SkyPredict: Airfare Price Prediction Machine Learning Model

Tanmay Jain
Dept. of Computer Science and Engineering
Netaji Subhas University of Technology
Delhi, India
tanmay.jain.ug22@nsut.ac.in

Kushagra Anand
Dept. of Computer Science and Engineering
Netaji Subhas University of Technology
Delhi, India
kushagra.anand.ug22@nsut.ac.in

Abhishek Verma
Dept. of Computer Science and Engineering
Netaji Subhas University of Technology
Delhi, India
abhishek.verma.ug22@nsut.ac.in

Dr. Gaurav Singal
Assistant Professor
Dept. of Computer Science and Engineering
Netaji Subhas University of Technology
Delhi, India
gaurav.singal@nsut.ac.in

I. Introduction

Airlines rely on complicated yield management algorithms in order to roll out dynamic pricing mechanisms, which can adjust airfares as real-time as possible and usually aims to maximize revenue or profit for each flight. This generally leads to extreme price volatility — where the cost of a ticket can increase or decrease rapidly over very short time scales. The more seasoned travelers among us may be aware of the basic principles around airfare trends, but for most infrequent fliers it can seem virtually non-navigable.

For many passengers, especially occasional flyers being a little lost isn't uncommon. Usually, you do assume that booking further in advance will always be the cheaper option. That includes the belief that prices won't decrease suddenly — but as you probably know, they can for various reasons (like last-minute seat releases or changes in demand). So, passengers are typically in a position playing activity by continuously checking fares over time in the hope of straining out that finest ticket. Checking airfare trends on a daily basis can be enough to stress out even the most unflappable traveler — and simply looking for prices months before you plan to travel won't save money since many of us end up paying more than we should have, under different booking circumstances but in the same seat. Worsening the situation, airline pricing practices continue to be notoriously opaque, despite being longtime leaders in airfare transparency for consumers.

The surge of online technology and the rise in numerous e-commerce platforms has made it easier for travelers to access flight details with a range of airfares, both domestic and international on different airlines worldwide. However, this transparency itself is a part of the problem due to airfare being highly predictable. This is because the ticket prices are constantly changing, but also due to a myriad of factors influencing those changes, which renders it nearly impossible for customers to figure out when they should buy their tickets. Hence the demand in intelligent systems forecasting real-time flight prices and advice to customers when booking at lowest price.

To solve this problem, machine learning technologies are developed to predict airfare reviews. The methods are based on historical prices and other various ticket cost factors in order to model when prices will rise or fall as well as how long you should wait before making your purchase.

In this research our goal is to make compare between linear regression, decision tree regressor with random grid search cv and bagging regressor vs random forest for preparing model that can predict flight prices accurately as well as precisely. The study does so by using the number of stops, source and destination place of a flight etc as attributes from data set to analyze different factors responsible for determining airfare. By looking at these features, the study is trying to account for something about each individual airline as well as overall market conditions that influence airfare fluctuations.

To do this, it begins by training every model in the study on the historical dataset, so they learn what kind of patterns are associated with fare changes. Model performance is compared with evaluation metrics like accuracy and precision, which ensure that the model can predict flight prices correctly under different scenarios. The aim, therefore, is to find the model that fits well with data and at same time generalizes so as not guaranteed on new/invisible instances of it while making good predictions.

It provides more price transparency, which allows consumers to make purchase decisions based on a clearer picture of how they will be charged for airline services than is the case. And it may also help airlines and travel agencies tune their pricing strategy for the best, hopefully to reduce booking uncertainty in airline reservation process (and potentially save travelers some money along with a stress-free journey.)

II. Literature Survey

Airline fare prediction has been the focus of multiple studies, each exploring various machine learning techniques to model and forecast dynamic ticket pricing. Below is a summary of notable research efforts in this field:

A. Predicting the Fare of a Flight Ticket with Machine Learning Algorithms

The Random Forest algorithm was identified as the most effective model, achieving an R^2 value of 0.81. This model demonstrated superior performance compared to Decision Tree (R^2 = 0.63) and Linear Regression (R^2 = 0.50). The study concluded that the high precision and low Mean Absolute Error (0.11) of Random Forest make it a reliable tool for predicting airfare prices, providing consumers with insights to optimize their purchasing decisions.

B. An Airfare Prediction Model for Developing Markets

Focusing on the emerging aviation market in Vietnam, this study developed a stacked prediction model using Random Forest and Multilayer Perceptron algorithms, achieving an R^2 value of 0.47. The interval between the purchase date and departure date was identified as the most significant factor influencing ticket prices. The research suggests integrating additional factors, such as flight delays and multi-stop flights, to further enhance model accuracy.

