|  |  |
| --- | --- |
| Flip Robo Technologies | LinkedInFlight Ticket Price Prediction   |  | | --- | | Submitted By,  Poovarasi Vijayan | |

**Acknowledgement:**

I wish to express my sincere thanks to the following companies, without whom I would have not got opportunity to work on this project; Data Trained Institute and Flip Robo Technology-Bangalore.

**Introduction:**

**Business Problem Facing:**

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.

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on –

1. Time of purchase patterns (making sure last-minute purchases are expensive).

2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases).

Here we are trying to help the buyers to understand the price of the flight tickets by deploying machine learning models. These models would help the sellers/buyers to understand the flight ticket prices in market and accordingly they would be able to book their tickets.

**Concept Background of the Domain Problem:**

The main aim of this project is to predict the price of flight tickets based on various features. The purpose of this project is to study the factors which influence the fluctuations in the airfare prices and how they are related to the change in the prices. Then using this information, build a system that can help buyers whether to buy a ticket or not. So that Machine Learning model is deployed for flight ticket price prediction and analysis. This model will provide the approximate selling price for the flight tickets based on different features.

**Review of Literature:**

As per the requirement of our client, data is scraped from [www.easemytrip.com](http://www.easemytrip.com), which is one of the platforms to book flight tickets. This project was built with a aim of gaining insights about various factors affecting the price of flight. This project is more about data exploration, feature engineering and pre-processing of data.

The goal of this project is to build an application which can predict the price of flight tickets with the help of the features available. In long term this would allow people to better explain and review their purchase in this digital world.

**Motivation for Problem Undertaken:**

Having lived in India for so many years if there is one thing that I had been taking for granted,

it’s that housing and rental prices continue to rise. Since the housing crisis of 2008, housing

prices have recovered remarkably well, especially in major housing markets. However, in the

4th quarter of 2016, I was surprised to read that Bombay housing prices had fallen the most in

the last 4 years. In fact, median resale prices for condos and coops fell 6.3%, marking the first

time there was a decline since Q1 of 2017. The decline has been partly attributed to political

uncertainty domestically and abroad and the 2014 election. So, to maintain the transparency

among customers and also the comparison can be made easy through this model. If customer

finds the price of house at some given website higher than the price predicted by the model, so

he can reject that house.

Having lived in India for so many years if there is one thing that I had been taking for granted,

it’s that housing and rental prices continue to rise. Since the housing crisis of 2008, housing

prices have recovered remarkably well, especially in major housing markets. However, in the

4th quarter of 2016, I was surprised to read that Bombay housing prices had fallen the most in

the last 4 years. In fact, median resale prices for condos and coops fell 6.3%, marking the first

time there was a decline since Q1 of 2017. The decline has been partly attributed to political

uncertainty domestically and abroad and the 2014 election. So, to maintain the transparency

among customers and also the comparison can be made easy through this model. If customer

finds the price of house at some given website higher than the price predicted by the model, so

he can reject that house.

This problem is taken based on the requirement of the client and also, with a curiosity to know how the flight of price changes frequently and on what basis it changes. And also, to get hands on experience and to know how it works on real bases.

**Analytical Problem Framing:**

**Mathematical/Analytical Modelling of the Problem:**

Effective Machine Learning model needs to be created which predicts the price of flight tickets. So ‘price’ is the target variable which is continues, so it is Regression problem where regression algorithm is used to predict the price of flight tickets.

Web Scraping is done to collect data from [www.easemytrip.com](http://www.easemytrip.com).

First, the analysis is started with importing the data, this dataset contains Null values in ‘price’ column as it is target variable all null values are dropped.

Once data is cleaned outliers and skewness are checked, if present they are removed, then Data Pre-processing, Standard Scaler is used to standardize the data and VIF is checked for multicollinearity in dataset.

Once this is all done then data is ready for modelling. As the target variable contains continuous data, Regression is used. 4 regressors – Decision Tree Regressor, Support Vector Regressor, KNeighbors Regressor, Linear Regression, 4 ensemble -Random Forest Regressor, AdaBoost Regressor, Gradient Boosting Regressor, XGB Regressor, 3 metrics – r2\_score, mean\_squared\_error, mean\_absolute\_error and 3 regularization – Lasso, Ridge, ElasticNet techinques are used in this project to build Regression model. From this the best model is identified and using Cross Validation technique is checked for overfitting and underfitting and Hypertuning is done to increase accuracy. Then, Finally the best model is saved.

**Data Source and their Format:**

The data is scrapped using selenium web scraping which is stored in xlsx format.

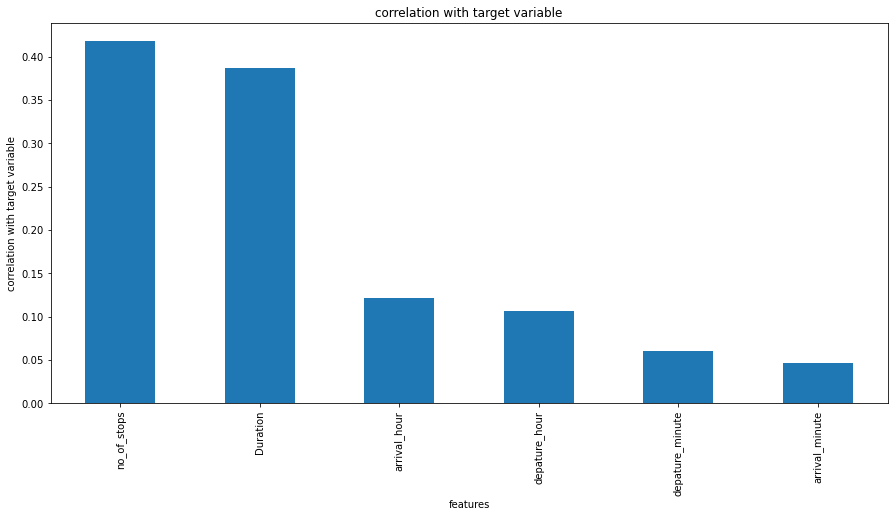
* Data contains 1829 entries each having 9 variables.
* Here the dataset is divided into independent and target variable.

