FLIGHT PRICE PREDICTION

REGRESSION MODEL

SEEP BANSAL |Machine Learning Model | 30-01-2022

ACKNOWLEDGEMENT

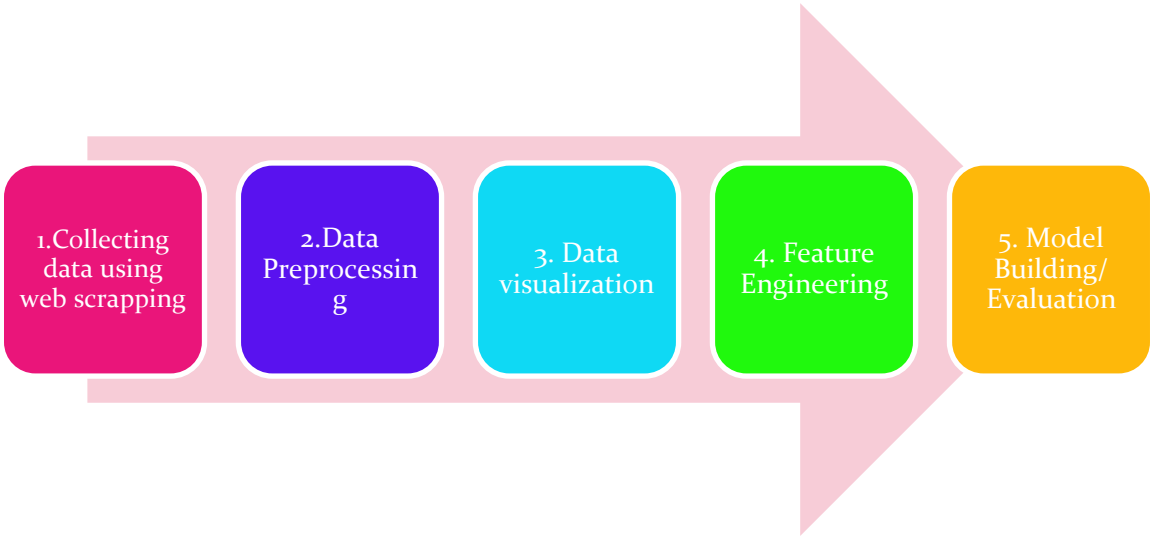
I would like to express my special thanks of gratitude to the team Data Trained and my Mentor MS Deepika Sharma for their exemplary guidance, monitoring and constant encouragement thought the journey of learning Data science and Machine learning techniques. I would also like to express my heartly gratitude to the support team of data trained for their constant support. Last but not the least, I would also like to thank the team of Flip Robo technologies for giving me this opportunity to work on this project and the mentors in Flip Robo Technologies who are constantly guiding me to enhance my knowledge and work. This project helped me not only to learn how to do proper research but also helped me in learning many new things.

INTRODUCTION

Nowadays, flights are considered as the major source of transportation i.e people do travel frequently via air. Flight price is something which is very hard to guess. The various airline companies use complex algorithms to predict the price of the flight. These companies have their own factors to decide the fares. The price of the flights increases or decreases unexpectedly. This usually happens as these airline companies wants to generate maximum revenue.

PROJECT SUMMARY

Determining the price of the flight is one of the toughest things to predict as we may see a price today and again check out the price of the same flight tomorrow and it will be a different story. In order to understand this problem, we will try to take the help of some machine learning algorithms. The various information regarding the flights like source, destination, number of stops, duration, arrival time and departure time is collected to predict the fluctuating fares. Various machine learning algorithms like Linear Regression, Support Vector Mechanism, Decision Tree Regressor, Random Forest Regressors, K Neighbours Regressor are implemented and evaluated to predict the price of the car across different cities in India. The best results are given by Random Forest Regressor. Though conventional Linear Regression also gave the good results with the advantage of significantly lower training time as compared to the aforementioned algorithm.



ANALYTICAL PROBLEM FRAMING

DATASET

The above dataset is the data of the Domestic flight fares collected from one of the most leading websites "www.yatra.com" and www.makemytrip.com" using the various web scrapping tools like beautiful soup and selenium. The dataset consists of 1576 rows and around 9 features are scrapped from the above-mentioned site. Features like Source, Destination, Duration, Date of journey Number of stops and price are collected. As we are supposed to predict the fares/price of the flight tickets, so price will be our output variable and features like Source, Destination, duration, Arrival time, Departure time, number of stops will be our independent features which will help us in predicting the price of the flights.

