FLIGHT PRICE PREDICTION

Submitted by:

Ishmeet Kaur Sahota

ACKNOWLEDGMENT

I want to express my sincere thanks to Flip Robo Techniques and my SME Khusboo Garg who helped me in every possible way that she could and guide me through new things. without whom I won't have been able to complete this project .

INTRODUCTION

• Business Problem Framing:-

To know 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).

• Review of Literature:-

In this project, I have use ***MakeMyTrip*** *website as a source to make my database. I have used web scraping (selenium) to collect my data. I have scraped the flights related information i.e:- Airline Name , Departure Place , Departure Time , Flight Duration , Arrival Place , Arrival Time , Total Stops and Price etc.*

• Motivation for the Problem Undertaken:-

To know how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time .This machine learning will help in Flight price valuation. It helps in understanding how to deal with prices when this kind of situation occur and what steps should be taken to fix these situations.

Analytical Problem Framing

• Mathematical / Analytical Modelling Problems:-

The dataframe contains several rows and columns containing all the necessary information. I have used the replace method to replace ('₹',' ') in price column. And i have deleted the extra space present in the "Flight_Duration" column . I have used several statistical and exploratory data visualizations for better understanding .

• Data Sources and their formats:-

The source of data is MakeMyTrip website. Then I stored the data in a dataframe. The Dataframe contains 1980 rows and 8 columns . containing all the details of Flights .

Dataframe contains several columns :-

- *Airline Name*
- *Departure Place*
- *Departure Time*
- *Flight Duration*
- *Arrival Place*
- *Arrival Time*
- *Total Stops*
- *Price*

• Data Preprocessing Done:-

I have used the replace method to replace ('₹',' ') in price column. And also i have deleted the extra space present in the “Flight_Duration” column. I have converted / Separated “Departure Time” and “Arrival Time”. Used a For Loop to separate the hours ”h” and mins “m” in the “Flight_Duration” column. I get Dummies of “Airline_Name” . used replace method on “Total_stops” column and replaced “Non stop” with 0 and “1 stop” with 1. Then applied concatenation method .

• Hardware and Software Requirements and Tools Used:-

I have used MakeMyTrip website to scrape the data of flights and to make the Dataframe. then I imported several libraries for further model building , EDA and data cleaning.

```
import selenium
import pandas as pd
from selenium import webdriver
import warnings
warnings.filterwarnings("ignore")
import time
from selenium.common.exceptions import StaleElementReferenceException ,
NoSuchElementException
import re
from selenium.webdriver.common.by import By
import requests
import seaborn as sns
from sklearn import metrics
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error
```

Model/s Development and Evaluation

• Identification of possible problem-solving approaches (methods):-

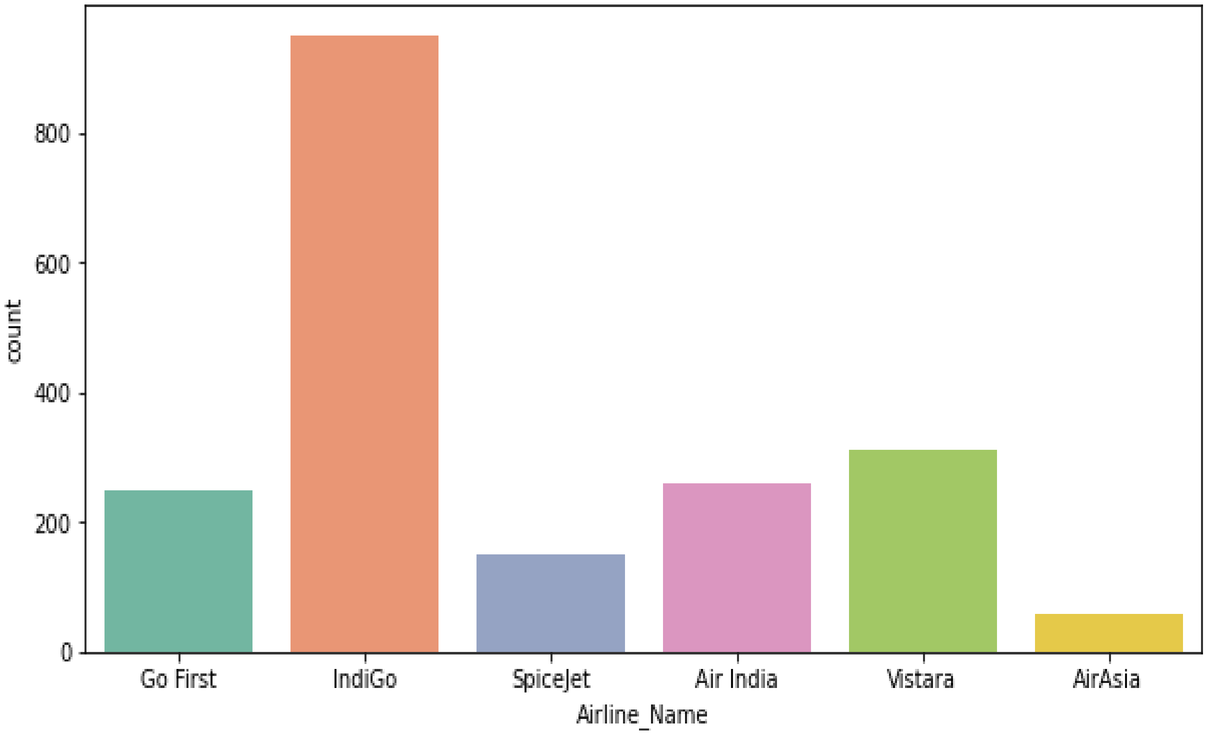
```
Import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline import
seaborn as sns
```

