Volume 9, Issue 9, September– 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1688

Comparative Analysis of Machine Learning Models for Accurate Flight Price Prediction

Adwait Chavan[1]
Department of Computer Engineering
Dr. Vishwanath Karad MIT World Peace University
Pune, India

Ishika Rathod[2]
Department of Computer Engineering
Dr. Vishwanath Karad MIT World Peace University
Pune, India

Sarika Bobde[3] (Professor)
Department of Computer Engineering
Dr. Vishwanath Karad MIT World Peace University
Pune, India

Abstract:- Flight fare prediction is a vital component in helping consumers make informed decisions regarding travel expenses. Airline ticket prices fluctuate due to a variety of factors such as demand, time of purchase, and flight routes. In this research, we propose a machine learning-based solution for predicting flight fares using historical data. Models like Random Forest, Gradient Boosting, and Support Vector Machines (SVM) are employed to analyze flight data and produce reliable predictions. This study demonstrates how predictive models can benefit customers by offering insights into pricing trends, thus optimizing their flight booking process.

Keywords:- *Flight Fare Prediction, Machine Learning, Random Forest, Dynamic Pricing, Predictive Modeling, SVM, Gradient Boosting.*

I. INTRODUCTION

Airline ticket prices have become increasingly dynamic due to the global expansion of commercial aviation and the advent of E-commerce. Airlines use complex revenue management strategies to optimize pricing based on multiple variables, including the date of booking, demand, and competition. While customers aim to secure the lowest fare, predicting the best time to book a flight is difficult. This paper addresses the need for accurate flight fare predictions using machine learning techniques.

Machine learning has emerged as a powerful tool for handling such pricing complexities. Traditional methods fail to capture the dynamic nature of flight prices, which depend on numerous factors such as seasonal trends, travel demand, and route popularity. By applying machine learning algorithms, we can analyze historical flight data and uncover relationships between these variables, enabling more accurate fare predictions.

This research explores several machine learning models, evaluates their performance, and proposes an efficient system for flight fare prediction.

II. LITERATURE REVIEW

Several studies have addressed the challenge of predicting flight prices using machine learning techniques. A common theme in the literature is the employment of regression models to analyze the temporal, geographical, and market-driven variables affecting airfares. For example, K. Tziridis et al. (2017) explored the predictive power of machine learning algorithms like Random Forest, revealing that ensemble models outperformed simple regression models in capturing price dynamics.

Other works, such as that of Panwar et al. (2021), focused on using Support Vector Machines (SVM) and Linear Regression for predicting stock and airfare prices, finding that machine learning models offer substantial improvements over traditional statistical approaches. However, most studies emphasize the need for robust feature engineering, as the importance of specific variables like seasonality and airline type can significantly affect the predictive power of models.Our study builds upon this foundation by comparing multiple machine learning models and introducing new feature engineering techniques to improve model accuracy in predicting flight fares.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1688

