Flight Prices Prediction

Abstract

As domestic air travel is getting more and more popular these days in India with various air ticket booking channels coming up online, travellers are trying to understand how these airline companies make decisions regarding ticket prices over time. Nowadays, airline corporations are using complex strategies and methods to assign airfare prices in a dynamic fashion. These strategies are taking into consideration several financial, marketing, commercial and social factors are closely connected with the ultimate airfare prices. Due to the high complexity of the pricing models applied by the airlines, it is very difficult for a customer to purchase an air ticket at the lowest price, since the price changes dynamically. For this reason, several techniques ready to provide the proper time to the customer to buy an air ticket by predicting the airfare price, are proposed recently. The majority of those methods are making use of sophisticated prediction models from the computational intelligence research field known as Machine Learning (ML). In this machine learning in python project there is only one module namely, User. User can login with valid credentials in order to access the web application. A traveller can access this module to get the future price prediction of individual airlines. The prediction will help a traveller to decide a specific airline as per his/her budget. Single entries of current or previous data can be made. This training set is used to train the algorithm for accurate predictions.

##### **Advantages**

* Traveller get the fare prediction handy using which it’s easy to decide the airlines.
* Saves time in searching / deciding for airlines.

##### **Disadvantages**

* Improper data will result in incorrect fare predictions.

1. Problem Definition:

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 travelers saying that flight ticket prices are so unpredictable. As data scientists, we are gonna prove that given the right data anything can be predicted. 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.

1. Data Analysis.

Size of training set: **10683** records

Size of test set: **2671** records

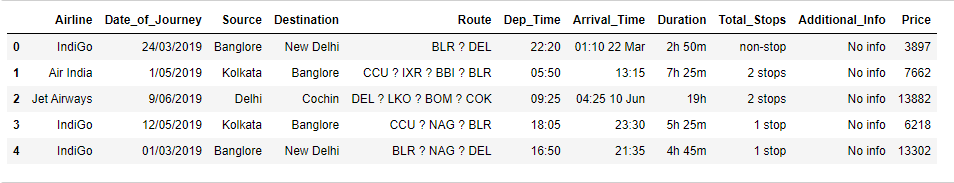
FEATURES:

* **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

We having two data set training and testing:

**Training:**

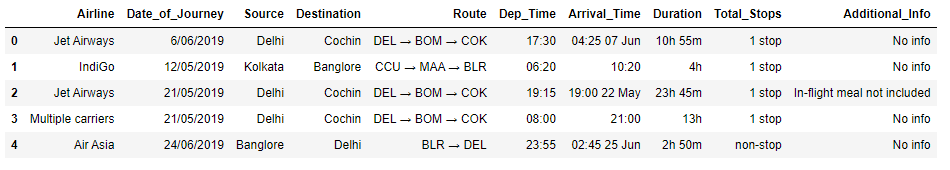
Training dataset has following dataset:



Training dataset has 10683 row and 11 column.our target variable is price.

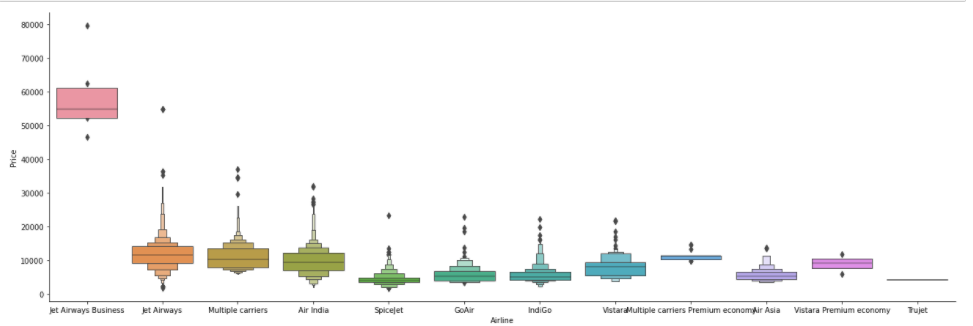
Testing:

Training dataset has following dataset



1. EDA

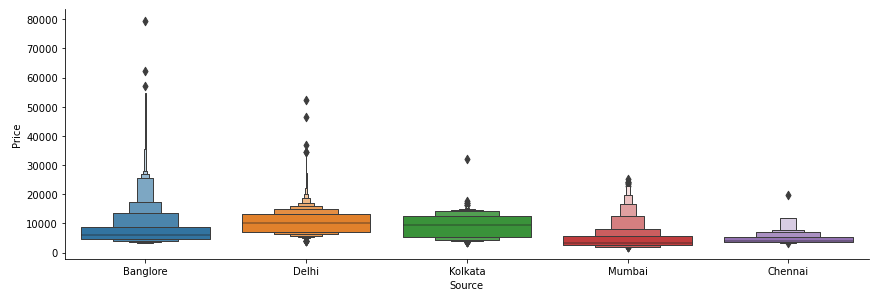
* Relationship between price and Airline



From graph we can see that Jet Airways Business have the highest Price.

