

RESEARCH ARTICLE



Interpretable Airfare Prediction Using Random Forests: Seasonal Variation, Statistical Testing, and a Structured Prediction Workflow

Swathi Ganesan^{1,*}, Amjad Alam¹, Sangita Pokhrel¹, Ganapathy Kumar², Shiv Kunwar³, Minel Somasiri⁴ and Nalinda Somasiri¹

¹Department of Computer and Data Science, York St John University, UK

²Independent Researcher, UK

³York St John University, UK

⁴Rickmansworth School, UK

Abstract: Airfare prediction remains a significant challenge in transportation economics, as fares are influenced by a combination of temporal, operational, and market-related factors, resulting in complex pricing dynamics. This study proposes a more interpretable data-driven framework that ties together statistical checking with ensemble machine learning to investigate fare variations. This study uses the US Airline Flight Routes and Fares (1993–2024) dataset from the Bureau of Transportation Statistics, followed by preprocessing, feature engineering, and inferential testing to examine the most influential factors for airfare dynamics. First, seasonal variation was evaluated using formal statistical tests to determine whether the quarterly differences in average airfare were statistically significant, thereby justifying the inclusion in the predictive model. The outcomes from this study show noticeable seasonal effects, with a one-way analysis of variance finding statistically significant variation across quarters ($F = 46.19$, $p < 0.001$). In addition, a structured prediction workflow has been built, so the system can accept user-specified inputs and return scenario-based airfare forecasts using the trained model. Overall, the framework offers a clear, usable approach for understanding seasonal shifts and operational patterns in airfare pricing.

Keywords: airfare forecasting, random forest regression, seasonality analysis, inferential validation, machine learning framework

1. Introduction

In this modern era of machine learning (ML)-driven decision-making, how good the data is and how carefully feature engineering is handled before the models are trained determine how dependable and interpretable the results are. Raw datasets are often not that consistent as they have missing entries; moreover, they have outliers that can quietly impact the learning signals, which in turn hurts predictive accuracy. This has led scientists to highlight that preprocessing of data and feature selection is the foundation of successful predictive modeling and has helped models become accurate and explainable [1]. Statistical and probabilistic methods used in preprocessing help machine learning systems address uncertainty, noise, and high-dimensional data structures, that are common challenges in real-world prediction tasks such as transportation and pricing analytics. The use of

feature engineering both enhances the performance of a model and makes the model interpretable. Houkan et al. [2] showed that the minimum redundancy maximum relevance (MRMR) performed better than principal component analysis (PCA) in terms of feature selection to classify industrial Internet of Things (IoT) intrusion cases as informative variables. Their findings emphasize the power of feature selection to maximize accuracy and reduce computational costs, which is useful in general to the field of ML. On the same note, Felahatpisheh et al. [3] proposed a hierarchical two-level genetic algorithm to implement a feature selection approach that was automated in a manner that would trade interpretability and predictive power in a multi-objective way. These innovations confirm the fact that feature refinement must have a systematic approach in order to create scalable, transparent, and computationally efficient learning models. Although ML has been more popular in predictive modeling, most of the available literature ignores the fact that prior to model training, seasonal and structural differences need to be verified by formal statistical

*Corresponding author: Swathi Ganesan, Department of Computer and Data Science, York St John University, UK. Email: s.ganesan@yorks.ac.uk

testing. The gap in this research is that by incorporating modeling data with the method of inferential analysis, the temporal variation is not only statistically validated but also computationally represented.

The increasing sophistication of modern data has also led to interest in explainable artificial intelligence (XAI). The studies by Khan et al. [4] and O'Brien Quinn et al. [5] state that by incorporating interpretability mechanisms into ML systems, stakeholders can bridge classical statistical inference and the modern predictive system. The literature highlights the impact of hybrid approaches like interpretable decision trees and post hoc explanation methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). This seems to boost transparency and show predictive ability. In our current work, we try to get interpretability together with statistical checking via analysis of variance (ANOVA) and Tukey's Honestly Significant Difference (HSD) with random forest feature importance insights. This has been done in order to preserve the airfare variation to be grounded in numbers rather than just intuition. At the same time, data-driven methods have been gaining increasing attention in transportation forecasting. So, Sreelekha and Midhunchakkaravarthy [6] showed how ML could be used in handling city traffic, and Yang and Chen [7] used empirical mode decomposition (EMD) together with ML to improve the fact, that is, how accurately air freight volume can be predicted.

