FLIGHT PRICE PREDICTION PROJECT

Submitted by:

SHANTANU

ACKNOWLEDGMENT

I want to express my great appreciation to my mentor for her valuable and constructive suggestions during the planning and development of this research work. Her willingness to give his time so generously has been very much appreciated.I would also like to thank the staff of the Data Trained for helping me with the problems I faced during the research work.Articles from the "Medium" platform werebeneficial during the whole process. It helped me clear my concepts.

INTRODUCTION

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

Objective:

This project contains two-phase-

Data Collection Phase

We scrapped more than 1500 rows of data. In this we scrapped s the data of flights from different websites (yatra.com, skyscanner.com, official websites of airlines, etc).. Generally, these columns are airline name, date of journey, source, destination, route, departure time, arrival time, duration, total stops and the target variable price.

Model Building Phase

After collecting the data, we built a machine learning model. Before model building, we did data pre-processing steps.

Followed the complete life cycle of data science. Include all the steps like.

1. Data Cleaning
2. Exploratory Data Analysis
3. Data Pre-processing
4. Model Building
5. Model Evaluation
6. Selecting the best model

Analytical Problem Framing

In the whole research process various mathematical, statistical and analytics modelling has been done. There has been reduction of the columns because few of them was not necessary for the problem solving like Id. And few of them was removed due to very less correlation with dependent variable.To fix the outliers we used z score method. After this also there was a lot of skewness in dataset so power transform has been used. To check the accuracyr2 score was used also for cross validation cross_val _score is used.

DATA/ DATA PREPROCESSING:

- The dataset contains 1792 rows and 8 columns
- Fare is our dependent variable.
- We created new features from old ones.
- All columns were object data types we converted necessary ones into int and float.
- There are no null values in the dataset.
- Trimmed few columns

Hardware and Software Requirements and Tools Used

- HP 5- i5 8th generation, 8gb ram, NVidia mx130 integrated graphic,
- JuypterNotebook/Google chrome
- Libraries and packeges used:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_selection import VarianceThreshold
from sklearn.feature_selection import mutual_info_regression
from sklearn.feature_selection import SelectPercentile
from sklearn.preprocessing import StandardScaler
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.preprocessing import power_transform
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import ExtraTreesRegressor
from sklearn.svm import SVR
from sklearn.model_selection import cross_val_score
from sklearn.ensemble import BaggingRegressor
from sklearn.ensemble import AdaBoostRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.model_selection import GridSearchCV
```

the library used here is sklearn,numpy,matplotlib,pandas and seaborn. The matpotplotlib and seaborn library has been used to make charts to visualize and understand the problem, correlation, outliers and many other things, the pandas and NumPy library is used to handle dataset and perform various tasks. The seaborn library is used for model building and cross validation of the models.

Model/s Development and Evaluation

The approach to solve this problem was to get the domain knowledge to understand the data better. Which values can be the part of the data and which is not? After exploring the data, it is found that though the data hasno missing value. It has extreme outliers and unrealistic value. We used Z-Score method to remove outliers. There was some skewness in the data, power transform method has been used so it dealt skewness. To check the accuracy, mean square error, mean absolute error, r2 score was used also for cross validation cross_val _score is used

Algoritham used for Traning and testing:

```
from sklearn.model_selection import train_test_split

from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

from sklearn.linear_model import LinearRegression

from sklearn.tree import DecisionTreeRegressor

from sklearn.neighbors import KNeighborsRegressor

from sklearn.ensemble import RandomForestRegressor

from sklearn.ensemble import ExtraTreesRegressor

from sklearn.svm import SVR

from sklearn.model_selection import cross_val_score

from sklearn.ensemble import BaggingRegressor

from sklearn.ensemble import AdaBoostRegressor

from sklearn.ensemble import GradientBoostingRegressor

from sklearn.model_selection import GridSearchCV
```

Performance of the model:

```
#spliting train test data
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.30, random_state=56)
x_train.shape,y_train.shape,x_test.shape,y_test.shape
```

```
((817, 32), (817,), (351, 32), (351,))
```

By the train test split method, 70 percent of the data has been taken for the model building while 30 percent of the data has been reserved for checking the model's performance.

