PROJECT TITLE

FLIGHT PRICE PREDICTION PROJECT

Submitted by:

MOHAMMED ZAHED N

**ACKNOWLEDGMENT**

Below are some of the sources from which code snippets have been helpful during the project completion

**Refernces:**

1. <https://www.scikitlearn.org>

2. <https://www.askpython.com>

3. <https://www.stackoverflow.com>

4. <https://www.geeksforgeeks.org>

**INTRODUCTION**

* Business Problem Framing/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 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.

.

* Conceptual Background of the Domain Problem

Some relevant columns in the dataset:

**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

**Data Analysis**

* Mathematical/ Analytical Modelling of the Problem

Many Statistical models were used also mathematical models necessary, Some of the used models are listed below.

Statistical Model:

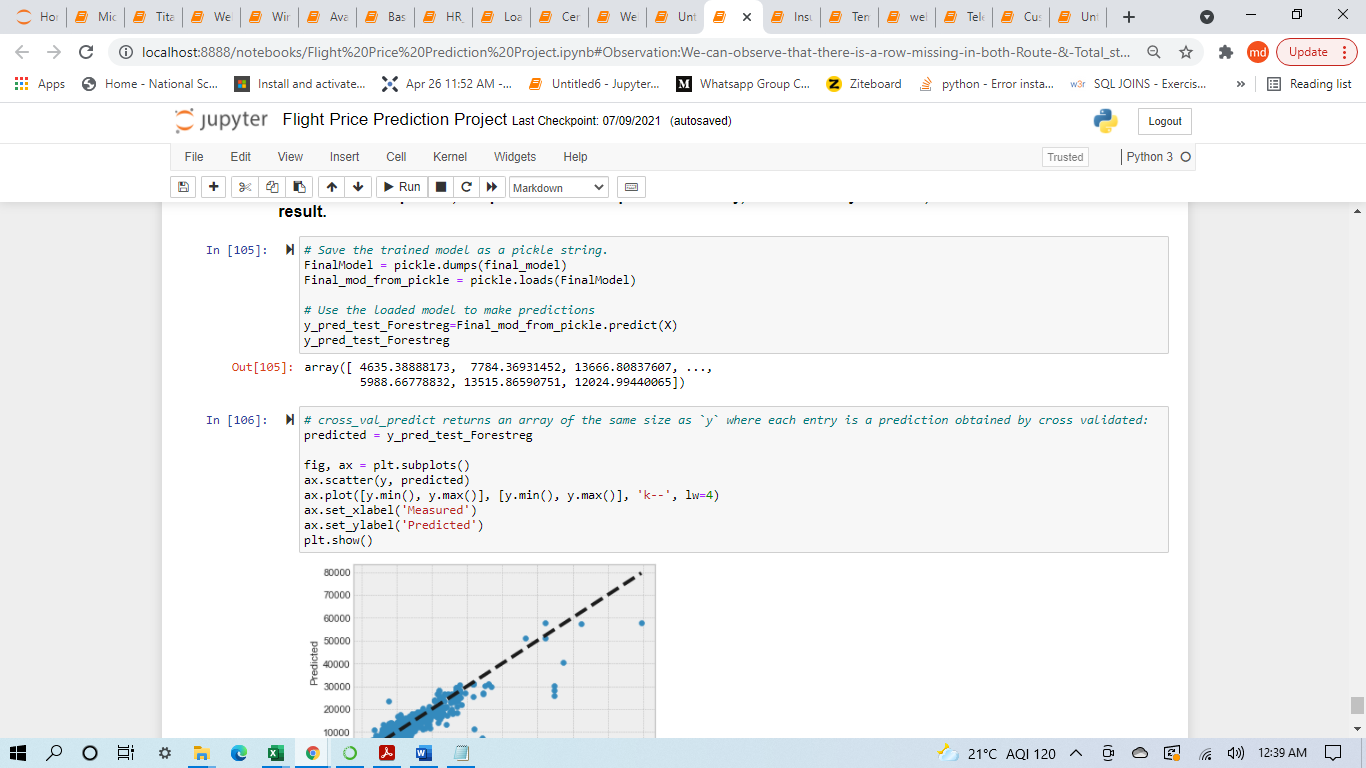
1. Pandas Profiling
2. Label Encoders
3. Co-relation metrics
4. Linear Regressor
5. Random Forest Regressor
6. Decision Tree Regressor
7. LogisticRegression
8. KNeighborsRegressor
9. GradientBoostingRegressor
10. SVR

* Data Sources and their formats

Most of the Code snippets and their formats are taken from below site

<https://www.geeksforgeeks.org>

Code on how to save a model using pickle library



* Data Preprocessing Done

1. Checking Missing Values & handling it

2. Data Description

3. Dropping Unnamed column as it does not contain any much significant data in it

4. Converting Categorical Columns Such as Type Column

5. Changing data types of 'Date\_of\_Journey','Dep\_Time' & 'Arrival\_Time' from object type to datetime.

6. Extracting day,month,hour & Mins data from 'Date\_of\_Journey','Dep\_Time' & 'Arrival\_Time' columns.

* Data Inputs- Logic- Output Relationships

Most of the data in the dataset was already Numerical except for few columns after data cleaning and processing sent to model as independent features and run Test & Train along with dependent feature as output variable.

**Independent Features->Dependent Feature->Model(Prediction)**

* Hardware and Software Requirements and Tools Used

List of hardware and software requirements along with the tools, libraries and packages used.

