**Project** ​ ​**Progress Report**

Group 5

**Submitted Date: 10/12/2016**

**Prepared by:**

Aadesh Pai

Achintya Sen

Dipendu Chanda

Mrunmayi Modak

Utkarsh Verma

**Data Set Information:**

Expedia is one of the world’s largest online travel agency (OTA) and powers search results for millions of travel shoppers every day. In this project we are using Expedia’s dataset that includes shopping and purchase data as well as information on price competitiveness. The data are organized around a set of “search result impressions”, or the ordered list of hotels that the user sees after they search for a hotel on the Expedia website. In addition to impressions from the existing algorithm, the data contain impressions where the hotels were randomly sorted, to avoid the position bias of the existing algorithm. The user response is provided as a click on a hotel or/and a purchase of a hotel room.

**Attribute Information:**

The following table contains the information about the attributes used.

|  |  |  |  |
| --- | --- | --- | --- |
| ***S.No***  ***.*** | **Column Name** | **Data**  **Type** | **Description** |
| *1* | srch\_id | Integer | The ID of the search |
| *2* | date\_time | Date/tim  e | Date and time of the search |
| *3* | site\_id | Integer | ID of the Expedia point of sale (i.e.  Expedia.com, Expedia.co.uk, Expedia.co.jp,  ..) |
| *4* | visitor\_location\_country\_id | Integer | The ID of the country the customer is located |
| *5* | visitor\_hist\_starrating | Float | The mean star rating of hotels the customer has previously purchased; null signifies there is no purchase history on the customer |
| *6* | visitor\_hist\_adr\_usd | Float | The mean price per night (in US$) of the hotels the customer has previously purchased; null signifies there is no purchase history on the customer |
| *7* | prop\_country\_id | Integer | The ID of the country the hotel is located in |
| *8* | prop\_id | Integer | The ID of the hotel |
| *9* | prop\_starrating | Integer | The star rating of the hotel, from 1 to 5, in increments of 1. A 0 indicates the property has no stars, the star rating is not known or cannot be publicized. |

