

NAME OF THE PROJECT

FLIGHT PRICE PREDICTION

Submitted by:

V TARAK RAM SAI

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

* Business Problem Framing

The focus of this project is to develop a machine learning model that predicts the fares of flight tickets. Data used in this article is scraped from yatra.com. This will be a regression problem since the target or dependent variable is the price (continuous numeric value).

* Conceptual Background of the Domain Problem

To solve this problem, we have been provided with prices of flight tickets for various airlines on the days of March 15 and March 22 of 2022 and between New Delhi and various cities, using which we aim to build a model which predicts the prices of the flights using various input features.

* Review of Literature

There is a need to develop a model for the consumers from which they can predict these prices and moreover analyse which feature is the most influential in determining these prices. In this project I have applied Machine learning algorithms on a dataset consisting of various factors which can influence the flight fares to prepare such a model.

* Motivation for the Problem Undertaken

Nowadays, the number of people using flights has increased significantly. It is difficult for airlines to maintain prices since prices change dynamically due to different conditions. That’s why we will try to use machine learning to solve this problem. This can help airlines by predicting what prices they can maintain. It can also help customers to predict future flight prices and plan their journey accordingly.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Any machine learning model should follow the below steps while dealing a business problem. They are:

**i.)** **Business Understanding:** The first step is to comprehend the research’s background, the problem description, and how the proposed project will achieve the goals.

**ii.**) **Data Understanding:** The second stage requires collection of data listed in the project resources. This involves in determining the data requirements and exploring key data attributes.

**iii.**) **Data Preparation:** The third stage involves data cleaning and should the handle the missing values in the data.

**iv.**) **Modelling:** This involves determining the modelling technique and testing the design.

**v.**) **Evaluation:** Here, we should evaluate the achieved results and should determine the performance of the model with best accuracy.

**vi.**) **Deployment:** The last stage is implementation of the model.

* Data Sources

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

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

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

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

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

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

**Date**: The date of the journey

**No\_of\_Stops**: Total stops between the source and destination.

**Price**: The price of the ticket

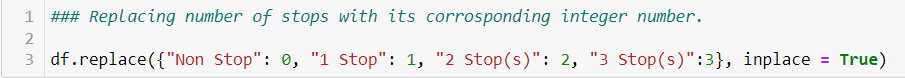
* Data Preprocessing

It entails converting raw data into comprehensible format that a machine learning model can understand. The data pre-processing involves data cleaning which involves handling missing values, transformation of data i.e. normalizing the data and data reduction which involves only required features.

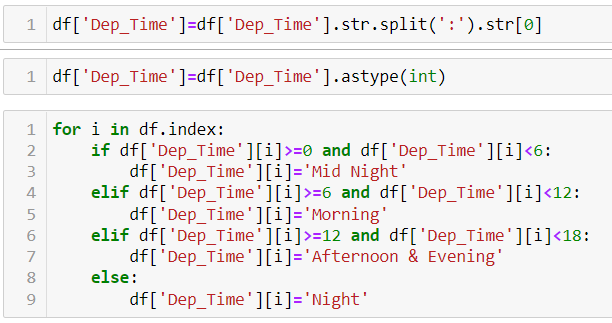


**Data Cleaning**

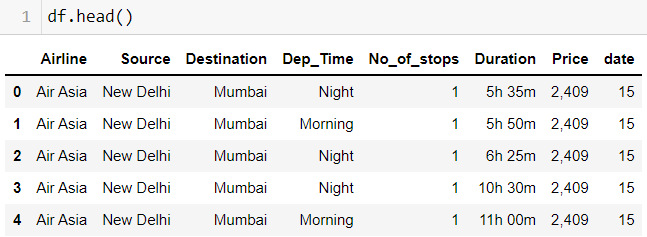
Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. Data cleaning is the process of preparing data for analysis by weeding out information that is irrelevant or incorrect. This is generally data that can have a negative impact on the model or algorithm it is fed into by reinforcing a wrong notion. Data cleaning not only refers to removing chunks of unnecessary data, but it’s also often associated with fixing incorrect information within the dataset and reducing duplicates.



We can divide the Departure timings into 4 quarters like Morning, Afternoon-Evening, Night and Mid-Night.



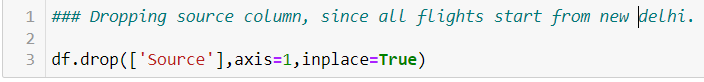
Then the dataframe will be as follows-



**Data Reduction**

The next step of data processing is data reduction. This is used to remove duplicate features present in the data i.e. unwanted features for prediction.

Dropping "source" column because all the rows has source as New Delhi and dropping "Data" because I created date column for better analysis. We can also delete Arrival Column, because Departure time and Duration gives the same information.







