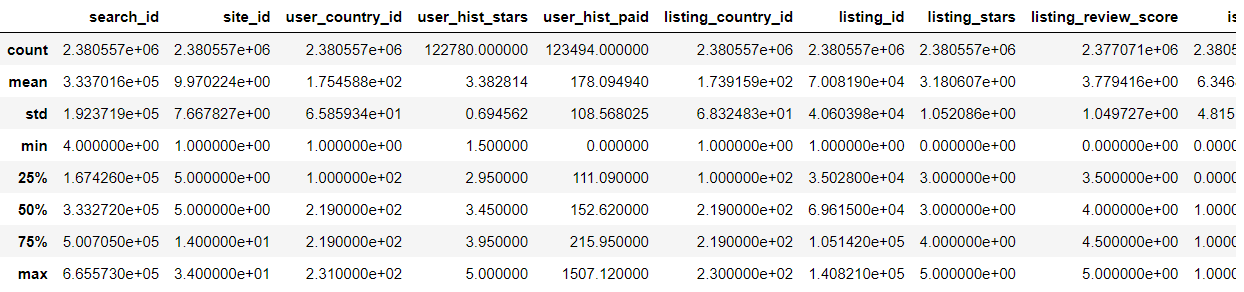
**Approach & Findings**

The data set has 2380557 rows and 54 columns. The data set has a lot of missing values.

I started to analyze the data set with **summary statistics** to understand the mean, standard deviation, minimum and maximum variables. Measures tells us where your data is centered at, or where a trend lies.

* Mean (also called the arithmetic mean or average).
* Median (the middle of a data set).
* Standard deviation (the square root of its variance)
* 25%/50%/75% (interquartile range (IQR))



Next steps are to clean the data set. To understand the missing values for each column in the data set, I tried to calculate the percentage of missing values in each column. Found, most of the columns have huge missing values. I had dropped those columns with more than 70% missing data.

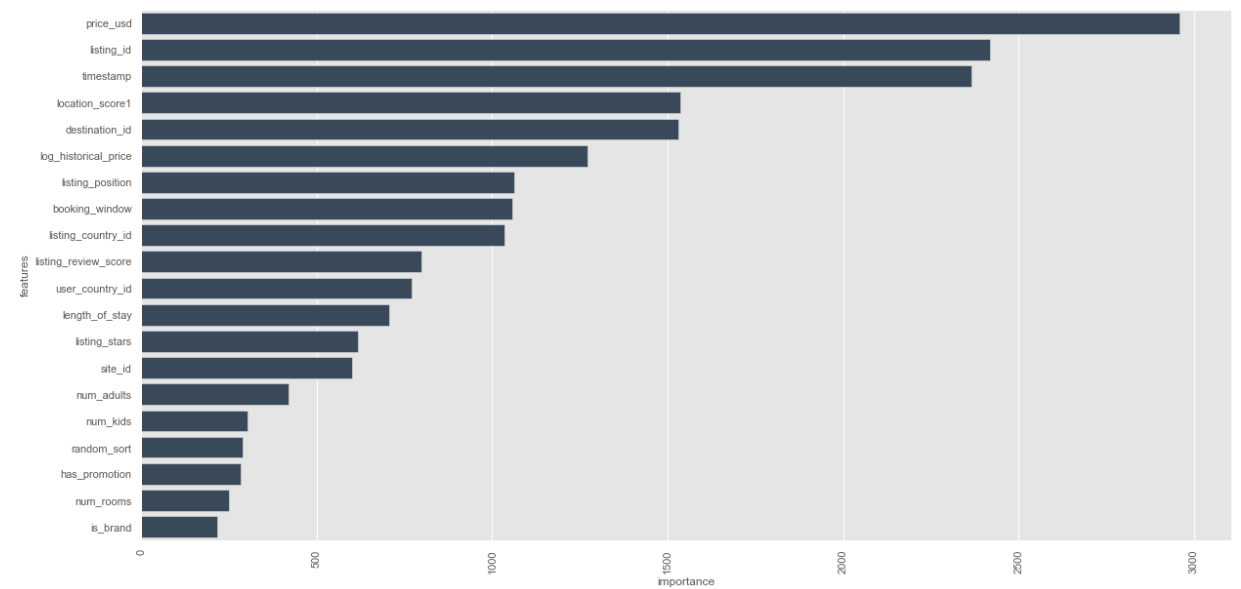
Next step is to analyze the columns one by one and try to impute the values for the missing values. But before that, I tried to analyze some of these columns and dropped the columns which are less important.

