Machine Learning Project

Predict The Flight Ticket Price

Name: R.Amshu Naik

Roll No: B20CS046

Introduction:

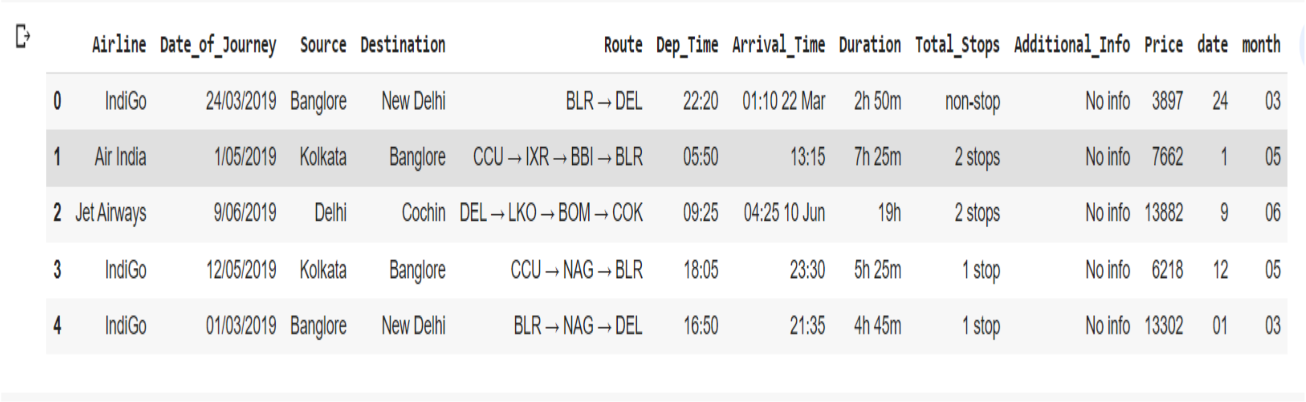
Nowadays, the number of people using flights has increased significantly. It is difficult for airlines to maintain prices due to complexity of the algorithm to calculate flight prices under various conditions. That's why we will try to use machine learning to solve this problem. This can help airlines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

Datasets:

The file ***Dataset.xlsx - Sheet1.csv*** is used as the training dataset.

The train dataset contains 10683 rows where each row represents a train-image with 11 columns containing :

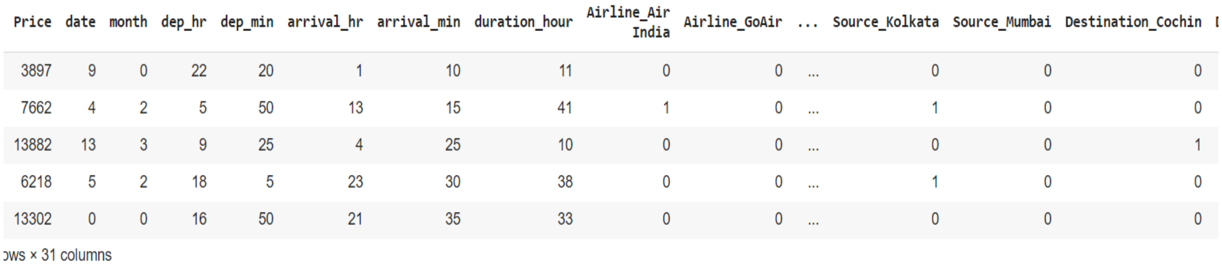
- One Airline column
- One Price column: continuous form .
- 9 latent vector columns



	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	date	month
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897	24	03
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662	1	05
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882	9	06
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218	12	05
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302	01	03

Data preprocessing:

- Splitted Date of journey column into 2 separate columns date and month.
- Splitted Dep_time and Arrival_time into 2 separate columns each of the Hour of departure and another the min of departure and same for Arrival time too.
- Splitted Duration column to only 2 seperate columns of hour of duration and to minutes of duration.
- Label encoding of columns “Routes”,”Additional_Info”,Total_stops”.
- One hot encoding of columns “Airline”,”Source”,”Destination”.
- Dropped all other columns except for “price” and the new feature column created.



Price	date	month	dep_hr	dep_min	arrival_hr	arrival_min	duration_hour	Airline_Air India	Airline_GoAir	...	Source_Kolkata	Source_Mumbai	Destination_Cochin
3897	9	0	22	20	1	10	11	0	0	...	0	0	0
7662	4	2	5	50	13	15	41	1	0	...	1	0	0
13882	13	3	9	25	4	25	10	0	0	...	0	0	1
6218	5	2	18	5	23	30	38	0	0	...	1	0	0
13302	0	0	16	50	21	35	33	0	0	...	0	0	0

ows × 31 columns

Methodology:

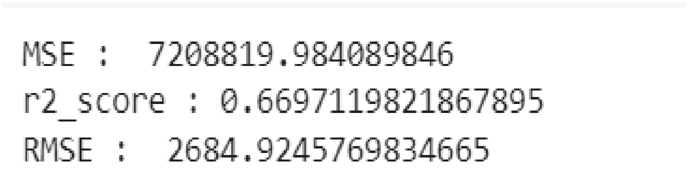
There are various classification algorithms present out of which we shall implement the following

- KNN
- Linear Regression
- Decision Tree Classifier
- Random Regression model

We calculated the r2 score, Accuracy, Root mean square error(RSME) and Mean square error (MSE) of the model.