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

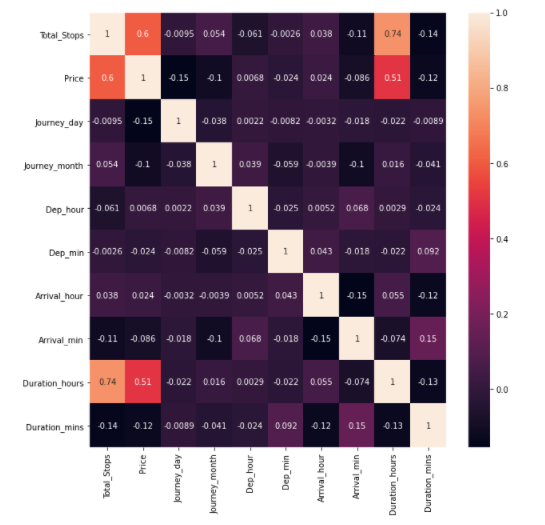
* Relationship between price and source



The flight which start from Bangalore has more price and flight start from Chennai charge less.

* Heatmap

Finds correlation between Independent and dependent attributes



Heat map show that the duration of hours and price has highest correlation.

Number of stop and price has positive correlation means number of stop increases prices is increasing.

4. Pre-Processing

Additional\_Info contains almost 80% no\_info so we drop additional\_info column.

Route and Total\_Stops are related to each other so we drop the route column.

Apply hot encoding to convert data in 0 and 1.

1. Building Machine Learning Models.

* Feature selection:

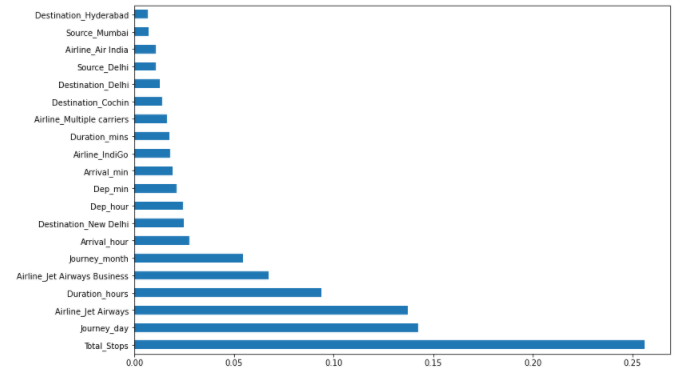


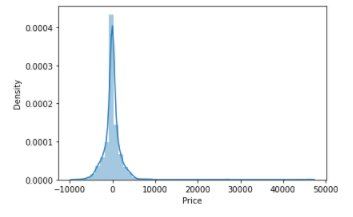
Fig show low to highly important feature.total\_stop is very important .Fitting model using Random Forest Split dataset into train and test set in order to prediction w.r.t X\_test If needed do scaling of data Scaling is not done in Random forest Import model Fit the data Predict w.r.t X\_test In regression check RSME Score Plot graph.

Random Forest

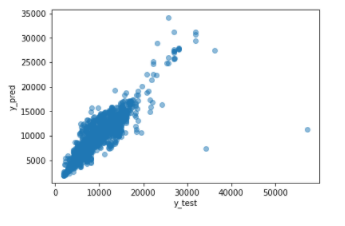
It is a supervised learning algorithm. The benefit of the random forest is, it very well may be utilized for both characterization and relapse issue which structure most of current machine learning framework. Random forest forms numerous decision trees, what’s more, adds them together to get an increasingly exact and stable expectation. Random Forest has nearly the equivalent parameters as a decision tree or a stowing classifier model. It is very simple to discover the significance of each element on the expectation when contrasted with others in this calculation.

The regular component in these techniques is, for the kth tree, a random vector theta k is produced, autonomous of the past random vectors theta 1, … , theta k-1 however with the equivalent distribution,while a tree is developed utilizing the preparation set and bringing about a classifier. x is an information vector. For a period, in stowing the random vector is created as the includes in N boxes where N is the number of models in the preparation set of information. In random split, choice includes various autonomous random whole numbers between 1 to K. The dimensionality and nature of theta rely upon its utilization in the development of a tree. After countless trees are created, they select the most famous class. These methodology are called as random forests.

After apply random forest model on training data it give 95% of score and testing data has 79% of score.



Here this graph form by using test data and predicted data .it for Gaussian distiribution.it means our model work properly.



On same data apply on scatter plot data show the increasing order.

We also use metrics and it give the r2 score is 79% which is also good.

**Hyper parameter Tuning**:

For hyper tuning we use random searchCV and our r2 score is increase 79% to81%

6. Concluding

In the proposed paper the overall survey for the dynamic price changes in the flight tickets is presented. This 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 in the computational intelligence feild that are evaluated before on different datasets are studied. Their accuracy and performances are evaluated and compared in order to get better result. For the prediction of the ticket prices perfectly differnt prediction models are tested for the better prediction accuracy. As the pricing models of the company are developed in order to maximize the revenue management. So to get result with maximum accuracy regression analysis is used. From the studies , the feature that influences the prices of the ticket are to be considered. In future the details about number of availble seats can improve the performance of the model.