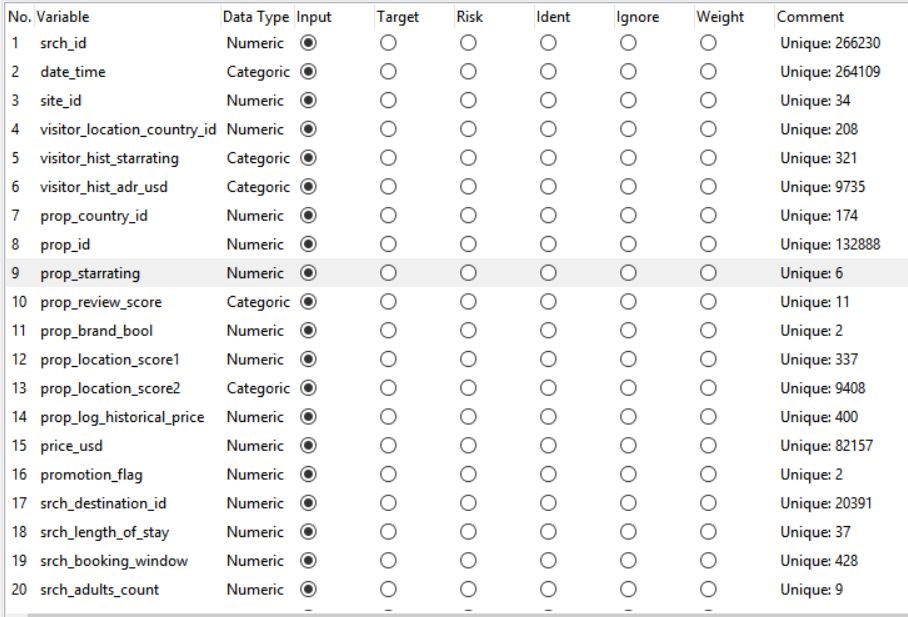
|  |  |  |  |
| --- | --- | --- | --- |
| *10* | prop\_review\_score | Float | The mean customer review score for the hotel on a scale out of 5, rounded to 0.5 increments. A 0 means there have been no reviews, null that the information is not available. |
| *11* | prop\_brand\_bool | Integer | +1 if the hotel is part of a major hotel chain;  0 if it is an independent hotel |
| *12* | prop\_location\_score1 | Float | A (first) score outlining the desirability of a hotel’s location |
| *13* | prop\_location\_score2 | Float | A (second) score outlining the desirability of the hotel’s location |
| *14* | prop\_log\_historical\_price | Float | The logarithm of the mean price of the hotel over the last trading period. A 0 will occur if the hotel was not sold in that period. |
| *15* | position | Integer | Hotel position on Expedia's search results page. This is only provided for the training data, but not the test data. |
| *17* | price\_usd | Float | Displayed price of the hotel for the given search. Note that different countries have different conventions regarding displaying taxes and fees and the value may be per night or for the whole stay |
| *18* | promotion\_flag | Integer | +1 if the hotel had a sale price promotion specifically displayed |
| *19* | gross\_booking\_usd | Float | Total value of the transaction. This can differ from the price\_usd due to taxes, fees, conventions on multiple day bookings and purchase of a room type other than the one shown in the search |
| *20* | srch\_destination\_id | Integer | ID of the destination where the hotel search was performed |
| *21* | srch\_length\_of\_stay | Integer | Number of nights stay that was searched |
| *22* | srch\_booking\_window | Integer | Number of days in the future the hotel stay started from the search date |
| *23* | srch\_adults\_count | Integer | The number of adults specified in the hotel room |
| *24* | srch\_children\_count | Integer | The number of (extra occupancy) children specified in the hotel room |
| *25* | srch\_room\_count | Integer | Number of hotel rooms specified in the search |
| *26* | srch\_saturday\_night\_bool | Boolean | +1 if the stay includes a Saturday night, starts from Thursday with a length of stay is less than or equal to 4 nights (i.e. weekend); otherwise 0 |
| *27* | srch\_query\_affinity\_score | Float | The log of the probability a hotel will be clicked on in Internet searches (hence the values are negative) A null signifies there are no data (i.e. hotel did not register in any searches) |
| *28* | orig\_destination\_distance | Float | Physical distance between the hotel and the customer at the time of search. A null means the distance could not be calculated. |
| *29* | random\_bool | Boolean | +1 when the displayed sort was random, 0 when the normal sort order was displayed |
| *30* | comp1\_rate | Integer | +1 if Expedia has a lower price than competitor 1 for the hotel; 0 if the same; -1 if Expedia’s price is higher than competitor 1; null signifies there is no competitive data |
| *31* | comp1\_inv | Integer | +1 if competitor 1 does not have availability in the hotel; 0 if both Expedia and competitor 1 have availability; null signifies  there is no competitive data |
| *32* | comp1\_rate\_percent\_diff | Float | The absolute percentage difference (if one exists) between Expedia and competitor 1’s price (Expedia’s price the denominator); null signifies there is no competitive data |

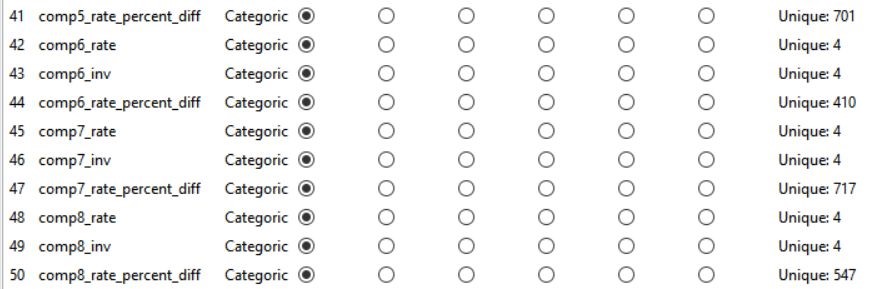
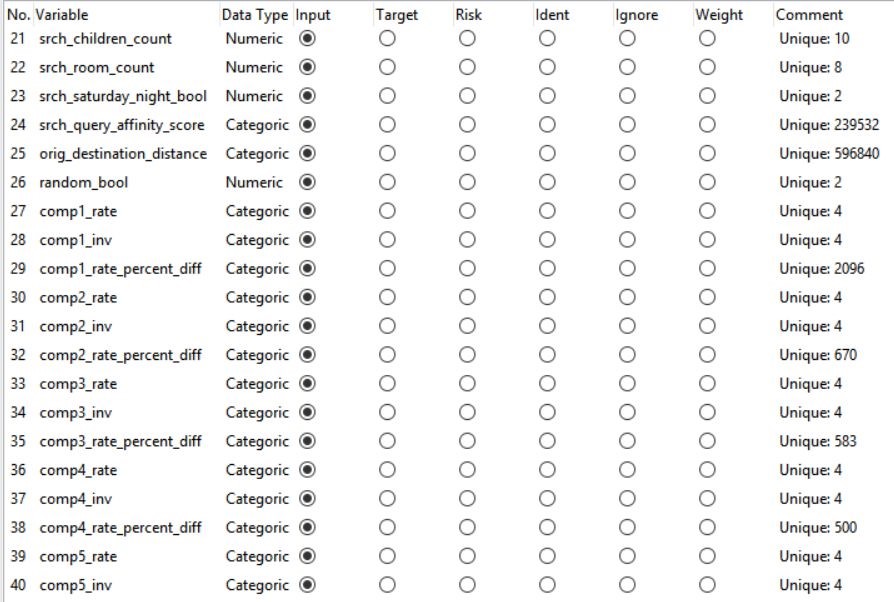
**Number of Observation:**

