Flight Price Prediction

1. PROBLEM DEFINITION

The given dataset consists of information of various Indian Airlines from year 2019 between the month of March to June. The main objective is to build a machine learning algorithm which predicts the flight price given the following parameters.

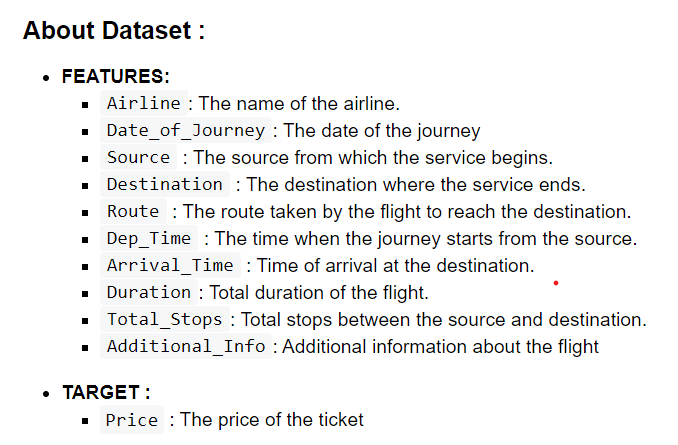
1. Introduction

In India, there are over 400 airports and airstrips, while 153 were operational. Passenger traffic amounted to over 115 million at airports across India in financial year 2021, out of which over 10 million were international passengers. Airline companies use various algorithms to predict flight prices on the basis of dynamically changing financial, marketing and social aspects.

Anyone who've booked an airplane ticket online knows that prices on a particular ticket is always constantly varying. Main objective is to analyse and build a dynamic machine learning model that can predict the flight prices on the basis of information of the flight provided by the airlines.

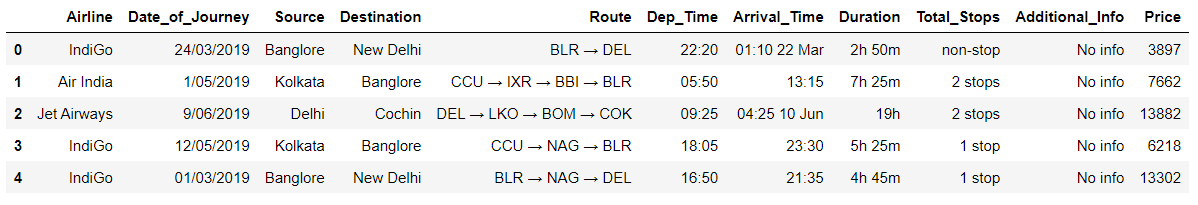
3. EXPLORATORY DATA ANALYSIS   
This step refers to drawing various insights from our dataset.

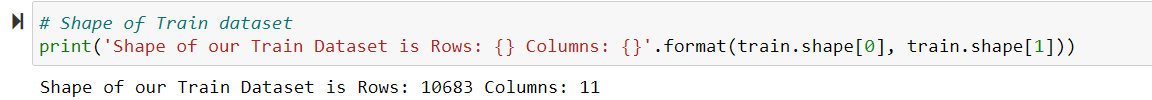
Dataset consists of the following variables:



Few observations about some of the variables:

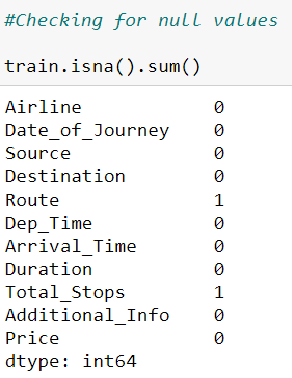
* ‘Price‘ will be our dependent variable and all remaining variables can be used as independent variables.
* ‘Total\_Stops‘ can be used to determine if the flight was direct or connecting.





Null and Duplicates.

- Upon analyzing our dataset, I came across a two null values which were dropped.



- There were exactly 220 duplicate fields which comprised of approximately 0.20% of our complete dataset; hence it seemed appropriate to drop the duplicates.