**Softwares:**

1. Ms Word: Ms Word for documentation purpose

2. Ms Excel: To view Dataset and to perform basic subtraction on columns for comparion

3. Jupiter Notebook(Anaconda):To run python code for building suitable model for prediction.

**Hardware:**

OS: Windows 10

RAM: 4GB

HDD: 1TB

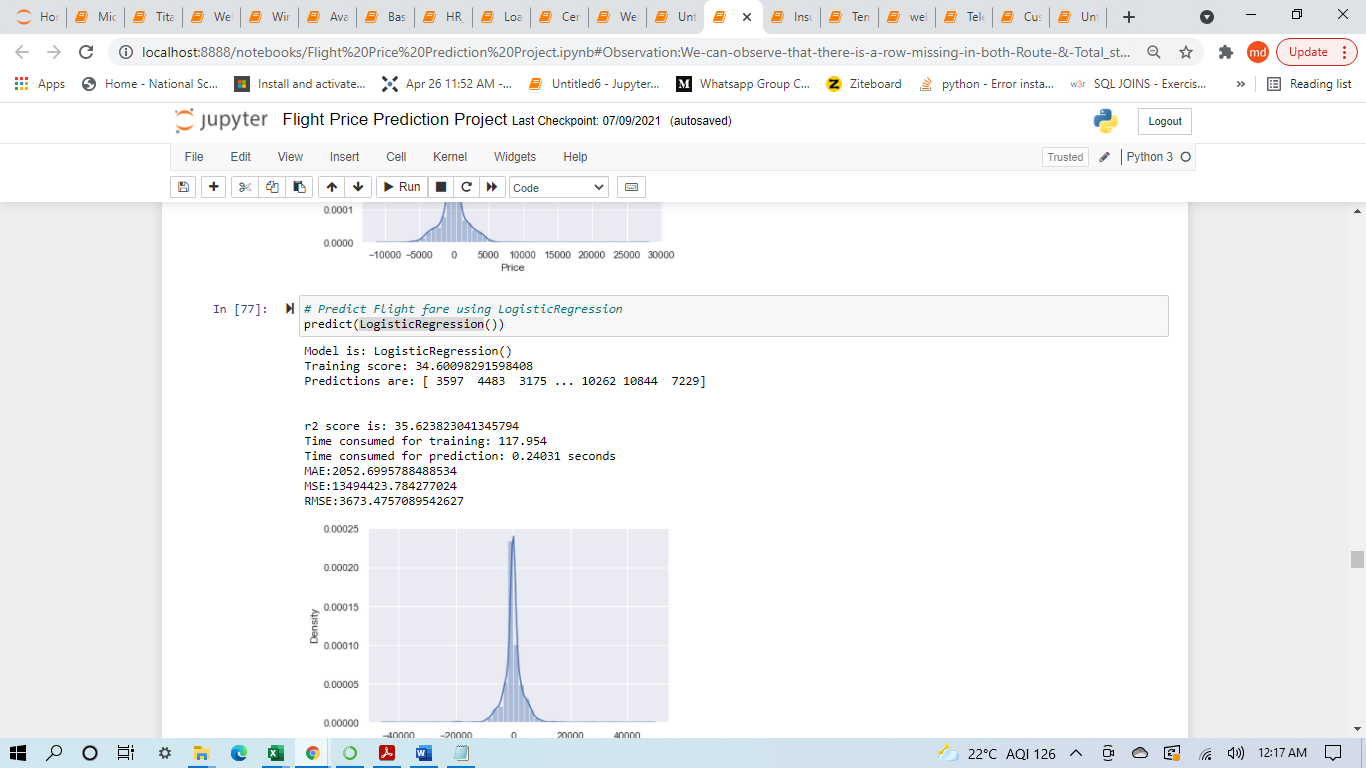
**Model/s Development and Evaluation**

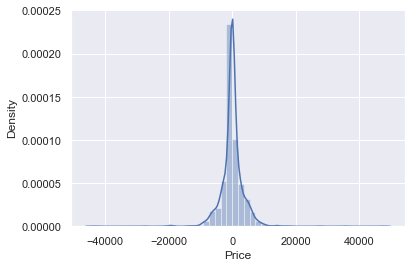
* Identification of possible problem-solving approaches (methods)
* Loading the data into Jupiter Notebook
* Checking the basic details (Null Value, Datatype, Shape etc.)
* Identifying the target and independent features and perform EDA (Univariate, Bivariate and Multivariate analysis) using Data Visualization and Statistical approach accordingly.
* Performing data cleaning, outliers handling, missing value imputation
* Building model
* Evaluating the built model for accuracy score
* Performing hyperparameter tunning to enhance the performance
* Evaluating the model again
* Make prediction
* Testing of Identified Approaches (Algorithms)

1. Random Forest Regressor
2. Decision Tree Regressor
3. LogisticRegression
4. KNeighborsRegressor
5. GradientBoostingRegressor
6. SVR

