**Flight Air Price Prediction**

A Machine Learning Approach

CS 668-Analytics Capstone Project

FINAL PROJECT

**SHANKAR NAYAK LAVUDYA**

# Abstract

The airline industry is ever changing, fluctuation in air prices due to multiple factors such as seasons, demand, fuel costs, and geopolitical events. A precise and accurate prediction of air ticket prices is crucial for both travellers and airlines to optimise their travel planning and revenue management. This study proposes a machine learning-based approach to predict air ticket prices, using the historical pricing data and relevant features. The methodology involved here is data processing, feature engineering and implementation of multiple machine learning models such as SVM, regressions based models like XGB, decision trees and random forest. These models capture the complex relationships influencing ticket prices. The dataset includes information about the flight routes, departure times, airlines, historical pricing, and other relevant factors.A very important analysis called feature analysis is conducted to identify the key drivers affecting ticket prices. After the analysis all the above mentioned machine learning models are used to build and evaluate predictive models. These models are then validated on historical data, and their performance is assessed based on metrics such as mean absolute error and root mean square error. This study and its results aim to provide deep insights into the factors influencing air ticket prices and to develop a reliable prediction model. Model that is being proposed here can contribute to better decisions making for both the airlines and the travellers which are the customers. Assistiance in price optimisation strategies, demand forecasting, can enhance the overall efficiency of the air travel industry and revolutionise its travel. Also this research can help us understand deeper into the complexities that involve int he airfare dynamics and serve as a founding stone for the future enhancements in the field of air ticket price prediction.Keywords—Air Ticket Price Prediction, Machine Learning, Regression Models, Feature Engineering, Historical, Pricing Data, Predictive Modelling, Feature Importance Analysis, Flight Routes, Seasonality, Air Travel Industry, Pricing Dynamics.

# I. INTRODUCTION

1. ***Background and Motivation:***

The airline industry is being characterised by its high pricing mechanism this makes the industry face constant challenges in adapting to dynamic nature that influence to the price of a flight. The air ticket prices are vulnerable all many changes which include the demand, seasonal variations, fuel costs, and geopolitical events. Due to all the mentioned reasons, a very reliable solution is required for the air ticket price prediction and hence this model becomes very evident. Travellers seek to optimise their travel expenses, on the other hand the airlines work hard to implement effective revenue management strategies.This study is motivated by the overarching goal of addressing all the complexities the hinder the prediction of airfare And try to contribute to more optimised and efficient air travel ecosystem.

1. ***Problem Statement:***

We humans have had so many significant technical and scientific advancements but predicting air ticket prices still remains one of the more complex tasks due to the vary unstable nature of the factors that influence it. The traditional pricing models often fall short in capturing the Complexities between the demand, other external events and operational costs. Hence my research identifies the need for a better and a robust, adaptable prediction model that can identify intricate details of the airline industry. The lack of accurate airfare predictions not only hampers travellers' ability to plan cost-effective journeys but also poses challenges for airlines in optimising revenue streams. So, by addressing such issues, this study will aim to provide a valuable contribution to the field of air ticket prices prediction.

# II. RELATED WORK

There has always been intensive research done as to find the most accurate system to predict air ticket price prediction with various methods and trying different approaches that I have explored in the literature survey. The earlier studies on this matter mostly relied on the traditional statistics model and attempted to predominantly correlate historical pricing and its trends with external factors And here came the issue, these models failed to adapt to the dynamic nature of the airline industry. Recent advancements in machine learning have made a new wave to research possible and with this shift towards more sophisticated predictive models capable of handling vast and complex datasets.

# III. METHODOLOGY

1. ***Data Collection:***

The dataset used in my study is a very detailed dataset that is derived from a very comprehensive compilation of all the historical airfare data and capturing a diverse range of flight routes, departure times, airlines, and associated pricing information. This dataset also spans a significant time frame, allowing for a very neat and detailed analysis of seasonal variations, market trends and the long-term patterns.Additionally, complementary datasets that have detailed external factors such as geopolitical events, economic indicators and fuel costs are integrated to provide a complete and holistic view of all the variables that can cause an influence in the air ticket prices. Also, the dataset is rigorously checked for data quality to ensure the reliable and accurate nature of it. With any anomalies or any mossing values addressed throughout appropriate measures

To extract data from the websites, Octoparse scraping tool was used. This data was collected in two parts. One was for the ecomomy class tickets and the other one was the business class tickets. And a total of 300261 distinct flights booking option data was extracted from these sites. This data was collected over a span of 50 days, from Feb 11 to march 31 2022. Another dataset was also created as a secondary source of data from ease my trip website. These datasets contained information about flight booking options from the website ease my trip for flight travel between India's top 6 metro cities. There are 300261 datapoints and 11 features after cleaning the dataset.

1. ***Data Pre - Processing:***

Data preprocessing is a very important phase for effective analysis of the dataset. It involved a series of steps that aimed to make a clean dataset which has refined data that can enhance the quality of subsequent model training. In the data preprocessing phase only the outliers are also identifies and addressing using robust statistical techniques to prevent from the final results to produces any deviance or get influenced by the outliers.Missing values were imputed judiciously, looking at the nature of the data and the potential of impact on predictive accuracy. Also data normalisation techniques were applied to bring features to a consistent scale, clearing all the issues associated with different ranges of numerical values. This resulted preprocessed dataset serves as a foundation for various feature engineering and for the model training, ensuring that the predictive models operate on a reliable and standardised input.

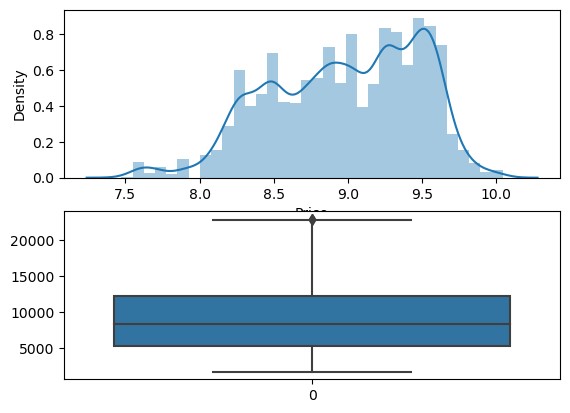
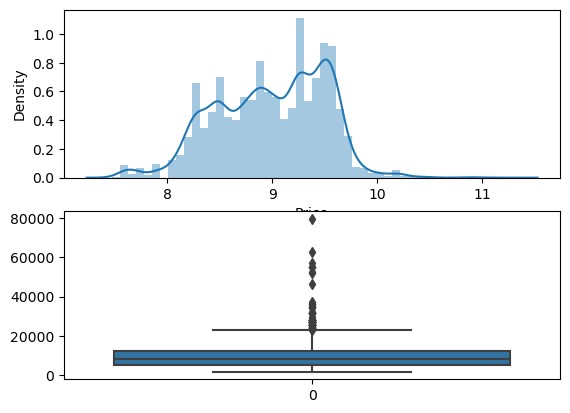
***1)Features:***

The following describes the different attributes of the cleaned dataset:Airline: The airline column contains the name of the airline company. This feature falls into a specific category and has six airlines.Flight: Flight keeps track of the flight code of the aircraft. It falls within a category.Source City: The city where the aircraft departs. It is a category characteristic with six distinct cities. Departure Time: By classifying time intervals into bins, this derived categorical feature was produced. It has six distinct time labels and keeps information about the departure time. Stops: A categorical characteristic with three unique values that stores the number of stops between the source and destination cities. Class: A categorical characteristic that gives information about seat class; it has two separate values: Business and Economy. Duration: A continuous feature that displays the total length of time it takes to travel between cities in hours. Days Left: This is a derived feature determined by dividing the travel date by the booking date. Target variable holds information about the ticket price.