	Unnamed: 0	Airline	Source	Date of Journey	Destination	Departure	Arrival	Duration	Stops	Price
0	0	SpiceJet	Mumbai	8 Mar 2022	Chennai	19:30	21:20	1h 50m	Non Stop	1,888
1	1	SpiceJet	Mumbai	8 Mar 2022	Chennai	06:00	08:00	2h 00m	Non Stop	1,888
2	2	Go First	Mumbai	8 Mar 2022	Chennai	05:50	07:40	1h 50m	Non Stop	1,890
3	3	IndiGo	Mumbai	8 Mar 2022	Chennai	16:25	18:15	1h 50m	Non Stop	1,890
4	4	IndiGo	Mumbai	8 Mar 2022	Chennai	21:25	23:15	1h 50m	Non Stop	1,890
...	...	...	...	...	...	...	...	...	...	...
1571	74	Vistara	New Delhi	24 Feb 2022	Kolkata	09:30	16:55	07 h 25 m	1 stop via Mumbai	4,868
1572	75	Vistara	New Delhi	24 Feb 2022	Kolkata	14:20	19:55	05 h 35 m	1 stop via Mumbai	4,868
1573	76	Vistara	New Delhi	24 Feb 2022	Kolkata	12:50	19:55	07 h 05 m	1 stop via Mumbai	4,868
1574	77	Spicejet	New Delhi	24 Feb 2022	Kolkata	09:25	14:20	04 h 55 m	1 stop via Bagdogra	4,898
1575	78	Go First	New Delhi	24 Feb 2022	Kolkata	11:10	15:50	04 h 40 m	1 stop via Bagdogra	5,692

Fig 1: Image of the dataset.

Feature Description

1. Airline: The name of the airline company listed on the website.
2. Source: From where the flight is taking off.
3. Destination: Where the flight lands.
4. Date of Journey: Date on which you are travelling.
5. Departure: Time at which the flight departs
6. Arrival: Time at which the flight arrives.
7. Duration: Travel time of the flight.
8. Stops: The number of layovers the flight will take.

DATA PREPROCESSING/DATA CLEANING

Preprocessing is one of the important steps in building a model. In this phase we usually deal with missing values if any or if there are any unrealistic values. In case of any irrelevant value, we will remove that data. In case if the data loss is huge then removing/dropping of data is not a good practice. We will try to improve the quality of data in this phase so that we can develop a model with high accuracy score. For the dataset, we will first of all fill the missing values if any, then we have most of the data in categorial form we will convert the data in numerical form so that we are able to fed the data into classification algorithms. We can also check if there exist any multicollinearities through VIF FACTOR calculation.

EDA concluding Remarks

- ✓ Typecasting of the various columns is done i.e datatypes of the columns were changed.
- ✓ New columns were created from the existing columns.
- ✓ Null values are removed.
- ✓ Label Encoding is done as there were many categorical values that were converted to numeric values.
- ✓ Skewness was removed.
- ✓ Scaling is done to bring all the variables on same scale.
- ✓ VIF factor is calculated to check the multicollinearities.

HARDWARE AND SOFTWARE REQUIREMENTS

Hardware Requirements: Hardware Requirements followed while developing this model:

- Intel core i5
- 11^{th} generation
- 16 GB RAM
- Windows 10

Software Requirements: Software required are:

- Anaconda Navigator (64-bit Graphical Installer)
- Juypter Notebook
- Microsoft Edge
- Knowledge of Python Language and Machine learning Algorithms