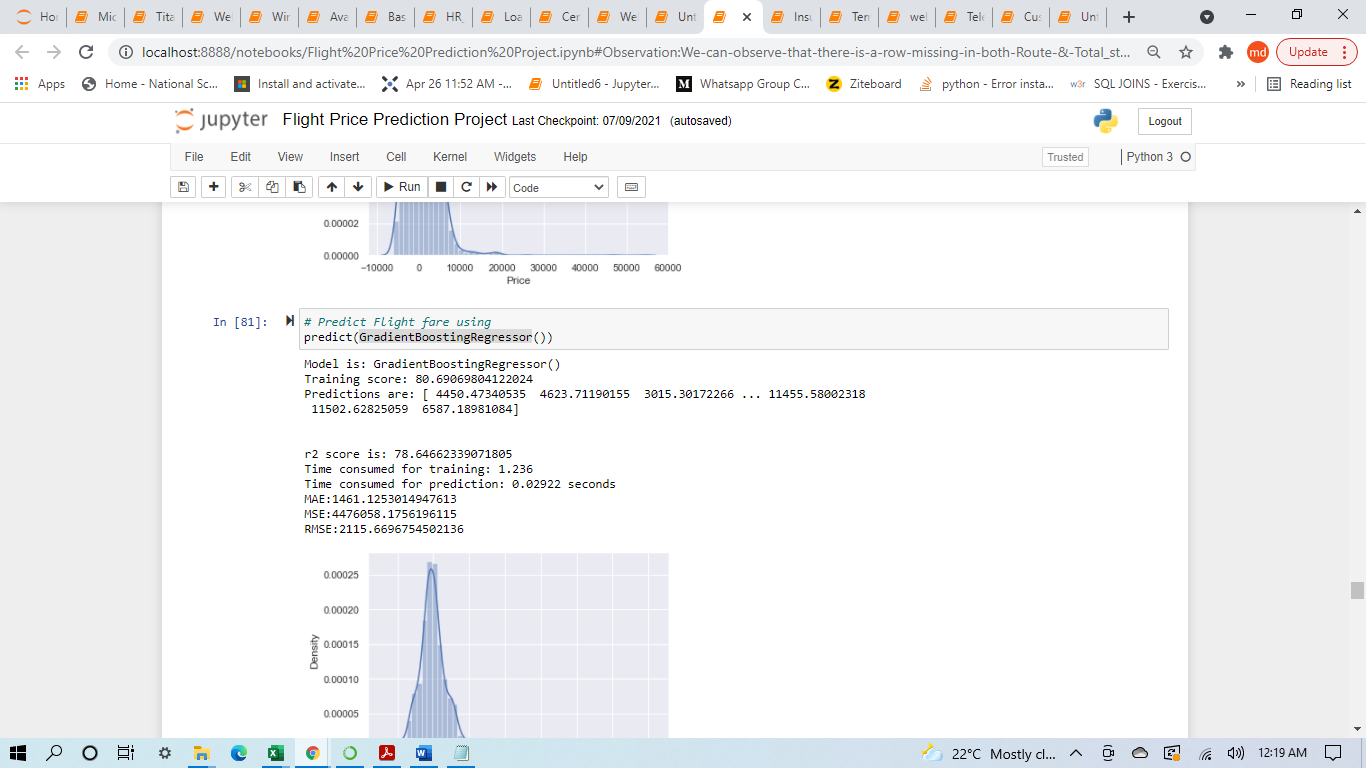
These were the algorithms that are used for training and testing purpose.

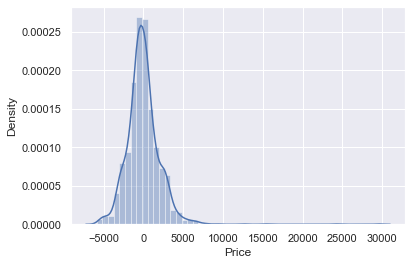
* Run and evaluate selected models
* Logistic Regression: The Accuracy found by using this Model is least which is 35.62%.



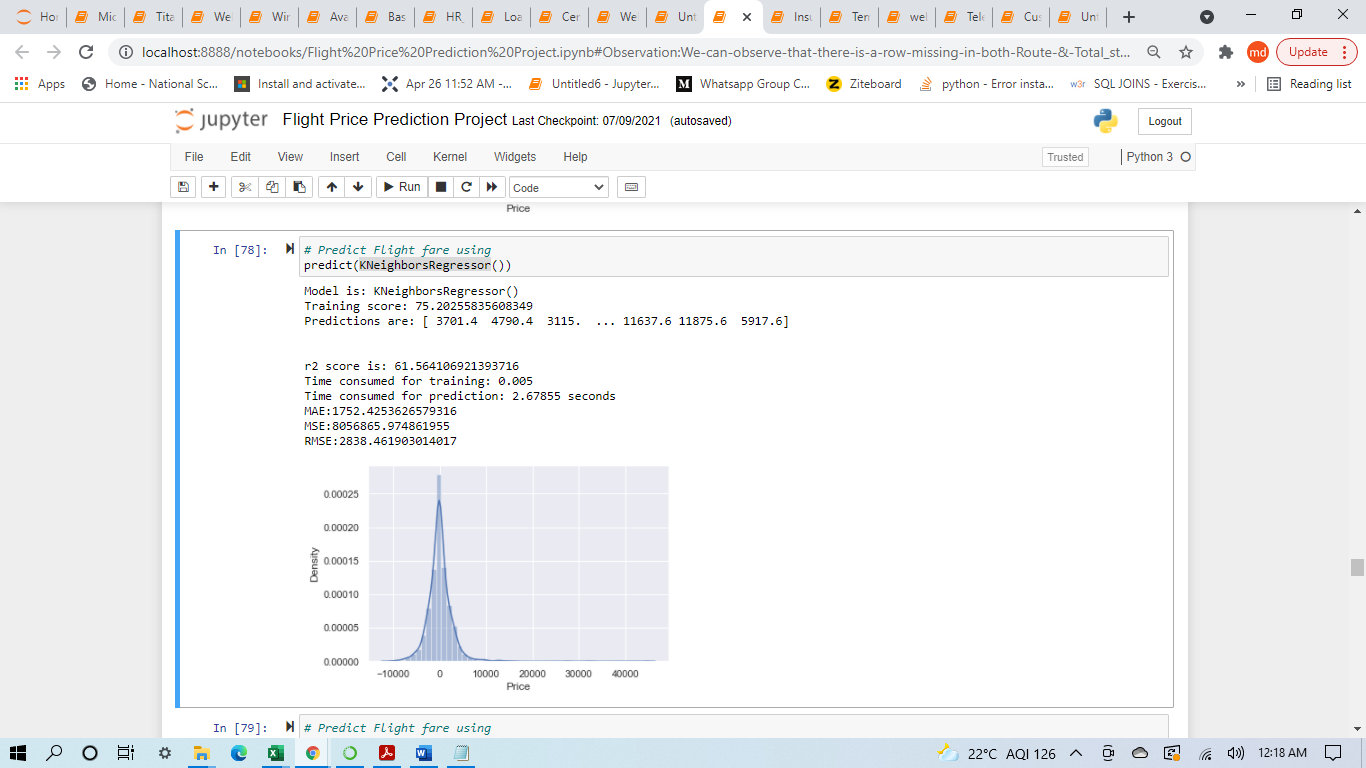


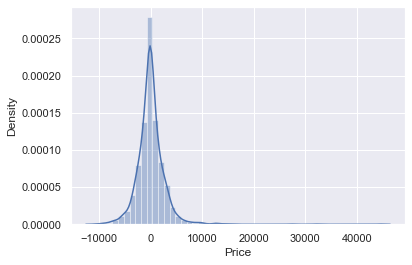
* Gradient Boosting Regressor: The Accuracy found by using this Model is least which is 78.64%.



