FLIP ROBO

PROJECT REPORT ON:

Flight Price Prediction Project

SUBMITTED BY

RAHUL RANJAN

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.

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:

- Testing of Identified Approaches (Algorithms)
- Key Metrics for success in solving problem under consideration
- Visualization
- Run and Evaluate selected models
- Interpretation of the Results

- Conclusion
  - Key Findings and Conclusions of the Study
  - Learning Outcomes of the Study in respect of Data Science
  - Limitations of this work and Scope for Future Work

1.INTRODUCTION

1.1 Business Problem Framing:

The tourism industry is changing fast and this is attracting a lot more travelers each year. The airline industry is considered as one of the most sophisticated industry in using complex pricing strategies. Now-a-days flight prices are quite unpredictable. The ticket prices change frequently. Customers are seeking to get the lowest price for their ticket, while airline companies are trying to keep their overall revenue as high as possible. Using technology it is actually possible to reduce the uncertainty of flight prices. So here we will be predicting the flight prices using efficient machine learning techniques.

When booking a flight, travelers need to be confident that they're getting a good deal. The Flight Price Analysis API uses an Artificial Intelligence algorithm trained on Amadeus historical flight booking data to show how current flight prices

compare to historical fares. More precisely, it shows how a current flight price sits on a *distribution* of historical airfare prices.

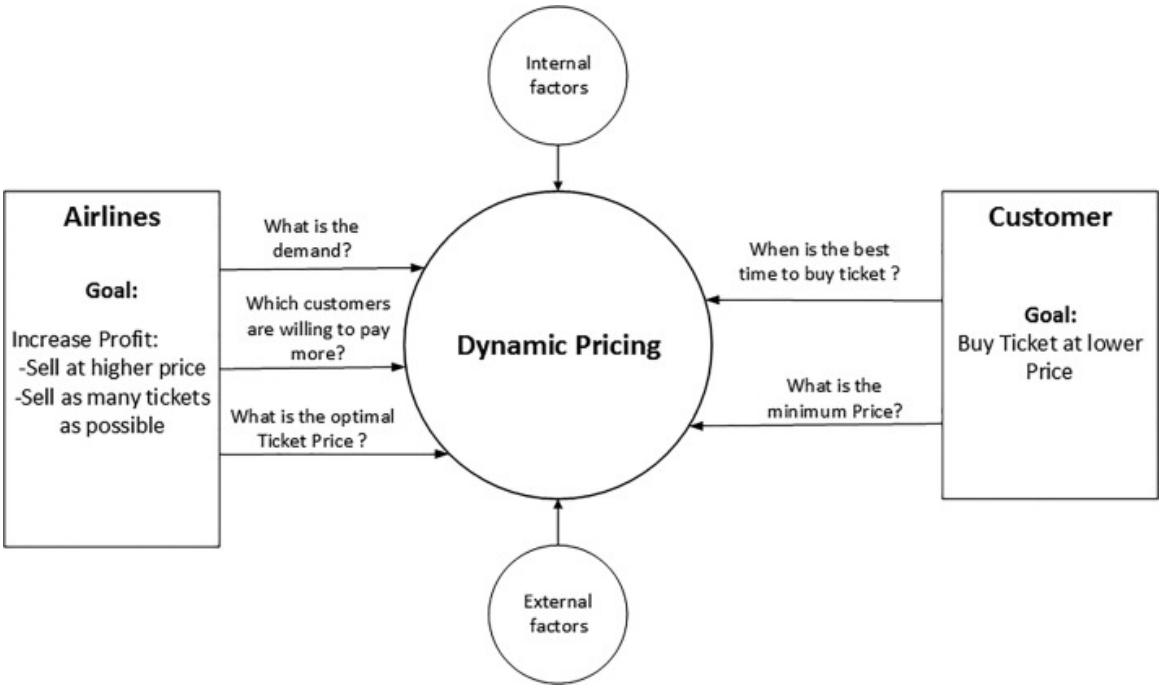
As retrieving price metrics through aggregation techniques and business intelligence tools alone could lead to incorrect conclusions – for example, in cases where have insufficient data points to compute specific price statistics – we used machine learning to forecast prices. This provides an elegant way to interpolate missing data and predict coherent prices. Moreover, we confirmed the forecast decisions using state of the art Explainable AI techniques.

- ## Conceptual Background of the Domain Problem

Flight prices are something unpredictable. It's more than likely that we spent hours on the internet researching flight deals, trying to figure an airfare pricing system that seems completely random every day. Flight price appears to fluctuate without reason and longer flights aren't always more expensive than shorter ones.

But now the question is how to know proper Flight price, for that I have built a Machine learning model which can predict the Flight price. Using various features like **Airline, Source, Destination, Arrival time, Departure time, Stops, Travelling date and the Price for the same travel**. So using all these previously known information and analysing the data I have achieved a good model that has **91.50 accuracy**. So let's understand what all the steps we did to reach this good accuracy.

