Capstone Project Final Report

Prediction of popular hotel class based on historical browsing data

Mentor: Ryan Rosario

Mentee: Tanzina Zaman

1 INTRODUCTION

1.1 Problem definition:

The aim of this project is to predict the popular hotel destinations based on hotel location, users' location, past browsing history etc. The dataset used in this project is collected from a Kaggle competition sponsored by Expedia. The client, Expedia wants to predict which hotel class is booked by the customer, based on customer's previous bookings and clicks. A click means a user clicked a link to explore hotel details in a hotel information site page. The service provider has predefined hotel class based on historical price, customer star ratings, geographical locations relative to city center, etc. New hotels which have no historical data are not considered in this case.

1.2 Business need:

Hundreds and thousands of people look for hotel either for work purpose or for vacation and numerous options appear when potential customers search for hotel that will fit their needs. The aim of this project is to help the client to personalize their recommendation system which will predict the popular hotel class a potential customer may chose based on the previous browsing history. This would make the hotel selection process easier for potential customers as they would get recommendation for hotels that will match their need and guide them toward right direction.

2 DATA DESCRIPTION

For this project, the 'training' dataset is used from the sponsored competition on Kaggle[1]. The dataset has observations from year 2013 and 2014 which is considered as the parent source of information in this project. This dataset contains over 37 millions of observation and 24 fields. Among the 24 fields, the hotel class label is considered as a target and rest are used as feature attributes. These fields include a timestamp as well as the following information: Expedia point of sell, customer country, region and city, the distance between the customer and the searched hotel,

whether or not the customer was browsing through a mobile device etc. These fields provide the information related to hotel search. Not every fields of the dataset have been used in the project. To understand the data fields, a quick glance at the Expedia website would be helpful.

The first thing that draws any customers attention while using Expedia website is 'Going to' tab. This tab provides the information of destination type, the corresponding continent, country and the market of a hotel. The second important feature of the site are 'check in' and 'check out' tab, which maps to two more features of the dataset, the date when the customers checked in and checked out respectively. The three tabs next to 'check out' tab, map to the number of adults, children and rooms specified during the hotel search. If the customer uses 'Add a flight' tab, then the transaction is considered as a package.

