Airfare Dynamics: Machine Learning Insights into Airline Fares and Passenger Behavior

Link to the demo and code: https://colab.research.google.com/drive/1sABuRX6OtXLJ7UAstXBde0tUso_do02w?usp=sharing

Advaithbarath Raghuraman Bhuvaneswari
MS in Cybersecurity and Privacy
New Jersey Institute of Technology
New Jersey, USA.
ar2728@njit.edu

Srivatsan Jayaraman
MS in Computer Science
New Jersey Institute of Technology
New Jersey, USA.
sj796@njit.edu

Aravind Kalyan Sivakumar
MS in Data Science Computational Track
New Jersey Institute of Technology
New Jersey, USA.
as4588@njit.edu

***Abstract*—Airline fares are influenced by a variety of factors, including route distances, passenger demand, market competition, and seasonal trends. These fluctuations make it challenging for consumers to secure optimal prices and for airlines to maintain profitability. The advent of machine learning in airfare prediction has provided innovative solutions for understanding these dynamics and assisting stakeholders in making informed decisions. This research leverages the Consumer Airfare Report dataset to develop reliable machine learning models for predicting airfare trends and analyzing market competition. A comprehensive evaluation of multiple machine learning algorithms—including Neural Network XGBoost, Random Forest and LightGBM, Regression—was performed to ensure the accuracy and robustness of the models. Performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and model accuracy were used to assess and compare their effectiveness. The Random Forest model might seems consistently outperformed others with the highest R^2 score and minimal error rates. However, due to large datasets and overfitting issues, LightGBM performs much better in this scenario. Advanced feature engineering techniques were applied to transform raw data into actionable insights, further enhancing the model's predictive power. This application exemplifies the potential of machine learning to address the complexities of airfare pricing and promote data-driven decision-making for all consumers.**

***Index Terms*—Machine Learning, Airfare Prediction, Model Deployment, Performance Metrics, Feature Engineering**

I. Introduction

Organizing air travel for leisure, business, or personal purposes often poses significant challenges due to the unpredictable nature of airline ticket pricing. Airlines typically offer discounted fares when tickets are purchased well in advance, but prices tend to escalate as the departure date nears. However, the timing of the booking is just one factor affecting ticket costs. Other variables, such as route demand, market competition, seasonal trends, and promotional strategies, also play a significant role in determining prices.

Consumers often face difficulty in finding the most affordable flight options because of the complex and opaque pricing mechanisms utilized by the airline industry. Airfare prices are highly volatile and subject to sudden changes influenced by factors like peak travel seasons, economic conditions, and shifts in consumer preferences. In this competitive landscape, navigating the frequent and unpredictable fluctuations in ticket prices becomes a daunting task for travelers seeking to secure cost-effective travel options.

Analyzing passenger behavior is integral to understanding trends within the aviation sector, as it directly influences fare structures and service offerings. By evaluating patterns in passenger volumes, travel preferences, and popular routes, airlines can forecast demand and adjust their strategies accordingly. For instance, an increase in passengers on certain routes may signify rising market demand, prompting adjustments in pricing or scheduling. Similarly, identifying seasonal travel trends enables airlines to optimize flight schedules and design targeted promotions. This study leverages machine learning techniques to examine passenger behavior, providing airlines with valuable insights to refine strategies and enhance service delivery.

To address these challenges and help travelers make informed decisions, machine learning (ML) techniques have emerged as an effective solution for airfare prediction. By analyzing historical data alongside variables such as flight routes, departure schedules, and seasonal patterns, ML models provide accurate and actionable fare predictions. These models empower travelers to make better booking decisions while enabling airlines to adapt to market trends and optimize their pricing strategies.

This research is centered around three primary features of airfare prediction aimed at providing actionable insights for both travelers and airlines. The first feature focuses on predicting flight fares for various airlines between two cities and analyzing how these fares change over specific intervals of time, offering valuable information about pricing trends. The second feature identifies the best-recommended fare for a chosen route, helping travelers make cost-effective booking decisions. The third feature delivers personalized airline fare predictions, enabling users to find airlines that best match their preferences and budget. In addition to these predictive capabilities, the research also examines passenger traffic and their behavioral responses to airfare trends, further enriching the system's applicability in understanding travel dynamics.

The primary goal of this study is to evaluate the performance of four different machine learning-based systems in predicting airfare, analyzing passenger traffic, and understanding user preferences. Each system is assessed based on key performance metrics to determine the most suitable model for implementing the identified features. The research further explores passenger behavior by leveraging historical data and engineered features to uncover preferences

and patterns. By integrating these insights into the prediction models, the system generates highly accurate, transparent, and user-friendly fare forecasts. These predictions empower travelers to make well-informed booking decisions and provide airlines with the tools to better understand market trends and passenger preferences, thereby enhancing the overall efficiency and transparency of the airline industry.

II. Literature Review

Degife et al. [1] proposed a hybrid model integrating Aspect-Based Sentiment Analysis (ABSA) with Gated Recurrent Units (GRU) to enhance airfare prediction. While their work focuses on leveraging customer sentiment to refine predictions, our research primarily emphasizes historical pricing data and airline-specific features to predict airfare trends. Unlike Degife et al., we do not incorporate qualitative data such as customer reviews, focusing instead on robust feature engineering and algorithmic optimization for quantitative datasets.

Upadhye et al. [2] developed a hybrid model combining K-means clustering with Decision Trees to optimize airfare forecasting. Their approach is particularly effective for handling extensive datasets and localizing predictions. However, our work extends beyond clustering and decision trees by comparing the efficacy of ensemble models like Random Forest and LightGBM, which are better suited for complex, high-dimensional datasets.