***2). Checking, handling outliers and categorical data:***

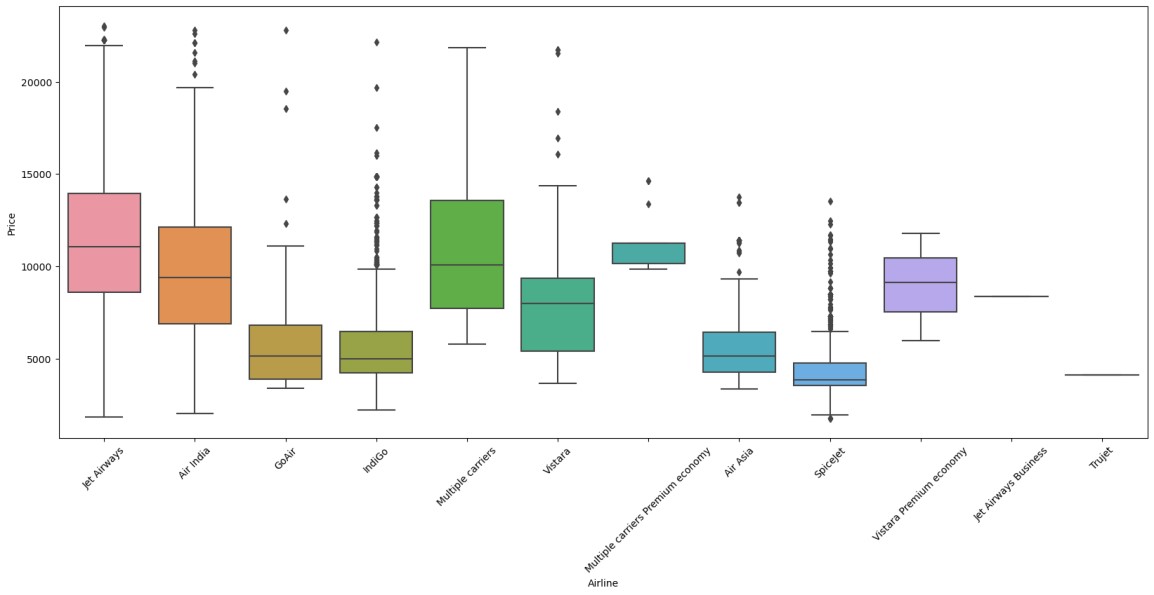
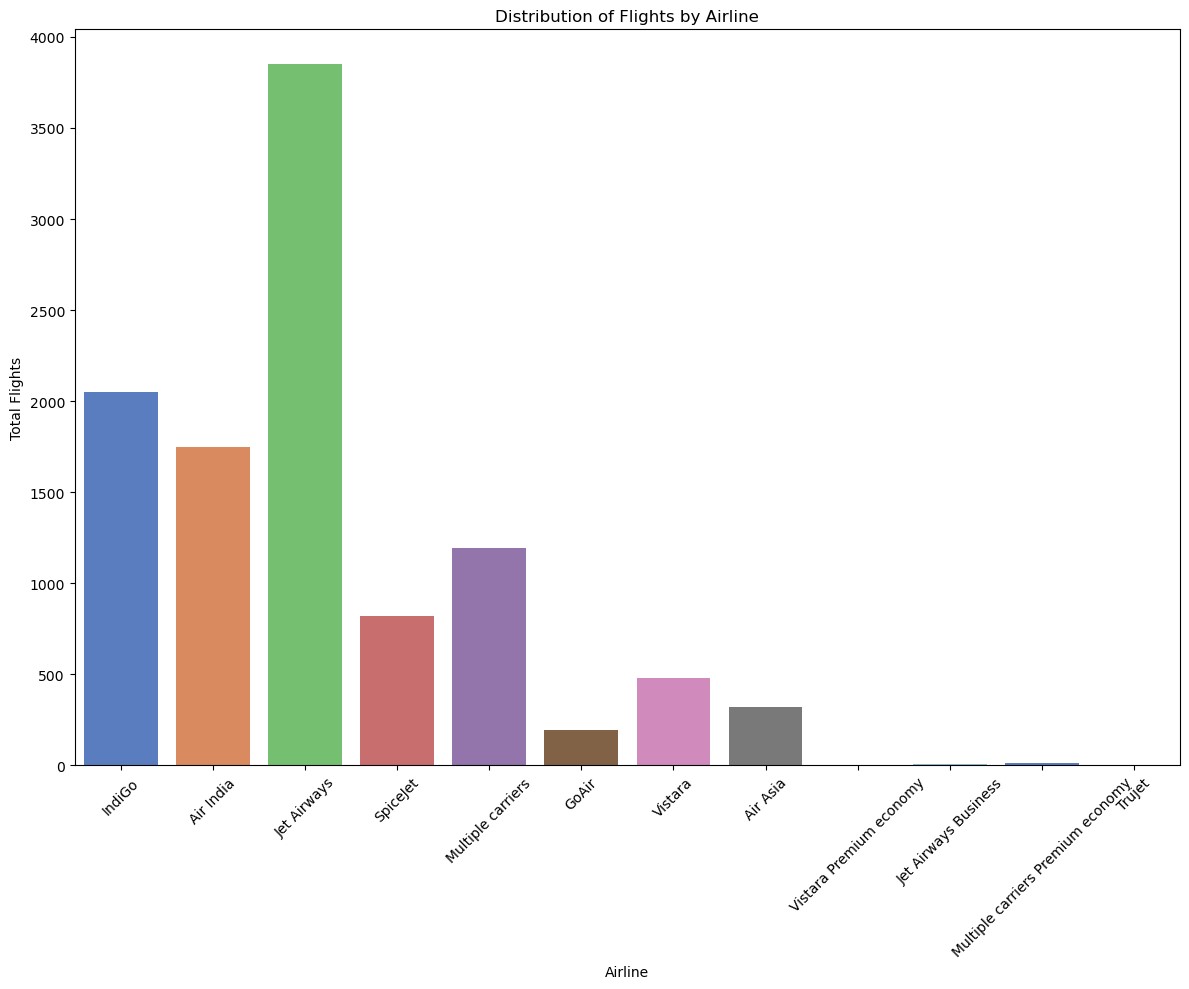
Visualized 'Price' distribution and outliers using a log-transformed histogram and boxplot. Detecting and understanding data patterns and extremes.Using IQR, outliers in 'Price' were identified (>23017), totaling 94. Replacing them with the median ensured data integrity.

Categorical data in 'Airline,' 'Source,' 'Destination,' and ' Categorical data in 'Airline,' 'Source,' 'Destination,' and 'Route' were encoded using one-hot and label encoding techniques, ensuring numerical compatibility. Handling missing values improved data integrity.

