FLIGHT FARE PREDICTION

Analyzing and Predicting Flight Prices Using Machine Learning

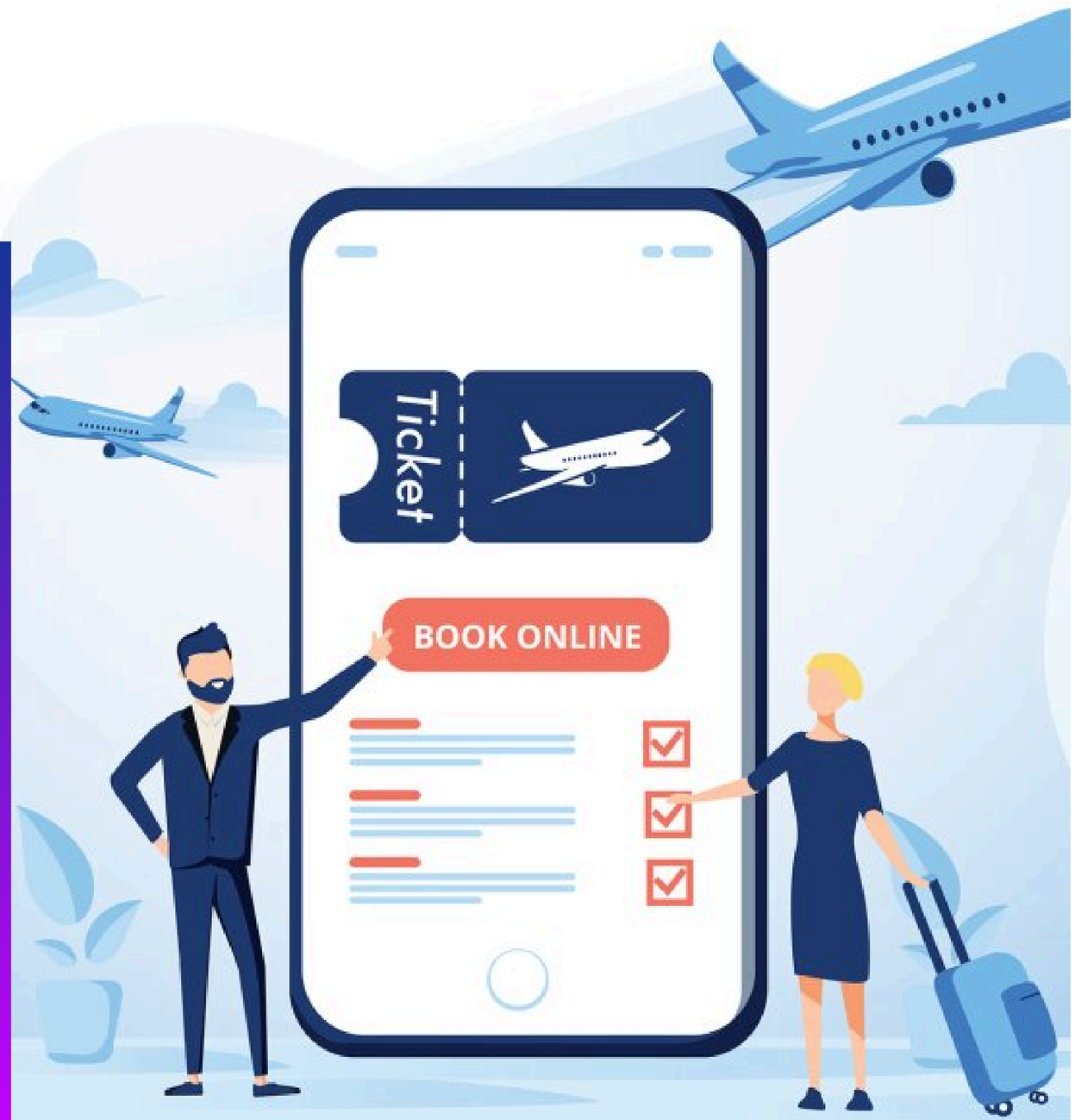
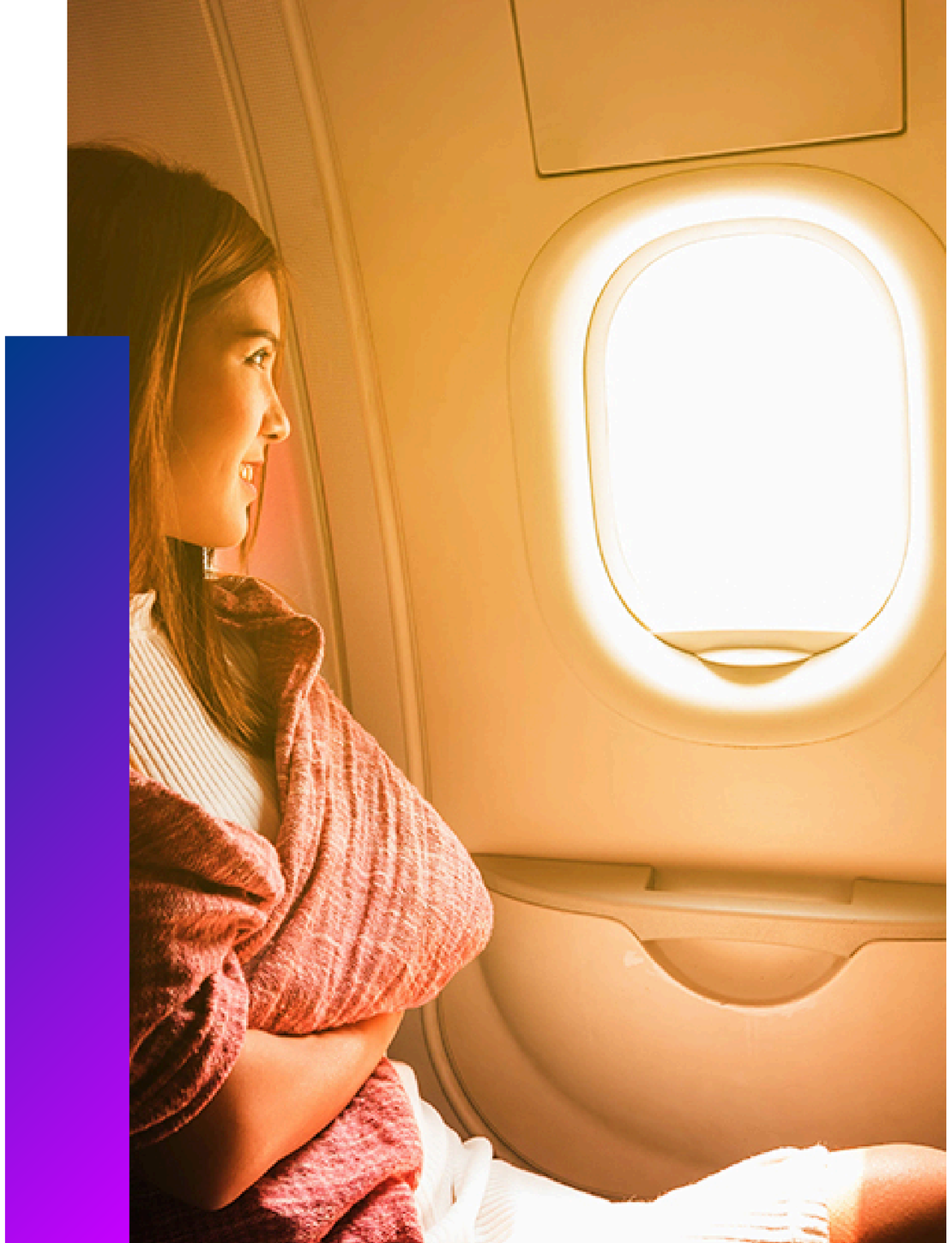