Guan [3] introduced an AI-driven model using GANs and LSTM networks for real-time airfare forecasting. Their real-time implementation focuses on dynamic adaptability, whereas our research emphasizes static data preprocessing and feature importance analysis to derive actionable insights for both airlines and consumers. Moreover, our approach ensures interpretable results, which is a challenge with GAN-based systems.

Chavan et al. [4] implemented a Random Forest Regressor model through a Flask web application, emphasizing accessibility. While Random Forest is one of the algorithms we compare, our research incorporates advanced hyperparameter optimization techniques and evaluates multiple ensemble methods to identify the best-performing model under various metrics, providing a more comprehensive evaluation

Degife et al. [5] highlighted GRU's ability to capture nonlinear trends in airfare pricing, outperforming traditional models like MLP and LSTM. In contrast, our research focuses on ensemble methods such as Random Forest and LightGBM, which have shown higher interpretability and scalability for large datasets. Additionally, we integrate a comparative analysis of feature importance to optimize model inputs.

Escañuela Romana et al. [6] analyzed price elasticity in U.S. domestic air travel using a quasi-experimental approach. While their study focuses on policy-oriented insights, our work prioritizes predictive modeling to assist in dynamic pricing and market competition analysis, offering tools for operational decision-making rather than policy implications.

Kalampokas et al. [7] proposed a multi-technique framework incorporating quantum machine learning (QML) for airfare prediction. Although QML achieves high accuracy, our research concentrates on practical, widely applicable machine learning techniques that balance performance with computational efficiency, making our solutions more accessible to industry stakeholders.

Sznajder et al. [8] combined multinomial logit and hedonic regression to analyze consumer preferences, focusing on ancillary fare options. While their work improves customer choice modeling, our research targets fare prediction accuracy by examining historical pricing trends and market competition without ancillary features, catering to both airlines and consumers.

Subramanian et al. [9] conducted a comparative analysis of seven machine learning models, highlighting Random Forest and MLP for their accuracy. Similarly, our work compares multiple models; however, we focus on ensemble methods and their hyperparameter tuning to maximize performance while addressing overfitting issues in large datasets.

Thilak et al. [10] explored Random Forest, Extra Tree Regression, and Linear Regression for price prediction, achieving 81.22% accuracy. Our research advances this by integrating LightGBM and XGBoost, achieving higher accuracy and robustness, especially in handling the temporal and spatial complexities of airfare data.

Fageda et al. [11] examined the implications of antitrust immunity on airline alliances. Unlike their market and policy-oriented study, our research provides operational insights by analyzing airfare dynamics and passenger behavior, supporting tactical decision-making for airlines.

Almansur et al. [12] investigated Delta Air Lines' vertical integration as a strategy for managing fuel cost volatility. While their research focuses on operational assets, our work is centered on optimizing airfare prediction models to enhance pricing strategies and improve customer satisfaction.

Atems, Bachmeier, and Williams [13] evaluated jet fuel prices as predictors of airline fares, finding limited influence on broader demand forecasting. While their study addresses macroeconomic variables, our work examines micro-level features such as distance, carrier type, and market competition, providing a detailed understanding of fare dynamics.

Tziridis et al. [14] evaluated machine learning models like MLP and Decision Trees for airfare prediction, achieving 88% accuracy. Our research builds upon this by comparing multiple ensemble models, which consistently outperform traditional models, and by incorporating feature engineering to enhance predictive power.

Groves et al. [15] introduced a machine learning algorithm for optimizing ticket purchase timing. While their work focuses on consumer decision-making, our research emphasizes predictive modeling for both consumers and airlines, offering insights into dynamic pricing and competitive strategies.

Groves et al. [16] analyzed the economic impacts of unbundling services like baggage fees. Their focus on consumer behavior through price discrimination differs from our data-driven approach, which evaluates historical airfare trends and market competition to optimize airline operations and fare prediction.

Groves et al. [17] analyzed how unbundling services, like baggage fees, impacts airline revenue and consumer behavior, emphasizing price sensitivity and revenue optimization. In contrast, our work focuses on predictive modeling of airfare trends and passenger behavior to provide actionable insights for pricing strategies.

Ref.No	Methodology	RMSE	MAE	R^2
[1]	XGBoost	0.872	0.894	0.939
[4]	Random Forest	0.914	0.921	0.933
[5]	LightGBM	0.872	0.894	0.939
[7]	Neural Network	0.851	0.842	0.842

Table 2: Comparison from Literature Review

Tabular comparison can be referred using Table 2 and Table 3 from the results section

III. PRELIMINARIES

This section provides a detailed description of the dataset used in this study, including its features and structure. Additionally, it outlines the performance evaluation metrics employed to assess and compare the accuracy of various machine learning models in predicting airfare trends and passenger behavior.

A. Dataset Description

The dataset used in our study has been downloaded from data.gov website, titled, "**Consumer Airfare Report: Table 1a - All U.S. Airport Pair Markets**" 1. The dataset consists of 2,45,956 observations with 19 features. More details are described in Table 1

The Consumer Airfare Report dataset is a thorough collection of airfare information, capturing intricate details regarding flight routes, pricing, and passenger counts. Covering multiple years, it consists of 245,956 records, with each entry reflecting specific market data between different city pairs within the airline network. This dataset features attributes such as the year and quarter of data collection, unique identifiers for cities and airports, distances between locations, average fare prices, and comprehensive details about the airlines servicing those routes.

Each record in the dataset provides vital insights into the airline industry, including:

- Identifiers for Cities and Airports: Unique codes for both the origin and destination cities, as well as their respective airports, enabling detailed analysis of specific routes.
- Passenger Numbers: The count of travelers on each route, serving as a key measure of demand and market potential.
- Fare Information: Average prices charged by both major and low-cost carriers, allowing for a comparative analysis of their pricing strategies.
- Market Share Data: Insights into the market shares of the largest and lowest-cost carriers on each route, facilitating an evaluation of competitive dynamics within the airline sector.