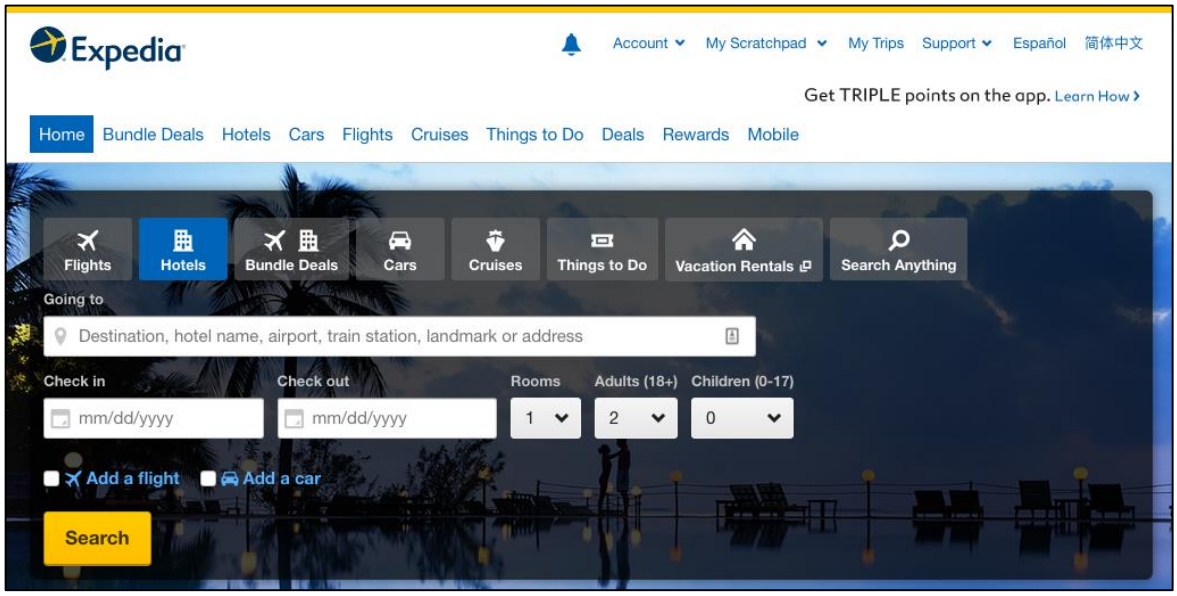

Figure 1: An example of Expedia interface to search or book hotels online

3 EXPLORATORY DATA ANALYSIS

The hotels in the dataset are divided into 100 class among which, ten most frequently booked hotels have been used in this project.

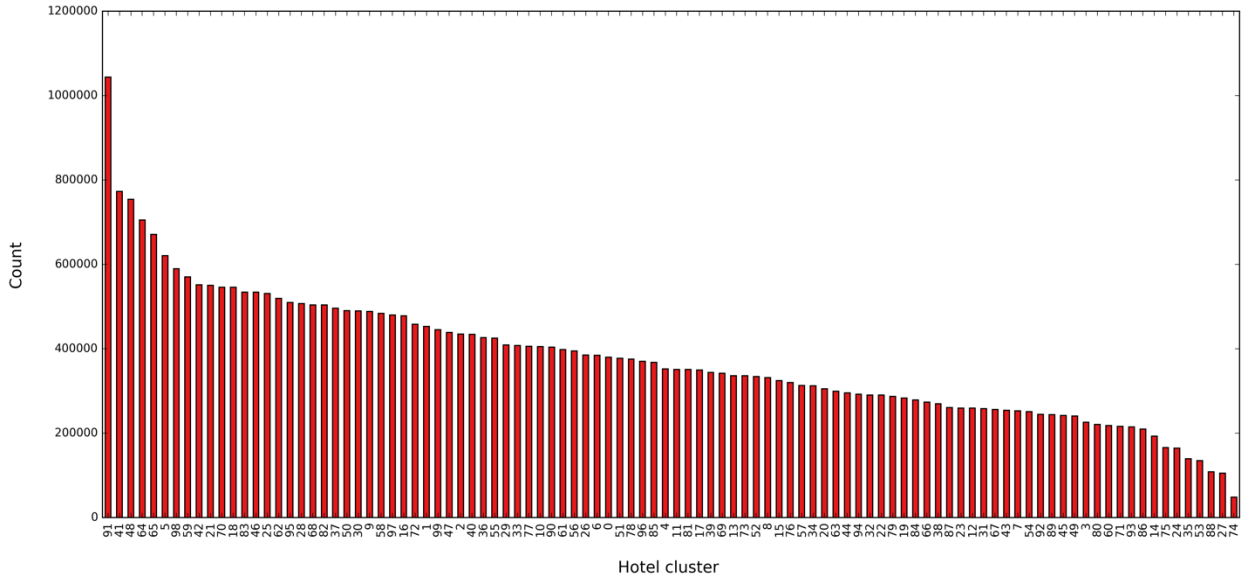

Figure 2: Frequency of hotel class

Among the top ten classes, the booking outcomes are not equally distributed. Such trend in the dataset is quite natural as people tend to browse a lot before they finally decide to book any hotel. The below bar-chart shows the percentage of 'Booked' and 'Not booked' events.

Figure 3: Percentage of 'Booked' and 'Not booked' events in the dataset

People tend to search for hotels almost throughout the year, but the figure 4 illustrates that the search trend gradually increases during summer season and lasts till holiday season in the end of year.