TABLE OF CONTENT

ABSTRACT

People who frequently travel through flight will have better knowledge on best discount and right time to buy the ticket. For the business purpose many airline companies change prices according to the seasons or time duration. They will increase the price when people travel more. Estimating the highest prices of the airlines data for the route is collected with features such as Duration, Source, Destination, Arrival, Departure. Features are taken from chosen dataset and in this paper, we have used machine learning techniques and regression strategies for prediction of the price wherein the airline price ticket costs vary overtime. We have implemented flight price prediction for users by using decision tree and random forest algorithms. Random Forests shows the best accuracy of 99% for predicting the flight price.

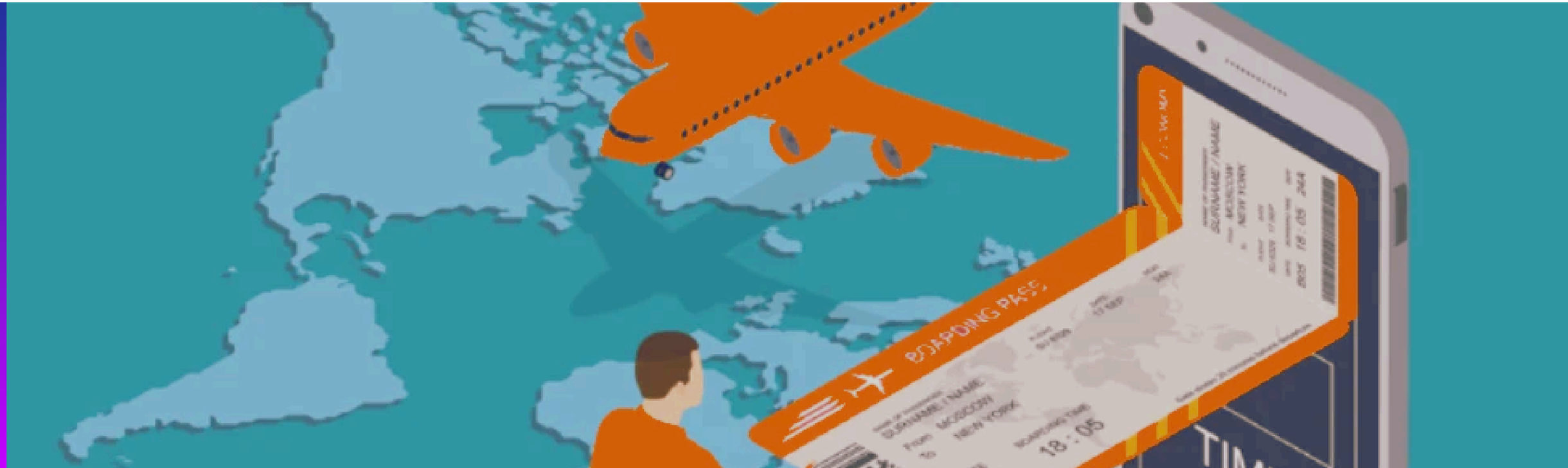


PROBLEM STATEMENT

Created a model which can predict price where the airline industry is highly competitive, with prices fluctuating based on a multitude of factors such as airline carrier, source and destination cities, flight duration, number of stops, flight class, and the number of days left until departure. Predicting flight prices accurately can benefit various stakeholders, including passengers, travel agencies, and airlines. Passengers can plan their travels and book flights at the best prices, while airlines and travel agencies can optimize their pricing strategies and improve customer satisfaction.

INTRODUCTION

Perfect time for purchasing plane ticket by the passenger’s view is difficult since passengers get very less information of future business price rates. Different models figure out future business price on plane and categorize the best time to obtain flight ticket. Airlines use different strategies of pricing for their tickets, later taking the decision on price because order shows higher value for the approximation models .Also, seating arrangements in flight which is not occupied shows the loss of the amount invested for the business airline companies and making them purchase the ticket to fill the seats for any price this would be the best idea to get profit in loss too. Passengers should be compatible with the airline companies to get adjusted for the increase and decrease of the price. Planes ticket prices changes as time passes, pulling out the elements which creates the difference. Reporting the correlated and models which is used to price the flight tickets. Then, using that information, building the model which helps passengers to make pull out the ticket to buy and predicting air ticket prices which progresses in the future. Duration, Arrival time, Price, Source, Destination and much more these are the attribute used for flight price prediction.



