Hybrid Travel Recommender System (HTRS)

Data Preprocessing and Exploratory Analysis

Our first step was to clean and pre-process the data and perform exploratory analysis to get some interesting insights into the process of choosing a hotel. The first thing we observed was that there were many users who have only searched for hotels and did not make any reservation. Moreover, the test data had only the users who made a reservation. Thus, we pruned the dataset by removing all the users who did not make any booking as these entries do not provide any indication of which hotel clusters those users prefer and this could possibly have interfered with making predictions.

	Feature	Description
1	date time	Timestamp
2	site name	ID of Expedia point of Sale
3	posa continent	ID of site's continent
4	user location country	ID of customer's Country
5	user location region	ID of customer's region
6	user location city	ID of customer's city
7	orig destination distance	Physical distance between a hotel and a customer

8	user id	ID of user
9	is mobile	1 for mobile device, 0 otherwise
10	is package	1 if booking/click was part of package, 0 otherwise
11	channel	ID of a marketing channel
12	srch ci	Check-in date
13	srch co	Check-out date
14	srch adults cnt	Number of adults
15	srch children cnt	Number of children
16	srch rm cnt	Number of rooms
17	srch destination id	ID of the destination
18	srch destination type id	Type of destination
19	hotel continent	Hotel continent
20	hotel country	Hotel country
21	hotel market	Hotel market
22	is booking	1 if a booking, 0 if a click
23	cnt	Number of similar events in the context of the same user sessiont
24	hotel cluster	ID of hotel cluster

Table 1: Features used for Training

From the remaining entries, we identified the searches by each user belonging to a specific type of destination. This gave us some useful information about which hotel cluster was finally chosen over other hotel clusters explored by the user. One important observation to note is that few users might be travel agents and could explore multiple type of destinations at the same time. This could also be true for few users who are planning multiple vacations at the same time.

That is why we considered the preferences of the users separately for each destination type he/she explored. Also, after a booking was made, subsequent searches by the user were treated separately. We describe aforementioned approach with an example below.

In table 2, the searches made by user 1 are shown. There are 2 types of destinations: 11938 and 8821. Based on the type of destination, we identify the hotel clusters that were rejected/selected.

For destination type 11938, cluster 52 was selected, and clusters 87 and 65 were rejected. Similarly, for destination type 8821, cluster 20 and cluster 35 were rejected. Now, we keep a track of the rejected clusters.

User ID	Hotel Cluster	Destination Type	Booking
54261	65	11938	0
54261	87	11938	0
54261	52	11938	1
54261	20	8821	0
54261	30	8821	0

Table 2: Subset of training set

Also, we use the check-in and check-out dates to find the duration of the stay for each of the entries in the training set. From the Figure 1, we see that all the clusters seem equally likely when the duration is short, but as the duration increases, certain hotel clusters are preferred over the others. This seems to be a good feature to identify the hotel cluster user would choose.

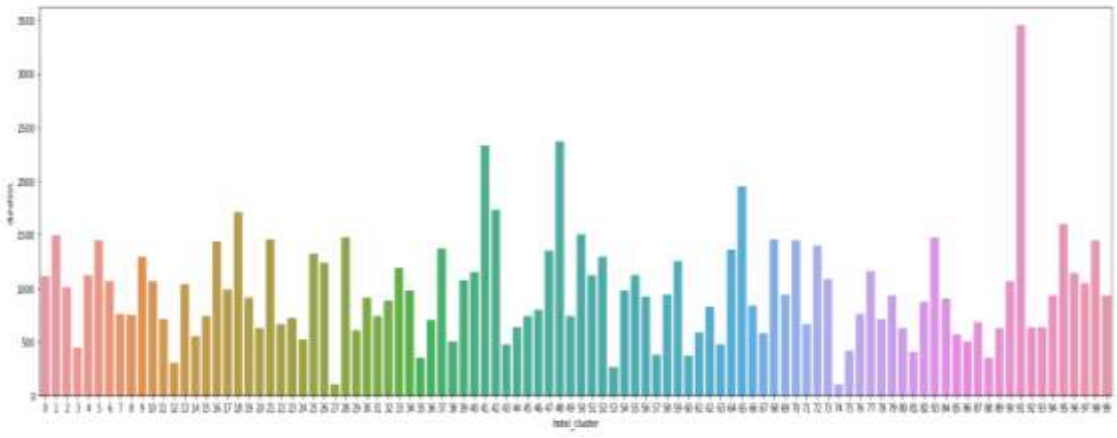

Figure 1: Duration of stay vs hotel cluster

Furthermore, as per the details provided by Expedia on the competition page, hotels tend to change their cluster seasonally. To capture this variation, we also included one-hot representation of the month from which user is seeking to start his/her stay. Finally, we make use of the latent features of the destinations provided in the dataset.

Finally, we make use of the latent features of the destinations provided in the dataset. However, since we have 149 latent features for each destination, we decided to apply PCA to extract the most relevant dimensions.