The dataset is highly significant across various fields. For airfare analysis and forecasting, it offers a valuable resource for researchers, airlines, and policymakers to examine pricing patterns, understand consumer behavior, and pinpoint critical factors that influence fare changes. By leveraging this data, stakeholders can create accurate predictive models that guide pricing strategies and capacity management.

B. Performance Evalution Parameters

1) ***R^2 Score****:* The formula for calculating the R2 score, also known as the coefficient of determination, is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

Where y_i represents the observed value of the target variable, $\hat{y}_i$ is the predicted value of the target variable, and $\bar{y}$ denotes the mean of the observed values of the target variable. Additionally, n represents the number of samples in the dataset. In simpler terms, R^2 is calculated as the proportion of variance in the dependent variable that can be predicted using the independent variables. It effectively measures the goodness of fit of the model, with higher values indicating a better fit to the data.

2) ***Mean Absolute Error:*** The Mean Absolute Error calculates the average difference between the calculated values and actual values. It is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

where n represents the number of samples, y_i represents the actual value for the sample, and $\hat{y}_i$ represents the predicted value for the sample.

3) ***Root Mean Squared Error****: The Root Mean Squared Error calculates the square root of the average of the square of the difference between the calculated values and actual values. It is given by:*

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

where n represents the number of samples, y_i represents the actual value for the sample, and $\hat{y}_i$ represents the predicted value for the sample.

4) ***Time:*** The time taken from running each and every model for each features

1 https://catalog.data.gov/dataset/consumer-airfare-report-table-1a-all-u-s-airport-pair-markets

Table 1: Dataset Description

Feature	Data Type	Description
Year	int	The year during which the data was recorded.
Quarter	int	The quarter of the year (e.g., Q1, Q2).
Citymarketid_1	int	Unique identifier for the market of city 1 (origin).
Citymarketid_2	int	Unique identifier for the market of city 2 (destination).
City1	object	Name of the origin city.
City2	object	Name of the destination city.
Airportid_1	int	Unique identifier for the airport corresponding to city 1.
Airportid_2	int	Unique identifier for the airport corresponding to city 2.
Airport_1	object	Name or code of the origin airport.
Airport_2	object	Name or code of the destination airport.
NSmiles	int	Distance between the two cities in miles.
Passengers	int	Number of passengers traveling on the route.
Fare	float	Average fare for the route.
Carrier_LG	object	Code of the largest carrier on the route.
Large_MS	float	Market share of the largest carrier on the route.
Fare_LG	float	Average fare charged by the largest carrier.
Carrier_Low	object	Code of the lowest-cost carrier on the route.
LF_MS	float	Market share of the lowest-cost carrier.
Fare_Low	float	Fare charged by the lowest-cost carrier.

Table. 1: Dataset Description

IV. PROPOSED METHODOLOGY

This section delineates the procedural sequence adopted for airfare prediction. The proposed methodology encompasses the following stages: 1.Features, 2. Data preprocessing, 3. Data Splitting, 4. Metaheuristic Optimization and 5. Model Implementation. These steps in the proposed methodology are illustrated in Fig. 1. Results will be discussed in the next section.

A. Features

*1) **Interval***

This feature focuses on how airline fares fluctuate over time. By analyzing fare changes at various intervals, such as daily, weekly, or monthly, the model can capture patterns related to seasonality, demand surges, and special events. For instance, airline prices often spike during holiday seasons, weekends, or peak travel times. This feature helps travelers identify the best times to book tickets for the lowest prices and provides airlines with data to refine dynamic pricing strategies.

*2) **Best_Fare***

This feature identifies the most economical fare for a given route by comparing pricing data from different airlines. It considers various parameters, such as travel dates, destinations, departure times, and available discounts, to provide users with recommendations for the cheapest flights. For customers, this feature simplifies decision-making and ensures cost-effective travel. For airlines, it offers insights into competitors' pricing strategies and supports competitive fare structuring.

*3) **Date_Carrier_Fare***

This feature integrates critical elements of fare prediction:

- **Date:** Captures temporal factors like travel season, day of the week, or time of year, which influence ticket prices.
- **Carrier:** Considers the airline operator, as fares can vary significantly depending on the carrier's reputation, service quality, and market position.
- **Fare:** Represents the target variable, which the machine learning model predicts based on historical data and derived patterns. Together, these components allow the model to provide nuanced predictions and enable users to make well-informed decisions when selecting flights.

*4) **Passenger_Traffic***

Passenger traffic refers to the volume of travelers on a specific route during a particular time. This feature helps identify high-demand routes, peak travel times, and underutilized connections.

By analyzing passenger traffic, the model can understand how demand influences ticket pricing. For airlines, this feature is valuable for capacity planning, optimizing flight schedules, and implementing targeted marketing campaigns. For travelers, it helps anticipate fare fluctuations due to demand-driven pricing.

*5) **Passenger_Behaviour***

Passenger behavior is a crucial factor in understanding airfare pricing and market trends. It includes aspects such as price sensitivity, booking patterns, and traveler preferences. While some passengers prioritize affordability, others value convenience, flexibility, or premium services.

External factors like economic conditions, travel restrictions, and global events also influence behavior, impacting demand and pricing strategies. Route-specific preferences highlight distinct patterns among business and leisure travelers.

Additionally, feedback and sentiment analysis offer insights into customer satisfaction and perceptions of pricing fairness. Understanding these trends helps airlines optimize pricing models, enhance customer satisfaction, and develop strategies tailored to diverse consumer needs.