STES FOR PROJECT BUILDING

01 Data Collection

02 Importing Necessary Library

03 Data Analysis & Infromation

04 Exploratory Data Analysis

05 Data Preprocessing

06 Model Building

07 Model Performance Metrics

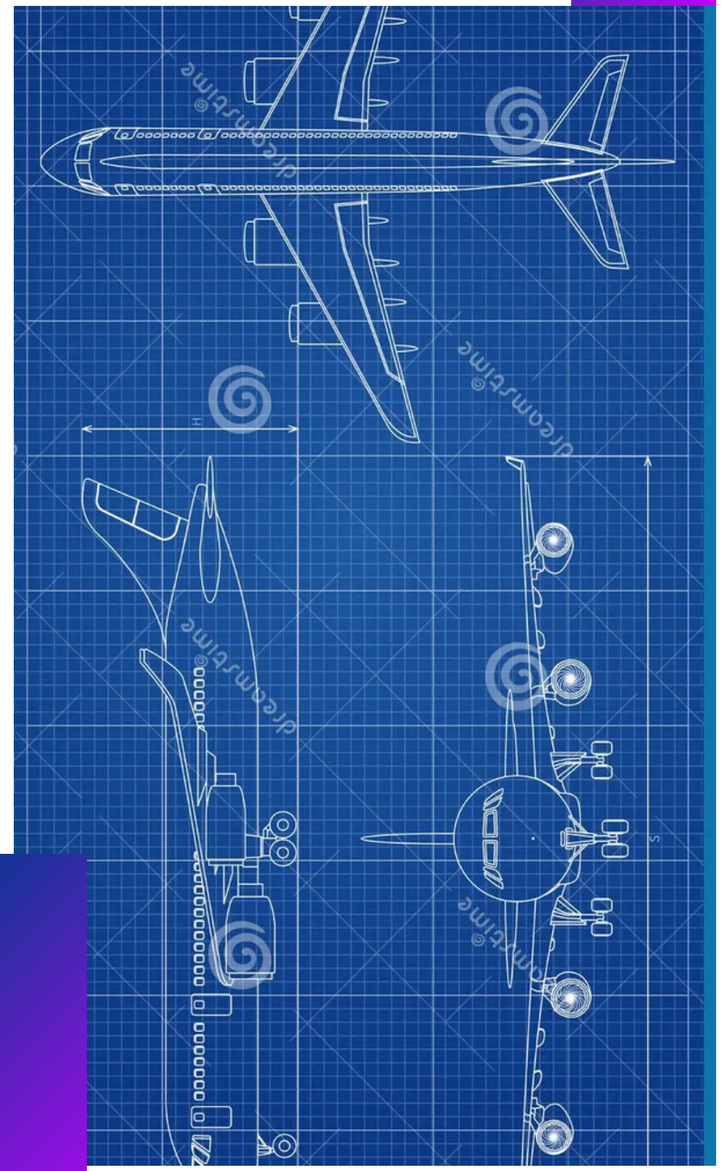
08 Testing Data & Accuracy Comparison

09 Conclusion

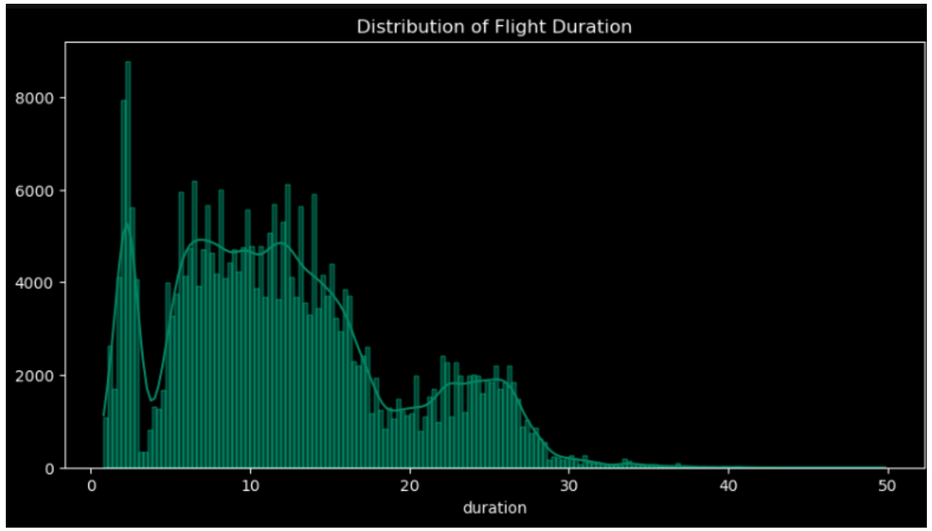
EXPLORATORY DATA-ANALYSIS

USED GRAPHS

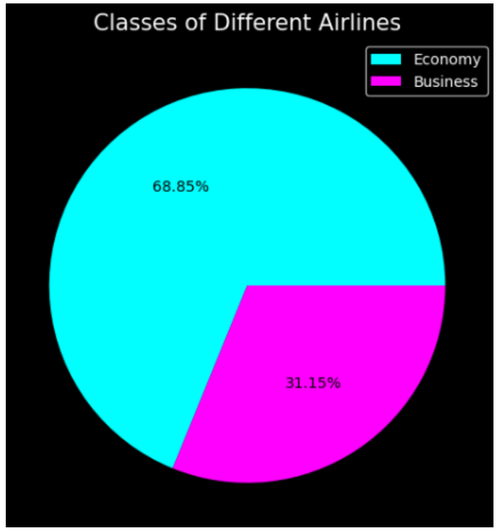
- Pandas Profiling
- Count Plot
- Histogram
- Pie Chart
- Hist-Plot
- Violin Plot
- Line Plot
- Box Plot
- Dis-Plot
- Pair Plot
- Scatter Plot
- Heat Map



