FLIP ROBO

PROJECT REPORT ON:

“Flight Price Prediction”

SUBMITTED BY

Ajit Madame

ACKNOWLEDGMENT

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. And I want to express my huge gratitude to Ms.Gulshana Chaudhary (SME Flip Robo), she is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to my academic team “Data trained” who are the reason behind what I am today. Last but not least my parents who have been my backbone in every step of my life. And also thank you for many other persons who has helped me directly or indirectly to complete the project.

Contents:

1.INTRODUCTION

1.1Business Problem Framing:

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 –

- 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)

Therefore, a predictive model to accurately predict Air fares is required to be made.

1.2 Conceptual Background of the Domain Problem

Predictive modelling, Regression algorithms are some of the machine learning techniques used for predicting Flight Ticket prices. Identifying various relevant attributes like Airline Brand, flight duration, source and destination etc are crucial for working on the project as they determine the valuation of air fare.

1.2 Review of Literature

A Research paper titled: “Airline ticket price and demand prediction: A survey” by Juhar Ahmed Abdella and online article titled: “Trying to Predict Airfares When The Unpredictable Happens” were reviewed and studied to gain insights into all the attributes that contribute to the pricing of flight tickets.

It is learnt that deterministic features like Airline Brand, flight number, departure dates, number of intermediate stops, week day of departure, number of competitors on route and aggregate features – which are based on collected historical data on minimum price, mean price, number of quotes on non-stop,1-stop and multi-stoppage flights are some the most important factors that determine the pricing of Flight Tickets.

- Airline ticket price and demand prediction: A survey -ScienceDirect
- Flight Price Predictor | American Express GBT (amexglobalbusinesstravel.com)

1.2 Motivation for the Problem Undertaken

With airfares fluctuating frequently, knowing when to buy and whento wait for a better deal to come along is tricky. The fluctuation in prices is frequent and one has limited time to book the cheapest ticket as the prices keep varying due to constant manipulation by Airline companies. Therefore, it is necessary to work on a predictive model based on deterministic and aggregate feature data that would predict with good accuracy the most optimal Air fare for a particular destination, route and schedule.

2.Analytical Problem Framing

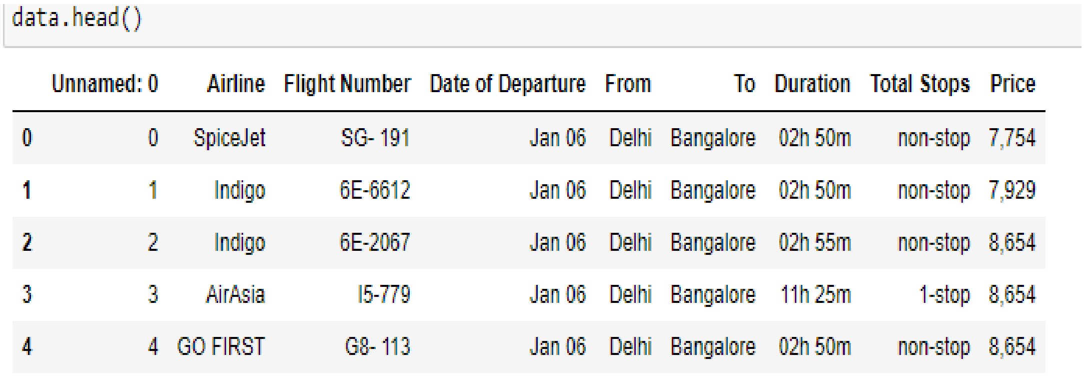
2.1 Mathematical/ Analytical Modelling of the Problem

Various Regression analysis techniques were used to build predictive models to understand the relationships that exist between Flight ticket price and Deterministic and Aggregate features of Air travel. The Regression analysis models were usedto predict the Flight ticket price value for changes in Air travel deterministic and aggregate attributes. Regression modelling techniques were used in this Problem since Air Ticket Price data distribution is continuous in nature.

In order to forecast Flight Ticket price, predictive models such as ridge regression Model, Random Forest Regression model, Decision tree Regression Model, Support Vector Machine Regression model and Extreme Gradient Boost Regression modelwere used to describe how the values of Flight Ticket Price depended on the independent variables of various Air

Fare attributes.

2.2 Data Sources and their formats



```
data.head()
```

	Unnamed: 0	Airline	Flight Number	Date of Departure	From	To	Duration	Total Stops	Price
0	0	SpiceJet	SG- 191	Jan 06	Delhi	Bangalore	02h 50m	non-stop	7,754
1	1	Indigo	6E-6612	Jan 06	Delhi	Bangalore	02h 50m	non-stop	7,929
2	2	Indigo	6E-2067	Jan 06	Delhi	Bangalore	02h 55m	non-stop	8,654
3	3	AirAsia	I5-779	Jan 06	Delhi	Bangalore	11h 25m	1-stop	8,654
4	4	GO FIRST	G8- 113	Jan 06	Delhi	Bangalore	02h 50m	non-stop	8,654

The Dataset was compiled by scraping Data for various Air Fareattributes and Price from https://www.easemytrip.com/

The data was converted into a Pandas Dataframe under variousFeature and Label columns and saved as a .csv file.

