FLIGHT PRICE PREDICTON PROJECT

Submitted by:

VAISHNAVI BALAJI

ACKNOWLEDGMENT

[1] Bachis, E., & Piga, C. A. (2011). “Low-cost airlines and online price dispersion. International Journal of Industrial Organization”, 29(6), 655– 667. doi:10.1016/j.ijindorg.2011.02.006Enis gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric,” Car Price Prediction Using Machine Learning”; (TEM Journal 2019)

[2] Groves, W. and Gini, M., 2021. “A Regression Model For Predicting Optimal Purchase Timing For Airline Tickets.”. Available at:https://conservancy.umn.edu/handle/11299/215872

[3] T. Wohlfarth, S. Clemencon, F. Roueff and X. Casellato, "A Data-Mining Approach to Travel Price Forecasting," doi: 10.1109/ICMLA.2011.11.

INTRODUCTION

- Business Problem Framing

It might be challenging to choose the best time to buy an airline ticket from the perspective of the traveller because they have very little knowledge about future business price rates. Different models forecast future air travel costs and classify the ideal window for booking tickets. Airlines use a variety of pricing systems for their tickets, making price decisions later since order displays a larger value for approximation models. Each of these factors is what makes the system challenging. Airlines must control demand since there are only so many seats that can be filled in planes. Assume that the airline may raise prices to slow down the rate at which seats fill when demand exceeds capacity.

Additionally, requiring business airline businesses to buy tickets to fill empty seats at any cost would be the best way to turn a profit while also incurring a loss due to the seating arrangements in empty flights. In order to compensate for price increases and decreases, passengers must be amenable to airline firms. Customers or passengers should make their own plans in order to take advantage of the greatest deals offered by various airlines and fly for a lower cost. As time goes on, the cost of airline tickets fluctuates, separating the factors that cause the difference.

Reporting the models that are connected and utilised to determine the cost of airline tickets. Using that data, a model was created to assist customers in selecting and purchasing tickets as well as forecast future increases in ticket costs. The attributes used for flight price prediction include duration, arrival time, price, source, destination, and a great deal more.

- Conceptual Background of the Domain Problem

In the realm of airline tickets, adversarial risk can be seen in two contexts: the conflict between customers and sellers, and the rivalry between different airlines offering the same service. While sellers aim to keep overall revenue as high as possible to maximise profit, buyers frequently want to pay the least amount feasible for their vacation. Each seller must simultaneously take into account changes in the prices of its rivals in order to maintain prices that are sufficiently competitive to generate enough (but not excessive) demand. Without taking into account both forms of antagonistic interactions, it is hard to adequately handle the issue of optimising decision making from the perspective of the buyer.

Sellers (airlines) invest a lot of money over a long period of time on hardware (planes), route agreements, and fixed infrastructure (airports, repair facilities). These long-term decisions' specifics are meant to roughly meet anticipated demand, but they frequently don't. Airlines employ the dynamic setting of prices as a tool to better match their unique supply and demand profiles in order to maximise revenue.

The knowledge imbalance between buyers and sellers is a major problem in the airline ticket purchasing area. Airlines can create models for anticipated future demand for each flight by mining enormous databases of historical sales data. Demand for a particular flight is likely to change over time and will also change depending on the airline's chosen pricing strategy. It is typically recommended for purchasers to purchase long in advance of a flight's departure because the prices have a tendency to skyrocket as the departure date gets closer. However, airlines frequently disregard this rule and lower prices to boost sales.

- Review of Literature

The previous work on enhancing prediction models for airline fares by utilising Machine Learning (ML) approaches focused on different features and trained the models on different types of Airlines. They are attempting to predict the price, which is a certain trend. Particularly, classifying flight costs into two divisions of components aids in the analysis of the impact on average plane costs. After 70 days, categorical cases for a flight are noticed, according to the authors' analysis of the airline profit using pricing modes. as soon as an aircraft takes off, and the chances for discounts also tend to grow. We find equal pricing strategies used by airline firms to effectively manage airline offers and demand to boost their financial performance through analysis. Results indicate that airlines are concerned about websites' seasonal price adjustments.

The significance of the relationship between online pricing and the actual price dispersion on a flight should be recognised. The authors have concluded that online pricing division is more strongly prevalent in lower business airline competition using prices from actual transactions.

First off, a plane's lowest price will not be available during the high price period since the passenger information is particularly scheduled and may need to be adjusted for the price. It is not necessary to use the attribute from the dataset that mentions good understanding of the subject. To check your comprehension of the subject, look at the finished model. Additional price restrictions are removed in the final effort to produce results that are closer to the correct answer.