**Data Pre-Processing:**

In Machine Learning, data pre-processing refers to the process of cleaning and organising raw data in order to make it appropriate for creating and training Machine Learning models. In other words, anytime data is received from various sources, it is collected in raw format, which makes analysis impossible. Data pre-processing is a crucial stage in Machine Learning since the quality of data and the relevant information that can be gleaned from it has a direct impact on our model's capacity to learn; consequently, we must pre-process our data before feeding it into our model. As a result, it is the first and most important stage in developing a machine learning model.

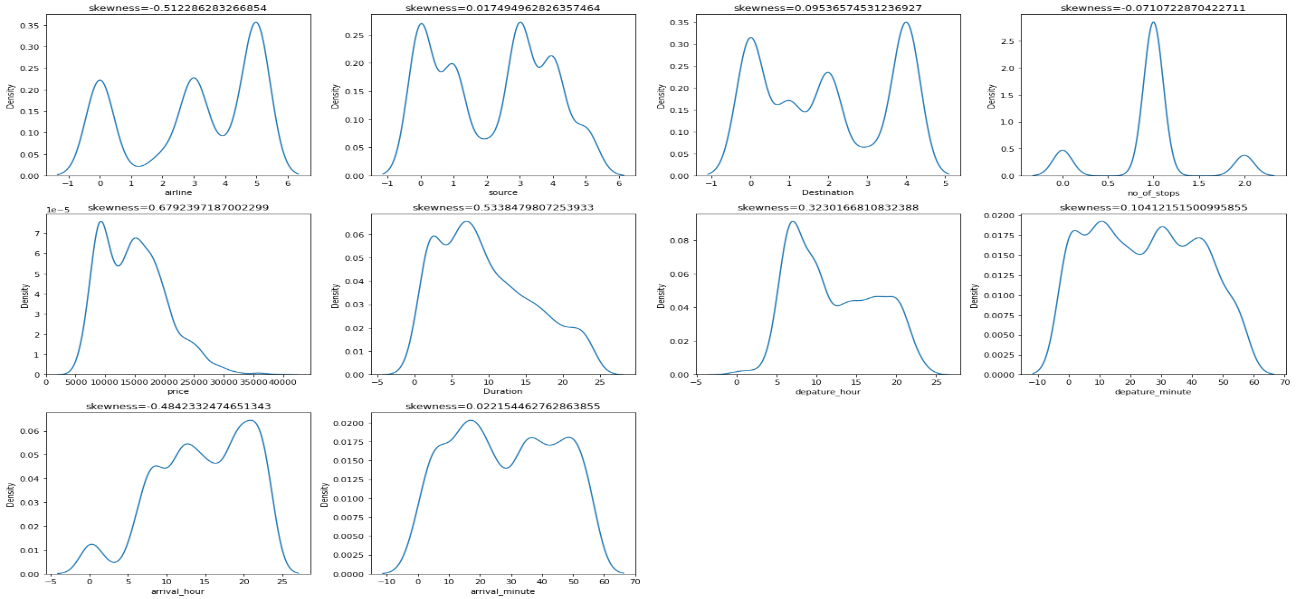
Some of the techniques used in this project are listed below:

1. Started with web scraping using selenium from [www.easemytrip.com](http://www.easemytrip.com) website. Different flights are scrapped using different locations. Then the scraped data is stored in xlsx format.
2. Then the required libraries are imported and then the dataset which is in xlsx format.
3. Dataset contains 1829 rows and 9 variables. Then the information of the columns is observed carefully.
4. Few columns like ‘Unnamed: 0’ is dropped as it does not contribute much in model building.
5. Then Null values is checked, as there are few null values in target column, null values are dropped.
6. Numeric data and Categorical data are then separated so that visualization is done with distplot for numeric data and countplot for categorical data.
7. Then all the categorical data is converted to numeric by using LabelEncoder so that analysis can be made in better way.
8. Data Description is made followed by correlation where positive and negative correlated data is checked.

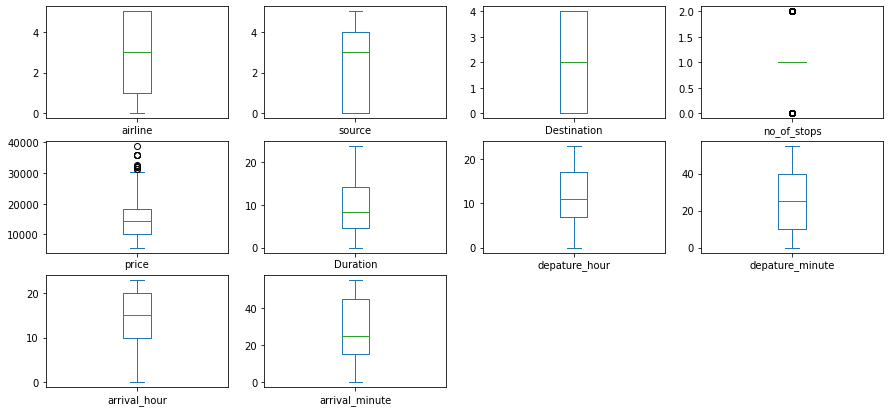


* 'no\_of\_stops',’duration’ is highly correlated with price which means if number of stops and duration decreases flight price increases.
* All other variables arrival\_hour, depature\_hour, depature\_minute, arrival\_minute are positively correlated with target variable.