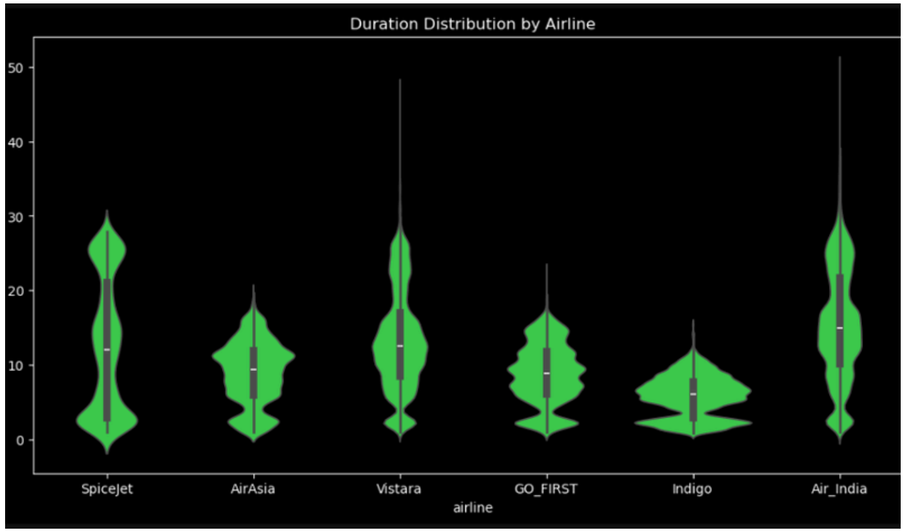
SOME OF THE VISUALIZATION



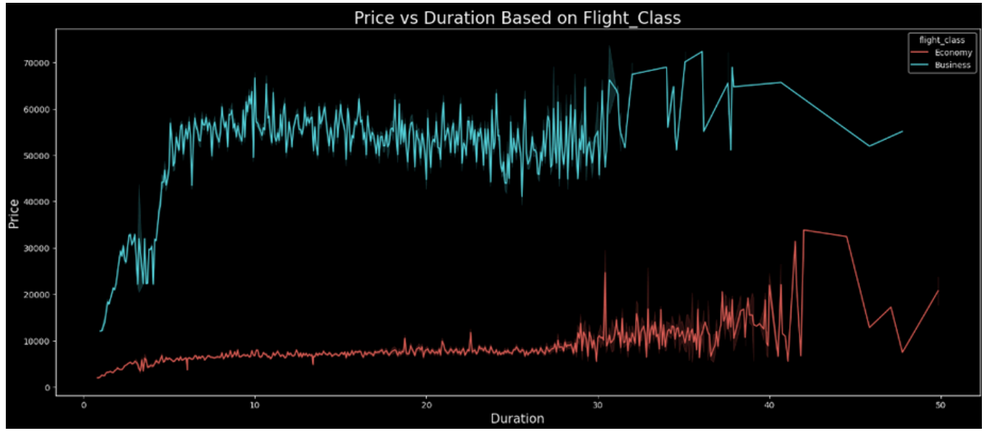
Hist-Plot Distribution By Count and Duration