The price feature enables clustering based on similarities in behaviour as well. By using this representation, the statistical analysis stage is followed by the newly formed groups. For each cluster, the probability distribution for observed paths has been computed. We learn a decision tree approach using prior data, visualising through several qualities to allocate new flight. To produce more precise forecasts, a version has also been taken into account that includes the initial points of the trajectory in the list of

predictor variables. Future data improvements should include pauses, prices from websites where travellers may buy flights, and data on price increases from 28 to 90 days.

- Motivation for the Problem Undertaken

Anyone who has purchased a plane ticket is aware of how costs may change suddenly. The price of the least costly ticket on a specific flight decreases with time. Raising prices on a flight that is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases is a common practise in an effort to maximise revenue based on time of purchase patterns (making sure last-minute purchases are expensive) and maintain the flight as full as they desire.

An airline business may be able to pre-plan flights and set reasonable prices for a route with the use of early demand forecasting. The majority of the current demand prediction models aim to forecast consumer demand for a single flight or route as well as the market share of a specific airline.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

This project consists of two phases- Data Collection Phase and Model Building Phase. In Data Collection phase, the flight ticket far

data is scraped from Expedia portal for the dates- 1st Feb 2023- 9th Feb 2023 from Hyderabad to the cities-Mumbai, Bangalore, Chennai, Kochi, Delhi and Ahmedabad using Web Driver and the data is exported in CSV format. The data is again imported for data analysis and model development and testing which gives an efficient model to predict the price of flight ticket fare.

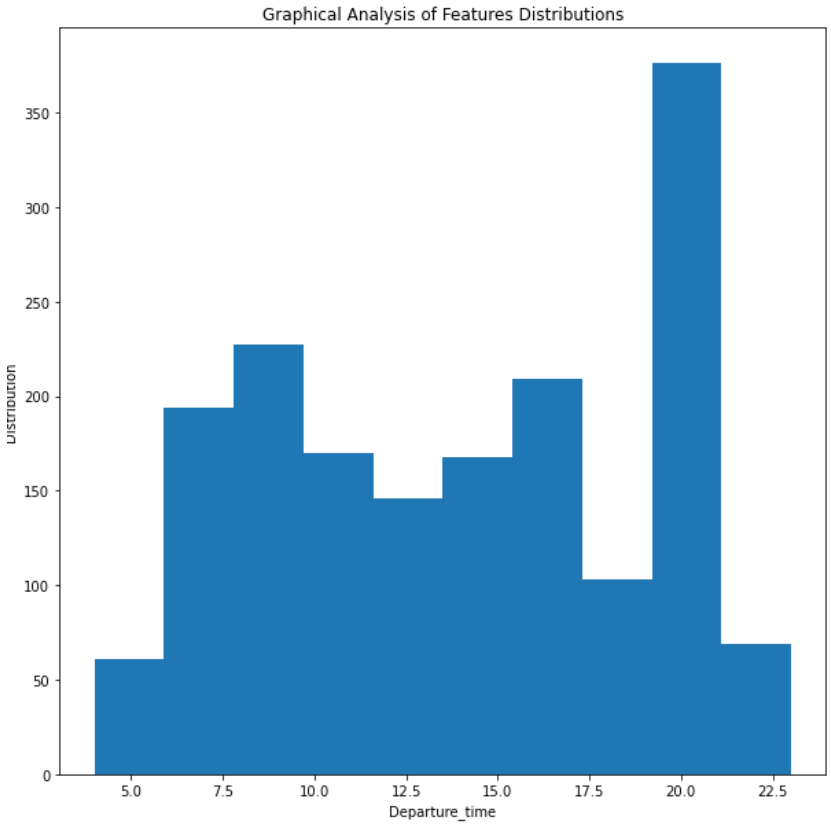
- Data Sources and their formats

For the data collection, data of flight ticket prices and other features are scraped from Expedia using Selenium Web Driver. The data of flight prices are scraped for flights from Hyderabad to the cities- Mumbai, Bangalore, Chennai, Kochi, Delhi, Ahmedabad for the dates 1st February 2023 to 9th February 2023. The parameters based on which the flight price data is collected are- Departure city, Arrival city, Departure time, Arrival time, Airline and layovers. In this project the target is the flight ticket price and the features are - Departure_time, Arrival_time, Departure_city, Arrival_city, Departure_date, Airline and layovers. After analysing the data, the following conclusions are made-