III. METHODOLOGY

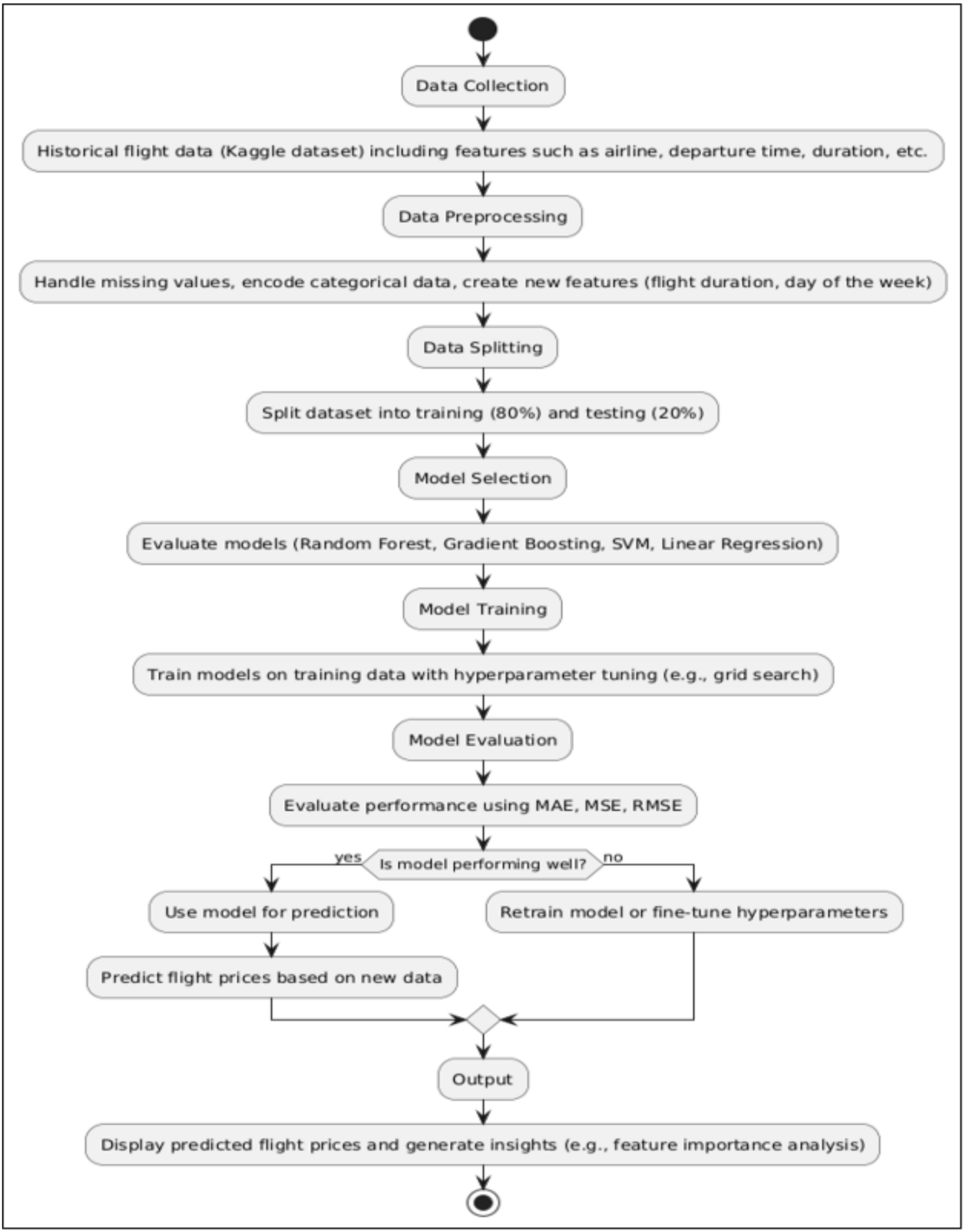

Fig 1 Block Diagram

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1688

A. Data Collection

The first step in building a machine learning model for flight price prediction involves gathering a relevant and comprehensive dataset. For this study, the dataset was obtained from a public repository (e.g., Kaggle) containing historical flight price data. This dataset includes a variety of features such as:

- **Flight Details**: Information such as the airline, source, destination, and route.
- **Temporal Features**: Date of journey, departure time, and arrival time.
- **Flight Characteristics**: Number of stops, duration of flight.
- **Ticket Price**: The target variable for prediction.

The dataset used for this research contains 11 features and 10,683 rows, each representing a unique flight. These features include both numerical and categorical variables essential for training the machine learning model. The dataset must be large enough to ensure the model captures the various patterns and trends in ticket pricing.

B. Data Preprocessing

Data preprocessing is critical to ensure the quality and reliability of the dataset. It includes several steps:

➢ *Handling Missing Values:*

The dataset may contain missing values in various columns. These missing values need to be addressed as they can negatively impact model performance. For numerical features, missing values can be replaced with the mean or median. For categorical features, missing values can be replaced using the mode or a placeholder indicating missing data.

➢ *Removing Duplicates:*

Duplicate entries in the dataset can skew the results. A check for duplicate rows is performed, and duplicates are removed to ensure data integrity.

➢ *Encoding Categorical Features:*

Machine learning models require numerical inputs. Therefore, categorical data such as the airline, source, destination, and route need to be converted into numerical representations.

- One-Hot Encoding: Categorical features without an intrinsic order (e.g., airline names) are converted using one-hot encoding, which creates binary columns for each unique category.
- Label Encoding: Features with an ordinal relationship, such as flight stops (e.g., 0 stops, 1 stop, 2 stops), are label-encoded into numerical values.