There's only one numerical feature in our dataset i.e. ‘Price’, whose statistical description is as follows:



- maximum price of a ticket : 79512.00

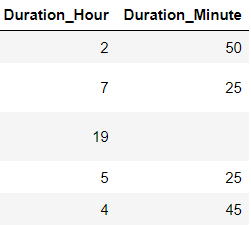
- minimum price of a ticket : 1759.00

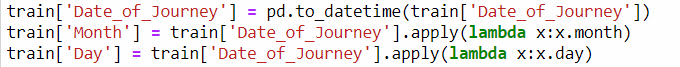
- average price of a ticket lies between : 8266.00 - 9026.79

1. Data Cleaning/ Feature Engineering :

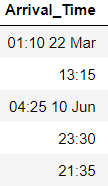
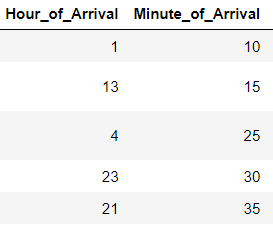
Our variables like Date\_of\_Journey, Arrival\_Time, Dep\_Time were comprised of date-time formats which needs to be handled before performing our analysis. So I've used pandas date-time function and lambda funtion to segregate:

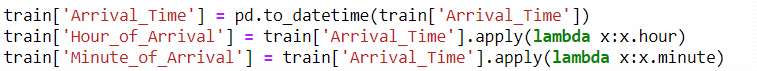
* Date and Month from Data\_of\_journey (Note : I'm not defining a new column for year because all our data is from same year i.e. 2019)

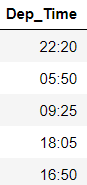
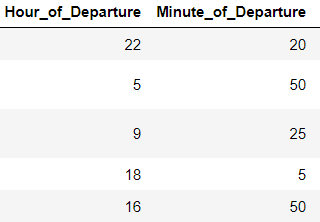


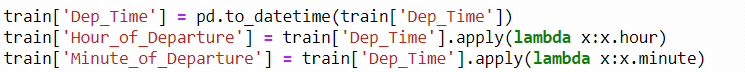
* Hour and Minute from Arrival\_Time

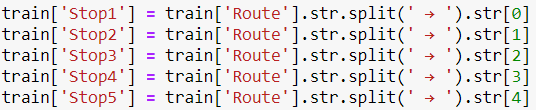


* Hour and Minute from Dep\_Time

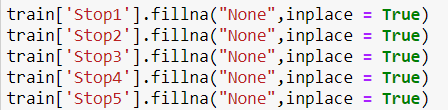
 

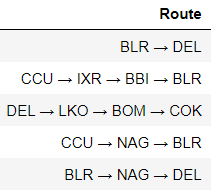
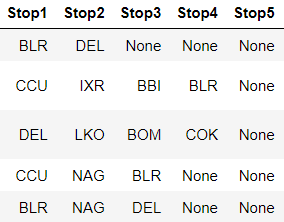


* In the "Route” column we either have direct flights from source to destination i.e. with 0 stops or at most we've 4 stops , I've separated all routes into stops.

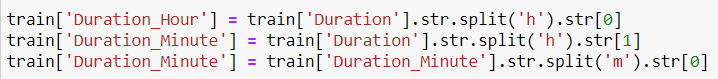


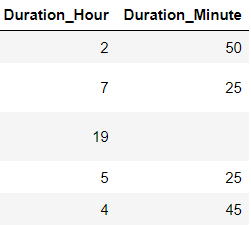
* Futher, I've replaced null in the new columns with None to identify there is no stopping point.