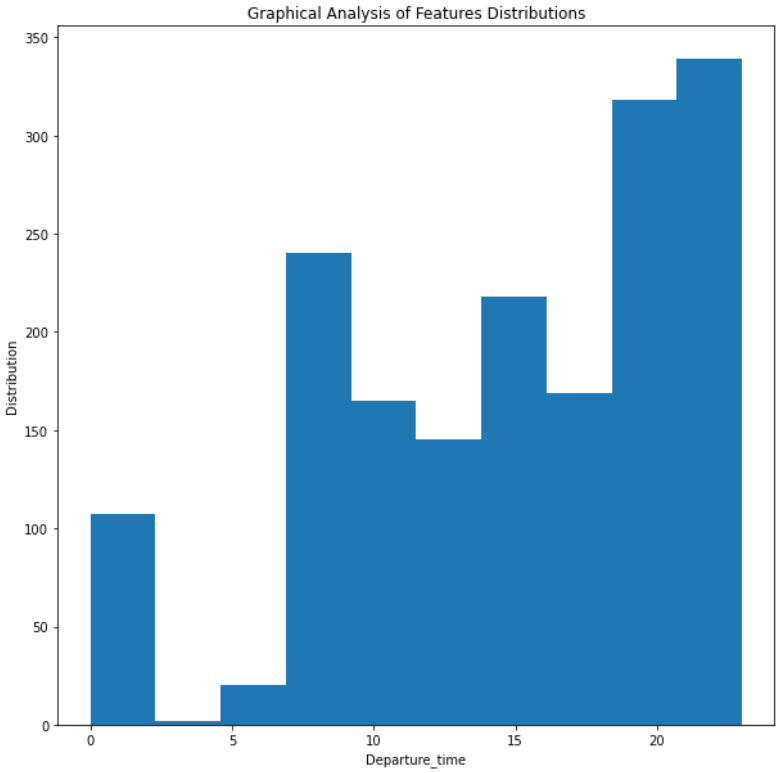
When graphical analysis was performed on the features of the data, the following points were observed:

- The dataset had 5346 rows and 8 columns.
- All the columns of the database are of object type, time, date and price column have to converted to integer type.

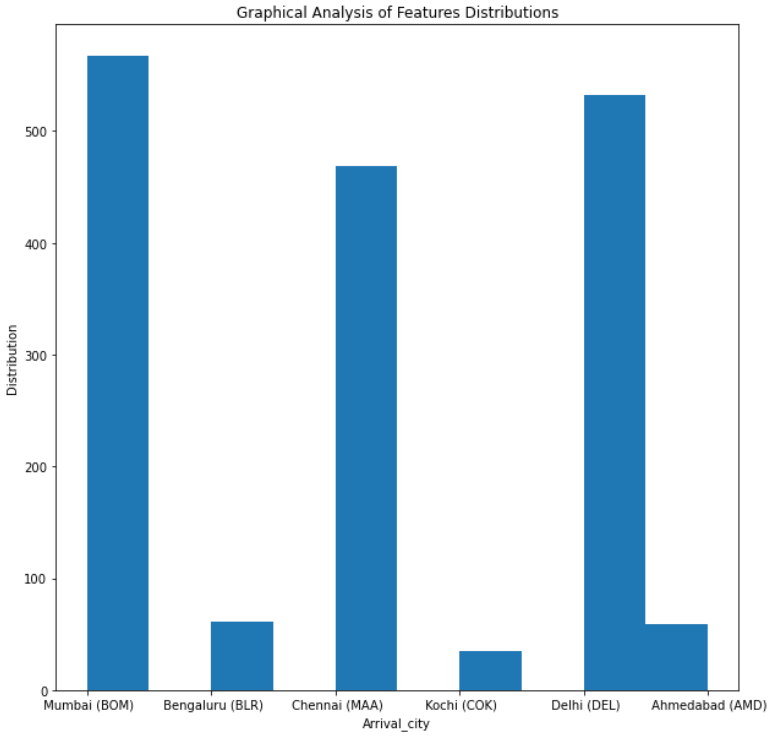
- Majority of the tickets are having departure time of 6-10 PM followed by 7-10AM and 4-5 PM.



