FLIGHT PRICE PREDICTION REPORT

INTRODUCTION

The Flight Price Prediction Project aims to create an intelligent system that predicts future flight prices based on a variety of factors. These may include historical flight price data, time of purchase, seasonality, flight duration, departure and arrival airports, airline, among other factors.

The goal of this project is to help consumers make more informed decisions when booking flights. It could potentially save them money and time, as they can determine the best time to book a flight.

This project will use machine learning and data analytics techniques to model and forecast flight prices. This involves training a predictive model on historical data, testing its performance, and deploying it for real-time price prediction.

The objective of the study is to analyse the flight booking dataset obtained from “Ease My Trip” website and to conduct various statistical hypothesis tests in order to get meaningful information from it.

The 'Linear Regression' statistical algorithm would be used to train the dataset and predict a continuous target variable. 'Easemytrip' is an internet platform for booking flight tickets, and hence a platform that potential passengers use to buy tickets. A thorough study of the data will aid in the discovery of valuable insights that will be of enormous value to passengers.

FEATURES

The various features of the cleaned dataset are explained below:
1) Airline: The name of the airline company is stored in the airline column. It is a categorical feature having 6 different airlines.
2) Flight: Flight stores information regarding the plane's flight code. It is a categorical feature.
3) Source City: City from which the flight takes off. It is a categorical feature having 6 unique cities.
4) Departure Time: This is a derived categorical feature obtained created by grouping time periods into bins. It stores information about the departure time and have 6 unique time labels.
5) Stops: A categorical feature with 3 distinct values that stores the number of stops between the source and destination cities.
6) Arrival Time: This is a derived categorical feature created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time.
7) Destination City: City where the flight will land. It is a categorical feature having 6 unique cities.
8) Class: A categorical feature that contains information on seat class; it has two distinct values: Business and Economy.
9) Duration: A continuous feature that displays the overall amount of time it takes to travel between cities in hours.
10)Days Left: This is a derived characteristic that is calculated by subtracting the trip date by the

booking date.
11) Price: Target variable stores information of the ticket price.

TECHNOLOGIES USED

The code trains several regression models, each with different algorithms and configurations. The models being trained are:

- **Linear Regression:**
  The code uses a Linear Regression model from a machine learning library (not explicitly mentioned in the code). Linear Regression is a simple and widely used statistical method for modeling the relationship between a dependent variable and one or more independent variables. In this case, it's being used to predict the output (y) based on the input features (X).
  The code I used creates and trains a Linear Regression model using the provided training data (X_train and y_train). It then evaluates the performance of the model on a separate test dataset (X_test and y_test) using some statistical measures. The results are collected and added to the existing data for further analysis or comparison with other models.
- **Decision Tree Regression:**
  A Decision Tree Regressor is a type of machine learning model used for regression tasks, where the goal is to predict continuous numeric values as outputs based on input features. It forms the basis for more sophisticated ensemble methods like Random Forests and Gradient Boosting.
- **Bagging Regressor:**
  Bagging Regressor is an ensemble learning method that improves the accuracy and robustness of regression models by combining predictions from multiple base regressors trained on different subsets of the training data. It is a powerful technique to reduce variance and enhance performance, making it popular in various regression applications.
- **Random Forest Regressor:**
  The Random Forest Regressor is an ensemble learning method that combines the predictions of multiple decision trees to perform regression tasks. It offers high accuracy, robustness, and feature importance analysis, making it a popular choice for a wide range of regression applications in machine learning.
- **XGBoost Regressor:**
  XGBoost Regressor, short for Extreme Gradient Boosting Regressor, is an advanced and powerful implementation of the Gradient Boosting algorithm for regression tasks. It is a widely used machine learning model known for its high predictive accuracy and efficiency.
- **K Neighbors Regressor:**
  K Neighbors Regressor is a simple and intuitive algorithm for regression tasks. It predicts continuous numeric values by averaging the output values of the K nearest neighbors in the feature space. While it has advantages such as simplicity and non-linearity, it may be computationally expensive for large datasets and requires careful data preprocessing to handle varying feature scales.
- **Extra Trees Regressor:**
  Extra Trees Regressor is an ensemble learning method based on randomized decision trees. It reduces the bias of individual decision trees by further randomizing the feature selection

process. The ensemble of Extra Trees provides high predictive accuracy and is computationally efficient. It is particularly useful when dealing with high-dimensional datasets and can be a valuable option for regression tasks in machine learning.

