Flight Price Prediction
(Regression Analysis)

Submitted by:
LAKSHITA KAIN
DATA SCIENCE INTERN

ACKNOWLEDGMENT

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude I give Data Trained and Flip Robo, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my project. Furthermore, I would also like to acknowledge with much appreciation the crucial role of our SME of Mr Keshav Bansal.

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.

1.1 PROBLEM DEFINITION:

Optimal timing for airline ticket purchasing from the consumer's perspective is challenging principally because buyers have insufficient information for reasoning about future price movements. In this project we simulate various models for computing expected future prices and classifying whether this is the best time to buy the ticket.

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)

1.2 OBJECTIVE:

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.

2. Literature Review

2.1 Machine Learning Algorithms

Machine learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from data. Machine learning methods are appropriate in big data since attempting to manually process vast volumes of data would be impossible without the support of machines. Machine learning in computer science attempts to solve problems algorithmically rather than purely mathematically. Therefore, it is based on creating algorithms that permit the machine to learn. However, there are two general groups in machine learning which are supervised and unsupervised. Supervised is where the program gets trained on pre-determined set to be able to predict when a new data is given. Unsupervised is where the program tries to find the relationship and the hidden pattern between the data.
Machine learning-based systems are growing in popularity in research applications in most disciplines.

However, some of them give better performance in certain circumstances. Thus, this thesis attempts to use regression algorithms to compare their performance when it comes to predicting values of a given dataset. The performance will be measured upon predicting house prices since the prediction in many regression algorithms relies not only on a specific feature but on an unknown number of attributes that result in the value to be predicted. House prices depend on an individual house specification. Houses have a variant number of features that may not have the same cost due to its location. For instance, a big house may have a higher price if it is located in desirable rich area than being placed in a poor neighbourhood.

2.1.1 Multiple Linear Regression

Multiple Linear Regression (MLR) is a supervised technique used to estimate the relationship between one dependent variable and more than one independent variables. Identifying the correlation and its cause-effect helps to make predictions by using these relations. To estimate these relationships, the prediction accuracy of the model is essential; the complexity of the model is of more interest. However, Multiple Linear Regression is prone to many problems such as multicollinearity, noises, and overfitting, which effect on the prediction accuracy.
Regularised regression plays a significant part in Multiple Linear Regression because it helps to reduce variance at the cost of introducing some bias, avoid the overfitting problem and solve ordinary least squares (OLS) problems. There are two types of regularisation techniques L1 norm (least absolute deviations) and L2 norm (least squares). L1 and L2 have different cost functions regarding model complexity.

2.1.2 Lasso Regression

Least Absolute Shrinkage and Selection Operator (Lasso) is an L1-norm regularised regression technique that was formulated by Robert Tibshirani in 1996 [6]. Lasso is a powerful technique that performs regularisation and feature selection. Lasso introduces a biasterm,butinsteadofsquaringtheslopelikeRidgeregression,theabsolute valueofthe slope is added as a penalty term. Lasso is defined as

$$L=Min(sum\ of\ squared\ residuals + \alpha * |slope|)$$

Where $Mi(sum\ of\ squared\ residuals)$ is the Least Squared Error, and $\alpha * |s\ ope|$ is the penalty term. However, alpha α is the tuning parameter which controls the strength of the penalty term. In other words, the tuning parameter is the value of shrinkage. $|slope|$ is the sum of the absolute value of the coefficients .
Cross-validation is a technique that is used to compare different machine learning algorithms in order to observe how these methods will perform in practice. Cross-validation method divides the data into blocks. Each block at a time will be used for testing by the algorithm, and the other blocks will be used for training the model. In the end, the results will be summarised, and the block that performs best will be chosen as a testing block . However, α is determined by using cross-validation. When α =0, Lasso becomes Least Squared Error, and when $\alpha \neq 0$, the magnitudes are considered, and that leads to zero coefficients. However, there is a reverse relationship between alpha α and the upper bound of the sum of the coefficients t. When $t \to \infty$, the tuning parameter α =0. Vice versa when α =0 the coefficients shrink to zero and $t \to \infty$. Therefore, Lasso helps to assign zero weights to most redundant or irrelevant features in order to enhance the prediction accuracy and interpretability of the regression model.