Dataset Description

The Independent Feature columns are:

- Airline: The name of the airline.
- Flight Number: Number of Flight
- Date of Departure: The date of the journey
- From: The source from which the service begins
- To: The destination where the service ends
- Duration: Total duration of the flight
- Total Stops: Total stops between the source and destination.

Target / Label Column:

- Price: The Price of the Ticket

2.3 Data Pre-processing Done

- Duplicate data elements in various columns: ‘Airline’,’From’,’To’,which had their starting letters in upper case and lower case were converted to data elements starting with uppercase letters.
- Data in column ‘Price’ was converted to int64 data type.
- Columns: Unnamed: 0(just a series of numbers) was droppedsince it doesn't contribute to building a good model for predicting the target variable values.
- The Date format of certain data elements in ‘Date of Departure’was changed to match the general Date format of majority of the data elements of the column.

2.4 Data Inputs - Logic - Output Relationships

- The Datasets consist mainly of Int and Object data type variables.The relationships between the independent variables and dependent variable were analysed.

2.5 Hardware, Software and Tool Used

Hardware Used:

Processor – Intel core i3

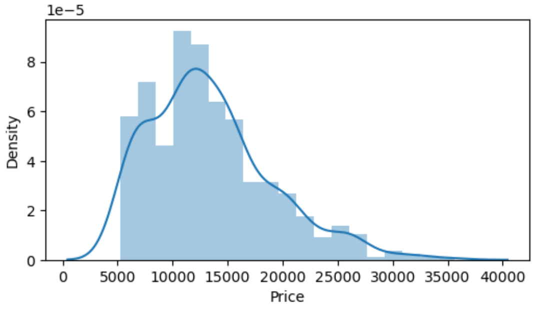
Physical Memory – 8 GB

Software Used:

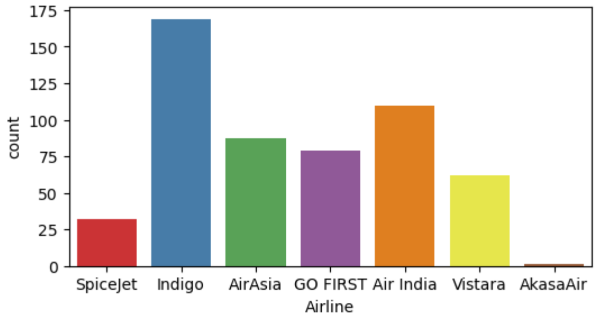
- Windows 10 Operating System
- Anaconda Package and Environment Manager
- Jupyter Notebook
- Python Libraries used: In Which Pandas, Seaborn, Matplotlib, Numpy and Scipy
- sklearn for Modelling Machine learning algorithms, Data Encoding, Evaluation metrics, Data Transformation, Data Scaling, Component analysis, Feature selection etc.

3.Data Analysis and Visualization

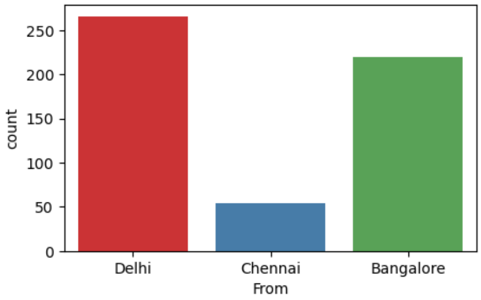
3.1 Univariate Visualization



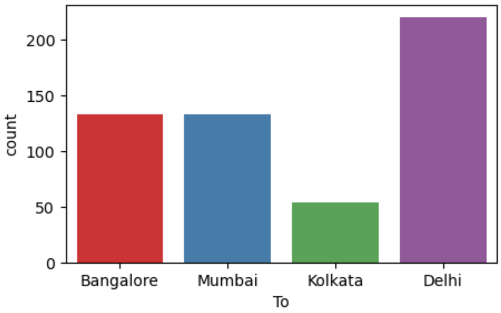
- Distribution is skewed and tails of from 15000 mark.



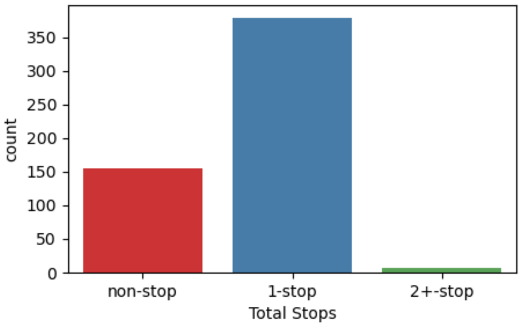
- IndiGo has the highest number of flights followed by Air India and Air Asia.



- Highest number of flights are from Delhi followed by Bangalore.