- **Ridge Regression:**
  Ridge Regression, also known as L2 regularization, is a linear regression technique used to handle multicollinearity and prevent overfitting in a linear regression model. It is an extension of the ordinary least squares (OLS) regression, where the goal is to predict continuous numeric values based on input features. Ridge Regression introduces a penalty term to the regression cost function, which helps to control the model's complexity and makes it more robust.
- **Histogram Gradient Boosting Regressor:**
  Histogram Gradient Boosting Regressor is an efficient and scalable ensemble algorithm for regression tasks. It uses histogram-based algorithms and weighted quantile sketches to improve efficiency while providing competitive performance compared to traditional Gradient Boosting algorithms. Its ability to handle large datasets and missing values makes it a valuable option for regression problems, particularly when efficiency and scalability are essential considerations.
- **Lasso Regression:**
  Lasso Regression is a linear regression technique that introduces L1 regularization to promote sparsity and feature selection in the model. It is particularly useful when dealing with datasets with many features and can automatically identify and keep only the most important features. Lasso Regression is widely used in machine learning when interpretability and feature selection are crucial considerations.
- **Gradient Boosting Regressor:**
  Gradient Boosting Regressor is a powerful and popular ensemble machine learning algorithm used for regression tasks. It is an extension of the Gradient Boosting algorithm, which combines the predictions of multiple individual models (typically decision trees) to make a final prediction.

CHALLENGES FACED

In this flight price prediction project, some of the challenges that may arise include:

Data Availability: Obtaining reliable and comprehensive historical flight data can be challenging, especially if you need data from various airlines and routes.

Data Quality: Ensuring the accuracy and consistency of the collected data can be difficult, as flight prices are influenced by numerous factors, such as seasonality, demand, and external events.

Feature Engineering: Selecting and engineering the right features that capture the relevant information for price prediction can be complex and time-consuming.

Dynamic Nature of Prices: Flight prices fluctuate frequently due to various factors, making it challenging to build a model that accurately captures these changes.

Model Selection: Choosing the appropriate machine learning algorithm and model architecture can be tricky, as different models may perform differently on flight price prediction tasks.

Overfitting: Avoiding overfitting is crucial to ensure the model generalizes well to unseen data, given the complexity of the flight price prediction problem.

Limited Historical Data: Having limited historical data might impact the model's ability to learn patterns effectively, especially if you're trying to predict prices for less popular or new routes.

External Factors: Various external factors, such as geopolitical events or natural disasters, can significantly influence flight prices, and incorporating these factors into the model can be challenging.

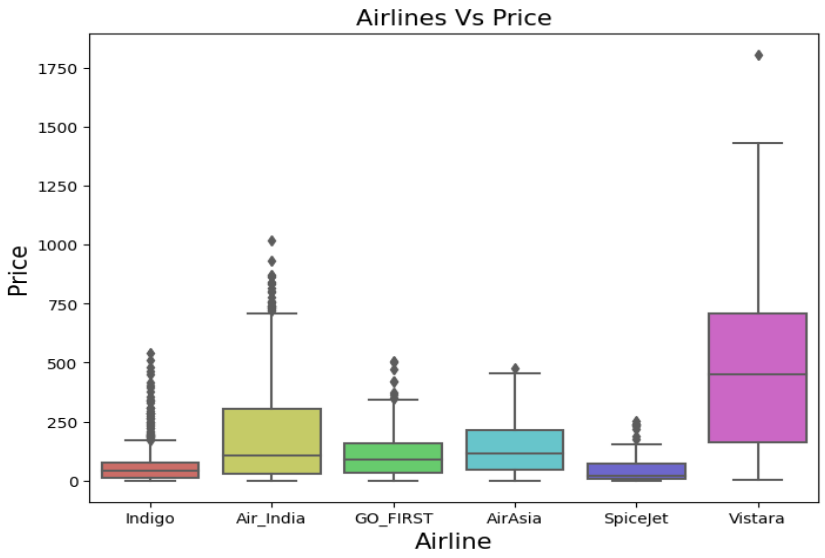
Interpretability: Building a model that is interpretable and understandable to users is important, especially if it is used in a consumer-facing application.

