**Flight Price Prediction**

**A picture containing sky, grass, outdoor, plane

Description automatically generated**

**Problem Statement:**

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable. Here you will be provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities.

Size of training set: **10683** records

Size of test set: **2671** records

FEATURES:

**Airline**: The name of the airline.

**Date\_of\_Journey**: The date of the journey

**Source**: The source from which the service begins.

**Destination**: The destination where the service ends.

**Route**: The route taken by the flight to reach the destination.

**Dep\_Time**: The time when the journey starts from the source.

**Arrival\_Time**: Time of arrival at the destination.

**Duration**: Total duration of the flight.

**Total\_Stops**: Total stops between the source and destination.

**Additional\_Info**: Additional information about the flight

**Price**: The price of the ticket

You have to use your skills as a data scientist and build a machine learning model to predict the price of the flight ticket.

**Problem Definition**

The flight ticket buying system is to purchase a ticket many days prior to flight take-off to stay away from the effect of the most extreme charge. Mostly, aviation routes do not agree this procedure. Plane organizations may diminish the cost at the time, they need to build the market and at the time when the tickets are less accessible. They may maximize the costs. So, the cost may rely upon different factors. To foresee the costs this venture uses AI to exhibit the ways of flight tickets after some time. All organizations have the privilege and opportunity to change its ticket costs at any time. Explorer can set aside cash by booking a ticket at the least costs. People who had travelled by flight frequently are aware of price fluctuations. The airlines use complex policies of Revenue Management for execution of distinctive evaluating systems. The evaluating system as a result changes the charge depending on time, season, and festive days to change the header or footer on successive pages. The aim of the airways is to earn profit whereas the customer searches for the minimum rate. Customers usually try to buy the ticket well in advance of departure date to avoid hike in airfare as date comes closer. But this is not the fact. The customer may wind up by giving more than they ought to for the same seat.

**Importing Library:**

**Graphical user interface

Description automatically generated with low confidence**

I am importing the whole library which I required for EDA, visualization, prediction and finding all matrices. The reason of doing this is that it become easier to use all the import statement at one go and we do not require to import the statement again at each point. We could find all the importing statement at one place without finding it on whole notebook and can update also.

**Loading Data Set**

**Graphical user interface, text, application

Description automatically generated**

Here I am loading the data set into a variable i.e., “df” and processing the first 5 rows. As in this data set most of the column are float in nature and type is of categorical value.

**Exploratory Data Analysis:**

Text

Description automatically generated



As seen in data set there are 10 objective columns and one integer column. Also, I am checking the shape of both data set as there are 10683 rows and 11 columns in train data set and in 2671 rows and 10 columns in test data set.

Text

Description automatically generated

Above I am checking the null values, as find there are no null values in the test data set.

Graphical user interface, text

Description automatically generated

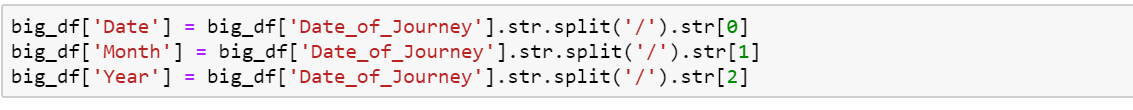
Above I am checking the null values, as find there are 2 null values in the train data set. Now I have dropped all the null values for better data set predictions.

**Feature Generation**

In this step we mainly work on the data set and do some transformation like creating different bins of columns, clean the messy data so that it can be used in our ML model. This step is particularly important because for a high prediction score you need to continuously make changes in it.

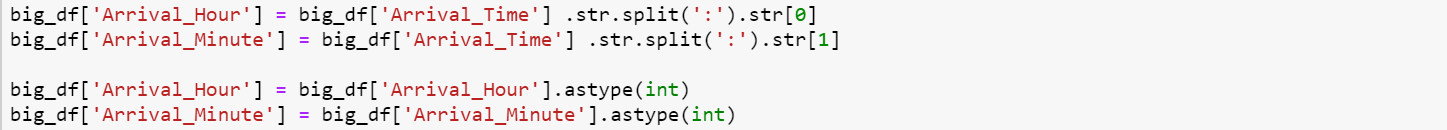
**Date\_of\_Journey:**

In the column ‘Date\_of\_Journey’, we can see the date format is given as dd/mm/yyyy and as you can see the datatype is given as object So there is two ways to tackle this column, either convert the column into Timestamp or divide the column into date, Month, Year. Here, I am splitting the columns.