****

* Similarly, split duration hour and minutes



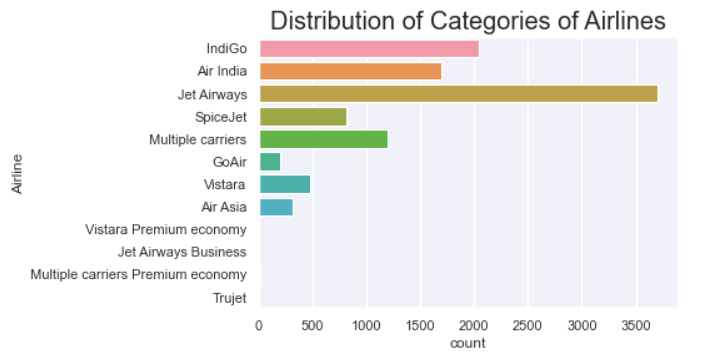


* After creating of new features I again checked for null values and dropped them.
* Lastly, I've dropped columns which were used to derive other columns i.e. "Date\_of\_Journer", "Arrival\_Time", "Dep\_Time", "Route" and "Duration".

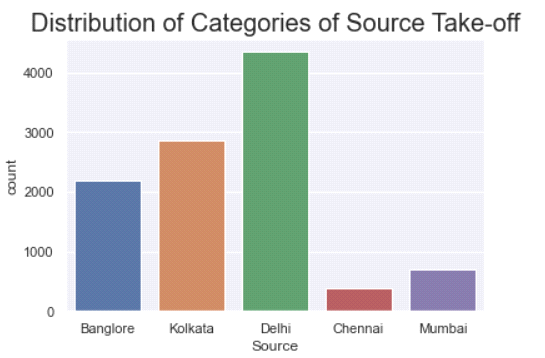
5. Visualizations

Key aspect of any data analysis is through visual representations. In this section, I've used matplotlib and seaborne to visualize our dataset.

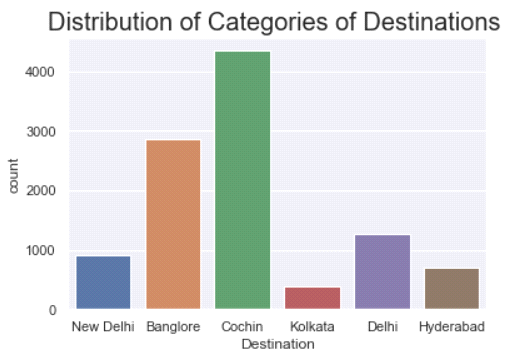
**Univariate Analysis**: To represent the data distribution of every feature



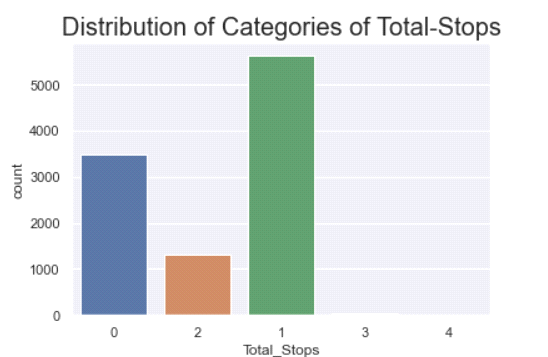
* It is to be noted that in our dataset particularly "Jet Airways" have made major appearance.



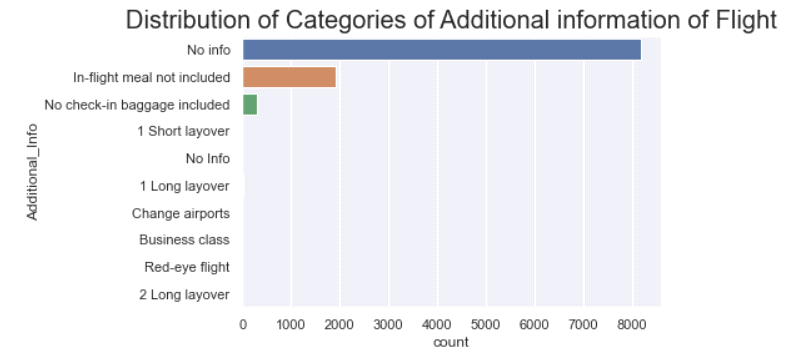
* There are basically 5 take- off points in our dataset out of which "Delhi" is mostly recurring.