C. A Framework for Predicting Airfare Prices Using Machine Learning

In this study, the Decision Tree (DT) algorithm emerged as the most effective method for predicting flight prices within Indian networks, achieving an accuracy of 89%, significantly surpassing Naive Bayes (45%) and Stochastic Gradient Descent (38%). The research underscored the importance of model selection and evaluation metrics, demonstrating DT's superior precision and recall in analysing complex datasets, particularly when including specific attributes such as airline, origin, and destination.

D. Machine Learning Modelling for Time Series Problem: Predicting Flight Ticket Prices

This study investigated the effectiveness of various machine learning models, with the AdaBoost-Decision Tree model emerging as the best performer. The research highlighted the HMM Sequence Classification algorithm, which showed a 31.71% improvement over a random purchase strategy for generalized routes. The use of clustering techniques such as K-Means and Expectation-Maximization (EM) helped enhance the models' handling of imbalanced datasets.

E. Airfare Analysis and Prediction Using Data Mining and Machine Learning

This study found that Naïve Bayes achieved the highest accuracy (84.01%) for predicting airline prices, followed closely by SVM (83.97%). The research identified factors such as days until departure, crude oil prices, and competition as critical variables in airfare fluctuations. The predictive models provided valuable insights into airline pricing strategies, particularly in the Indian domestic air travel market

F. Flight Price Prediction Using Machine Learning

The study evaluated the models comparing them to are Artificial Neural Network (ANN), Linear Regression, Decision Tree, and Random Forest, evaluated using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). Results show that Decision Tree performs best with the lowest RMSE (0.0061) and MAPE (0.5202), while ANN has the highest RMSE (0.0084) and MAPE (0.6832), suggesting Decision Tree's superior prediction accuracy. Future work could integrate advanced techniques such as genetic algorithms to further enhance accuracy.

G. Airline Price Prediction Using Machine Learning

This research evaluated various algorithms, including Gradient Boosting, KNN, SVR, and Random Forest. Gradient Boosting achieved the lowest standard deviation (62.75), indicating consistent predictions, while KNN exhibited the highest variability. These findings provide valuable insights into the performance of different models for predicting airline ticket prices in a dynamic pricing environment

H. A Survey on Flight Pricing Prediction Using Machine Learning

This paper provides a comprehensive review of factors affecting dynamic flight pricing, such as the day of the week and time of day. Several models, including Random Forest, KNN, Multilayer Perceptron, SVM, and Gradient Boosting, were evaluated, with Random Forest achieving an R^2 of 0.67 and Multilayer Perceptron closely following with 0.65. The research highlighted the effectiveness of regression analysis and metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE) in improving model accuracy. Future studies may benefit from incorporating additional features like seat availability to further optimise predictions.

I. Predicting Flight Prices in India

In the paper analyses a dataset of approximately 10,000 flight records with features such as booking date, airline, source, destination, duration, number of stops, and departure and arrival times. Several machine learning algorithms were evaluated, including Linear Regression, Decision Tree, and Random Forest, for their effectiveness in predicting flight prices. Linear Regression achieved around 60% accuracy, highlighting its limitations in handling non-linear relationships. Decision Tree performed better with an accuracy of approximately 70%, as it managed to capture non-linear patterns in the data. The Random Forest model outperformed the others with an accuracy of around 85%, benefiting from its ensemble structure that captures complex interactions between features. The study underscores the significance of feature engineering and model selection, with Random Forest identified as the most effective algorithm for predicting flight prices in this context.

J. Airline Fare Prediction Using Machine Learning Algorithms

This study analysed three datasets from various sources to estimate flight prices. Seven machine learning models were evaluated, including K-Nearest Neighbours (KNN) Regression, Linear Regression, Lasso Regression, Ridge Regression, Decision Tree Regression, Stacking Regression, and Random Forest Regression. Among these, Random Forest Regressor outperformed the others, achieving an average R^2 score of 0.8 across all datasets. The research emphasised the significance of data visualisation in identifying key factors that affect airline fares, such as airline type, airports, time of day, and number of stops. Future improvements could incorporate data like seat availability to enhance prediction accuracy.

III. Data Collection

Data collection is a central component of this project, as high-quality, recent information is essential for building accurate models. To gather airfare data, we used a combination of web scraping tools and custom Python scripts, allowing us to efficiently capture large volumes of current information from various online sources. Specifically, we extracted data from customer-centric travel websites, ensuring a diverse dataset that reflects real-world pricing trends and consumer behaviour. This resulted in three distinct datasets, each sourced independently, providing a comprehensive foundation for our analysis and model development. By leveraging these varied sources, we aim to build a model that is both robust and well-aligned with the latest market dynamics.

