Expedia Hotel Recommendations

****

**Aim of the project**

Goal of this competition is to predict the booking outcome (hotel cluster) for a user event, based on their search and other attributes associated with that user event.

**Data provided**

The train and test datasets are split based on time: training data from 2013 and 2014, while test data are from 2015. The data are split base on time as well. Training data includes all the users in the logs, including both click events and booking events. Test data only includes booking events.

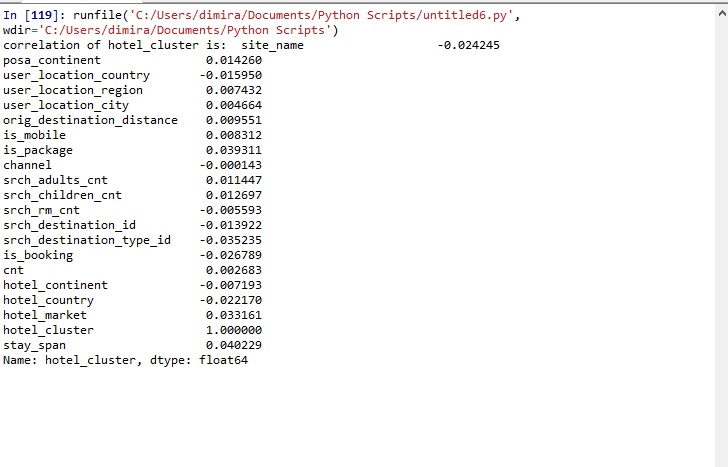
destinations data consists of features extracted from hotel reviews text.

**Exploring and cleaning the data**

Train data has additional column hotel\_cluster that is the goal of the project to predict. Train data has more than 37.0 million rows and test data has more than 2 million rows. Data was too big to handle on an 8gb RAM as train data took too much time to load. We created a new train data csv file with 50,000 lines using random sampling.

In train and test data lot of columns has NaN value, so first task is to fill the value of columns with either the mean of the column or minimum value in the column.

After cleaning the data let’s explore the co-relation of hotel\_clusters with other data and the result show less than 1% of correlation between hotel\_clusters and other variables.



The correlation shows that hotel\_clusters is not linearly related to any data, so any linear machine learning algorithms like linear regression will not work on the data.

**Generating new data set and extracting feature using PCA and Gradient boost variable importance**

On non-linear data the best and simplest model is decision tree, but before applying any model let’s explore what other data can be created from preliminary data set and what data can be removed. In destination csv file d1-d149 columns have latent descriptions of the region, so instead of using all the columns for the region, better to use 3 columns for the data. PCA (Principal component analysis) will reduce the number of columns in a matrix while trying to preserve the same amount of variance per row. After creating new data on destination region and srch\_destination\_id, join the new data with the train data. A new span data in hours can be created from check-in and check-out data.

After creating new data set datetime, check-in and check-out are of no use so removing this data from the dataset.

Let’s explore further that which features are more important, so we feed all our train data into Gradientboost classifier variable importance with hotel\_clusters as predictor variable. From the output we can visualize that srch\_adults\_cnt, srch\_children\_cnt, srch\_rm\_cnt, is\_mobile, is\_package, channel,cnt value equals to 0 showing these parameters we can remove from the data set.

All the destinations and srch\_ destination\_id parameters show value more than 0.1 mean that hotel\_clusters value has dependencies on these parameters.

**Dividing data into train and validation data and fitting a Random Forest classifier**

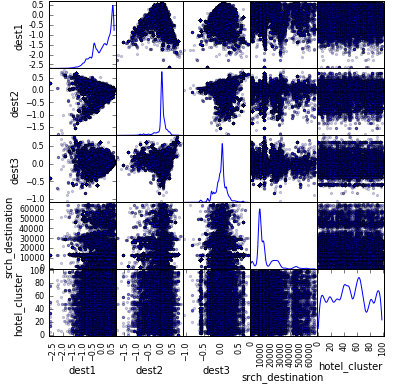
Now we have generated a new data set by removing datetime and other redundant features, we divide our train data into two parts train data and validation data. Train data is 80% while validation data is 20%.

Now we will train our data using a Random Forest Classifier, we will use 3-fold cross validation across the training set to generate a reliable error estimate. We’ll first initialize the model and compute cross validation scores on train data.

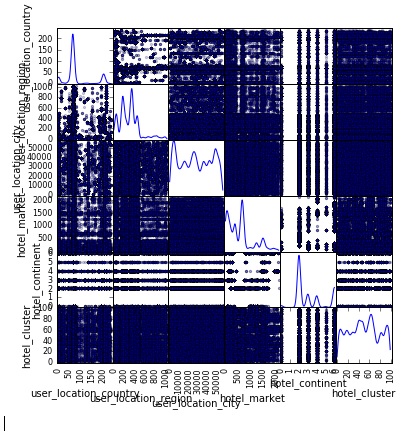


**Clustering destination and visitor location data**

From the Random forest classifier poor result, it is clear that machine learning will not give better result in this problem, so let’s explore how destination and user data are related to hotel\_clusters.



Scatter graph between srch\_destination, dest1,dest2, dets3 and hotel\_cluster

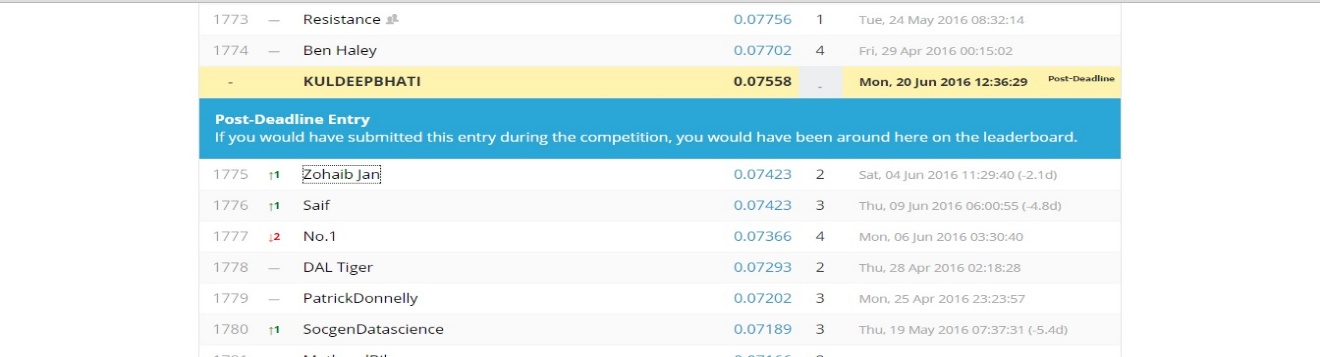
****

Scatter graph on user location,country,location, hotel market and hotel\_clusters

From the data it is clear that certain destinations are more searched and particular destinations are more searched, same can be said for user\_location and hotel\_market. So If we group destination data and user \_location data to match data with particular hotel\_cluster we can make better predictions. So our result on predicted result came at 0.22 on validation data set.



**Kaggle result on test data**

****

**Conclusion**

Although result was not extra ordinary but further work could be done on sub setting particular destination data, user\_location and hotel\_clusters. Finding correlation between sub set data and applying machine learning on that.