1. Outliers and Skewness is checked and removed in order to avoid bias while model building.



Keeping +/- 0.5 as skew value, all columns have less than .5.



* Outliers are removed by using z-score method.

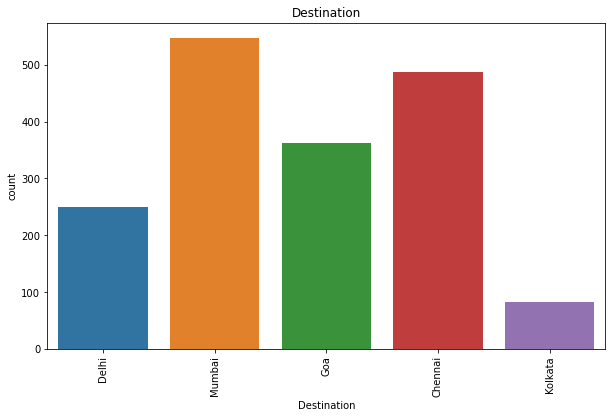
1. Standard Scaler is used for data standardization and VIF is used to check Multicolinearity is checked to find if any data variable is correlated with each other and it is removed.
2. Once all these processes are done then data is ready for model building where various Machine Learning models is used to check the accuracy of data.

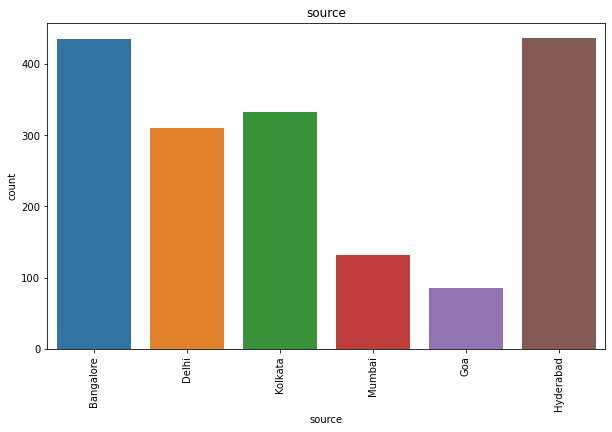
**Data Input Logic Output Relations:**

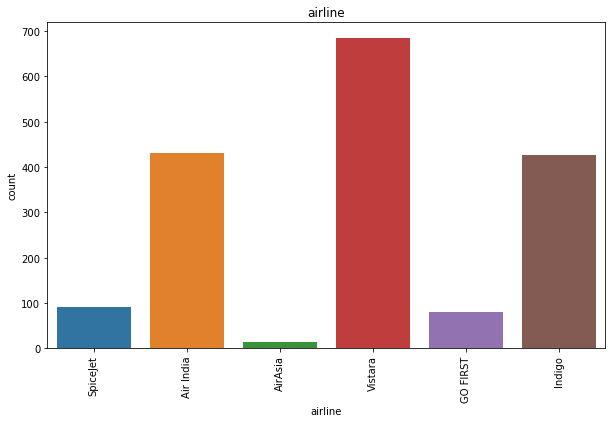
Data visualization is used to find the relation between input and the output variable.

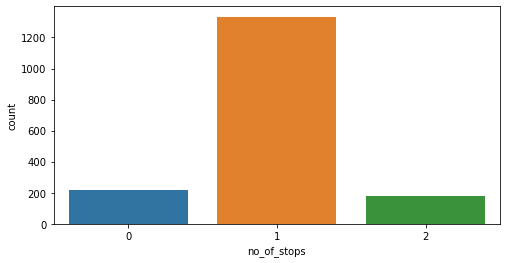
**Data Visualization:**

**Visualization of Categorical Variables:**





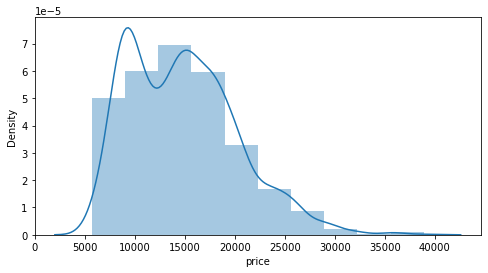




**Key Observations:**

1. Many people travel to Mumbai and Chennai followed by Goa.
2. Source is mostly in Bangalore followed by Hyderabad.
3. Vistara is mostly prefered airlines followed by AirIndia and Indigo
4. Our dataset contains many one stop airplanes compared to non-stop.