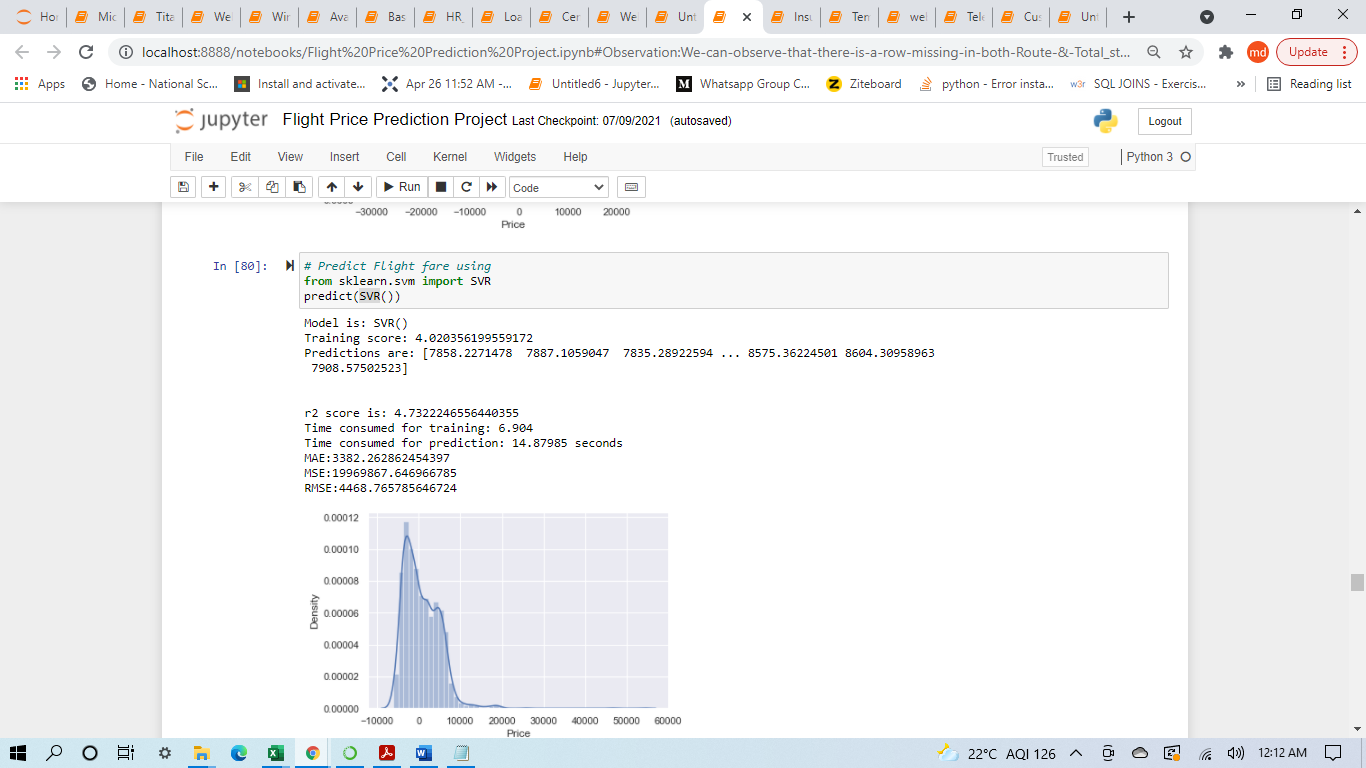


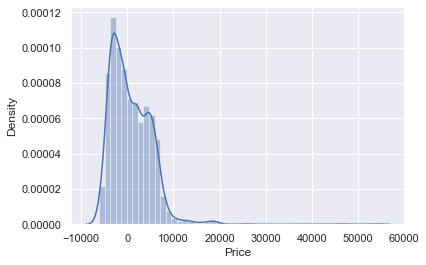
* KNeighborsRegressor: The Accuracy found by using this Model is least which is 61.56%.