Nowadays, the number of people using flights has increased significantly. It is difficult for airlines to maintain prices since prices change dynamically due to different 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.



- Review of Literature

Customer retention survives when the companies can fulfill customer expectations and additionally maintain it in long-term relationships to ensure long-term buying decisions [13–15]. The topic of customer retention is argued in business economics commonly within the perspective of relationship marketing, which considers customer relationships as one of the primary concerns with the long-term objective of developing and maintaining them [16–18]. Many previous studies indicated that companies should always manage customer satisfaction to achieve the retention stage. According to [19] "satisfaction is an overall customer attitude towards a service provider". In [20], authors added that satisfaction is an emotional reaction regarding what customers expect and what they receive, including the fulfillment of needs and goals. Customer retention states a desired outcome in the future to satisfaction, so long-term of relationship is demonstrated by satisfaction. Although customer satisfaction does not guarantee repurchase, it still plays a vital role in ensuring customer retention. While many studies on customer retention had long focused on customer satisfaction, additional factors are stated as an influence in customer retention, such as trust and commitment. [21], in "The Commitment-Trust Theory of Relationship Marketing," which is the most influential Relationship Marketing, suggests that the center of successful relationship marketing is the relationship of commitment and trust. They urged the importance of commitment and trust that leads to build a positive correlation between company and customers and encourage efficiency, productivity, and effectiveness. The degree of trust between service provider and customer is significantly influenced by the quality of the service, which results in an effective commitment to the provider, and

Customer retention survives when the companies can fulfill customer expectations and

additionally maintain it in long-term relationships to ensure long-term buying decisions [13–15]. The topic of customer retention is argued in business economics commonly within the perspective of relationship marketing, which considers customer relationships as one of the primary concerns with the long-term objective of developing and maintaining them [16–18]. Many previous studies indicated that companies should always manage customer satisfaction to achieve the retention stage. According to [19] "satisfaction is an overall customer attitude towards a service provider". In [20], authors added that satisfaction is an emotional reaction regarding what customers expect and what they receive, including the fulfillment of needs and goals. Customer retention states a desired outcome in the future to satisfaction, so long-term of relationship is demonstrated by satisfaction. Although customer satisfaction does not guarantee repurchase, it still plays a vital role in ensuring customer retention. While many studies on customer retention had long focused on customer satisfaction, additional factors are stated as an influence in customer retention, such as trust and commitment. [21], in "The Commitment-Trust Theory of Relationship Marketing," which is the most influential Relationship Marketing, suggests that the center of successful relationship marketing is the relationship of commitment and trust. They urged the importance of commitment and trust that leads to build a positive correlation between company and customers and encourage efficiency, productivity, and effectiveness. The degree of trust between service provider and customer is significantly influenced by the quality of the service, which results in an effective commitment to the provider, and

People and real estate agencies buy or sell houses, people buy to live in or as an investment and the agencies buy to run a business. Either way, we believe everyone should get exactly what they pay for. over-valuation/under-valuation in housing markets has always been an issue and there is a lack of proper detection measures. Broad measures, like house/Real-estate price-to-rent ratios, give a primary pass. However, to decide about this issue an in-depth analysis and judgment are necessary. Here's where machine learning comes in, by training an ML model with hundreds and thousands of data a solution can be developed which will be powerful enough to predict prices accurately and can cater to everyone's needs.Real Estate has become more than a necessity in this 21st century, it represents something much more nowadays. Not only for people looking into buying Real Estate but also the companies that sell these Estates. Real Estate Property is not only the basic need of a man but today it also represents the richness and prestige of a person. Investment in real estate generally seems to be profitable because their property values do not decline rapidly. Changes in the real estate price can affect various household investors, bankers, policymakers, and many. Investment in the real estate sector seems to be an attractive choice for investments. Thus, predicting the real estate value is an important economic index.

An attempt has been made in this article to review the available literature in the

area of microfinance. Approaches to microfinance, issues related to measuring social
impact versus profitability of MFIs, issue of sustainability, variables impacting sustainability, affect of regulations of profitability and impact assessment of MFIs have been summarized in the above article. We hope that the above review of
literature will provide researchers a platform for further research and help the
An attempt has been made in this article to review the available literature in the
area of microfinance. Approaches to microfinance, issues related to measuring social
impact versus profitability of MFIs, issue of sustainability, variables impacting sustainability, affect of regulations of profitability and impact assessment of MFIs have been summarized in the above article. We hope that the above review of
literature will provide researchers a platform for further research and help the
An attempt has been made in this article to review the available literature in the
area of microfinance. Approaches to microfinance, issues related to measuring social
impact versus profitability of MFIs, issue of sustainability, variables impacting sustainability, affect of regulations of profitability and impact assessment of MFIs have been summarized in the above article. We hope that the above review of
literature will provide researchers a platform for further research and help the
An attempt has been made in this article to review the available literature in the
area of microfinance. Approaches to microfinance, issues related to measuring social
impact versus profitability of MFIs, issue of sustainability, variables impacting sustainability, affect of regulations of profitability and impact assessment of MFIs have been summarized in the above article. We hope that the above review of
literature will provide researchers a platform for further research and help the
It is hard for the client to buy an air ticket at the most reduced cost. For this few procedures are explored to determine time and date to grab air tickets with minimum fare rate. The majority of these systems are utilizing the modern computerized system known as Machine Learning. The model guesses airfare well in advance from the known information. This framework is proposed to change various added value arrangements into included added value arrangement heading which can support to solo gathering estimation.