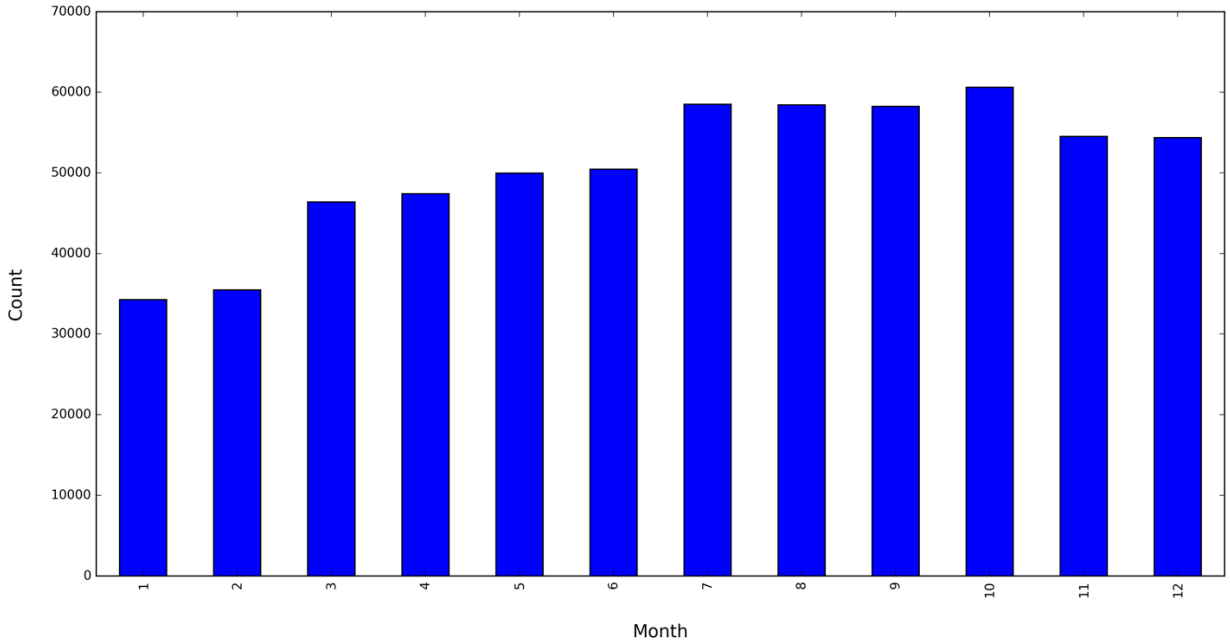

Figure 4: Trend of hotel booking events varies over month.

People also tend to search more during the start of a week, rather than end of the week. In figure 5, the trend of booking events on the days of week is presented.