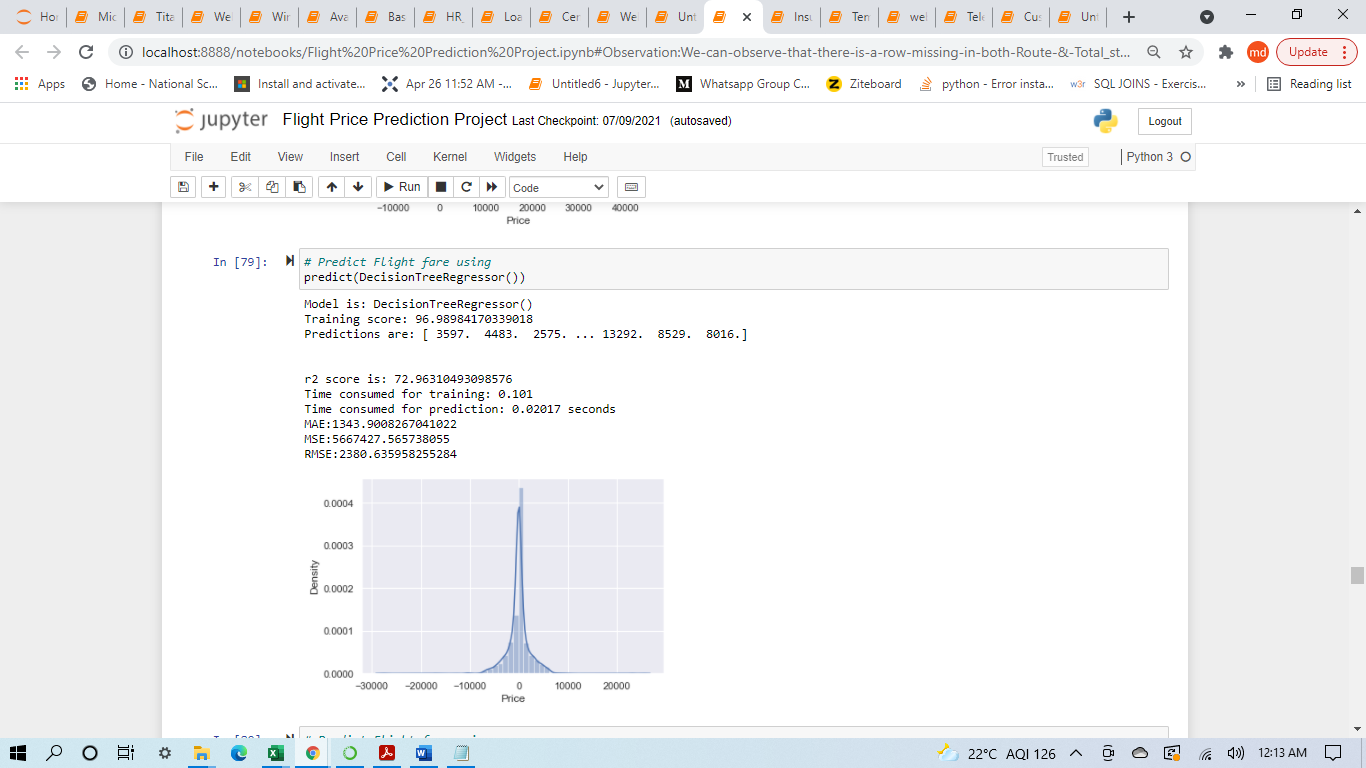


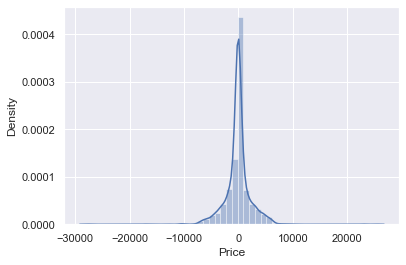
* Support Vector Regressor: The Accuracy found by using this Model is least which is 47.32%.



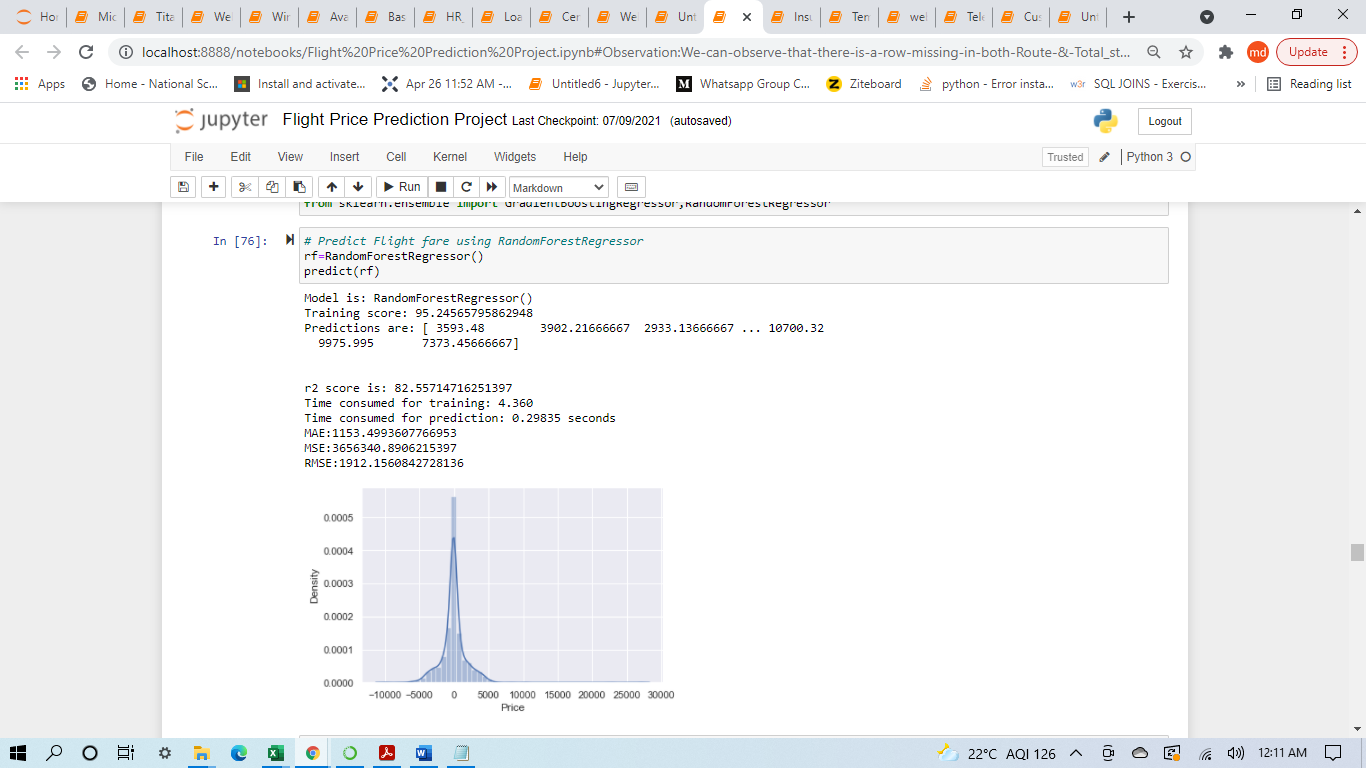


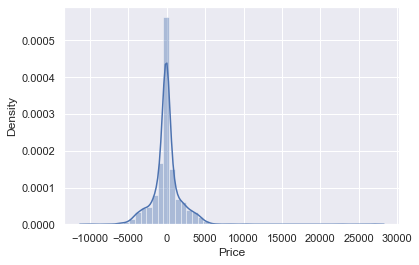
* Decision Tree Regressor: The Accuracy found by using this Model is 72.96%.





