***Prediction on flight price***

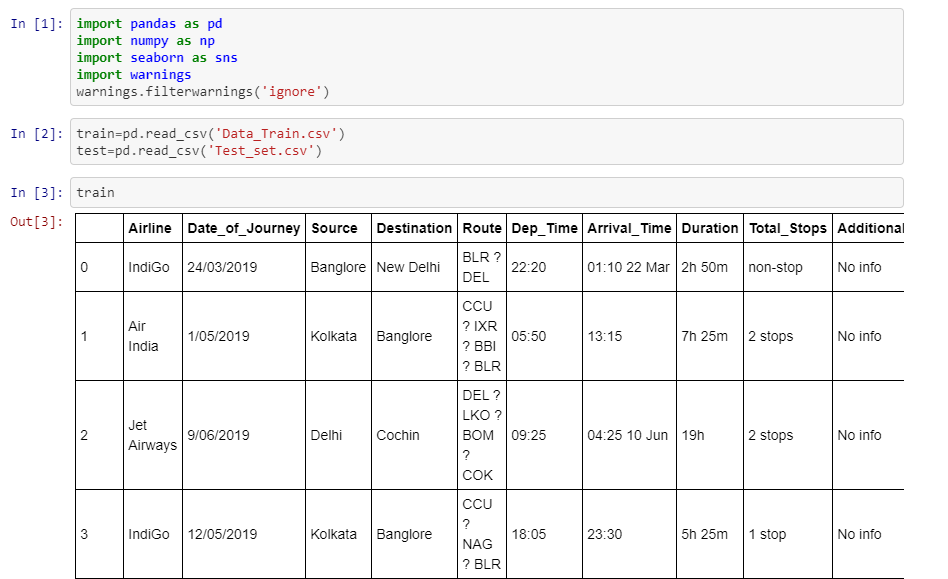
1. **Problem Definition:**

According to the survey, flight ticket prices change during the morning and evening time of the day, holidays or festival season. There are several different factors on which the price of the flight ticket depends. Considering the features such as departure time, the number of days left for departure and time of the day it will give the best time to buy the ticket. Using this information, we can build a system that can help buyers whether to buy a ticket or not.

1. **Data Collection:**

Data Collection is the most important aspect of any project. More and better the data we get, the better our model will perform. We will be using two datasets- Train data and Test data. Test data is similar to Train data without Price column.

* Import relevant libraries in Python
* Import the train and test datasets



Dataset is combination of both categorical and numerical values, also we can see some special characters being used. The data needs to be transformed before applying it to our model.

1. **Data Analysis:**

In this step we work on transforming the data like removing null values, clean the messy data so that it can be used to train the model.

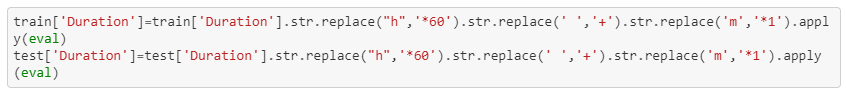
* **Date\_of\_Journey:**

In ‘Date\_of\_Journey’ column the date format is given as dd/mm/yyyy and the datatype is object. Here we divide the column into date, month and drop ‘Date\_of\_Journey’ column in train and test datasets.



* **Duration:**

Duration column is converted into int datatype by replacing the string ‘h’ with \*60,’m’ with \*1 and applying ‘eval’. Now the duration is calculated in minutes

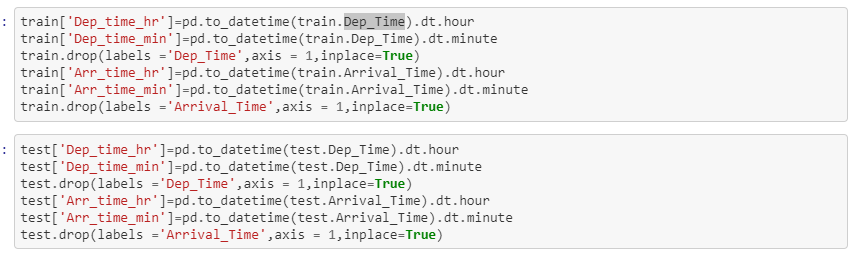


* **Arrival\_Time:**

Arrival\_Time column is combination of both time and month, but we take only time details and split into ‘hours’ and ‘minutes’.