```
#creating function
def model(name):
    model=name()
    model.fit(x_train,y_train)
    predict=model.predict(x_test)
    print("""mean squared error is:
    """,mean_squared_error(y_test, predict))

    print("The mean absolute error is: ", mean_absolute_error(y_test,predict))


    print("""r2 score is:

    """,r2_score(y_test,predict))

    print("cross_val_score", cross_val_score(model,x,y,cv=5).mean())
```

The code above has been used to speed up the model training and its evaluation process. Here the function name model ha been created which take the name of model as argument.

```
#LineraRegression
mQdel( LineanRegression)

mean squared e nor is
     31é60é9.2223878595
The mean absolute enron is: 1433.866794lll)B72
r2 score is'

     6. 25 1934025 13999397
c ros s_val_s core 0. 016G5777862732844
```

```
#decisiontreeregressor
mOdel(DecisionTreeRegressor)

mean squared e nor is
     2363036.64Plé42B84
The mean absolute error is: 839.0975d48S36356
r2 scone is:

     0.44691726018038
cross_val_score -8.10927074293391207
```

```
#randomforestregressor
ma de l ( Rand omF o r esl R egr es s on)

nean squa red e°ror is
     1783233 . .^960.^773 G
The mean abs a1 ul e e rror i s : 82G . 3791080769895
r2 s co re is :

     0.5826235105°21618
cross_val_score 0.18439391680321432
```

```
#extratreesreressor
model(ExtraTreesRegressor)

mean squared error is:
     1836607. 35 19434037
The mean abs a1ut e e rror  i s :  735. 0998276676108
r2 s co re is :

     0.5668543757490115
cross_val_score 0.267G53989G3^2B817
```

```
model(KNeighborzFegressor)

    2B35y56.&e57507fi8
The mean absolute error is:  991.4515580736545

     0.5235191675749986
<rozs val score B.1064265Bb750B12^7
```

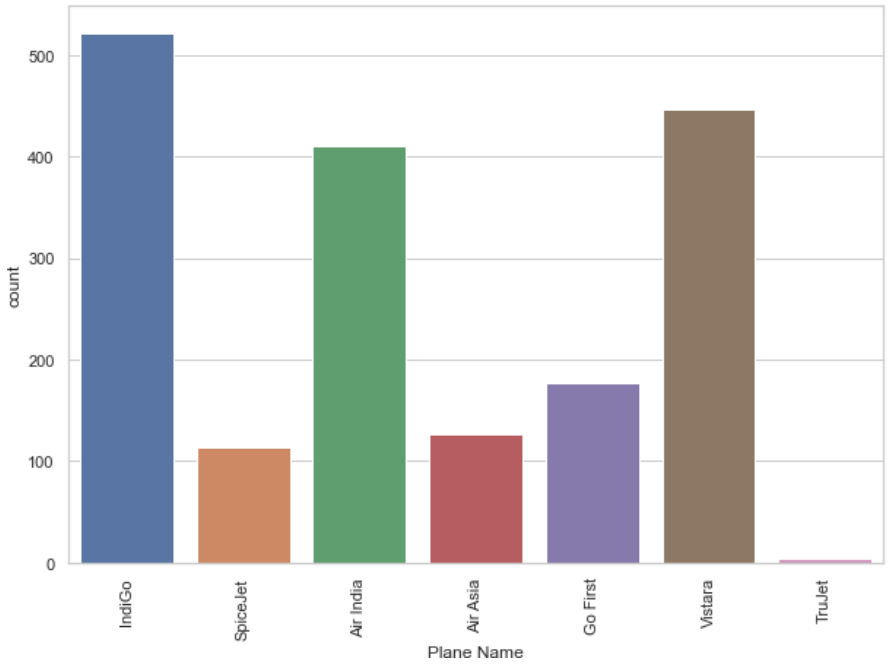
```
from zKlearn.ensemble import BaggingRegressOr
from sklearn.ensemble import AdaBoostRegressor
from sklearn.ensemble import GradientBOostingRegressor

model(BaggingRegressor)

mean squared error is:

The mean absolute error is: 855.^6B267BC54662
r2 score is

     0.5548223096347795
bros; val score 0.13421416Ba833171
```

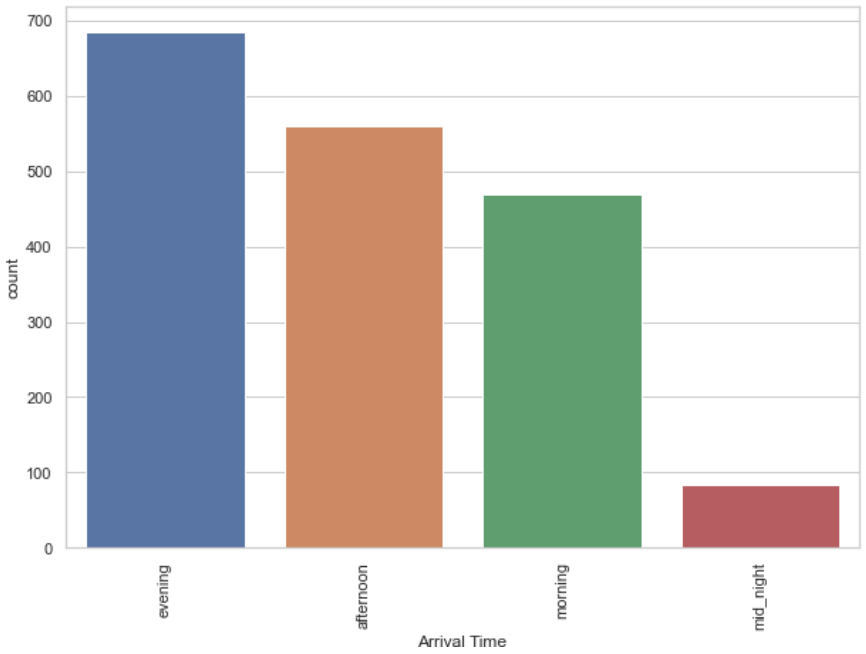
```
model(AduBOOztRegnussOn)

The mean absolute error is: 1371.042d0BS522315
r2 score is:

<ross_val_score B.1355557127445a
```

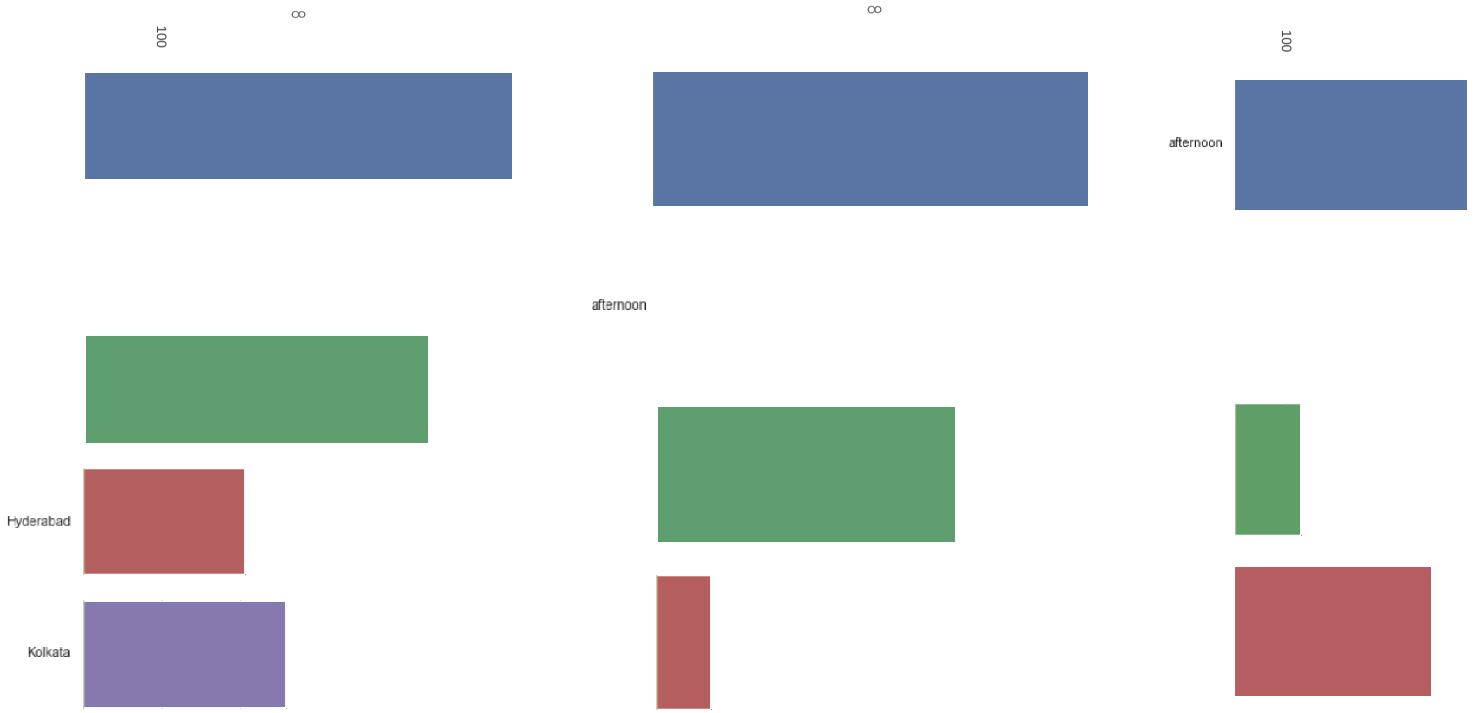
```
aC d' e tSoo<' . qs <q' °> so
node1 (G ad Tgp IB OOH I ggK go â OK )