Libraries Used:

```
import pandas as pd   #importing the libraries
import numpy as np
import warnings
warnings.filterwarnings("ignore")
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression,Lasso,Ridge
from sklearn.metrics import mean_squared_error, mean_absolute_error,r2_score
from sklearn.model_selection import train_test_split,cross_val_score,GridSearchCV
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from scipy.stats import zscore
from sklearn.tree import DecisionTreeRegressor
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import roc_curve,roc_auc_score
from sklearn.model_selection import GridSearchCV
import statsmodels.api as sm
from scipy import stats
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

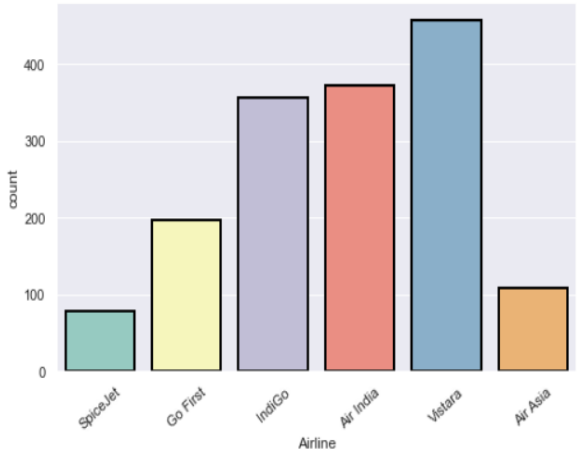
Fig 2: Libraries Used

- Import and export of data takes place with the help of pandas
- All the numerical operations are carried out with the help of numpy library.
- Matplotlib.pyplot and seaborn libraries helps in graphical representation of data.
- Warnings library is used to ignore the unwanted warnings
- Sklearn library helps in importing all the machine learning algorithms and evaluation matrix that are required.
