

Submitted by:

Ankit Prasad

**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 were beneficial 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 areairline 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 accuracy r2 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

* Acer Aspire 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 has no 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

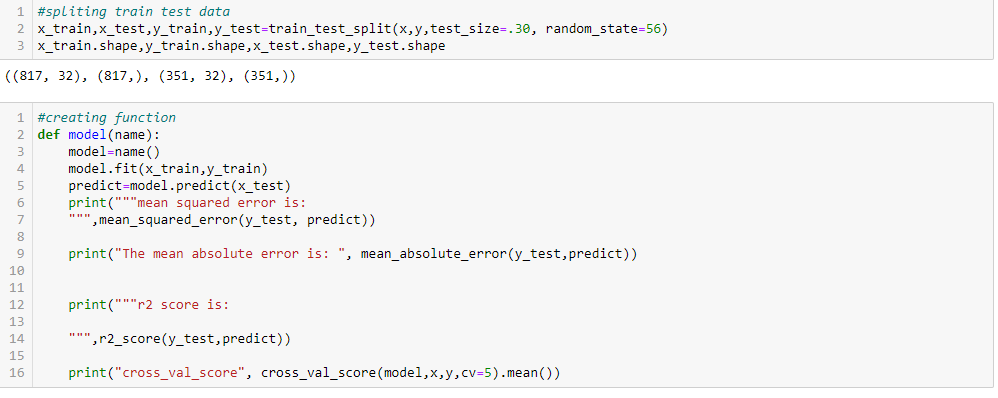
from sklearn.ensemble import BaggingRegressor

from sklearn.ensemble import AdaBoostRegressor

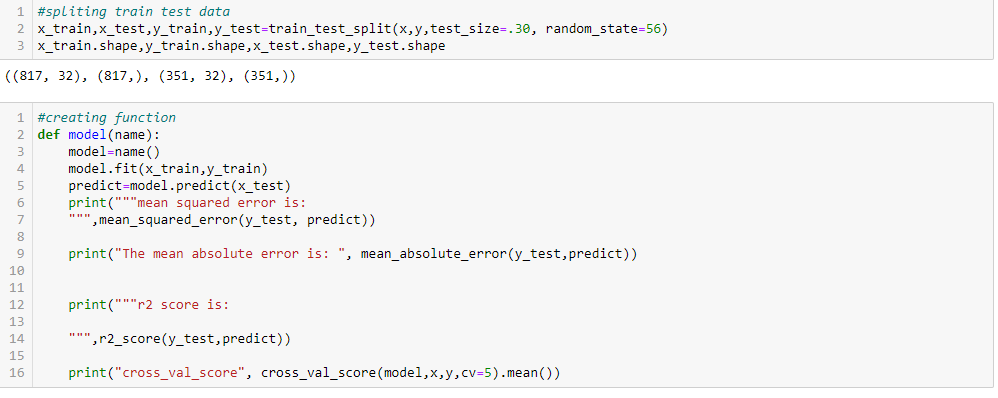
from sklearn.ensemble import GradientBoostingRegressor