**Total\_Stops:**

This column is combination of number and a categorical variable like ‘1 stop’. So, we need only the number details from this column, so we split that and take the number details only also we change the ‘nonstop’ into ‘0 stop’ and convert the column into integer type.



**Total\_Stops:**

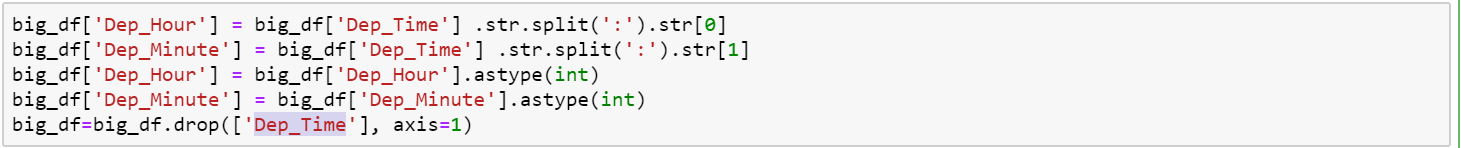
This column is combination of number and a categorical variable like ‘1 stop’. So, we need only the number details from this column, so we split that and take the number details only also we change the ‘nonstop’ into ‘0 stop’ and convert the column into integer type

Graphical user interface, text, application

Description automatically generated

**Dep\_Time:**

As same as ‘Arrival time’. we split this column also in hour and minute and convert it into integer



**Route:**

The ‘Route’ columns mainly tell us that how many cities they have taken to reach from source to destination. This column is especially important because based on the route they took will directly affect the price of the flight, So We split the Route column to extract the information. Regarding the ‘Nan’ values we replace those ‘Nan’ values with ‘None’.

Text

Description automatically generated

**Chart, bar chart, histogram

Description automatically generated**

There are a greater number of flights of Jet Airways.

Jet Airways Business, Vistara Premium economy, Trujet have almost negligible flights.

# Handling Categorical Data

One can find many ways to handle categorical data. Some of them categorical data are,

Nominal data - data are not in any order --> **OneHotEncoder** is used in this case.

Ordinal data- data are in order --> **LabelEncoder** is used in this case.

Text

Description automatically generated

I can see that Jet Airways Business have the highest Price.

Apart from the first Airline almost all are having similar median.

**As Airline is Nominal Categorical data, we will perform One Hot Encoding**

Table

Description automatically generated

**Prepare categorical variables for model using label encoder.**

To convert categorical text data into model-understandable numerical data, we use the Label Encoder class. So, all we must do, to label encode a column is import the Label Encoder class from the sklearn library, fit and transform the column of the data, and then replace the existing text data with the new encoded data.

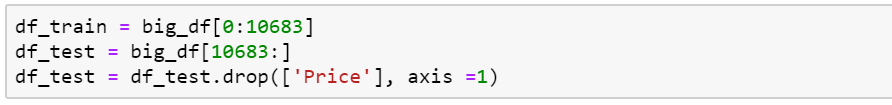
A picture containing text

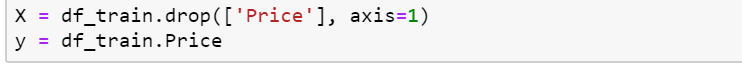
Description automatically generated

Label encoding of Categorical variables

**Divide the data set into test and train.**

Now that all our data is numerical after label encoding, so we split the data into test and train and drop the price column from the test set because we must predict the price with our test data set.





**Build Model**

The goal in this step is to develop a benchmark model that serves us as a baseline, upon which we will measure the performance of a better and more tuned algorithm. We are using different Regression Technique and comparing them to see which algorithm is giving better performance than other and at the end we will combine all of them using Stacking and see how our model is predicting.

**Important feature using Extra Trees Regressor**

Graphical user interface, text, application

Description automatically generated

plot graph of feature importance’s for better visualization

Chart

Description automatically generated

# Random Forest.

Split dataset into train and test set to prediction w.r.t X\_test.