**Imputation** is one of the most important things in data analysis. Imputation is the process used to determine and assign replacement values for missing, invalid or inconsistent data that have failed edits. I started imputing values for some important columns such as listing\_review\_score. I calculated the mean of the variable and tried to put the mean in missing places.

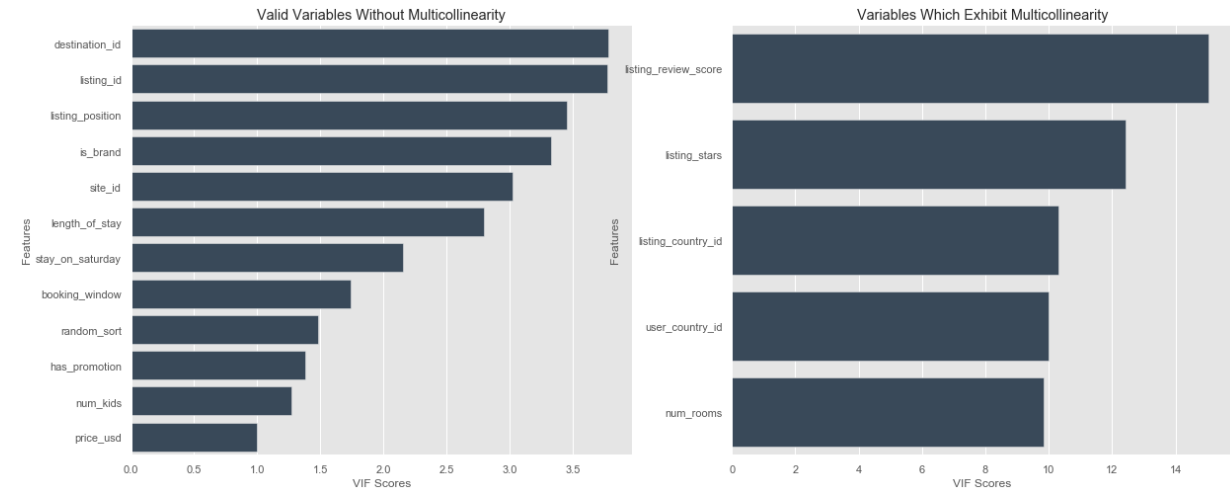
**Exploratory Data Analysis (EDA)** is the crucial process of using summary statistics and graphical representations to perform preliminary investigations on data in order to uncover patterns, detect anomalies, test hypotheses, and verify assumptions.

Data normalization, also known as **feature scaling**, standardizes a data set’s range of attributes, which might vary considerably. As we can see, the dataset comprises Travel website and some important variables are binary variables ranging from 0 and 1. Scaling the features simplifies the computing processes involved in model training and improves accuracy.

I started with feature selection. Below is the figure showing the features and its importance.



Then I tried to see the multi collinearity between the variables by a heat map and also Variance inflation factor. **Multicollinearity** is a statistical phenomenon in which predictor variables in a logistic regression model are highly correlated. Multicollinearity can cause unstable estimates and inaccurate variances which affects confidence intervals and hypothesis tests. The existence of collinearity inflates the variances of the parameter estimates, and consequently incorrect inferences about relationships between explanatory and response variables. Below is the figure displaying the VIF. Although, we can drop all the Multicollinearity variables. But, for now I have kept it as it is.



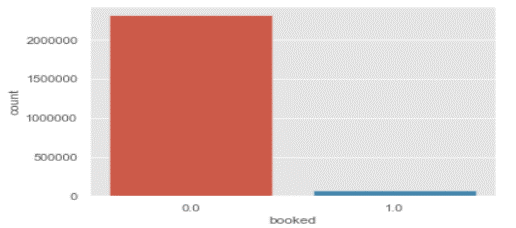
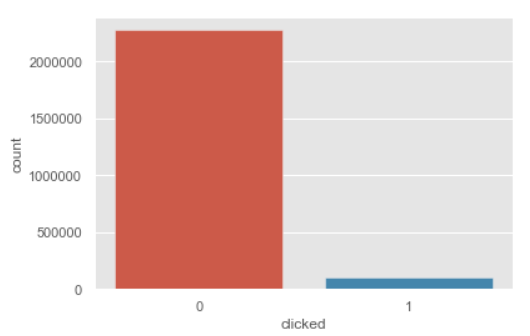
Let’s do a Univariate analysis for some the variables.

Started with “booked” and also tried to understand the distribution of this variable first. As, it is a binary variable so the distribution won’t help us understand.

