

FLIGHT PRICE PREDICTION

Submitted by: Harneet Kaur rehsi

ACKNOWLEDGMENT

I want to thank my SME Khusboo Garg and Flip Robo Techniques to express our sincere thanks to the following people, without whom we would not have been able to complete this project. I have scrapped the data from google flights which is an online flight booking website where u will find out all the necessary details regarding flights.

INTRODUCTION

- **Business Problem Framing:** 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. And a model to predict the fares of flights.
- **Review of Literature:** In this project, we need to scrap flight-related information i.e airline name, date of journey, source, destination, route, departure time, arrival time, duration, total stops, and the target variable price, etc. information. From online websites. There are several websites available regarding flight booking etc (yatra.com, skyscanner.com, official websites of airlines, etc). many more. I have scrapped data using one of the Data scraping technique: Selenium. I have scrapped my entire data from the online website name Google Flights
- **Motivation for the Problem Undertaken:** objective behind making this project is to book a flight ticket knowing how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time. And a model to predict the fares of flights. This model will then be used to understand how exactly the price varies with the variables. The flight price rise according to some situations as well.

Analytical Problem Framing

- **Mathematical/ Analytical Modeling of the Problem:** In this project, the dataset contains several rows and columns containing all the necessary information. I have used replace method to replace nan values with some meaningful data. Splits the hours and minutes time of departure and arrival and duration accordingly. In The dataset i have used several statistical and exploratory data visualization for better understanding and model building for predictions
- **Data Sources and their formats:** The data contain 1720 rows and 8 columns respectively. containing all the necessary details. Data is scrapped or originated from Google Flights which is an online flight booking website and all the necessary details regarding flights.

	Unnamed : 0	Airlines	Departure_Time	Route	Arrival_Time	Flight_Duration	Total_Stops	Price
0	0	GO FIRST	10:40	DEL-GOI	1:20	2hr 40min	Nonstop	₹15,984
1	1	SpiceJet	12:05	DEL-GOI	2:25	2hr 20min	Nonstop	₹16,383
2	2	IndiGo	12:25	DEL-GOI	2:50	2hr 25min	Nonstop	₹16,383
3	3	Vistara	11:10	DEL-GOI	1:50	2hr 40min	Nonstop	₹17,040
4	4	GO FIRST	5:25	DEL-GOI	5:25	12hr	1 stop	₹16,246

- Data Preprocessing Done:** Data contains some time based columns (Departure time, Arrival time, and Flight duration) with the help of DateTime I have separated minutes and hours separately in different columns. To clean the data I have used the fillna method to fill nan values with mean. To fill NaN values present in the particular dataset. Data also contain columns with object-type data which we need to convert into numerical or encode them. So that our machine learning model understands them because the machine learning model understands only numeric type data. I have changed some columns' data types into integers as well. with the help of replace method, I have replaced Currency symbols and commas present in the price column. and also used a Label Encoder to encode all the categorical or object data present in Airlines column into some label so that our machine learning model will understand the information present.
- Hardware and Software Requirements and Tools Used:** I have used Data Scrapping Technique Selenium for scrapping the data then I have used and imported several libraries like pandas , numpy , matplotlib.pyplot ,seaborn , siciklearning models etc:

- 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 pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- %matplotlib inline
- import warnings
- warnings.filterwarnings('ignore')
- import seaborn as sns
- import datetime
- from sklearn.preprocessing import LabelEncoder
- from sklearn import metrics
- from sklearn.model_selection import train_test_split
- from sklearn.linear_model import LogisticRegression
- from sklearn.tree import DecisionTreeRegressor
- from sklearn.ensemble import RandomForestRegressor
- from sklearn.preprocessing import StandardScaler
- from sklearn.neighbors import KNeighborsRegressor

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 warnings
 - warnings.filterwarnings('ignore')

- import seaborn as sns
- import datetime
- from sklearn.preprocessing import LabelEncoder