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

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)

So, we have to work on a project where we collect data of flight fares with other features and work to make a model to predict fares of flights.

1.4 Motivation for the Problem Undertaken

Flight Price Prediction project help tourists to find the right flight price based on their needs and also it gives various options and flexibility for travelling. Different features (airline, source, destination, departure and arrival timeings, Journey date etc.) helps to understand the flight price variations. Using it airlines also get benefits and required passengers. Also they will get benefit in scheduling also.

2.Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem

As a first step I have scrapped the required data from makemytrip website. I have fetched data for different source and destinations and saved it to csv format.

In this perticular problem I have Price as my target column and it was a continuous column. So clearly it is a regression problem and I have to use all regression algorithms while building the model. There was no null values in the dataset. Since we have scrapped the data from yatra website the raw data was not in the format, so we have use feature engineering to extract the required feature format. To get better insight on the features I have used ploting like distribution plot, bar plot, strip plot and count plot. With these ploting I was able

to understand the relation between the features in better manner. I did not found any skewness or outliers in the dataset. I have used all the regression algorithms while building model then tunned the best model and saved the best model. At last I have predicted the Price using saved model.

2.2 Data Sources and their formats

The data was collected from makemytrip.com website in csv format. The data was scrapped using selenium. After scrapping required features the dataset is saved as csv file.

Also, my dataset was having 2260 rows and 10 columns including target. In this perticular datasets I have object type of data which has been changed as per our analysis about the dataset. The information about features is as follows.

Features Information:

- Airline: The name of the airline.
- Journey_date: The date of the journey
- Source: The source from which the service begins.
- Destination: The destination where the service ends.
- Route: The route taken by the flight to reach the destination.
- DepartureTime: The time when the journey starts from the source.
- Arrival Time: Time of arrival at the destination.
- Stops: Total stops between the source and destination.
- Price: The price of the ticket

- ## Data Preprocessing Done

  - As a first step I have scrapped the required data using selenium from makemytrip website.

- And I have imported required libraries and I have imported the dataset which was in csv format.
- Then I did all the statistical analysis like checking shape, nunique, value counts, info etc…..
- While checking for null values I found there was a row full of null values in the dataset and I dropped that row as it will not help our analysis.
- I have also droped Unnamed:0 column as I found it was the index column of csv file.
- Next as a part of feature extraction I converted the data types of datetime columns and I have extracted usefull information from the raw dataset. Thinking that this data will help us more than raw data.

2.4 Data Inputs- Logic- Output Relationships

- Since I had numerical columns I have plotted dist plot to see the distribution of skewness in each column data.
- I have used bar plot for each pair of categorical features that shows the relation between target and independent features.
- I have used strip plot to see the relation between numerical columns and target column.
- I can notice there is a good relationship between maximum columns and target.

• Hardware and Software Requirements and Tools Used

While taking up the project we should be familiar with the Hardware and software required for the successful completion of the project. Here we need the following hardware and software.

Hardware required: -

1. Processor — core i5 and above

2. RAM — 8 GB or above

3. SSD — 250GB or above

Software/s required: -

1.Anaconda

Libraries required :-

```
#importing required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt
import warnings
warnings.filterwarnings('ignore')
```

To run the program and to build the model we need some basic libraries as follows:

- **import pandas as pd**: **pandas** is a popular Python-based data analysis toolkit which can be imported using import pandas as pd. It presents a diverse range of utilities, ranging from parsing multiple file formats to converting an entire data table into a numpy matrix array. This makes pandas a trusted ally in data science and machine learning.
- **import numpy as np**: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
- **import seaborn as sns:** Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python.

Visualization is the central part of Seaborn which helps in exploration and understanding of data.

- **Import matplotlib.pyplot as plt:** matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.
- from sklearn.preprocessing import LabelEncoder
- from sklearn.preprocessing import StandardScaler
- from sklearn.ensemble import RandomForestRegressor
- from sklearn.tree import DecisionTreeRegressor
- from sklearn.ensemble import GradientBoostingRegressor
- from sklearn.ensemble import ExtraTreesRegressor
- from sklearn.neighbors import KNeighborsRegressor as KNN
- from sklearn.ensemble import BaggingRegressor
- from sklearn.metrics import classification_report
- from sklearn.metrics import accuracy_score
- from sklearn.model_selection import cross_val_score

With this sufficient libraries we can go ahead with our model building.

