**FLIGHT CHARGES PREDICTION -SOLUTION WITH DATA SCIENCE**

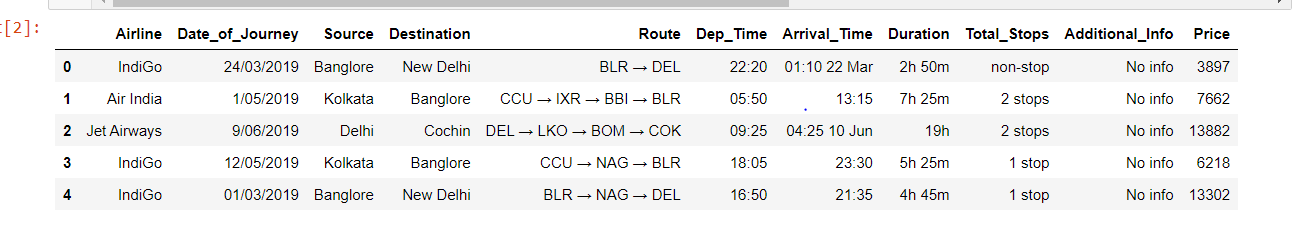
**During my Learning face of Data science, the project which I felt close to by heart is “Prediction of Flight charges”**

**PROBLEM STATEMENT:** As most of us faced this problem to book a flight for our journey. There is no means we could predict flight charge in next few minutes, above and all suggestion from well-wishers and friends are so different trying to predict the parameters causing the price charges and to book ticket in lowest price possible

Some suggestion , I heard from few of my friends are Don’t use the website a lot, Divide you bookings into many small bookings, Use the later days, Don’t books on weekends, try booking back seats, try booking flights at odd hours, try booking 1 seat at a time, try using this/that website , try using these Airlines , Use in this month for this city and that month for that city…..and it goes on and on…..

Are all these factors considered equally for the PRICE CHANGE A TICKET IN A FLIGHT

**The DATA set I used is shown in short**



This data set has (10683, 11) 10683 rows of the data and 11 features

**DATA ANALYSIS**

Airline-Airlines operating this flight

Date\_of\_Journey- Data of journey

Source-Start location

Destination-End Location

Route-Route the flight

Dep\_Time- When do you start your journey

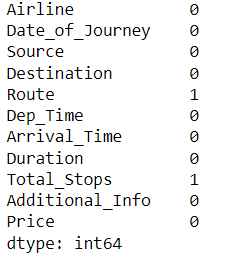
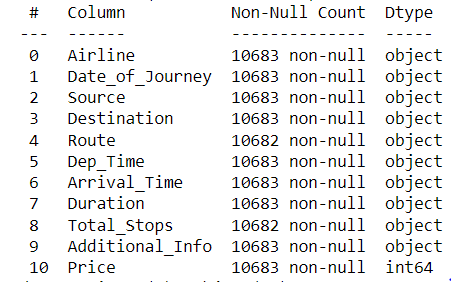
Arrival\_Time-When do you reach the end location