**Handling Outliers**

### Instead of eliminating the outliers, we have replaced the outliers higher than the upper whisker by the value of upper whisker and the outliers lower than the lower whisker by the value of lower whisker.

### 

**Encoding**

Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the different models.

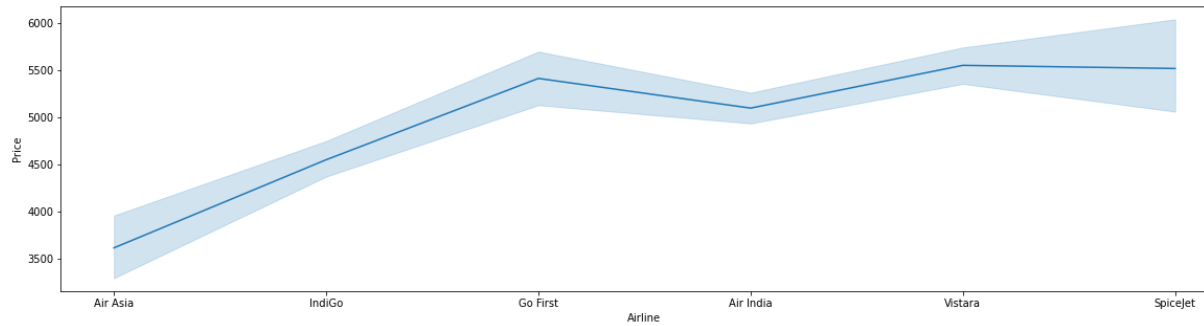
### 

**Scaling**

Feature scaling is **a method used to normalize the range of independent variables or features of data**. To convert data into a distribution with a mean of 0 and standard deviation of 1, we will use a standard scalar.

### 

* Data Inputs- Logic- Output Relationships
* The Relation between all the features in the dataset are determined in the below graphs as follows



### Go First and Vistara are the costliest airlines compared to others.

### 

### Number of Air India and IndiGo flights are more in the dataset.

### 

### The flights on March 22nd are cheaper than rates on March 15th.

### 

### Most of the Ticket prices are in the range of 4000 - 6000.

### 

### There are more flights to Bangalore and Chennai from Delhi.

### 

### There are more flights in Day Time.

### 

### The flight ticket prices at night time are cheaper.

* Hardware and Software Requirements and Tools Used

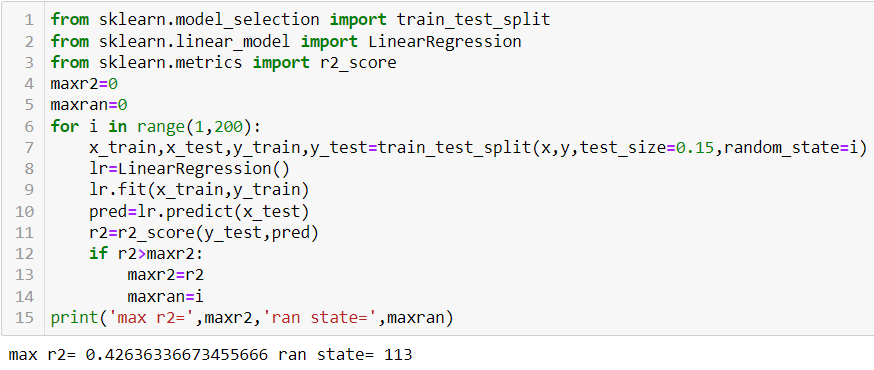
The hardware requirements for the project includes a laptop with at least 4GB RAM. This project uses a Jupyter Notebook as a code editor. The Machine Learning models are implemented using python version 3.7 with libraries like numpy, pandas, matplotlib, seaborn and sklearn.

**Model/s Development and Evaluation**

The independent variables are declared in x and the dependent variable i.e. ‘Price’ is declared in y as follows-



The below code is done for choosing the Random state variable. We should do testing by using any of the three regression algorithms.



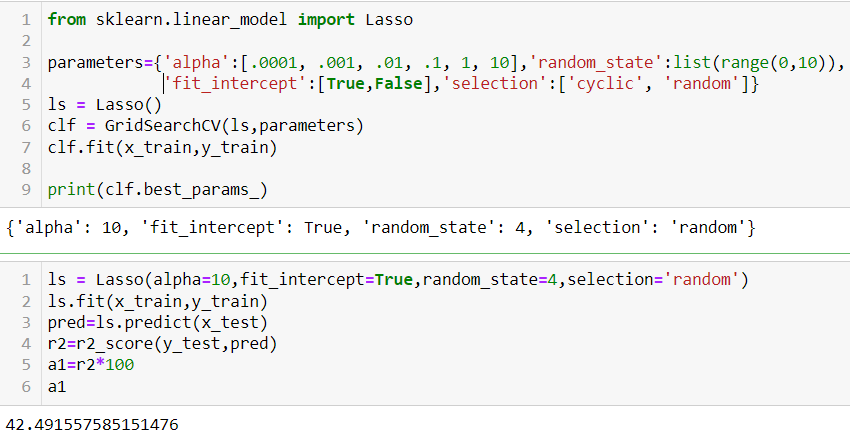


**Hyper Parameter Tuning**

In machine learning, hyper parameter optimization or tuning is **the problem of choosing a set of optimal hyper parameters for a learning algorithm**. A hyper parameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