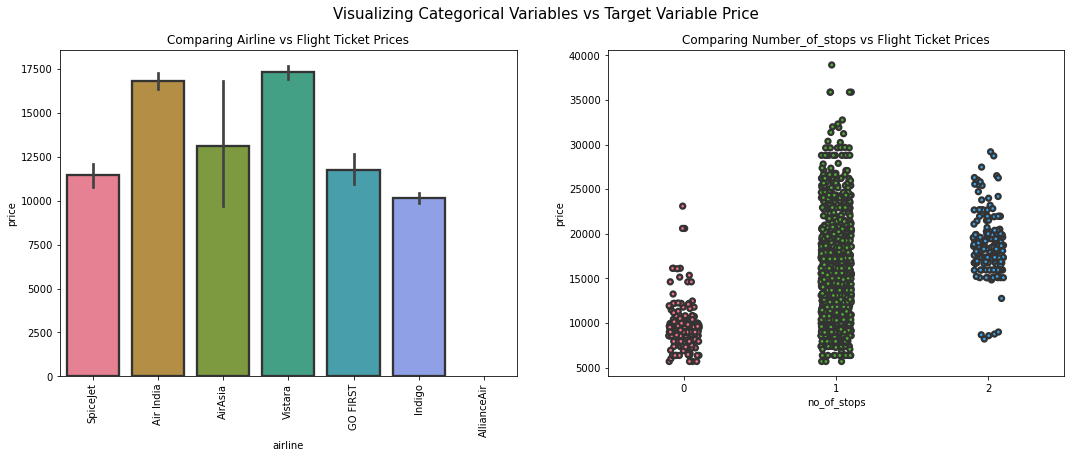
**Visualization of Numerical columns:**

****

**Key Observations:**

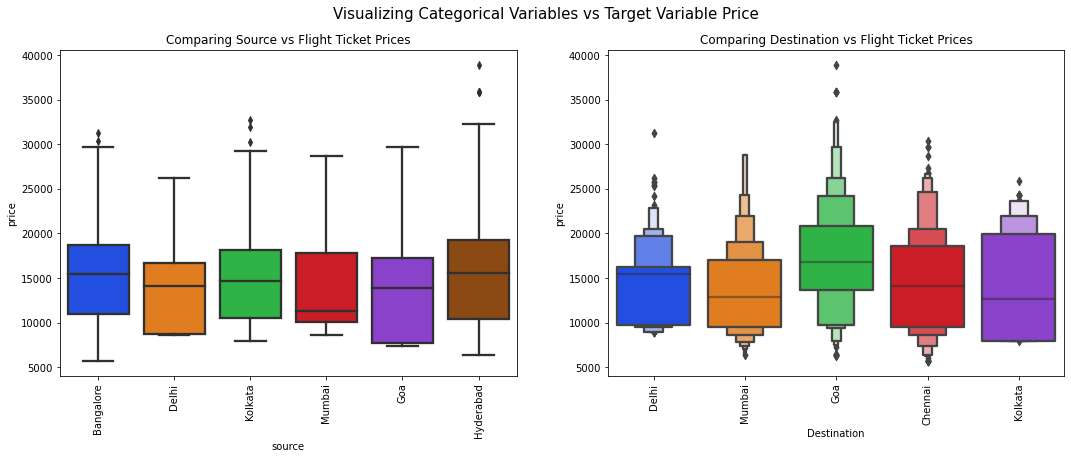
1. Mostly the price ranges from 5000 to 20000 and it extends upto 40000

**Bivarient Analysis:**

****

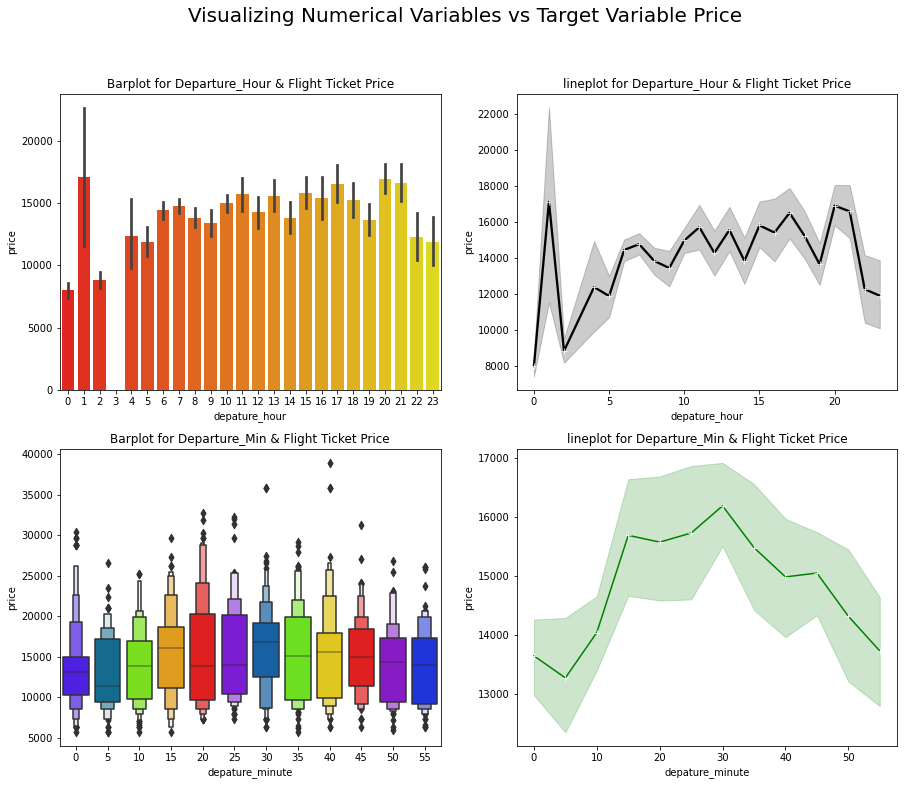
**Key Observations:**

1. Airline vs Price: From the bar plot we can notice "Vistara" and "Air India" airlines have highest ticket prices compared to other airlines.
2. Number\_of\_stops vs Price: From the strip plot we can notice the flights which have 1 stop between source and destination have highest ticket prices compared to others.