Simultaneously, Guan [8] presented a generative AI architecture for real-time airfare forecasting and emphasized how AI can change the state of dynamic and uncertain pricing. All these studies demonstrate the fact that the aviation and transportation industry offers a good platform to experiment with the cutting-edge ML methodologies that unite both statistical understanding and computational scalability. For analysts and planners, it is important to understand that they should not just rely on numerical indicators but should note that interpretability is now a key requirement of aviation analytics for predictive performance.

Probabilistic modeling has been a stronger method to deal with uncertainty and data anomalies. Hence, Li et al. [9] have proposed a Bayesian outlier detection framework to identify anomalous observations in structural health monitoring without compromising data integrity.

Similarly, Barsce et al. [10] introduced a reinforcement-learning algorithm with Bayesian optimization as an automated method of hyperparameter optimization, focusing on the efficiency of demand forecasting models. Together, these studies show a growing integration of Bayesian statistical methods with ML techniques to boost predictive robustness; however, uncertainty is still present. A similar trend is also visible within the healthcare domain, in which Talaat [11] made use of deep learning, more specifically convolutional neural networks, to estimate cardiovascular risk using patient-based health indicators. The above research in References [10, 11] yielded near-perfect R-squared values. An example covered in References [11, 12] is closer to domain-specific; however, it does highlight the extensive adaptability of ML methodologies across many other areas like healthcare and IoT security systems. Abdella et al. [13] conducted a broad review within the domain of airfare prediction covering both airline-centric and consumer-focused predictive models. Their results indicate that existing methods are restricted by a limited feature space, usually leaning on old pricing records and trip dates. On the same side, they ignore more complex context and exogenous factors. This makes the case that we really need better modeling frameworks, ones that are more encompassing, more interpretable, and more resilient for airfare prediction. Notably,

there is still a clear research gap where studies don't really bring in seasonality inference together with interpretable ML, all inside one unified modeling framework, especially for domestic airfare forecasting.

To address this gap, the present study proposes a structured framework that blends an interpretable random forest model with rigorous statistical validation, seasonal decomposition, and scenario-based predictive analysis. The methodology includes data preprocessing, including data cleansing and outlier detection, statistical techniques such as hypothesis testing (ANOVA and Tukey's HSD), and the Random Forest Regressor for predictive modeling. The goal is to build a dependable, interpretable, and generalizable inference setup that can capture both seasonal and operational dynamics shaping airline pricing behavior. This in turn helps to take a transparent, data-driven, and modern approach to airfare prediction. This combined approach contributes not only to the transportation analytics literature but also to a broader set of studies along with rigor, clarity, and transparency. It also helped in data preparation and probabilistic logic, which substantially improves the integrity disclosure and overall performance of ML systems. The airfare prediction has recently been studied using both econometric and ML approaches, but few have given a comprehensive treatment including seasonal inference, competitive dynamics, and model transparency. This encourages a more combined assessment of how things have been done. On the whole, the research provides a clear and statistically justified approach to modeling the seasonal behavior of airfares, supported by a structured prediction framework.

2. Literature Review

Airfare modeling has developed in three broad directions: econometric methods, ML methods, and hybrid probabilistic models. Airfare prediction has increasingly become a critical research area due to fluctuations in demand, seasonality, economic variables, and market competition. Conventional statistical methods often struggle to effectively model the nonlinear and dynamic behavior of airfare trends, encouraging researchers to adopt more advanced ML-based approaches. Early comparative studies show that traditional statistical techniques can fall short, especially when dealing with high-dimensional inputs or with data that are irregularly arranged. For example, Liu et al. [14] showed that ML models perform better than conventional statistical methods when linked to forecasting air cargo volumes. This is because these models may pick up intricate, nonlinear interactions, suggesting that airfare prediction might benefit from similar approaches.