Real-Time Updates: If the prediction model is intended for real-time applications, ensuring it can handle real-time updates and respond to price changes promptly can be a challenge.

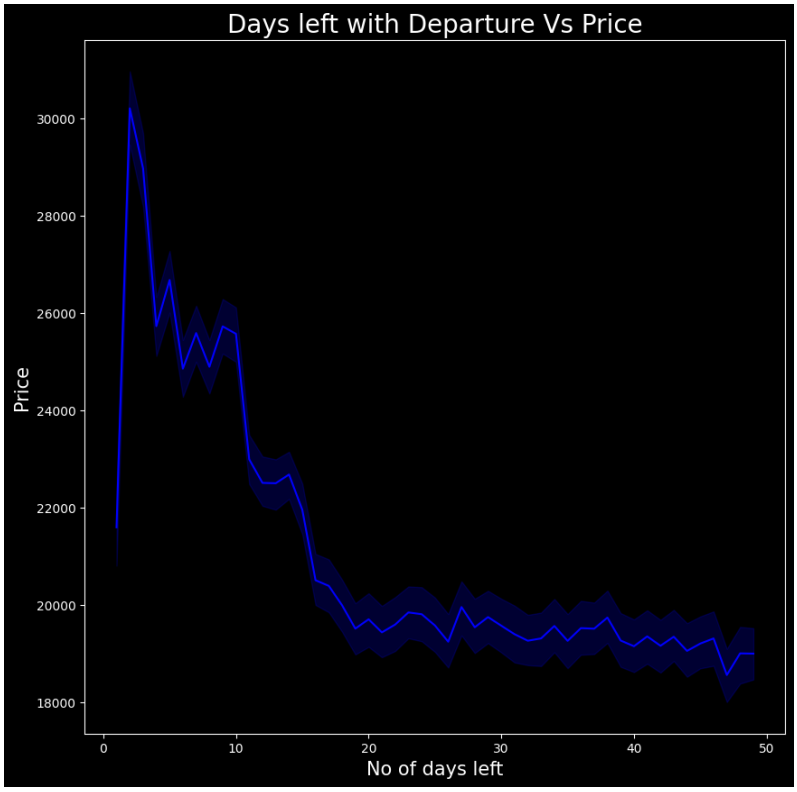
Addressing these challenges often requires a combination of data preprocessing, feature engineering, model selection, and continuous monitoring and updating of the model to maintain its accuracy and relevance over time.

FINAL OUTCOME

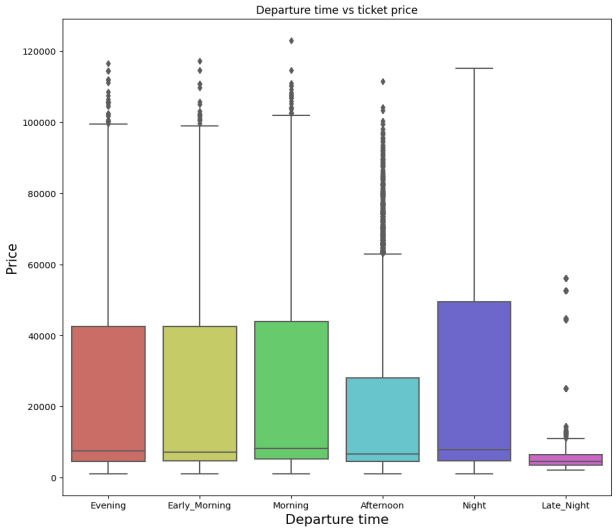
The aim of our study is to answer the below research questions:
a) Does price vary with Airlines for the same source_city to destination_city?
ANS- Yes, the price can vary based on different airlines operating between the same source and destination cities.