If needed do scaling of data

Scaling is not done in Random Forest.

Import model.

Fit the data.

Predict w.r.t X\_test.

In regression check RSME Score

Plot graph

Graphical user interface, text, application, email

Description automatically generatedRandom Forest is a method of artificial intelligence, which is an integrated learning method that uses bagging algorithm combined with many decision trees for classification, regression, or other tasks. Random forest is a forest composed of many decision trees, and the decision trees are independent of each other. Random forest, proposed by Leo Bierman, uses a bootstrap resampling technique to randomly extract k samples randomly from the original training sample set N to generate a new set of training samples. Then generate k classification trees to form random forest, and the classification result of new data depends on the classification tree voting.

# Hyperparameter Tuning

Choose following method for hyperparameter tuning.

RandomizedSearchCV --> Fast

GridSearchCV

Assign hyperparameters in form of dictionary.

Fit the model.

Check best parameters and best score.

**Randomized Search CV**

Number of trees in random forest

Graphical user interface

Description automatically generated with low confidence

Graphical user interface, text, application

Description automatically generated

Random Forest Regressor giving Maximum Accuracy as compared to other Regressor algorithm. In this type of dataset Feature Engineering is the most crucial think. We can see how we have handled the categorical and numerical data and how I build different ML model on the same dataset. We also check the RMSE score of e model so that we can understand how it should perform in our test dataset.

**Results**

* The trend of flight prices varies over various months and across the holiday.
* There are two groups of airlines: the economical group and the luxurious group. Spicejet, AirAsia, IndiGo, Go Air are in the economical class, whereas Jet Airways and Air India in the other. Vistara has a more spread-out trend.
* The airfare varies depending on the time of departure, making timeslot used in analysis an important parameter.
* The airfare increases during a holiday season. In our time, during Diwali the fare remained high for all the values of days to departure. We have not considered holiday season as a parameter now, since we are looking at data for a few months.
* Airfare varies according to the day of the week of travel. It is higher for weekends and Monday and slightly lower for the other days.
* There are a few times when an offer is run by an airline because of which the prices drop suddenly.

**Conclusion**

Dynamic pricing is one of the most common pricing strategies implemented by the airline industry to adjust ticket prices in response to various internal and external factors such as changes in demand, competitor promotions, ability of users to buy, availability of seats and others. Airlines need to predict changes in these factors to implement a dynamic pricing scheme that dynamically adjusts ticket prices to increase their profit. On the other hand, customers are also interested to forecast how ticket prices would change in the future to be able to buy tickets at lower prices. Therefore, researchers have developed various prediction models both for airlines and customers to help them deal with dynamic pricing. The two most common methods proposed for airlines are demand prediction and price discrimination which we collectively refer to as Airlines side models. Customer side modes involve optimal ticket purchase time prediction models and ticket price prediction models. There is a trade-off between money saving by customer and increasing revenue by companies. As customers become more strategic by using customer side tools, it becomes more difficult for the airlines to apply dynamic pricing and to generate profit and vice versa. Therefore, there is a need for a prediction model that can predict the optimal ticket prices that can bring mutual benefit both for customers and airlines.

Based on what we have presented, we can infer that ticket price prediction and demand prediction research is at an infancy stage. There is room for improvements in several areas including predicting exact value of ticket prices/demand, dataset issues, limited the number of features, lacking generality, better prediction techniques and performance and complexity issues. Most of the research conducted in this area do not predict the exact value of a ticket price or the demand.

These models work in such a way that for each ticket query the customer performs, the model generates a binary signal indicating either to buy or to wait. However, the models do not predict the exact value of a ticket price in advance. Moreover, the maximum performance achieved so far is 75% which is not always acceptable. Nevertheless, there are few studies which attempted to predict the exact value of ticket prices. However, the used models in these studies suffer from computational overhead as it is computationally more intensive than predicting the optimal purchase time.

In the area of demand prediction, the most notable work predicts quarterly route demand but cannot work for short term prediction. The other models in suggested for demand prediction only estimate the percentage increment or decrement in demand for a flight based on price elasticity. Another important topic that is not yet explored well is related to the development of a price discrimination model. None of the previous studies propose a technique for price discrimination but they rather focus on proving the existence of price discrimination in airlines pricing strategies.