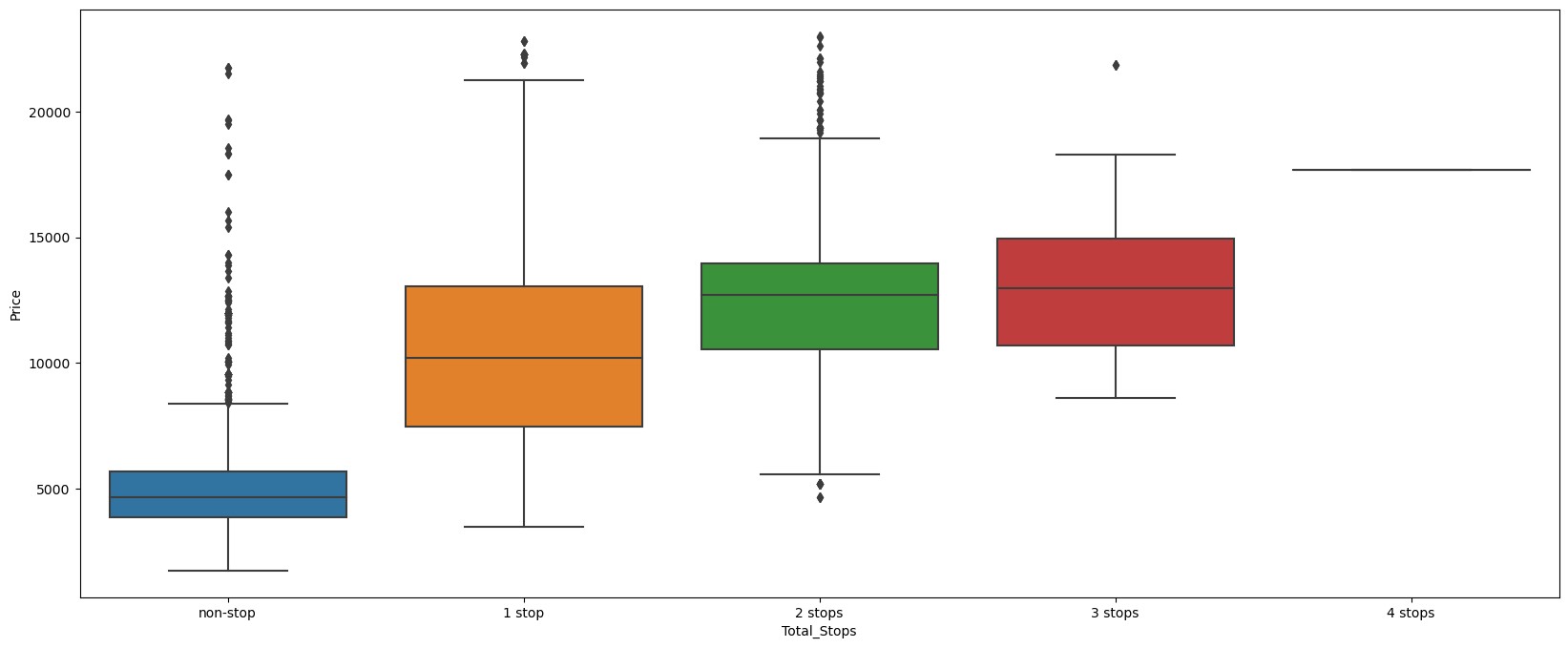
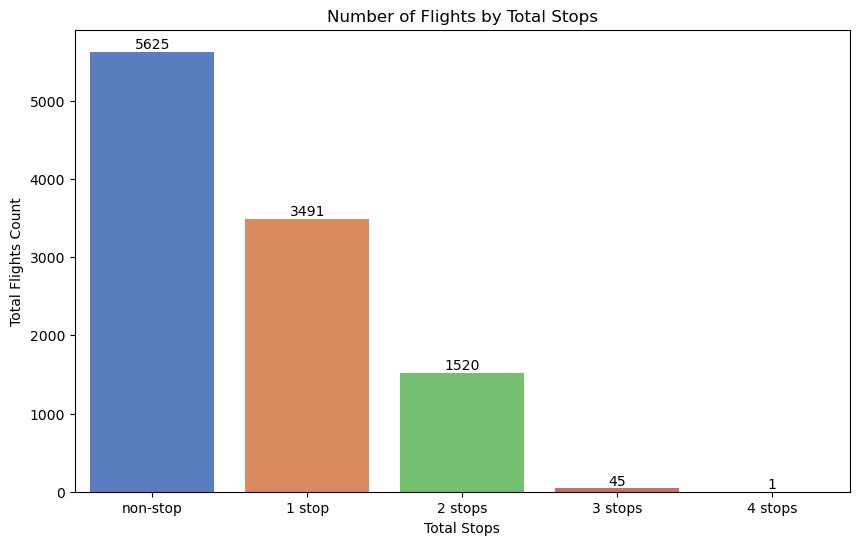


***C. Analysis:***

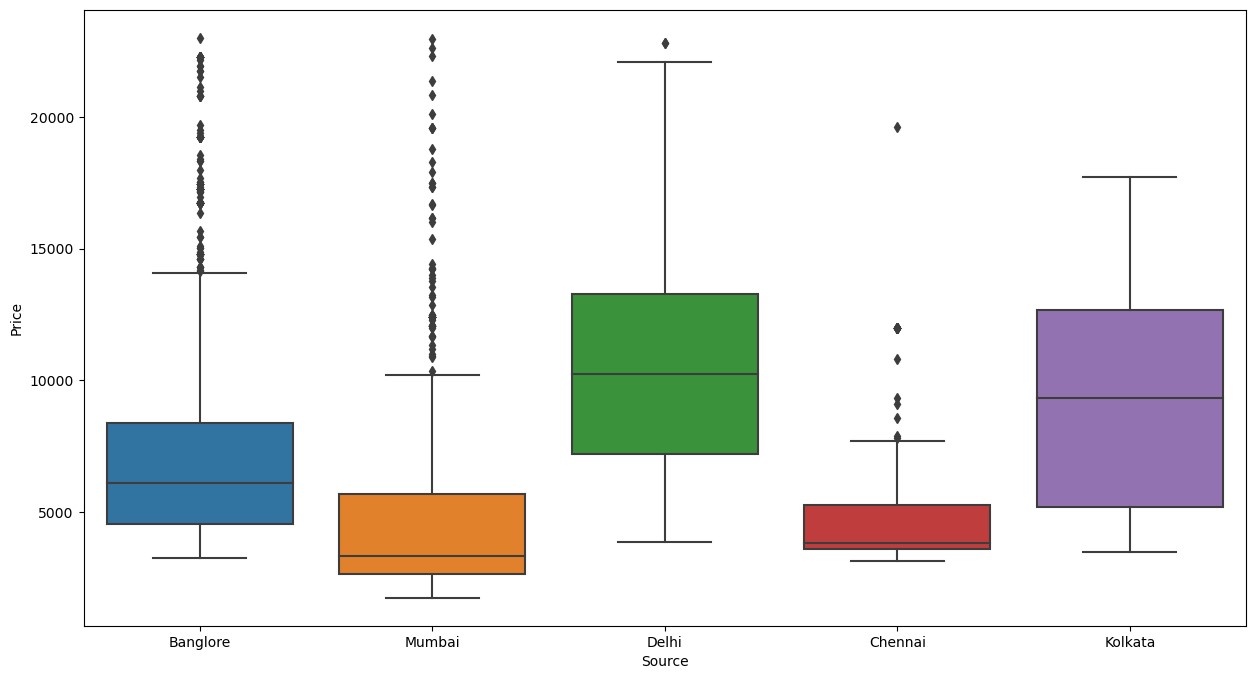
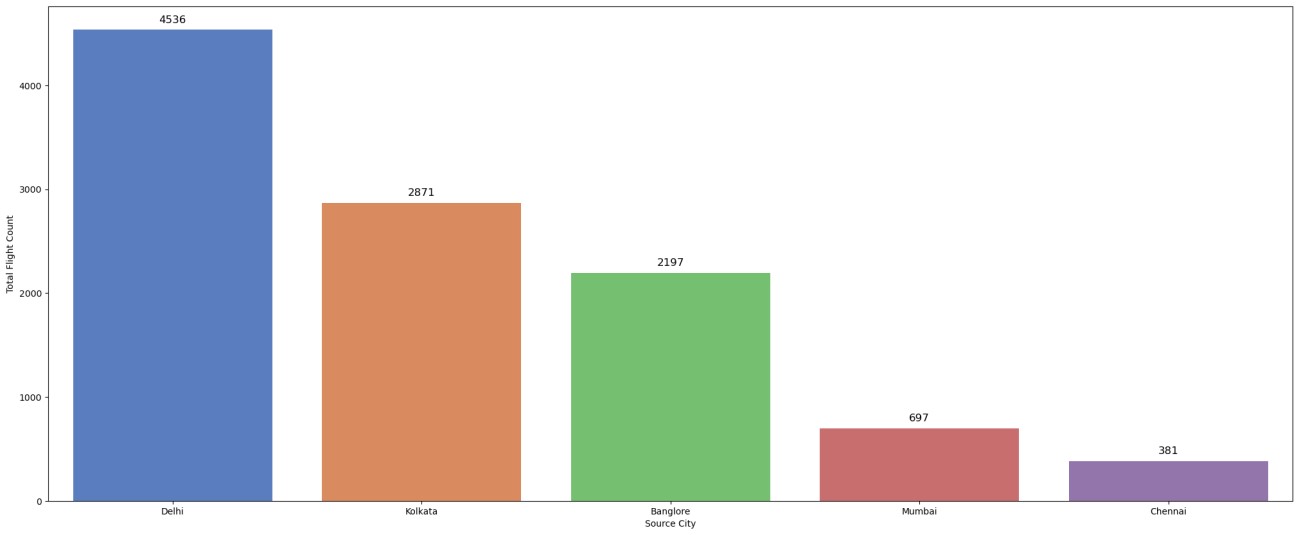
**Airline vs Price:** Jet Airways has the highest median price. The boxplot visualizes price distribution among different airlines.