3.Data Analysis and Visualization

3.1 Identification of possible problem-solving approaches (methods)

- Since the data collected was not in the format we have to clean it and bring it to the proper format for our analysis. Since there was no outliers and skewness in the dataset no need to worry about that. We have dropped all the unnecessary columns in the dataset according to our understanding. Use of Pearson's correlation coefficient to check the correlation between dependent and independent features. Also I have used Standardisation to scale the data. After scaling we have to check multicolinearity using VIF. Then followed by model building with all Regression algorithms.

• Testing of Identified Approaches (Algorithms)

Since Price was my target and it was a continuous column with improper format which has to be changed to continuous float datatype column, so this perticular problem was Regression problem. And I have used all Regression algorithms to build my model. By looking into the r2 score and error values I found ExtraTreesRegressor as a best model with highest r2_score and least error values. Also to get the best model we have to run through multiple. Below are the list of Regression algorithms I have used in my project.

- RandomForestRegressor
- ExtraTreesRegressor
- GradientBoostingRegressor
- DecisionTreeRegressor
- KNN
- BaggingRegressor

• Key Metrics for success in solving problem under consideration

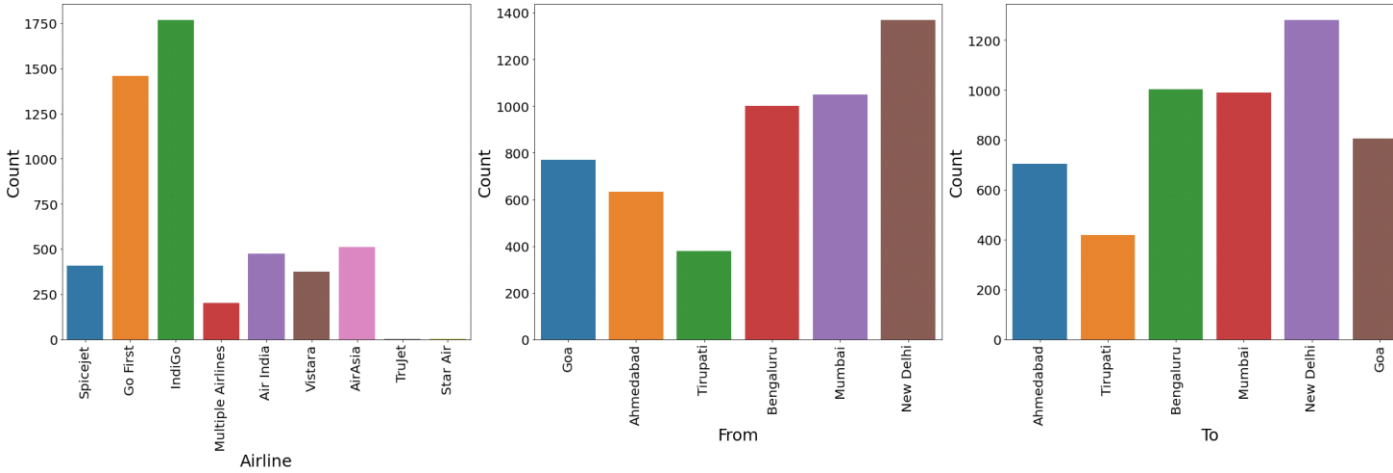
I have used the following metrics for evaluation:

- I have used mean absolute error which gives magnitude of difference between the prediction of an observation and the true value of that observation.

- I have used root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions.
- I have used r2 score which tells us how accurate our model is.
-

3.4 Visualizations

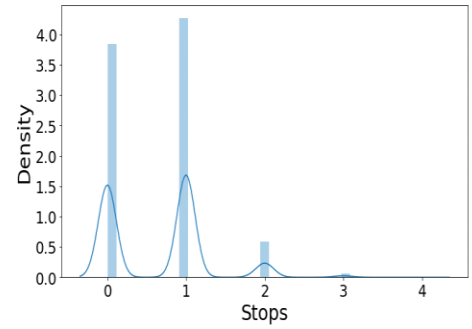
I have used bar plots to see the relation of categorical feature with target and I have used 2 types of plots for numerical columns one is disp plot for univariate and strip plot for bivariate analysis.