used replace method and also use a for loop to separate “h” and “m” from Flight_Duration.

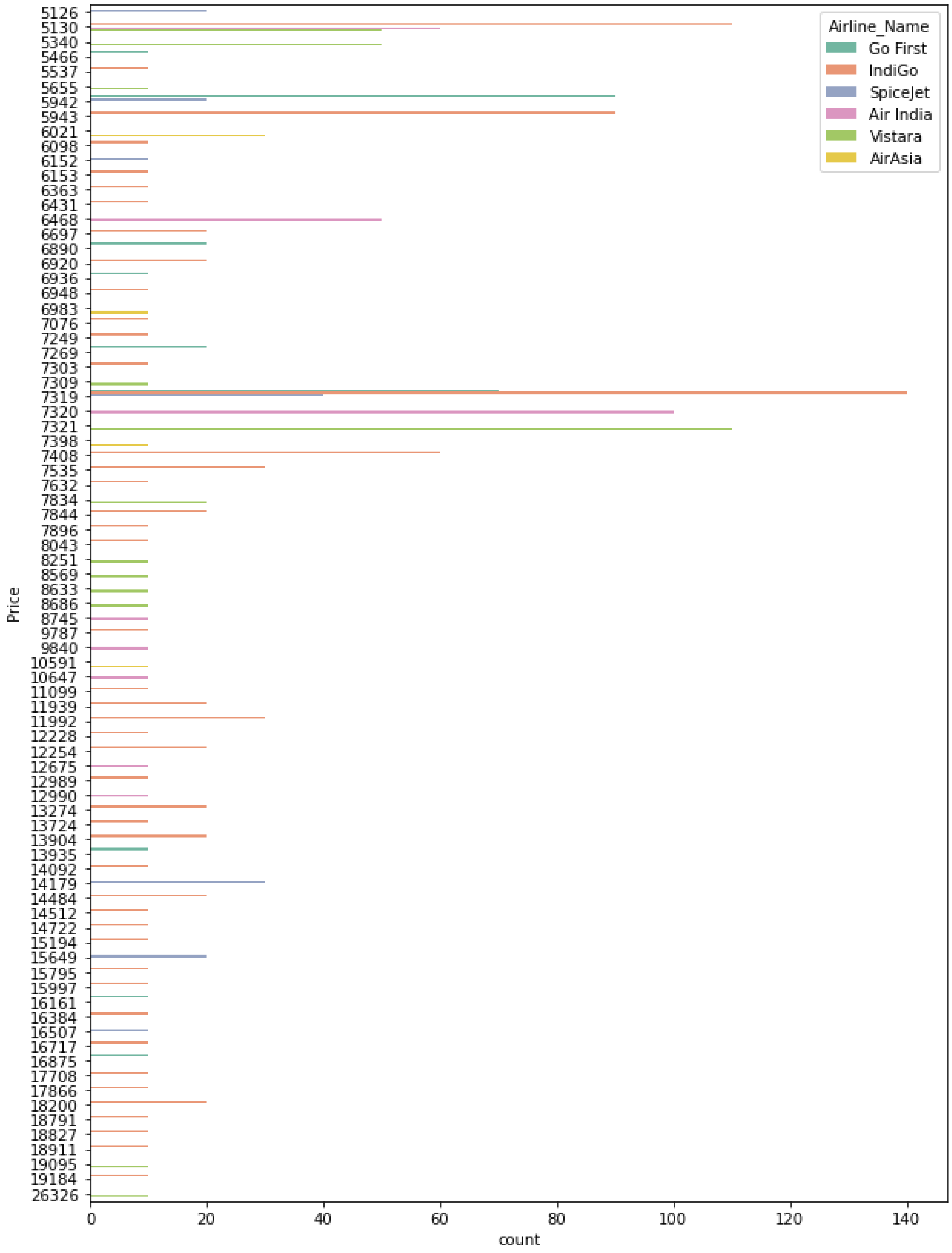
•Testing of Identified Approaches (Algorithms):-

```
from sklearn import metrics
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error
```