More recent research has been trying to boost predictive performance using time-series decomposition strategies. For example, a study by Gürler and Özçalıcı [15] introduced a hybrid framework that blends EMD with ML algorithms. The results from this research show a noticeably improved accuracy by breaking the signal into intrinsic mode functions. This method is especially relevant for airfare prediction because time-related volatility occurs a lot, along with irregular fluctuations. This seems to appear without warning.

Hybrid modeling is tuned specifically for airfare prediction as it shows a lot of promise. The study in References [16] did a comparative evaluation of hybrid ML models for flight fare prediction, which concluded that combining different algorithms boosts predictive accuracy instead of leaning on just one estimator. This aligns with forecasting research more broadly, where ensemble and multilevel methods are now used more frequently,

mainly because pricing behavior is inherently complex. On the other side, pricing elasticity still plays a major role in airfare prediction. The demand for air travel responds strongly to price changes in accordance with research on elasticity-driven fare optimization in US airline markets [17]. This also supports the inclusion of passenger behavior and economic sensitivity in the predictive framework, if improved reliability is the desired outcome.

Recent developments in deep learning have further advanced prediction abilities with the use of transportation datasets. A study introduced in Reference [18] related to an uncertainty-aware temporal graph convolutional network. This captures spatiotemporal dependencies in a useful way. Hence, this fundamental concept of modeling dynamic temporal relationships can be applied to the fluctuating patterns observed throughout airline fares in practice, even though the work has been implemented on traffic speed predictions.

Improvements in predictive performance are closely connected with the optimization of model choice and hyperparameters, despite the inherent complexity of the process. A study by Wang et al. [19] proposed a framework that integrates reinforcement learning with Bayesian optimization to automate the algorithm selection process. The results of this study showed better performance outcomes. These kinds of automated optimization methods are especially relevant in airfare prediction, where multiple different models must be continuously fine-tuned to follow shifting market conditions.

Feature engineering and data preprocessing remain essential for developing robust predictive models. A comprehensive review of feature selection methods across various domains highlights that the identification of relevant features significantly improves both model interpretability and predictive performance [20]. Complementing this, Bala and Behal [1] highlighted the importance of preprocessing steps such as data cleaning, normalization, and dimensionality reduction, all of which are critical for managing complex, heterogeneous airfare datasets.

The reported work was complemented by another study by Houkan et al. [2], who showed that MRMR is more effective at finding the necessary features to detect an anomaly than PCA, with a maximum of twelve features revealing up to 99.9% accuracy. The results of their study demonstrate that the elimination of the strategic features can not only improve the performance but also reduce the complexity of computation, an idea that is directly applicable to airline fare modeling, in which multidimensional data are typical.

Furthermore, modern studies are becoming more inclined toward interpretable and probabilistic modeling. The studies by Khan et al. [4] and O'Brien Quinn et al. [5] claimed that interpretability should co-exist with performance and suggested frameworks that incorporate classical statistics with XAI solutions, including SHAP and LIME.

Barsce et al. [10] applied Bayesian optimization to reinforcement learning in the forecasting field and reduced the computational expenses while preserving the accuracy of the selected model and hyperparameter optimization. In a similar way, the article by Li et al. [9] used Bayesian outlier detection to enhance regression modeling in high-speed railway monitoring, which showed the viability of probabilistic reasoning to ensure the integrity of data. These advances mark a transition to hybrid statistical ML models with predictive accuracy, interpretability, and robustness. In the field of aviation forecasting and prediction, more recent literature has brought new frameworks that build upon deep learning and data decomposition techniques.

Guan [8] used generative AI to develop a real-time web-based prediction system for airfare that focuses on flexibility to market changes by integrating data continuously. The research by Yang and Chen [7] suggested mixing EMD with ML algorithms in order to predict air freight, and in practice, it boosts accuracy because the method can decompose temporal subcomponents hidden inside the time-series stream. These kinds of applications demonstrate that AI-driven models are being used to cope with the nonlinear behavior, the uncertainty, and also with the time-related complexity that often appears in aviation data.