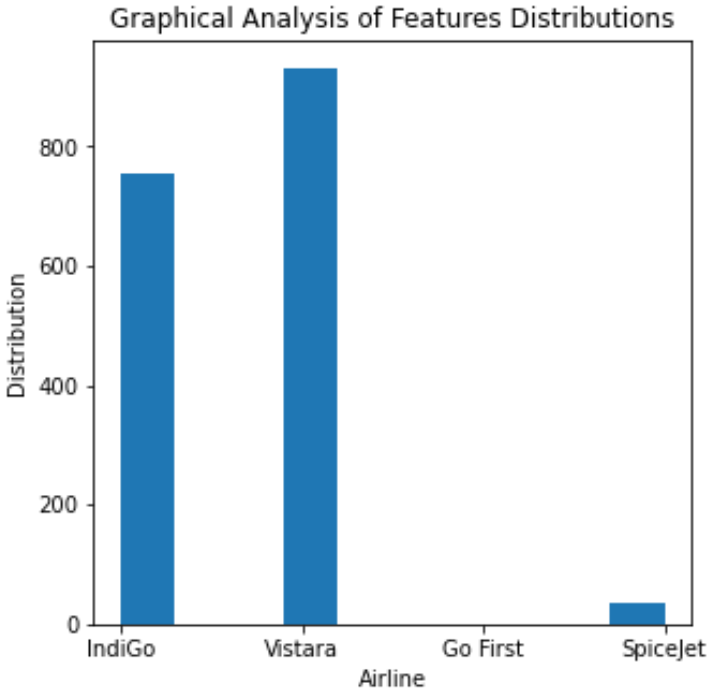
- Majority of the tickets are having arrival time of 6-9PM followed by 7-10AM.



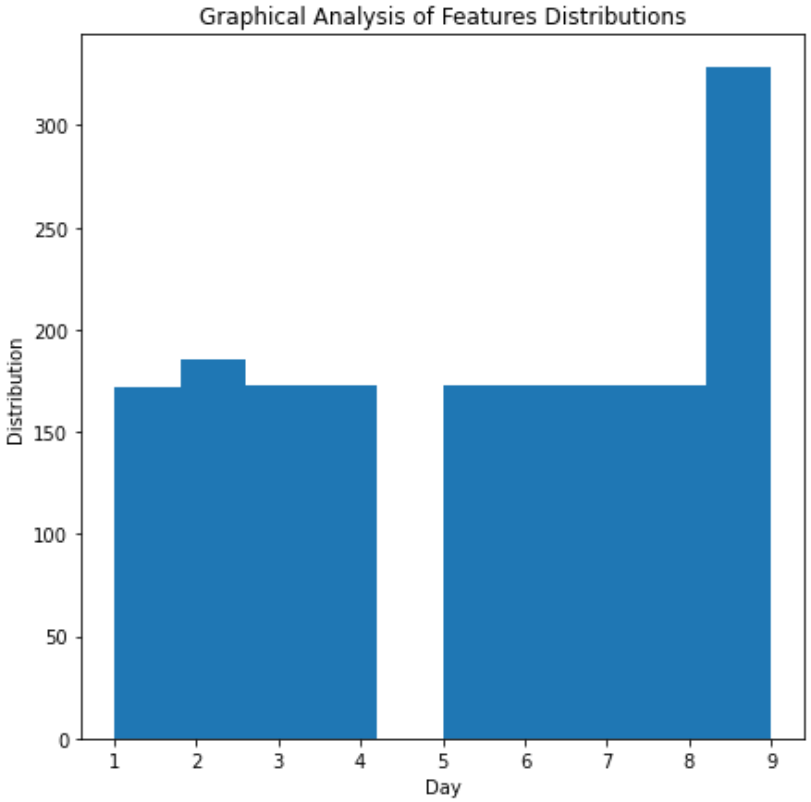
- The flights are departing from Hyderabad to the mentioned cities-Mumbai, Bangalore, Chennai, Kochi, Delhi and Ahmedabad.
- Most of the flights are arriving at Mumbai followed by Delhi and Chennai.



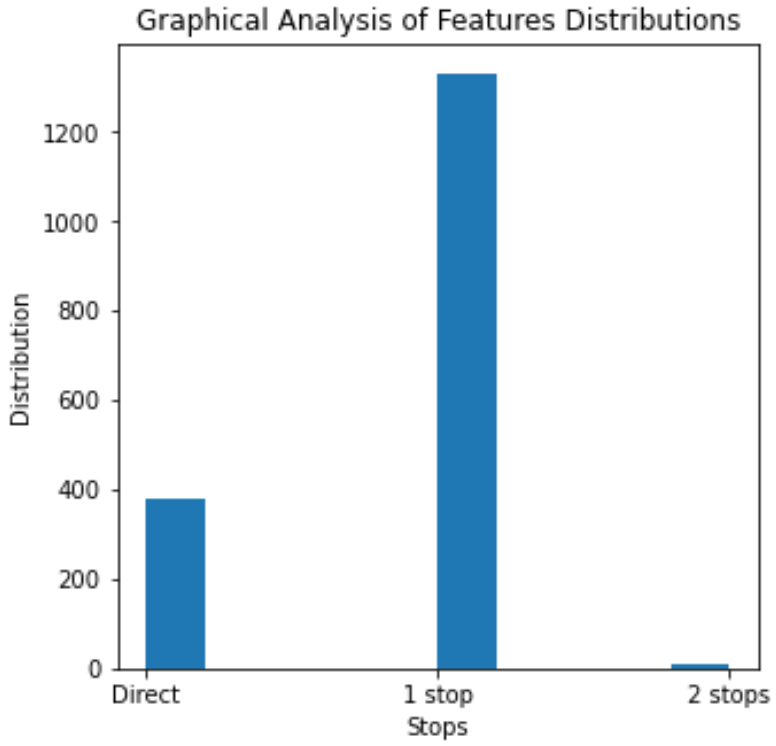
- Most of the flights are of airline-Vistara followed by Indigo and Spicejet and least is Go First.



