Overview

Flight ticket prices are often seen as unpredictable, fluctuating significantly within short time spans. However, using data science techniques, we aim to demonstrate that given the right data, these prices can indeed be predicted. This report analyzes flight ticket prices for various airlines between March and June 2019, covering multiple routes and factors influencing prices.

Dataset

Training set: 10683 records
Test set: 2671 records

Features

Airline: The name of the airline.
Date_of_Journey: The date of the journey.
Source: The origin of the flight.
Destination: The destination of the flight.
Route: The flight path taken.
Dep_Time: Departure time.
Arrival_Time: Arrival time.
Duration: Total flight duration.
Total_Stops: Number of stops during the flight.
Additional_Info: Miscellaneous information about the flight.
Price: Ticket price (target variable).

Data Preprocessing

Data Loading

The dataset was loaded using pandas

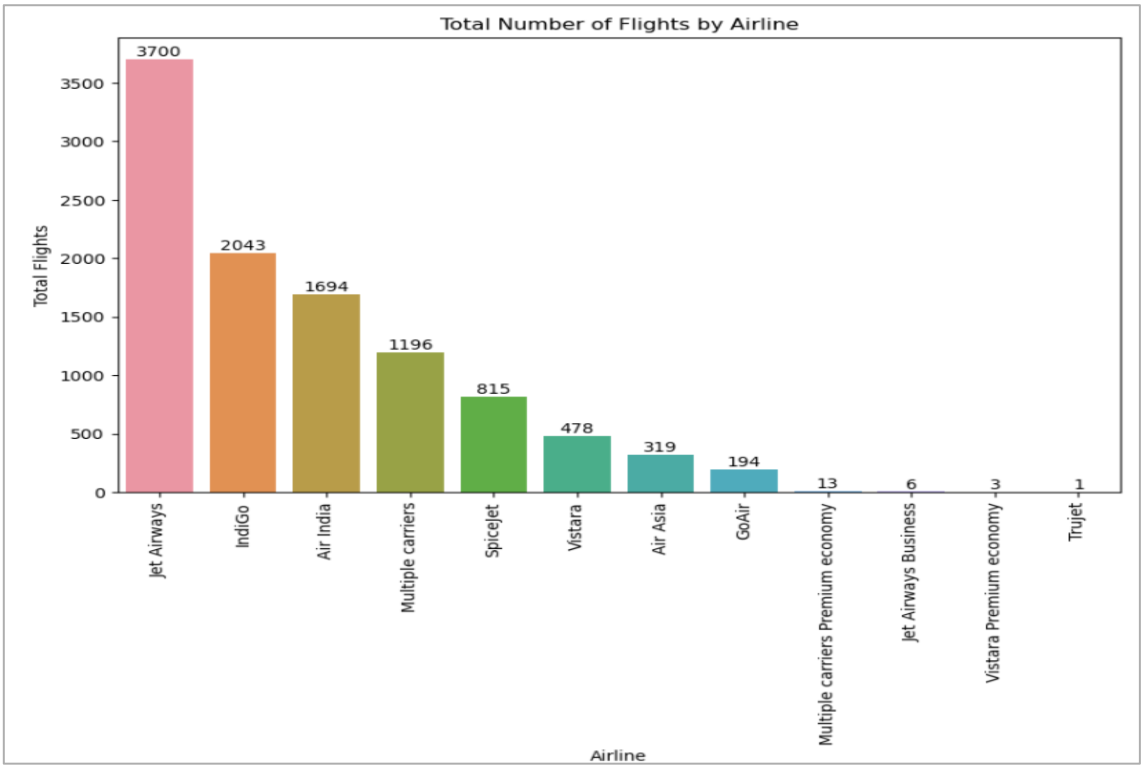
Data Cleaning and Feature Engineering

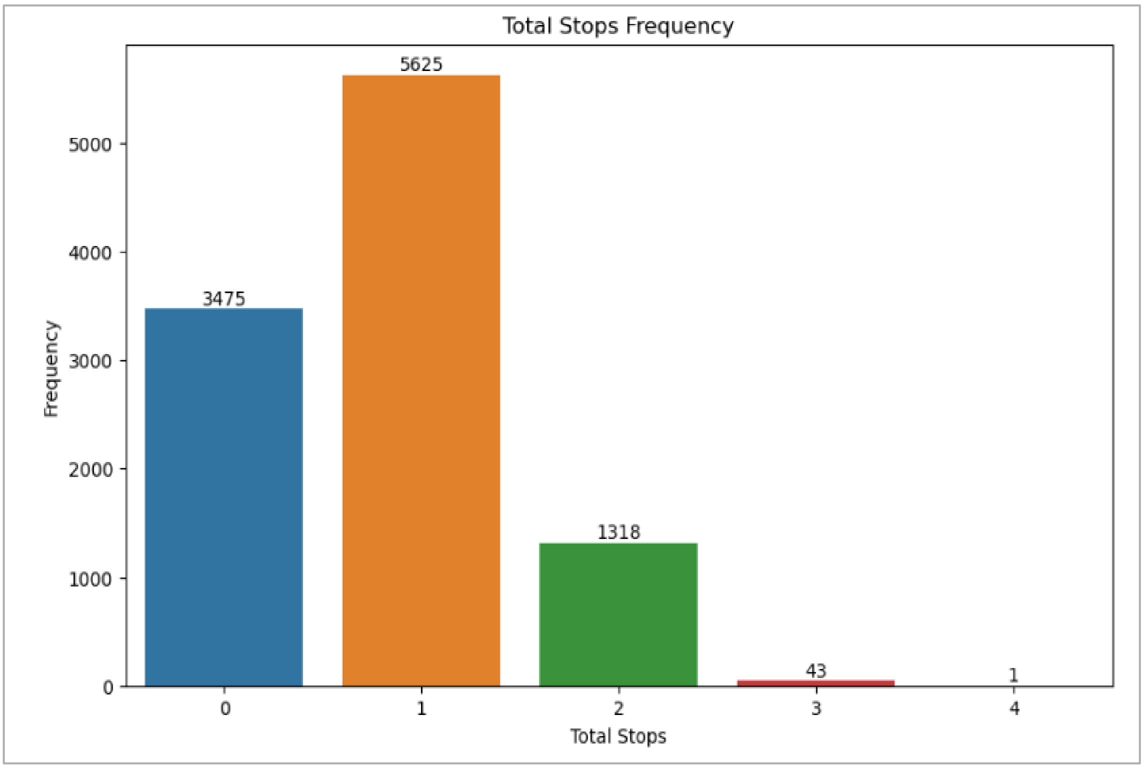
- Date and Time Features
    - ⇒ Date_of_Journey: Split into Day, Month, and Year.
    - ⇒ Dep_Time: Converted into Dep_hour and Dep_min.
    - ⇒ Arrival_Time: Split similarly into Arrival_hour and Arrival_min.
    - ⇒ Duration: Parsed to extract hours and minutes.
- Categorical Features
    - ⇒ Airline, Source, Destination: Converted into categorical variables using one-hot encoding.
- Stops
    - ⇒ Total_Stops: Converted from string to integer.

