Capstone Project: Expedia Hotel Recommendation

### What is the problem you want to solve?

Predict the hotel-cluster for the customer.

### Who is your client and why do they care about this problem?

Client is Expedia in this case and can be used by any Hotel aggregators.

They care about the customization to provide powerful personalized digital experience to the customer.

### What data are you going to use for this? How will you acquire this data?

Data is available at Kaggle, <https://www.kaggle.com/c/expedia-hotel-recommendations>

The following are the data provided:

* **train.csv** - the training set
* **test.csv** - the test set
* **destinations.csv** - hotel search latent attributes
* **sample\_submission.csv** - a sample submission file in the correct format

See next page for the detailed attributes related to each dataset.

### In brief, outline your approach to solving this problem (knowing that this might change later).

* Get and clean the data
* What is Hotel-cluster?
  + The hotels are grouped based on a set of attributes (price, star ratings and distance from city center etc.)
* How to contextualize / personalize for a customer?
  + Has the user booked earlier?
    - If yes then follow that lead
    - Otherwise, use search query info: number of days stay, adult and children count to formulate the rules.
  + Use EDA techniques to visualize the patterns.
* Define and build a predictor model with the above inference.

### What are your deliverables?

* R Code in the github
* Slidedeck

Detailed data fields:

**train/test.csv**

|  |  |  |
| --- | --- | --- |
| **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** | **Checkin 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** | **Numer of similar events in the context of the same user session** | bigint |
| **hotel\_cluster** | **ID of a hotel cluster** | int |

**destinations.csv**

|  |  |  |
| --- | --- | --- |
| **Column name** | **Description** | **Data type** |
| **srch\_destination\_id** | ID of the destination where the hotel search was performed | int |
| **d1-d149** | **latent description of search regions** | double |