- **Testing of Identified Approaches (Algorithms)**

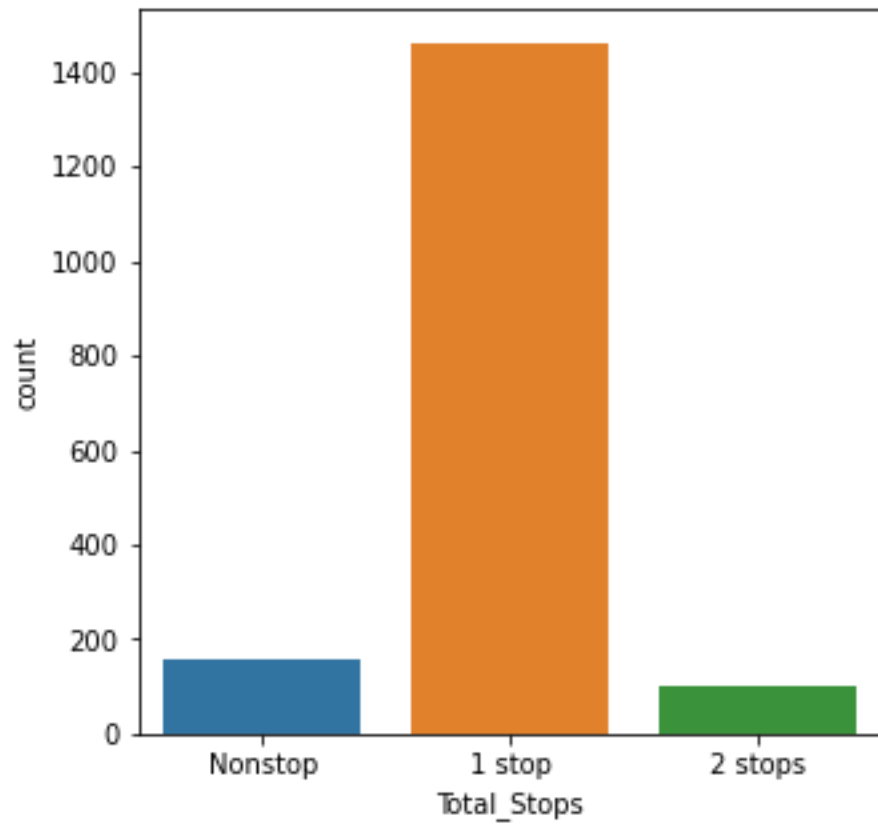
- from sklearn import metrics
- from sklearn.model_selection import train_test_split
- from sklearn.linear_model import LogisticRegression
- from sklearn.tree import DecisionTreeRegressor
- from sklearn.ensemble import RandomForestRegressor
- from sklearn.preprocessing import StandardScaler
- from sklearn.neighbors import KNeighborsRegressor

- **Run and Evaluate selected models:** I have used Four different Machine Learning Regression Models here.