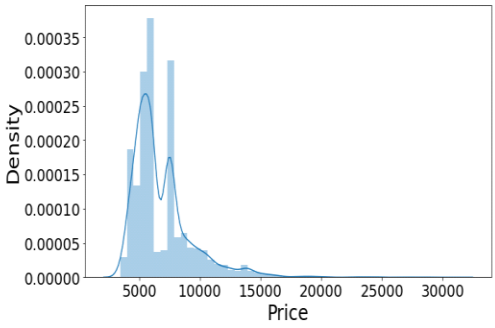


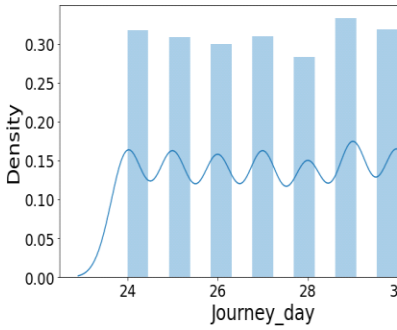
- **<u>Univariate Analysis for Categorical columns:</u>**

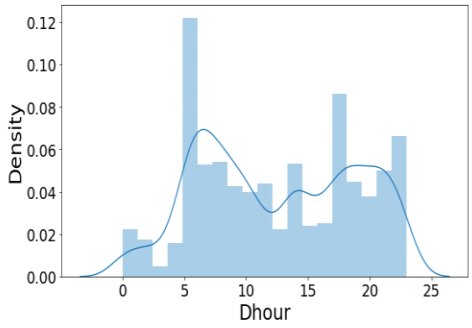
<u>Observations:</u>

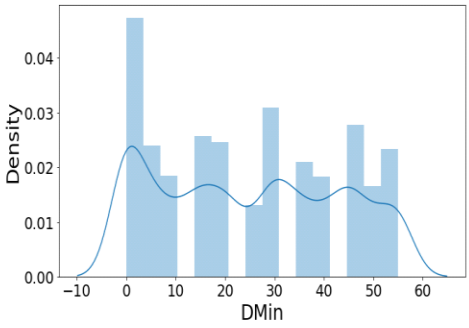
- Indigo has maximum count which means most of the passengers preferred Indigo for there travelling.
- New Delhi has maximum count for source which means maximum passengers are choosing New Delhi as there source.
- New Delhi has maximum count for Destination which means maximum passengers are choosing New Delhi as there Destination.

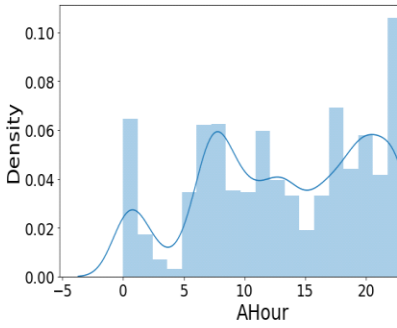


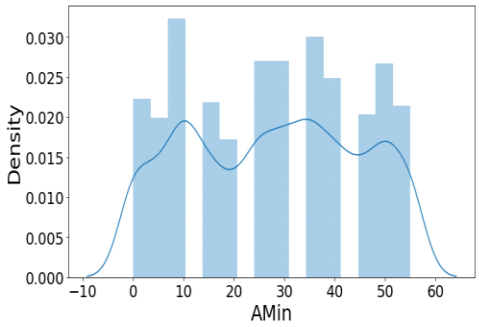












- **Univariate analysis for Numerical column:**

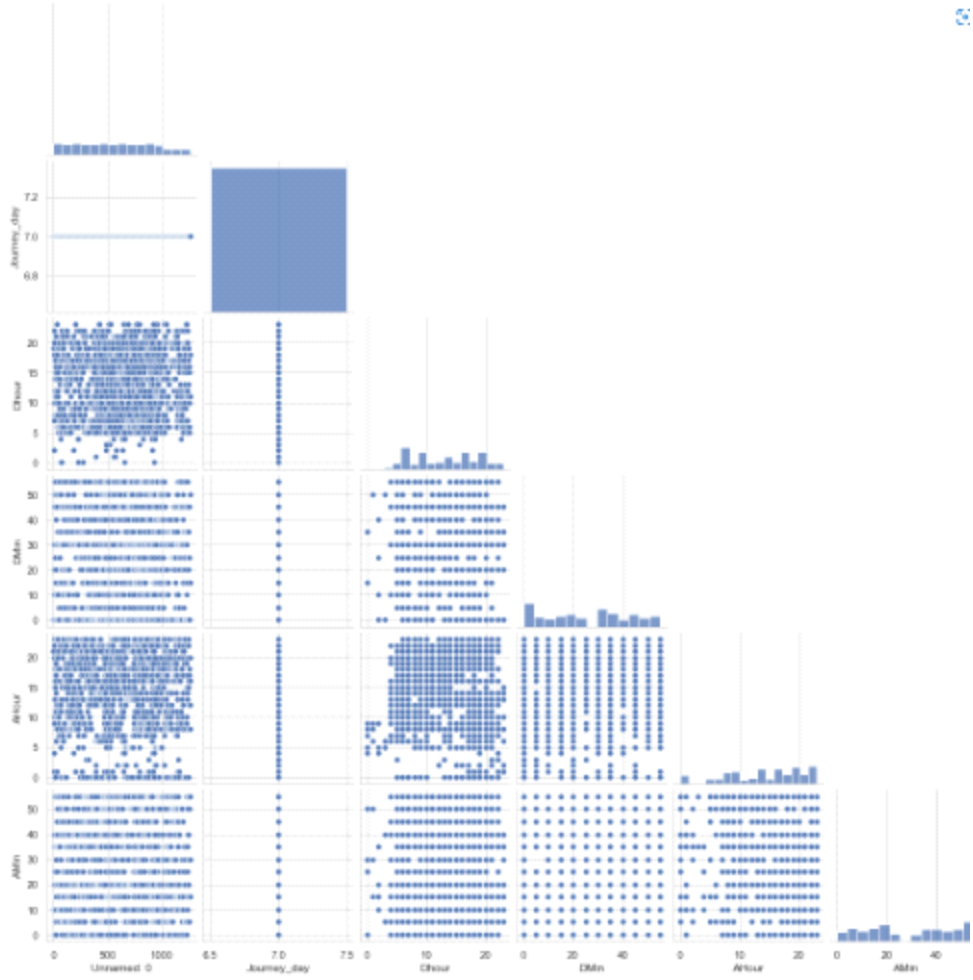
Observations:

- There is no skewness in any of the numerical columns.

- **Multivariate Analysis:**

```
sns.pairplot(df, corner=True)
```

```
<seaborn.axisgrid.PairGrid at 0x24aaa8c4850>
```



Above are the pair plots of each pair of features.

5. Run and Evaluate selected models

1. Model Building:

- **RandomForestRegressor:**

RandomForestRegressor:

```
RFR=RandomForestRegressor()
RFR.fit(X_train,y_train)
pred=RFR.predict(X_test)
R2_score = r2_score(y_test,pred)*100
print('R2_score:',R2_score)
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print('root_mean_squared_error:',np.sqrt(metrics.mean_squared_error(y_test,pred)))
```

```
R2_score: 91.28932641391731
mean_squared_error: 1929.2517662241887
mean_absolute_error: 21.825442477876106
root_mean_squared_error: 43.923248584595704
```

- RandomForestRegressor has given me 91.28% r2_score, but still we have to look into multiple models.

- **ExtraTreesRegressor:**

ExtraTreeRegressor:

```
ETR=ExtraTreesRegressor()
ETR.fit(X_train,y_train)
pred=ETR.predict(X_test)
print('R2_score:',r2_score(y_test,pred))
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print('root_mean_squared_error:',np.sqrt(metrics.mean_squared_error(y_test,pred)))
```

```
R2_score: 0.9150272905789147
mean_squared_error: 1881.9870600294985
mean_absolute_error: 17.782581120943952
root_mean_squared_error: 43.38187478693721
```

- ExtraTreesRegressor is giving me 91% r2_score.

- **GradientBoostingRegressor:**

Gradient Boosting Regressor:

```
GBR=GradientBoostingRegressor()
GBR.fit(X_train,y_train)
pred=GBR.predict(X_test)
print('R2_score:',r2_score(y_test,pred))
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print('root_mean_squared_error:',np.sqrt(metrics.mean_squared_error(y_test,pred)))
```

```
R2_score: 0.8099680487753101
mean_squared_error: 4208.85335578435
mean_absolute_error: 43.52111282761848
root_mean_squared_error: 64.87567614895701
```

- GradientBoostingRegressor is giving me 80.99% r2_score.

- **DecisionTreeRegressor:**

DecisionTreeRegressor:

```
DTR=DecisionTreeRegressor()
DTR.fit(X_train,y_train)
pred=DTR.predict(X_test)
print('R2_score:',r2_score(y_test,pred))
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print('root_mean_squared_error:',np.sqrt(metrics.mean_squared_error(y_test,pred)))
```

```
R2_score: 0.8498653045197262
mean_squared_error: 3325.2035398230087
mean_absolute_error: 17.073746312684367
root_mean_squared_error: 57.66457786044227
```

- DecisionTreeRegressor is giving me 84.98% r2_score.

- **KNN:**

KNN:

```
knn=KNN()
knn.fit(X_train,y_train)
pred=knn.predict(X_test)
print('R2_score:',r2_score(y_test,pred))
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print('root_mean_squared_error:',np.sqrt(metrics.mean_squared_error(y_test,pred)))
```

```
R2_score: 0.5740266111957524
mean_squared_error: 9434.51622418879
mean_absolute_error: 64.21297935103244
root_mean_squared_error: 97.13143787769637
```

- KNN is giving me 57.40% r2_score.

- **BaggingRegressor:**

Bagging Regressor:

```
BG=BaggingRegressor()
BG.fit(X_train,y_train)
pred=BG.predict(X_test)
print('R2_score:',r2_score(y_test,pred))
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print('root_mean_squared_error:',np.sqrt(metrics.mean_squared_error(y_test,pred)))
```

```
R2_score: 0.8963079966637146
mean_squared_error: 2296.584513274336
mean_absolute_error: 23.15457227138643
root_mean_squared_error: 47.92269309287966
```

- BaggingRegressor is giving me 89.63% r2_score.