A. Data Collection

For this project, we initially explored publicly available datasets from Kaggle. However, we found that many datasets, including one prominent dataset from 2018, were outdated and thus unrepresentative of current airfare trends. While we considered adjusting the older data to account for inflation, this approach proved insufficient due to the rapid and complex nature of airfare fluctuations. Ultimately, we opted to gather fresh, comprehensive data directly from the web using custom Python scripts.

Over the course of June, July, and August 2024, we collected real-time airfare data from popular Indian travel platforms such as EaseMyTrip and MakeMyTrip. The scraped data focuses on flight details between major metro cities across India, including Delhi, Mumbai, Bangalore, Hyderabad, Kolkata, and Chennai. This data includes approximately 65,000 records, providing an extensive view of domestic airfare dynamics across these high-traffic routes. Key data attributes encompass airline, flight route, source and destination cities, number of stops, layovers, additional flight details, fare prices, departure and arrival timings, and total flight duration.

This method of collection ensures that our dataset is both current and reflective of real-world pricing trends, capturing the nuances of seasonality, demand patterns, and price volatility in the Indian domestic air travel market. These datasets form a solid foundation for constructing models that accurately represent current airfare trends and are well-suited for predictive analysis in a rapidly changing industry.

B. Data Cleaning and Pre-processing

To prepare the data for analysis, we performed a comprehensive cleaning process to enhance its reliability and relevance. First, we addressed null and missing values; since these were minimal, we removed them without impacting the dataset's integrity. Next, we adjusted data types to ensure consistency and, where necessary, split certain attributes to increase their interpret-ability.

Additionally, we reviewed and refined the dataset by adding certain calculated columns and removing others that were redundant or less informative. Outlier detection and removal were also conducted to reduce noise and highlight meaningful patterns within the data.

Following these steps, the data scraped from EaseMyTrip and MakeMyTrip for Indian metro flights over September, October, November and December 2024 is now fully cleaned and ready for in-depth analysis and model building, providing a strong foundation for accurate trend representation.

C. Data Visualization

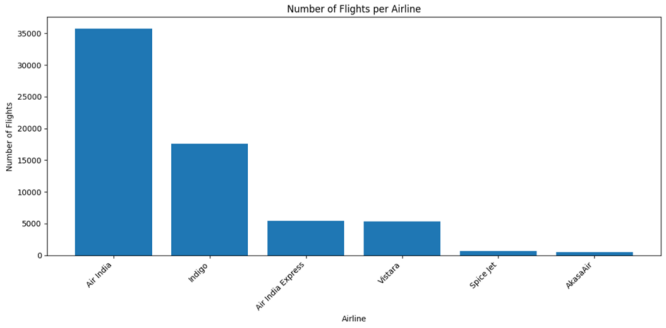

Fig. 1. Number of flights per airline (entire dataset).

The bar chart in Figure 1 illustrates the number of flights operated by various airlines, providing an overview of the distribution of flights across different carriers. The airlines shown include Air India, Indigo, Air India Express, Vistara, Spice Jet, and AkasaAir. Air India has the highest number of flights, significantly outnumbering the other airlines, with over 35,000 flights. Indigo follows as the second most frequent operator with around 20,000 flights, while the other airlines, such as Air India Express, Vistara, Spice Jet, and AkasaAir, show considerably fewer flights.

This visualization helps highlight the market share of each airline in terms of flight frequency, with Air India and Indigo being the major contributors in this dataset. Subsequent figures will further analyse other aspects of airline performance, passenger statistics, and operational metrics to provide a comprehensive view of the aviation industry trends in this study.

Figure2 presents a scatter plot showing the distribution of ticket prices across different airlines, allowing for a comparison of fare variability among carriers. Each point represents

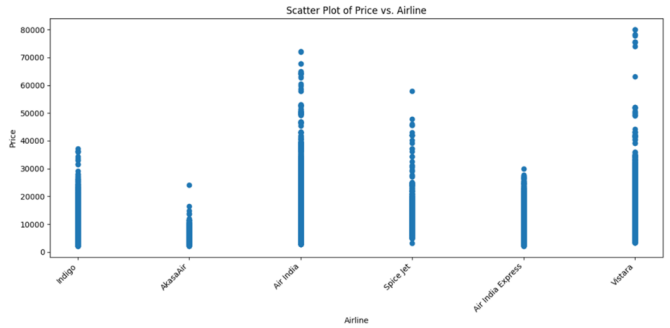

Fig. 2. Scatter Plot of Price vs. Airline.

an individual ticket price for a specific airline, plotted along the y-axis. The plot reveals a wide range of ticket prices for certain airlines, with some airlines exhibiting a larger dispersion in prices. For instance, Air India and Vistara display prices extending up to Rs.80,000, while other airlines like Air India Express have a more limited price range.