* There are 6 destinations out of which "Cochin" is the most recurring.



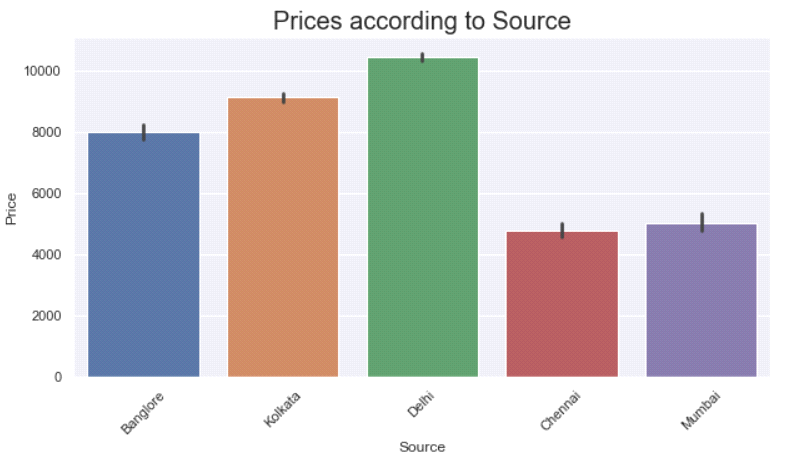
* Most Flights atleast have 1 stop in between the source and destination.



* In our dataset most Airlines have not provided any additional information.

**Bivariate Analysis:** To analyze how “price” is affected by every feature.









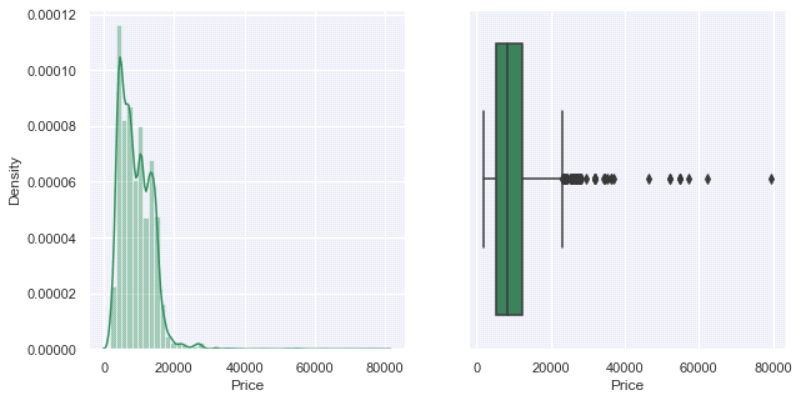
Insights:

* Jet Airways Business flight prices are really expensive whereas Trujet’s are inexpensive.
* Flights taking – off from Delhi are expensive.
* Flights landing in New Delhi are expensive
* Overall, a flight to or from Delhi costs more than other cities.
* Higher the number of stops higher will be the flight prices.

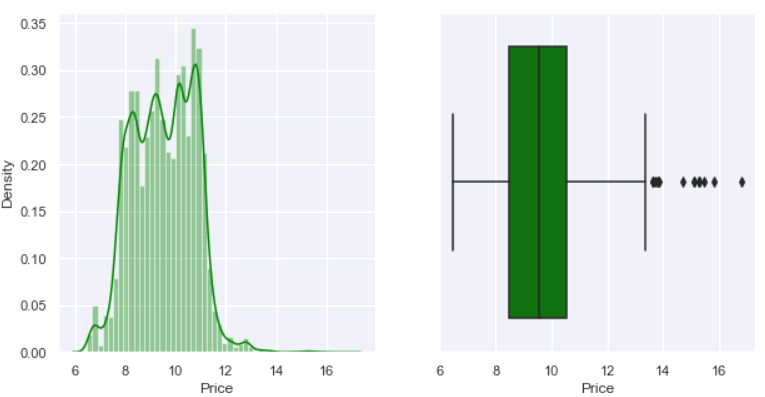
1. Data Preprocessing

6.1 Removing Skewness:

We only have one continuous variable i.e. "Price", which is right - skewed. To normalize the target variable I've used square-root transformation.