Cross Validation score of different model:

```
root_mean_squared_error: 47.922693092879 66
```

```
: from sklearn.model_selection import cross_val_score

scr = cross_val_score(DTR, x, y, cv=5)
print("Cross Validation score of DecisionTreeRegressor model is:", scr.mean())

scr = cross_val_score(RFR, x, y, cv=5)
print("Cross Validation score of RandomForestRegressor model is:", scr.mean())

scr = cross_val_score(ETR, x, y, cv=5)
print("Cross Validation score of ExtraTreesRegressor model is:", scr.mean())
```

```
Cross Validation score of DecisionTreeRegressor model is: 0.4525408340668454
Cross Validation score of RandomForestRegressor model is: 0.7077323343164922
Cross Validation score of ExtraTreesRegressor model is: 0.8201903135586122
```

```
: scr = cross_val_score(GBR, x, y, cv=5)
print("Cross Validation score of GradientBoostingRegressor model is:", scr.mean())

scr = cross_val_score(BG, x, y, cv=5)
print("Cross Validation score of KNeighborsRegressor model is:", scr.mean())
```

```
Cross Validation score of GradientBoostingRegressor model is: 0.6423831958502506
Cross Validation score of KNeighborsRegressor model is: 0.6825476426134864
```

```
By looking into the model r2_score and error i found ExtraTreesRegressor as the best model with highest r2_score and least
errors.
```

- **By looking into the model r2_score and error I found ExtraTreesRegressor as the best model with highest r2_score and least errors.**

3. Hyper Parameter Tunning:

Hyper Parameter Tuning:

```
#importing necessary libraries
from sklearn.model_selection import GridSearchCV
```

```
parameter = {'max_features':['auto','sqrt','log2'],
             'min_samples_split':[1,2,3,4],
             'n_estimators':[20,40,60,80,100],
             'min_samples_leaf':[1,2,3,4,5],
              'n_jobs':[-2,-1,1,2]}
```

```
from sklearn.model_selection import RandomizedSearchCV
```

```
RCV = RandomizedSearchCV(ExtraTreesRegressor(), parameter, cv=5, n_iter=10)
RCV.fit(X_train, y_train)
```

```
RandomizedSearchCV(cv=5, estimator=ExtraTreesRegressor(),
                   param_distributions={'max_features': ['auto', 'sqrt',
                                                         'log2'],
                                        'min_samples_leaf': [1, 2, 3, 4, 5],
                                        'min_samples_split': [1, 2, 3, 4],
                                        'n_estimators': [20, 40, 60, 80, 100],
                                        'n_jobs': [-2, -1, 1, 2]})
```

```
RCV.best_params_
```

```
{'n_jobs': 1,
 'n_estimators': 20,
 'min_samples_split': 3,
 'min_samples_leaf': 2,
 'max_features': 'auto'}
```

```
Best_mod=ExtraTreesRegressor(max_features='auto',min_samples_leaf=2,min_samples_split=2,n_estimators=80,n_jobs=1)
Best_mod.fit(X_train,y_train)
pred=Best_mod.predict(X_test)
print('R2_Score:',r2_score(y_test,pred)*100)
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print("RMSE value:",np.sqrt(metrics.mean_squared_error(y_test, pred)))
```

```
R2_Score: 88.3525800755326
mean_squared_error: 2579.6863169268413
mean_absolute_error: 25.63804966343268
RMSE value: 50.79061248820338
```

- **I have choosed all parameters of ExtraTreesRegressor, after tunning the model with best parameters I have incresed my model accuracy .**

- **Saving the model and Predictions:**

- I have saved my best model using .pkl as follows.

Saving New Model:

```
# Saving the model using .pkl
import joblib
joblib.dump(Best_mod,"Flight_Price.pkl")
```

```
['Flight_Price.pkl']
```

Predicting Flight Price for test dataset using Saved model of trained dataset:

```
# Loading the saved model
model=joblib.load("Flight_Price.pkl")

#Prediction
prediction = model.predict(X_test)
prediction
```

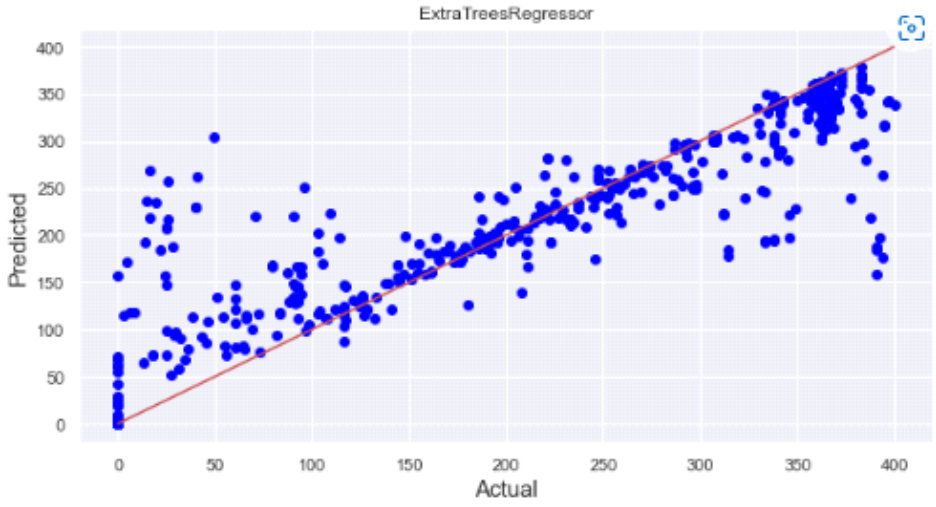
```
pd.DataFrame([model.predict(X_test)[:],y_test[:]],index=["Predicted","Actual"])
```