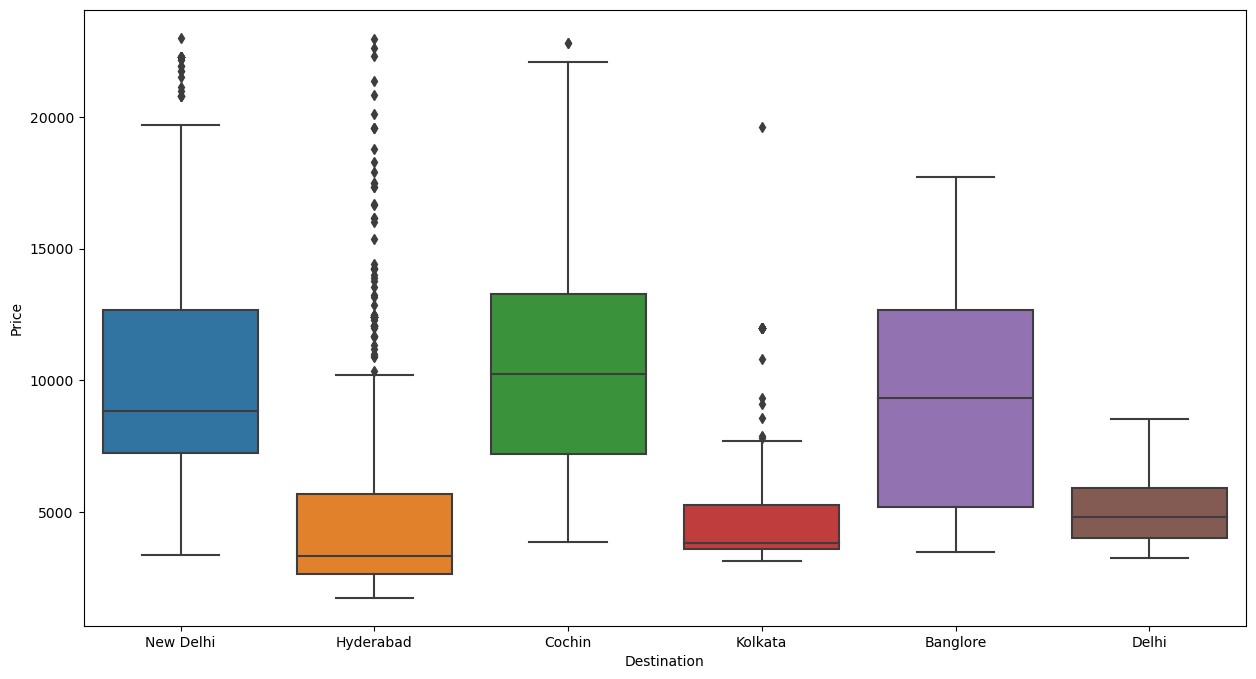
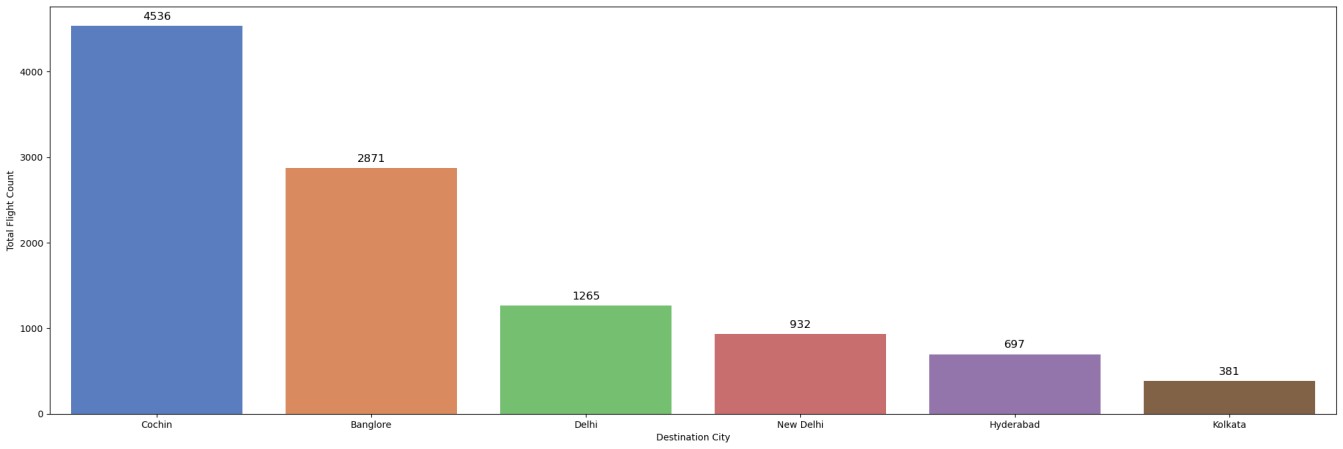
**Total Stops vs Price:** Median price increases with a higher number of stops, demonstrated in the countplot and boxplot.



**Departure City vs Price:** Bengaluru (Banglore) has the highest flight count. The boxplot shows price variations among departure cities. Cochin has the highest flight count. The boxplot reveals price distribution across destination cities.