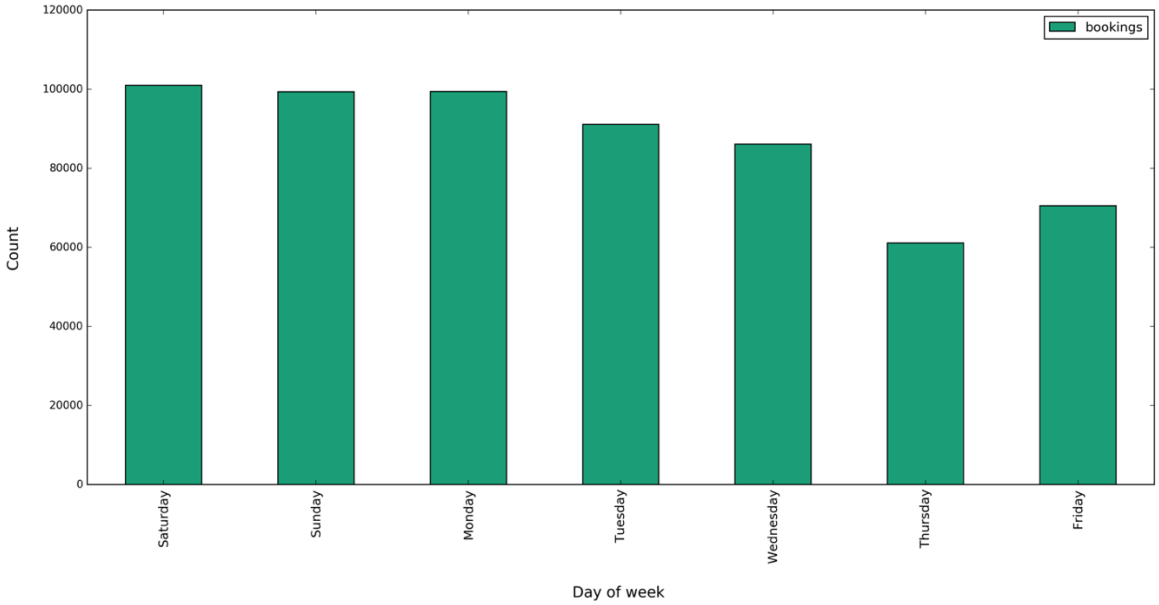

Figure 5: Trend of booking event over the days of week

4 DATA ANALYSIS

The aim of the project is to predict which popular hotel the customer will choose based on historical data which includes browsing and booking records, demographic information of customers, hotel details etc. As the purpose is to predict whether the customer will select a specific hotel class or not, the ultimate target is to identify the class label. Logistic regression, one of the most popular techniques to identify class is used here.

The data analysis method can be described into 2 steps. First, to train 10 different logistic regression models for each of the hotel class selected. Once all the models are trained, the next step is to test whether the trained model can predict the ID of actual hotel class in test set correctly. The probability of assigning a class label to any observation in test set is calculated and the highest probability indicates the true assignment. Then the accuracy of model can be determined comparing the actual and predicted class.

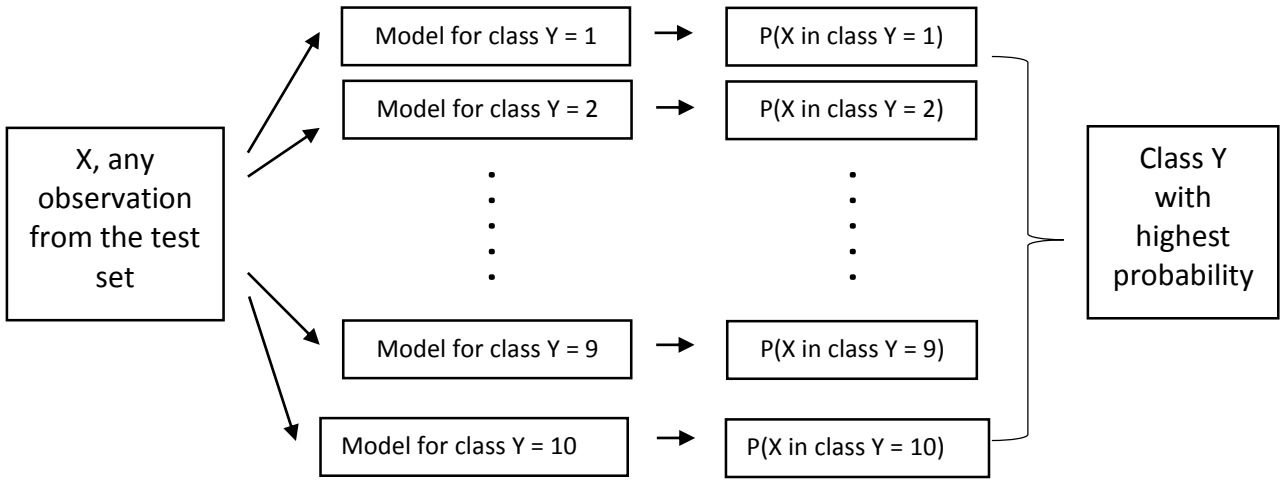

Figure 6: process flow for hotel class prediction

4.1 Data wrangling:

Couple of preprocessing techniques are followed before training the models while ten of most popular hotel classes are chosen for analysis. The ID of the classes are 64, 65, 98, 59, 5, 41, 42, 48, 21 and 91. Due to selecting top ten class, the dataset has been reduced to 6.8 million observations from 37 million. As the given dataset is scaled down to a smaller set, it is now split into training and testing set in 70-30 ratio.