from sklearn.model\_selection import GridSearchCV

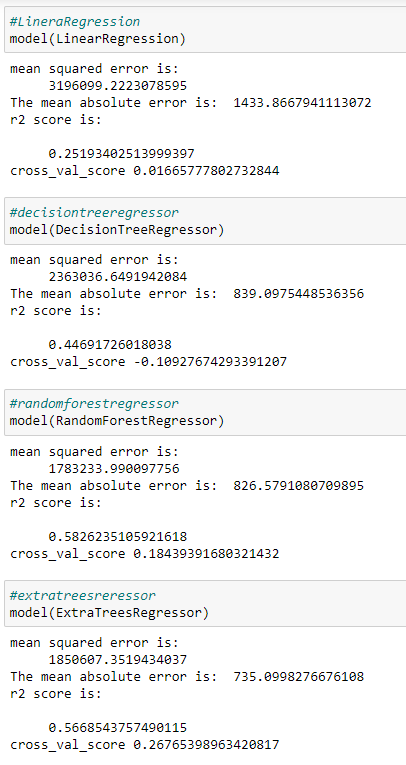
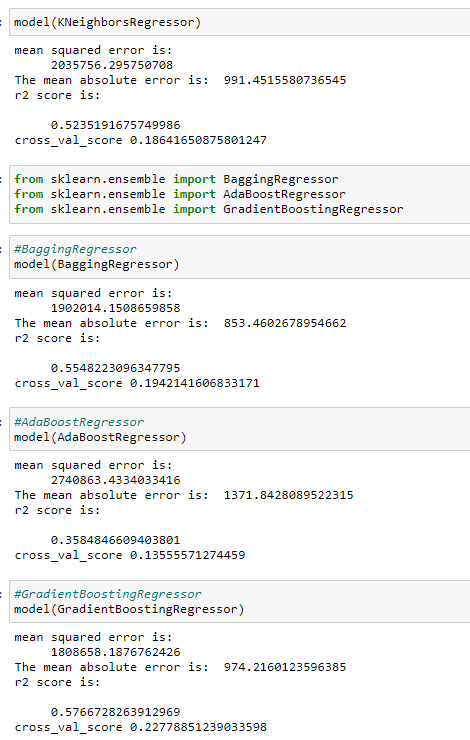
**Performance of the model:**



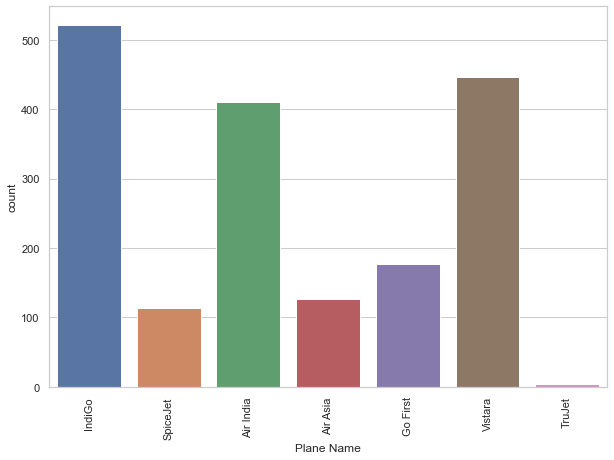
By the train test split method, 70 percent of the data has been taken for the model building while 30 percent of the data has been reserved for checking the model's performance.

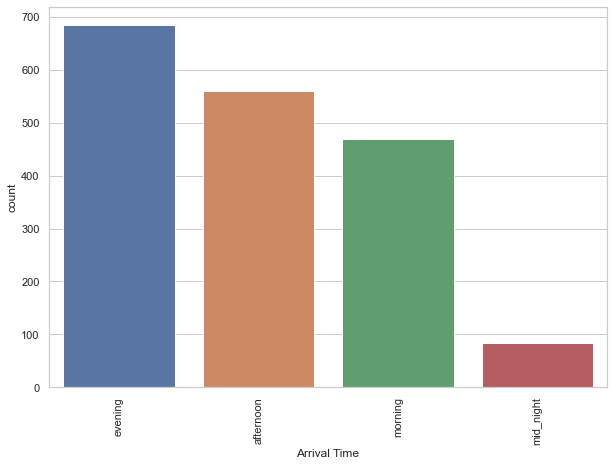


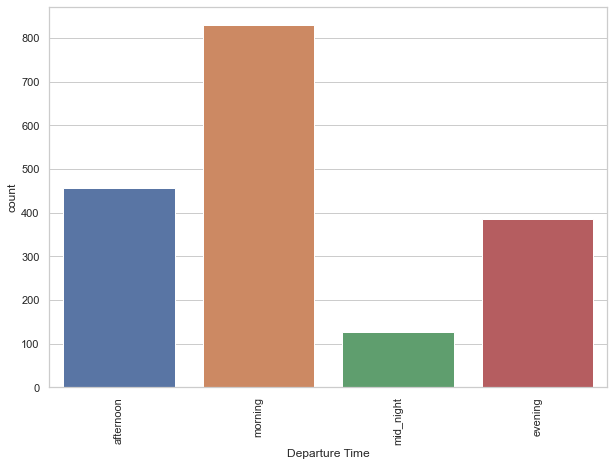
The code above has been used to speed up the model training and its evaluation process. Here the function name model ha been created which take the name of model as argument.