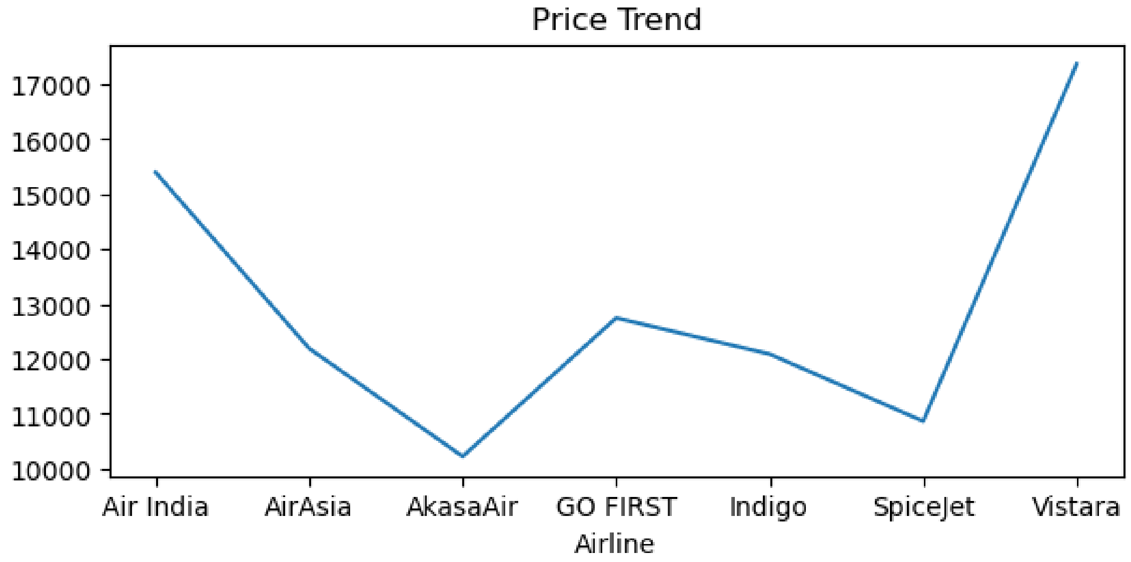


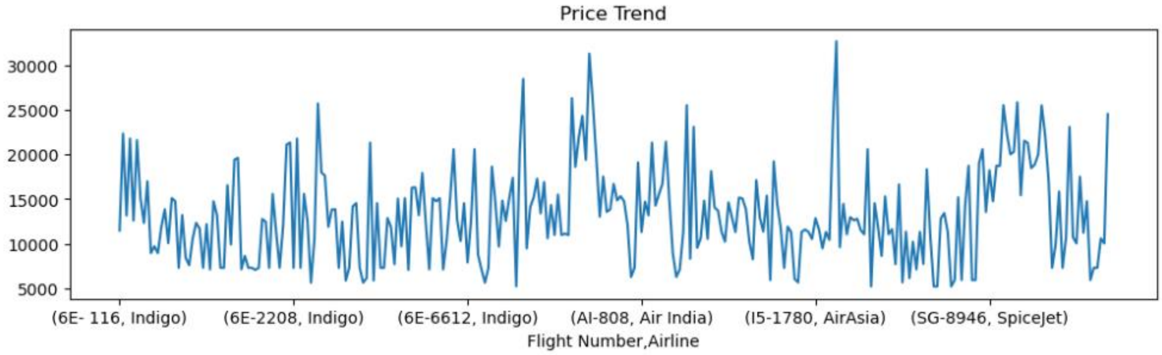
- Delhi is the most popular destination followed by Mumbai, Bangalore.



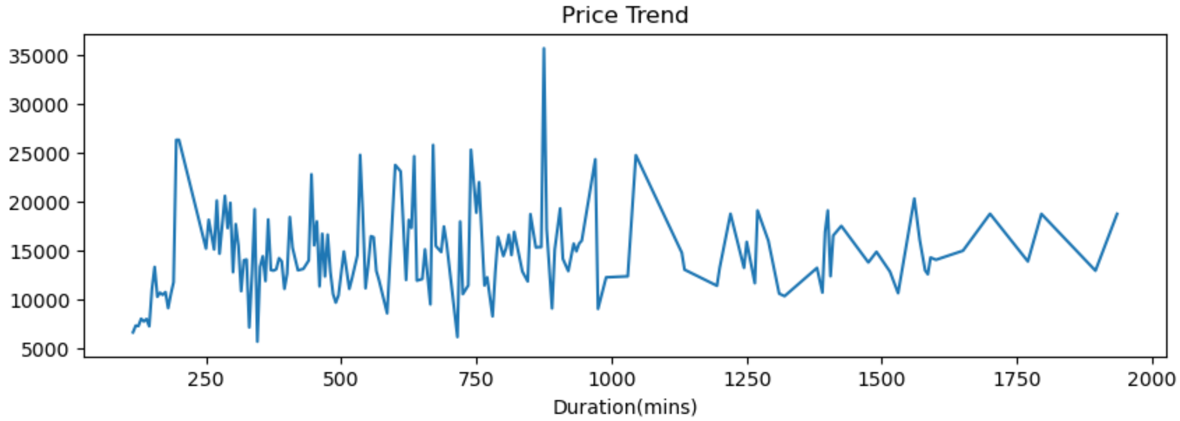
- Highest number of flights have only 1 stop between source and destination while 2nd highest number of flights are non stop.