Visualizations

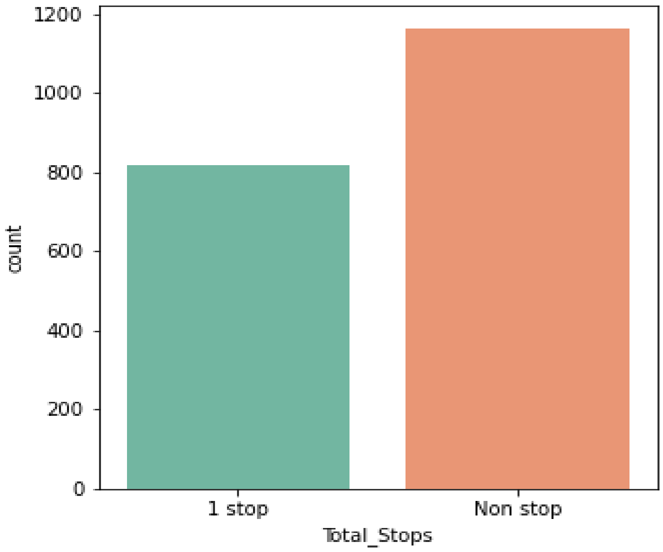


As u can see IndiGo is the most prefered Airline and AirAsia is the least prefered Airline.

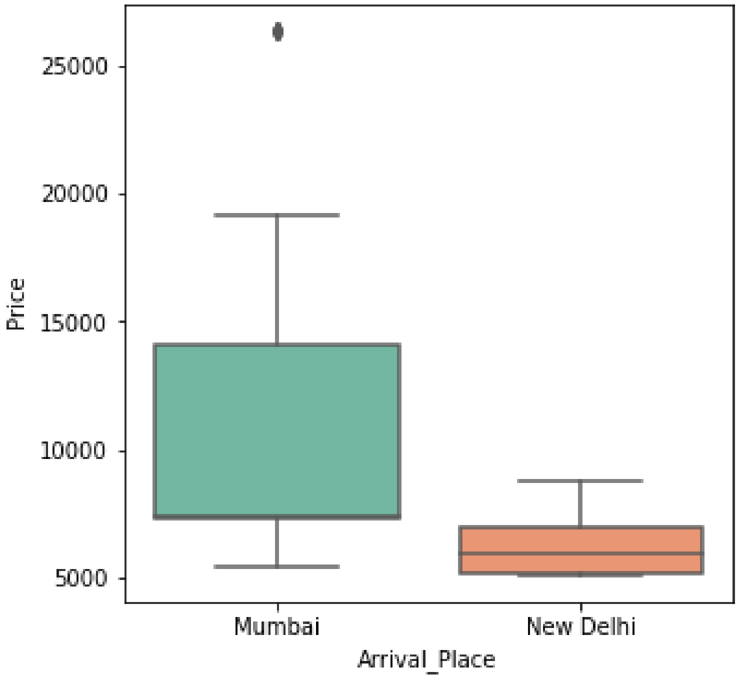


Checking the relationship between Airline Name and their Prices.