Even with these advances, there are still a bunch of research gaps left. A lot of the earlier literature leans on fixed prediction models that really don't catch seasonality properly as well as quarterly changes. It also tends to underplay feature explainability which is in practice, a vital part to build usable airline pricing model. Furthermore, many studies focus on scenario-based predictive ability, for example, when estimating fares under changing temporal or carrier conditions. Here, the present study tries to cover the gap by blending careful statistical testing, like ANOVA and Tukey's HSD, with ML using a Random Forest Regressor.

The overall framework improves interpretability, better captures seasonal variations, and also provides a data-driven approach in predicting airfare. This work supports and contributes more to transportation economics. This framework uses predictive analytics to be transparent and dependable. Moreover, it is easier to understand through the combination of probabilistic reasoning with ensemble learning.

Though there have been many developments and advancements in the area of airfare forecasting and transportation analytics, most of the existing studies have focused more on improving predictive accuracy through ML or hybrid modeling approaches. For example, studies in References [14, 15] demonstrate the effectiveness of ML and decomposition-based techniques in forecasting transportation-related outcomes. On the same side, studies in Reference [16] highlight the benefits of using hybrid models in airfare prediction. Similarly, deep learning-based approaches [18] and optimization frameworks [19] have been proposed to enhance model performance and adaptability. Many of these studies end up giving limited attention to formal statistical validation of the underlying temporal patterns, especially when it comes to seasonal variation. The interpretability part is usually not the main aim, since most existing work leans on predictive accuracy rather than a clear explanation of how the model behaves [4, 5]. Additionally, here scenario-based prediction refers to the ability to evaluate hypothetical future conditions, which is relatively underexplored in the literature. On the other hand, the present study tries to tackle these issues by integrating statistical validation, that is, ANOVA and Tukey's HSD, with an interpretable random forest and feature importance analysis. This combined approach enables both rigorous validation of seasonal effects and transparent identification of key drivers of airfare variation. Additionally, the inclusion of a structured prediction workflow supports scenario-based airfare estimation, enhancing the practical applicability of the framework for decision-making in airline pricing and policy analysis.

3. Methodology

This study follows a systematic methodological framework starting with data preparation, statistical validation, model building, interpretability, and then scenario-based prediction. The dataset, US Airline Flight Routes and Fares (1993–2024), from the Bureau of Transportation Statistics (BTS) contains numbers for

domestic routes. The dataset contains information related to passenger counts, average ticket prices, carrier category, and market share. In our study, “Fare” is the dependent variable used in both the predictive and inferential analyses. Here, “fare” refers to the average airfare for each route-quarter entry in the BTS dataset. As a result, the random forest model in our study will help to predicts airfare by considering temporal, operational, and carrier-related factors, whereas one-way ANOVA assesses on the fact whether the target value changes significantly over quarters. The independent variables include route distance, number of passengers, load factor (lf_ms), carrier characteristics (carrier_lg, carrier_low), and temporal variables (year, quarter). This methodological structure follows a sequential framework in which each stage builds upon the previous one, ensuring robustness in predictive modeling and interpretability [20].

3.1. Data cleaning and preprocessing

The dataset used for this research did contain unusual records, noise, and missing values, which could affect the model performance; therefore, data preprocessing has been done. To handle missing values, approach suggested in the paper by Jaiswal et al. [21] has been carried out, that is, ML-based imputation method, and it did produce reliable results than conventional statistical techniques.

Using a distribution-based approach, outlier detection was also carried out, which was further verified through regression residual analysis. This step was important for this research as these outliers will affect the model accuracy and reliability. The importance of this step is supported by Li et al. [9], who state that regression reliability is improved by identifying abnormal observations with the use of Bayesian outlier detection. Hence,

anomalous fare records, mainly caused by promotional pricing or data entry inconsistencies, were filtered out prior to model training.