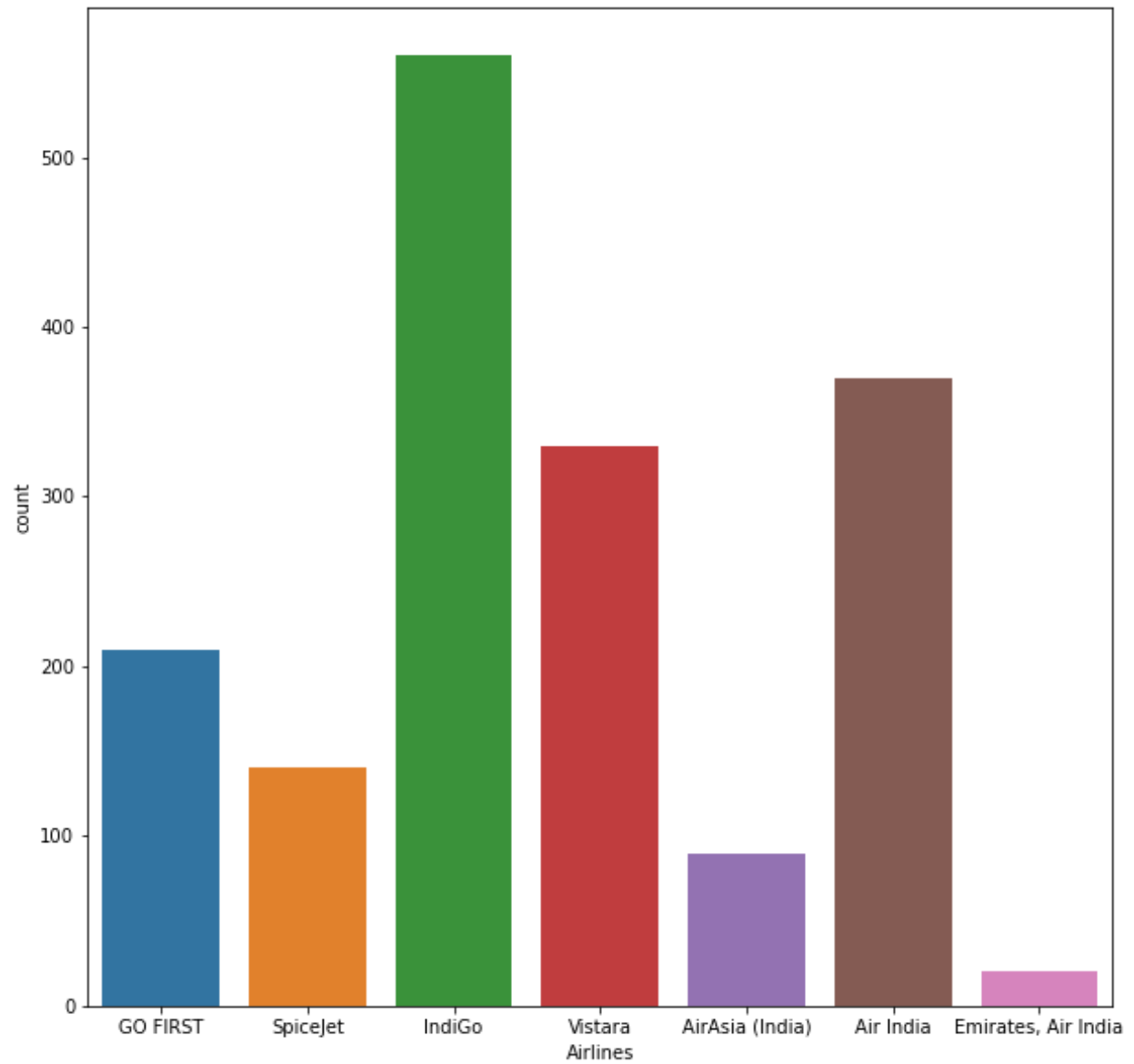
-  Logistic Regression
-  Decision Tree Regressor
-  Random Forest Regressor
-  KNeighbors Regressor

Here Decision Tree Regressor and KNN regressor are fitting best with a score of 100% . and Logistic Regression and Random Forest Regressor with a Score of 99%.