Next, we visualize the correlation matrix between the features of the training set in Figure 2 and observe and observe following things:

- ***hotel_cluster*** does not seem to have a strong (positive or negative) correlation with any other feature. Thus, methods which model linear relationship between features might not be very successful.

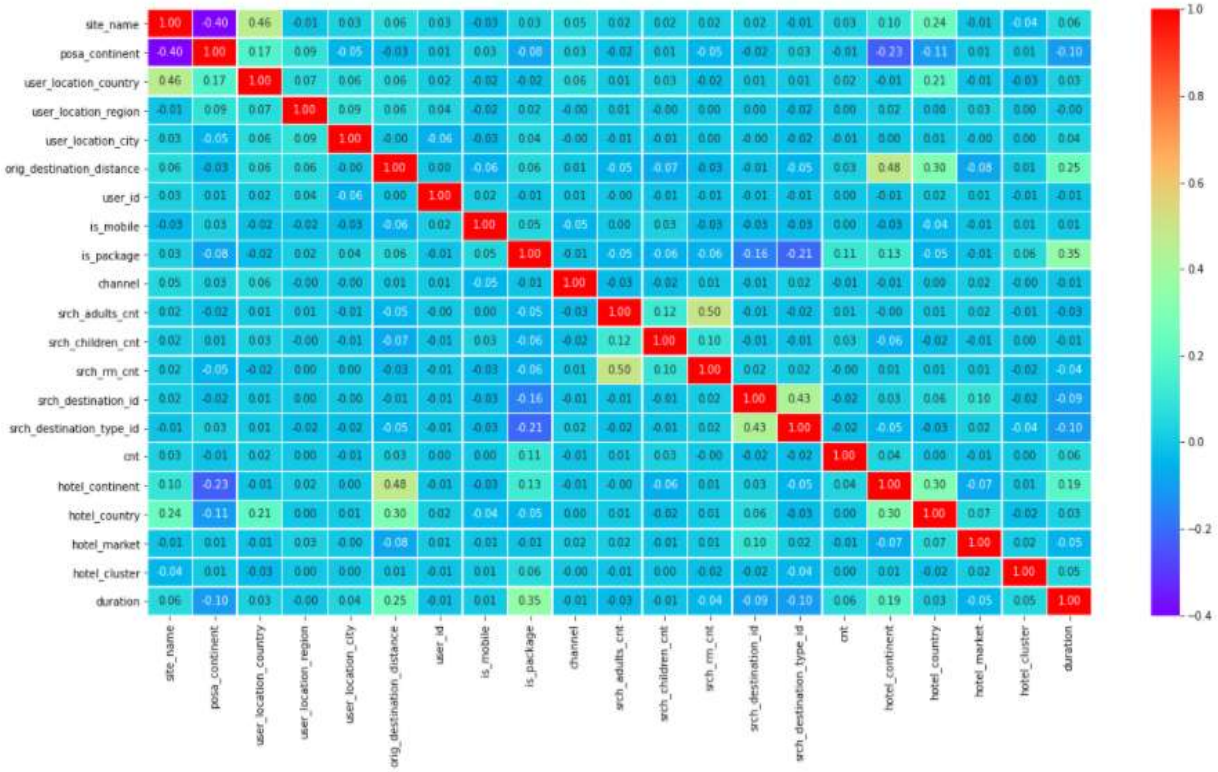

Figure 2: Correlation matrix of features

- ***orig_destination_distance*** has a positive correlation with duration (constructed using srch_ci and srch _co), which means people who are planning for a long trip tend to go far away from the place of origin.

- ***hotel_continent*** and ***posa_continent*** (which is from where the booking is done) are negatively correlated. This means that people tend to go to continents different from theirs for vacations.

- ***duration*** seems to have a strong positive correlation with ***is_package***. This means that people who tend to book hotel for longer duration usually choose hotels as a part of a package.

- ***srch_destination_id*** has a strong correlation with ***srch_destination_type_id***. This is expected as each destination would have an associated type; for example, vacation spot, city, etc.

- ***duration*** is also positively correlated with ***hotel_continent*** which means certain continents are preferred for longer duration of stay.

- ***srch_rm_cnt*** has a very strong correlation with ***srch_adults_cnt***, and to an extent, with ***srch_children_cnt*** also. This is expected as people tend to take rooms based on how many families/ couples are there.

Feature Engineering

- We applied some "Feature Engineering Techniques" to prepare input data to produce results. The date time, check-in date and check-out date columns can't be used directly. Therefore, some features such as month, year, number of searches, trip duration and number of bookings, solo trip or family trip, were extracted from data.

- We have 149 latent features for each destination; we decided to apply PCA to extract the most relevant dimensions.

- Expedia on the competition page, hotels tend to change their cluster seasonally. To capture this variation, we also included one-hot representation of the month from which user is seeking to start his/her stay.