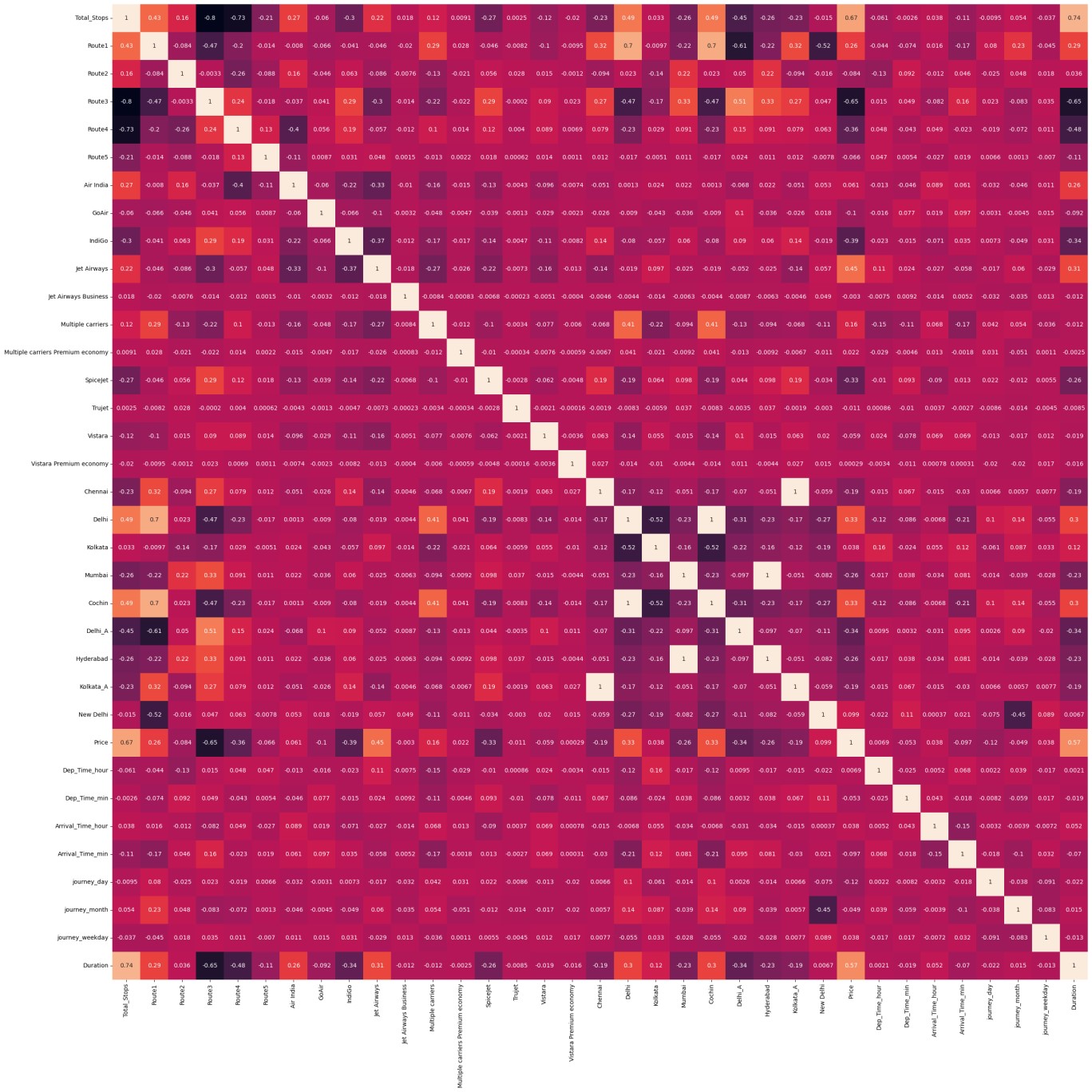


**Destination City vs Price:** Among the destinations, Cochin stands out with the highest flight count, followed by Banglore. The bar plot visualizes the distribution of flight counts across different destination cities. In terms of price, the boxplot reveals significant price variations among these destinations. New Delhi and Hyderabad exhibit relatively higher prices compared to other cities. This insight suggests that the destination city plays a crucial role in determining flight prices.



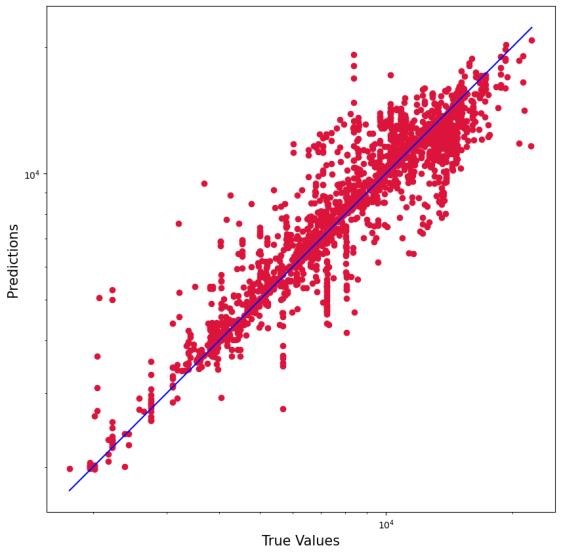
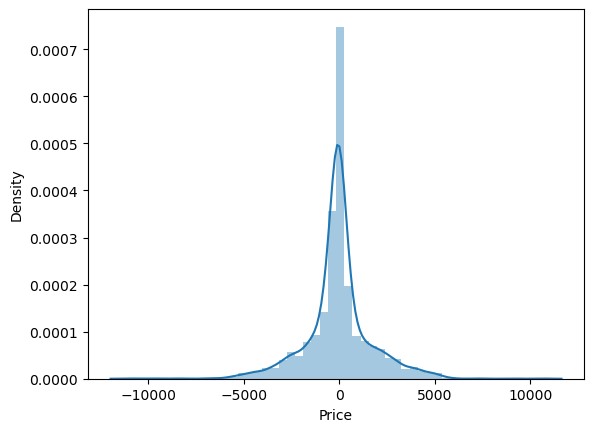
1. ***Feature selection:***

Mutual information scores reveal feature importance. High scores, like Route2 and Duration, indicate strong relationships with the target variable. Mutual Information is performed to identify the most important features for predicting flight ticket prices. Correlation analysis is also conducted to identify relationships between features.

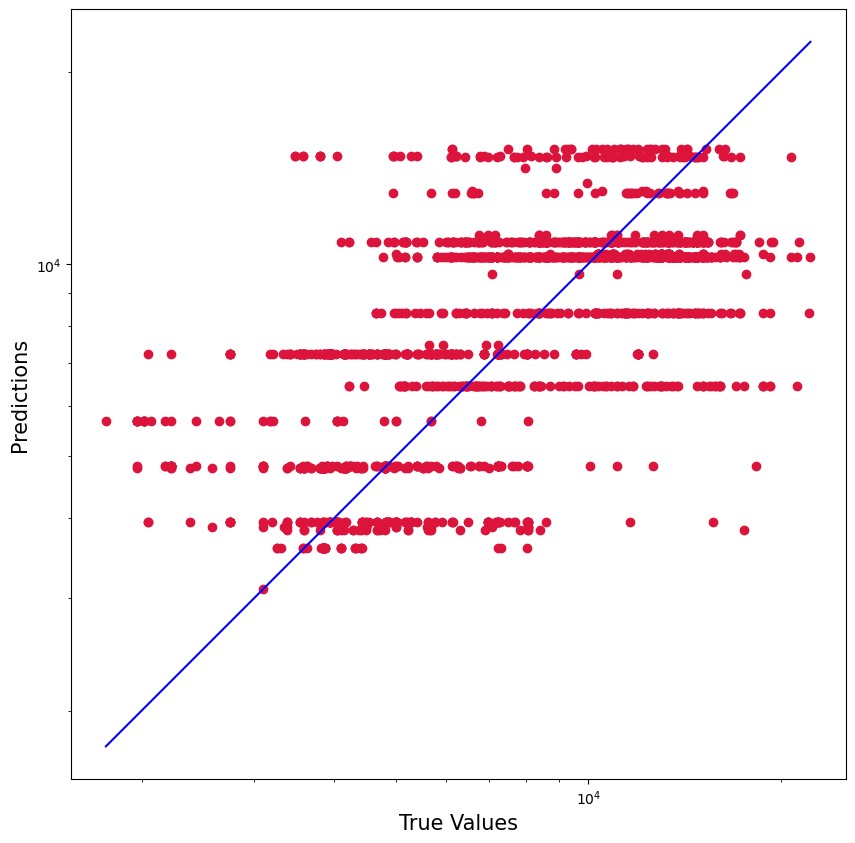
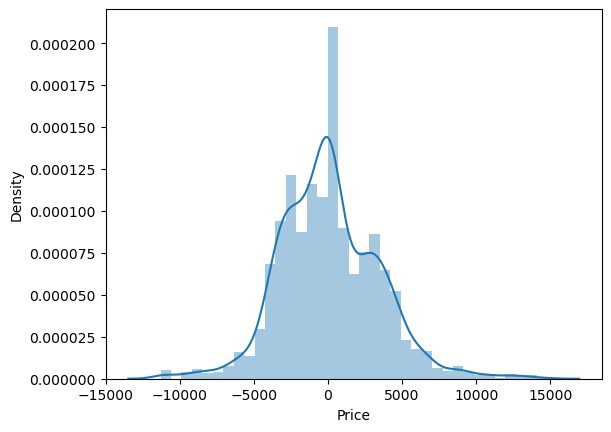