* Random Forest Regressor: The Accuracy found by using this Model is 82.55%.





* Key Metrics for success in solving problem under consideration

**Hyper Parameter Tuning** was used to maximise the output of the model but there was not much significant change observed compared to normal accuracy achieved from the selected Model.

In conclusion, the metrics for our best models on the training set after our hyperparameter fine-tuning are the following:

**Grid-search:**

**RMSE: 1563.73**

**accuracy: 87.04%**

**Randomized Search:**

**RMSE: 1328.04**

**accuracy: 91.36%**

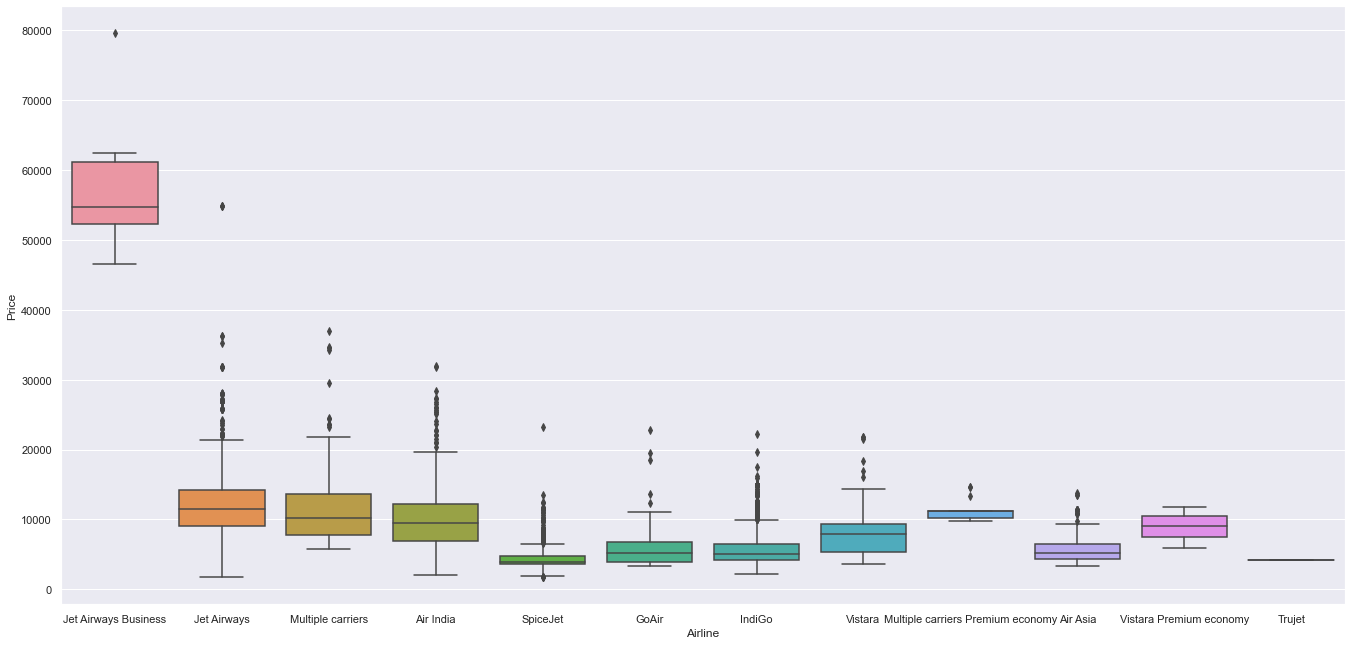
This of course does not mean that the Grid search is inherently inferior to Randomized -search.

Its just that in this notebook I used the information provided by the former to refine the latter.

In any case, this time I will consider the model obtained by the Randomized search as the best one.

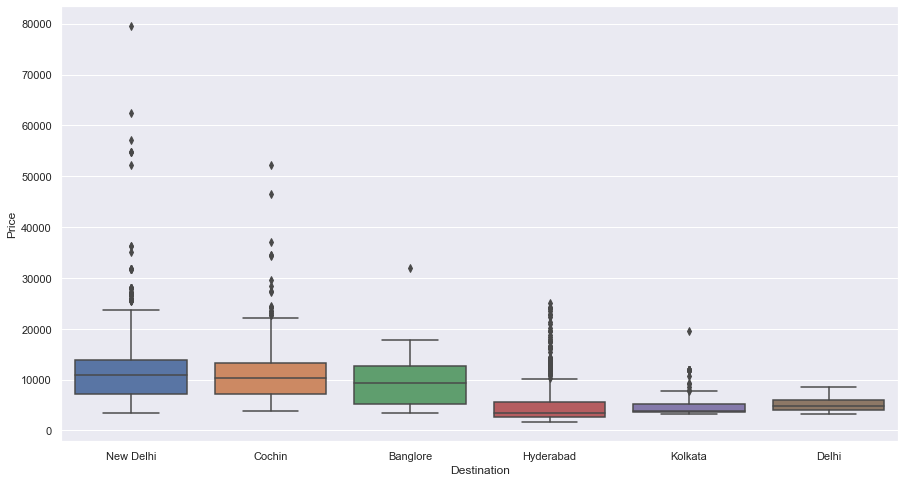
* Visualizations

1. This is a Airline vs Price Analysis Boxplot.



**Observation:** From graph we can see that passengers with Total\_stops equal to 3 & 4 stops have the highest Price

1. This



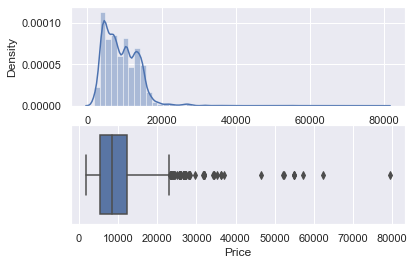
**Observation:** From graph we can see that people traveling back to Delhi (Destination) have the highest Price