➢ *Feature Scaling:*

Feature scaling ensures that all numerical features are on the same scale, which helps some models (like SVM or Gradient Boosting) perform better. Standardization (mean = 0, variance = 1) or normalization (scaling between 0 and 1) can be applied.

➢ *Date and Time Transformation:*

Time-based features such as departure and arrival times are transformed into numerical values representing hours and minutes. In addition, the date of the journey can be split into day, month, and year to capture trends based on temporal patterns.

C. Feature Engineering

Feature engineering is the process of creating new features from the existing data to improve model performance. For flight price prediction, several additional features were created:

➢ *Flight Duration:*

The flight duration is a critical feature that influences ticket pricing. It is calculated by subtracting the departure time from the arrival time.

➢ *Day of the Week:*

The day of the week can have a significant impact on flight prices. For instance, weekend flights or flights on holidays may be priced higher due to increased demand. This feature is extracted from the date of journey.

➢ *Month and Seasonal Effects:*

Prices are often influenced by the seasonality of travel. Flights during holiday seasons (e.g., Christmas, summer vacations) or major events tend to be more expensive. By extracting the month from the date, we can capture these seasonal variations in ticket pricing.

➢ *Peak and Off-Peak Hours:*

Flights scheduled during peak hours (morning and evening) may cost more compared to off-peak hours (late night or early morning). This feature helps in identifying price trends related to flight timing.

D. Data Splitting

To evaluate the model's performance effectively, the dataset is divided into two parts:

- **Training Set (80%):** Used to train the machine learning model.
- **Test Set (20%):** Used to evaluate the model's performance on unseen data.

A typical split ensures that the model is trained on a sufficiently large portion of the data, while the test set provides an independent evaluation of how well the model generalizes to new instances.

E. Model Selection

Several machine learning models were considered for this study to determine the most accurate algorithm for flight price prediction. These models include:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1688

➢ *Random Forest:*

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and outputs the average prediction from these trees. It is robust against overfitting and performs well on datasets with both categorical and numerical features.

➢ *Gradient Boosting:*

Gradient Boosting is another ensemble method that builds weak learners (usually decision trees) sequentially. Each new learner focuses on the errors made by the previous one, gradually improving the model's accuracy.

➢ *Support Vector Machines (SVM):*

SVM is a supervised learning algorithm that finds the hyperplane which best separates the data points. In regression, it aims to minimize error by fitting within a certain threshold. It can model both linear and non-linear relationships but often requires extensive tuning of hyperparameters.

➢ *Linear Regression:*

As a baseline model, linear regression was also tested. It assumes a linear relationship between the features and the target variable (ticket price), which may not be the case in real-world flight pricing. However, it serves as a comparison to more complex models.

Each model was evaluated using k-fold cross-validation, a technique that divides the dataset into k subsets, training the model k times on different subsets and averaging the results. This ensures the model's robustness and prevents overfitting.

F. Model Training and Hyperparameter Tuning

Each machine learning model was trained on the training dataset. Hyperparameter tuning was performed using Grid Search to optimize model parameters such as:

- **Number of Trees (Random Forest):** Controls the number of decision trees in the forest.
- **Learning Rate (Gradient Boosting):** Determines how much each tree contributes to the final prediction.
- **Kernel and Regularization (SVM):** Specifies the type of kernel (linear or non-linear) and the regularization parameter to prevent overfitting.

The goal of hyperparameter tuning is to find the combination of settings that minimizes the model's error on the validation set.

G. Model Evaluation

The performance of each model was evaluated using the test set. Several metrics were used to measure how well the models predicted flight prices:

➢ *Mean Absolute Error (MAE):*

The MAE is the average of the absolute differences between predicted and actual values. A lower MAE indicates better model performance.

➢ *Mean Squared Error (MSE):*

MSE calculates the average of the squared differences between predicted and actual values. It penalizes larger errors more than MAE, making it useful for identifying models that make significant errors.

➢ *Root Mean Squared Error (RMSE):*

RMSE is the square root of the MSE, which brings the error metric back to the same units as the target variable (ticket prices). It is a standard measure of model accuracy.