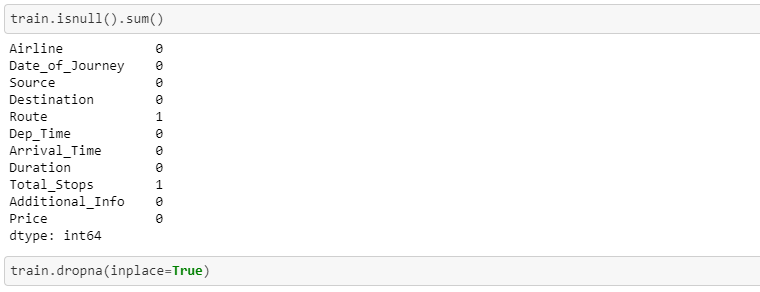
* **Dep\_Time:**

Similar to Arrival\_Time , this column is converted to integer by splitting into ‘hours’ and ‘minutes’.



* Repacing / Droping of missing Values

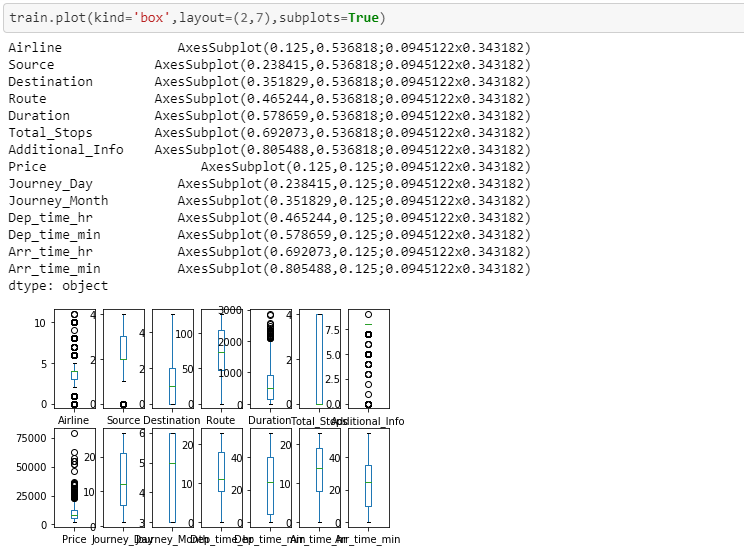
It is important to handle missing data. The values can be identified by the ‘isnull()’. After identifying the null values, it depends on each case whether to fill the missing value with the mean, median or the mode, data drop the entry.



1. **EDA Remarks:**

EDA is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

* Visualising the dataDuring the exploration phase, we try to understand what patterns and values our data has. We’ll be using different types of visualizationsand statisticaltestings to back up our findings. We also convert categorical text data into numerical data using different plot

****

It shows outliers present in the distribution.

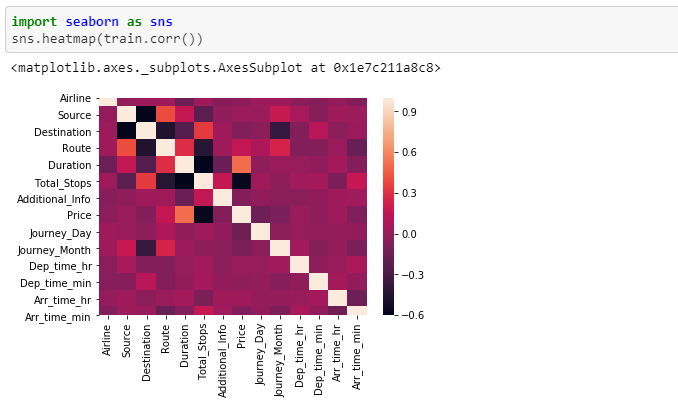
* Removing outliers

Outliers are the unusual values in the dataset. They cause measurement errors and sampling problems while building a predictive model. They can be removed using zscore