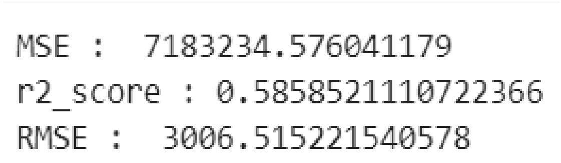
Implementation of classification algorithms:

1. **KNN Means:** K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It stores all the available cases and classifies the new data or case based on a similarity measure.



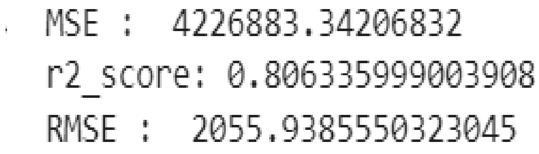
```
MSE :  7208819.984089846
r2_score : 0.6697119821867895
RMSE :  2684.9245769834665
```

2. **Linear Regression**: linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables .More specifically, that y can be calculated from a linear combination of the input variables (x).



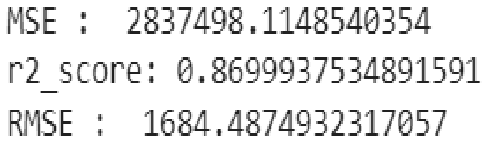
```
MSE :  7183234.576041179
r2_score : 0.5858521110722366
RMSE :  3006.515221540578
```

3. **Decision Tree Classifier:** Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete.



```
MSE :  4226883.34206832
r2_score: 0.806335999003908
RMSE :  2055.9385550323045
```

4. **Random Regression Model:** A Random Forest is an ensemble technique capable of performing regression with the use of multiple decision trees and a technique called Bootstrap and Aggregation, known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.



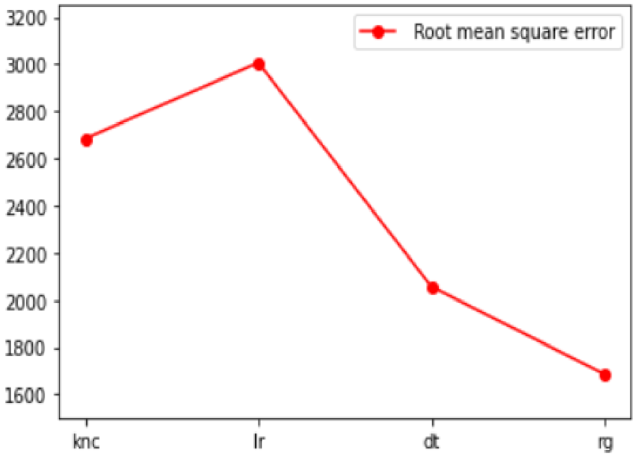
```
MSE :  2837498.1148540354
r2_score: 0.8699937534891591
RMSE :  1684.4874932317057
```

Comparison between models:

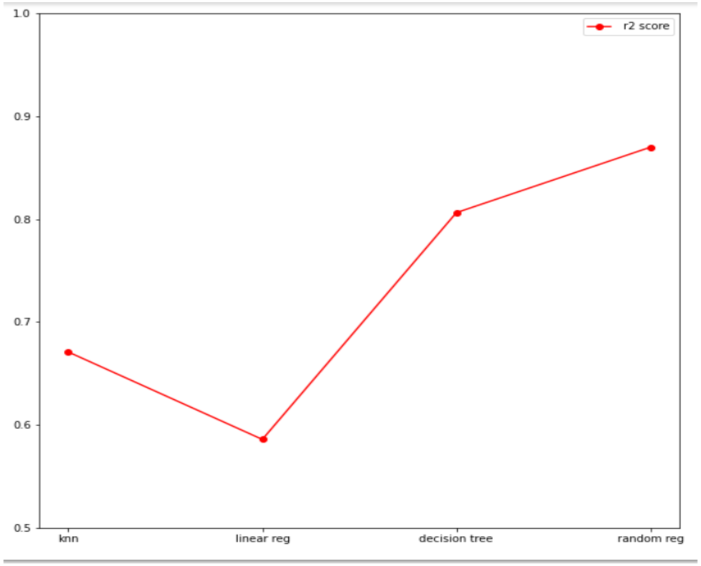
- *Tabular representation of evaluation metrics for models used:-*

Models	KNN	Linear Regression	Decision Tree	Random Regression tree
Accuracy	42.35%	50.55%	83.56%	83.56%
r2_score	2684.924	0.585852	0.8063	0.86994
MSE	7208819.98	7208819.98	4226883.342	2837498.11
RMSE	2684.924	3006.515	2055.9385	1684.487

- *Comparison of RMSE value of all the 4 models.*



Comparison of r2 score value of all the 4 models.



Conclusion:

- I am using RMSE and r2 score value as evaluation metric for the dataset.
- If the RMSE value is low, then that regression module is a good module .
- If the r2 score value is high for a module, then that module is good for our dataset.
- From the above 2 comparison graphs between the 4 models we used in our dataset, the RMSE value of the random regression model is low and its r2_score value is high ,so among these 4 models the random Regression model is the best and suitable model for our dataset.

Result:

'Random forest Regression model' is a suitable model for our dataset among the other 3 models used.

- ➔ RMSE value- 1684.487
- ➔ R2 score - 0.86994
- ➔ Accuracy- 83.56%