- Scipy library helps VIF factor calculation and zscore which is used for removal of outliers.

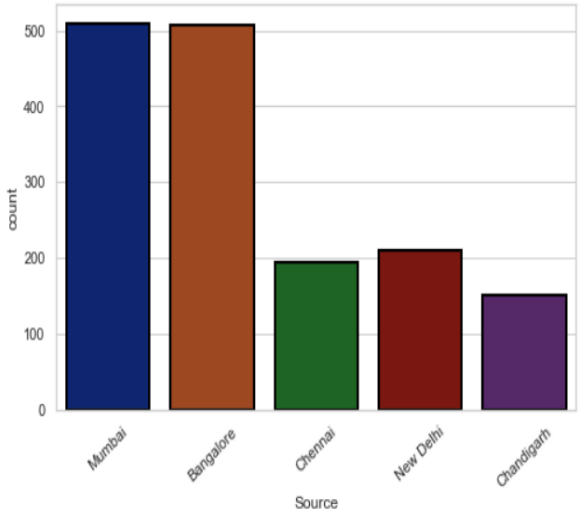
MODEL DEVELOPMENT/MODEL EVALUATION

VISUALIZATION

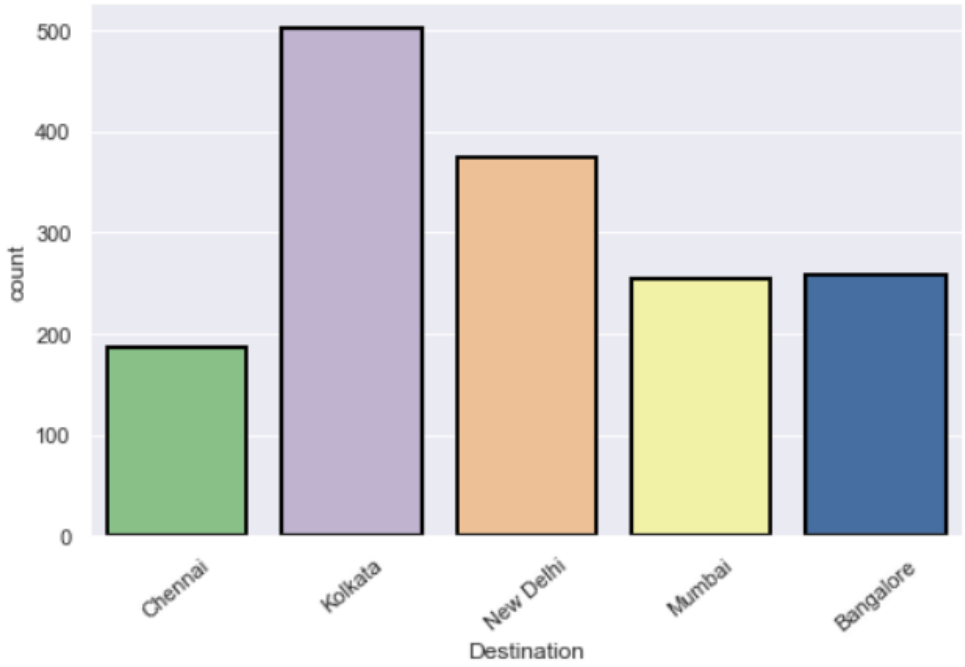
Following are some of the visualizations/Graphical Representations created to interpret the data:



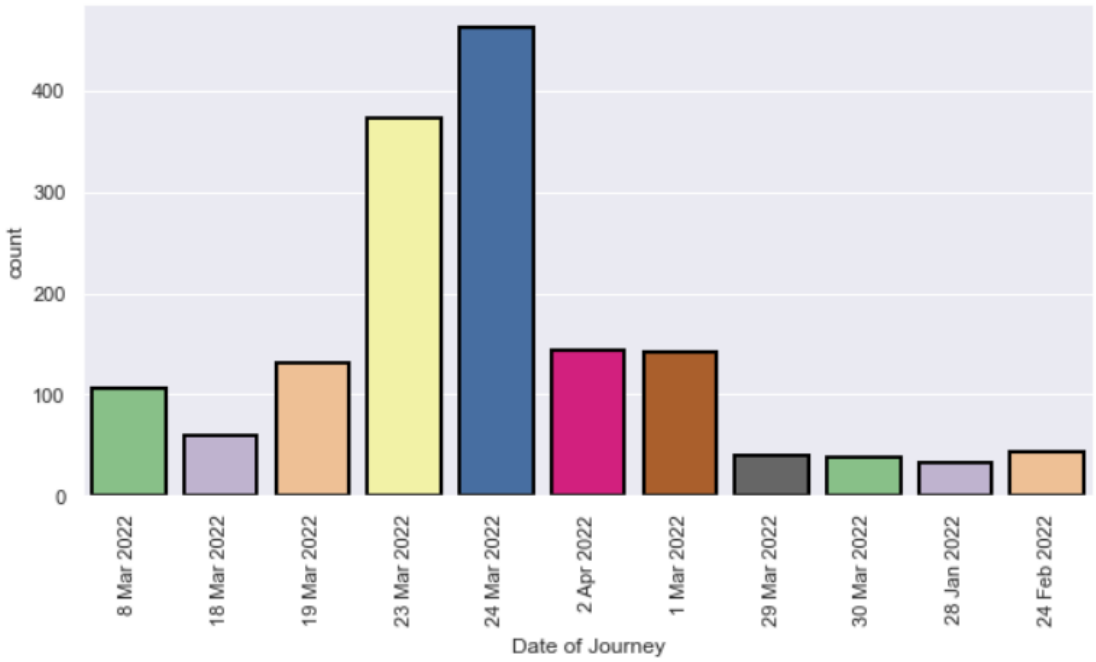
From the above plot we can see that the Vistara number of flights count is very high i.e majority of vistara flights are listed on the site filowed by AirIndia then Indigo. Spicejet has least number of listed flights.



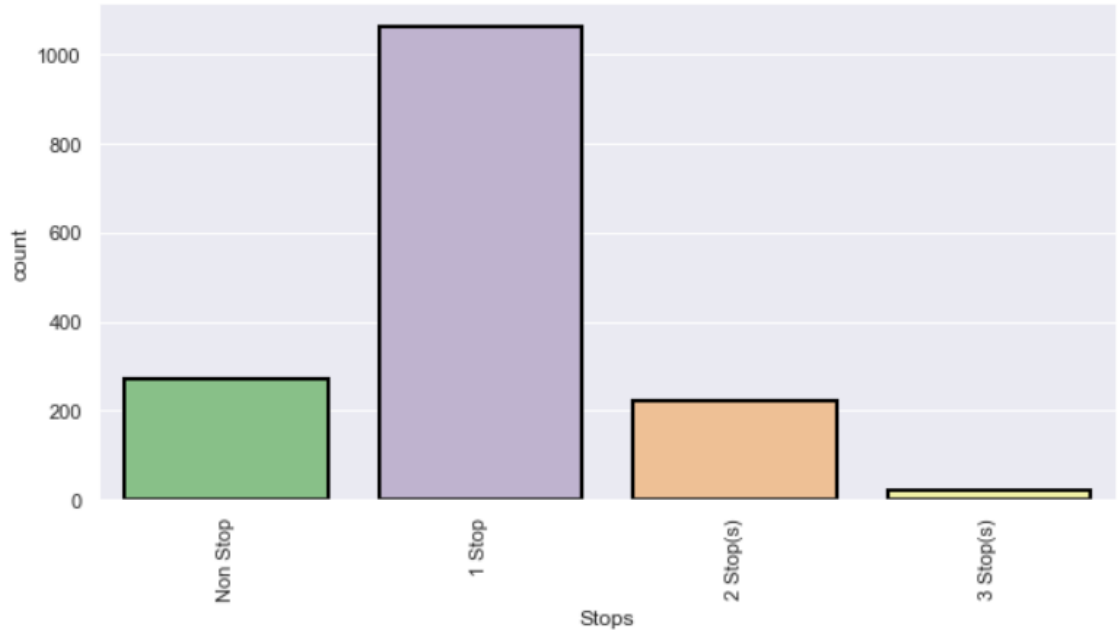
Mumbai and Bangaluru airport followed by New Delhi has the highest number of flights taking off from the airport listed on the site. As these are the capital cities of india and the airports in these are listed among top 50 airports in the global list.



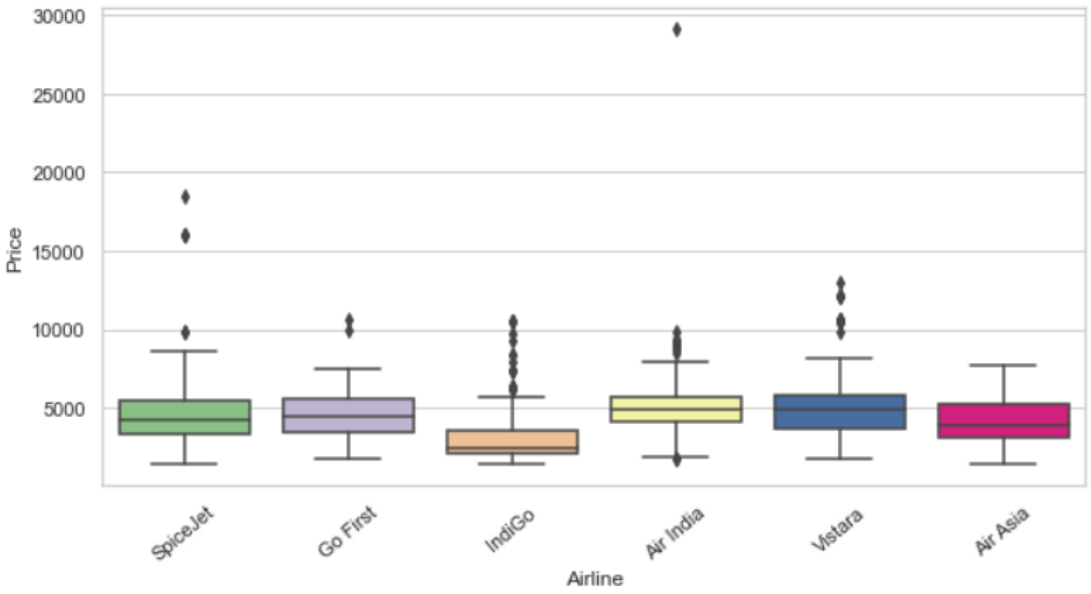
From the data collected the majority of flights lands in Kolkata and new Delhi....



Data collected for 23rd and 24th march have highest number of flights listed on the website.

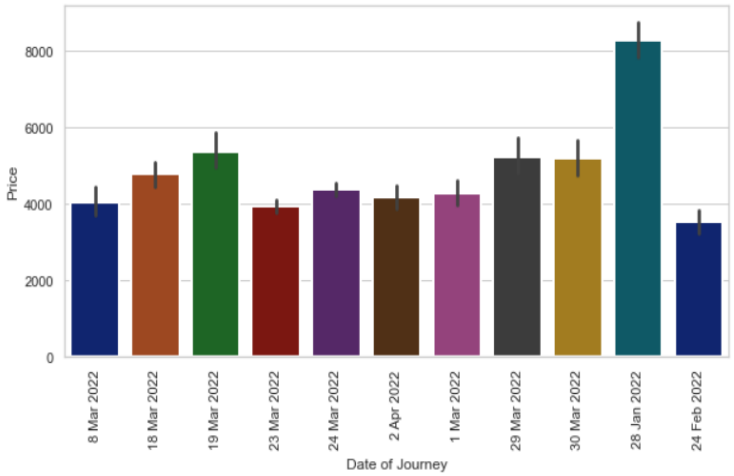