Duration -Journey time

Total\_Stops-Any stops in your journey

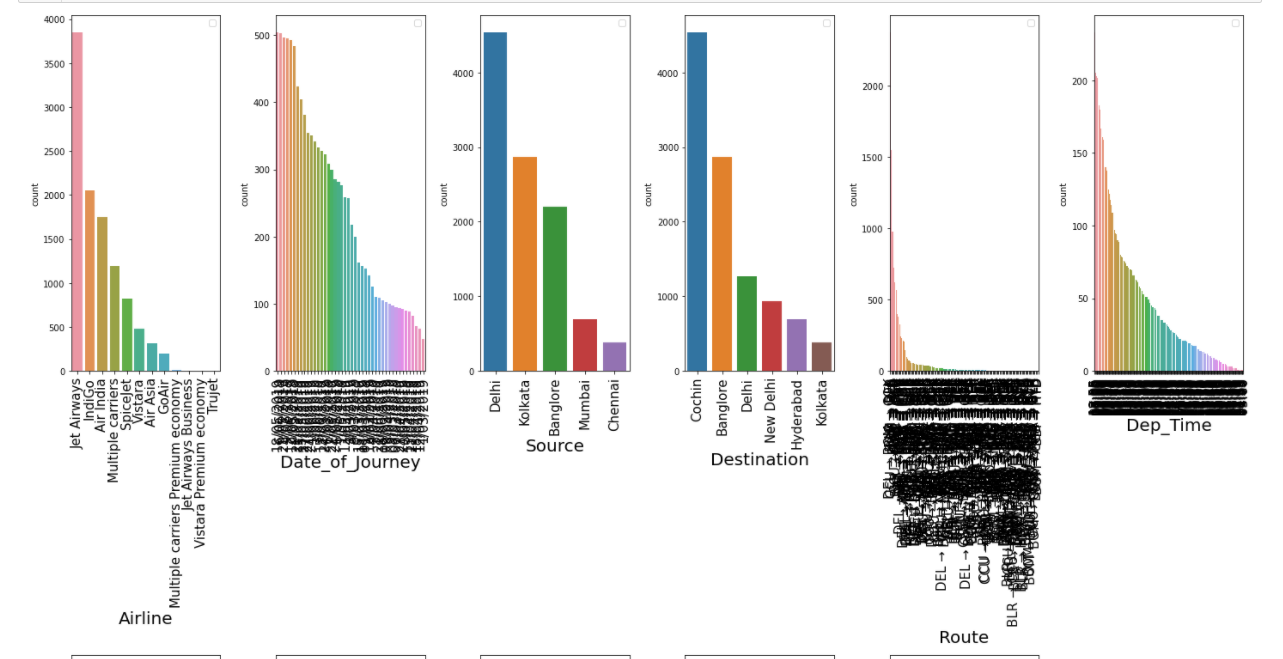
Additional\_Info-Additional comments form the Airlines

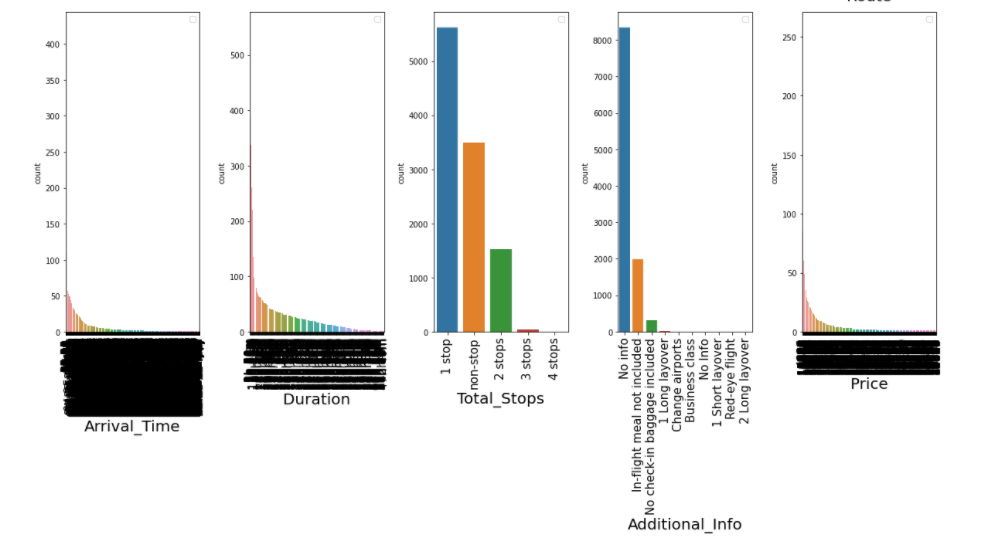
**Price-What is the price of the Ticket? This would be the Label/Target**



The Data Set have 2 missing data information in2 columns Route and total stops

Let’s understand by visualizing all the categories of data in every column how there are and how the data is distributes.





The Data set have more information from Jetways, then goes Indigo and very less information available from Trujet,

* Just keep these count percentage feeding into the model, this might influence the prediction accuracy in our test data

Check out Source and Destination places, total stops, Additional comments in this graph

Other graphs might not add value at this point.