****

**Key Observations:**

1. Source vs Price : From the box plot we can observe the flights from Hyderabad are having somewhat higher prices compared to other sources.
2. Destination vs Price: From the boxen plot we can notice that the flights travelling to Goa have higher flight ticket prices.



**Key Observations:**

1. Departure\_Hour vs Price: From the bar plot and line plot we can see that there are some flights departing in the early morning 1 AM having most expensive ticket prices compared to late morning flights. We can also observe the flight ticket prices are higher during afternoon (may fluctuate) and it decreases in the evening.
2. Departure\_Min vs Price: The boxen plot and line plot gives there is no significant difference between price and departure min.

**Tools Used:**

1. Python 3.8
2. Numpy
3. Pandas
4. Matplotlib
5. Seaborn
6. Data Science
7. Machine Learning

**Model/s Development and Evaluation:**

**Identification of possible problem solving approach:**

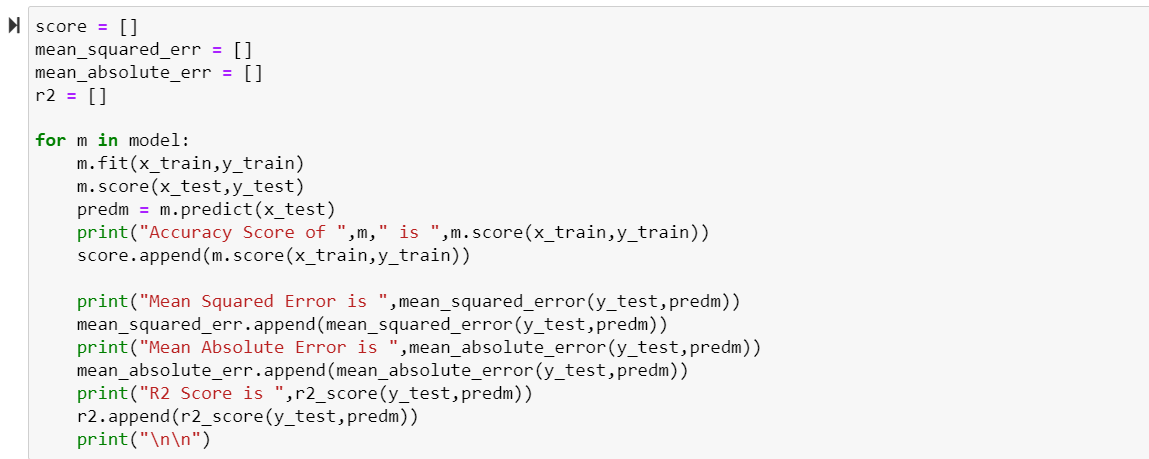
In this project, both Statistical and Analytical methods are used, in which Data Pre-processing, Exploratory Data Analysis is used after ensuring that data is cleaned. Here Target column is price, as it is continuous data, Regression algorithm is used. This project consists of 4 regressors – Decision Tree Regressor, Support Vector Regressor, KNeighbors Regressor, Linear Regression, 3 ensemble -Random Forest Regressor, AdaBoost Regressor, Gradient Boosting Regressor, XGB Regressor, 3 metrics – r2\_score, mean\_squared\_error, mean\_absolute\_error and 3 regularization – Lasso, Ridge, ElasticNet techniques. Out of which Random Forest Regressor gave high accuracy which is also chosen as best model in which there is least difference between r2 score and cv. Then with this Hypertuning is done in which accuracy is increased to 95%. Once all this is done, best model is saved using joblib. Then by using this saved model the output is determined which predicted the flight ticket price.

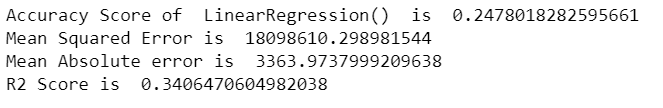
**Test of Identified Approaches:**

The approaches followed in this project are

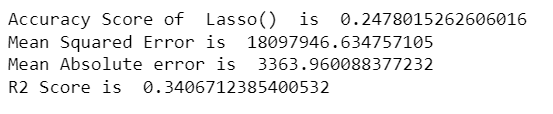
1. Find the best random state of a model.
2. Use all the other models to find accuracy score, mean squared error, mean absolute error and r2 score.
3. Then find Cross Validation of all models to find the best accuracy, which is the least difference between r2 score and cv score.
4. With the best model accuracy, we need to do hyper tuning using GridSearchCV.
5. Then finally saving the best model and by keeping this we need to predict the price of flight.