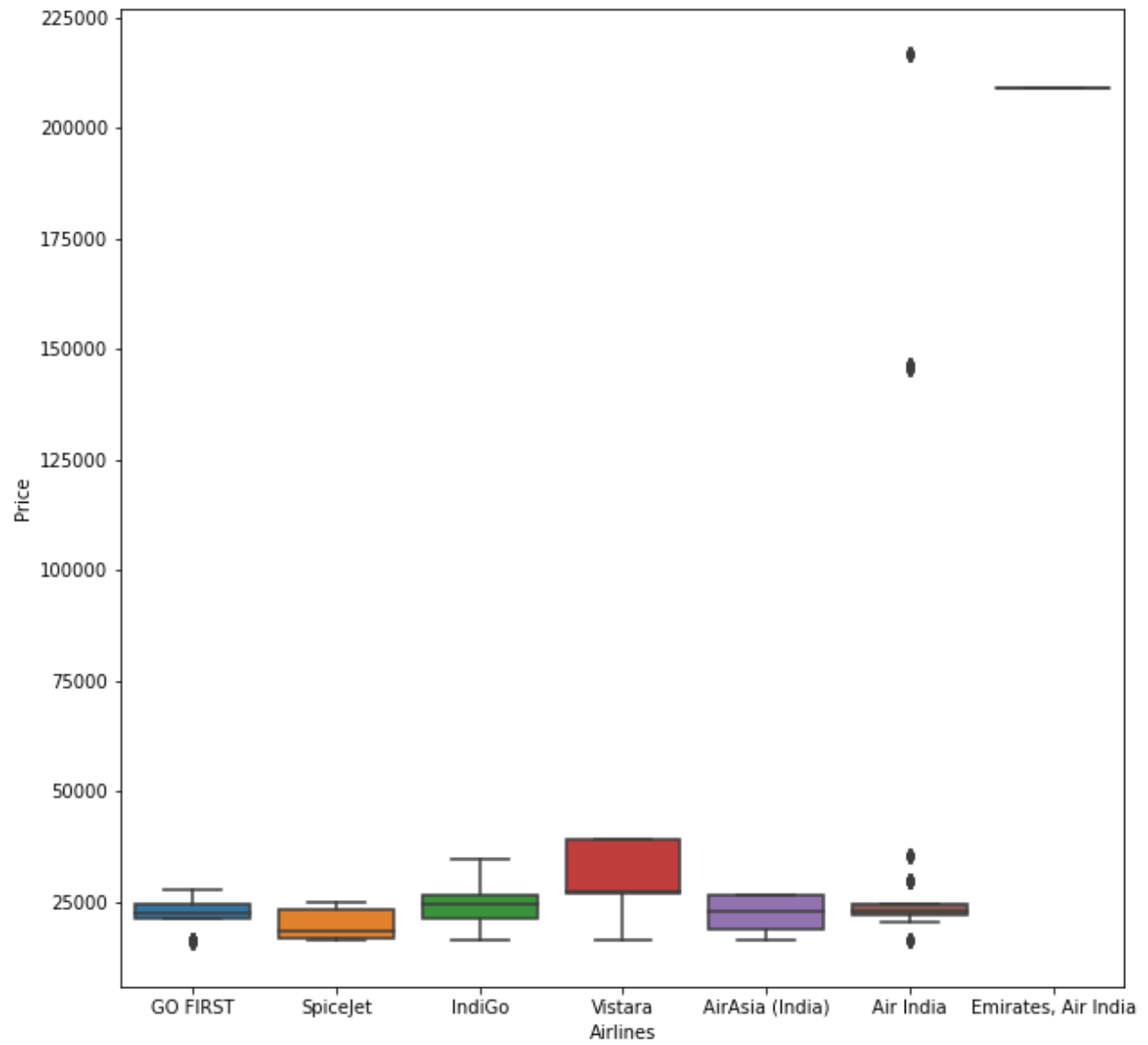
Visualizations



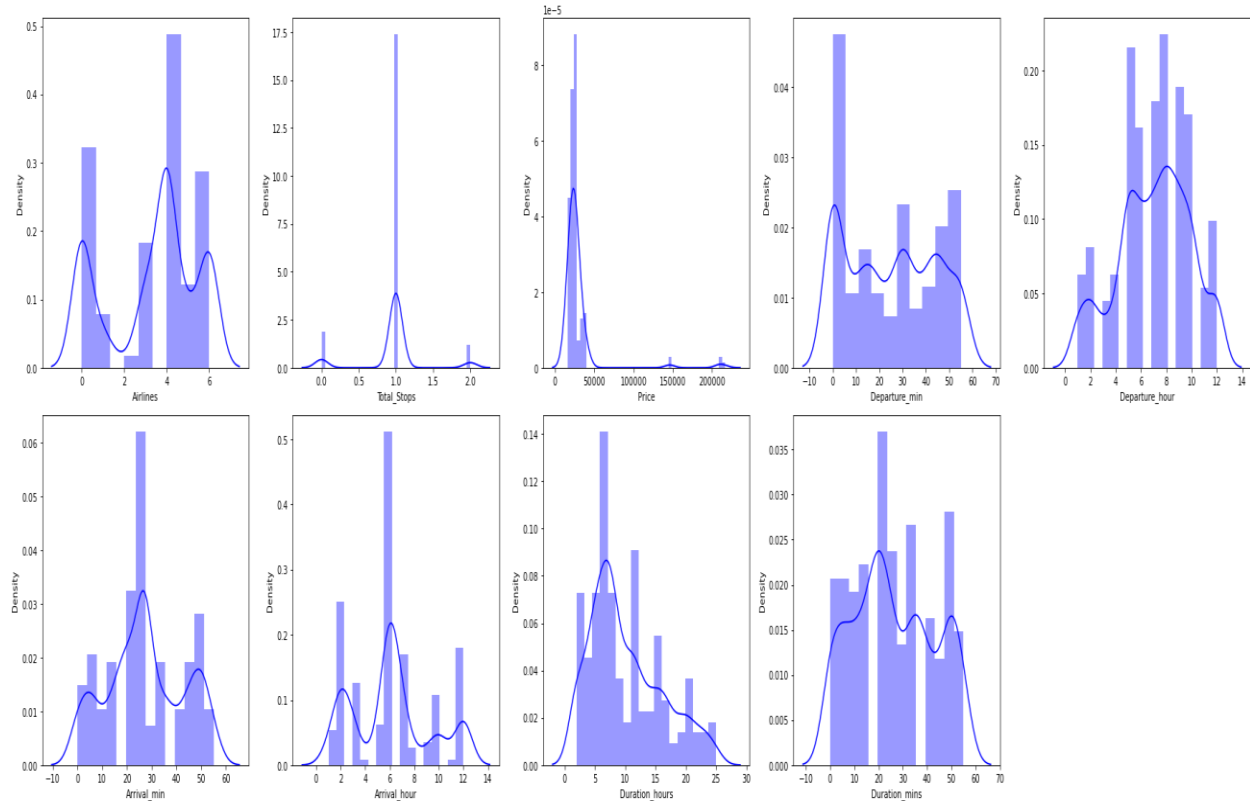
Countplot showing the number of stops



Countplot showing Different types of Airlines



Boxplot showing the relationship between Different types of Airlines with respect to price.



Distplot is used to check if the data is Skewed or not.

- Interpretation of the Results:** After visualizing the data I have concluded that most of the flights took 1 stop. There are different types of Airlines. And Boxplot makes us understand the relationship between Airlines and price respectively. Distplot tell us that our data is not Skewed. After using preprocessing methods to convert the data or encode the data and scale the data so that our machine learning model will understand the data properly. I have tried different machine learning models the score are a little different in all the models.

Logistic Regression model: 99% score value
 Random Forest Regression model: 99% score value
 Decision Tree Regression model: 100% score value
 K Neighbors Regression model: 100% score value

CONCLUSION

- **Key Findings and Conclusions of the Study:** After visualizing the data I have concluded that most of the flights took 1 stop. There are different types of Airlines. And Boxplot makes us understand the relationship between Airlines and price respectively. Distplot tellu that our data is not Skewed. After using preprocessing methods to convert the data or encode the data and scale the data so that our machine learning model will understand the data properly. The model that I have built give “100” score. I have tried different machine learning models the accuracy rate is a little different in all the models.

- **Learning Outcomes of the Study in respect of Data**

Science: Data contains some time-based columns (Departure time, Arrival time, and Flight duration) with the help of DateTime I have separated minutes and hours separately in different columns. To clean the data I have used the fillna method to fill nan values with mean. To fill NaN values present in the particular dataset. Data also contain columns with object-type data which we need to convert into numerical or encode them. So that our machine learning model understands them because the machine learning model understands only numeric type data. I have changed some columns’ data types into integers as well. with the help of replace method, I have replaced Currency symbols and commas present in the price column. and used a Label Encoder to encode all the categorical or object data in Airline column into some label so that our machine learning model will understand the information present. After visualizing the data I have concluded that most of the flights took 1 stop. There are different types of Airlines. And Boxplot makes us understand the relationship between Airlines and price respectively. Distplot tell us that our data is not Skewed. After using preprocessing methods to convert the data or encode the data and scale the data so that our machine learning model will understand the data properly. I have tried different machine learning models the accuracy rate is a little different in all the models.

Logistic Regression model: 99% score value
Random Forest Regression model: 99% score value
Decision Tree Regression model: 100% score value
K Neighbors Regression model: 100% score value