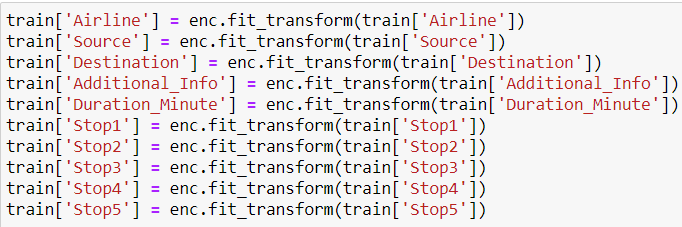
After applying square- root transformation :



6.2 Encoding Categorical Data

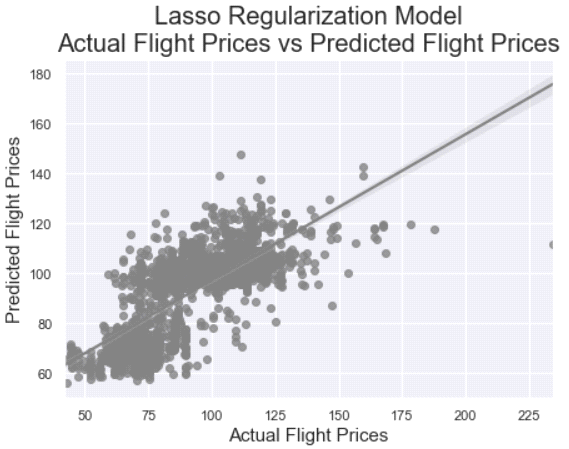
Finally, I've encoded our categorical features using label encoder and further split our dataset in ratio of 80% training and 20% testing .

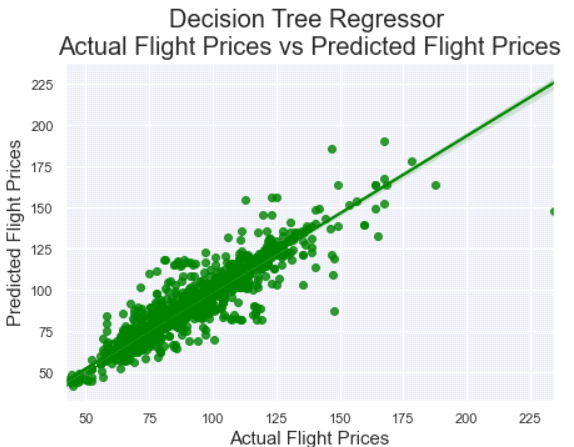
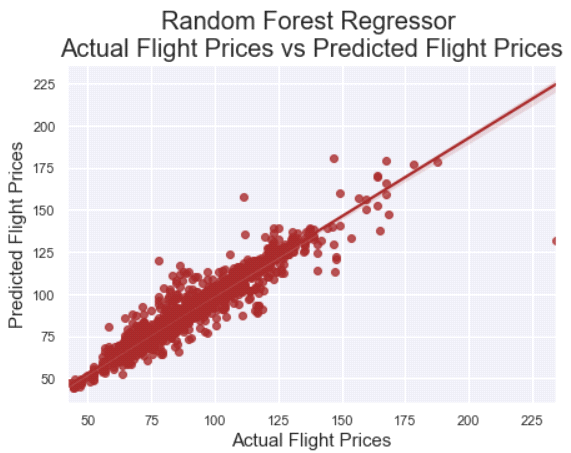