Throughout the process of features selection, the variables that still have non-zero coefficients after the shrinking process are selected to be part of the regression model. Therefore, Lasso is powerful when it comes to feature selection and reducing the overfitting.

2.1.3 Ridge Regression

The Ridge Regression is an L2-norm regularised regression technique that was introduced by Hoerl in 1962 [9]. It is an estimation procedure to manage collinearity without removing variables from the regression model. In multiple linear regression, the multicollinearity is a common problem that leads least square estimation to be unbiased, and its variances are far from the correct value. Therefore, by adding a degree of bias to the regression model, Ridge Regression reduces the standard errors, and it shrinks the least square coefficients towards the origin of the parameter space [10]. Ridge formula is:

$$R=Mi(sum\ of\ squared residuals + \alpha * slope^2)$$

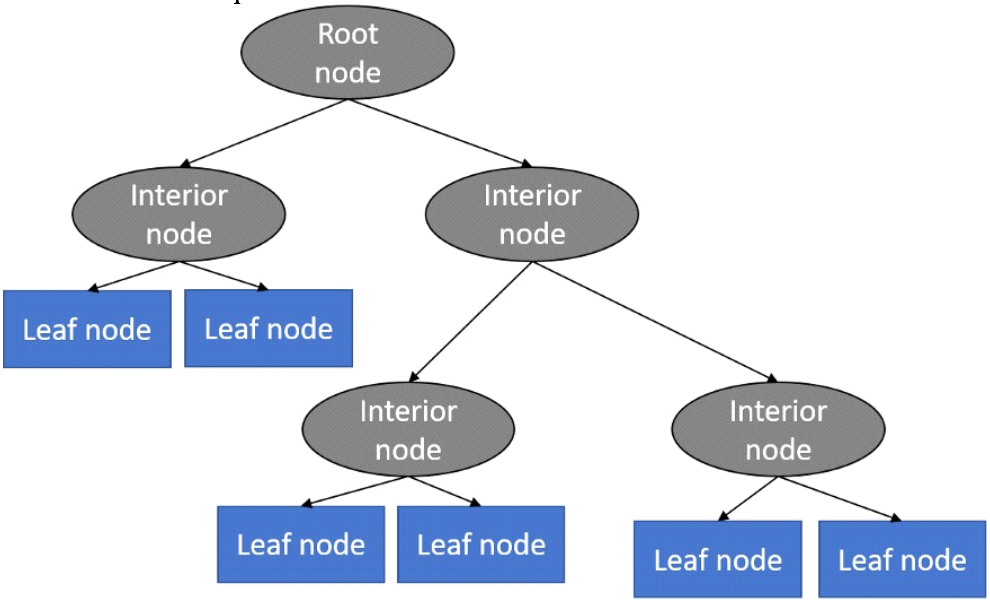
Where $Mi(sum\ of\ squared\ residuals)$ is the Least Squared Error, and $\alpha * slope^2$ is the penalty term that Ridge adds to the Least Squared Error.

When Least Squared Error determines the values of parameters, it minimises the sum of squared residuals. However, when Ridge determines the values of parameters, it reduces the sum of squared residuals. It adds a penalty term, where α determines the severity of the penalty and the length of the slope. In addition, increasing the α makes the slope asymptotically close to zero. Like Lasso, α is determined by applying the Cross-validation method. Therefore, Ridge helps to reduce variance by shrinking parameters and make the prediction less sensitive.

2.1.4 Decision Tree Regressor

Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application.

It is a tree-structured classifier with three types of nodes. The Root Node is the initial node which represents the entire sample and may get split further into further nodes. The Interior Nodes represent the features of a data set and the branches represent the decision rules. Finally, the Leaf Nodes represent the outcome. This algorithm is very useful for solving decision-related problems.