The feature attributes of the dataset can be divided into two types: categorical feature and quantitative feature. The continent, country, city etc. of customers and hotels are categorical feature whereas, the booking status, the number of rooms booked, number of adults and children etc. are quantitative variables. Besides, the timestamp in the dataset is provided in the format of 'YYYY-MM-DD hh:mm:ss', which is hard to interpret during training any classifier. This feature captures the information regarding year, month, day and day of the week for each observation. For this reason, 4 more features were created from every timestamp and the total number of attribute features increased to 27.

Here, the categorical and quantitative features are treated separately. If categorical values are converted to numerical, this would not be able to capture true meaning. For this reason, for each categorical feature, dummy variables are created and then combined together into a singled data frame. But some labels of some categorical features from training set are not present in test set, and similar things have been noticed for test set too. To make both training and test set consistent, the labels were dropped from training set, which are absent in test set. Then, numerical features are added to the newly formed data frame containing categorical variables. In this way, the volume of data has been increased. To handle the memory issue, sparse matrices have been used. Thus the training and test set are ready for the next step.

4.2 Method

After data management and feature generation, logistic regression models are created with stochastic gradient descent (SGD) learning for each hotel class, with 12 features out of 27. These 12 features are related to the location of customers, location of hotels, booking history, number of people considered for reservation. Features which has missing data are not considered here. During training the logistic regression model, for each hotel class in the training set is converted into a binary classification set. For example, for hotel class ID 64, the observations with hotel class 64 are considered as 1 and rest are 0. This step is repeated 10 times for each of the hotel class and 10 different classifiers are trained.

Then for each model, evaluation metrics such as precision, recall, mean F1 score, true positive rate and false positive rate are calculated. Using precision and recall score, precision-recall curve is plotted for each individual classifier. Similarly, using true positive rate and false positive rate, receiver operating characteristics (ROC) curve is generated and associated area under the curve (AUC) for each class is calculated. Finally, the predictive performance of ensemble classifier is estimated from the confusion matrix, which is created for 10 individual classifiers.

5 RESULT

The ability of any model to classify the hotel class with a good accuracy is reflected from the measure of AUC of ROC curve, precision-recall curve and confusion matrix. The plotted ROC curve and associated AUC for each of the classifier is a reliable metric to evaluate how accurately the model could predict the hotel class. ROC is a potentially powerful metric as it is invariant against class skew of the dataset. In ROC space, the false positive rate (FPR) is on the x-axis and the true positive rate (TPR) on the y-axis.

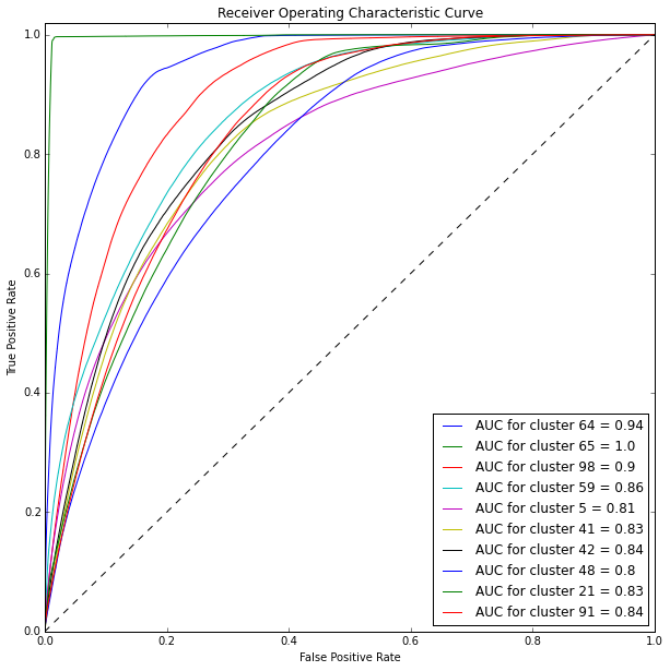

Figure 7: ROC curve for all hotel class

The FPR measures the fraction of negative examples that are misclassified as positive. The TPR measures the fraction of positive examples that are correctly identified. For perfect prediction, the