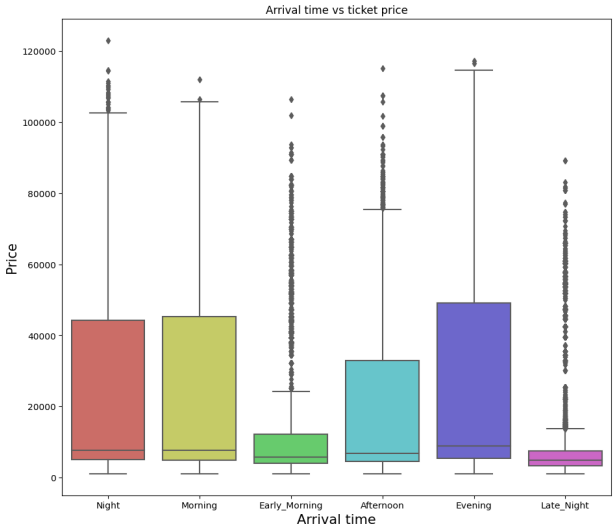


b) How is the price affected when tickets are bought just 1 or 2 days before departure?
ANS- Typically, ticket prices tend to be higher when purchased closer to the departure date due to increased demand and limited availability.



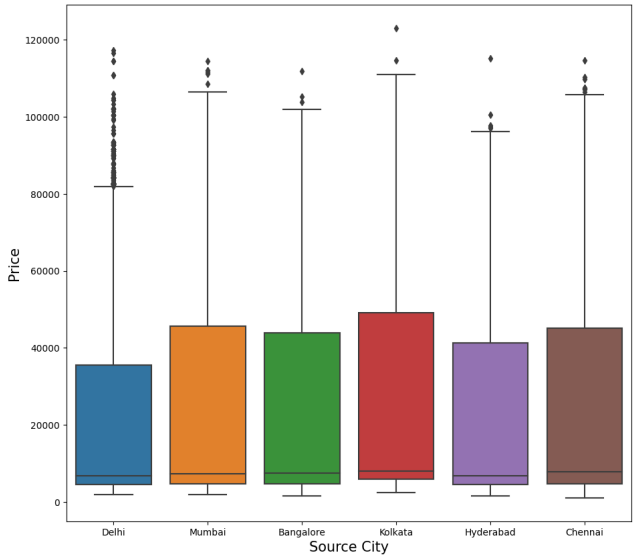
c) Does the ticket price change based on the departure time and arrival time?
ANS- Yes, the departure time and arrival time can impact ticket prices, as certain time slots may be more popular or in higher demand, resulting in price fluctuations.

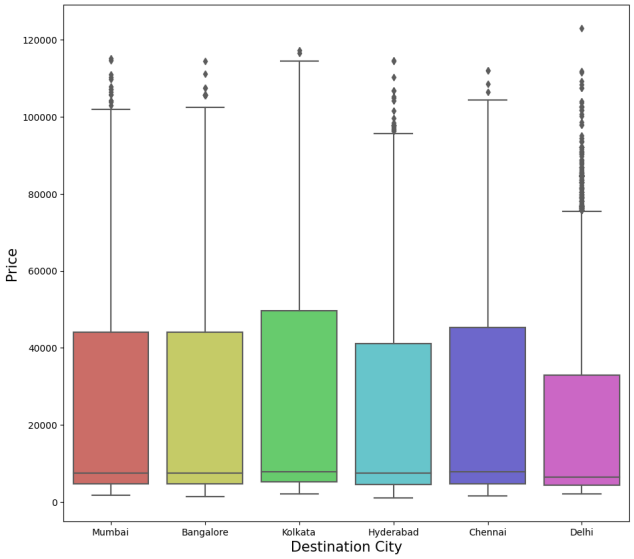




d) How the price changes with change in Source and Destination?