1. Hexbin Plot: Arrival\_Time\_hour Price Analysis as both are numerical columns will use hexbin plot.

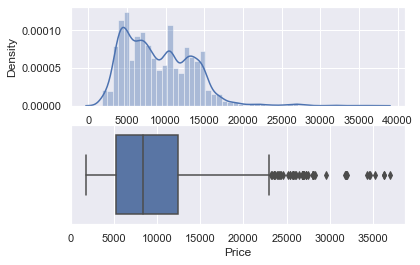


**Observation:** We can observe the above plot most of the price lies between 6000-20000

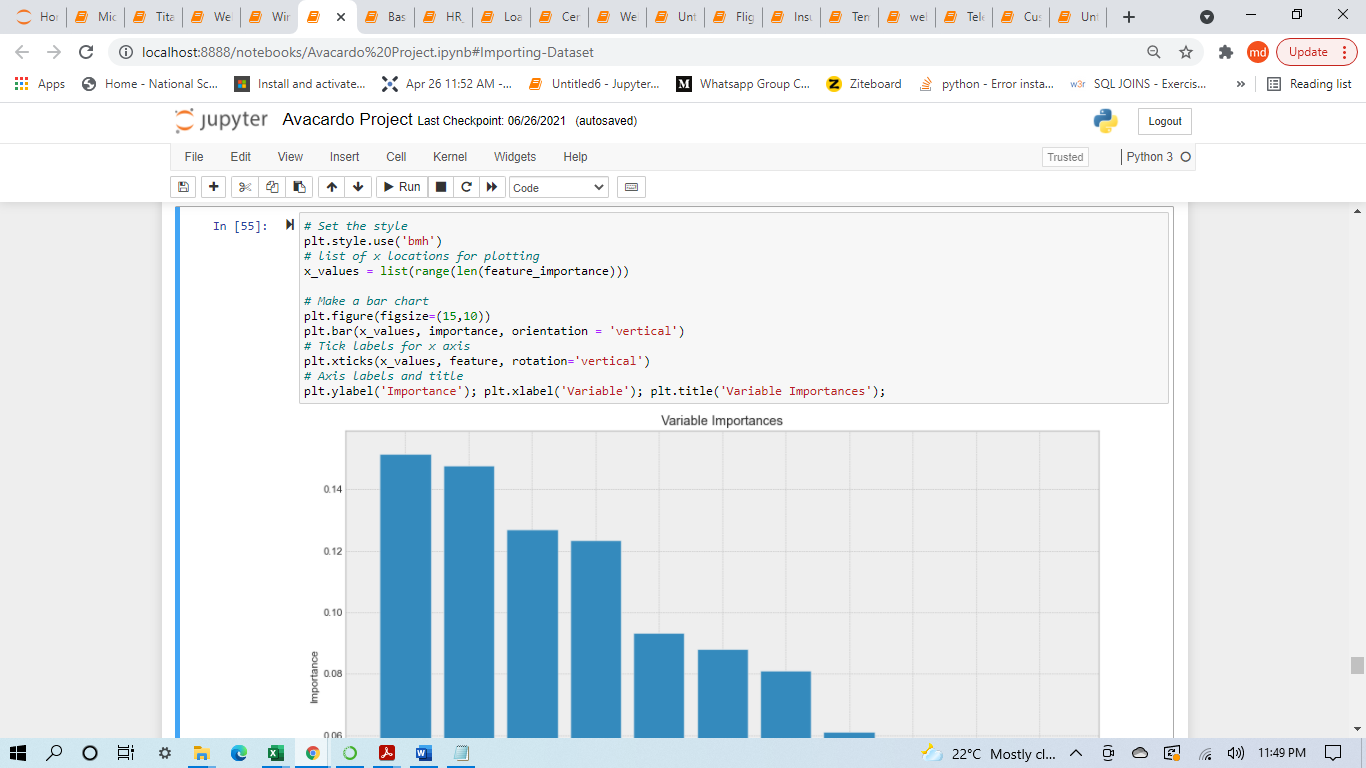
**Outliers:** Check For Outliers



**Handling outliers:** As there is some outliers in price feature,so we replace it with median.



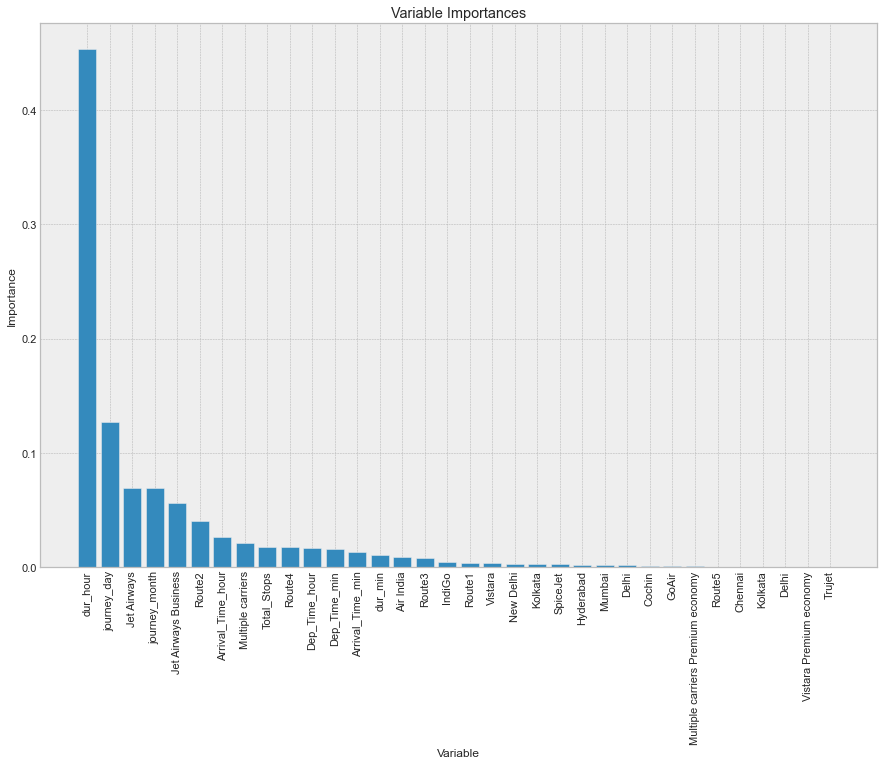
* Feature Importance



**Observation:** A future aim may be to cut the less relevant features (lets say we can drop large bags in terms of importance),

estimate a new model and compare it with the old ones.

I reckon it would lose predictive power, but on the other hand it would improve in terms of training speed.

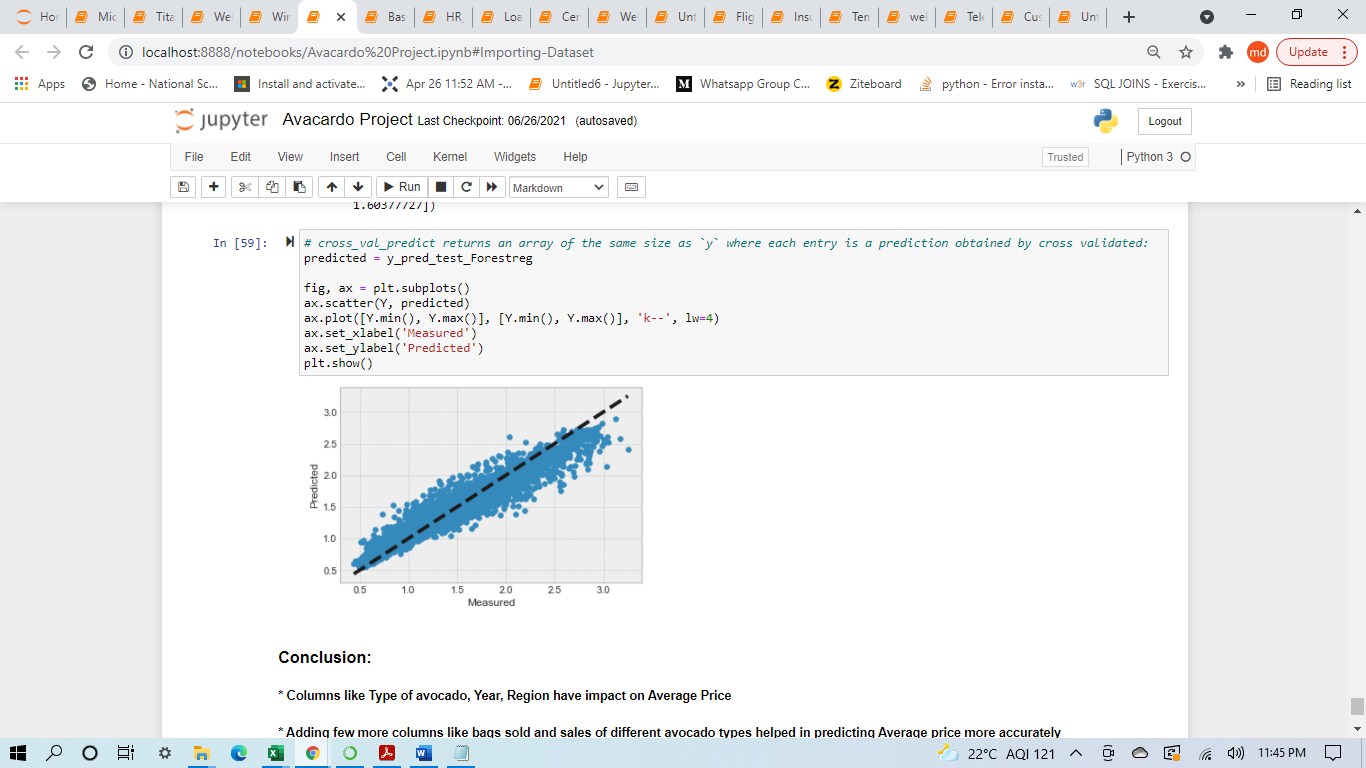


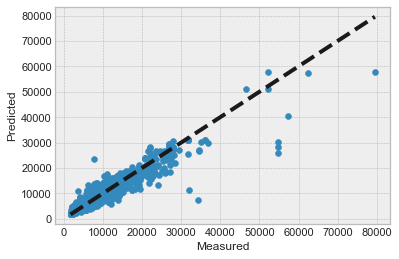
**In conclusion, these are my final considerations on the model:**

1. The best feature to reliably predict the Flight fare price is Type of dur\_hour. After dur\_hour there is a group of 6 features including: journey\_day,Jet Airways,journey\_month,Jet Airways Business,Route2,Arrival\_Time\_hour & Multiple carriers.

2. Trujet is the least relevant feature in this cluster.

* Predicted vs Actual





**CONCLUSION**

* Key Findings and Conclusions of the Study

1. Columns like dur\_hour,journey\_day,Jet Airways,journey\_month,Jet Airways Business,Route2,Arrival\_Time\_hour & Multiple carriers have effect on Flight Price.
2. Choosing a Airline can significantly effect the flight fare and predicting it.
3. The Actual vs Predicted plot clearly indicates the predicted values are almost linear hence performance of model is considerably Good
4. Random Forest Regressor model predicts the Flight price more accurately than linear regression model