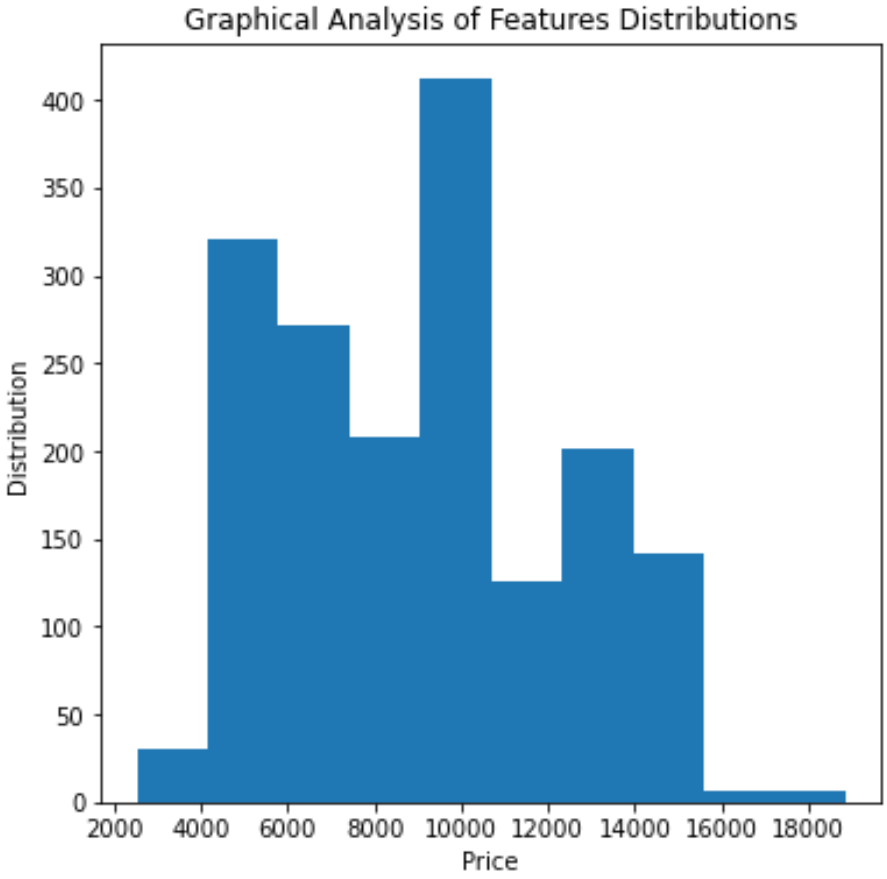
- The number of flights on 9th Feb 2023 is the highest followed by 2nd Feb.



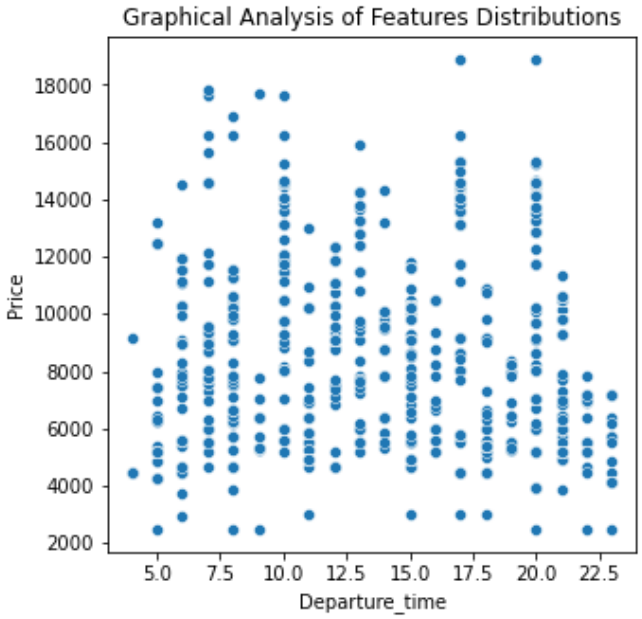
- 1300 flights are having one stop followed by 300 flights are direct.