Pie-Chart Count By Classes Of Different Airlines



Violin Plot Distribution By Airline & Duration

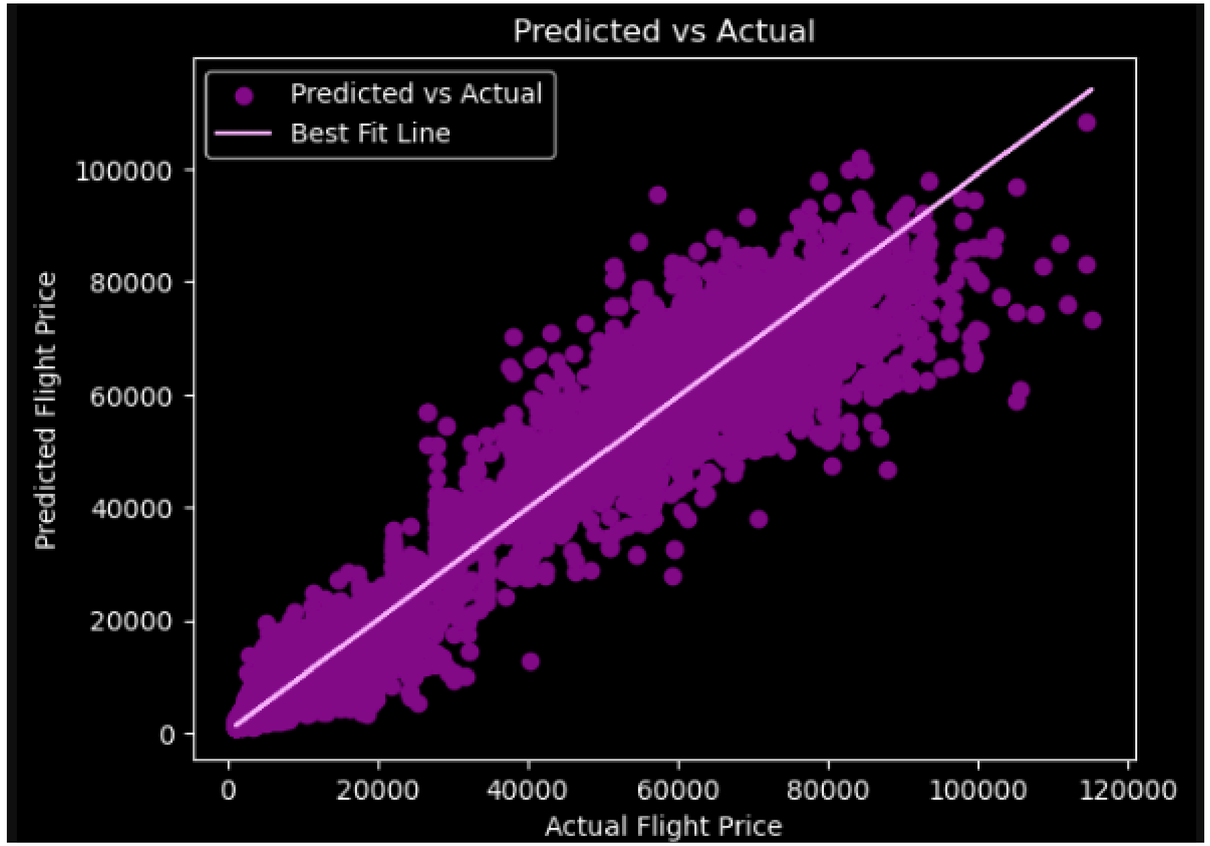


Line Chart Based on Flight Class & Price

The distribution of the categorical columns "airline","source_city","destination_city" and "flight_class" from a dataset is shown.
Furthermore, Pandas Profiling has been used to do a thorough exploratory data analysis (EDA) that offers a comprehensive overview of the dataset, including descriptive statistics, correlations, missing values, and feature distribution.
Plots of various kinds have been made to provide more in-depth understanding of the data.
With the help of these visualisations, one can gain a thorough grasp of the dataset and identify underlying patterns, linkages, and possible research topics. Plots and statistical summaries together can help us better prepare the data for model training and increase the precision and dependability of our forecasts.

ALGORITHMS USED

- Standard Scaler
- Linear Regression
- K-Neighbors Regressor
- Random-Forest Regressor
- Decision-Tree Regressor
- XGB Regressor
- Cat-Boost Regressor
- AdaBoost Regressor
- Lasso
- Ridge
- Elastic-Net
- Polynomial Feature