This visualization provides insights into pricing patterns across airlines, indicating that some airlines cater to a wider range of price points. In contrast, others maintain a narrower fare band. Such pricing dynamics could be influenced by factors like flight distance, class of service, and time of booking, which will be analysed in subsequent sections.

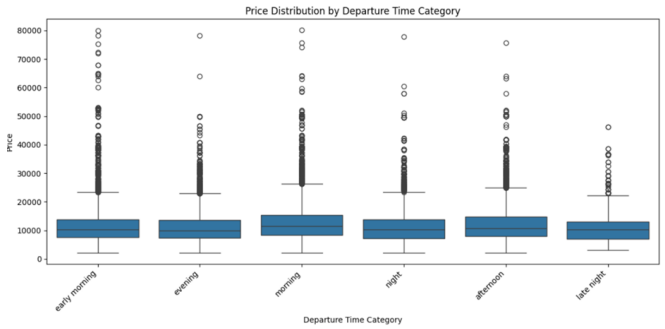

Fig. 3. Distribution of Flight Duration

The boxplot represents the distribution of flight prices across different departure time categories: early morning, evening, morning, night, afternoon, and late night. The central line within each box shows the median price for flights within each time category, while the upper and lower edges of the box represent the interquartile range (IQR). Whiskers extend to depict the range of prices within 1.5 times the IQR from the quartiles, while any points beyond this range are considered outliers and are shown as individual circles.

This visualization reveals that median prices are relatively consistent across all departure time categories. However, all categories exhibit a considerable number of outliers, indicating that certain flights are significantly more expensive than the typical range. Notably, flights departing in the late night and early morning show fewer extreme outliers compared to other categories, potentially suggesting a lower frequency of high-priced flights during these times.

This distributional analysis of prices by departure time aids in identifying patterns and anomalies, which can be critical for understanding how departure timing may impact flight pricing.

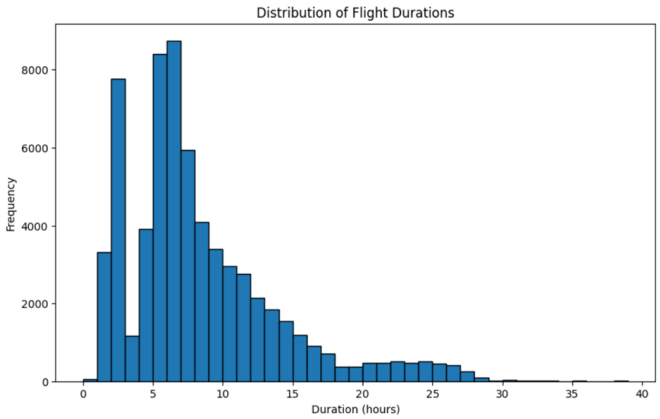

Fig. 4. Scatter Plot of Price vs. Airline.

Presents a histogram depicting the distribution of flight durations in hours. The x-axis represents the duration in hours, while the y-axis indicates the frequency of flights within each duration interval. The distribution is right-skewed, with the majority of flights concentrated in the lower duration range between 0 to 10 hours.

Most flights fall within the 4 to 6-hour range, indicating that these durations are the most common in the dataset. As the duration increases beyond 10 hours, the frequency of flights decreases progressively, with very few flights exceeding 20 hours. A small number of flights with exceptionally long durations (up to nearly 40 hours) are also present, possibly due to long layovers or multi-leg itineraries.

This visualization provides insight into the typical length of flights within the dataset and highlights the variability in flight durations, which may be influenced by factors such as layovers, route distances, and airline schedules.

IV. Machine Learning Models

A. Multiple Linear Regression Model

The Multiple Linear Regression model is a statistical technique used to predict continuous outcomes based on multiple predictor variables. In this study, we employed a Multiple Linear Regression model to analyse the relationship between various parameters and price.

$$\begin{aligned}\text{Price} = \beta_0 &+ \beta_1 \cdot \text{Stops} + \beta_2 \cdot \text{Duration} + \beta_3 \cdot \text{Airline} \\ &+ \beta_4 \cdot \text{Dest City} + \beta_5 \cdot \text{Arrival Time Category} \\ &+ \beta_6 \cdot \text{Src City} + \beta_7 \cdot \text{Departure Time Category} \\ &+ \beta_8 \cdot \text{Days Left} + \epsilon\end{aligned}$$

The performance of the Multiple Linear Regression model was evaluated using various metrics, including the coefficient of determination (R-squared), mean squared error (MSE), and mean absolute error (MAE). The results indicate that the

model explains a significant portion of the variation in prices and provides accurate predictions.