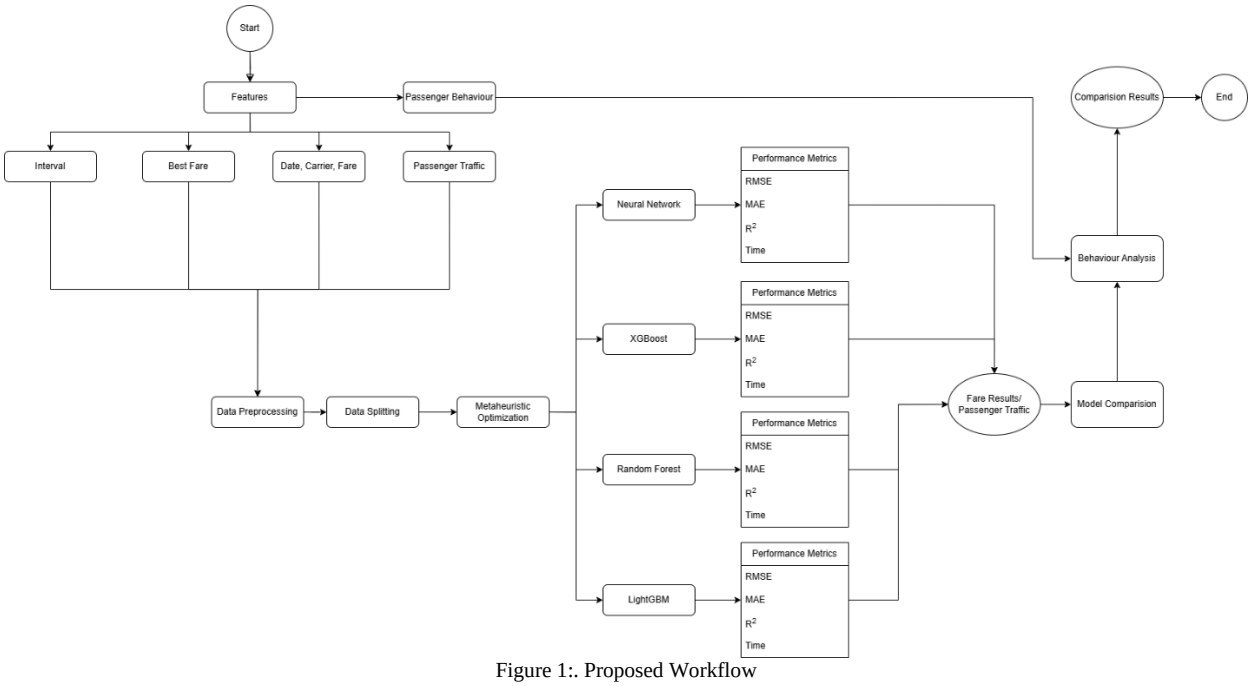

Figure 1:. Proposed Workflow

B. Data Preprocessing

This involves cleaning and organizing the dataset by removing irrelevant columns and incomplete rows categorical variables like airline carriers and cities are transformed into numerical formats using encoding methods like one-hot or label encoding numeric features such as fares and distances are normalized or scaled to enhance model performance while outliers are detected and addressed to prevent biased predictions these procedures guarantee that the data is tidy reliable and suitable for precise airfare prediction modeling

C. Data Splitting

it is crucial to assess a models performance and ability to ge neralize in predicting airfare by dividing the dataset into training and testing subsets typically in an 8020 or 7030 ratio the training set helps the model learn patterns between input features eg cities airlines dates and the target variable airfare while the testing set evaluates its predictive accuracy on new data cross-validation can also be employed to ensure consistent performance and prevent overfitting with attention given to maintaining a representative distribution of key variables in both subsets for effective handling of various pricing scenarios

D. Metaheuristic Optimization

This is crucial in adjusting hyperparameters of machine learning models for airfare prediction enhancing both accuracy and efficiency methods such as genetic algorithms particle swarm optimization and bayesian optimization cleverly search the hyperparameter space to find the best configurations for parameters like learning rate estimators or tree depth these techniques are particularly beneficial for managing extensive intricate datasets affected by seasonal and demand fluctuations through the optimization of hyperparameters models can enhance their accuracy scalability and robustness ultimately leading to dependable predictions and effective performance in practical scenarios

E. Model Implementation

1) Neural Network: Neural Networks are a type of machine learning model inspired by the structure of the human brain. They consist of layers of interconnected neurons that process input data, learn patterns, and make predictions. In airfare prediction, Neural Networks excel at capturing complex relationships between features, such as the interaction between travel dates, routes, and airlines. The model's ability to handle nonlinear relationships and large datasets makes it suitable for identifying subtle trends and patterns in fare fluctuations. By employing multiple hidden layers and activation functions, Neural Networks can adapt to intricate data structures, making them a powerful tool for accurate airfare predictions.

$$z = \sum_{i=1}^{n} w_i x_i + b$$

$$a = \sigma(z)$$

2) XGBoost: XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable gradient-boosted decision tree algorithm. It is renowned for its speed, performance, and ability to handle large datasets with high dimensionality. XGBoost uses an ensemble approach, building a series of decision trees sequentially, with each tree correcting the errors of the previous one. In airfare prediction, XGBoost is particularly effective in handling categorical features like airlines and routes, as well as temporal data like travel dates. Its regularization techniques help prevent overfitting,

ensuring robust predictions. Additionally, XGBoost's optimization for parallel processing makes it a preferred choice for fast and accurate predictions in large-scale airfare datas

$$L(\theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

3) Random Forest: Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. Each tree in the forest is trained on a random subset of data and features, and the final prediction is obtained by aggregating the outputs of all trees. In the context of airfare prediction, Random Forest is highly effective due to its ability to handle diverse types of data, such as categorical variables like airlines and continuous variables like fare amounts. It provides reliable predictions even when dealing with noisy or missing data. Its interpretability, through feature importance scores, also helps identify key factors influencing airfare, making it a valuable tool for analysis.