To ensure that categorical features of airline data are compatible with numerical variables, normalization and encoding techniques have been implemented. As recommended by Bomrah et al. [22], feature transformations and scaling need to be used to minimize bias. It is also necessary to stabilize model performance across diverse data inputs. This aligns with Alvaro et al.’s review of data preparation, which highlights data cleaning, transformation, integration, and feature preparation as important stages for improving the quality and usability of data before ML modeling [23].

Figure 1 illustrates the relationships between key predictors in the dataset, supporting the selection of relevant variables for modeling and confirming the presence of meaningful associations prior to feature engineering.

Figure 2 shows the distribution of average airfare (USD), highlighting the right-skewed shape and presence of extreme values, which informed the outlier removal strategy using the interquartile range method.

3.2. Feature engineering and encoding

Interpretability and model performance improvement came from feature engineering. The load factor (lf_ms) and the large carrier market share (large_ms) remained unchanged because they still reflect key operational plus market-related traits. The temporal variables, like quarter and year, also remained unchanged. This approach was used to address both seasonal fluctuations and longer-term price patterns. Categorical features, such as carriers, airports, and city pairs, were one-hot encoded using the

Figure 1
Relationship between features

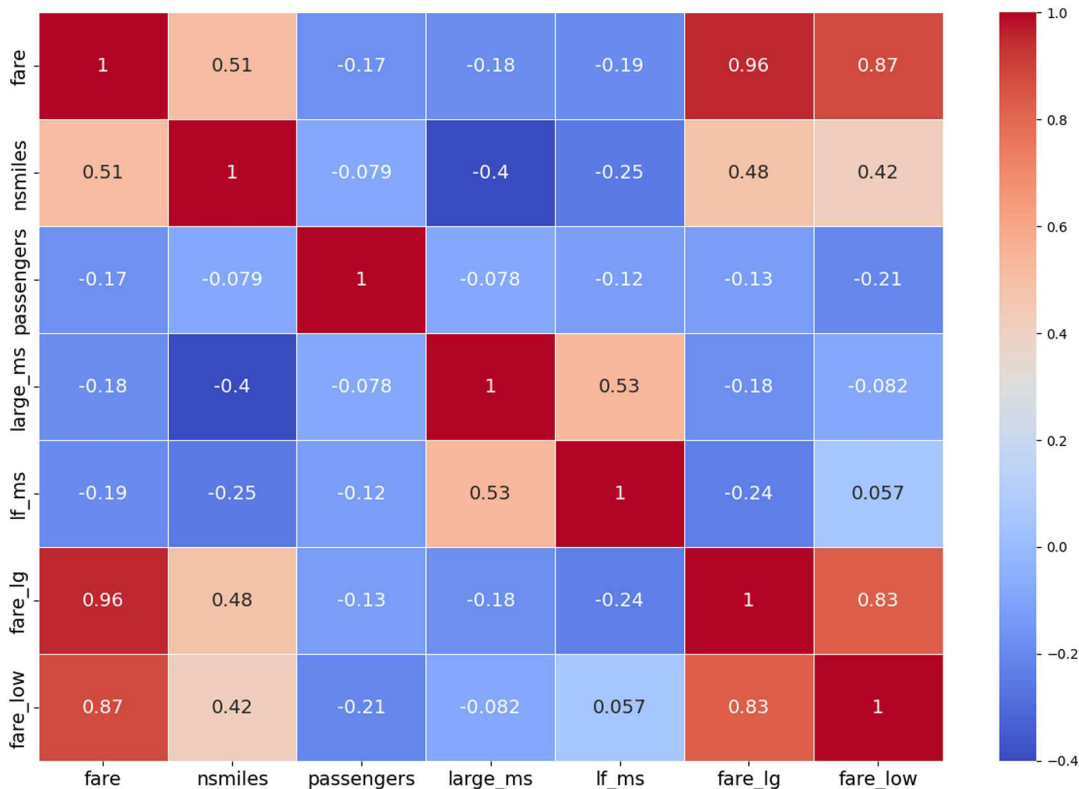
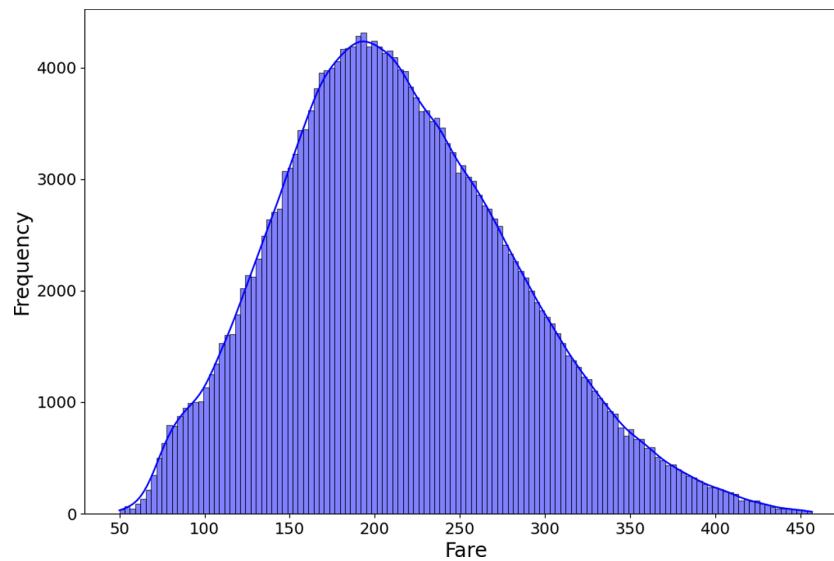