B. Decision Tree Regressor

The Decision Tree Regressor model is a type of supervised learning algorithm used for regression tasks. It works by recursively partitioning the data into smaller subsets based on the most important features and building a tree-like model.

The Decision Tree Regressor model is defined as: Price = f(Departure Time Category, Stops, Duration, Airline, Dest City, Arrival Time Category, Src City, Days Left)

where Price is the continuous outcome variable, Departure Time Category, Stops, Duration, Airline, Dest City, Arrival Time Category, Src City and Days Left are the input features, and f is the decision tree function. The performance of the Decision Tree Regressor model was evaluated using various metrics, including the mean squared error (MSE), mean absolute error (MAE), and R-squared. The results indicate that the model provides accurate predictions and is robust to noisy data.

C. Random Forest Regressor

The Random Forest Regressor model is an ensemble learning algorithm used for regression tasks. It combines multiple decision trees to produce a more accurate and robust prediction model.

D. K Neighbours Regressor

K-Nearest Neighbour (KNN) regression is a non-parametric, supervised machine learning technique used for both regression and classification tasks. It works by predicting the output for a new data point by averaging the outputs of its K nearest neighbours from the training dataset. KNN is considered non-parametric because it makes no assumptions about the underlying data distribution. The algorithm calculates the distance between the new data point and each point in the training set using distance metrics such as Euclidean, Manhattan, or Hamilton distance. After calculating the distances, the K nearest neighbours are selected, and the predicted value is typically the average of the values from those neighbours.

E. Extra Trees Regressor

The Extra Trees Regressor, short for Extremely Randomized Trees Regressor, is an ensemble learning method that builds multiple decision trees and aggregates their predictions for improved accuracy and robustness. It is similar to the Random Forest Regressor, but with a key difference: Extra Trees introduces even more randomness by splitting nodes based on random thresholds rather than optimizing each split for the best outcome. This results in a faster, often more diverse model that can perform well with less risk of overfitting.

F. XGB Regressor

The XGB Regressor, or Extreme Gradient Boosting Regressor, is a powerful and efficient implementation of gradient boosting tailored for regression tasks. It is part of the XGBoost library, which provides a highly optimized, flexible, and scalable machine learning model that combines the predictive power of multiple decision trees in an ensemble. XGB Regressor is known for its high accuracy and speed, making it popular in competitive machine learning and real-world applications.

G. Bagging Regression

Bagging Regression, or Bootstrap Aggregating Regression, is an ensemble technique that aims to improve model stability and accuracy by combining the predictions of multiple base regressors. It works by creating multiple subsets of the training data through random sampling with replacement (bootstrap sampling), training individual models on these samples, and then averaging their predictions. This approach reduces the variance of the final model and enhances performance, especially for complex models prone to overfitting, like decision trees.

H. Ridge Regression

Ridge Regression, also known as L2 regularization, is a linear regression technique that applies a penalty to the magnitude of the model's coefficients to prevent overfitting, especially useful when predictors are highly correlated. Unlike Lasso, Ridge does not set coefficients to zero but rather shrinks them closer to zero, making it suitable for cases where all features contribute to the outcome.

I. Lasso Regression

Lasso Regression, or Least Absolute Shrinkage and Selection Operator, is a type of linear regression that uses regularization to enhance model performance, especially when dealing with high-dimensional data. Lasso adds a penalty equal to the absolute value of the magnitude of coefficients to the loss function, which encourages the model to reduce some coefficients to exactly zero, effectively selecting a simpler model with fewer predictors.

V. Result Analysis

In this study, a single dataset is used to develop machine learning models aimed at accurately predicting flight ticket prices. The dataset is carefully cleaned and pre-processed to ensure the quality of input data for model training and testing. A variety of machine learning algorithms are implemented in Python3, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, K-Nearest Neighbours Regressor, Extra Trees Regressor, Gradient Boosting Regressor, XGBoost Regressor, Bagging Regressor, Ridge Regression, and Lasso Regression.