Majority of the flights have 1 layover before reaching the final destination.



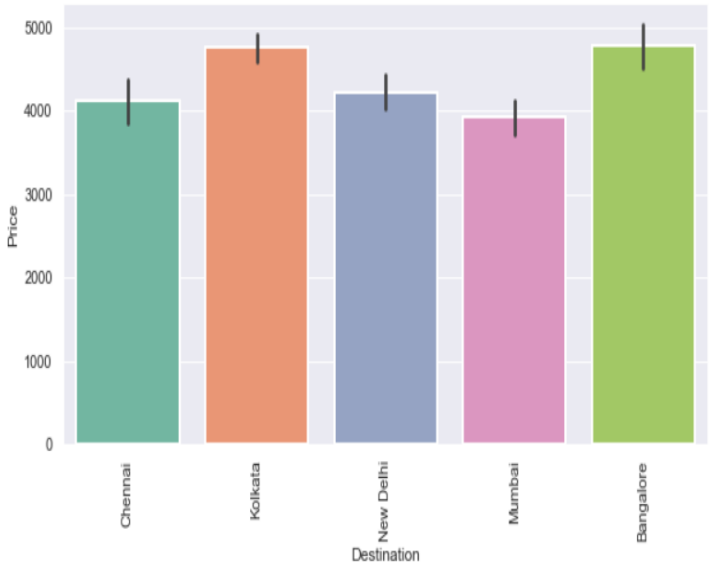
If we look at the above boxplot,

1. Indigo airlines offer the cheapest services as compare to any other airlines.
2. AirIndia has a flight fare which touches 30000.
3. AirIndia followed by Vistara is offering expensive services as comapre to other airlines.

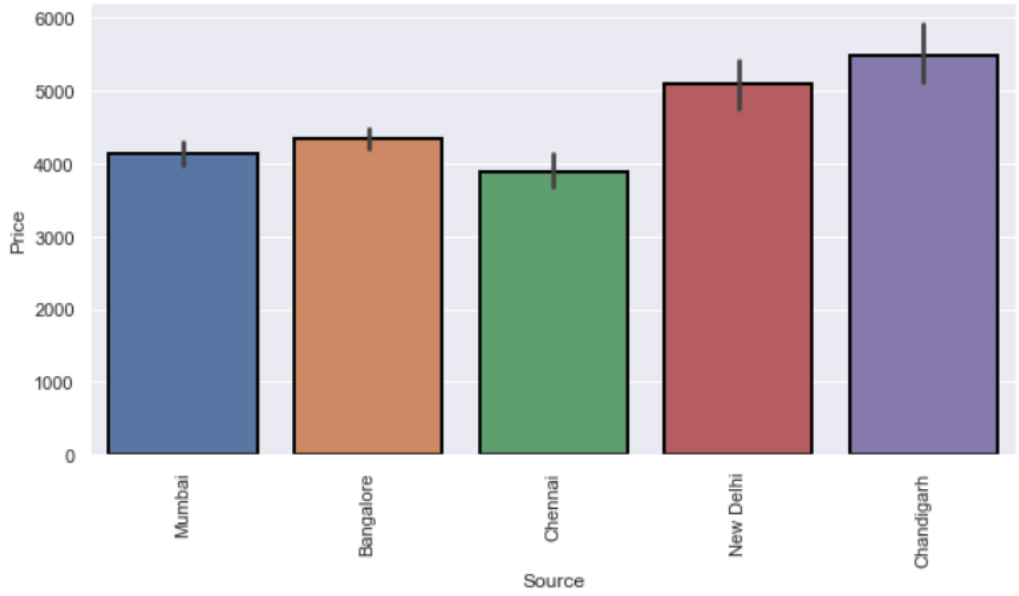


In the above barplot,

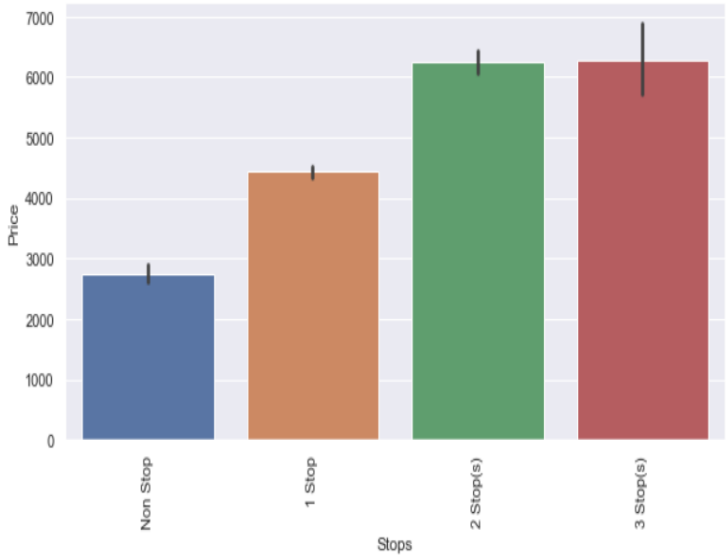
1. As the date was approaching i.e 28th January the flight fares were most expensive for the same.
2. 19th march which is saturday, flights are expensive as compared weekdays even though the flights are listed in march month.
3. If we look at consecutive week days i.e 29th and 30th march there is no much difference in the change of price.
4. If we look at consecutive week days i.e 18th and 19th march which is weekend there is much difference in the change of price i.e price increases on weekends.
5. Flights are cheaper on weekdays as compare to weekends.



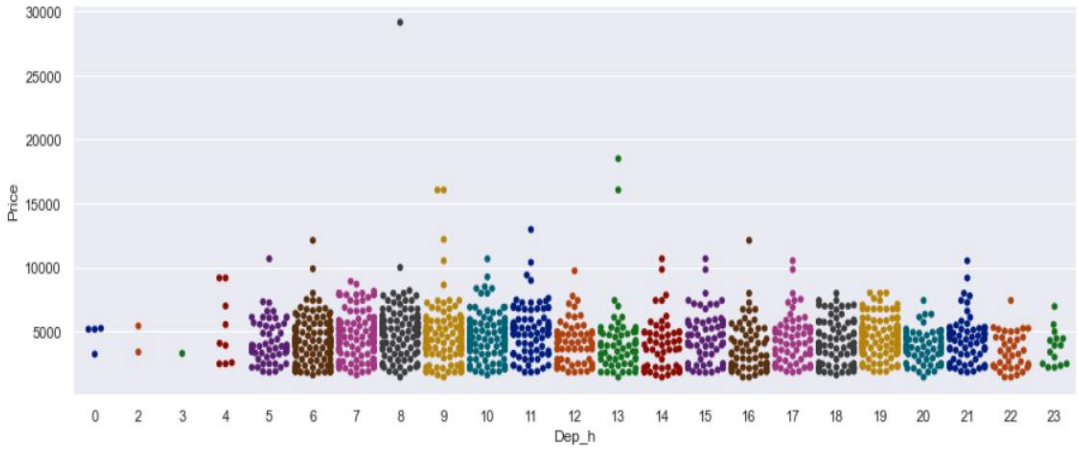
In the above plot we can see that kolkata nad bengalore are the expensive destinations but other factors do come into picture like what is the source and the day on which we are booking the flights.