Figure 2
Distribution of average airfare



drop_first=True setting. The aim here is to reduce multicollinearity in a more controlled way. A few recent systematic reviews note that strong feature selection tends to make interpretability better and usually boosts downstream results for cross-disciplinary work [24]. Feature scaling and normalization were not required because random forest is a tree-based ensemble method. After cleaning and encoding, the dataset went through the train-test split step, that is, a training subset (80%) and a test subset (20%). This facilitated the assessment of model generalization. For a fair evaluation, performance was assessed using standard regression metrics, that is, mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R^2). Applying these evaluation criteria ensures comparability with baseline approaches and enhances the trustworthiness of the performance assessment.

3.3. Model development using random forest

Random forest has been chosen for this research because of its decision tree-based architecture and robustness. Another reason for choosing it is that it has the capacity to capture a nonlinear interaction in a natural way. The aim was to improve predictive stability and to reduce variance in the model outcomes with the use of this model.

To find a balance between computational time and accuracy, key parameters including minimum samples per leaf, maximum depth, and number of estimators have been used as optimization factors. To maintain consistency across the whole framework, this study has used a random forest model to predict the same target variable as in the statistical analysis, that is, average airfare (USD) across all stages.

Random forest has been selected in this study due to its strong performance and robustness in handling mixed data types, nonlinear relationships, and feature interactions that are often present in transportation datasets.

Compared with a traditional linear regression model, random forest does improve generalization, and this is because it can capture complex patterns without the need for explicit feature transformations. It also reduces overfitting through ensemble averaging. Overall, it captures complex patterns without needing explicit feature transformations. Also, in comparison to single

decision trees, the random forest model does perform well as it reduces overfitting through ensemble averaging, improving generalization. On the other side, some alternative models such as neural network and gradient boosting good achieve higher predictive performance, but it requires extensive hyperparameter tuning. This leads to the reduction of interpretability. Hence, the use of random forest has helped in balancing predictive accuracy with interpretability and statistical validation. This makes random forest a suitable and effective modeling choice.

Standard regression measures such as MAE, MSE, RMSE, and R^2 have been used to assess the model performance in this paper. The above selected error metrics were employed to have a thorough evaluation of predicted dependability. From the results, the most significant determinant of the airfare dataset was found to be load factor, route distance, and quarter, in line with feature importance analysis.