3.2 Bivariate Visualization

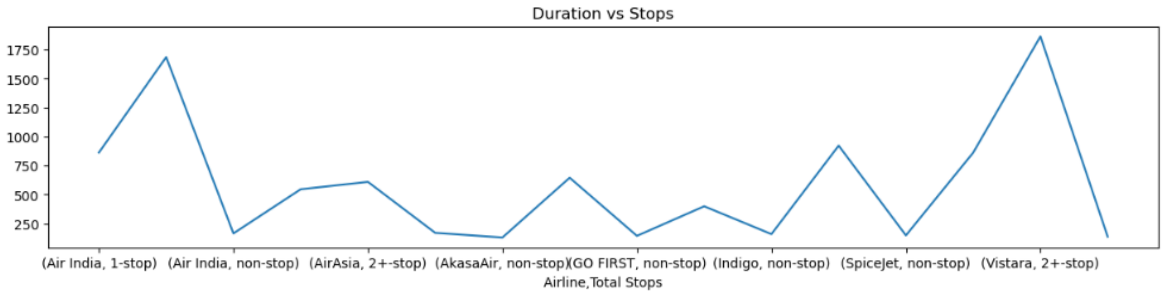




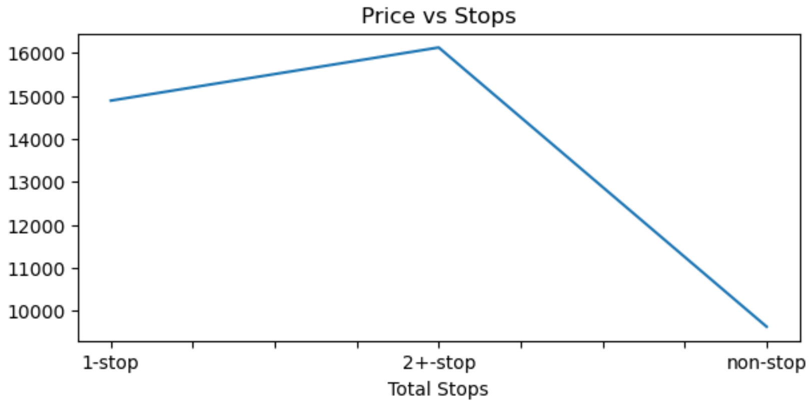
IndiGo and SpiceJet offer air tickets at the most affordable prices on average, whereas Air Asia, Air India are the most expensive on average.



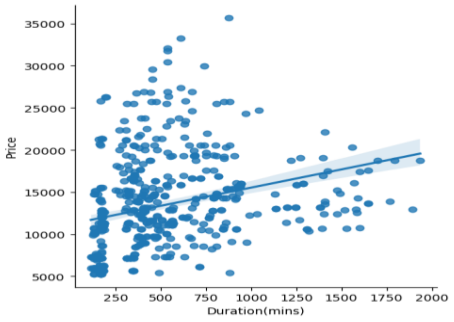
- we can see, mostly duration of flight is maximum lies in between 250 to 750 and they have 15000 to 25000 prices of ticket.



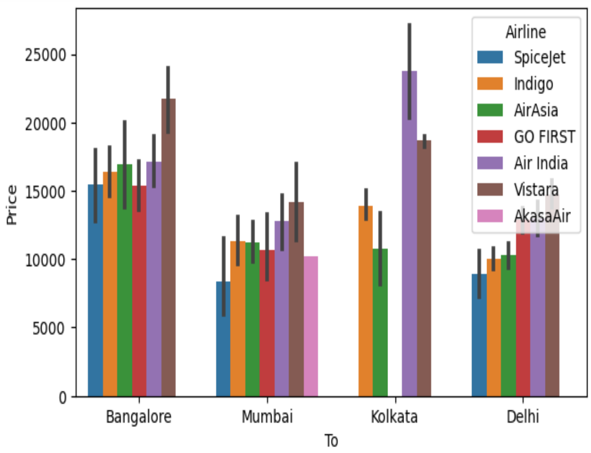
- It can be observed that Number of Stops impact the travel time of Airlines.



- It can be observed that Number of Stops impact the Air Ticket Pricing of Airlines.

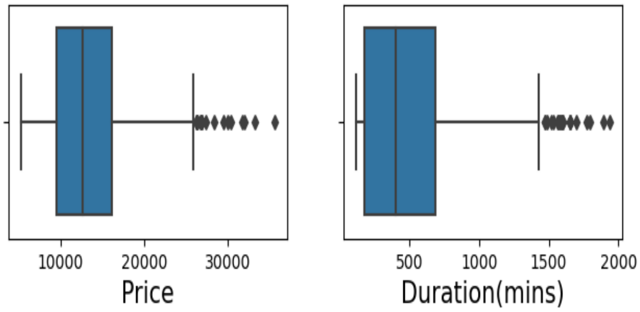


- There is a linear relationship between Price and flight duration.



- Bangalore,Kolakata are the most expensive destinations while Mumbai and Delhi are the most affordable destination.
- SpiceJet and Indigo provide most affordable Airtickets to the destinations.

Checking for Outliers



There are considerable outliers in the columns. Outliers were Removed using Z score method which resulted in a totaldata loss of 1.48%, which is within acceptable range.