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x)$$

4) LightGBM: LightGBM (Light Gradient Boosting Machine) is a gradient-boosting framework designed for speed and efficiency, especially with large datasets. It uses a histogram-based approach to split data, significantly reducing computation time and memory usage. LightGBM is well-suited for airfare prediction as it can handle categorical data directly, identify complex patterns, and scale effectively for high-dimensional datasets. Its ability to prioritize relevant features through gradient-based one-sided sampling ensures that it focuses on critical factors affecting airfare. With its fast training and robust predictive capabilities, LightGBM is an excellent choice for building scalable airfare prediction models.

$$L = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \lambda \sum_{j=1}^{m} w_j^2$$

V. Result

The graphs provided in Figure 2 display comparisons of four performance metrics (RMSE, MAE, R^2, and Time) across four models (Neural Network, XGBoost, Random Forest, and LightGBM) for different tasks (Interval, BestFare, DateCarrier, Passenger).

The Root Mean Squared Error (RMSE) measures the average deviation of the predictions from the actual values. Neural Network consistently exhibits the highest RMSE, indicating poor prediction accuracy compared to the other models. XGBoost performs better than Neural Network, with moderate RMSE values across tasks. However, Random Forest achieves the lowest RMSE in all tasks, showing exceptional accuracy. LightGBM performs slightly worse than Random Forest but still outperforms Neural Network.

Random Forest has relatively high training times, second only to Neural Network, likely due to its computationally intensive tree-building process. LightGBM strikes a balance, offering low training times comparable to XGBoost while maintaining strong predictive performance. This makes XGBoost and LightGBM ideal for time-sensitive applications. It emerges as the most accurate model, consistently achieving the lowest RMSE and MAE values and near-perfect R^2 scores. It is best suited for tasks where accuracy is paramount.

XGBoost and LightGBM are computationally efficient and deliver strong predictive performance, making them excellent choices for time-sensitive or resource-constrained scenarios. Neural Network, however, lags behind in both accuracy and efficiency, making it less favorable unless further tuning or a specific use case justifies its application.

1) Why prefer LightGBM, not Random Forest for Dynamic price comparision

Random Forest, while powerful and accurate, has notable drawbacks that impact its practicality in certain scenarios. One major limitation is its overreliance on dominant features, as observed in the feature importance analysis where the model heavily depends on the dynamic_fare variable as given in Figure 3.

This over-dependence poses a risk if this feature is missing, biased, or inaccurate, which can significantly degrade the model's predictions. Additionally, Random Forest's computational cost is relatively high due to the ensemble nature of constructing and evaluating numerous decision trees, making it less efficient for larger datasets or time-sensitive applications.

Another drawback is its lack of interpretability, as the predictions are derived from a collection of decision trees, making it harder to explain the reasoning behind certain outputs. This becomes a critical limitation in industries like finance or healthcare, where transparency is essential. Furthermore, Random Forest sometimes ignores potentially valuable features, as it tends to prioritize dominant predictors.

For example, features like fare_per_mile and remaining_seats are almost ignored in this case, reducing the model's flexibility in situations where the primary feature (e.g., dynamic_fare) might be unreliable or unavailable. Lastly, while Random Forest reduces overfitting compared to single decision trees, it is still prone to overfitting when the number of trees is high or when there are imbalanced feature contributions.