3.4. Statistical validation of seasonal effects

ANOVA, along with Tukey's HSD tests, was used to check the seasonal variation. Before building the predictive model, an ANOVA was performed on the same target variable, namely, average airfare (USD), to see if fare levels are different enough across the quarters. This step was carried out to ensure that the statistical validation step aligns with subsequent predictive modeling. By confirming these seasonal fare changes were statistically meaningful before training the random forest model, seasonal patterns were identified in measurable effects rather than random noise. Hence, there is a clear methodological use of inferential analysis that identifies the features to include along with the most effective ML model.

3.5. Conceptual automated prediction workflow

A structured prediction workflow has been created in this paper to help in estimating scenario-based airfare and to enhance practical applicability. In the current implementation, predefined input values such as route, quarter, year, carrier details, passenger volume, and route distance are supplied by the user. These data are then transformed into a dataframe and encoded to fit the training features schema, followed by training with a random

forest model in order to produce a forecast. Our conceptual architecture for the automated prediction assistant has been presented in Figure 3. However, a structured manual input pipeline has been used, which is in line with the trained random forest model to provide predictions. Therefore, the workflow should be understood as a conceptual interface for structured prediction and not a fully automated system.

3.6. Scenario-based prediction framework

A scenario-based prediction component was put together to estimate future airfare values, under user-set temporal, route, and carrier configurations. The model produced scenario-based predictions of fares by providing hypothetical values of quarter, year, carrier characteristics, and operational characteristics. In other words, this simulation helped in comparing projected fare outcomes across seasons, different routes, and carrier categories, thereby yielding useful results for airlines, travel platforms, and policymakers. These scenario run-throughs are basically an example of applying the model for both operational calls and strategic choices. Advanced spatiotemporal optimization studies in related transportation systems further highlight the importance of modeling dynamic, time-coupled dependencies in predictive frameworks [25].

To ensure analytical reliability, transparency, and consistency, the methodology has adopted procedures, such as data cleaning, statistical validation, model building, and interpretability evaluation. The resulting framework offers a scalable, interpretable, and statistical reasoning along with ML-based approach in predicting seasonal and operational dynamics of airline fare behavior.

3.7. Experimental environment and reproducibility

To support reproducibility, the experimental workflow was put together in a Python-based Jupyter Notebook environment. The whole analysis ran using open-source libraries, such as Pandas and NumPy for handling data, SciPy and Statsmodels for statistical

testing, Scikit-learn for model development and evaluation, and Matplotlib plus Seaborn for the visualization part.

The predictive model itself was built using Scikit-learn’s Random Forest Regressor. While training, it was set up with 100 decision trees (`n_estimators = 100`) and used `random_state = 42` to keep things reproducible. For speed and parallel execution, `n_jobs` has been set as `-1` in order to make the computations faster. A bunch of other random forest hyperparameters were just left at the default Scikit-learn Random Forest Regressor settings, for example, `criterion = squared_error`, `max_depth = None`, `min_samples_split = 2`, `min_samples_leaf = 1`, `max_features = 1.0`, and `bootstrap = True`. With these defaults, the ensemble develops into fully grown trees, and the aggregation across many estimators helps reduce prediction variance. This smooths the output by combining multiple individual predictions.