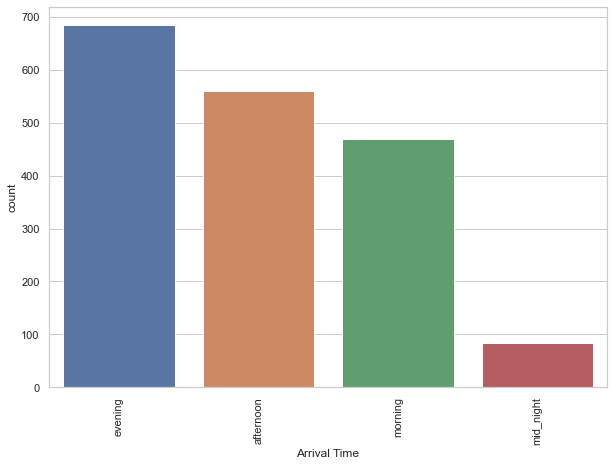
** **

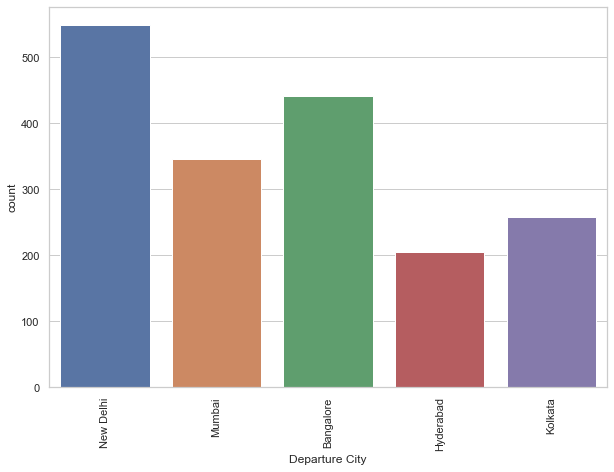
**Visualization:**

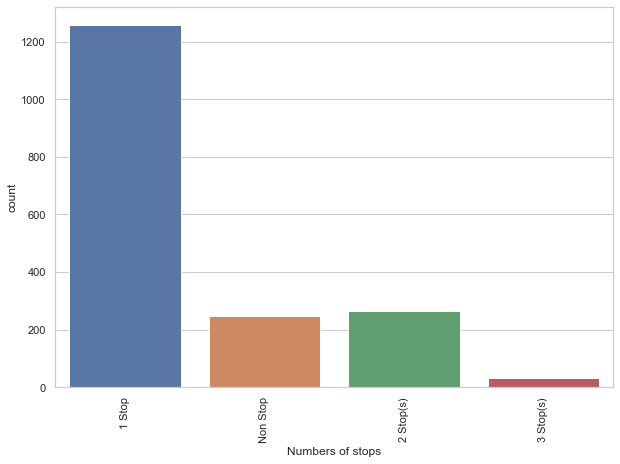