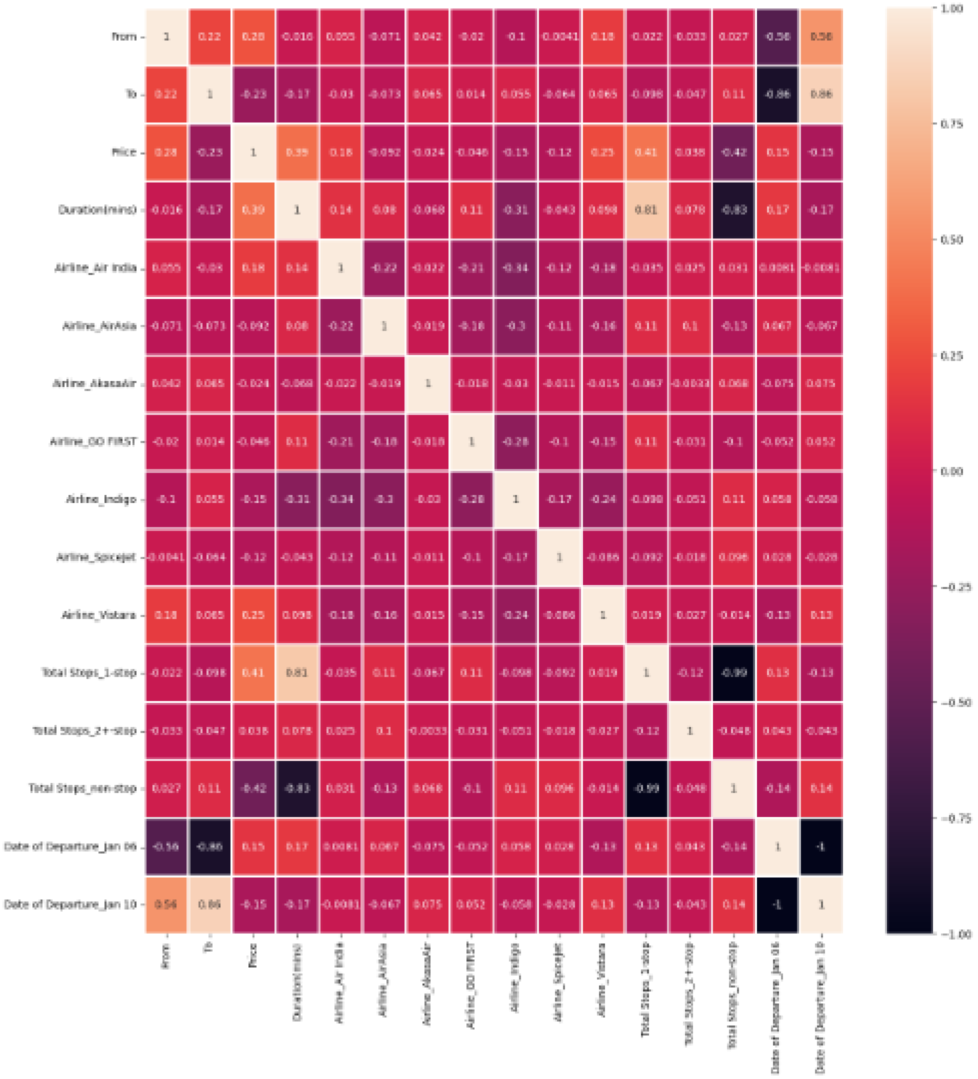
Data Normalization

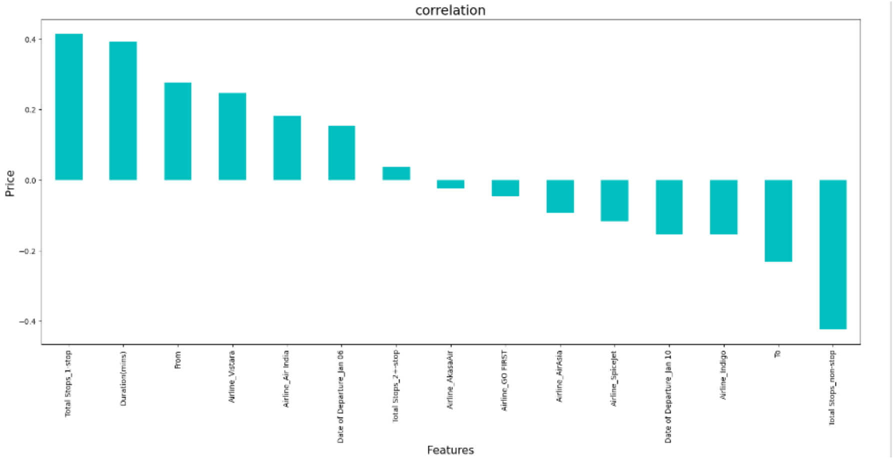
Data in Column ‘Duration(mins) was normalized using PowerTransformer technique.

Encoding Categorical Columns

Categorical Columns were encoded using Label Encodingtechnique and get_dummies() technique.

Finding Correlation between Feature and Targetcolumns

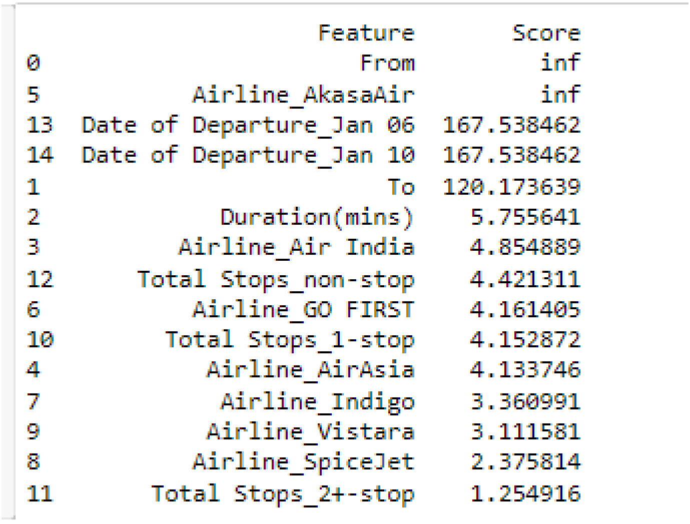




- It is observed thatTotal Stops_1-stop, Duration(min) and From have the highest positive correlation with Price, while Total - - Stops_non-stop,to have the highest negative correlation with Price.