mean squared error is:
     1808658.1876762426
The mean absolute error is:  974.2160123596385

    0.5I667£SZ63C1£36C
bros vol score Bñ2778651229032538
```

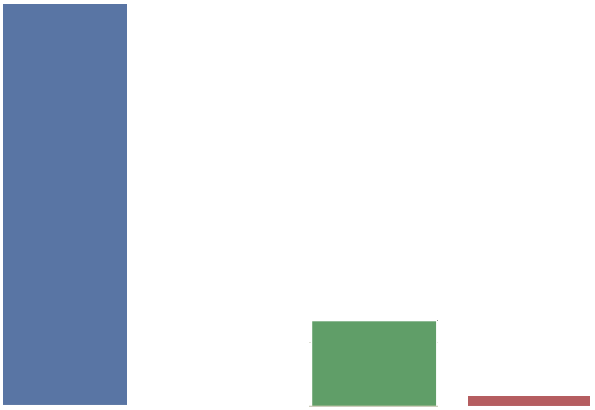
```
#Rondoing-oi°esI:regresson is I:he besI aodeL as Ohe RNSLE 'is inoxiuin
#se I:I:'i ng parameters fi-or hypei°porainetei° Sunn'i ng
parameter={
            "cr it er i on" : [ "mse™ "mae" § ,
            ' max_features ' : [ ' auto '    ' sqrt'   "1og2" § ,
            ' min_s amples_sp1it ' : [ 2 5, GB, 15§ ,
             ' min_s amples_leaT ' : [ 1 2, 5, 16a }


#us Eng Grp dSeai°chCV fi-or Hypes° pai°aine I:er I:unn'ing
-from sklea rn.mode1_se1ection import GrTdSea rchCV
qcs=G r ids ear chCV( Randomg-or estRegres sor( ) parameter cv=5)


gcv.fit(x_traim,y_traim)

: GrTdSearchCV{cv-S, estTwtor=Randomg- ores tRegressor ( )
                param    rid=( ' crTt erson ' :    'mse"   'mae'§
                                ' nax_f eat ures' : [ ' a uto,'      'sq rt ,'     ' log2 ', ]
                                'mT n_sanp1es_1eaf ' : [1, 2, s, ie],
                                'mTn_sanp1es_sp1Tt ':   2 5 GB, 1s] })


: #chech'i ng bet I: porainetei°s
  gcv . best_par ams_

: {'criterion'  : 'mae',
   'nax_features': 'sqrt',
   'nin_samples_leaf': l,
   'nin_samples_split':  2}


model=RamdomForestRegressor(criterion="mse™,max_features="auto",min_samples_leaf=l,min_samples_split=10)
mode1. fit ( x_t ra in y_t r ain )
pred=mode1. pred Act ( x_t e s t )

print ("" "mean squared error is :
      " "" , mean_squared_error y_test , pred) )

print ("The mean ab s o1ut e error is : " mean_absolute_error(y_test pred ) )


  p r i nt ( " r 2 s co re i s : " , r2_sc ore( y_te st p r ed ) )