Dry Run on Test data:6622629 observation Original Train data file 10 mil (approx.)

**Data Summary:**

Input Variables





**Individual Variable Exploratory Analysis:**

Using a sample dataset from the train dataset we were able to perform a descriptive analysis. Please find the findings below regarding the same. **Distribution of continuous variables:**

Here we are trying to find the distribution of numeric variables using histogram. Below are the variable descriptions and the distribution chart:

1. Property Star Rating

|  |  |  |
| --- | --- | --- |
| prop\_starrating | Integer | The star rating of the hotel, from 1 to 5, in increments of 1. A 0 indicates the property has no stars, the star rating is not known or cannot be publicized. |



From the distribution it can be inferred that hotels are booked depending on star rating. Three stars and 4 stars hotels are mostly preferred.

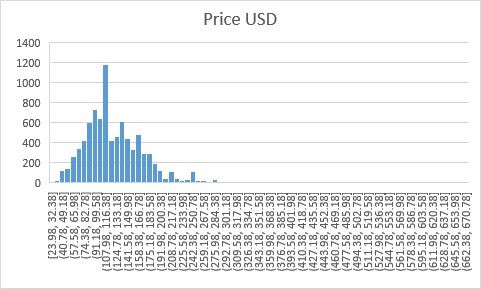
1. Property Review Rating

|  |  |  |
| --- | --- | --- |
| prop\_review\_score | Float | The mean customer review score for the hotel on a scale out of 5, rounded to 0.5 increments. A 0 means there have been no reviews, null that the information is not available. |



1. Price USD

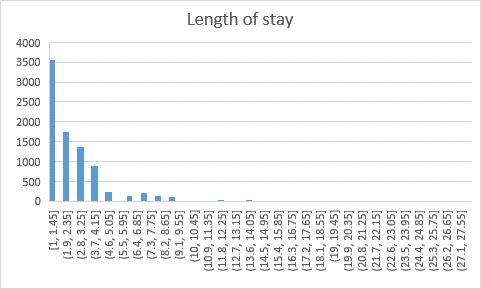
|  |  |  |
| --- | --- | --- |
| price\_usd | Float | Displayed price of the hotel for the given search. Note that different countries have different conventions regarding displaying taxes and fees and the value may be per night or for the whole stay |



Mostly price are in range of 40$ - 285$ per night.

1. Length of stay

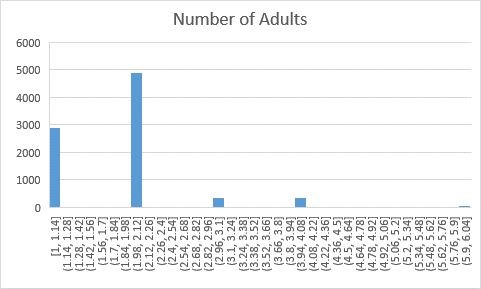
|  |  |  |
| --- | --- | --- |
| srch\_length\_of\_stay | Integer | Number of nights stay that was searched |



It can be inferred that mostly the hotels are getting book for one to three days, maximum for 13 days.

1. Number of Adults

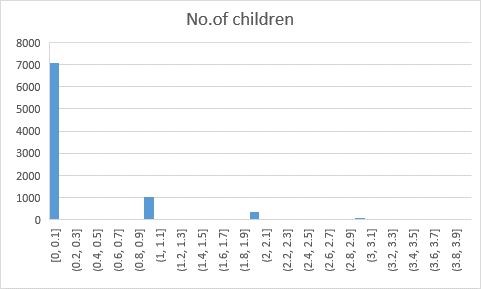
|  |  |  |
| --- | --- | --- |
| srch\_adults\_count | Integer | The number of adults specified in the hotel room |



Maximum 4 Adults can accommodate in one room.