****

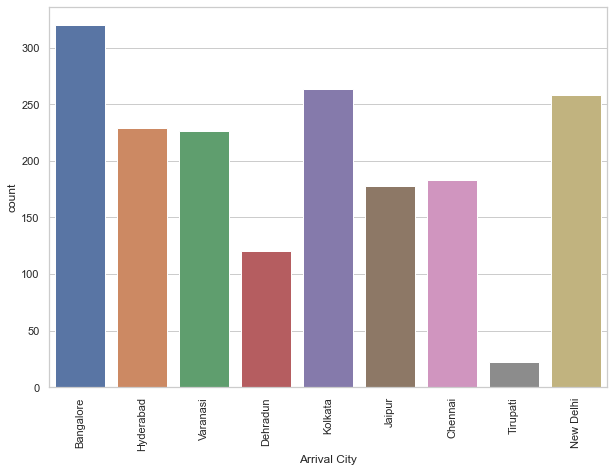
****

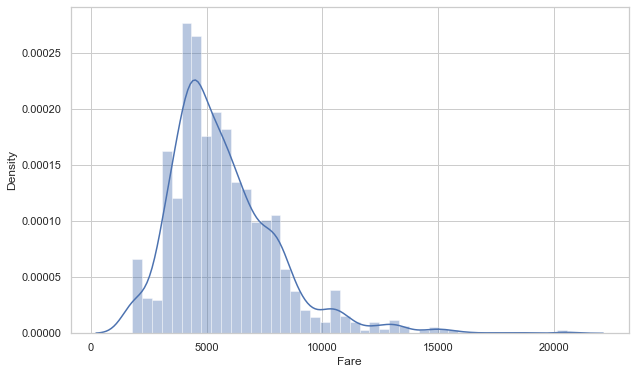
****

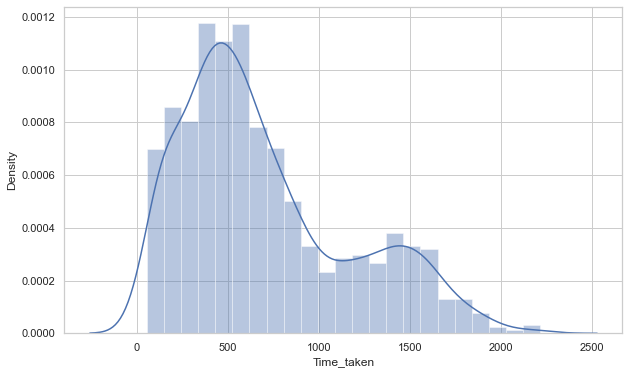