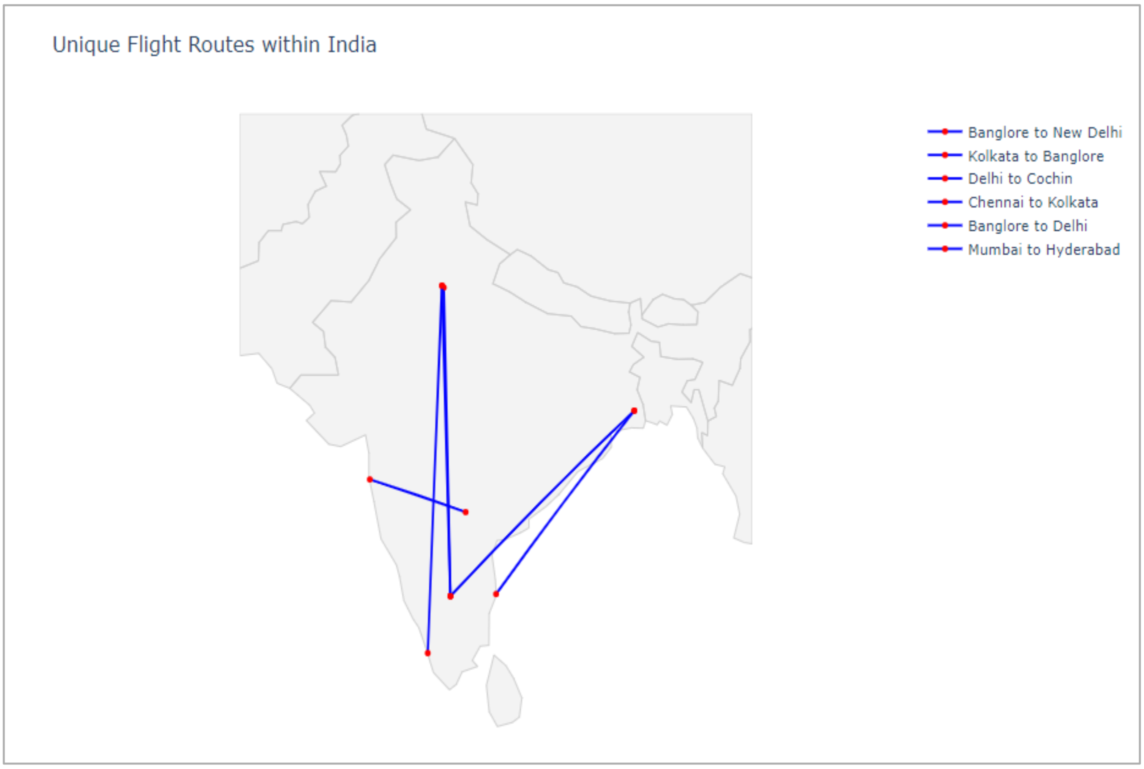
Final Data Preparation

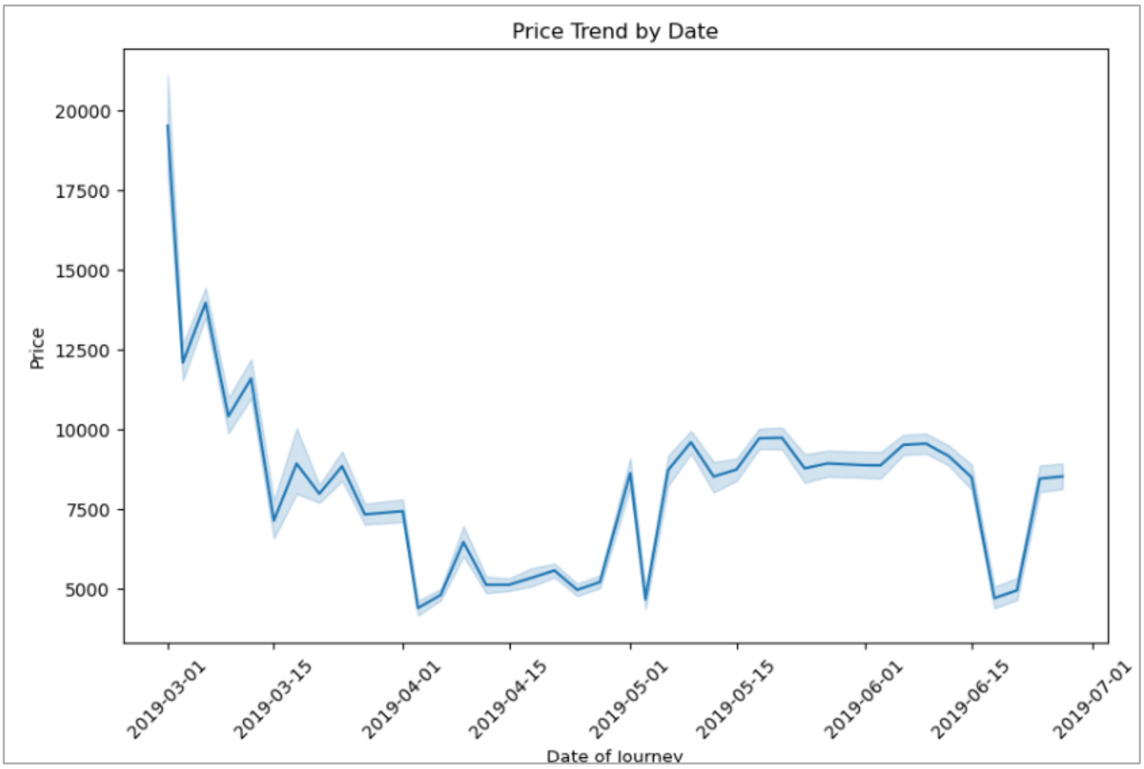
Combining all features into a final dataframe.