4. Models Development and Evaluation

4.1 Features Selection

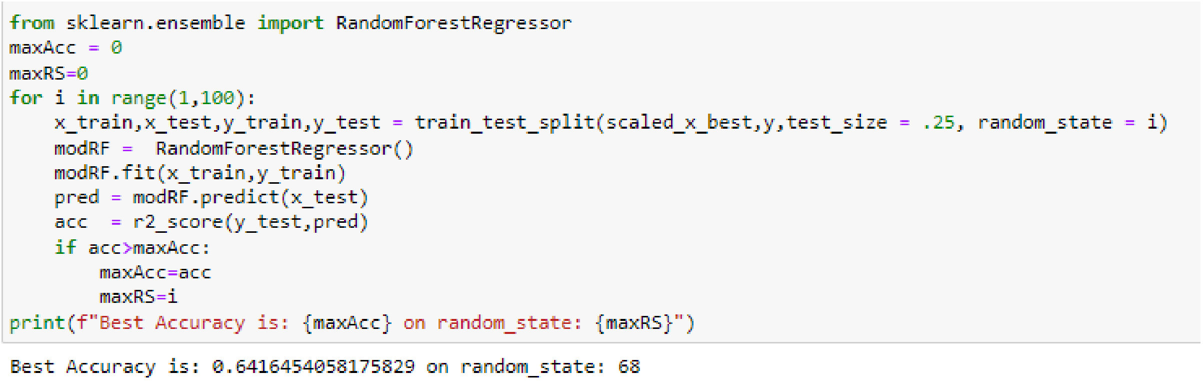


```
                     Feature        Score
0                       From          inf
5            Airline_AkasaAir          inf
13  Date of Departure_Jan 06   167.538462
14  Date of Departure_Jan 10   167.538462
1                         To   120.173639
2             Duration(mins)     5.755641
3          Airline_Air India     4.854889
12       Total Stops_non-stop    4.421311
6           Airline_GO FIRST     4.161405
10         Total Stops_1-stop    4.152872
4            Airline_AirAsia     4.133746
7             Airline_Indigo     3.360991
9            Airline_Vistara     3.111581
8           Airline_SpiceJet     2.375814
11        Total Stops_2+-stop    1.254916
```

Using SelectKBest and f_classif for measuring the respective ANOVA f-score values of the columns, the best features were selected. Using StandardScaler, the features were scaled by resizing the distribution values so that mean of the observed values in each feature column is 0 and standard deviation is 1. From sklearn.model_selection’s train_test_split, the data was divided into train and test data. Training data comprised 75% oftotal data where as test data comprised 25% based on the bestrandom state that would result in best model accuracy.

4.2 The model algorithms used

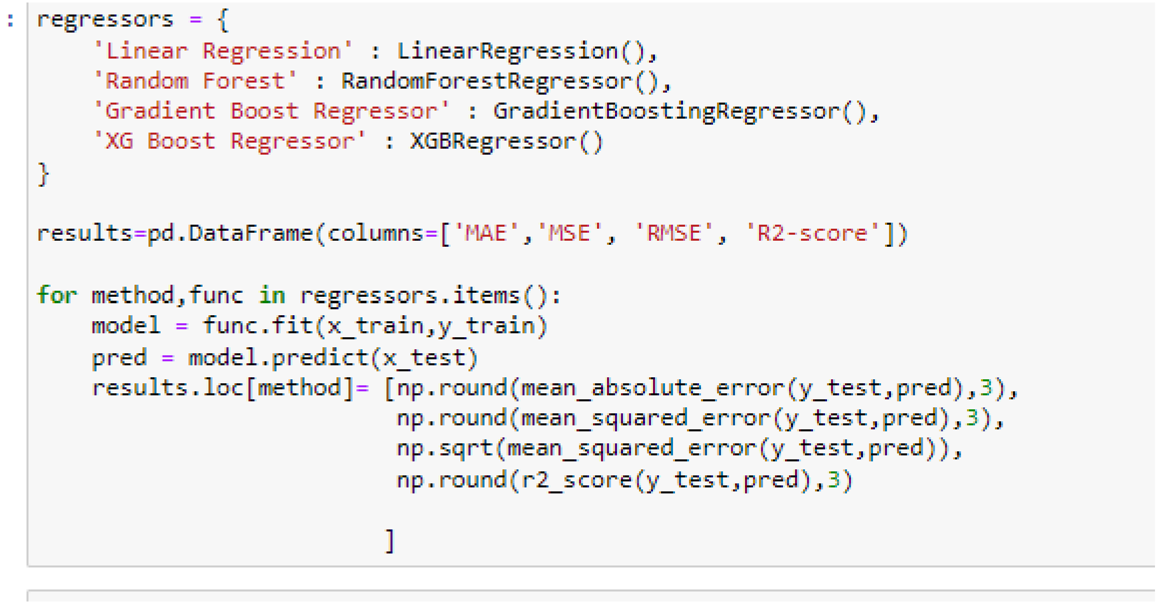
Finding Best Random State



```
from sklearn.ensemble import RandomForestRegressor
maxAcc = 0
maxRS=0
for i in range(1,100):
    x_train,x_test,y_train,y_test = train_test_split(scaled_x_best,y,test_size = .25, random_state = i)
    modRF =  RandomForestRegressor()
    modRF.fit(x_train,y_train)
    pred = modRF.predict(x_test)
    acc  = r2_score(y_test,pred)
    if acc>maxAcc:
        maxAcc=acc
        maxRS=i
print(f"Best Accuracy is: {maxAcc} on random_state: {maxRS}")
```

```
Best Accuracy is: 0.6416454058175829 on random_state: 68
```