**EXPLORATORY DATA ANALYSIS (EDA)**

LET’S STEPS DOWN WHAT ALL THE DATA SET NEEDS

**Step1) There is missing information-Need Imputing/deleting**

**Step2) All the columns are object types which need Encoding to feed into Machine learning Algorithm**

**Step3) The Data and time columns need to be defined correctly for the correct prediction**

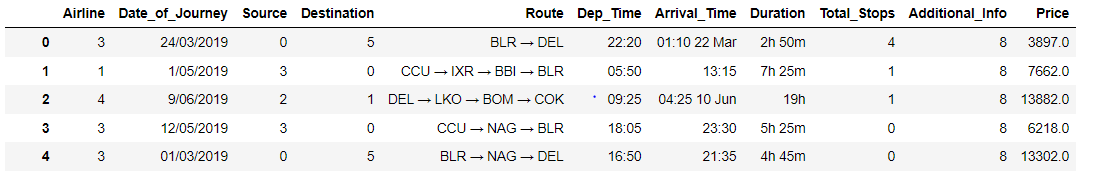
**Step1) There is missing information-Need Imputing/deleting**

This missing data on Route and Total Stops is encoded as unknow and moved forward

**Step2) All the columns are object types which need Encoding to feed into Machine learning Algorithm**

Encoding and defining the columns Airlines, Source, Destination, Total Stops, Additional Info by

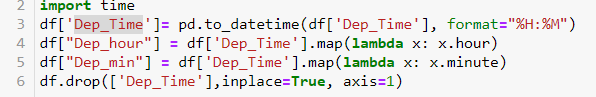
**Label Encoder** from SKLEARN PREPROCCING LIBRARY (Encode target labels with value between 0 and n\_classes-1.; This transformer should be used to encode target values, *i.e.* y, and not the input X.)

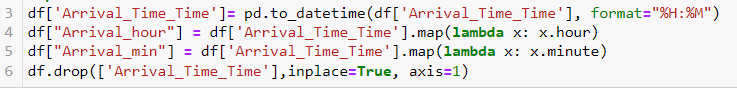


Here is my new data set encoded in some columns

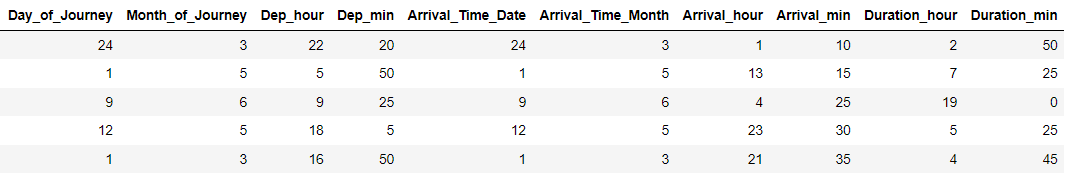
**Step3) The Data and time columns need to be defined correctly for the correct prediction**

**Defining Time and date columns like, Date\_of\_Journey, Dep\_Time, Arrival\_Time**





These are the new features generated from all the pre-available features

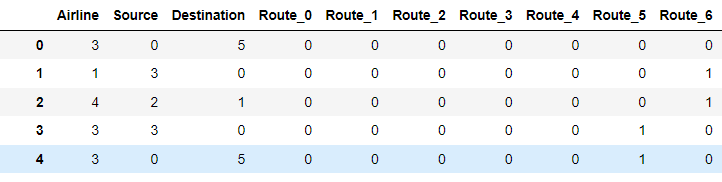


This Defining the format of the time, date data is very important as, there would keep the integrity of the data without disrupting the requirement of type aspect of Machine Learning Algorithm.

Encoding and defining the columns Route was done by

**Binary Encoder** from SKLEARN PREPROCCING LIBRARY (Encode target categorical variables, similar to one hot, but stores categories as binary bitstrings.)

I choose this to keep the integrity of the categories of the route of the flight as to be sure it is not distorted.



CORRELATION OF THE LABEL WITH THE FEATURES

When correlation is checked all the features contributes to the Label but there are all weekly correlated to the label/target variable

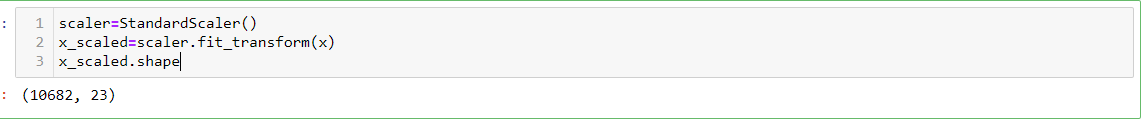
**PREPOSSING PIPELINE**

**SCALE THE DATA:**

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model.

Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

Its available in Sklearn prepossessing



**PRINCIPLE COMPONENT ANALYSIS:**

This is the reason I choose to go with PRINCIPLE COMPONETS ANALYSIS which reduces the dimensionality of a complex problem influenced by many features.