A Decision Tree can be defined as a model : φ=X ⟼Y

Where any node t represents a subspace $X_t \subseteq X$ of the input space and internal nodes t are labelled with a split s_t taken from a set of questions Q. However, to determine the best separation in Decision Trees, the Impurity equation of dividing the nodes should be taken into consideration, which is defined as:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

2.1.5 Random Forest Regressor

Random Forest is a model that constructs an ensemble predictor by averaging over a collection of decision trees. Therefore, it is called a forest, and there are two reasons for calling it random. The first reason is growing trees with a random independent bootstrap sample of the data.The second reason is splitting the nodes with arbitrary subsets of features . However, using the bootstrapped sample and considering only a subset of the variables at each step results in a widevarietyoftrees.ThevarietyiswhatmakesRandomForestmoreeffectivethan individual Decision Tree.

To improve prediction performance, Random Forest acquires out-of-bag (OOB) estimates, whichisbasedonthefactthat,foreverytree,approximately$1/e \approx 0.367 or 36\%$ofcasesarenot in the bootstrap sample [15]. There are several advantages to using OOB. One advantage is that the complete original example is used both for constructing the Random Forest classifier and for error estimation. Another advantage is its computational speed, especially when dealing with large data dimensions .

2.1.6 Adaptive Boosting Regressor

Adaptive boosting (AdaBoost) is an ensemble algorithm incorporated by Freund and Schapire (1997), which trains and deploys trees in time series. Since then, it evolved as a popular boosting technique introduced in various research disciplines. It merges a set of weak classifiers to build and boost a robust classifier that will improve the decision tree's performance and improve accuracy.

The mathematical presentation of the AdaBoost classifier at a high level is described of thefollowing:

Considerthattrainingdata(x1andy1),...,(xmandym)arethexi2X,yi2{1,+1}. Then theparametersofAdaBoostclassifierareinitialized:D1(i)=1/mfori=1,..., m. For t = 1,..., T:

- Train weak learner using distribution Dt.
- Aim: select ht with low weighted error:

$$\epsilon = \Pr_i \sim D_t[h_t(x_i) \neq y_i]$$

Select at = 1

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \cdot D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Hence, for i = 1, . . . ,m:

whereZtisanormalizationfactor(chosensothatDt+1willbeadistribution).Whichgeneratesan output of the final hypothesis as following:

H(x) = si.

3. Methodology

3.1 Data Collection

Data has been scrapped using Selenium webdriver with python. Selenium is a powerful tool for controlling web browsers through programs and performing browser automation.

Our data has been retrieved from https://www.yatra.com which consists of Airlines Flight information over the span of 16 days (from 01-12-2021 to 16-12-2021) such as Airlines Name, Date of Departure, Time of Departure, Time of Arrival, Duration of Flight, Number of Stops, Meals included and Price of the ticket.

3.1.2 About Dataset

□ Shape of our datset : records - 2255, features - 11

We have all object data type features, although some of our features have misinterpreted data types such as Duration, Price and Date should be of numeric data type not categorical. "Unnamed : 0" is an necessary column for our analysis therefore we drop it.

We don't have any null values but we have 119 duplicated fields in our obtained dataset.