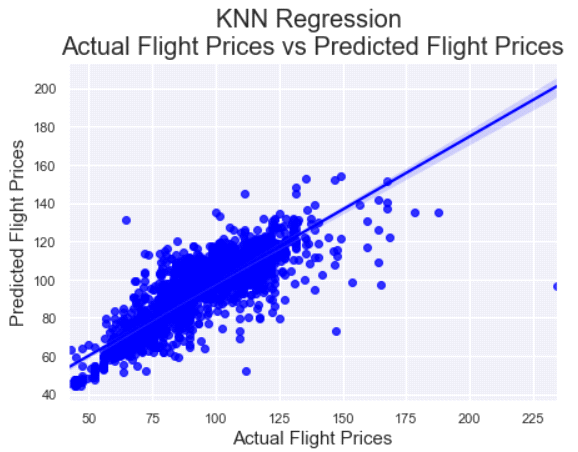
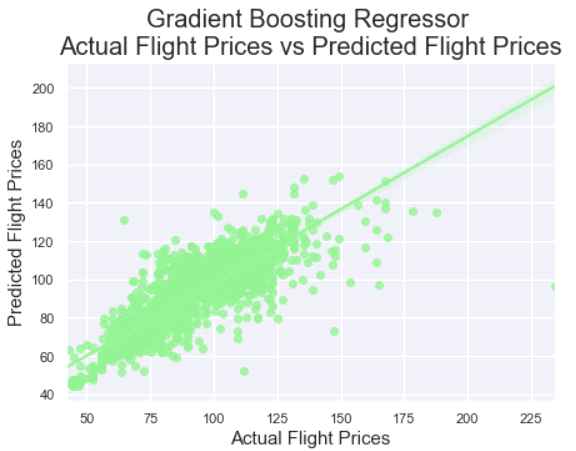


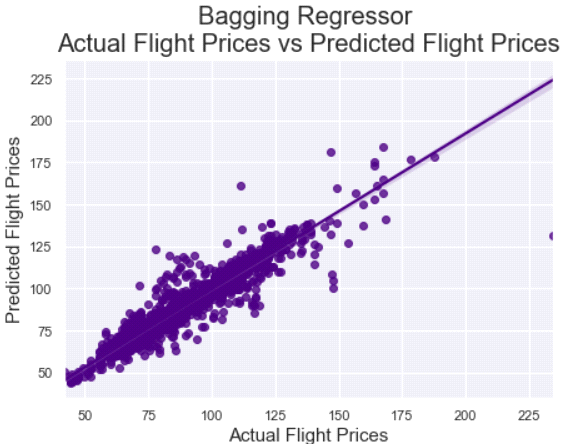
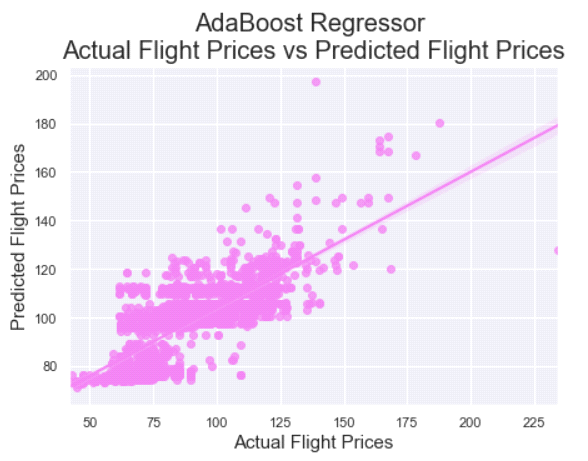
1. Models Used

I've used Linear Regression, Lasso Regularization, Decision Tree Regressor, KNN Regressor and ensemble models like Random Forest Regressor, Gradient Boosting, Bagging ,Adaboost. Out of all the model ensembles models performed really well. Here are some visualization of model performances and their fitted regression line covering the datapoints.