****

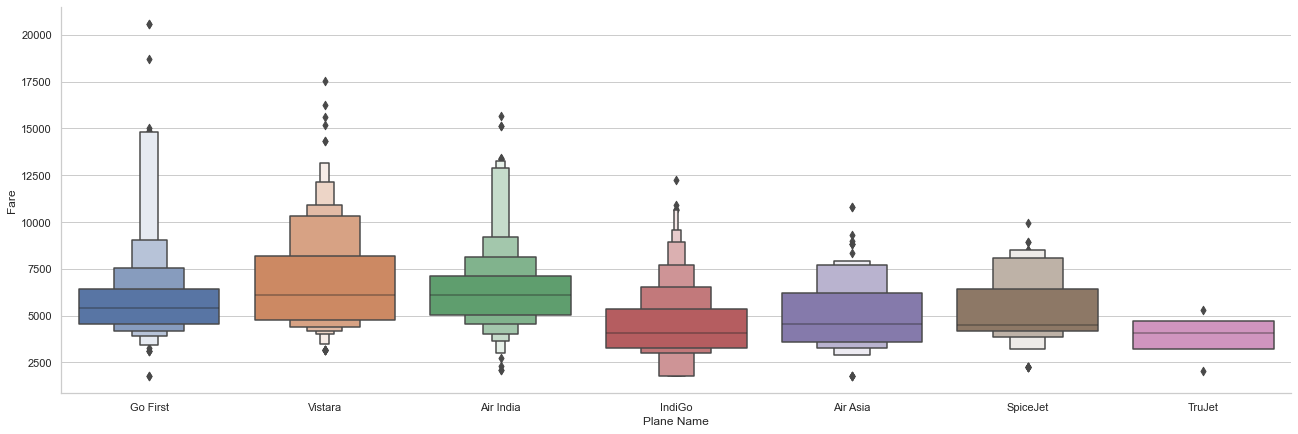
****

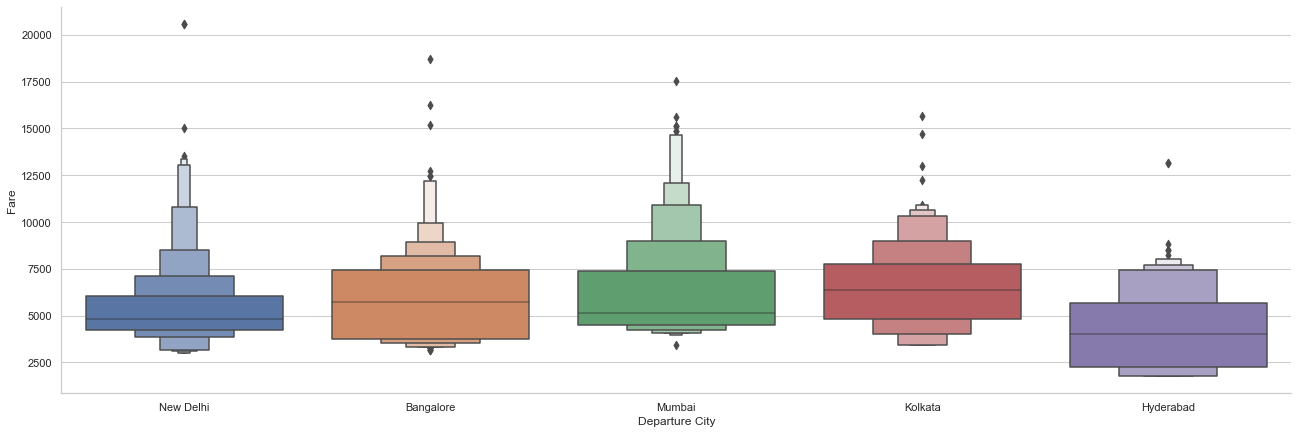
****

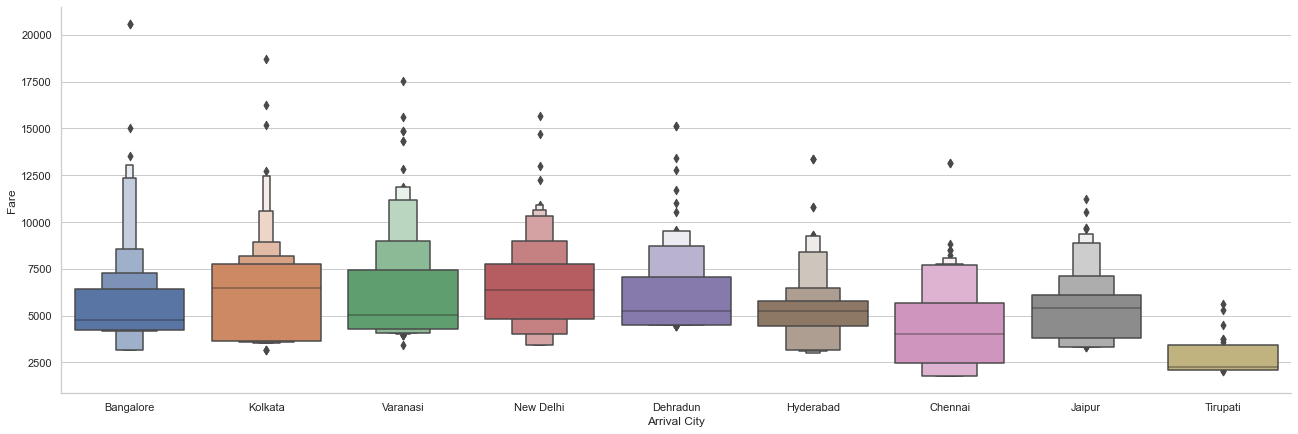
****