3.2 Data Preprocessing

3.2.1 Feature Engineering

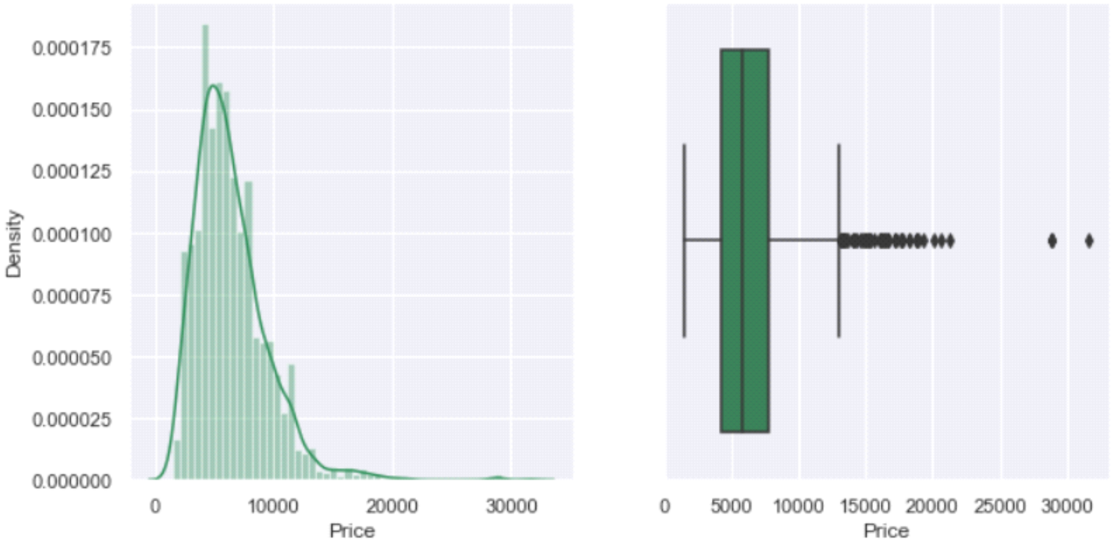
From our Date we've extracted seperate columns for day, month and year. Althought the data collected is from over 16 days of same month and year. Hence we drop, Month and Year.

From our Duration hour column. I've seperated duration hour and minutes into different columns

Similarly, Takeoff time and Arrival time was seperated with takeoff_hour, takeoff_minute, arrival_hour and arrival_minute respectively.

3.3 Outliers

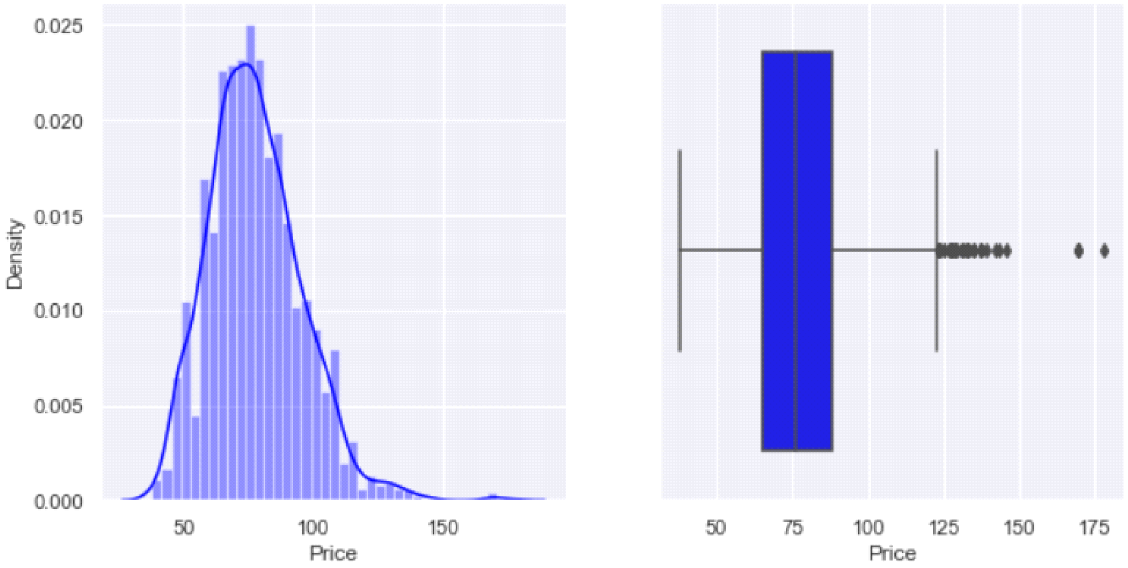
Outliers are noisy data that they do have abnormal behaviour comparing with the rest of the data in the same dataset. Outliers can influence the prediction model and performance due to its oddity. There are three types of outliers, which are point, contextual, and collective outliers. Our Target variable is right skewed : Price



Visualizing outliers using histogram and box - plot

3.3.1 Square - Root Transformation

The square root, x to x^(1/2) = sqrt(x), is a transformation with a moderate effect on distribution shape: it is weaker than the logarithm and the cube root. It is also used for reducing right skewness, and also has the advantage that it can be applied to zero values. Note that the square root of an area has the units of a length. It is commonly applied to counted data, especially if the values are mostly rather small.