A. Feature Selection and Feature Extraction

To determine the optimal feature set for predicting flight prices, a systematic feature engineering and selection approach was implemented. Initially, the predictive power of each feature was examined individually by calculating the accuracy achieved when using only one feature at a time. Features tested in this manner included the number of stops, duration, airline, destination city, arrival time category, source city, and departure time category, days left.

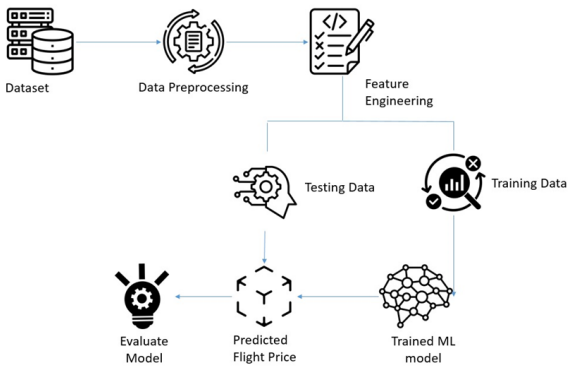

Fig. 5. Machine Learning Architecture Implemented

TABLE I
SINGLE FEATURE SELECTION

Feature	R2_Score	MAE	RMSE
airline	0.1906	3586.8051	5092.5364
src_city	0.015069	4116.6636	5617.7893
dest_city	0.027707	4086.9855	5581.6303
duration	0.230977	3395.5679	4964.0056
days_left	0.030803	4063.02269	5573.3102
depparture_time_category	0.013338	4119.6955	5622.7210
arrival_time_category	0.01785	5609.8269	5609.8269

Subsequently, the number of features was progressively increased, combining two, three, and four features in various configurations, and the accuracy of each combination was evaluated using the k-nearest neighbors (KNN) algorithm. This iterative process allowed for a detailed assessment of each feature's contribution to the model's predictive performance. The results indicated that the combination

TABLE II
DOUBLE FEATURE SELECTION

Features	R2_Score	MAE	RMSE
airline, src_city	0.2229	3522.4487	4989.8721
airline, dest_city	0.2321	3445.9872	4960.3558
airline, duration	0.3894	2932.3085	4423.2347
airline, days_left	0.2397	3438.0053	4935.6084
airline, departure_time_category	0.2068	3541.6584	5041.2092
airline, arrival_time_category	0.2118	3524.3500	5025.4597
src_city, dest_city	0.0643	3997.6398	5475.3871
src_city, duration	0.2935	3194.8353	4757.7507

of four specific features—source city (src_city), destination city (dest_city), duration, and departure time slot (departure_time_category)—yielded the highest accuracy of 96.5%. This suggests that these features collectively capture critical information for accurate prediction, likely due to their combined influence on travel time, route, and pricing patterns.

This feature selection approach demonstrates that targeted feature engineering can significantly enhance model accuracy by identifying the most relevant predictors. The final selected feature set of src_city, dest_city, duration, and departure_time_category forms the basis of the model, maximizing predictive accuracy and providing insights into the key factors that drive flight prices.

The data is then split into training and testing subsets, where the training data is used to fit the models, and the testing data is utilized to evaluate model prediction accuracy.

TABLE III
TRIPLE FEATURE SELECTION

Features	R2_Score	MAE	RMSE
airline, src_city, dest_city	0.3101	3224.4995	4701.3726
airline, src_city, duration	0.4648	2682.7768	4140.450
airline, src_city, days_left	0.2898	3336.3794	4770.1079
airline, src_city, departure_time_category	0.2566	3434.6565	4880.4990
airline, src_city, arrival_time_category	0.2542	3419.8431	4888.2354
days_left, day_of_week, month	0.0306	4063.0226	5573.3102

VI. ERROR ANALYSIS

A. R^2 or Coefficient of Determination

R^2 quantifies a regression model's ability to explain variance in the target variable relative to a baseline, often the mean of the observed data. As a scale-free metric ranging from 0 to 1, R^2 allows for meaningful comparison across models and datasets, with higher values indicating a better fit. Tracking R^2 over time enables researchers to measure model improvements, showing whether adjustments increase the model's explanatory power.

$$R^2 = 1 - \frac{R_{\text{ss}}}{T_{\text{ss}}}$$
$$R_{\text{ss}} = \text{Sum of squares of residuals}$$
$$T_{\text{ss}} = \text{Total sum of squares}$$

B. Mean Absolute Error

The Mean Absolute Error (MAE) measures a regression model's performance by calculating the average absolute difference between predicted and actual values. Unlike R^2, MAE is an absolute metric that reflects the model's average prediction error in the same units as the target variable, making it straightforward to interpret. Lower MAE values indicate higher accuracy, and tracking MAE over time can reveal whether model adjustments are reducing errors and improving predictive performance.