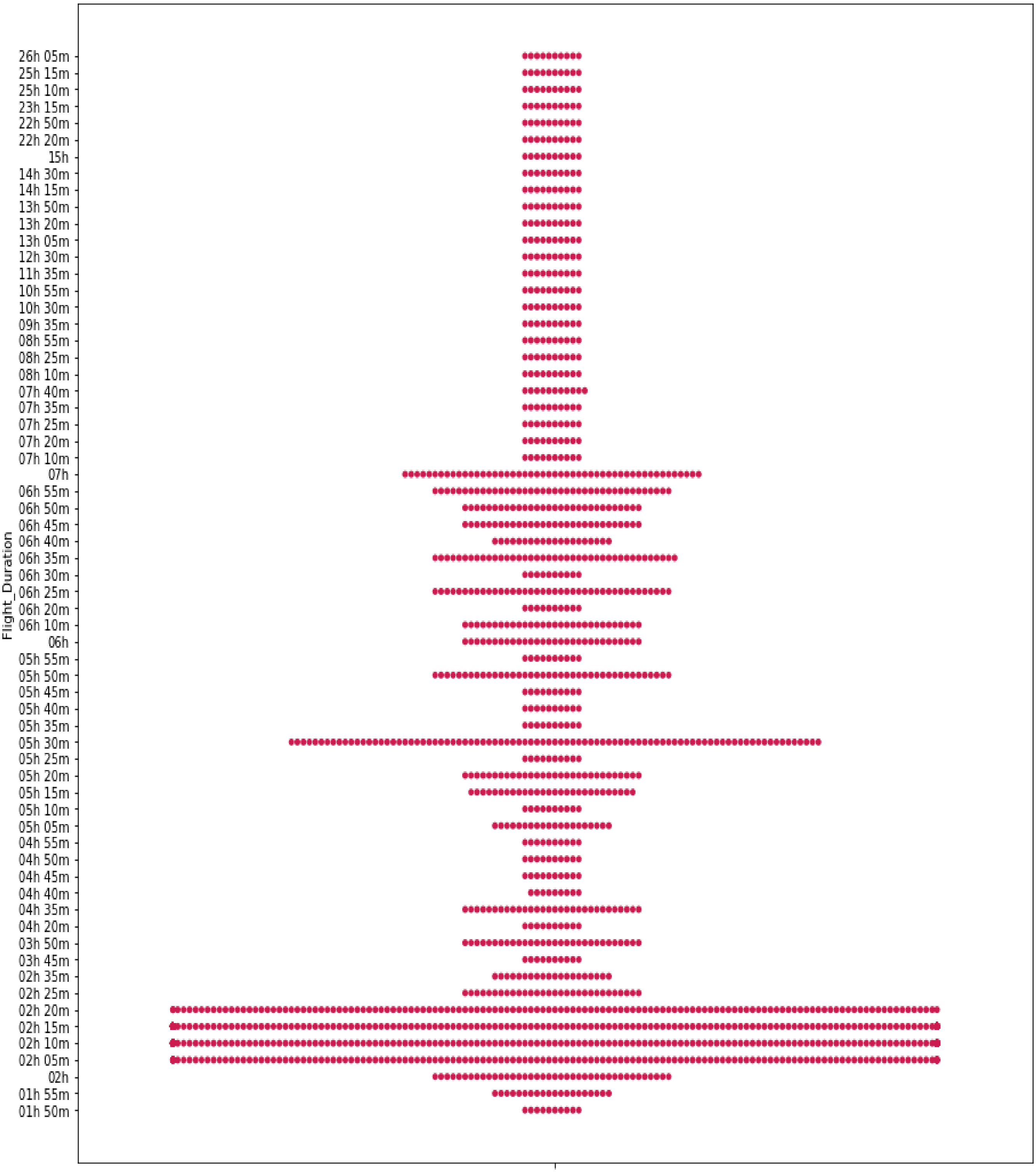
Here u can Non Stop flights are more preferable as compared to 1 stop flight.



Checking the relationship between Arrival Place and Price



Used Swarmplot to check the flight Duration



• *Interpretation of the Results:-*

After visualizing the data I have concluded that. IndiGo is the most prefered Airline and AirAsia is the least prefered Airline. IndiGo offers more price ranges the customers staring from the Cheapest to Business class and AirAsia offers less prices ranges . Non stop flights are prefered more as compared to 1 stop flights. Also show the flight Durations that starts from 1h 50m to 26h 05m.

I have put 4 Regression models :-

LogisticRegression
KNeighborsRegressor
DecisionTreeRegressor
RandomForestRegressor

LogisticRegression gives 67% of accuracy and KNeighborsRegressor, DecisionTreeRegressor and RandomForestRegressor gives 98% of accuracy.

Conclusion

• *Key Findings and Conclusions of the Study:-*

After applying visualizations techniques I concluded that. ***.*** *IndiGo is the most prefered Airline and AirAsia is the least prefered Airline. IndiGo offers more price ranges the customers staring from the Cheapest to Business class and AirAsia offers less prices ranges . Non stop flights are prefered more as compared to 1 stop flights. Also show the flight Durations that starts from 1h 50m to 26h 05m. and after applying all the statistical techniques and data cleaning techniques .I have put 4 Regression models :-*

LogisticRegression
KNeighborsRegressor
DecisionTreeRegressor
RandomForestRegressor

And concluded that LogisticRegression gives 67% of accuracy and KNeighborsRegressor, DecisionTreeRegressor and RandomForestRegressor gives 98% of accuracy.

• *Learning Outcomes of the Study in respect of Data Science :-*

In the Dataframe I have used the replace method to replace ('₹',' ') in price column. And also i have deleted the extra space present in the “Flight_Duration” column. I have converted / Separated “Departure Time” and “Arrival Time”. Used a For Loop to separate the hours ”h” and mins “m” in the “Flight_Duration” column. I get Dummies of “Airline_Name” . used replace method on “Total_stops” column and replaced “Non stop” with 0 and “1 stop” with 1. Then applied concatenation method . I have applied different statistical operations and different plots like Swarmplot, Boxplot, and Countplot for visualization and to understand it better. And applied 4 different models on the data as well.