1. ***Models:***

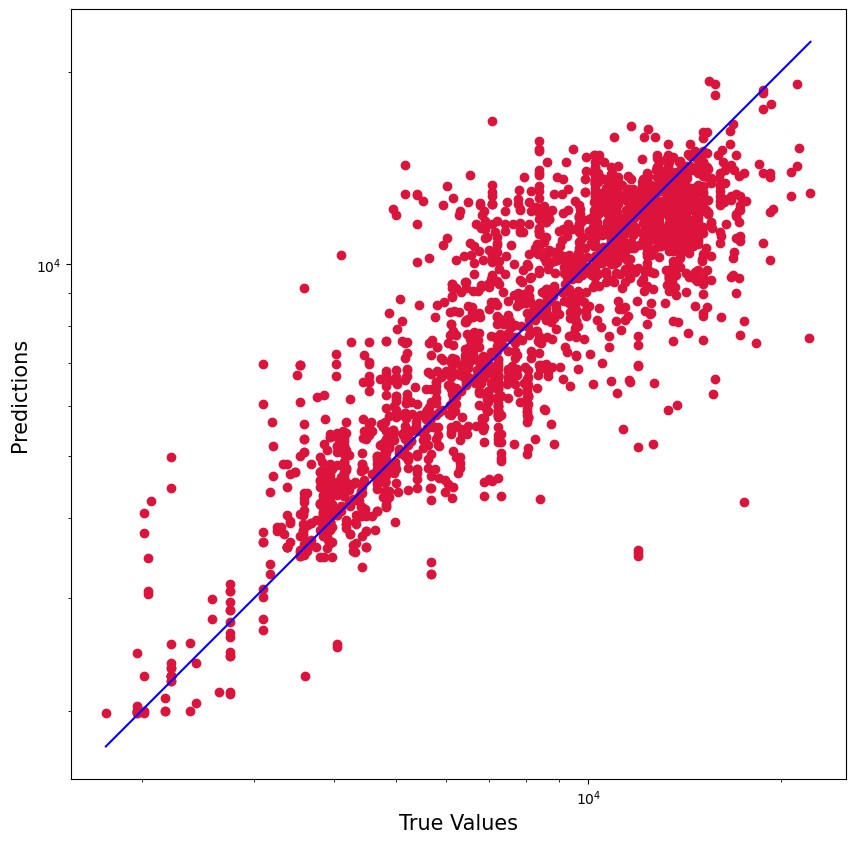
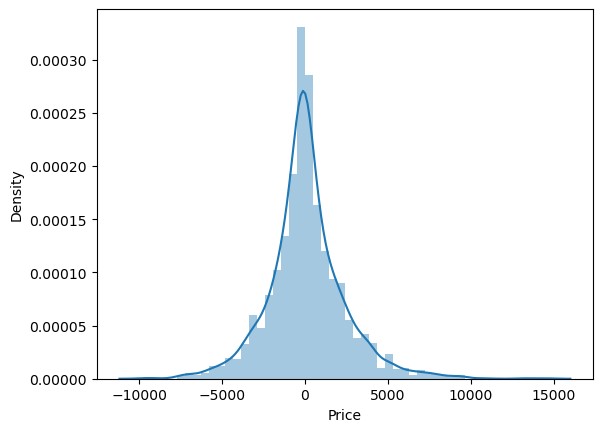
* **Random Forest Regressor**:
* Score: 94.7%
* Performance: R2 score of 81.4%, MAE of 1079.85, RMSE of 1709.03.
* Insight: Robust model capturing complex relationships, evident in high R2 score and low errors.



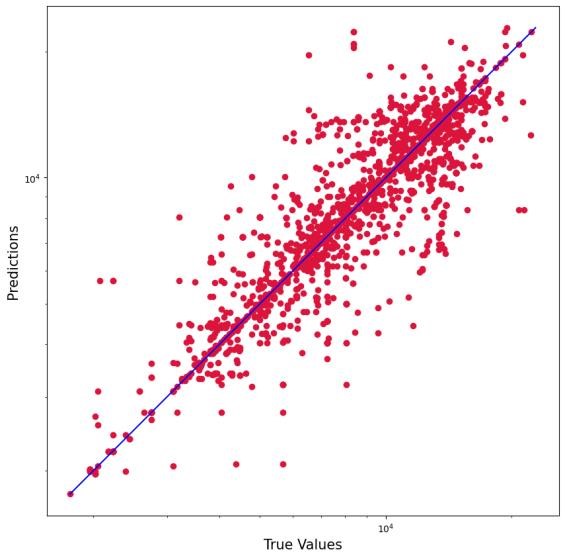
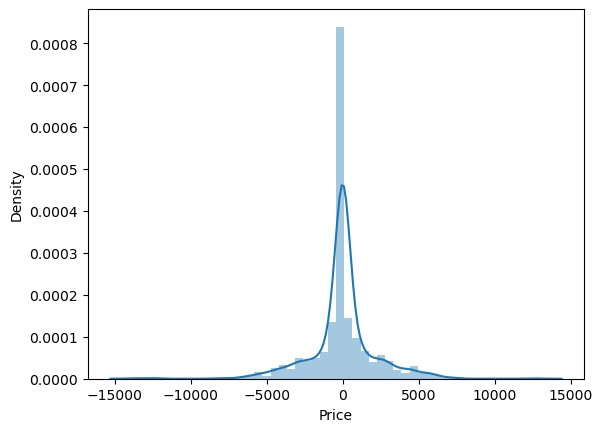
* **Logistic Regression**:
* Score: 7.3%
* Performance: R2 score of 27.8%, MAE of 2543.11, RMSE of 3365.75.
* Insight: Limited performance indicating model may not be suitable for this regression task.



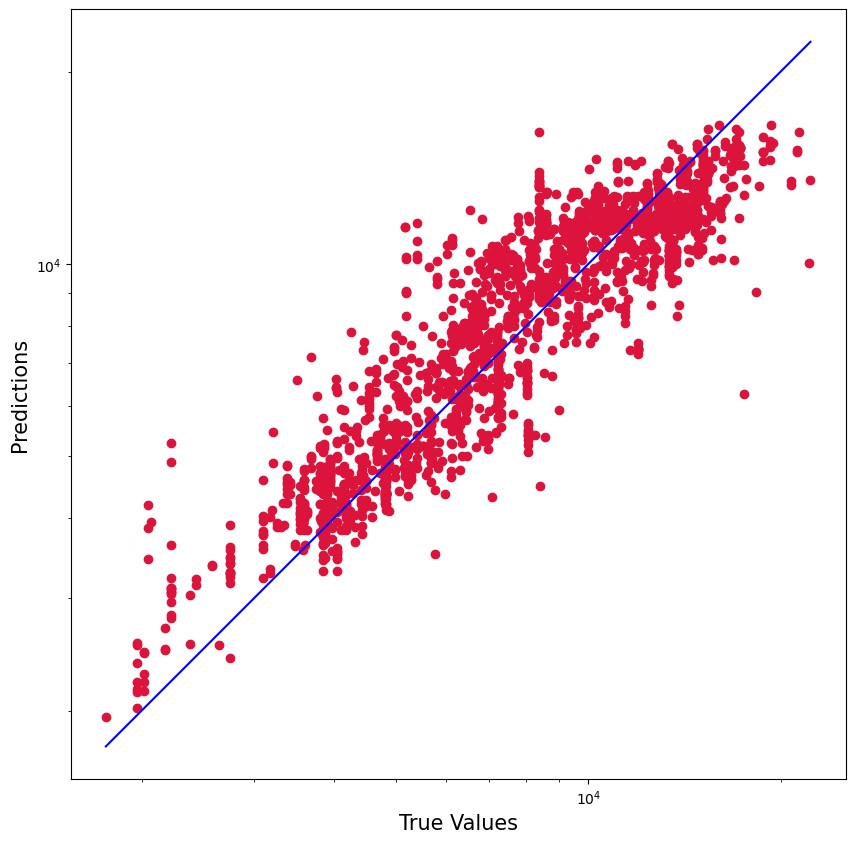
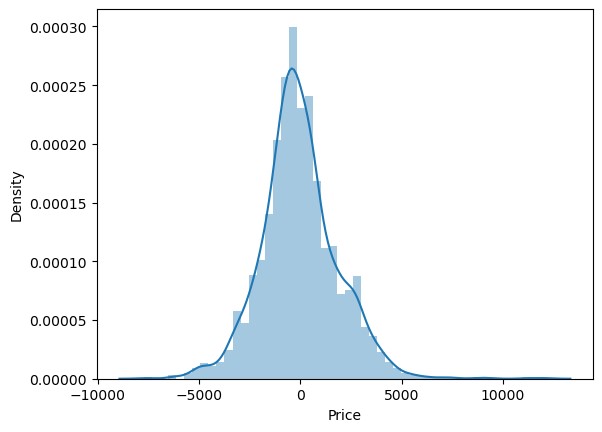
* **KNeighbors Regressor**:
* Score: 76.7%
* Performance: R2 score of 64.9%, MAE of 1617.96, RMSE of 2344.24.
* Insight: Moderate performance, capturing some relationship but with higher errors compared to Random Forest.