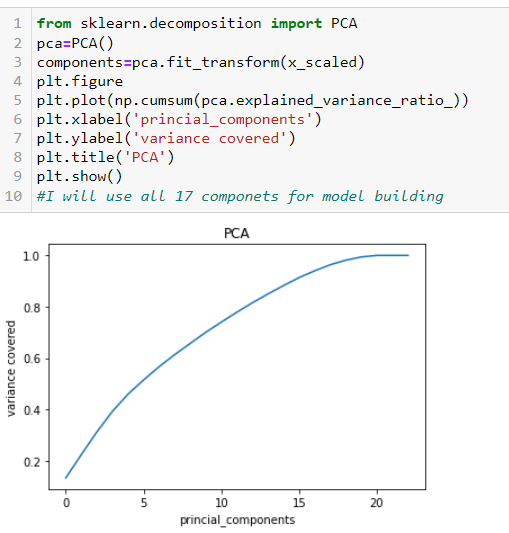
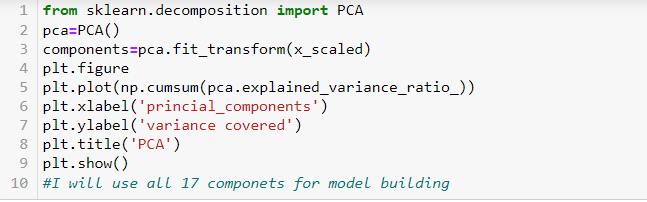
**As we know this is the issue with FLIGHT CHARGES PREDICTION, IT IS INFLUENCED BY MULTIDEMENTIONAL FEATURES**

**PCA is applied for dimensionality reduction to improve classification/regression accuracy by selecting the optimal set of lower dimensionality features.**

**Typical Applications of PCA**

* **Data Visualization.**
* **Data Compression.**
* **Noise Reduction.**
* **Data Classification.**
* **Image Compression.**
* **Face Recognition.**

**PCA is available in sklearn decomposition**

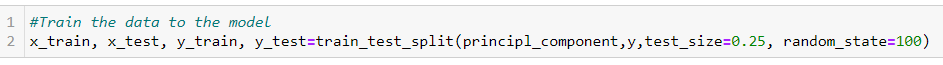
 

This Graphs indicated that there are 17 Principal components which would cover 100 of variance covered in the data.

This graph indicated to be that I would feed only 17 components into the system , This also saves money and time feeding 21 components.

**TRAIN THE MODEL:**





Here we can also leverage the random state value, to improve the accuracy,

**BUILDING MACHINE LEARNING ALGORITHM:**



I used multiple MACHINE LEARNING ALGORITHMS TO PREDICT THE PRICE OF FLIGHT, which is a regression problem

THE BEST MODEL WHICH SHOWED BEST ACCURARCY WAS **RANDOM FOREST**

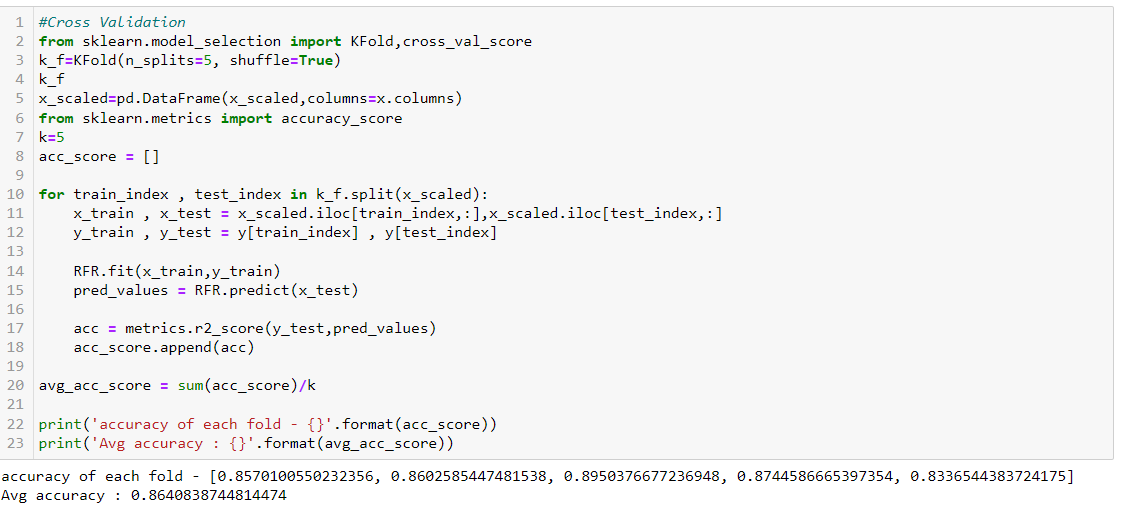
Random Forest grows multiple decision trees which are merged together for a more accurate prediction.

