**Maniya Yash Rajeshbhai**
**B20CS033**

# Flight Ticket Price Prediction

## Introduction :

We will be analysing the flight fare prediction using Machine Learning dataset using essential exploratory data analysis techniques then will draw some predictions about the price of the flight based on some features such as what type of airline it is, what is the arrival time, what is the departure time, what is the duration of the flight, source, destination and more.

## **Dataset** : flight_fare_dataset

**Airline**: This column has all the types of airlines like Indigo, Jet Airways, Air India, and many more.
**Date_of_Journey**: This column will let us know about the date on which the passenger's journey will start.
**Source**: This column holds the name of the place from where the passenger's journey will start.
**Destination**: This column holds the name of the place to where passengers wanted to travel.
**Route**: Here we can know about what is the route through which passengers have opted to travel from his/her source to their destination.
**Arrival_Time**: Arrival time is when the passenger will reach his/her destination.
**Duration**: Duration is the whole period that a flight will take to complete its journey from source to destination.
**Total_Stops**: This will let us know in how many places flights will stop there for the flight in the whole journey.
**Additional_Info**: In this column, we will get information about food, kind of food, and other amenities.
**Price**: Price of the flight for a complete journey including all the expenses before onboarding. This is the target/label column.

## Methodology :

Since the data has a lot of non-numerical data, we tried to convert the meaningless bunch of strings to meaningful numbers in preprocessing.

- Using the Routes, we calculated the the number of stops and accordingly updated the strings in the Total_Stops column by numbers.
- Arrival_Time had dates in some rows, so trimmed some of the parts to erase those dates.
- Converted Duration of flights from string to number of minutes.
- Then Label Encoded all the remaining features except the target variable.
- Divided the dataset into *training : validation : testing = 0.525 : 0.175 : 0.3* datasets.

Then we train 6 different models on the training dataset and test them on validation sets.

1. **Decision Tree Regressor** : It is a model that 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.

2. **Random Forest Regressor** : Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

3. **Gradient Boosting Regressor :** Gradient boosting Regression calculates the difference between the current prediction and the known correct target value. This difference is called residual. After that Gradient boosting Regression trains a weak model that maps features to that residual.

4. **Linear Regressor :** Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

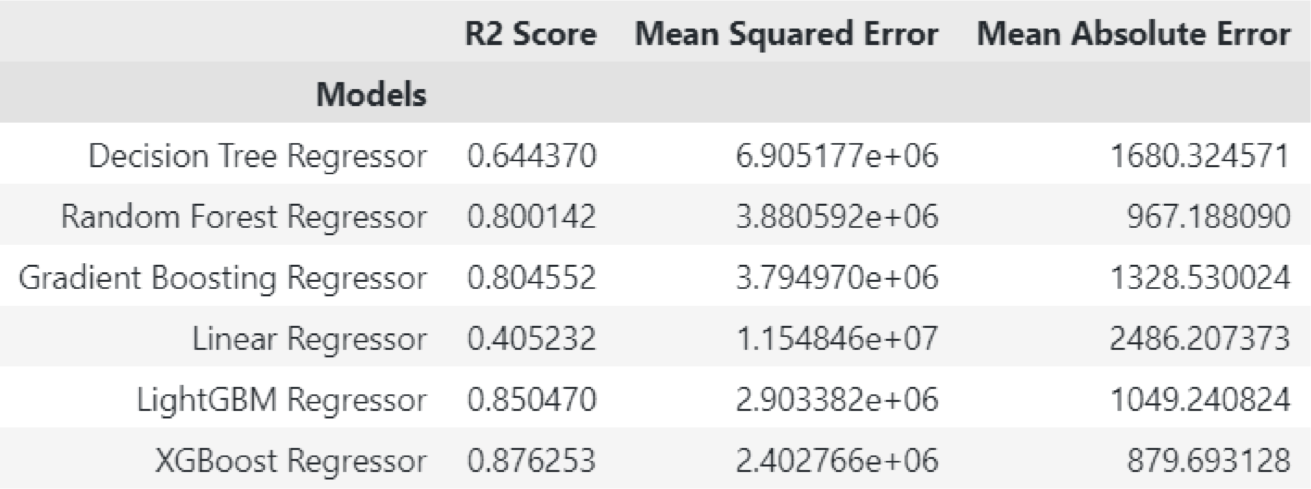
5. **LightGBM Regressor :** LightGBM is a gradient boosting framework based on decision trees to increase the efficiency of the model and reduce memory usage. It falls under the ensemble methods.

