otherwise | Int |
| 10 | is\_package | 1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise | Int |
| 11 | Channel | ID of a marketing channel | Int |
| 12 | srch\_ci | Checkin date | Object |
| 13 | srch\_co | Checkout date | Object |
| 14 | srch\_adults\_cnt | The number of adults specified in the hotel room | Int |
| 15 | srch\_children\_cnt | The number of (extra occupancy) children specified in the hotel room | Int |
| 16 | srch\_rm\_cnt | The number of hotel rooms specified in the search | Int |
| 17 | srch\_destination\_id | ID of the destination where the hotel search was performed | Int |
| 18 | srch\_destination\_type\_id | Type of destination | Int |
| 19 | hotel\_continent | Hotel continent | Int |
| 20 | hotel\_country | Hotel country | Int |
| 21 | hotel\_market | Hotel market | Int |
| 22 | is\_booking | 1 if a booking, 0 if a click | Int |
| 23 | cnt | Numer of similar events in the context of the same user session | Int |
| 24 | hotel\_cluster | ID of a hotel cluster | int |

**Preliminary Observation:**

* Most of the columns have IDs specified with no significant information on what that particular ID signifies.
* Nothing is linearly correlated with the target (hotel\_clusters), meaning we can't use fast machine learning techniques like linear regression.
* There are 100 different clusters, and according to the competition admins, the boundaries are fairly fuzzy, so it will likely be hard to make predictions. As the number of clusters increases, classifiers generally decrease in accuracy.
* Bundling hotel clustering and finding out the relevant information on trend makes more sense than the hotel cluster prediction.
* Should the whole data or only the booking data, excluding the “just click” data. As the booking data will give more information and significant conclusion can be drawn out of it.