$$\text{MAE} = \frac{\text{SAE}}{n}$$
$$\text{SAE} = \text{Sum of Absolute Errors}$$
$$n = \text{Total number of points in data sets}$$

C. Root Mean Squared Error

MSE evaluates a regression model's accuracy by calculating the average of the squared differences between predicted and actual values. As a quadratic metric, MSE penalizes larger errors more heavily, making it especially sensitive to outliers. Lower MSE values indicate a more accurate model, and tracking MSE over time can reveal improvements in predictive

performance as model adjustments reduce the magnitude of errors.

$$\text{MSE} = \frac{1}{n}\sum(y_i - y_o)^2$$
$$\text{RMSE} = \sqrt{\text{MSE}}$$
$$y_i = \text{Predicted value}$$
$$y_o = \text{Actual value}$$
$$n = \text{Number of observations}$$

D. Mean Absolute Percentage Error

MAPE assesses a regression model's accuracy by calculating the average absolute percentage difference between predicted and actual values. As a percentage-based metric, MAPE is scale-independent, allowing for easy comparison across datasets with different units or magnitudes. Lower MAPE values indicate better model accuracy, and tracking MAPE over time helps determine if model adjustments are consistently reducing relative errors in predictions.

$$\text{MAPE} = \frac{1}{n} \times 100 \times \sum \left| \frac{y_o - y_i}{y_o} \right|$$
$$y_i = \text{Predicted value}$$
$$y_o = \text{Actual value}$$
$$n = \text{Number of observations}$$

TABLE IV
COMPARISON OF EVALUATION METRICS

Model	R2_Score	MAE	RMSE	MAPE
Multiple Linear Regression	0.3048	3290.613	4179.708	31.22%
Decision Tree Regressor	0.6144	1647.143	3514.948	13.67%
Random Forest Regressor	0.7588	1515.948	2779.826	12.77%
K Neighbours Regressor	0.5038	2541.273	3987.226	22.35%
Extra Trees Regressor	0.7210	1496.134	2989.931	12.49%
Gradient Boosting Regressor	0.4400	2577.210	4235.991	20.43%
Bagging Regressor	0.7357	1594.880	2909.765	13.42%
Ridge Regression	0.3047	3290.623	4719.750	31.21%
Lasso Regression	0.3048	3290.591	4719.713	31.22%

VII. CONCLUSION

To estimate the dynamic fare of flights, data was scraped from a travel website, gathering information from three different sources. Data visualization uncovered several valuable insights from this dataset. Seven different machine learning algorithms were tested to develop a predictive model. Due to data limitations—considering the dataset was sourced from an online travel platform—only a certain level of information was accessible. The model's accuracy and reliability were assessed through evaluation metrics. Among the algorithms tested, K-Nearest Neighbours (KNN) achieved the highest accuracy, outperforming other models in predicting airline fares. This suggests that KNN is highly effective for this application

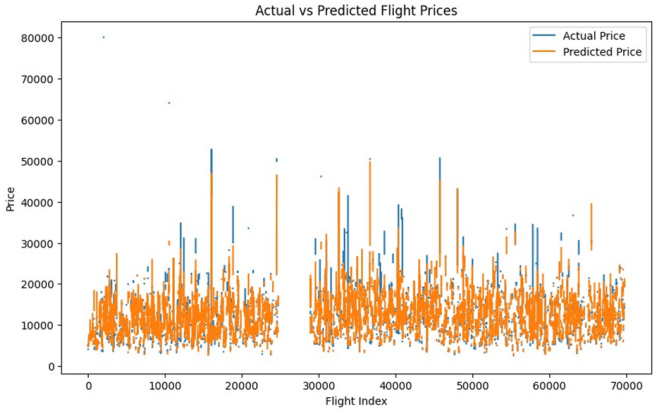

Fig. 6. Line plot for Actual vs Predicted Flight Prices

TABLE V
COMPARISON OF EVALUATION METRICS

Model	Accuracy
Multiple Linear Regression	0.1232
Decision Tree Regressor	0.9846
Random Forest Regressor	0.9273
K Neighbours Regressor	0.8613
Extra Trees Regressor	0.9432
Gradient Boosting Regressor	0.2732
Bagging Regressor	0.9301
Ridge Regression	0.1320
Lasso Regression	0.1234