6. **XGBoost Regressor :** XGBoost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energising machine learning model performance and computational speed.

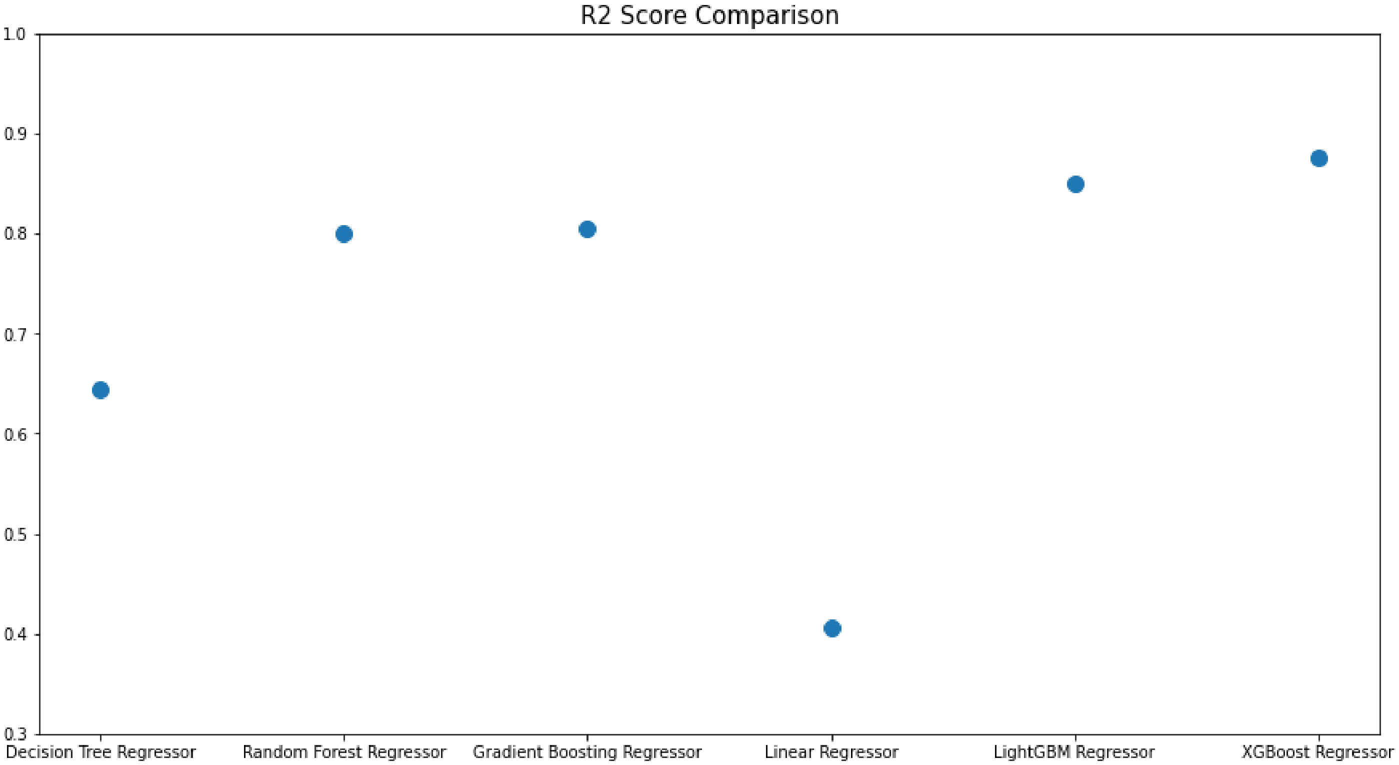
## Comparing the validation Results :