**Run and evaluate of selected model:**

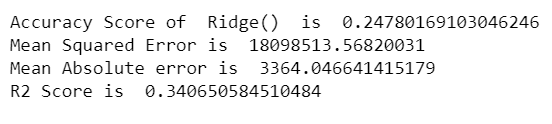
****

****

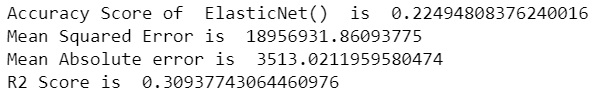
Linear Regression gives 2.4% accuracy and r2\_score 3.40

****

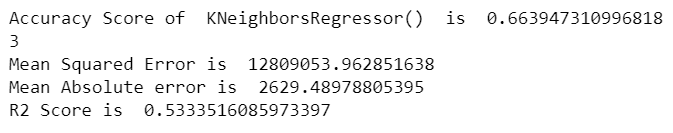
Lasso gives 2.4% accuracy and r2\_score 3.40

****

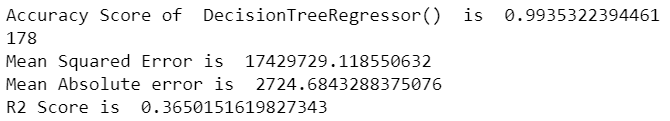
Ridge give 2.4% accuracy score and r2\_score 3.40

****

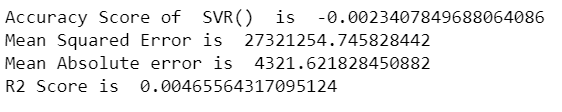
ElasticNet gives 2.2% accuracy score and r2\_score 3.09

****

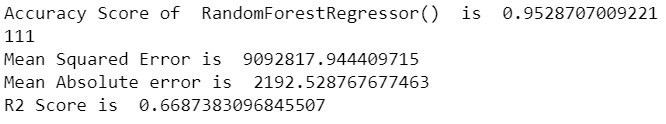
KNeighborsRegressor gives 6.63% accuracy and r2 score 5.33

****

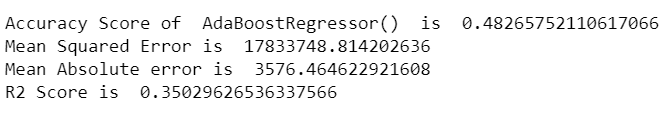
DecisionTreeRegressor gives 99.3% accuracy and r2\_score 3.65

****

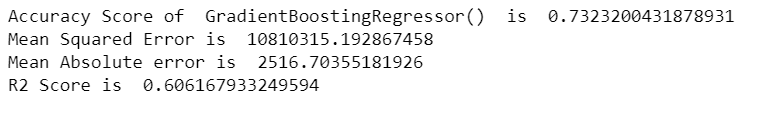
SVR gives 2.03% accuracy and 0.046 r2\_score

****

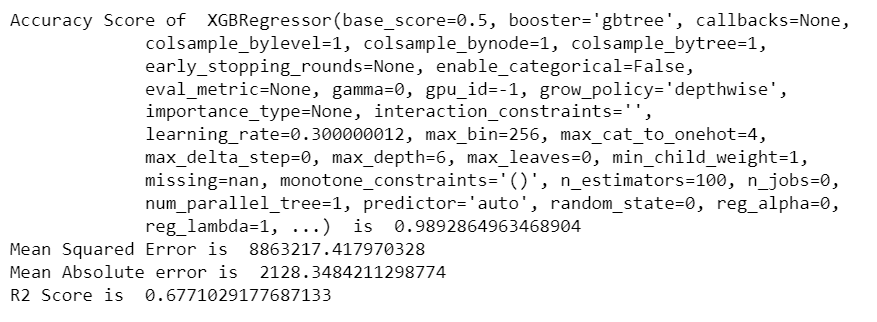
RandomForestRegressor gives 95% accuracy and 66.8 r2\_score