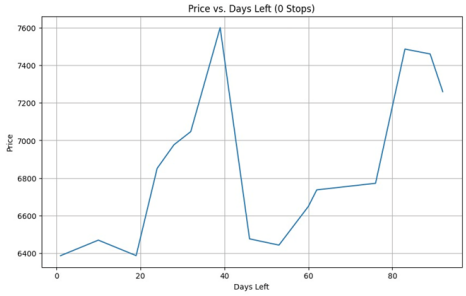

Fig. 7. Plot for price vs days left

Fig. 8. Scatter plot for Actual vs Predicted Flight Prices

VIII. Scope of Improvement

One promising direction for improving predictive accuracy in flight price modeling is the integration of additional data features. For instance, real-time seat availability data could provide insight into demand fluctuations, allowing the model to respond more accurately to price changes driven by capacity constraints. This data, when combined with other external factors—such as fuel prices, weather conditions, or even major events and public holidays—can add a layer of context to the predictions. Including these dynamic factors would allow the model to recognize broader patterns in consumer demand that influence pricing, especially during peak travel seasons or special events.

Another area for enhancement is dataset quality and diversity. Expanding the dataset to include a wider variety of routes, airlines, and geographic regions would create a more comprehensive model that generalizes well across different contexts. A larger, heterogeneous dataset could better capture the variations in pricing strategies used by different airlines or for different routes. Additionally, having access to real-time data updates on these variables would enable the model to stay relevant in fast-changing market conditions, reducing the chances of overfitting to outdated data trends.

Model optimization is another critical factor for enhancing predictive performance. Conducting extensive hyperparameter tuning, experimenting with advanced algorithms like ensemble methods, or integrating deep learning models for complex pattern recognition can lead to substantial accuracy gains. Hyperparameter tuning could optimize model parameters for each feature's importance, improving the model's sensitivity to price determinants. Advanced models, such as recurrent neural networks (RNNs), could also be effective, especially if time-series data on prices becomes more detailed, helping the model account for temporal dependencies in ticket pricing.

Lastly, refining the evaluation criteria and adding interpretability features could also improve the model's application in real-world settings. For example, using cross-validation techniques across varied subsets can give a more accurate picture of the model's performance. Implementing interpretability tools like SHAP values or feature importance plots would also allow users to understand how different variables impact predictions, increasing trust in the model's results and making it more user-friendly for commercial deployment.

Prediction-based services are now common across various fields, from stock forecasting tools for brokers to tools like Zestimate for housing values. A similar service could benefit the aviation industry, assisting customers in making more informed ticket purchase decisions.

References

[1] M. Mulkalla, Deepika and A. Joshi, "Predicting the Fare of a Flight Ticket with Machine Learning Algorithms," 2022 International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 2022

[2] V. H. Vu, Q. T. Minh and P. H. Phung, "An airfare prediction model for developing markets," 2018 International Conference on Information Networking (ICOIN), Chiang Mai, Thailand, 2018.

[3] Supriya Rajankar, Neha Sakharkar, 2019, A Survey on Flight Pricing Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 08, Issue 06 (June 2019)

[4] R. R. Subramanian, M. S. Murali, B. Deepak, P. Deepak, H. N. Reddy and R. R. Sudharsan, "Airline Fare Prediction Using Machine Learning Algorithms," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India

[5] Joshi, Achyut & Sikaria, Himanshu & Devireddy, Tarun. (2017). Predicting Flight Prices in India.

[6] Alapati, Naresh & Prasad, B V V S & Sharma, Aditi & Kumari, G.R.P. & Veeneetha, S.V. & Srivalli, N. & Lakshmi, T. & Sahitya, D.. (2022). Prediction of Flight-fare using machine learning. 134-138. 10.1109/ICFIRTP56122.2022.10059429.

[7] Ankita Panigrahi, Rakesh Sharma, Sujata Chakravarty, Bijay Paikaray, Harshvardhan Bhoyar. Flight Price Prediction Using Machine Learning. In Sarika Jain 0001, Sven Groppe, Nandana Mihindukulasooriya, editors. Volume 3283 of CEUR Workshop Proceedings, pages 172-178, CEUR-WS.org, 2022.

[8] Chawla, Bhavuk, Ms. Chandandeep Kaur and I II Data Preparation. "Airfare Analysis And Prediction Using Data Mining And Machine Learning." (2017).

[9] Lu, Jun. (2017). Machine learning modeling for time series problem: Predicting flight ticket prices. 10.48550/arXiv.1705.07205.

[10] A Framework for Predicting Airfare Prices Using Machine Learning Heba Mohammed Fadhil, Mohammed Najm Abdullah, Mohammed Issam Younis Iraqi Journal of Computers, Communications, Control and systems engineering 2022, Volume 22, Issue 3, Pages 81-96