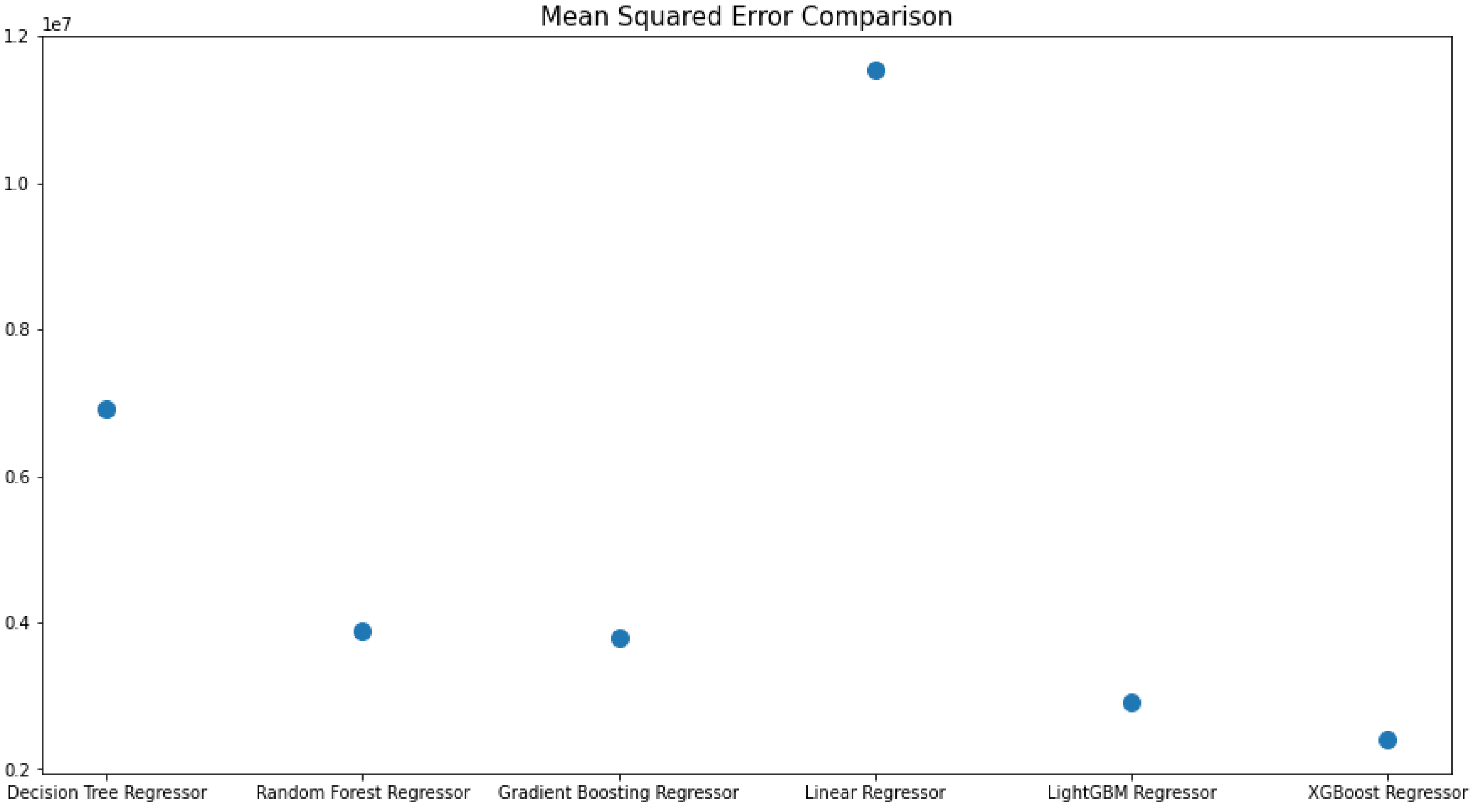
We see that XGBoost and LightGBM Regressors perform best on the validation sets. This was quite obvious as they use boosting techniques on a large number of Decision Trees [ we chose 450 n_estimators ].