	0	1	2	3	4	5	6	7	8	9	...	668	669	670	671	672	673	67
Predicted	0.0	110.351042	341.685417	0.0	342.030208	362.177083	241.824792	0.0	218.69494	0.0	...	0.0	346.873333	0.0	0.0	350.086458	182.505208	0.
Actual	0.0	117.000000	396.000000	0.0	357.000000	362.000000	252.000000	0.0	217.00000	0.0	...	0.0	358.000000	0.0	0.0	356.000000	181.000000	0.

2 rows × 678 columns

- Now loading my saved model and predicting the price values.

```
plt.figure(figsize=(10,5))
plt.scatter(y_test, prediction, c='blue')
p1 = max(max(prediction), max(y_test))
p2 = min(min(prediction), min(y_test))
plt.plot([p1, p2], [p1, p2], 'r')
plt.xlabel('Actual', fontsize=15)
plt.ylabel('Predicted', fontsize=15)
plt.title("ExtraTreesRegressor")
plt.show()
```



Plotting Actual vs Predicted,To get better insight.Red line is the actual line and blue dots are the predicted values.

- Plotting Actual vs Predicted, To get better insight. Red line is the actual line and blue dots are the predicted values.

- ## Interpretation of the Results

  - The dataset was scrapped from makemytrip website.
  - The dataset was very challenging to handle it had 10 features with 2260samples.
  - Firstly, the datasets was having a complete row as nan values, so I have dropped that row.
  - And there was huge number of unnecessary entries in all the features so I have used feature extraction to get the required format of variables.

- And proper ploting for proper type of features will help us to get better insight on the data. I found both numerical columns and categorical columns in the dataset so I have choosen strip plot and bar plot to see the relation between target and features.
- I did not found any outliers or skewness in the dataset.
- Then scaling dataset has a good impact like it will help the model not to get biased. Since we did not have outliers and skewness in the dataset so we have to choose Standardisation.
- We have to use multiple models while building model using dataset as to get the best model out of it.
- And we have to use multiple metrics like mse, mae, rmse and r2_score which will help us to decide the best model.
- I found ExtraTreesRegressor as the best model with 91.50% r2_score. Also I have improved the accuracy of the best model by running hyper parameter tunning.
- At last I have predicted the used flight price using saved model. It was good!! that I was able to get the predictions near to actual values.

4.CONCLUSION

4.1 Key Findings and Conclusions of the Study

In this project report, we have used machine learning algorithms to predict the flight prices. We have mentioned the step by step procedure to analyze the dataset and finding the correlation between the features. Thus we can select the features which are correlated to each other and are independent in nature. These feature set were then given as an input to seven algorithms and a hyper parameter tunning was done to the best model and the accuracy has been improved. Hence we calculated the performance of each model using different performance metrics and compared them based on those metrics. Then we have

also saved the best model and predicted the flight price. It was good the the predicted and actual values were almost same.

4.2 Learning Outcomes of the Study in respect of Data Science

I found that the dataset was quite interesting to handle as it contains all types of data in it and it was self scrapped from makemytrip website using selenium. Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed. New analytical techniques of machine learning can be used in flight price research. The power of visualization has helped us in understanding the data by graphical representation it has made me to understand what data is trying to say. Data cleaning is one of the most important steps to remove unrealistic values and null values. This study is an exploratory attempt to use seven machine learning algorithms in estimating flight price prediction, and then compare their results.

To conclude, the application of machine learning in predicting flight price is still at an early stage. We hope this study has moved a small step ahead in providing some methodological and empirical contributions to crediting online platforms, and presenting an alternative approach to the valuation of flight price. Future direction of research may consider incorporating additional used flight data from a larger economical background with more features.

- ## Limitations of this work and Scope for Future Work

- First drawback is scrapping the data as it is a fluctuating process.
- Followed by raw data which is not in format to analyse.
- Also, we have tried best to deal with improper format data and null values. So it looks quite good that we have achieved a accuracy even after dealing all these drawbacks.

- Also, this study will not cover all Regression algorithms instead, it is focused on the chosen algorithm, starting from the basic ensembling techniques to the advanced ones.