mean squared error is:
      1W3T62.  ST 598T91S
The  can  abso1ut e error  is :   8B1. 76751871519B5
rd score is: B . 6B12 243131166367
```

Visualization:

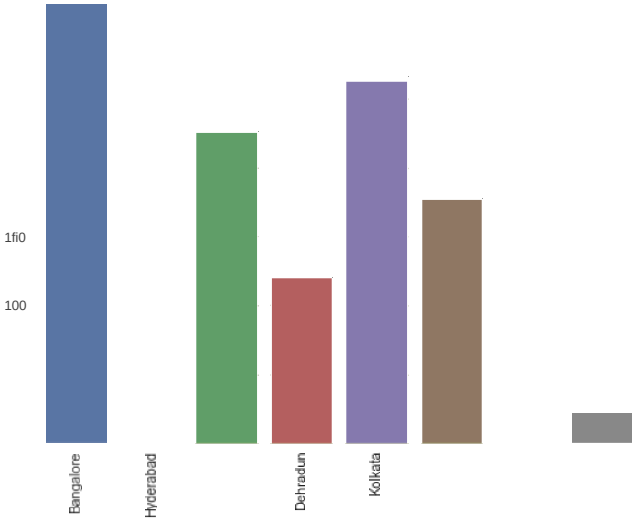


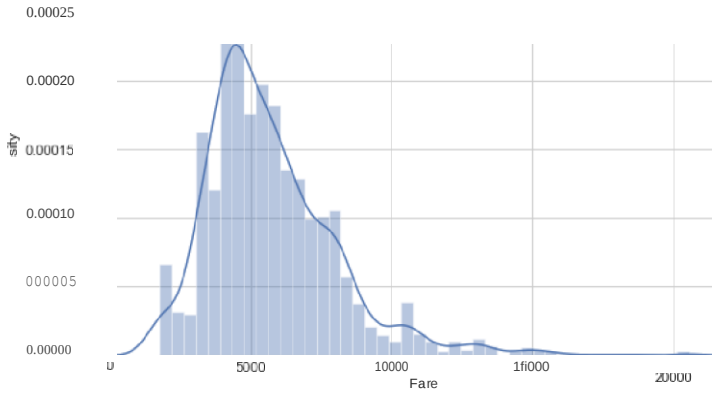


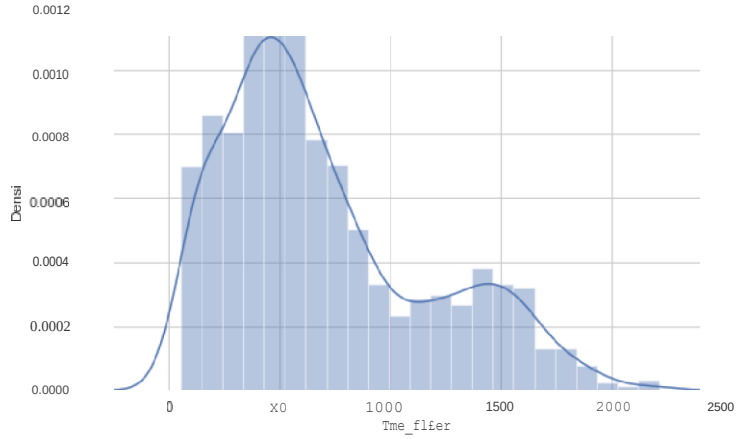

100
afternoon
Hyderabad
Kolkata

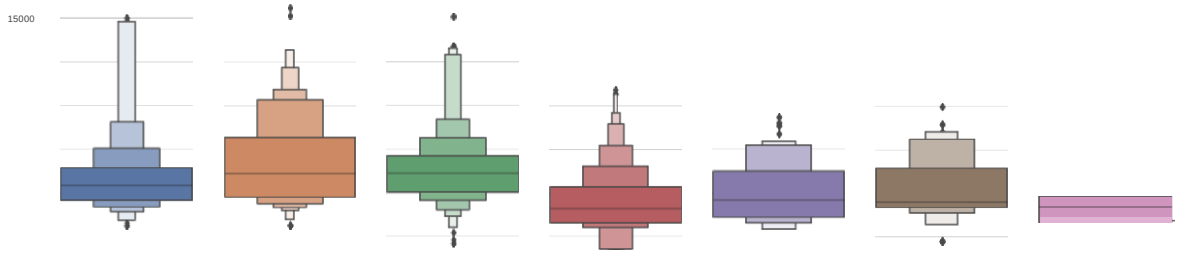


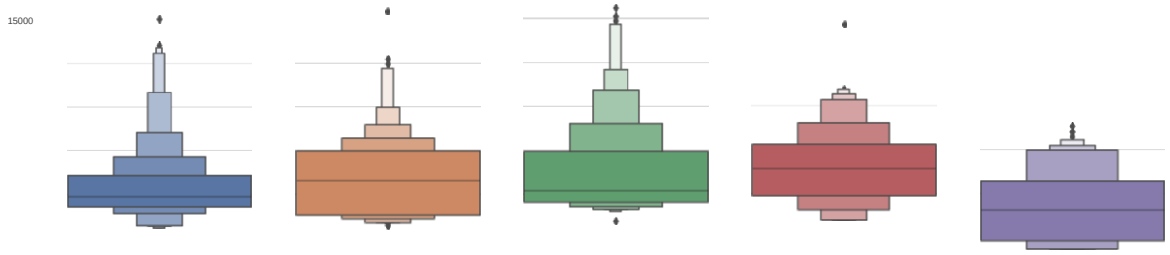
Numbers of stops

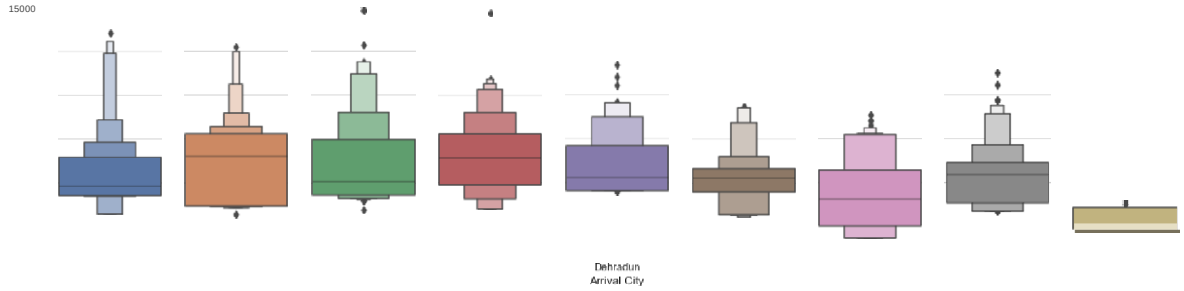


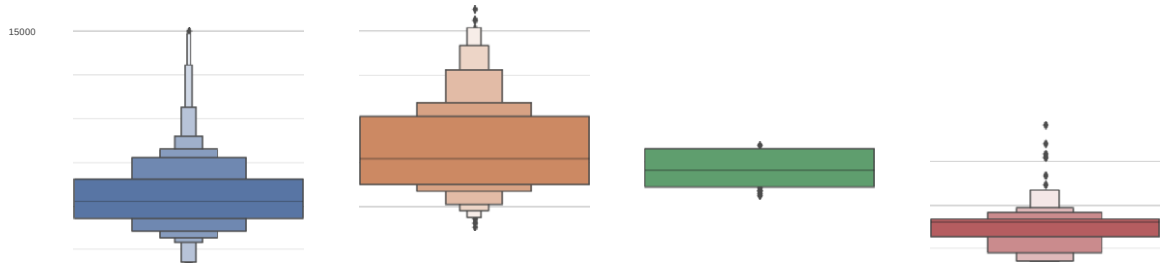

0.00025
0.00020
0.00015
0.00010
000005
0.00000
sity
5000
10000
Fare

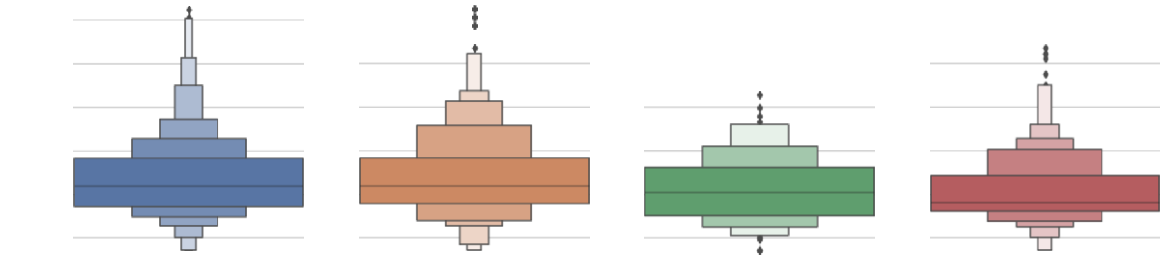

0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Densi
0
X0
1000
1500
2000
2500
Tme_fl£er

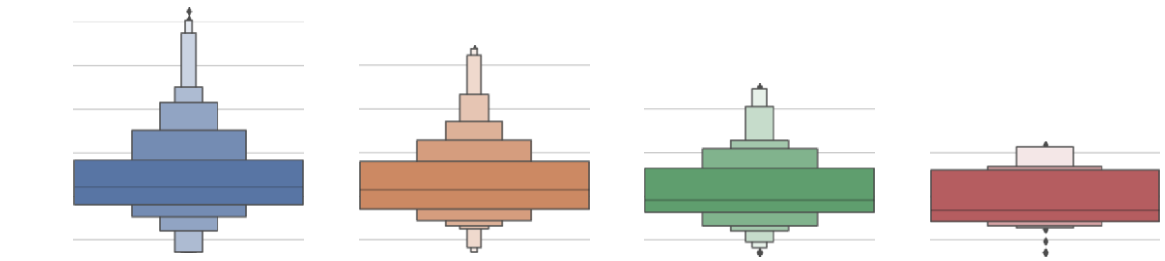

15000

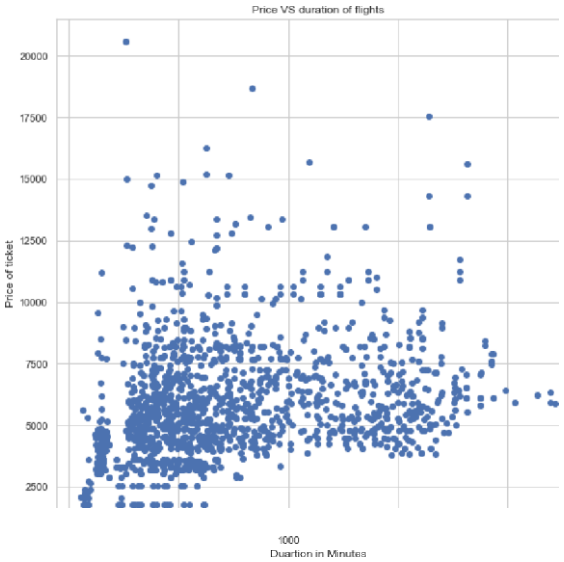

15000

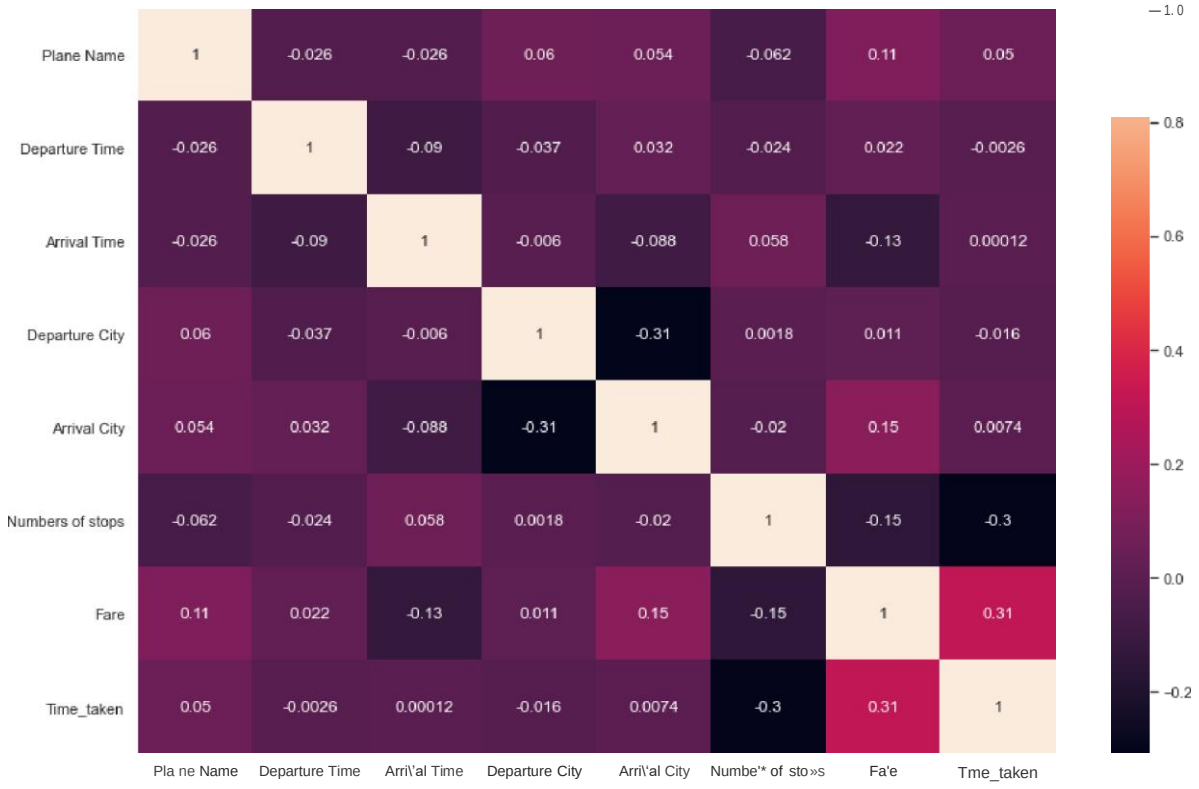

15000
Dehradun
Arrival City


15000






Price VS duration of flights