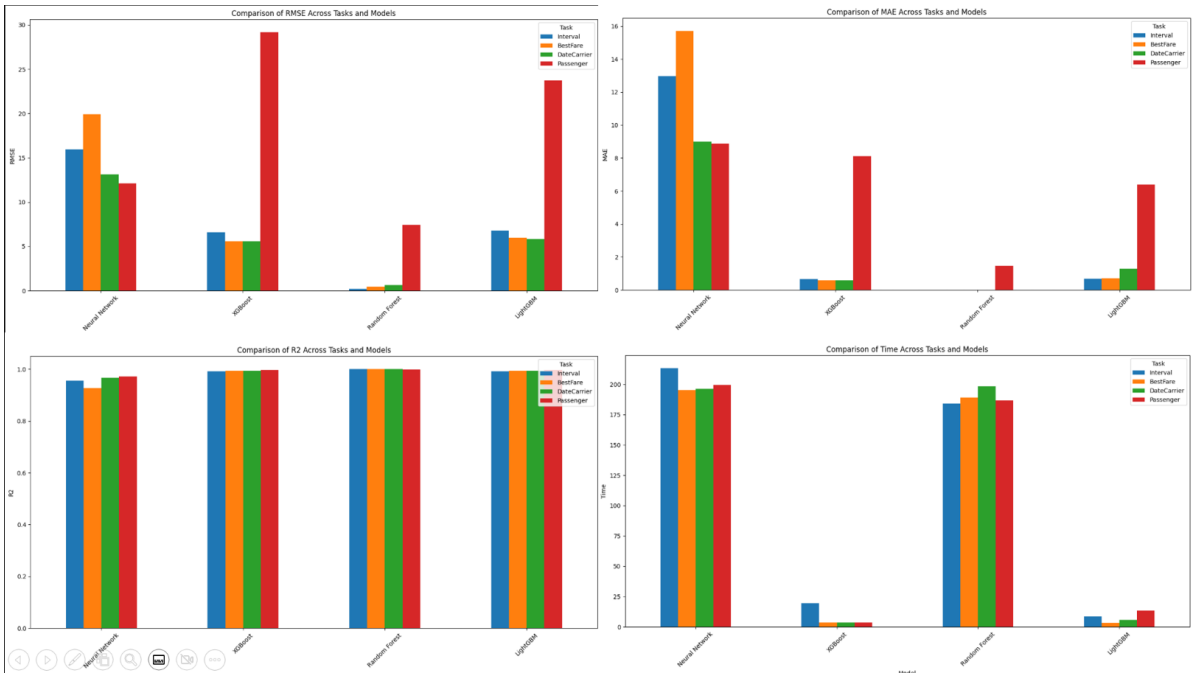

Figure2: Comparison of RMSE, MAE, R2 and Runtime

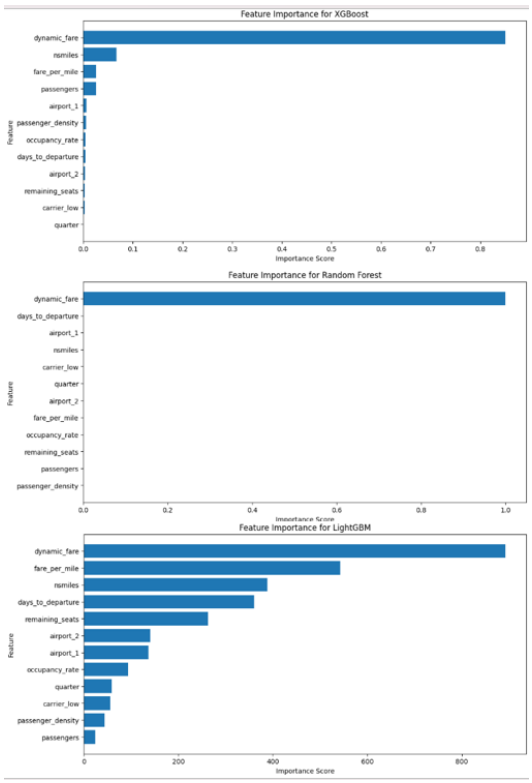

Figure 3: Feature Importance for Interval, Best fare and Date Carrier Fare

In contrast, LightGBM offers a more balanced approach by effectively utilizing a wider range of features and avoiding over-reliance on any single one. It is computationally efficient, scalable for large datasets, and quicker for both training and prediction, making it more suitable for real-world applications. Considering these factors, LightGBM emerges as the better model for this task, as it strikes a balance between accuracy, efficiency, and robustness, particularly in dynamic and data-intensive scenarios. While Random Forest achieves excellent performance in this dataset, its limitations make it less adaptable compared to LightGBM for broader, practical use cases.

2) XGBoost for Passenger Traffic

The passenger traffic analysis of feature importance across the models—XGBoost, Random Forest, and LightGBM—highlights that certain features like nsmiles (distance), fare_per_mile (cost efficiency), and days_to_departure (timing of booking) play a dominant role in predicting passenger traffic. XGBoost emerges as the best-performing model for this task, primarily due to its balanced handling of these key features. For instance, XGBoost assigns the highest importance to nsmiles, recognizing the critical role distance plays in determining passenger behavior. Additionally, it gives significant weight to fare_per_mile and passenger_density, underscoring its ability to capture both cost efficiency and traffic flow patterns on specific routes as shown in Figure 4.

In comparison, Random Forest, while also emphasizing nsmiles and fare_per_mile, tends to underweight passenger_density, which is vital for understanding traffic variations across different routes. LightGBM, on the other hand, places higher importance on days_to_departure, which is essential but might overshadow other critical metrics like passenger_density. These differences demonstrate that XGBoost's feature importance distribution is more aligned with real-world determinants of passenger traffic, offering a nuanced understanding of how fares, distance, and timing influence demand.

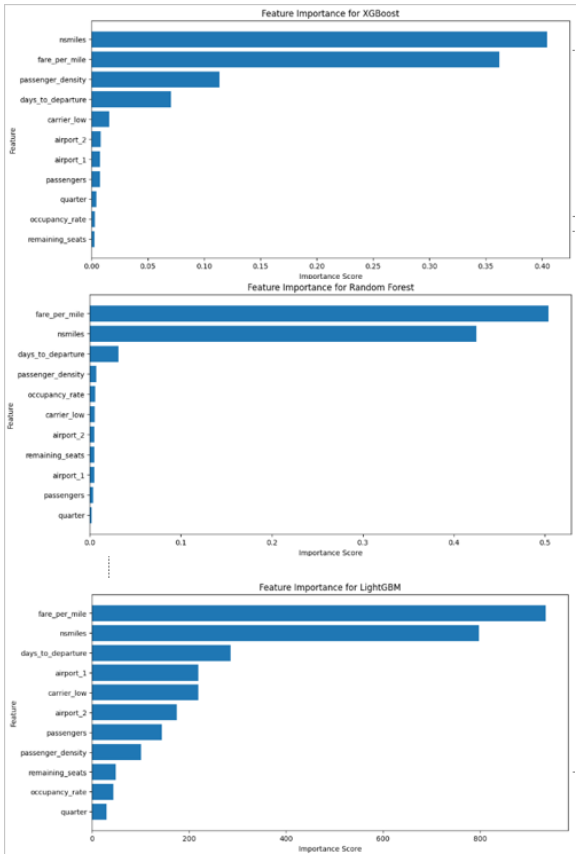

Figure 4: Feature Importance for Passenger Behavior

Moreover, XGBoost's ability to model complex, nonlinear relationships between features gives it a significant advantage in accurately predicting passenger traffic. It maintains superior performance metrics, including high R2R^2R2 scores and competitive RMSE and MAE values, reflecting its precision and reliability. Additionally, XGBoost strikes a balance between computational efficiency and accuracy, requiring relatively less training time compared to Random Forest while delivering comparable, if not better, predictive power.

In conclusion, XGBoost is the best model for passenger traffic prediction due to its superior handling of key features, ability to model nonlinear relationships effectively, and balanced computational efficiency. This makes it highly suitable for providing accurate and actionable insights into passenger behavior, ensuring robust performance in real-world scenarios.