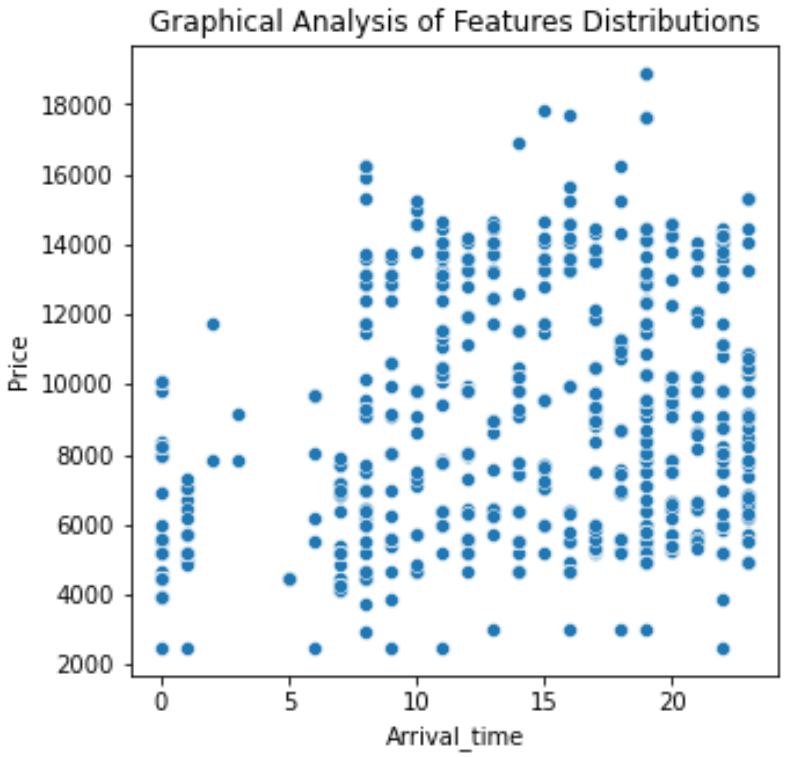
- The pricing of 430 tickets is between 9000-10000 followed by 320 tickets having a price range of 4000-5000.
- There are less tickets with price range of 15000-18000 and 2000-4000.



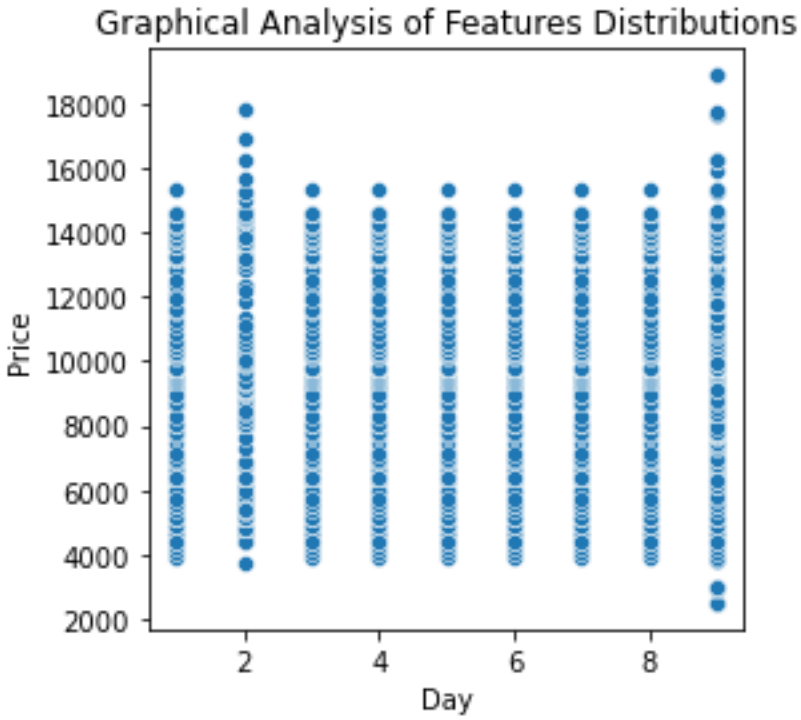
- Data Inputs- Logic- Output Relationships
  - The pricing is high for peak-time departure and price is comparatively less for early morning and late-night flights.