Price of ticket
20000
17500
15000
12500
10000
7500
5000
2500
1000
Duartion in Minutes


Plane Name
Departure Time
Arrival Time
Departure City
Arrival City
Numbers of stops
Fare
Time_taken
1 -0.026 -0.026 0.06 0.054 -0.062 0.11 0.05
-0.026 1 -0.09 -0.037 0.032 -0.024 0.022 -0.0026
-0.026 -0.09 1 -0.006 -0.088 0.058 -0.13 0.00012
0.06 -0.037 -0.006 1 -0.31 0.0018 0.011 -0.016
0.054 0.032 -0.088 -0.31 1 -0.02 0.15 0.0074
-0.062 -0.024 0.058 0.0018 -0.02 1 -0.15 -0.3
0.11 0.022 -0.13 0.011 0.15 -0.15 1 0.31
0.05 -0.0026 0.00012 -0.016 0.0074 -0.3 0.31 1
Pla ne Name
Departure Time
Arri\'al Time
Departure City
Arri\'al City
Numbe'* of sto»s
Fa'e
Tme_taken
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Observations:

#indigo provides the greatest number of services while Trujet provides least number of services.

#most flights depart in the morning

#most flight arrives in the evening

#from the collected data New Delhi is the has the most departed flight

#from the collected data New Delhi is the has the most arrived flight

#majority of flights have one stops

#the prices lie b/w 5k-7k. also there are few flights with higher price

#majority of the flights takes average 500-800 minute

#go fist has the highest fare and the lowest also

the flights that departs from Delhi has more fare than others

the flights that arrive in Bangalore has more fare than others

#the flight with 1 stop is more costly

#fights that departs in the morning costs more

##the flights that lands in afternoon has more fair

#fare shows a liner relationship with time

#fare is highly correlated with time taken

#there are some outliers in Time taken

CONCLUSION

This paper showed the model training process for the prediction of the fare Price. One of the objectives of the paper was to check the important variable for the prediction of the price and how these variables describe the price. Through model training and evaluating its performance. RandomForest proved to be as best model. As the difference between the r2score and cross validation score was minimum. This project has increased my understanding of the concept. During the research I came across various challenges and while solving them I learned a lot of new things. For example. How to plot different charts. For example, I learned how to plot subplot. I learned new libraries and how to use them. I explored various methods for feature selection. Also, I came to understand how can multicollinearity can cause problem during the model training.The limitation of the solution provided is that the data carried a lot of unrealistic values. Apart from that my laptop took to much time while running certain command where I lost a lot of precious time.