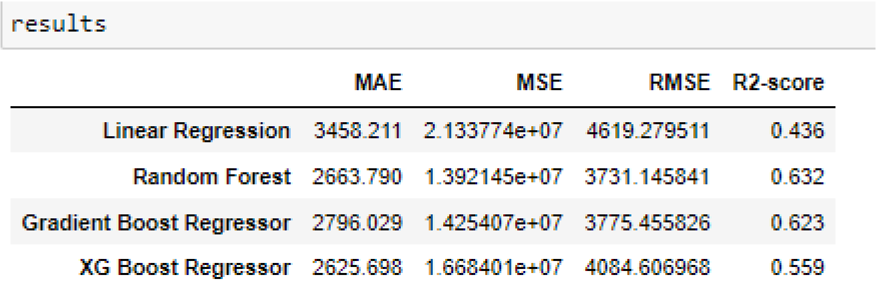
Trained the Models



```
regressors = {
    'Linear Regression' : LinearRegression(),
    'Random Forest' : RandomForestRegressor(),
    'Gradient Boost Regressor' : GradientBoostingRegressor(),
    'XG Boost Regressor' : XGBRegressor()
}

results=pd.DataFrame(columns=['MAE','MSE', 'RMSE', 'R2-score'])

for method,func in regressors.items():
    model = func.fit(x_train,y_train)
    pred = model.predict(x_test)
    results.loc[method]= [np.round(mean_absolute_error(y_test,pred),3),
                          np.round(mean_squared_error(y_test,pred),3),
                          np.sqrt(mean_squared_error(y_test,pred)),
                          np.round(r2_score(y_test,pred),3)

                         ]
```

Analyzing Accuracy of The Models

Mean Squared Error and Root Mean Squared Error metrics wereused to evaluate the Model performance. The advantage of MSEand RMSE being that it is easier to compute the gradient. As, wetake square of the error, the effect of larger errors become more pronounced than smaller error, hence the model can now focus more on the larger errors.

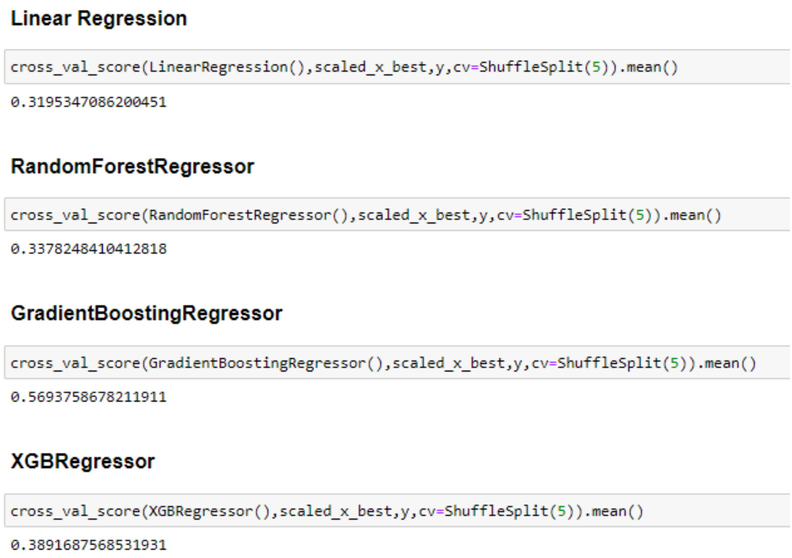


```
results
```

	MAE	MSE	RMSE	R2-score
Linear Regression	3458.211	2.133774e+07	4619.279511	0.436
Random Forest	2663.790	1.392145e+07	3731.145841	0.632
Gradient Boost Regressor	2796.029	1.425407e+07	3775.455826	0.623
XG Boost Regressor	2625.698	1.668401e+07	4084.606968	0.559

Cross validation is a technique for assessing how the statistical analysis generalises to an independent data set.It is a technique for evaluating machine learning models by training several modelson subsets of the available input data and evaluating them on the complementary subset of the data.

Using cross-validation, there are high chances that we can detectover-fitting with ease. Model Cross Validation scores were then obtained for assessing how the statistical analysis generalises to an independent data set. The models were evaluated by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.