After performing transformations to reduce skewness.

3.4 Encoding using Dummy Predictors:

A dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. pandas.get_dummies() is used for data manipulation. It converts categorical data into dummy or indicator variables. Our Dataset has 4 categorical features (Airlines,Source, Destination, Meal_Provided) which were all converted to give us dummy predictor for each category.

Our final dataset had 31 predictors.

3.5 Training–Test Set Split

For all the models fitted in this study, we split the balanced data into 80% or the training set and 20% for the testing set (validation).

4. Hardware and Software Requirements and Tools Used

4.1 Software Requirements:

- Python
- Anaconda (Jupyter Notebook)
- Python Libraries (Scikit - Learn, Imblearn, Pandas, Numpy , etc)

4.2 Hardware Requirements:

- Minimum 4 GB RAM
- Intel Core-i3 or above processor

5. Results

In this section, I present analytic results of the various machine learning models adopted in this research. All model diagnostic metrics in this paper are based on the validation/test set. In order to compare the machine learning models they have been tested on data where the price of the properties are known. Therefore, the data set is split into training data and testing data. Training data is, as the name suggests, used to train the machine learning model and constitutes the larger share of the split data set. When the model is fit with the training data the price of the flight are known to the model in order for it to learn from the data. The testing data is used to get a measurement of how well the machine learning model predicts flight prices. The machine learning model is given the test data but without the price of the flight in order to predict the price for them given the various features for the flight tickets. The predicted price is then compared to the actual price in the test data.

5.1 Error metrics

For measuring how good predictions the model makes, four error metrics have been used. Mean absolute error (MAE), Mean squared error (MSE), Median absolute error (MedAE) and Coefficient of determination (R2).

5.1.1 R-Squared Score

In statistics, the coefficient of determination, denoted R2 or r2 and pronounced "R squared", is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

$$\mathrm{R}^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

5.1.2 Mean absolute error (MAE)

Mean absolute error measures the prediction error by taking the mean of all

absolute values of all errors, that is:

$$MAE = \frac{\sum_{i=0}^{n} |y_i - \hat{y}_i|}{n}$$

Where n is the number of samples, y are the target values and ^y are the predicted values. A MAE closer to 0 means that the model predicts with lower error and that the prediction is better the closer the MAE is to 0.

5.1.3 Mean squared error (MSE)

Mean squared error is similar to MAE, but the impact of a termis quadratically proportional to its size. It measures the prediction error by taking the mean of all squared absolute values of all errors, that is:

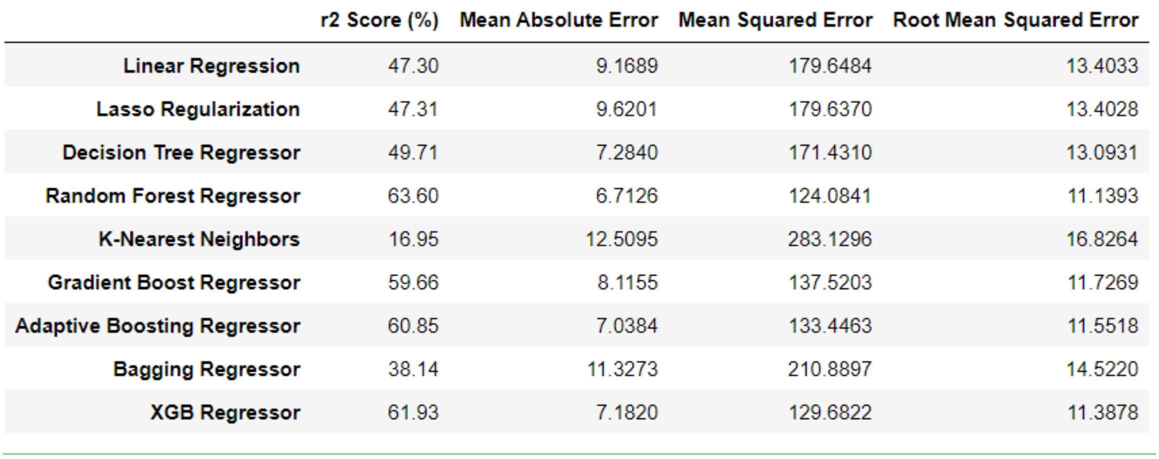
$$MSE = \frac{\sum_{i=0}^{n}(y_i - \hat{y_i})^2}{n}$$

As for MAE, values close to 0 are better.

5.1.4 Root Mean squared error (RMSE)

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2/N]^{1/2}$$

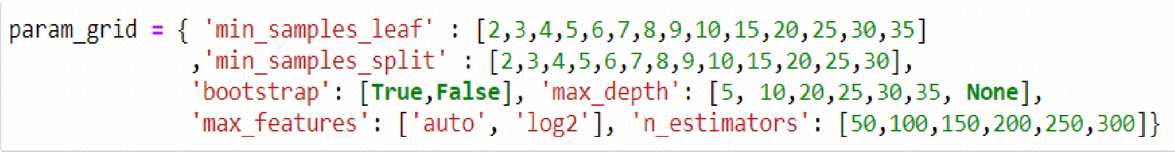


	r2 Score (%)	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	47.30	9.1689	179.6484	13.4033
Lasso Regularization	47.31	9.6201	179.6370	13.4028
Decision Tree Regressor	49.71	7.2840	171.4310	13.0931
Random Forest Regressor	63.60	6.7126	124.0841	11.1393
K-Nearest Neighbors	16.95	12.5095	283.1296	16.8264
Gradient Boost Regressor	59.66	8.1155	137.5203	11.7269
Adaptive Boosting Regressor	60.85	7.0384	133.4463	11.5518
Bagging Regressor	38.14	11.3273	210.8897	14.5220
XGB Regressor	61.93	7.1820	129.6822	11.3878

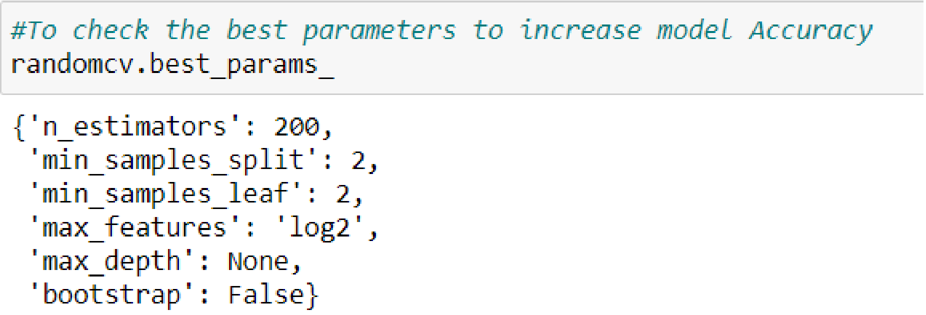
fig : Models used and their evaluation metrics

5.2 Tuning of Hyperparameters

In this subsection, we present the optimal hyperparameters to tune our best model i.e. Random Forest Regression model with the following hyperparameters:



```
param_grid = { 'min_samples_leaf' : [2,3,4,5,6,7,8,9,10,15,20,25,30,35]
              ,'min_samples_split' : [2,3,4,5,6,7,8,9,10,15,20,25,30],
              'bootstrap': [True,False], 'max_depth': [5, 10,20,25,30,35, None],
              'max_features': ['auto', 'log2'], 'n_estimators': [50,100,150,200,250,300]}
```