****

AdaBoostRegressor gives 48% accuracy score and 35.02 r2\_score

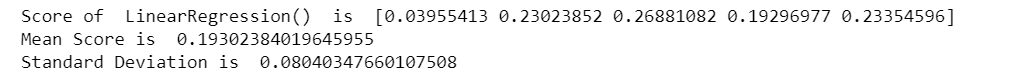
****

Gradient Boosting Regressor gives 73% accuracy and 60.61 r2\_score

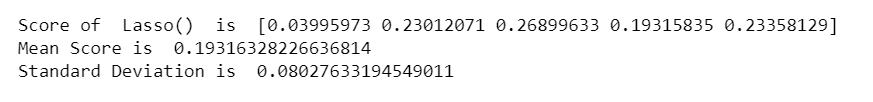
****

XGB Regressor gives 98.9% accuracy score and 67.7 r2\_score

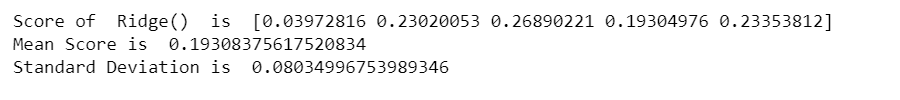
**Cross Validation:**

****

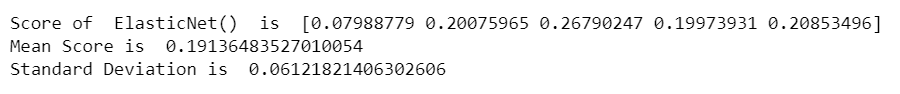
CV Score of LinearRegressor is 0.08040

****

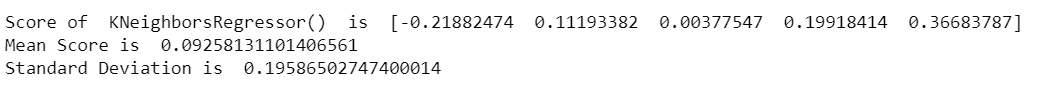
CV Score of Lasso is 0.0802

****

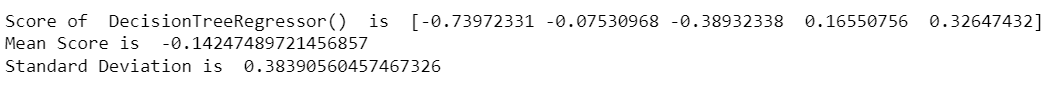
CV Score of Ridge is 0.08034

****

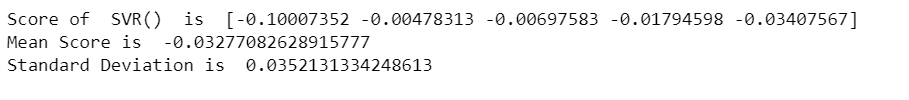
CV Score of ElasticNet is 0.06121

****

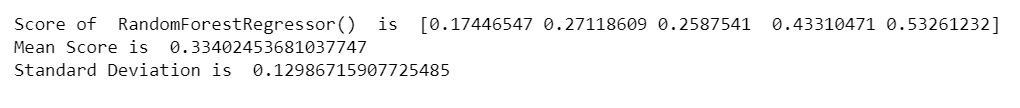
CV Score of KNeighborsRegressor is 0.09258

****

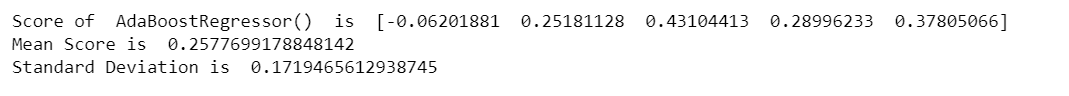
CV Score of DecisionTreeRegressor is 0.1424