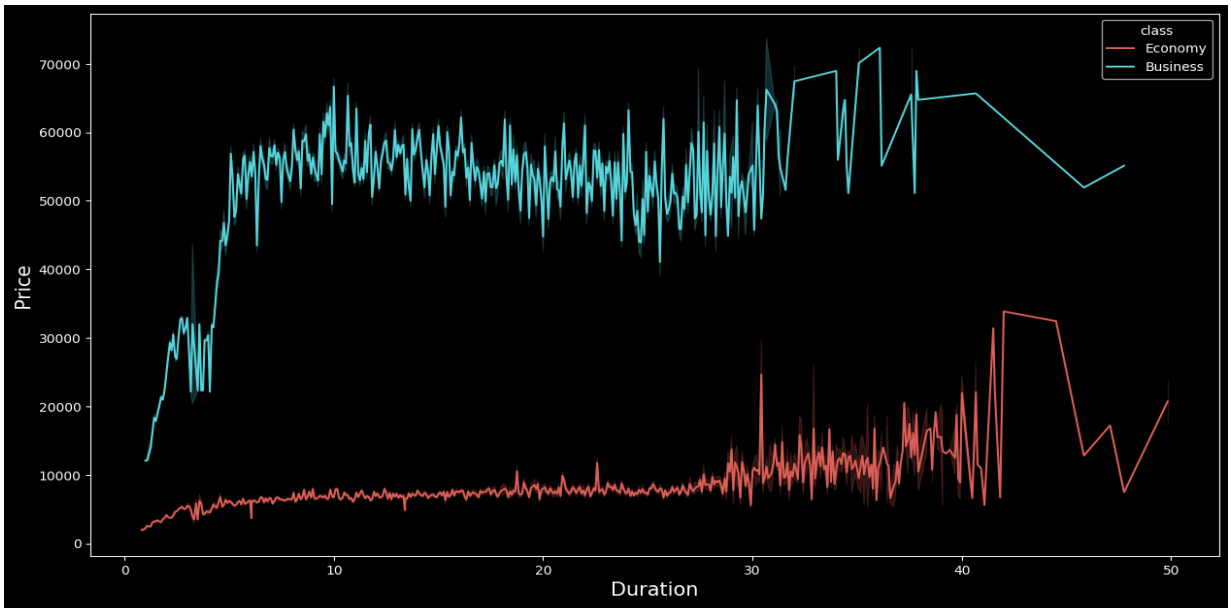
ANS- A Flight price prediction application which predicts fares of flight for a particular date based on various parameters like Source, Destination, Stops & Airline. In short, distance and demand are pivotal factors that determine flight ticket prices. Travel is just like any other valuable commodity, and it's priced so that when demand is low, prices are low, to stimulate sales. Likewise, when demand is high, prices are high to capitalize on the interest.





e) How does the ticket price vary between Economy and Business class?

ANS- For Business Class, the price increase with flight duration can be more significant. Business Class offers enhanced amenities and services, such as more spacious seating, better meals, and additional perks. These added luxuries contribute to the higher cost of Business Class tickets on longer flights



PRACTICAL USAGE

Practical usage of flight price prediction can benefit both travelers and the travel industry. Here are some practical applications:

1. **Optimal Booking Time:** Flight price prediction can help travelers determine the best time to book their tickets. By analyzing historical data and pricing trends, prediction models can suggest the optimal time to secure the lowest fares[1].

2. **Budget Planning:** Knowing the expected price range of flights in advance allows travelers to plan their travel budget more effectively. Price prediction tools can provide insights into future price fluctuations, enabling travelers to make informed decisions[4].

3. **Promotions and Discounts:** Airlines and travel agencies can utilize flight price prediction to offer targeted promotions and discounts. By identifying periods of low demand or predicting price drops, they can attract more customers and optimize revenue[5].

4. **Revenue Management:** Airlines can use price prediction models to optimize their revenue management strategies. By forecasting demand and adjusting prices accordingly, airlines can maximize their seat occupancy and profitability[6].

5. **Competitive Analysis:** Price prediction tools can help travel agencies and airlines analyze their competitors' pricing strategies. By monitoring and comparing prices, they can adjust their own pricing to stay competitive in the market[1].

6. **Market Insights:** Flight price prediction can provide valuable market insights to airlines and travel industry stakeholders. By analyzing historical data and trends, they can gain a better understanding of customer behavior, market dynamics, and pricing patterns[6].

It is important to note that while flight price prediction can provide valuable guidance, it is not a guarantee of the exact ticket price. Factors such as sudden changes in demand, external events, and market conditions can still impact prices. However, utilizing prediction tools can increase the chances of finding the best deals and optimizing travel plans.