Feature	Model	RMSE	MAE	RÂ²	Time (s)	Best Model
	Neural Network	15.9522	12.9706	0.9562	213.04	
	XGBoost	6.6145	0.6542	0.9925	19.69	
	Random Forest	0.2043	0.0046	0.99999	184.03	
Interval	LightGBM	6.7914	0.6703	0.9921	8.46	Best
	Neural Network	19.9373	15.6921	0.9269	194.99	
	XGBoost	5.5708	0.5774	0.9943	3.68	
	Random Forest	0.4566	0.0086	0.99996	189.16	
Best Fare	LightGBM	5.9471	0.7148	0.9935	3.33	Best
	Neural Network	13.1459	9.0061	0.9682	196.33	
	XGBoost	5.5708	0.5774	0.9943	3.67	
	Random Forest	0.6334	0.01	0.99993	198.38	
Date Carrier Fare	LightGBM	5.8181	1.2718	0.9938	5.63	Best
	Neural Network	12.1264	8.8794	0.9729	199.23	
	XGBoost	29.1514	8.1138	0.9967	3.65	Best
	Random Forest	7.4328	1.474	0.99979	186.69	
Passenger Traffic	LightGBM	23.7294	6.3825	0.9978	13.5	

Table 3: Predicting the best model based on comparison

3) Passenger Behavior Analysis

The passenger behavior analysis provides insightful trends that can guide strategies as given in Figure 5. Over half of the passengers, around 52.96%, exhibit price sensitivity, indicating that competitive pricing and discounts are essential to cater to this majority. Conversely, 47.87% of passengers are less price-sensitive, suggesting opportunities to offer premium services or additional amenities for an upsell. Regarding booking behavior, a significant 83.16% of passengers prefer to book early, which emphasizes the importance of early-bird discounts and promotions to secure bookings well in advance. Medium-term and last-minute bookings account for 9.09% and 7.75%, respectively, underscoring the need for flexible pricing strategies to accommodate urgent travelers as shown in Table 4.

When it comes to route preferences, medium-haul routes are the most popular, attracting 51.88% of passengers, followed by long-haul routes at 30.16% and short-haul routes at 18.45%. This suggests airlines should focus on optimizing their regional services while balancing resources for long- and short-haul demands. In terms of holiday and peak-period preferences, a significant majority of 73.70% prefer off-peak travel, highlighting the need for off-peak pricing incentives to attract this segment. Meanwhile, 26.29% of passengers prioritize peak-period travel, creating opportunities to introduce premium pricing and enhanced services for this group.

In conclusion, these insights underscore the importance of adopting a segmented strategy. Airlines can optimize revenue by tailoring dynamic pricing for price-sensitive and premium customers, promoting early-bird discounts, prioritizing medium-haul services, and designing targeted offers for both off-peak and peak travelers. These strategies not only align with passenger preferences but also improve customer satisfaction and operational efficiency.

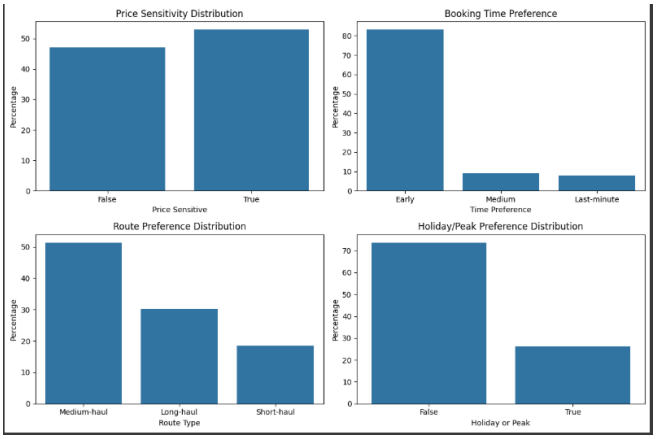

Figure 5: Passenger Behavior Analysis

Category	Sub-category	Proportion
Price Sensitivity	TRUE	52.9621
	FALSE	47.0379
Booking Time Preference	Early	83.160493
	Medium	9.053996
	Last-minute	7.785511
Route Preference	Medium-haul	51.38096
	Long-haul	30.16388
	Short-haul	18.45516
Holiday/Peak Preference	FALSE	73.700046
	TRUE	26.299954

Table 4: Passenger Behavior Analysis Table

VI. Conclusion

The analysis of predictive models and passenger behavior reveals critical insights into optimizing airline operations and pricing strategies. Random Forest stands out as the most accurate model with the lowest RMSE and MAE values and near-perfect R^2, making it ideal for tasks where accuracy is paramount. LightGBM balances accuracy and computational efficiency, making it a strong choice for dynamic pricing tasks by utilizing a diverse range of features and avoiding over-reliance on a single one. XGBoost excels in passenger traffic prediction due to its ability to handle nonlinear relationships, prioritize key features like distance and timing, and deliver competitive performance with low computational costs. Insights from passenger behavior analysis suggest the importance of early-bird discounts, dynamic pricing for last-minute travelers, and strategic optimization of medium-haul routes, along with targeted offers for off-peak and peak travelers. These strategies align with passenger preferences, improving customer satisfaction and operational efficiency.

VII. Future Work

Future efforts can be focused on optimizing models through advanced tuning, exploring hybrid approaches combining Random Forest, LightGBM, and XGBoost, and validating them with real-time datasets for robustness. Expanding feature engineering to include variables like seasonal trends and demographics, and using clustering for passenger segmentation, can refine strategies. Tools like SHAP and LIME should be used to enhance model interpretability. Prioritizing scalability and deployment will ensure seamless integration into real-time systems, enabling dynamic pricing and route optimization. These advancements will help airlines boost revenue, streamline operations, and deliver tailored customer experiences.

VIII. Referrences

[1] Degife, D. and Lin, Z., ” Predicting Airfare Prices Using GRU and Aspect-Based Sentiment Analysis,” IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 2, pp. 243-253,(2024). https://doi.org/10.3390/app14104221

[2] Upadhye, A., Lakhani, R., Pendam, N., Bari, A., Deulkar, A., ” Kmeans clustering and decision tree model for airfare predictions,” Journal of Machine Learning Applications, vol. 45, no. 6, pp. 123-130, (2024). ISSN: 2073-607X,2076-0930.