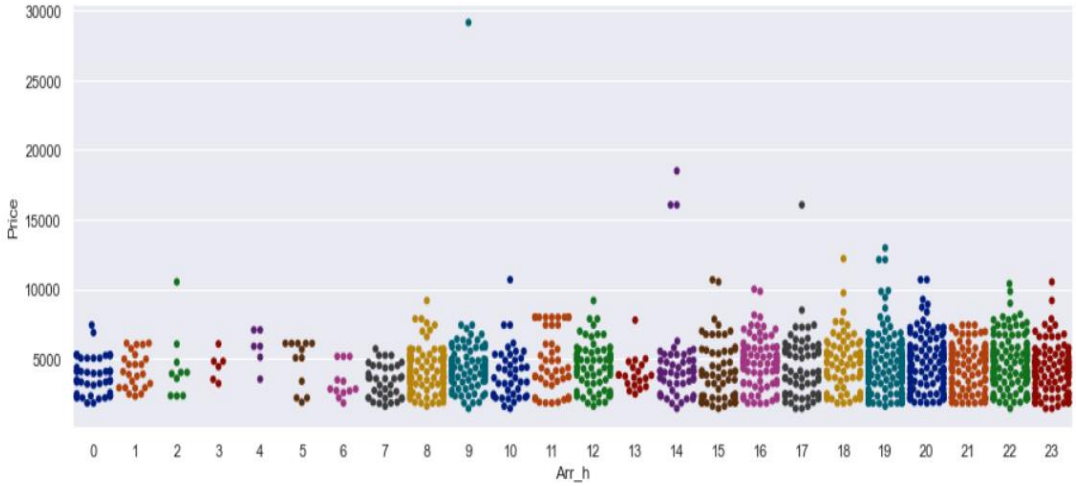
Chnadigarh offers high flight prices as compare to other airport.



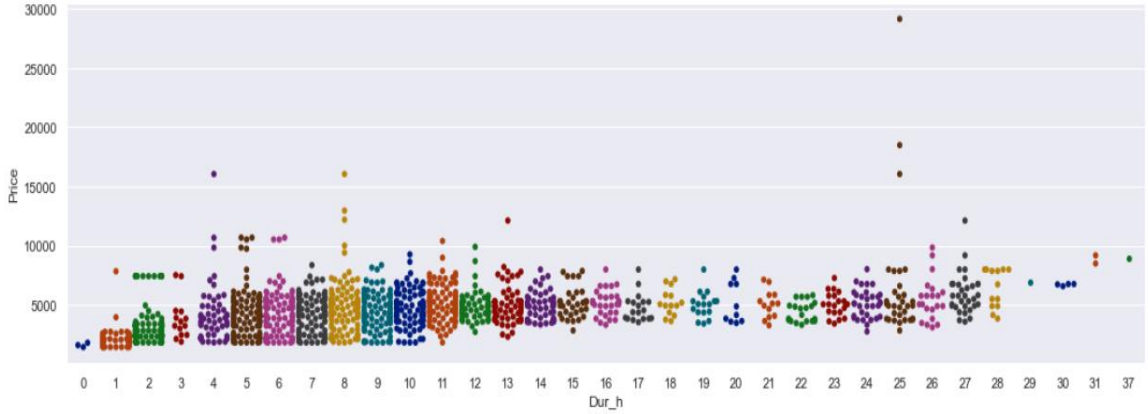
Flights with 2 and 3 layovers are expensive as compared to Non stop flights. But again we cannot conclude this from the above data as this is a very small data others factors like date of journey, source and destination.



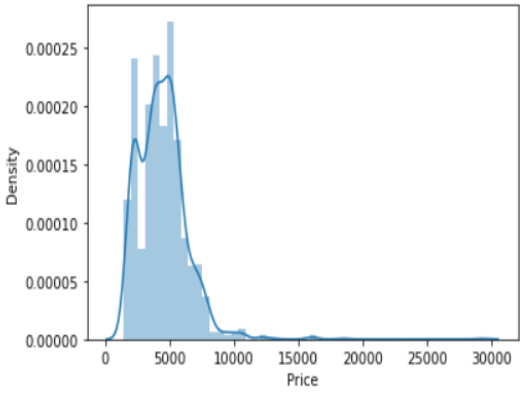
From the above plot we say that, the flights that depart at 7am to 11am are expensive whereas the flights that depart late night are cheaper.



From the above plot we can say that flights that arrive between 2pm to 7pm are expensive.



If the duration of the flight is between 4 to 8 hours then the flight may cost expensive.



Distribution of price we can see that some of the flights are highly expensive which makes the presence of outliers and makes data skewed.

In the above heatmap,

1. Duration hour is contributing positively in predicting the price.
2. Stops is contributing least in predicting the price.
3. duration and stops are negetively co related.

ALGORITHMS USED For Training and Testing the Model:

- Linear Regression Model
- Decision Tree Regressor
- Random Tree Forest Regressor
- K Neighbors Regressor
- Support Vector Regressor

REGRESSION EXPERIMENTS

We have performed various Regression algorithms like Linear Algorithm, Decision Tree Algorithm, Support vector Regressor, K Neighbors Regressor, Random Tree Regressor to check the MSE, MAE, RMSE. The main algorithm which has predicted the best results is the Random Forest Tree Regressor. Random forest regression is an ensemble learning technique which takes multiple algorithms or same algorithm multiple times and put together a model that's more powerful than the original.

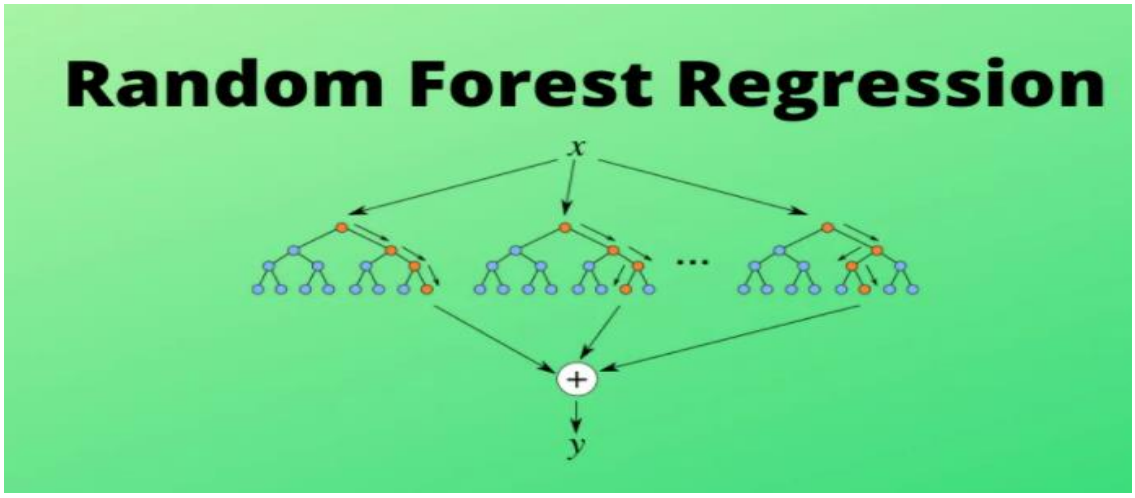


FIG: Random Forest Regression

EVALUATION

Evaluation of model plays a very important role in evaluating the performance of any Regression. The metrices that are evaluated here are the R2 score, MSE, MAE, RMSE.Time taken to test the model on dataset plays a very crucial role.

Here, for Random Tree Forest Regressor the accuracy score we are getting is 99% and the MSE is around 2618141.9 , MAE is 691.06 and RMSE is 1618.06 which is least in any of the model build and the computational time is very less.