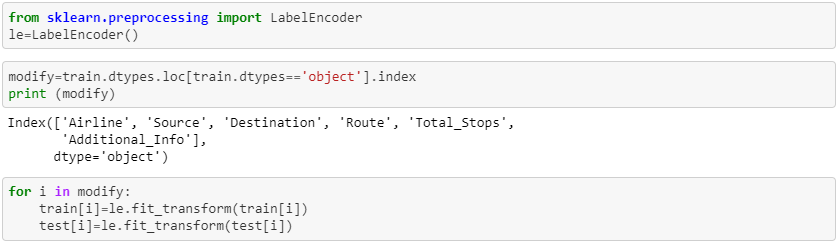
* Correlation

Correlation is a measure of how strongly one variable depends on another var-iable. It is an important tool in building machine learning models. Predictors which are uncorrelated with the objective variable are probably good candidates to trim from the model. Correlation coefficients are a useful tool for exploring relationships within your data .



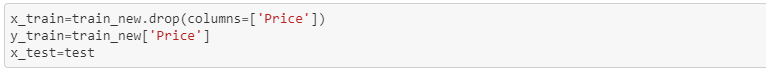
* Converting Categorical data into numerical

As we cannot calculate anything with string values, we have to convert these values into numeric data. This can be done using Label Encoder.



* Splitting the data set

Now that all our data is numerical after label encoding so we split the data into x\_train , y\_train and x\_test.



1. **Pre-processing Pipeline:**

A data pipeline architecture is a system that captures, organizes, and routes data so that it can be used to gain insights. Raw data contains too many data points that may not be relevant. Data pipeline architecture organizes data events to make reporting, analysis, and using data easier. It is primarily applied to help data improve targeted functionality. This allows for easier tuning and better access to the configuration of the entire model. They are cyclical and iterative as every step is repeated to continuously improve the accuracy of the model and achieve a successful algorithm.

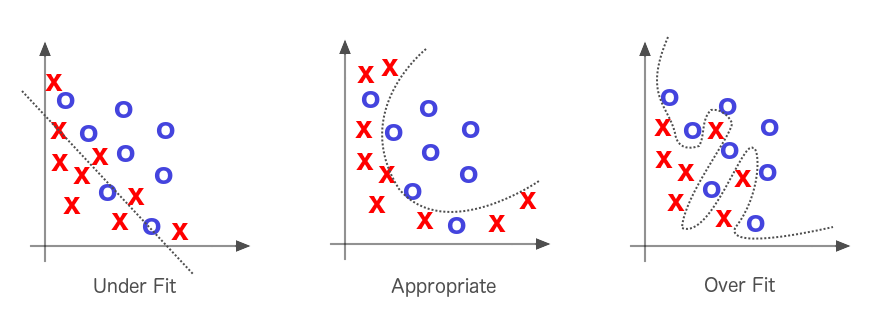
1. **Model Training:**

Our goal is to develop a benchmark model which measures the performance of a better and more tuned algorithm. To develop the model many conventional machine lear-ning algorithms are evaluated. We are using different Regression Technique and comparing them to see which algorithm is giving better performance.