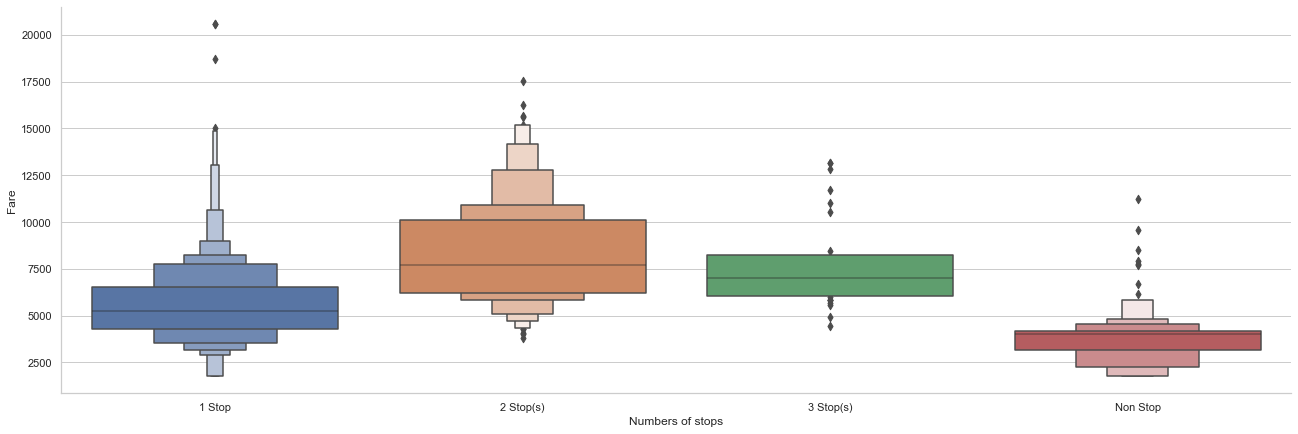
****

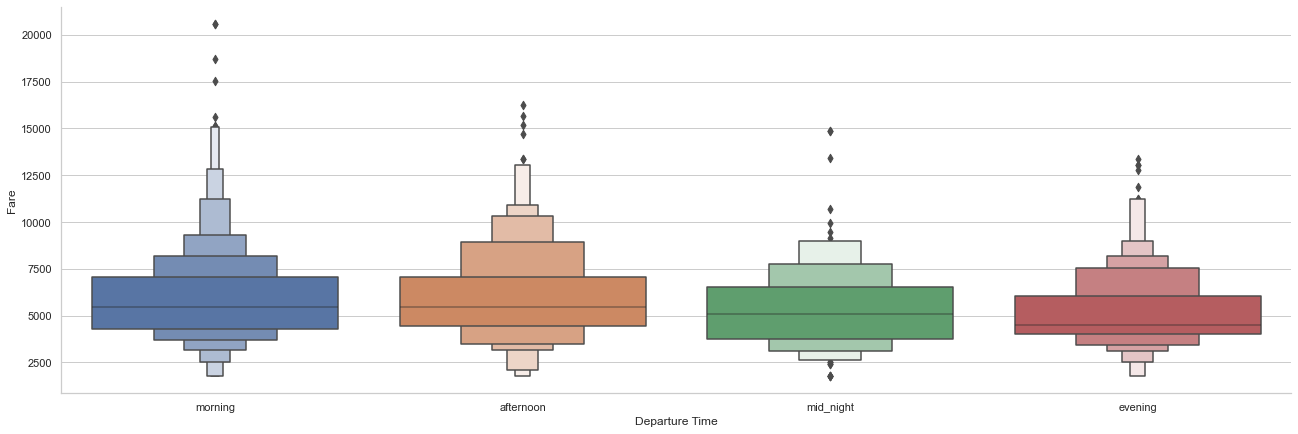
****

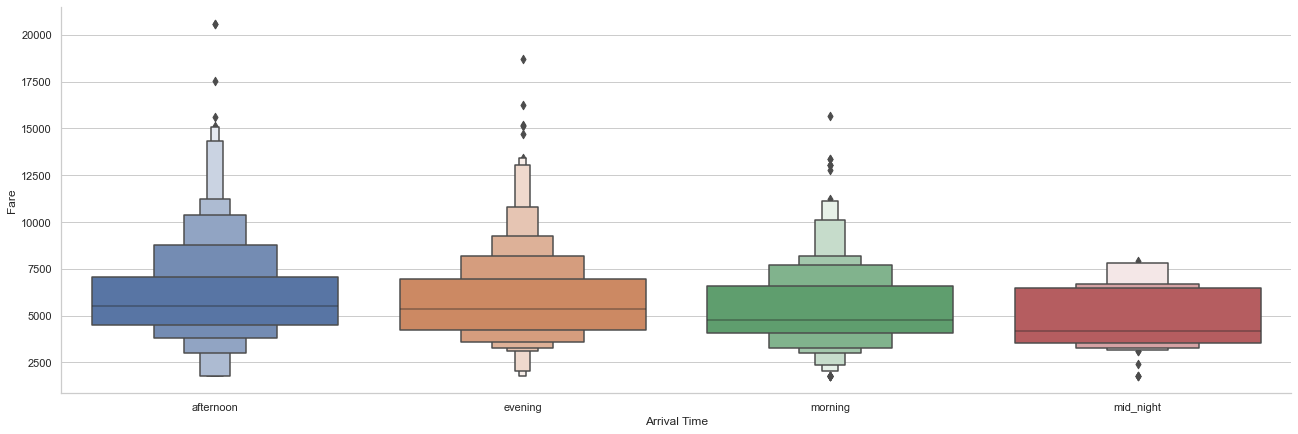
****

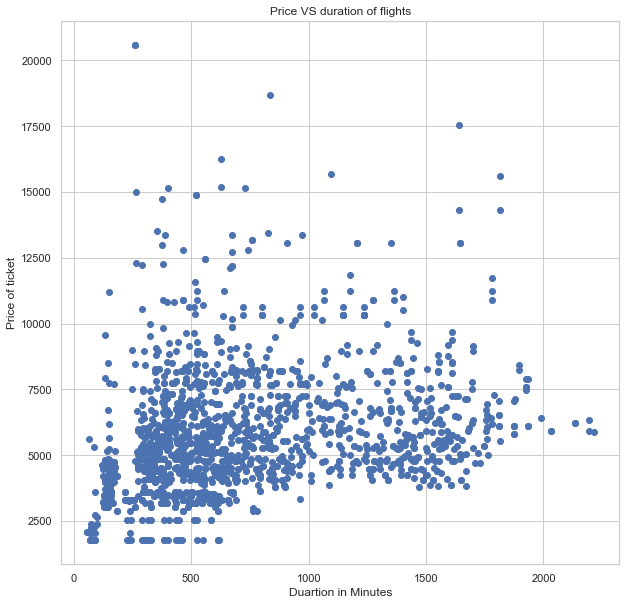