Let’s do a bivariate analysis of the “booked” variable with the “timestamp”. We can see the monthly distribution over the months for the booked variable.

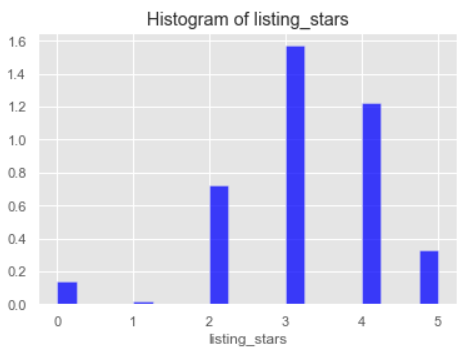
Also, I had done a bivariate analysis of the “clicked” variable with the “timestamp”.

Let’s understand the binary distribution of the “booked” variable by plotting a bar graphs. We can see that only 2.7% is been booked from the whole data set.

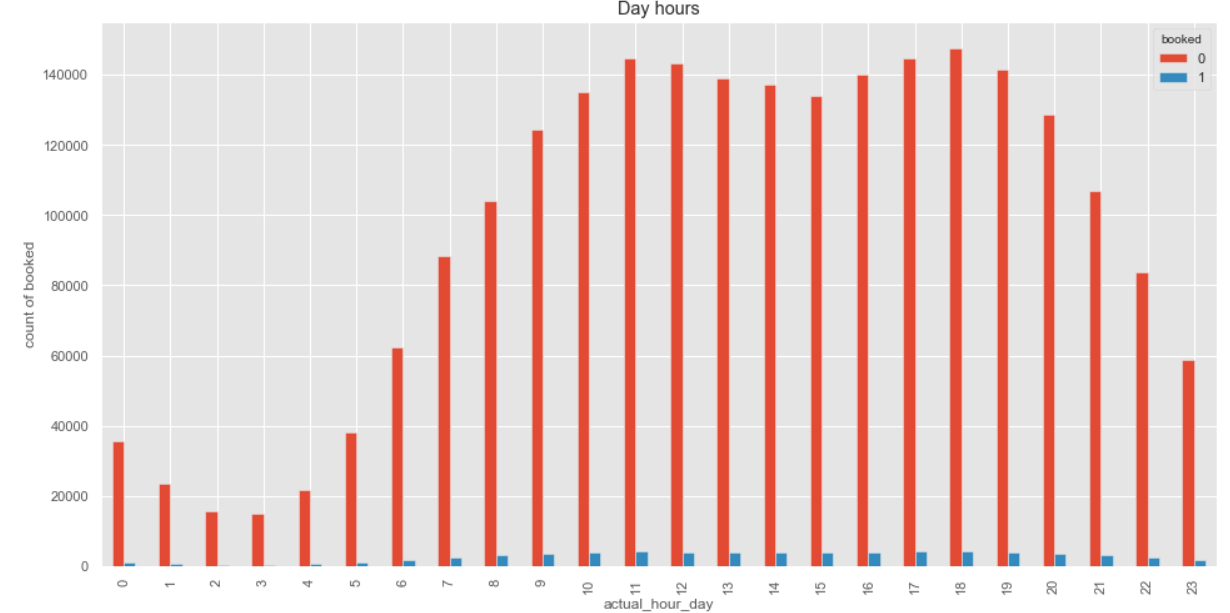
 

Also, only 4.45% is the clicking rate from the data set.

Let’s see the distribution of the variable called listing\_stars through a histogram. We can see that most people rated 3 and then 4. But listing\_review\_score on the other hand has 4 and 5 as most rated.

I also tried to see the distribution of the booked data over the hours. The blue bar charts which are small ones are booked and the red long bars are not booked.



Model Development

I have used Logistic Regression and also some Ensemble Algorithm. Now, let us know the importance of these algorithm and also why I choose them.

Logistic Regression Model

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

The main motive behind choosing logistic regression is because of its dependent variable which is binary in nature.

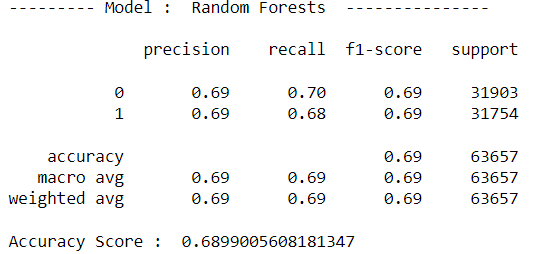
I have assumed “booked” as the dependent variable. Now I will split the data into training and test sets. Here I have used 70 per cent of data as training and 30 per cent as testing.