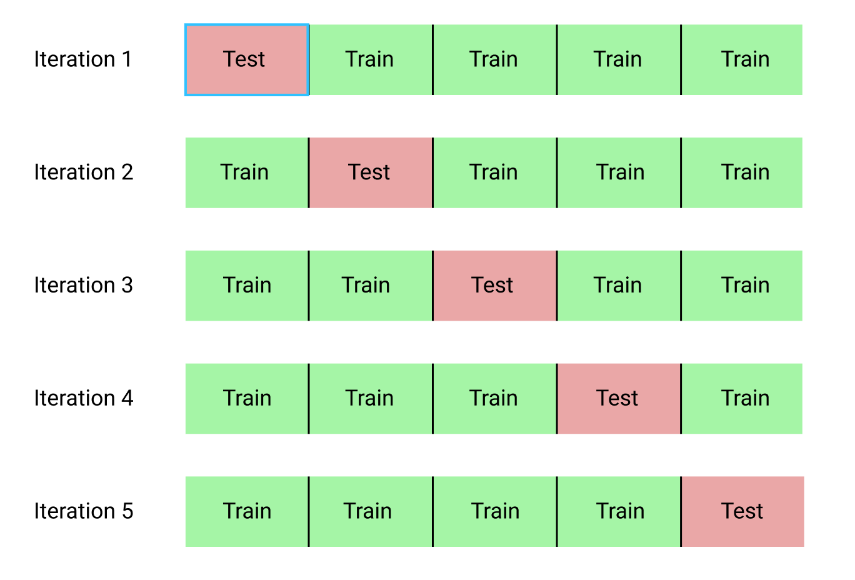
* **Cross validation**

Once we are done with training our model, we just can’t assume that it is going to work well on data that it has not seen before. A Good Model is not the one that gives accurate predictions on the known data or training data but the one which gives good predictions on the new data and avoids overfitting and underfitting. To evaluate the performance of any machine learning model we need to test it on some unseen data. Based on the models performance on unseen data we can say weather our model is Under-fitting,Over-fitting or Well generalised.

* **Overfitting** a model result in good accuracy for training data set but poor results on new data sets. Such a model is not of any use in the real world as it is not able to predict outcomes for new cases.
* **Underfitting** occurs when the model or the algorithm does not fit the data well enough. Underfitting is often a result of an excessively simple model which do not contribute much to the predictor variable. Cross validation (CV) is one of the technique used to test the effectiveness of a machine learning models.



In general we use K-Fold cross validation. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

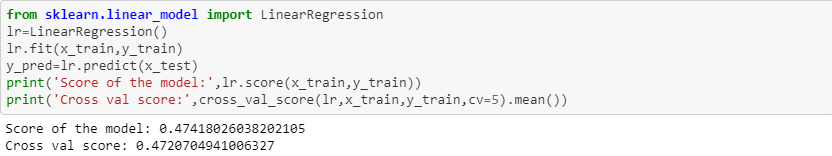


* **Linear Regression** :

**Linear Regression** is a machine learning algorithm based on **supervised learning.**It performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

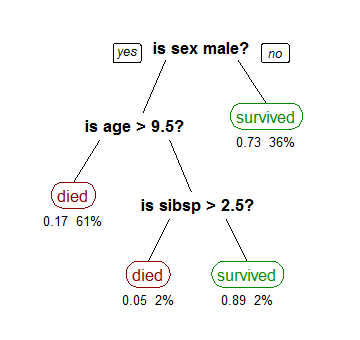
**y = a\_0 + a\_1 \* x --> Linear Equation**

Here motive of the linear regression algorithm is to find the best values for a\_0 and a\_1.



* **Decision Tree Regressor:**

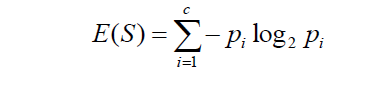
A decision tree is one of most frequently and widely used supervised machine learning algorithms that can perform both regression and classification tasks. For each attribute in the dataset, the [decision tree](https://en.wikipedia.org/wiki/Decision_tree_learning) algorithm forms a node, where the most important attribute is placed at the root node. For evaluation we start at the root node and work our way down the tree by following the corresponding node that meets our condition or "decision". This process continues until a leaf node is reached, which contains the prediction or the outcome of the decision tree.



There are two primary characteristics in the decision tree calculation. One is Information Gain and another is the Gini index.

1) Information Gain is the proportion of Change in entropy. Higher the entrop y more the instructive substance, where the entropy is a proportion of vulnerability of arbitrary variable.