****

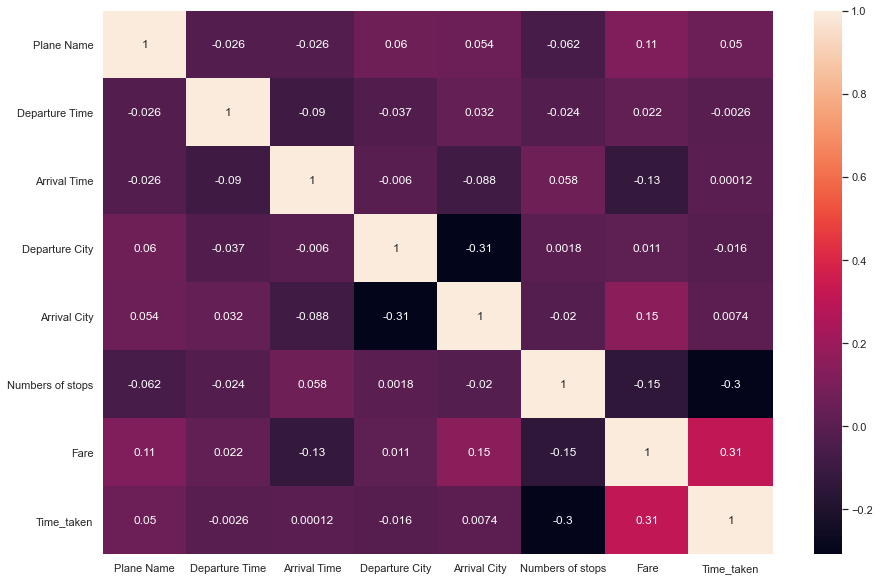
****

****

****

****

****

****

**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.