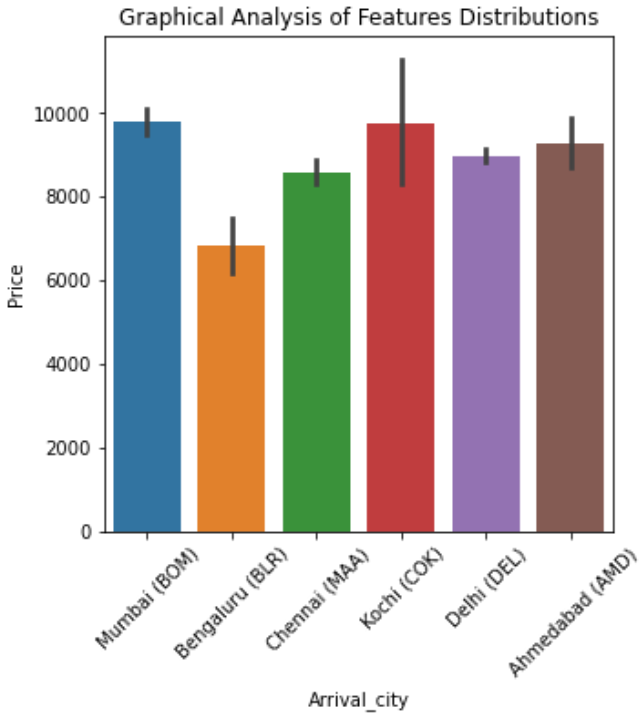
  - The pricing is less for mid-night and early morning(till 7 AM) arrivals.



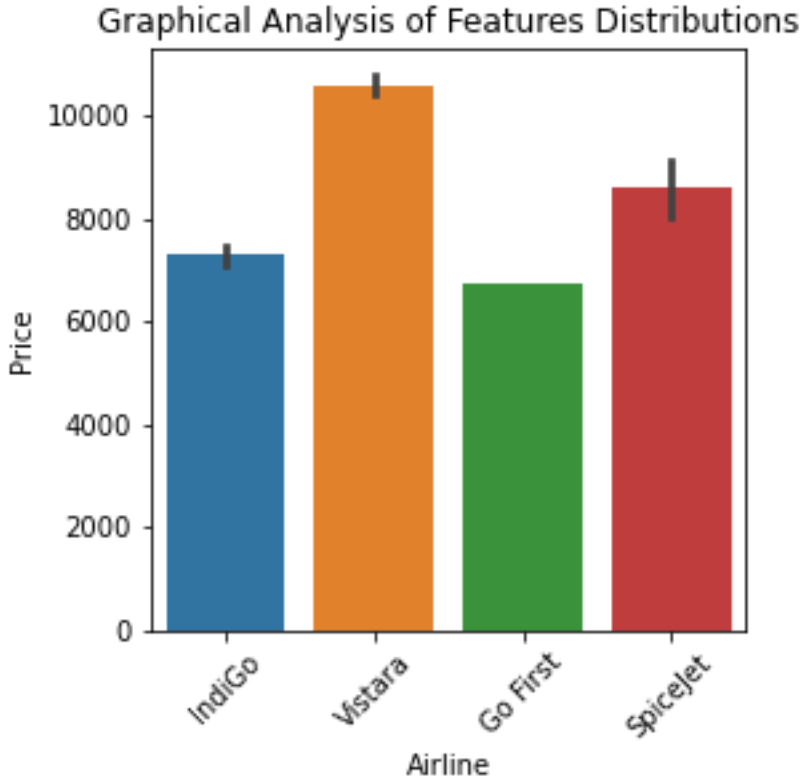
- 2nd and 9th Feb have the high-ticket fare.
- 1st,3rd,4th,5th,6th,7th and 8th have almost same pricing.



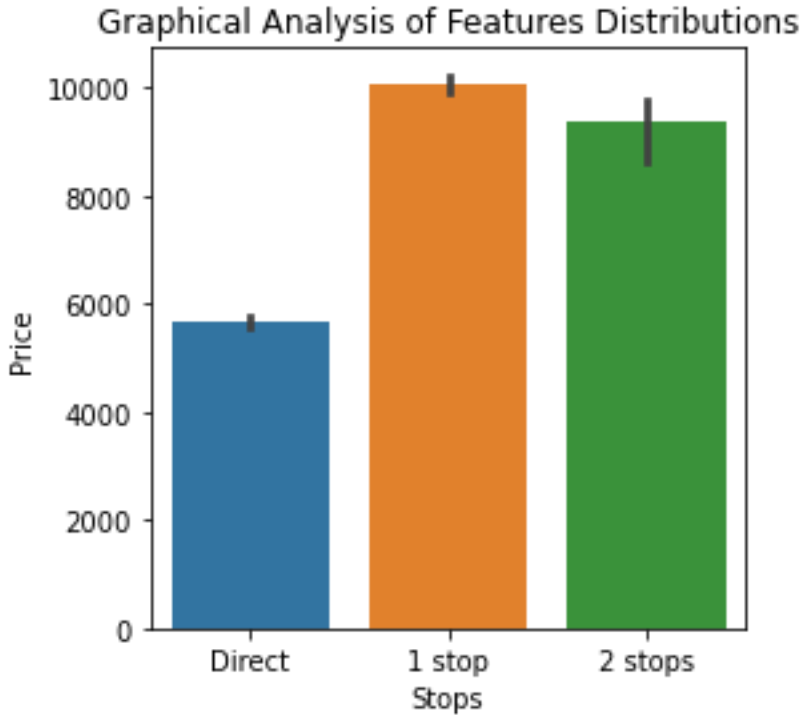
- From Hyderabad the flights to Mumbai and Kochi are higher price range followed by Ahmedabad, Delhi, Chennai and Bangalore.