Linear Regression

```
cross_val_score(LinearRegression(),scaled_x_best,y,cv=ShuffleSplit(5)).mean()
```

```
0.3195347086200451
```

RandomForestRegressor

```
cross_val_score(RandomForestRegressor(),scaled_x_best,y,cv=ShuffleSplit(5)).mean()
```

```
0.3378248410412818
```

GradientBoostingRegressor

```
cross_val_score(GradientBoostingRegressor(),scaled_x_best,y,cv=ShuffleSplit(5)).mean()
```

```
0.5693758678211911
```

XGBRegressor

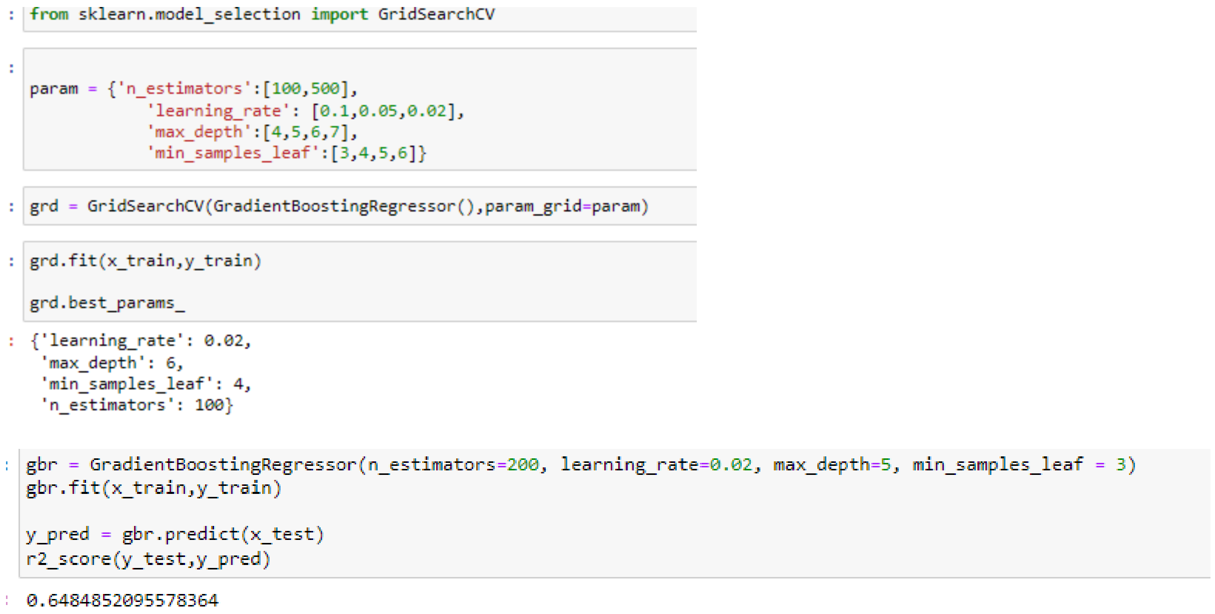
```
cross_val_score(XGBRegressor(),scaled_x_best,y,cv=ShuffleSplit(5)).mean()
```

```
0.3891687568531931
```

4.3 Interpretation of the results

Based on comparing Accuracy Score results with Cross Validation results, it is determined that Gradient Boosting Regressor is the best model.

4.4 Hyperparameter Tuning



```
: from sklearn.model_selection import GridSearchCV
```

```
:
  param = {'n_estimators':[100,500],
              'learning_rate': [0.1,0.05,0.02],
              'max_depth':[4,5,6,7],
              'min_samples_leaf':[3,4,5,6]}
```

```
: grd = GridSearchCV(GradientBoostingRegressor(),param_grid=param)
```

```
: grd.fit(x_train,y_train)

  grd.best_params_
```

```
: {'learning_rate': 0.02,
   'max_depth': 6,
   'min_samples_leaf': 4,
   'n_estimators': 100}
```

```
: gbr = GradientBoostingRegressor(n_estimators=200, learning_rate=0.02, max_depth=5, min_samples_leaf = 3)
  gbr.fit(x_train,y_train)

  y_pred = gbr.predict(x_test)
  r2_score(y_test,y_pred)
```

```
: 0.6484852095578364
```

5. Conclusions

5.1 Key Finding and Conclusions

Based on the in-depth analysis of the Flight Price Prediction Project, TheExploratory analysis of the datasets, and the analysis of the Outputs of the models the following observations are made:

- Air Fare attributes like Date,Month,Duration,Total Stops etc play a bigrole in influencing the used Flight price.
- Airline Brand also has a very important role in determining the usedFlight Ticket price.
- Various plots like Barplots,Countplots and Lineplots helped invisualising the Feature-label relationships which corroborated the importance of Air Fare features and attributes for estimating FlightTicket Prices.
- Due to the Training dataset being very small, only very small amount ofthe outliers was removed to ensure proper training of the models.
- Therefore, Random Forest Regressor, which uses averaging to improve the predictive accuracy and controls over-fitting. performed welldespite having to work on small dataset and produced good predictions that can be understood easily.

5.2 Limitation of this works and scope for future works

A small dataset to work with posed a challenge in building highly

accurate models. This project also relied heavily on historical data andwas unable to account for various other factors that influence demandand ticket pricing like pandemic status affecting demand, government regulations on air travel, shifting in routes, weather conditions, etc.

Most airline companies also do no publicly make available their ticket pricing strategies, which makes gathering price and air fare related datasets using web scraping the only means to build a dataset for building predicting models.

Availability of more features and a larger dataset would help build bettermodels.