****

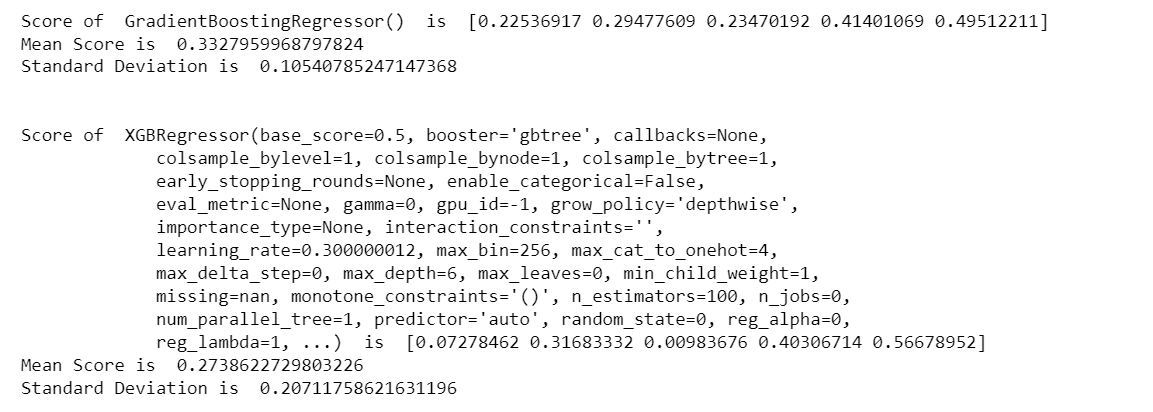
CV Score of SVR is 0.03277

****

CV Score of RandomForestRegressor is 0.334

****

CV Score of AdaBoostRegressor is 0.25770

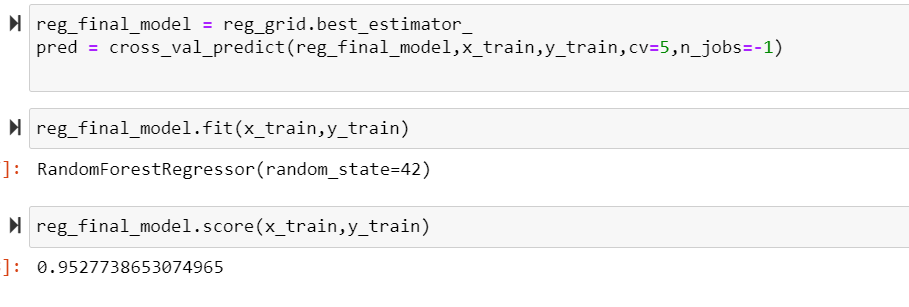
****

CV Score of GradientBoostingRegressor is 0.332

CV Score of XGBRegressor is 0.273

**HyperTuning:**

****

****

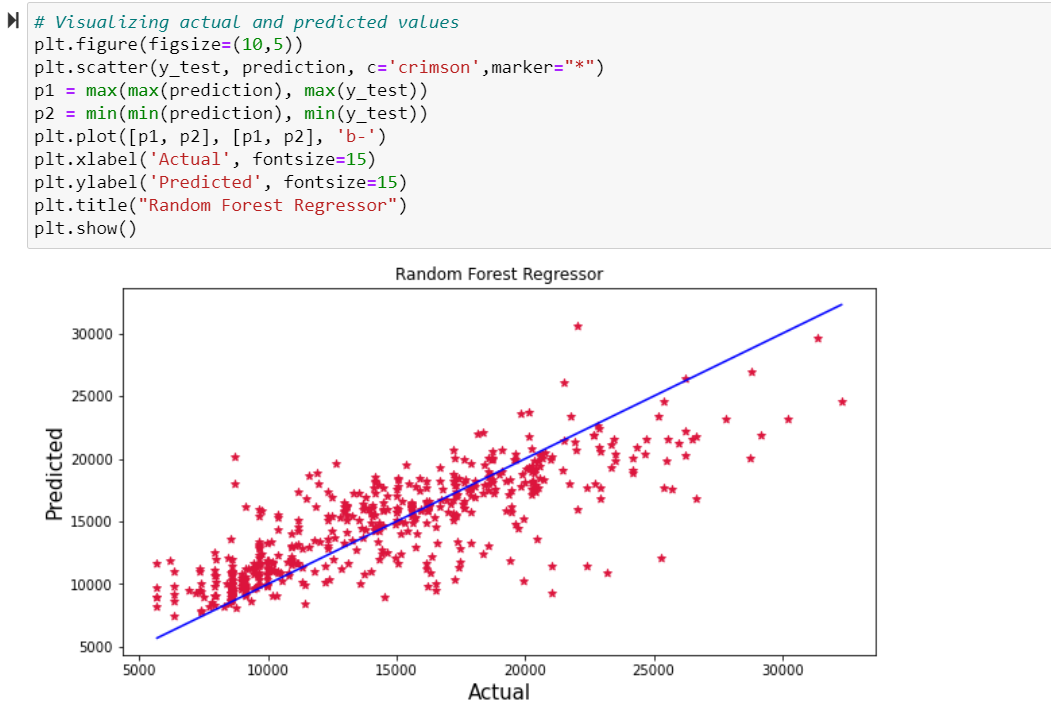
After Hypertuning we have 95% accuracy score which is improved.

**Saving the best model:**

The best model is saved using pickle.

****

Graphical representation of actual and predicted values

****

The graph shows how our final model is mapping. The plot gives the linear relation between predicted and actual price of the flight tickets. The blue line is the best fitting line which gives the actual values/data and red dots gives the predicted values/data.

**Key Metrics for success in solving problem under consideration:**

Some of key metrics for the evaluation of this project are

1. R2\_score
2. Mean\_squared\_error
3. Mean\_absolute\_error
4. Cross\_validation
5. Hypertuning

**R2 Score:**

R-squared is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. Closer the r2\_square value to 1 better the model fits.

**Mean Squared Error:**

The average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated is measured by the MSE of an estimator (of a process for estimating an unobserved variable). MSE is a risk function that represents the squared error loss's anticipated value. The Root Mean Squared Error is abbreviated as RMSE.

**Mean Absolute Error:**

MAE is a statistic that assesses the average magnitude of mistakes in a set of forecasts without taking into account their direction. It's the average of the absolute differences between forecast and actual observation over the test sample, where all individual deviations are given equal weight.

**Cross Validation:**

Cross-validation aids in determining the model's overfitting and underfitting. The model is constructed to run on several subsets of the dataset in cross validation, resulting in numerous measurements of the model. If we fold the data five times, it will be separated into five parts, each representing 20% of the whole dataset. During the Cross-validation, the first part (20%) of the 5 parts will be left out as a holdout set for validation, while the rest of the data will be utilised for training. We'll acquire the initial estimate of the dataset's model quality this way.

Further rounds are produced in the same way for the second 20% of the dataset, which is kept as a holdout set while the remaining four portions are utilised for training data during the process. We'll acquire the second estimate of the dataset's model quality this way. During the cross-validation procedure, these stages are repeated to obtain the remaining estimate of model quality.

**Hypertuning:**

Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process. These hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.

**Interpretation of the Results:**

In univariate analysis count plots is used to visualize the counts in categorical variables and distribution plot to visualize the numerical variables. In bivariate analysis bar plots, strip plots, line plots, box plots, and boxen plots are used to check the relation between label and the features, pair plot is used to check the pairwise relation between the features. The heat map and bar plot helped to understand the correlation between dependent and independent features. Detected outliers and skewness with the help of box plots and distribution plots respectively.

After all pre-processing and visualization, it is found that flight ticket price is mainly based on number of stops and duration of flight.

**Conclusion:**

**Key Findings:**

The Flight ticket price goes up and down and there is no fixed time. Some flights have early morning and day flights with expensive flight ticket price and mid night ticket price with cheap price. The last-minute flights are expensive. Indigo and SpiceJet have almost same cost of ticket price.

**Learning outcomes of study in Data Science:**

While working on this project many things like the features of flights, the flight ticket selling web platforms are learnt and got idea about how the machine learning models helps to predict the price of flight tickets.

The challenges faced are while scrapping the real time data from easemytrip.com website, it took so much time to gather data. Finally, our aim was achieved by predicting the flight ticket price and built flight price evaluation model that could help the buyers to understand the future flight ticket prices.

**Limitations of this work and scope of Future work:**

The main drawback of this study is the low number of records that have been used. In the dataset, data is not properly distributed in some of the columns many of the values in the columns are having string values. Due to some reasons the models may not make the right patterns and the performance of the model also reduces. So that issues need to be taken care.

The greatest shortcoming of this work is the shortage of data. Anyone wishing to expand upon it should seek alternative sources of historical data manually over a period of time. Additionally, more varied set of flights should be explored, since it is entirely possible that airlines vary their pricing strategy according to the characteristics of the flight. Finally, it would be interesting to compare our system's accuracy against that of the commercial systems available today (preferably over a period of time).

**References:**

A few external references have been used to complete this project successfully. Below are the sources

1. <https://scikit-learn.org/>
2. <https://github.com/>
3. <https://analyticsvidhya.com/>