Visualizations






Unique Flight Routes within India
Banglore to New Delhi
Kolkata to Banglore
Delhi to Cochin
Chennai to Kolkata
Banglore to Delhi
Mumbai to Hyderabad


Price Trend by Date
Price
20000
17500
15000
12500
10000
7500
5000
2019-03-01
2019-03-15
2019-04-01
2019-04-15
2019-05-01
2019-05-15
2019-06-01
2019-06-15
2019-07-01
Date of Journey

- **Observation:**
  - ⇒ Jet Airways seems to be the airline with the most number of flights offered, followed by IndiGo and Air India.
  - ⇒ Jet Airways Business has the highest price distribution(Avg - ₹58358), while Trujet having the lowest(Avg - ₹4140).
  - ⇒ The total stop frequency in the India appears to be higher for flights with zero or one stop compared to flights with more than two stops.
  - ⇒ The price trend fluctuated a lot, with the greatest price of over ₹20,000 in March and the lowest price of approximately ₹4,000 in April.
  - ⇒ The maximum number of flights(4345) are from Delhi to Cochin and minimum number of flights(381) from Chennai to Kolkata.

Feature Engineering

New features were created, and unnecessary columns were dropped after encoding. Correlation between features was also checked to understand their relationships.

Feature Importance

Using feature importance analysis, Total_stops was identified as the most significant feature,
followed by Day and Airline_Jet_Airways

Model Building

Several regression models were tested:

1. Linear Regression
2. Polynomial Regression
3. Ridge Regression
4. Decision Tree Regression
5. Gradient Boosting Regression
6. XGBoost Regression
7. Random ForestRegression

Model Performance

Model	MAE	MSE	RMSE	R2 Score
Linear Regression	1998.639306	8.68E+06	2945.796236	0.583808
Polynomial Regression	1682.40661	6.09E+06	2468.386729	0.707777
Ridge Regression	1999.825146	8.67E+06	2944.926607	0.584054
Decision Tree Regression	1405.1077	5.97E+06	2442.734247	0.713819
Gradient Boosting Regression	1537.088597	5.15E+06	2268.34933	0.753221
XGBoost Regression	1132.428776	3.39E+06	1841.272342	0.837399
Random Forest Regression	1171.100148	3.86E+06	1965.544611	0.814709

Key Findings

Duration and Total_Stops are significant predictors of flight prices.
Airline, Source, and Destination categorical variables also contribute substantially.
XGBoost Regression provides the best result

Conclusion

The analysis demonstrates that flight ticket prices can be effectively predicted using machine learning models. Among the tested models, the XGBoost Regression performed the best, making it a suitable choice for predicting flight prices based on the given features.

Future Work

Incorporate additional features such as seasonality and demand-supply factors.
Explore deep learning models for potentially better performance.
Real-time price prediction system integration.