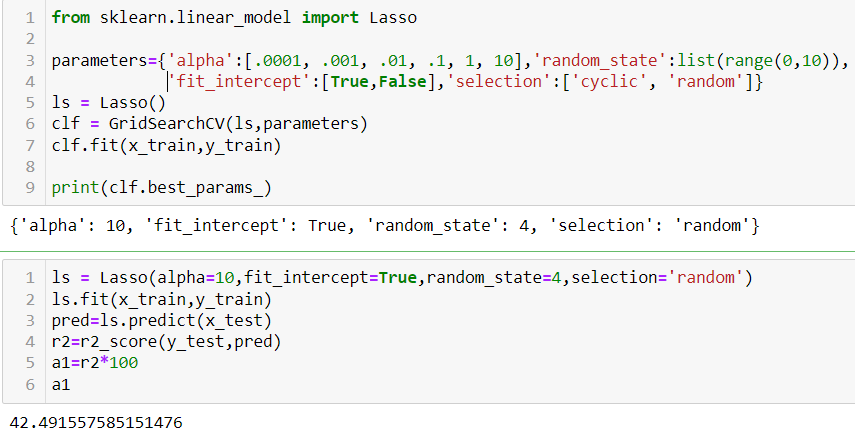
**Linear Regression**

Linear Regression attempt to model the relationship between two variables by fitting a linear equation to observed data. The other is considered to be dependent variable.



**Lasso Regression**

The “LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multi-collinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.



**Ridge Regression**

A Ridge regressor is basically a regularized version of Linear Regressor. The regularized term has the parameter ‘alpha’ which controls the regularization of the model i.e. helps in reducing the variance of the estimates.



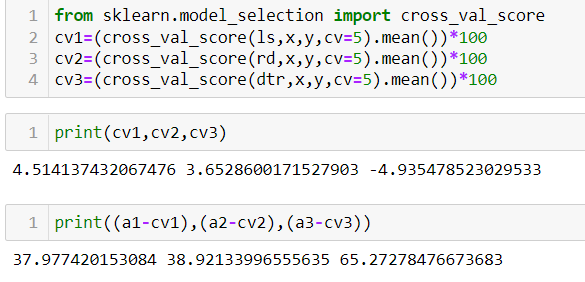
**Decision Tree Regression**

Decision tree builds regression or classification models in the form of a tree structure. It **breaks down a dataset into smaller and smaller subsets** while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.



**Cross Validation**

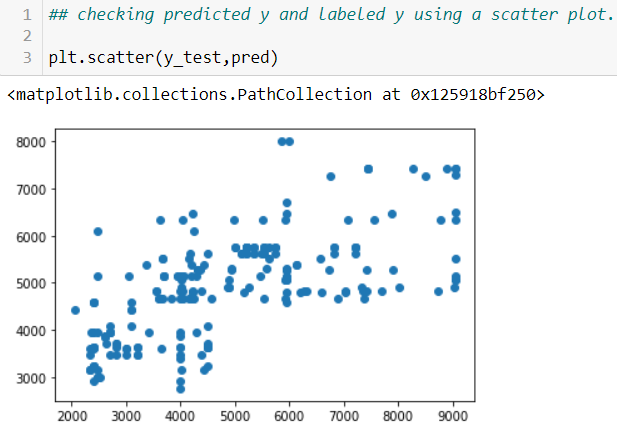
Cross Validation is a **very useful technique for assessing the effectiveness of your model**, particularly in cases where you need to mitigate overfitting.



### We can choose Lasso Regression as our model since its cross validation score and accuracy score are almost similar.

### 

**Relation b/w Predicted output and Actual Output**



**CONCLUSION**

* Flight price prediction can be a challenging task due to the various types of attributes that should be considered for the accurate prediction. The major step in the prediction process is collection of data and pre-processing of the data. In this research, the model is built to normalize, standardize and clean data to avoid unnecessary noise for machine learning algorithms.
* Data cleaning is one of the processes that increases prediction performance. The main attributes that contribute to the price of flight ticket are the type of airlines, no of stops between source and destination, duration of journey and date of journey. The above mentioned are the main attributes that influence the ticket price.
* Limitations of this work and Scope for Future Work

For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. To correct for overfitting in Random Forest, different selections of features and number of trees will be tested to check for change in performance.

However, the drawback of the proposed system is that it consumes much more computational resources than single machine learning algorithm.