AUC for ROC curve would be 1. Here, all the models show the ability for nearly perfect prediction and specifically, class ID 65 has AUC of 1, which is the maximum among all the classifier models.

In this dataset, the booking outcome is largely skewed and Precision-recall (PR) curve is often used in cases with a large skew in the class distribution. In PR space, recall is on x-axis and precision on y-axis. Recall is same as TPR but precision measures the fraction of examples classified as positive that are truly positive. Typically, precision and recall are inversely related. However, the goal in precision-recall curve is to be in the upper right corner, whereas, for ROC is to be in upper left corner. In this case, not all the models were successful to achieve that goal.

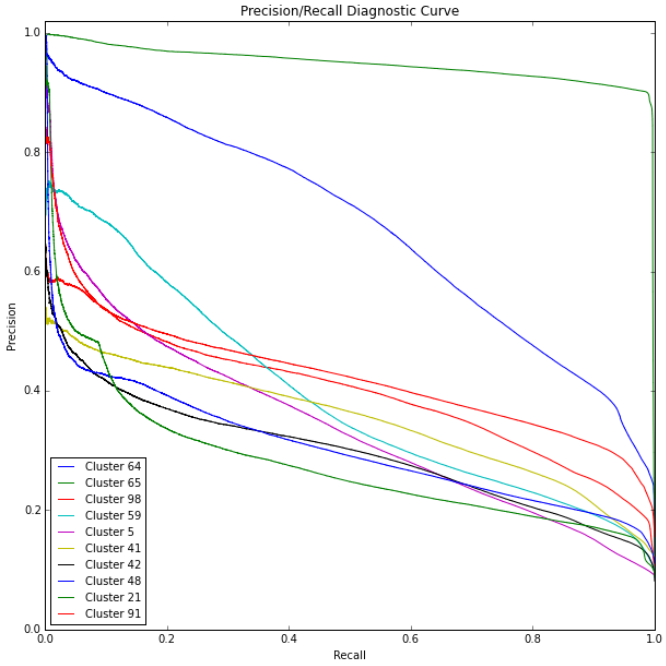

Figure 8: Precision-recall curve for all hotel class

Both ROC and PR curves above show the evaluation score for each classifier. The confusion matrix below describes the performance of the ensemble of 10 classifiers. A confusion matrix contains information about actual and predicted classification done by classifier. Here, the ensemble classifier decides which class to assign for each test row based on the computed probability. Along the diagonal, from upper left to lower right indicates the truly predicted class by the classifier. From this confusion matrix, it is clear that the ensemble classifier could classify

the hotel class for hotel ID 21 with maximum accuracy, however for hotel ID 59 and 5, there are some misclassification.