Now let’s have a look at the accuracy of the model which is around 97%.

Ensemble Techniques

Ensemble models is a machine learning approach to combine multiple other models in the prediction process. Those models are referred to as base estimators. The main objective behind choosing this technique is because of the Features noise and bias. Ensemble model relies heavily on one or a few features while making a prediction.

Let’s try with Ensemble Algorithms – Random Forest and see the accuracy.



**Answers to the given Questions.**

Answer 1. (b) The median and the mean both measure central tendency. Standard deviation is the square root of its variance, 25%/50%/75% are the interquartile range (IQR).

From the data set and the summary, we can say that the data set is highly imbalanced. A lot of variables are binary in nature.

Answer 1. (c) Multicollinearity can cause unstable estimates and inaccurate variances which affects confidence intervals and hypothesis tests. The existence of collinearity inflates the variances of the parameter estimates, and consequently incorrect inferences about relationships between explanatory and response variables. Tried to differentiate between Multicollinear variable and non Multicollinear variables which is show in above figure.

Answer 1. (d) There are few very common problems which makes generalization difficult and these can be seen in our data set. Now let’s understand what generalization means. Generalization refers to how well the concepts learned by a machine learning model apply to specific examples not seen by the model when it was learning. Overfitting and underfitting are these common problems. In our case, we can see a clear cut of overfitting problem due to its complex structure.

Answers1. (e) We can see, we have overfitting problem. There are two important techniques that you can use when evaluating machine learning algorithms to limit overfitting:

* Use a resampling technique to estimate model accuracy.
* Hold back a validation dataset.

The most popular resampling technique is k-fold cross validation. It allows you to train and test your model k-times on different subsets of training data and build up an estimate of the performance of a machine learning model on unseen data. But in our scenario, I have used balancing strategy to down sample the data.

Answer1. (f) Data is the most valuable asset for any business. Data security, or information security, includes the practices, policies and principles to protect digital data and other kinds of information. Data security is based on three foundational principles — confidentiality, integrity, and availability.

Yes, there are some data privacy and security issues in our dataset. They are variables related to Users, Ratings, competitors’ information’s etc.

Protecting sensitive information requires far more than implementing basic security technologies such as an antivirus solution and a firewall. Modern strategies include identity and access management, data discovery and classification, change management, and user and entity behavior analytics.

Answer2. (a) Click-through rates - 0 – 0.955433 and 1 – 0.044567

Conversion rates – 0 – 0.982842 and 1 – 1.598090

Answer2. (b) Property review score and property star rating do not have a significant effect on conversion rate. Also studies have shown that room price and hotel size are negatively associated with conversion rate. It was also found that a high rank in search listings, a high number of recommendations and location rating increase online hotel bookings.

Some of the statistical techniques to formally compare the effects of different levels of these variables are Univariate and Bivariate analysis. Some of them are shown with the code document.

Answer3. (a) I have chosen “booked” variables to model and train a machine learning model of your choice.

Answer3. (b) (a) I have chosen booked variable by eliminating booked value which has more than 90% missing values. Then out of booked and clicked, booked is more likely a dependent variable then clicked.

(b) My initial variable was booked variable.

(c) Various techniques were applied during the pre-processing of the data such as cleaning the raw data, Imputation techniques for the missing variables, dropping the variables which were not useful in the analysis, using multicollinearity techniques and also doing some exploratory analysis.

These techniques are very much useful in any data analytics process to normalize the data set to be useful to develop a machine learning model.

If these techniques are not used then the model will face a lot of problems such as overfitting or underfitting. Also mistakes, redundancies, missing values, and inconsistencies all compromise the integrity of the set, we need to fix all those issues for a more accurate outcome.

(d) I have chosen logistic regression first and then used ensemble machine learning techniques.

I choose logistic regression because of the variables in the data set is binary in nature. The other technique i.e., ensemble machine learning techniques is a machine learning approach to combine multiple other models in the prediction process. Those models are referred to as base estimators. The main objective behind choosing this technique is because of the Features noise and bias. Ensemble model relies heavily on one or a few features while making a prediction.

(e) Mostly I have chosen accuracy, precision, recall and f1 score as the evaluation metrics.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

(f) The model is quite complex because of the noise in the data, also we can see a lot of missing data in the data set.

If the model is too complex to be directly interpreted, how would you approach this problem? (Ambiguous Question)