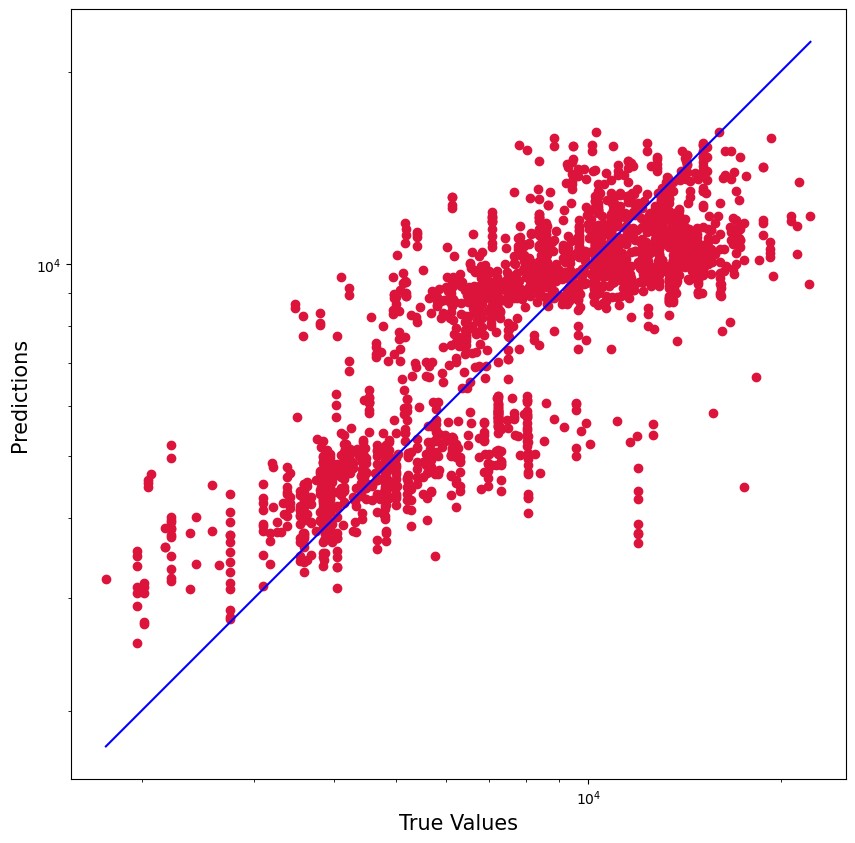
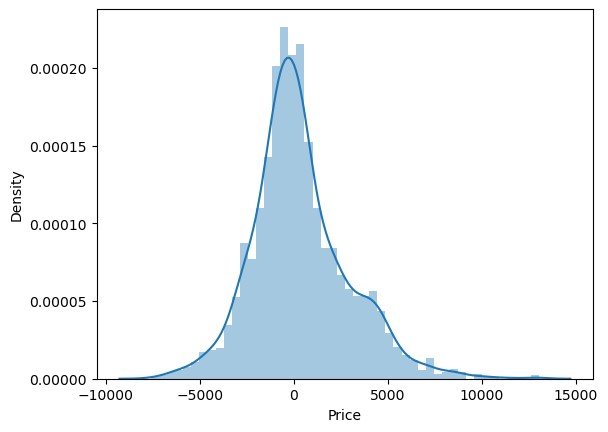
* **Decision Tree Regressor**:
* Score: 96.1%
* Performance: R2 score of 70.9%, MAE of 1212.25, RMSE of 2135.89.
* Insight: High training score with decent R2 score, indicating potential overfitting.



* **Gradient Boosting Regressor**:
* Score: 77.8%
* Performance: R2 score of 77.2%, MAE of 1416.10, RMSE of 1889.86.
* Insight: Similar performance to Random Forest, capturing complex relationships with lower errors.



* **Support Vector Regressor (Linear Kernel)**:
* Score: 57.0%
* Performance: R2 score of 56.8%, MAE of 1904.61, RMSE of 2597.72.
* Insight: Limited performance compared to other models, indicating linear relationships may not be sufficient for capturing the data's complexity.



***F. Hypertuning parameteres:***

**Random Forest Regressor Hyperparameter Tuning:**

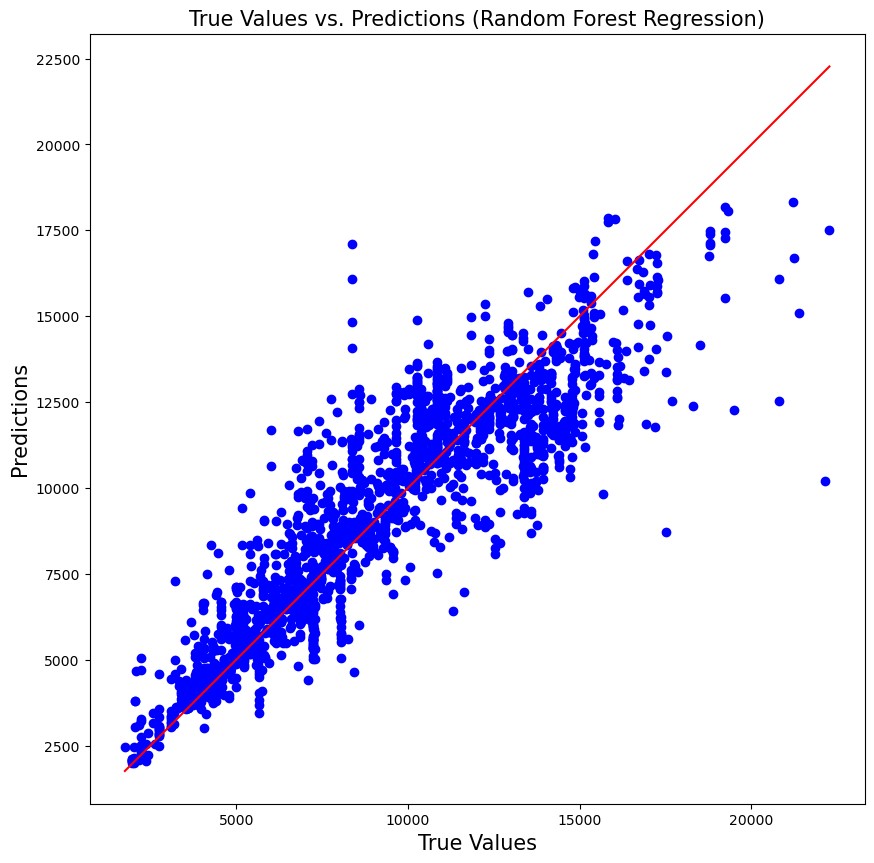
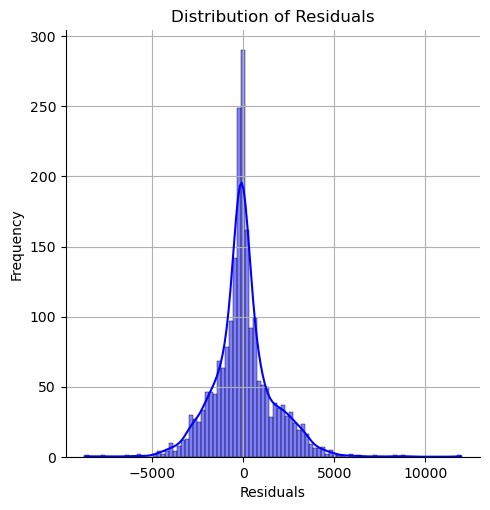
• Randomized Search: o Parameters Searched: ▪ 'n\_estimators': [100, 120, 150, 180, 200, 220]