```
final=RandomForestRegressor(max_depth=19, max_features='log2', oob_score=False,bootstrap=False)
final.fit(x_train1,y_train1)
sw1=final.score(x_train1,y_train1)
#print("Coffecient is: ",rfr.coef_)
#print("Intercept is: ",rfr.intercept_)
print("Score is: ",sw1)
pred=final.predict(x_test1)
print("Mean Squared Error is:",round(mean_squared_error(y_test1,pred),2))
print("Mean Absolute Error is:",round(mean_absolute_error(y_test1,pred),2))
print("R2 Score is:",round(r2_score(y_test1,pred),2))
print("RSME",np.sqrt(mean_squared_error(y_test1,pred)))
```

```
Score is:  0.9999172088394229
Mean Squared Error is: 2618141.9
Mean Absolute Error is: 691.06
R2 Score is: 0.5
RSME 1618.0673336561476
```

CONCLUSION, LIMITATIONS AND FUTURE SCOPE

Though it is very tough to understand the price fluctuations of the flight fares, by developing this model and performing different algorithms was aimed to get the different perspectives. The various data visualization techniques were used. Data was analyzed from different point of views many preprocessing techniques like scaling, label encoding etc were followed. The relation between different features were examined and the best model Random Forest Regressor with the best results were used to predict the best fares.

The above model is used to predict the fares of the flights. However, this was relatively a small dataset with around 1600 rows was used make a strong inference and the data was collected form only 2 site.

Gathering more data from different sites or form the same site could lead to more better results. Collecting more features like whether meals are included or not, insurance is inclusive or not could actually help is in building a robust model.

Another room of improvement is more data preprocessing and more data cleaning techniques to be followed which could help is reducing RMSE. We could actually build this model with different scaling technique like min-max scaler which may give us better results or we could also try different encoding technique which may enhance the results.