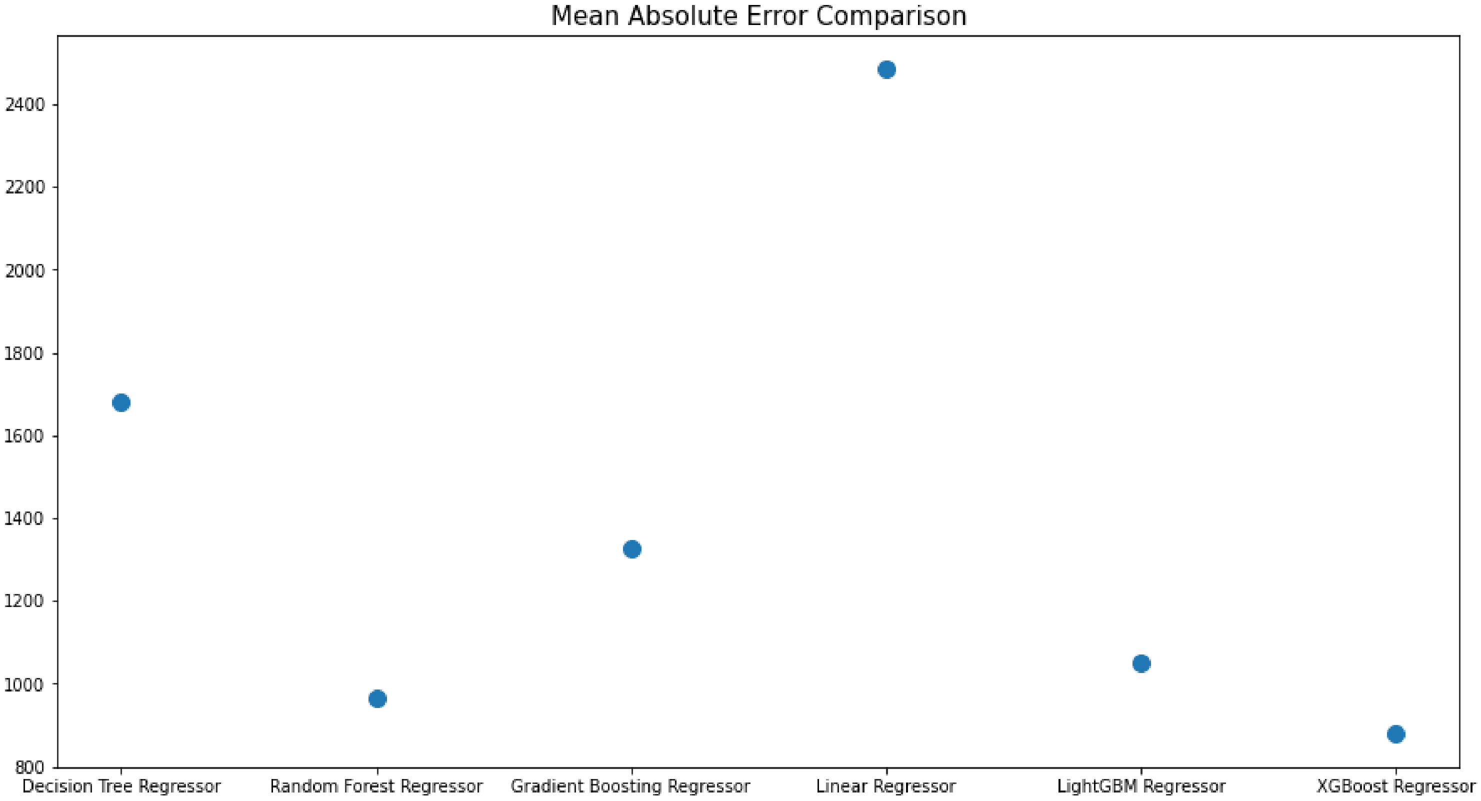
Linear Regression performs worst and the reason might be because the dataset is not linearly separable. The Decision Tree too performs poorly as a single tree usually cannot train all the parameters up to a sufficient level. If we increase the complexity of tree, we overfit the data. But if we keep a simpler tree, we might underfit. So we need a large number of trees and bagging applied collectively to increase R-squared score and decrease the MSE.



| | R2 Score | Mean Squared Error | Mean Absolute Error |
|---|---|---|---|
| **Models** | | | |
| Decision Tree Regressor | 0.644370 | 6.905177e+06 | 1680.324571 |
| Random Forest Regressor | 0.800142 | 3.880592e+06 | 967.188090 |
| Gradient Boosting Regressor | 0.804552 | 3.794970e+06 | 1328.530024 |
| Linear Regressor | 0.405232 | 1.154846e+07 | 2486.207373 |
| LightGBM Regressor | 0.850470 | 2.903382e+06 | 1049.240824 |
| XGBoost Regressor | 0.876253 | 2.402766e+06 | 879.693128 |




Mean Squared Error Comparison
1e7
1.2
1.0
0.8
0.6
0.4
0.2
Decision Tree Regressor
Random Forest Regressor
Gradient Boosting Regressor
Linear Regressor
LightGBM Regressor
XGBoost Regressor

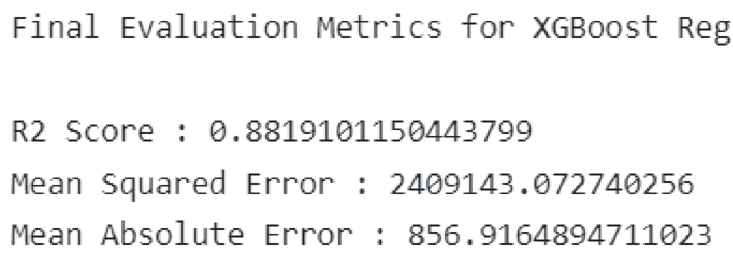

Mean Absolute Error Comparison
2400
2200
2000
1800
1600
1400
1200
1000
800
Decision Tree Regressor
Random Forest Regressor
Gradient Boosting Regressor
Linear Regressor
LightGBM Regressor
XGBoost Regressor

## Final Model Evaluation :

We use XGBoost Regressor as our final model as it performs best on the validation data with a R2-Score of 0.8763.