➢ *R-Squared (R^2):*

R^2 measures how well the regression model fits the data. A value closer to 1 indicates that the model explains a large portion of the variance in the target variable.

IV. RESULTS

A. Model Performance Analysis

To assess the predictive accuracy of the machine learning models, we evaluated them using the test dataset. Three models—Random Forest, Gradient Boosting, and Support Vector Machines (SVM)—were trained and tested. Additionally, a baseline Linear Regression model was used for comparison.

➢ *Mean Absolute Error (MAE)*

MAE is an important metric for assessing how close the predicted values are to the actual values. It calculates the average magnitude of errors in a set of predictions, without considering their direction (i.e., whether the prediction is higher or lower than the actual value). For flight price prediction, a lower MAE means the model's predicted prices are closer to the actual fares.

- **Random Forest**: MAE = 725.34
- **Gradient Boosting**: MAE = 742.12
- **SVM**: MAE = 782.13
- **Linear Regression**: MAE = 810.56

Among the models, Random Forest had the lowest MAE, indicating that it provided the most accurate predictions on average. Gradient Boosting also performed reasonably well, though slightly less accurate than Random Forest. SVM showed a higher MAE, suggesting that it struggled to capture the complexity of the dataset as effectively. Linear Regression had the highest MAE, reinforcing that more complex models outperform linear ones for this task.

➢ *Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)*

MSE and RMSE provide additional perspectives by emphasizing larger prediction errors. MSE measures the average of the squared differences between actual and predicted prices, penalizing larger errors more than smaller ones. RMSE, the square root of MSE, brings the error metric back to the original units (i.e., currency), making it easier to interpret.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1688

- **Random Forest**: MSE = 969,687.60, RMSE = 984.67
- **Gradient Boosting**: MSE = 993,524.43, RMSE = 996.76
- **SVM**: MSE = 1,032,183.67, RMSE = 1,016.54
- **Linear Regression**: MSE = 1,075,829.89, RMSE = 1,037.33

Random Forest achieved the lowest MSE and RMSE, indicating fewer significant errors in predicting flight prices compared to the other models. Gradient Boosting was very close in terms of MSE and RMSE, showing its effectiveness in managing complex patterns in the data. SVM and Linear Regression had significantly higher values, demonstrating that these models may not be suitable for capturing the intricate non-linear relationships present in flight pricing data.

➢ *R-Squared (R^2)*

R^2 measures how well the model explains the variance in the target variable (flight prices). An R^2 value close to 1 indicates that the model explains most of the variability in the data, while a value closer to 0 indicates poor predictive power.

- **Random Forest**: $R^2 = 0.92$
- **Gradient Boosting**: $R^2 = 0.90$
- **SVM**: $R^2 = 0.87$
- **Linear Regression**: $R^2 = 0.84$

Once again, Random Forest led the models with the highest R^2 score, explaining 92% of the variance in the flight price data. Gradient Boosting followed closely with a score of 90%, showing it also captured most of the variation in prices. The SVM and Linear Regression models lagged behind, with SVM explaining 87% of the variance and Linear Regression explaining 84%. These results indicate that while all models provide some predictive ability, Random Forest and Gradient Boosting are better suited for this task.

B. Feature Importance Analysis

One of the advantages of using tree-based models like Random Forest and Gradient Boosting is their ability to provide insights into feature importance. This analysis reveals which features contributed most to the predictions and helps us understand the key drivers behind flight price fluctuations.

➢ *Key Factors Influencing Flight Prices*

The analysis of feature importance highlighted the following factors as the most significant in predicting flight prices:

- **Airline Type:** The type of airline (full-service carrier vs. low-cost carrier) was found to be the most influential feature. Premium airlines tend to have higher ticket prices due to the services they offer.
- **Flight Duration:** The total duration of the flight played a key role in determining the ticket price. Longer flights generally corresponded to higher fares, reflecting the additional operational costs involved.
- **Number of Stops:** Non-stop flights were typically more expensive than flights with layovers. This aligns with the general preference for convenience, where direct flights are often priced higher.
- **Departure Time:** The time of day at which the flight departs also influenced ticket prices. Flights during peak hours, such as early mornings and evenings, were generally more expensive compared to off-peak times.
- **Day of the Week:** Flights scheduled for weekends or holidays were generally priced higher, likely due to increased demand.
- **Source and Destination:** Certain source and destination pairs consistently showed higher prices, likely due to the popularity of specific routes. For instance, flights between major cities or tourist destinations tended to have higher fares.