- Vistara airlines have higher price range followed by Spiceject, Indigo and Gofirst.



- 1-stop flights are higher in price followed by 2-stops and direct flights are in less price range.



- Departure time and arrival time have the high positive correlation with price.

- Hardware and Software Requirements and Tools Used

Hardware requirements:

- Operating system- Windows 7,8,10
- Processor- dual core 2.4 GHz (i5 or i7 series Intel processor or equivalent AMD)
- RAM-4GB

Software Requirements

- Selenium Webdriver
- Python
- Pycharm
- Jupyter Notebook
- Chrome

1) Pandas- In pandas library the functions used are-
    - Pd.read function is used to load the data.
    - The DataFrame's information is printed via the info() method. The data includes the total number of columns, their labels, data kinds, memory use, range index, and the number of cells in each column (non-null values)
    - The DataFrame's data is described via the describe() method. If the DataFrame includes numerical data, each column's description will provide the following details, count is the total amount of non-empty values, mean – a measure of average value, the standard deviation, or std, min is the lowest value, 25 percent – the 25 percentile*, the fifty percentile is 50%, the 75% percentile* is at 75%, max is the highest value.
    - The isnull() method returns a DataFrame object where all the values are replaced with a Boolean value True for NULL values, and otherwise False.

- Pandas DataFrame.dtypes attribute return the dtypes in the DataFrame. It returns a Series with the data type of each column.
- DataFrame from Pandas. The DataFrame's dtypes are returned via the dtypes attribute. A Series with the data types of each column is what is returned.

2) Python scripts can be used to create 2D graphs and plots using the Matplotlib module. With features to control line styles, font attributes, formatting axes, and other features, it offers a module called pyplot that makes things simple for plotting.
3) A package called Seaborn uses Matplotlib as its foundation to plot graphs. In order to see random distributions, it will be used.
4) A free machine learning library for Python is called Scikit-learn. It supports Python's NumPy and SciPy libraries as well as a number of methods, including support vector machines, random forests, and k-neighbors.

Model/s Development and Evaluation

- Testing of Identified Approaches (Algorithms)

The features and target are split using the train_test_split function in sklearn, and a loop is defined to obtain the random state value for the function which yields best results.

The algorithms which were used in training and testing the model are:

1. Linear Regression
2. Random Forest Regression
3. Decision Tree Regression
4. K Neighbours Regression
5. Support Vector Regressor
6. Gradient Boosting Regressor
7. XGB Regressor
8. Ada Boost Regressor

- Run and evaluate selected models

  When the features and target were split using the train_test_split function and results were analysed for different random state values, the following observations were made:

  - Linear Regression model, Support Vector Machine and Adaboost Regressor have very low R2 score and explained variance score.
  - Random Forest Regressor and Decision Tree Regressor are having very good r2 score and explained variance score for random_state value of 2.
  - Random forest model has a cross-validation score of 0.93 and Decision Tree model has cross-validation score of 0.9.

- Key Metrics for success in solving problem under consideration

  The R2 score and Explained Variance Score were the two metrics employed in this research. An indicator of how much variance in a dependent variable is explained by one or more independent variables is the R-Squared statistic in a regression model. The fraction of the variability of a machine learning model's prediction is measured using the explained variance. It is, in a nutshell, the discrepancy between the anticipated value and the expected value. Understanding how much information we can lose when reconciling the dataset is a crucial subject.

- Interpretation of the Results

  The random forest model is selected as the efficient model for predicting the flight ticket fare, the model is further optimized using hyper-parameter tuning.

CONCLUSION

- Random Forest Model is the effective model for the case study since it has a good R2 score and explained variance score.