1. Number of Children

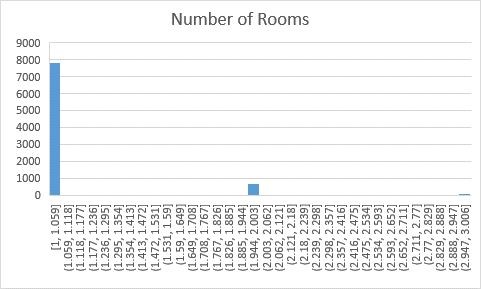
|  |  |  |
| --- | --- | --- |
| srch\_children\_count | Integer | The number of (extra occupancy) children specified in the hotel room |



Mostly one child is accompanying in one room with 3 being the maximum.

1. Number of Rooms

|  |  |  |
| --- | --- | --- |
| srch\_room\_count | Integer | Number of hotel rooms specified in the search |



Almost all the search is directed to single rooms in hotel with three being the maximum but few.

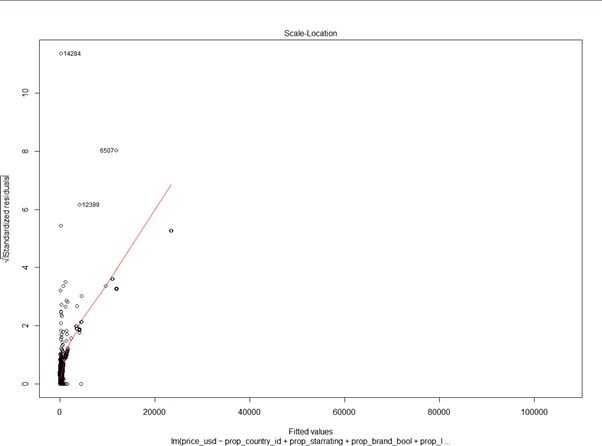
**Objective:**

After analyzing the test data we were able to discover varied possible models to show prediction based on multiple target variables. Performing a descriptive analysis we were able to find key factor variables which play an important role in deriving predictions relating the model. We have also plotted the distributions of the continuous numeric predictor attributes using the histograms in order to understand the spread of these attributes in our dataset.

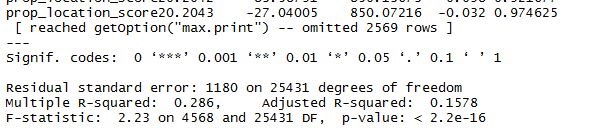
As such, having the best ranking of hotels (“sort”) for users with the best integration of price competitiveness gives an OTA the best chance of winning customer satisfaction. The primary goal of this analysis is to improve the OTA performance in fetching the best results compared to the competitor rate difference.

We have modelled few linear regression models in finding insights in the data. Below are the few mentioned: **Regression model 1:**

*fit<-lm(price\_usd~prop\_country\_id+prop\_starrating+prop\_brand\_bool+prop\_location\_sco re1+prop\_location\_score2+promotion\_flag,data = test\_data1,rm.na=TRUE)*



This shows the prediction of price\_usd w.r.t. prop\_country\_id, prop\_starrating, prop\_brand\_bool, prop\_location\_score1, prop\_location\_score2, promotion\_flag variables.



As the data is very large we are working on the best fitting attributes and unbiased sample generation by random sampling. We were able to reduce the sample size to 30 thousand from 10 million instances. Multiple characteristic of the data is yet to be found.

**Summary:**

We have preprocessed the data and run a basic regression model to predict the price of the hotel booking based on multiple models. We will do further iteration of linear regression variable by variable until high correlation is obtained. We will perform cluster analysis as well as run decision trees algorithm. The primary process flow is still being constructed to find insight into the data.

**Divide the dataset:**

We started by dividing our dataset into Train, Validation and Test set with 70, 15, 15 ratio.

**Preprocessing:**

#We created a set with equivalent number of gross\_bookings\_usd values, which is our final Y variable.

# New dataframe with only the useful columns

#Changing the value from NULL to NA's

#Impute NA values of these 2 columns. : prop\_location\_score2 and prop\_review\_score

#Find Correlations of all the input variables and removed 2 correlated column.

**Models:**

#Decision Tree

#Random forest

#Linear Model

# Logistic Regression