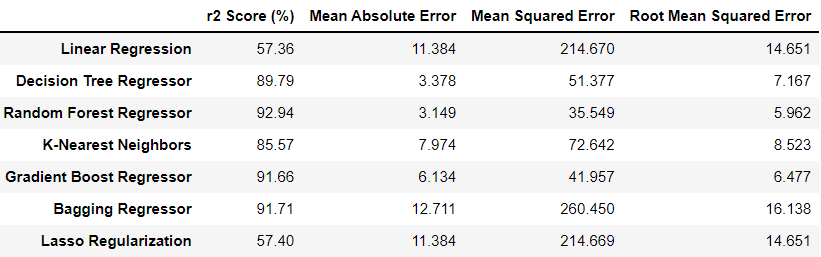
 

Summation of all the evaluation metrics and how our models performed :



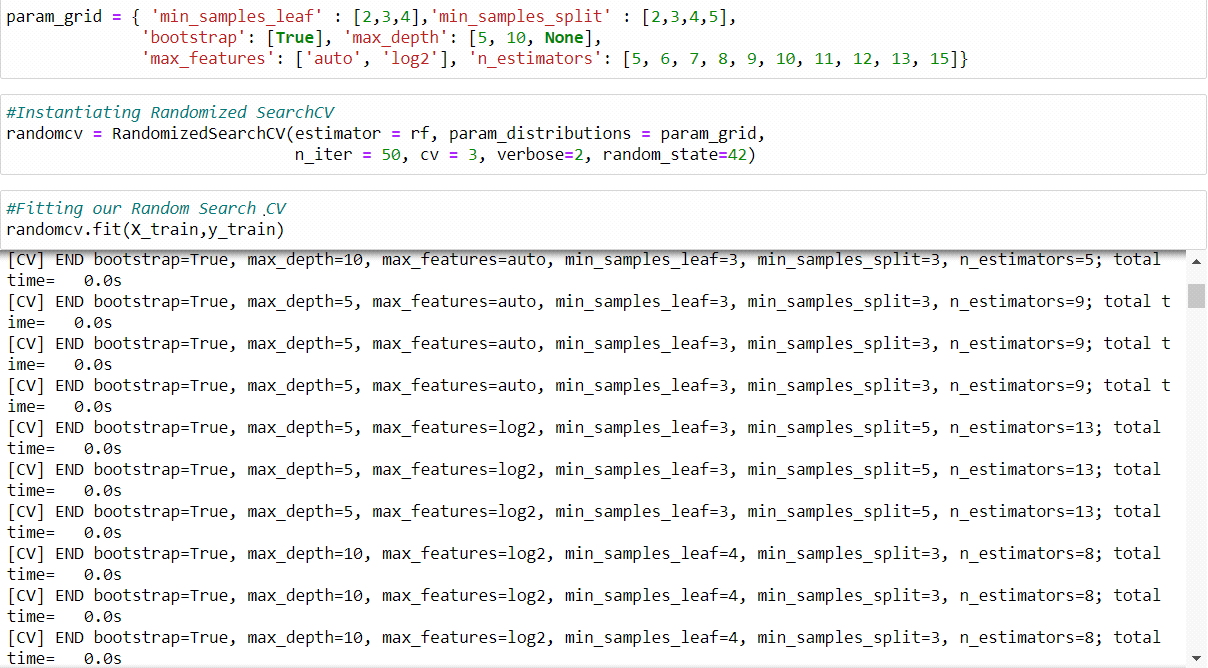
The least performing model in terms of r-squared score is Linear Regression model. .However, note that the best performing models were the machine learning ensemble regressor (random forest, Bagging, and Gradient Boost). Random Forest slightly outperformed the Gradient classifier and bagging classifier.

1. Hyperparameter Tuning

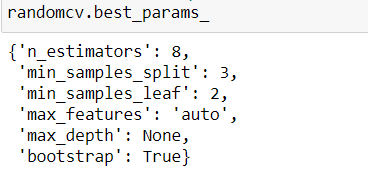
Hyperparameter tuing is performed to optimize the parameters of a model to give us the best results. It is used to tackle model underfitting as well as model overfitting problem.

Here, I've used RandomizedSearchCV to optimize parameters of my best model i.e. Random Forest Regressor model for this particular dataset.