[3] Guan, L., ” A generative AI model for real-time airfare forecasting using GAN and LSTM,” Journal of Web Engineering, vol. 23, no. 2, pp. 299-314, (2024). https://doi.org/10.13052/jwe1540-9589.2325

[4] Chavan, R., Makhar, S., Kulkarni, P., Kulkarni, R., ” Flight Price Prediction for Enhanced Recommendations via Machine Learning Web Application, ” in Proceedings of the 2nd International Conference on Sustainable Computing and Smart Systems, IEEE Xplore, pp. 1036-1037, (2024). https://doi.org/10.1109/ICFIRTP56122.2022.10059429

[5] Degife, D., Lin, Z., ” GRU model for nonlinear airfare price prediction IEEE Transactions on Neural Networks and Learning Systems, vol. 34, (2023). https://doi.org/10.3390/app13106032

[6] Esca˜nuela Romana, M., Torres-Jim´enez, J., Carbonero-Ruz, M., ” Airfare demand elasticity estimation in U.S. domestic air travel,” Journal of Transport Economics, vol. 20, pp. 67-77, (2023). https://doi.org/10.1007/s11293-023-09779-4

[7] Kalampokas, T., Tziridis, K., Vrochidou, E., Nikolaou, A., Kalampokas, N., Papakostas, G. A., ” A Holistic Approach on Airfare Price Prediction Using Machine Learning Techniques,” IEEE Access, vol. 11, pp. 3274669, (2023). 1109/ACCESS.2023.3274669

[8] Sznajder, B., Ratliff, R., Kaya, A., ” A heuristic for incorporating ancillaries into air choice models with personalization (Part 2: integrated multinomial logit and hedonic regression models),” Journal of Revenue and Pricing Management, vol. 22, pp. 140‑151, (2023). https://doi.org/10.1057/s41272-022-00400-y

[9] Sznajder, B., Ratliff, R., Kaya, A., ” A heuristic for incorporating ancillaries into air choice models with personalization (Part 1:estimating preferences using hedonic regression), ” Journal of Revenue and Pricing Management, vol. 22, pp. 122‑139, (2023). https://doi.org/10.1057/s41272-022-00399-2

[10] Subramanian, S., Deepak, V., Murali, K., ” Airline Fare Prediction Using Machine Learning Algorithms,” in Proceedings of the Fourth International Conference on Smart Systems and Inventive Technology (ICSSIT-2022),https://doi.org/10.1109/ICSSIT53264.2022.9716563

[11] Thilak, S.J., Chittate, A.R., Benny, B.P., Khan, T.A., Paulose, E., Kouatly, R., ” A Comparison Between Machine Learning Models for Airticket Price Prediction,” 2022 3rd International Informatics and Software Engineering Conference (IISEC), IEEE, (2022) https://doi.org/10.1109/IISEC56263.2022.9998230

[12] Fageda, X., Flores-Fillol, R., Lin, Y., ” Vertical differentiation and airline alliances: The effect of antitrust immunity,” Regional Science and Urban Economics, vol. 81, (2020). https://doi.org/10.1016/j.regsciurbeco.2020.103517

[13] Almansur, K., Megginson, W., Pugachev, A., ” Vertical integration as an input price hedge: The case of Delta Air Lines and trainer refinery, ” Financial Management, vol. 49, pp. 179-206, (2020). https://doi.org/10.1111/fima.12260

[14] Atems, B., Bachmeier, L., Williams, A., ” Jet Fuel Prices and Airline Fares: An Autoregressive Analysis of U.S. Air Travel Demand,” Applied Economics Letters, vol. 26, no. 11, pp. 877‑882, (2019).

[15] Tziridis, I., Kalampokas, G., Papakostas, G., ” Airfare Prices Prediction Using Machine Learning Techniques,” in Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO), (2017). https://doi.org/10.23919/EUSIPCO.2017.8081365

[16] Groves, T., Gini, C., ” Product Unbundling in the Travel Industry: The Economics of Airline Bag Fees, ” Journal of Economy Management Strategy, vol. 24, no. 3, pp. 457-484, (2015). https://doi.org/10.1016/j.jconsbeh.2015.02.001

[17] Groves, T., Gini, C., ” On Optimizing Airline Ticket Purchase Timing, ” in Proceedings of the ACM 2015, (2015). http://dx.doi.org/10.1145/2733384

IX. Acknowledgement of contributions

The authors equally contributed to this project and wish to outline their respective roles in its development:

Advaithbarath Raghuraman Bhuvaneswari (ar2728):
Contributed to the "Objective of the Project," "Problem Statement," and "Dataset Description" in Phase I. Led data preprocessing and exploratory data analysis (EDA) during Phase II, including visualizations such as "Passenger Volume Over the Years." In Phase III, worked on documenting insights, refining the "Conclusions" section, and formatting the final report.

Srivatsan Jayaraman (sj796):
Participated in writing the "Introduction" and structuring the project framework during Phase I. In Phase II, worked on feature engineering and implemented machine learning models like Random Forest and LightGBM. Contributed to model evaluation and "Future Work" in Phase III, ensuring consistency across the report.

Aravind Kalyan Sivakumar (as4588):
Assisted with the "Objective of the Project" and "Dataset Description" in Phase I. Conducted literature review and implemented statistical analyses and visualizations, including "Fare vs. Distance" and "Market Share of Largest Carriers," in Phase II. During Phase III, contributed to creating the performance comparison table and integrating sections into a cohesive final document.

Each author’s contribution was critical to the successful completion of this project, and all authors acknowledge their equal efforts in achieving the outlined objectives.