* 'max\_features': ['auto', 'sqrt'] ▪ 'max\_depth': [5, 10, 15, 20] o Best Parameters: ▪ 'n\_estimators': 1800
* 'min\_samples\_split': 10
* 'min\_samples\_leaf': 1
* 'max\_features': 'sqrt'
* 'max\_depth': 30
* 'bootstrap': False • Model Evaluation:

o Performance Improvement:

▪ R2 score increased from 81.38% to 83.76%.

* Distribution Plot: o Improved fit evident in the distribution plot, aligning actual and predicted values.

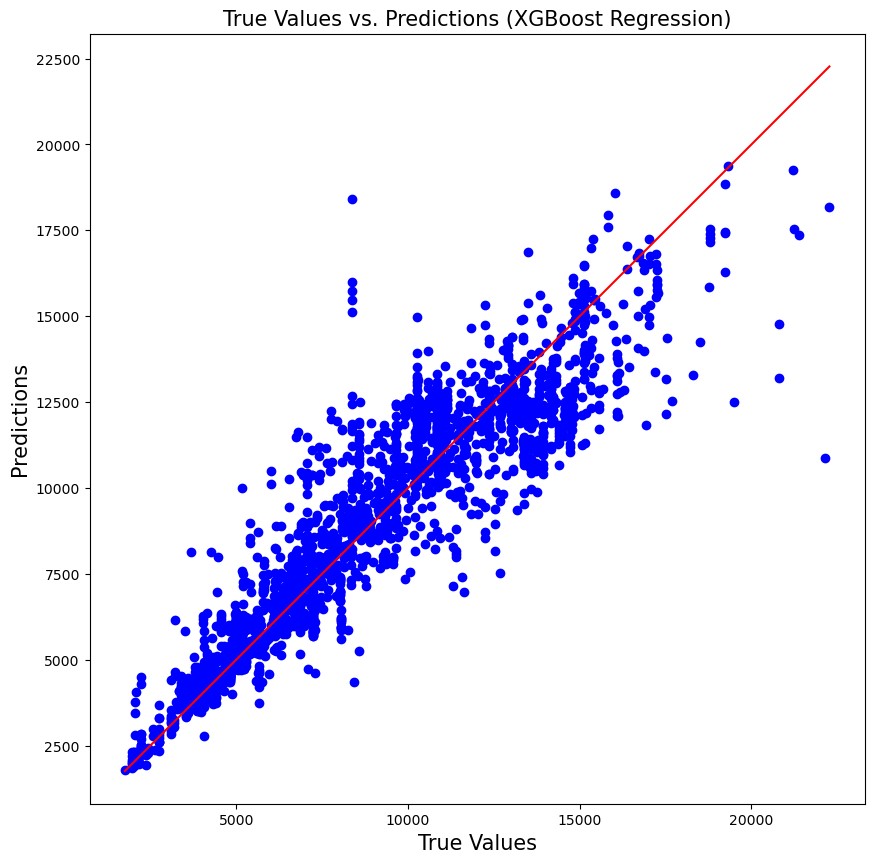
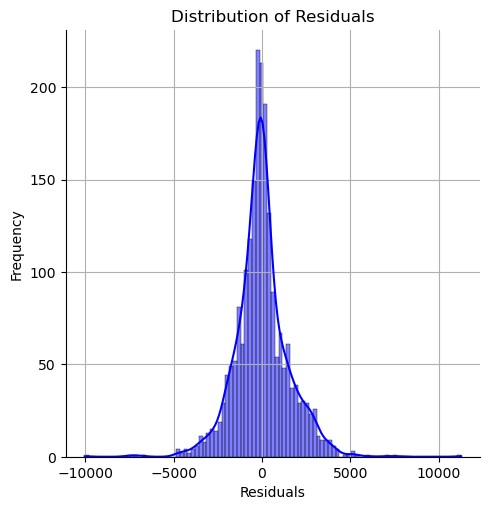


**XGBoost Regressor Hyperparameter Tuning:**

* XGBoost Model: o Parameters Used:
* Default parameters with learning\_rate=0.1. o Best Hyperparameters:
* 'subsample': 0.7
* 'reg\_lambda': 0.1
* 'reg\_alpha': 0
* 'n\_estimators': 500
* 'min\_child\_weight': 3
* 'max\_depth': 5
* 'learning\_rate': 0.05
* 'gamma': 0.2
* 'colsample\_bytree': 0.7

• Model Evaluation: o Performance: ▪ R2 score: 85.37%

* MAE: 1065.62
* MSE: 2295324.44
* RMSE: 1515.03 • Distribution Plot: o Good alignment in the distribution plot, indicating improved model fit.



**Model Comparison Conclusions:**

* The XGBoost Regressor outperformed the Random Forest Regressor in terms of R2 score, MAE, MSE, and RMSE.
* XGBoost showed better performance in capturing the complex relationships within the data compared to Random Forest.
* Both models demonstrated improvements after hyperparameter tuning, indicating the importance of optimizing model parameters for better performance.

# IV. RESULTS

In the attempt of building an optimal regression model for the air ticket price prediction a very comprehensive evaluation was done for various models and algorithms. The initial assessments showed that random forest regressor performed with a remarkable performance of r2 score of 81%, and showed an improved performance with 83.10%, this shows how important fine-tuning is and how it can increase the accuracy and optimization.

With random forest we also performed in parallel, XGBoost regressor, another powerful algorithm and this algorithm yielded an r2 score of 85.36%, hence performing better than Random Forest Regressor. This comparison shows the significance of model selection and fine tuning the parameters and this in return can help us build a stable yet dynamic airline ticket prices system.

## V. CONCLUSION AND FUTURE WORK

After a thorough analysis of predicting airline ticket prices using various machine learning models, the study reveals the effectiveness of both Random Forest Regressor and XGBoost Regressor. Random Forest Regressor achieved a commendable R2 score of 83%, while XGBoost Regressor outperformed with an impressive R2 score of 85%.These results underscore the importance of model selection and parameter tuning in achieving accurate predictions. While Random Forest Regressor demonstrates robust predictive capabilities, XGBoost Regressor offers even higher precision, highlighting the potential for further improvements.

Looking ahead, future research could explore advanced techniques such as ensemble methods, neural networks, or more sophisticated optimization algorithms to enhance prediction accuracy. Additionally, incorporating additional features or data sources may further refine the models and provide travelers with even more reliable information for their travel planning needs. Continued refinement and exploration of diverse modeling approaches will be crucial for advancing the accuracy of airline ticket price predictions and enhancing the overall travel experience for users.

## VI. REFERENCES

EaseMyTrip dataset: [Dataset source or website link]

Scikit-learn Documentation Matplotlib Documentation

Seaborn Documentation

XGBoost Documentation