The Mathematical formula for Entropy is as follows -

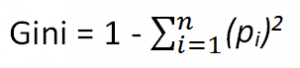


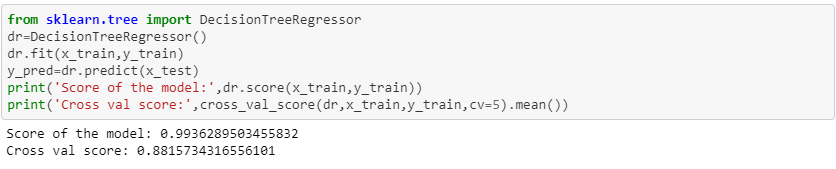
Where ‘Pi’ is simply the frequentist probability of an element/class ‘i’ in our data.

Next we need a metric to measure the reduction of this disorder in our target variable/class given additional information( features/independent variables) about it. This is where Information Gain comes in. Mathematically it can be written as:



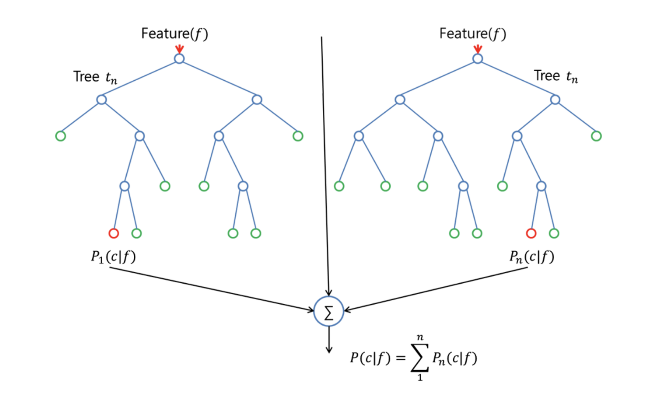
2) Gini Index is a component that measures how frequently an arbitrarily picked component would be mistakenly distinguished. It implies a characteristic with a lower Gini index ought to be liked. The Gini Index is calculated by subtracting the sum of the squared probabilities of each class from one. The Mathematical formula for Gini Index is as follows -

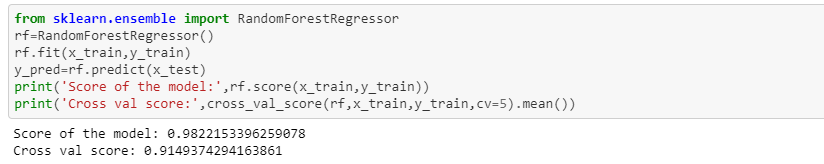




* **Random Forest Regressor:**

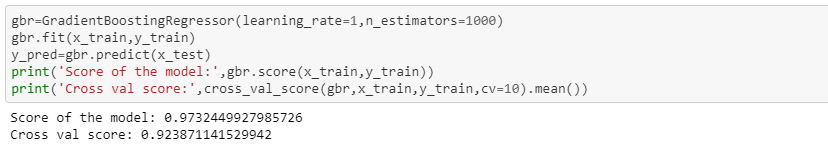
Random Forest is a supervised learning algorithm. It uses the ensemble learning technique(Ensemble learning is using multiple algorithms at a time or a single algorithm multiple times to make a model more powerful) to build several decision trees at random data points. Then their predictions are averaged. Taking the average value of predictions made by several decision trees is usually better than that of a single decision tree.





* **Gradient Boo sting Regressor:**

Gradient Boosting Regressors are ensemble decision tree regressor models. It is a type of inductively generated tree ensemble model. At each step, a new tree is trained against the negative gradient of the loss function, which is analogous to (or identical to, in the case of least-squares error) the residual error. Gradient boosting models work best when all of the input features have been normalized to have zero mean and unit variance.



We also check the RMSE score of each model so that we can understand how it should perform in our test dataset.

1. **Conclusion:**

In this blog the overall survey for the dynamic price changes in the flight tickets is presented which gives the information about the highs and lows in the airfares according to the days, weekend and time of the day that is morning, evening and night. Also the machine learning models are evaluated on different datasets are studied. Their accuracy and performances are evaluated and compared in order to get better result. From the above different Regression Technique we can see Gradient boosting model is performing really good in regards to all .Finally we will use this to predict our test data.