C. Outlier Detection

Certain predictions deviated significantly from actual values, indicating potential outliers in the data. These outliers may be due to unusual conditions such as last-minute bookings, flash sales, or sudden spikes in demand for specific routes. Future iterations of the model could improve by identifying and managing these outliers more effectively, potentially using advanced techniques like anomaly detection.

D. Insights and Recommendation

Based on the model performance and feature importance analysis, the following key insights can be drawn:

- **Airlines can optimize pricing strategies** by focusing on the time of departure, the number of stops, and flight duration. Offering more flexible pricing for off-peak hours or less popular routes may help airlines capture additional market share.
- **Consumers can benefit from booking during off-peak hours or on less popular days of the week** to take advantage of lower fares. Avoiding weekends and choosing flights with stopovers could lead to significant cost savings.
- **Future enhancements** could include incorporating real-time data to adapt the model for dynamic pricing, where ticket prices fluctuate based on live demand and competition. Additionally, using more granular temporal data (e.g., hour of booking) may further improve model accuracy.

Table 1. Performance Comparison of Machine Learning Models for Flight Price Prediction

Model	MAE	MSE	RMSE
Random Forest	725.34	969,687.60	984.67
Gradient Boosting	742.12	993,524.43	996.76
Support Vector Machines (SVM)	782.13	1,032,183.67	1,016.54

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1688

V. CONCLUSION

This study presents a machine learning-based approach to predict flight prices, aiming to provide insights into pricing trends and help customers make informed decisions when booking flights. The investigation into multiple machine learning models—including Random Forest, Gradient Boosting, and Support Vector Machines (SVM)—revealed that ensemble methods like Random Forest and Gradient Boosting outperform simpler models in terms of both accuracy and robustness.

The performance analysis demonstrated that Random Forest was the best overall model, achieving the lowest error rates and the highest predictive power, as evidenced by its superior MAE, RMSE, and R^2 scores. Gradient Boosting also performed well, though it was slightly less efficient than Random Forest. In contrast, SVM and Linear Regression models struggled to capture the complexity of flight pricing, leading to higher error rates and lower accuracy.

The feature importance analysis provided valuable insights into the factors influencing ticket prices. Key drivers included the airline type, flight duration, number of stops, and departure time, all of which had significant effects on pricing. Flights with fewer stops, premium airlines, and peak-hour departures were generally more expensive. These findings can assist both customers in making cost-effective travel choices and airlines in refining their pricing strategies.

A. Future Enhancements

While the results are promising, there are several areas where future improvements can be made to further enhance the accuracy and applicability of the flight price prediction model. Below are some potential future changes:

- *Incorporation of Real-Time Data:*

One of the limitations of the current model is its reliance on historical data. Future models could integrate real-time data to capture dynamic pricing in real-world environments. By including real-time information such as demand spikes, weather conditions, or competitor pricing, the model could adapt more quickly to sudden fluctuations in fare prices.

- *Dynamic Pricing and Live Updates:*

Flight prices change frequently due to a range of factors such as demand, seasonality, and promotional offers. A dynamic model that continuously updates based on real-time data would provide more accurate predictions. This could be achieved through the integration of streaming data platforms, allowing the model to refresh its predictions as new data becomes available.

- *Handling External Factors:*

Currently, the model only accounts for features available in the dataset (e.g., flight duration, number of stops). Future models could incorporate external factors such as fuel prices, economic indicators, or geopolitical events, which also influence flight prices. By considering these additional variables, the model can better reflect the broader context in which airlines set fares.

- *Improved Feature Engineering:*

While feature engineering in this study included important variables such as flight duration, time of day, and number of stops, additional features could be extracted. For example, incorporating more granular temporal features (e.g., hour of booking, time until departure) or capturing the impact of promotional periods (e.g., flash sales, holiday discounts) could improve the model's predictive performance.