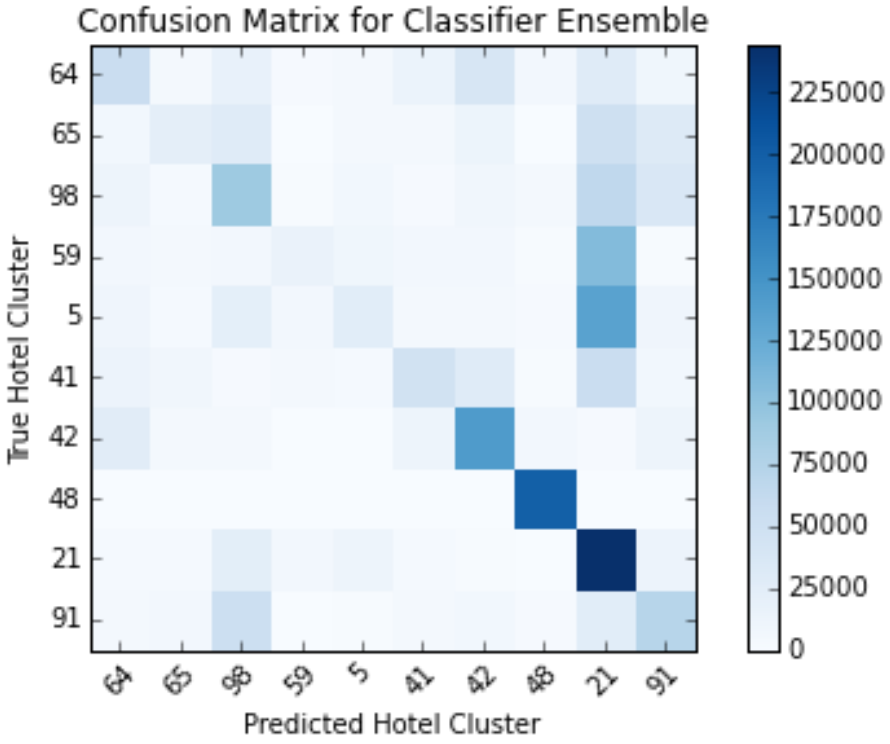

Figure 9: Confusion matrix for all hotel class

6. Conclusion:

In this analysis, we looked at a potential way to propose a method to make it easier for service provider to predict hotel class for a customer who has been using their website for searching hotels. Currently the service provider, Expedia is using search parameters to adjust their hotel recommendation, but unable to personalize enough for each user. Solving this multiclass classification problem will help the service provider to make the search results more personalized. Other than hotel recommendation, the service provider may also use this model to place advertisements associated to the hotel location like vacation package, restaurant deals etc. into their website.

Appendix A

[1]The dataset can be downloaded from the following link:

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

Figure 1: The example of training set

	0	1	2
date_time	2014-11-22 22:00:24	2014-10-13 15:25:05	2014-07-10 23:26:18
site_name	30	2	25
posa_continent	4	3	2
user_location_country	195	66	23
user_location_region	991	174	48
user_location_city	47725	46432	4924
orig_destination_distance	NaN	110.51	NaN
user_id	1048	3313	3972
is_mobile	1	0	1
is_package	0	0	0
channel	9	1	9
srch_ci	2015-06-26	2014-10-24	2014-08-13
srch_co	2015-06-28	2014-10-26	2014-08-14
srch_adults_cnt	2	2	2
srch_children_cnt	0	0	1
srch_rm_cnt	1	1	1
srch_destination_id	8803	11835	8278
srch_destination_type_id	1	1	1
is_booking	0	0	0
cnt	1	1	1
hotel_continent	3	2	2
hotel_country	151	50	50
hotel_market	69	633	368
hotel_cluster	59	17	63

Table 1: Field name and description of train / test set

	Field name	Description
1	date_time	Timestamp
2	site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)
3	posa_continent	ID of continent associated with site_name
4	user_location_country	The ID of the country the customer is located
5	user_location_region	The ID of the region the customer is located
6	user_location_city	The ID of the city the customer is located
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
8	user_id	ID of user
9	is_mobile	1 when a user connected from a mobile device, 0 otherwise
10	is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise
11	channel	ID of a marketing channel
12	srch_ci	Check-in date
13	srch_co	Checkout date
14	srch_adults_cnt	The number of adults specified in the hotel room
15	srch_children_cnt	The number of (extra occupancy) children specified in the hotel room
16	srch_rm_cnt	The number of hotel rooms specified in the search
17	srch_destination_id	ID of the destination where the hotel search was performed
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	Numer of similar events in the context of the same user session
24	hotel_class	ID of a hotel class