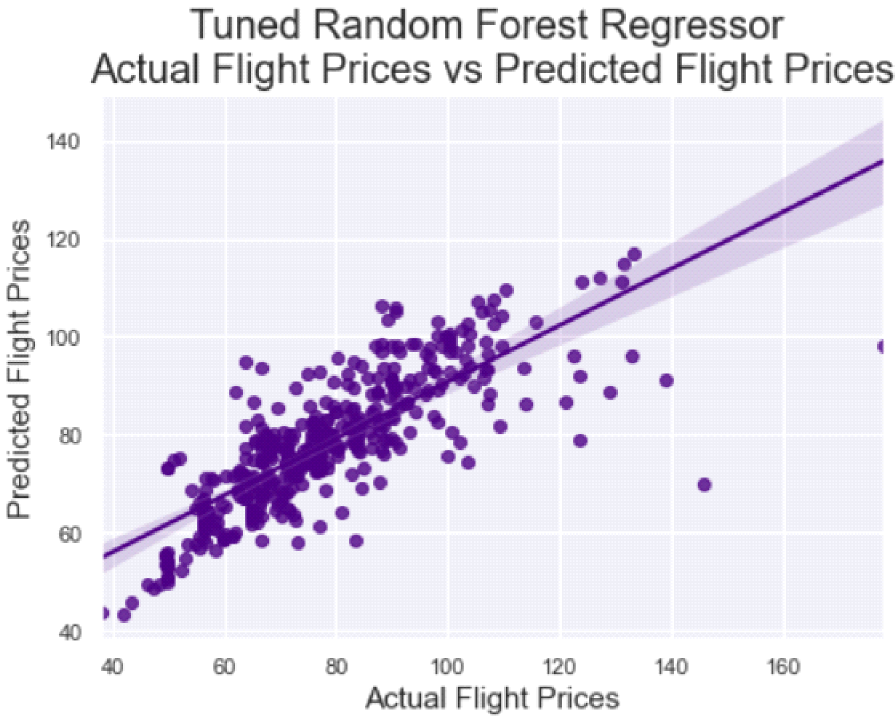


```
#To check the best parameters to increase model Accuracy
randomcv.best_params_
```

```
{'n_estimators': 200,
 'min_samples_split': 2,
 'min_samples_leaf': 2,
 'max_features': 'log2',
 'max_depth': None,
 'bootstrap': False}
```

Best params

We've deployed Randomized Search CV for Hyperparameter Tuning of our model with 3 folds for each of 300 candidates, totalling 900 fits.



We concluded the following error metrics for our final tuned model :

- R2 Score for Tuned Lasso Regression Model: 0.6517
- Mean Absolute Error for our Tuned Lasso Regression Model: 6.9662
- Mean Squared Error for our Tuned Lasso Regression Model: 118.7438
- Root Mean Squared Error for our Tuned Lasso Regression Model: 10.8969

6. Conclusions:

Our Ensemble models performed well than Linear Regression, Lasso Regularization. Although, there's still scope for improvement. Random Forest Regression Performed better than all other models, whereas, K-Nearest Neighbours performed really poorly.

Flight prices almost always remain constant or increase between the major cities
There are two groups of airlines: the economical group and the luxurious group. SpiceJet, AirAsia, IndiGo, Go Air are in the economical class, whereas Jet Airways and Air India in the other. Vistara has a more spread out trend.

7. Future Scope

A Flight reservation system can be further developed which can book tickets on the basis of needs as per clients. Flight reservation system shoots up the sales of an airline company and gives a competitive edge.

When we book online flight tickets it gives us the freedom to navigate as per our budget requirements and more.

For further analysis, More routes can be added and the same analysis can be expanded to major airports and travel routes in India.

The analysis can be done by increasing the data points and increasing the historical data used. That will train the model better giving better accuracies and more savings.