The logic behind the Random Forest model is that multiple uncorrelated models (the individual decision trees) perform much better as a group than they do alone. When using Random Forest for classification, each tree gives a classification or a “vote.” The forest chooses the classification with the majority of the “votes.” When using Random Forest for regression, the forest picks the average of the outputs of all trees.

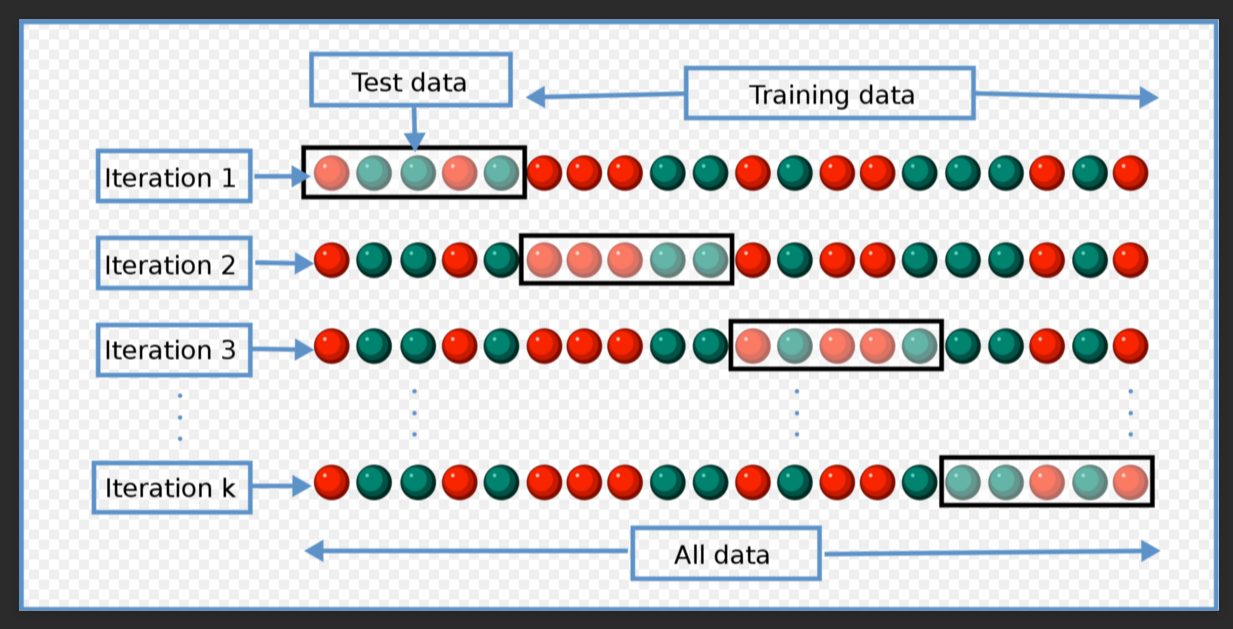
The key here lies in the fact that there is low (or no) correlation between the individual models—that is, between the decision trees that make up the larger Random Forest model. While individual decision trees may produce errors, the majority of the group will be correct, thus moving the overall outcome in the right direction.

**CROSS VALIDATION TECHNIQUE: I have chosen K-Fold validation scores**

**TO CHECK THE ROBASTNESS OF THE MODEL**



This K fold Validation technique, would split the data set into desired sets of train and test model and set the accuracy of the model

Illustration of the technique

**HYPERTUNNING THE MODEL-Grid search CV**

This HYPERTUNNING IS THE TECHNIQUE, where we try to change and iterate the model with various parameter to fetch the best accuracy and robust model.

**CONCLUSION**

**SAVED THE MODEL, which would predict the prices of flight charges with accuracy of 81%**