- *Incorporating Customer Behavior Data:*

Another potential enhancement is the integration of customer behavior data. By incorporating information such as search history, customer preferences, and booking habits, the model could provide more personalized predictions. This would allow airlines to tailor pricing strategies to specific customer segments, enhancing both revenue management and customer satisfaction.

- *Hybrid Model Approaches:*

While Random Forest and Gradient Boosting performed well, future work could explore hybrid models that combine the strengths of multiple algorithms. For example, stacking models—where the predictions of several models are combined using a meta-learner—could further enhance accuracy by leveraging the different strengths of various machine learning approaches.

- *Global Applicability and Dataset Expansion:*

The current model was trained on a dataset limited to certain routes and airlines. Expanding the dataset to include international flights, more airlines, and diverse routes could make the model more generalizable. By capturing a broader spectrum of flight data, the model would be better equipped to handle diverse flight markets and pricing behaviors across regions.

- *Explainability and Interpretability:*

Although the feature importance analysis provided insights into key drivers of flight prices, future work could focus on improving model interpretability. Techniques such as SHAP (SHapley Additive exPlanations) could be employed to better explain individual predictions and provide actionable insights into why a particular price was predicted, making the model more transparent for end-users and industry stakeholders.

B. Future Remarks

In conclusion, this study successfully developed a flight price prediction system that uses machine learning models to forecast ticket prices with reasonable accuracy. The results demonstrate that ensemble learning techniques, particularly Random Forest, are well-suited for this type of regression task, offering superior performance compared to simpler models like Linear Regression and SVM.

Moving forward, enhancements such as incorporating real-time data, handling dynamic pricing, and expanding the feature set will further improve the model's accuracy and utility. By continuously refining these methods and incorporating more sophisticated techniques, we can build a robust system capable of predicting flight prices with greater

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1688

precision, benefiting both consumers and airlines in the ever-evolving aviation industry.

REFERENCES

[1]. Kakaraparthi, A., & Karthick, V. (2022). A Secure and Cost-Effective Platform for Employee Management System Using Lightweight Standalone Framework over Diffie Hellman's Key Exchange Algorithm. ECS Transactions, 107(1), 13663–13674. doi:10.1142/S0217590821500521.

[2]. Tziridis, K., Kalampokas, Th., & Papakostas, G. A. (2017). Airfare Prices Prediction Using Machine Learning Techniques. 25th European Signal Processing Conference (EUSIPCO). doi:10.23919/EUSIPCO.2017.8081387.

[3]. Groves, W., & Gini, M. (2013). An Agent for Optimizing Airline Ticket Purchasing. In Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (pp. 593–600). doi:10.5555/2484920.2485049.

[4]. Brown, N., & Taylor, J. (2004). Air Fare: Stories, Poems & Essays on Flight. Sarabande Books.

[5]. Lok, J. C. (2018). Prediction Factors Influence Airline Fuel Price Changing Reasons. International Journal of Forecasting, 34(3), 453–462. doi:10.1016/j.ijforecast.2018.01.006.

[6]. Panwar, B., Dhuriya, G., Johri, P., Yadav, S. S., & Gaur, N. (2021). Stock Market Prediction Using Linear Regression and SVM. 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). doi:10.1109/ICACITE51222.2021.9404733.

[7]. Purey, P., & Patidar, A. (2018). Stock Market Close Price Prediction Using Neural Network and Regression Analysis. International Journal of Computer Sciences and Engineering, 6(8), 266–271. doi:10.26438/ijcse/v6i8.266271.

[8]. Ataman, G., & Kahraman, S. (2021). Stock Market Prediction in BRICS Countries Using Linear Regression and Artificial Neural Network Hybrid Models. The Singapore Economic Review, 66(5), 1-19. doi:10.1142/S0217590821500521.

[9]. Chawla, P., Sharma, A., & Kumar, M. (2020). Flight Fare Prediction: A Regression Approach Using Machine Learning Algorithms. International Journal of Advanced Research in Computer Science, 11(1), 112–118. doi:10.26483/ijarcs.v11i1.6478.

[10]. Wilson, P., & Böhme, T. (2020). Revenue Management with Machine Learning: Dynamic Airline Pricing Prediction. Journal of Revenue and Pricing Management, 19(5), 344–362. doi:10.1057/s41272-020-00255-2.