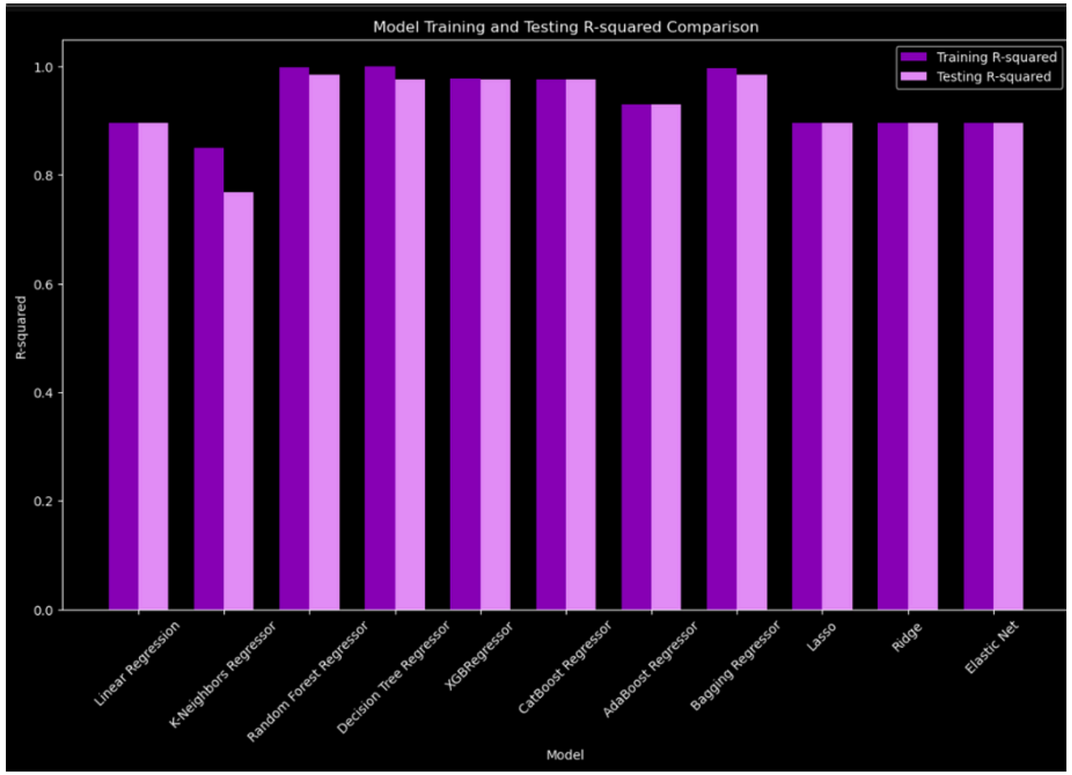


Created a model where the performance of the flight price prediction model is visually shown by the "Predicted vs. Actual" scatter plot. The x-axis shows the actual flight prices, and the y-axis shows the expected flight prices. Each point on the plot represents a pair of real and predicted flight prices. The best fit line, represented by the diagonal purple line, shows where the points would fall in the event that the forecasts were exact.

The dense clustering of points around the best fit line in the plot indicates a substantial connection between the actual and anticipated values. This implies that the model can predict flight costs for a broad range of variables with accuracy.

Slight point dispersion, especially at higher price points, however, suggests that although the model works well in general, there is considerable variation in its forecasts for more costly trips.

In summary, the figure illustrates how well the model predicts flight prices overall, and it suggests more research and improvement of forecasts for more expensive trips in order to increase accuracy and dependability across the board.



Model	Training R-squared	Test R-squared
Linear Regression	0.895300	0.895000
K-Neighbors Regressor	0.850200	0.768700
Random Forest Regressor	0.997400	0.985000
Decision Tree Regressor	0.999300	0.975800
XGBRegressor	0.977300	0.976500
CatBoost Regressor	0.975700	0.975200
AdaBoost Regressor	0.929700	0.930500
Bagging Regressor	0.996800	0.983800
Lasso	0.895289	0.894975
Ridge	0.895289	0.894975
Elastic Net	0.895289	0.894975

CONCLUSION

With the highest Test R-squared value of 98%, the Random Forest Regressor performs better when forecasting flight costs, suggesting improved performance on fresh data. our model is quite appropriate for our application since it captures intricate relationships in the data in an effective manner. Although the Decision Tree Regressor exhibits great accuracy, its propensity to memorize training data makes it susceptible to overfitting, which impairs performance on unseen data.

In light of these findings, the Random Forest Regressor is advised due to its exceptional generalizability and accuracy. But it's crucial to keep an eye out for any overfitting by routinely assessing the model's performance on fresh, varied datasets. To further improve forecast accuracy and robustness, an ensemble method or stacking models could be investigated. To ensure the model's continuous relevance and performance in the ever-changing airline business, regular updates and retraining with fresh data are required.

THANK YOU