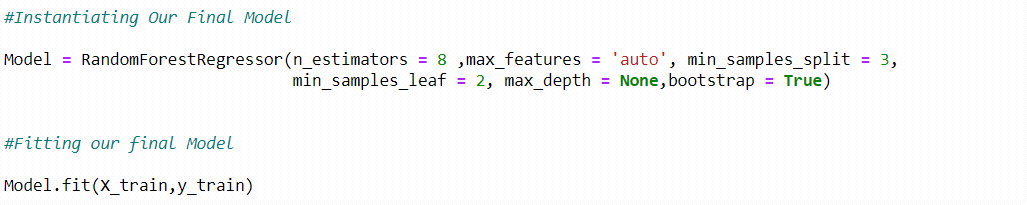
The following parameter are passed to randomized search cv to select the best parameter for model optimization.



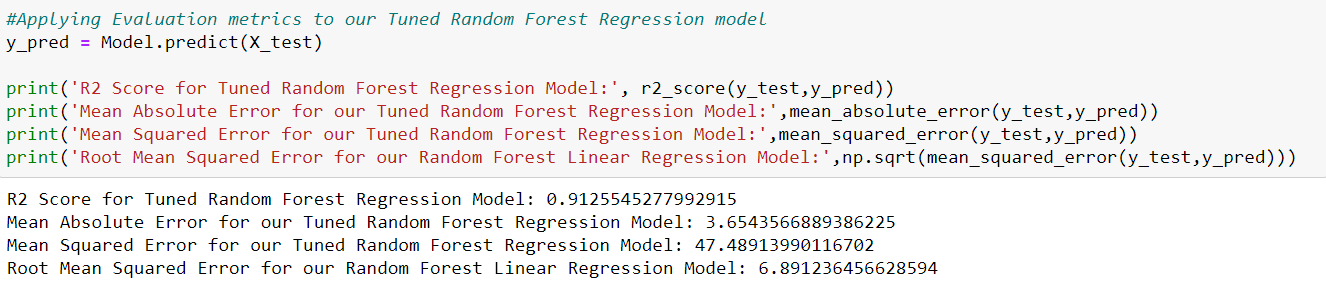
best parameters idenfied :



Finally, we re-instantiate and fit our model with the given best parameters.

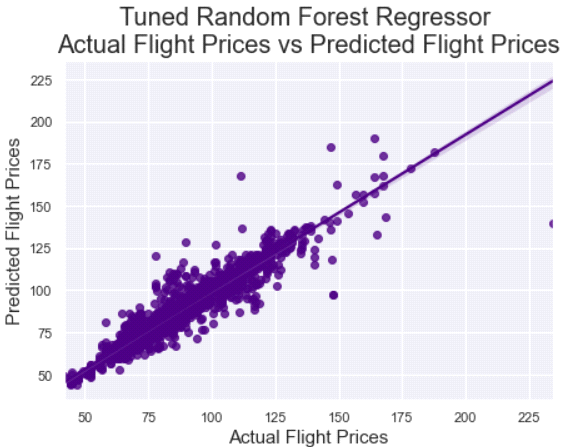


Finally performing evaluation metrics on our tuned Random Forest Regressor model



Evaluation metrics for the final model. It is seen that our model was little overfitted. The final r-squared score of the model is 91.71%, Along with MAE - 3.654, MSE - 47.4891 and RMSE - 6.891

Visualizing our final Regression model. We can clearly see how well the regression line is passing through majority of data points.



9. Conclusion

This article evaluated various Machine Learning Regression Algorithms to predict flight prices from the given various Indian Airlines Dataset. We compared different regresssion algorithms on the basis of their performances and evaluation metrics such as R- squared, MAE, MSE and RMSE evaluation metrics.

It's evident that tree-based algorithms have outperformed other algorithms. Random Forest Regressor have performed really well on our dataset.

I hope this Article helped you grasp some of the applied machine learning techniques in Data cleaning and preprocessing, applying machine learning algorithms, identifying an evaluation metric, hyperparameter tuing and to sum up building a machine learning regression model from scratch.

For any further queries please feel free to reach out!

Thank you and happy learning!