Parameters of the model :

n_estimators = 450
max_depth = 5
learning_rate = 0.1
min_child_weight = 1

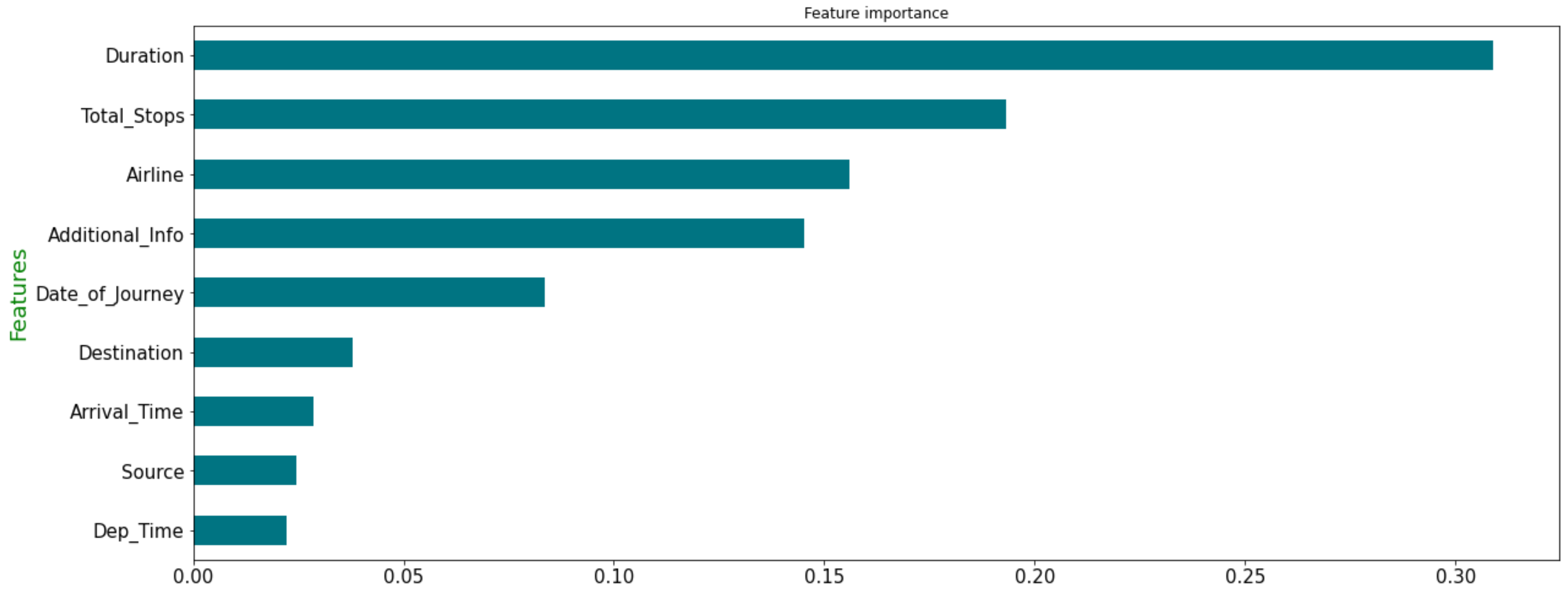


```
Final Evaluation Metrics for XGBoost Regressor on Testing Dataset:

R2 Score : 0.8819101150443799
Mean Squared Error : 2409143.072740256
Mean Absolute Error : 856.9164894711023
```

Finally we get an R2 Score of 0.8819 on the testing dataset with our final model.

On visualizing the Feature Importance of the Model, we see that Duration plays the major role. It is quite obvious that flights with larger duration will have higher cost prices.
Also the total number of stops in a journey and the Airline in which the person is travelling have a major impact on the Flight Ticket Price.


Feature importance
Features
Duration
Total_Stops
Airline
Additional_Info
Date_of_Journey
Destination
Arrival_Time
Source
Dep_Time