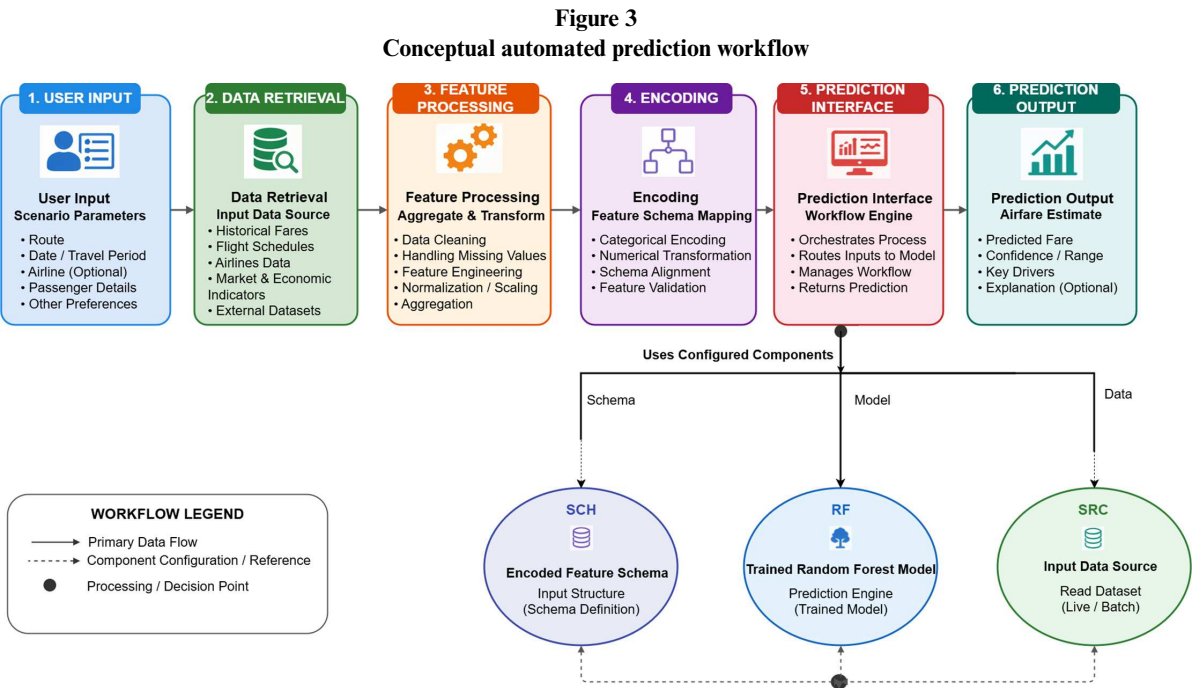
All experiments were carried out inside the same notebook-based setup, so the preprocessing, statistical analysis, model training, evaluation, visualization, and the scenario-based prediction stayed consistent. The full implementation was also released so that others can rerun the experimental pipeline reproducibly. This will help in future validation and also with the expansion of our study in an easier manner.

4. Result

On our held-out test set, the use of random forest regression showed strong predictive performance, demonstrating its applicability for modeling the intricate relationship governing the dynamics of airline fare. Standard regression metrics were used to assess performance and were interpreted in relation to baseline methods, that is, linear regression and conventional statistical models. With an overall MAE of 18.81, MSE of 786.40, RMSE of 28.04, and an R^2 of 0.838, the model successfully explained the variation in average airfare (USD) of roughly 83.8%.

4.1. Long-term trends and fare structure

To evaluate how pricing behavior shifts over time and to validate the temporal variables used in the predictive model, this



section examines long-term airfare trends and structural patterns found in the dataset. Understanding these temporal dynamics will help confirm the applicability of temporal variables utilized in the predictive model and in comprehending how fare behavior changes over time. Highlighting these primary airfare trends provides an insight into underlying market changes.

Figure 4 shows the long-term pricing trend from 1993 to 2024. It can be observed that there is a significant drop in the early 2000s and then a steady increase after 2010. Hence, the results illustrate that the year variable has turned out to be one of the best predictors in the feature importance ranking.

4.2. Influence of distance and market structure

Alongside seasonal trends, the relationship between operational variables and airfare provides a comprehensive view of airline pricing behavior. This section examines the influence of

route distance and market structure on fare variation. It also provides evidence of nonlinear relationships to justify the deployment of ensemble-based ML models.

The hexbin plot in Figure 5 shows a strong positive, nonlinear relationship between distance (nsmiles) and fare. Dense clusters in the 500–1200 mile range indicate where most domestic traffic occurs. The nonlinearity supports the choice of a tree-based model capable of capturing such complex relationships.

Figure 6 illustrates the changing market share of large and low-cost carriers over time but also reflects the evolving competitive structure of the airline market. It can be observed that pricing tactics become more dynamic with increased competition. The rise in low-cost carrier presence after the year 2016 has contributed to larger fare dispersion. The pattern observed in Figure 6 indicates that market structures should be taken into account in addition to operational factors while analyzing airfare behavior.

Figure 4
Average fare by year

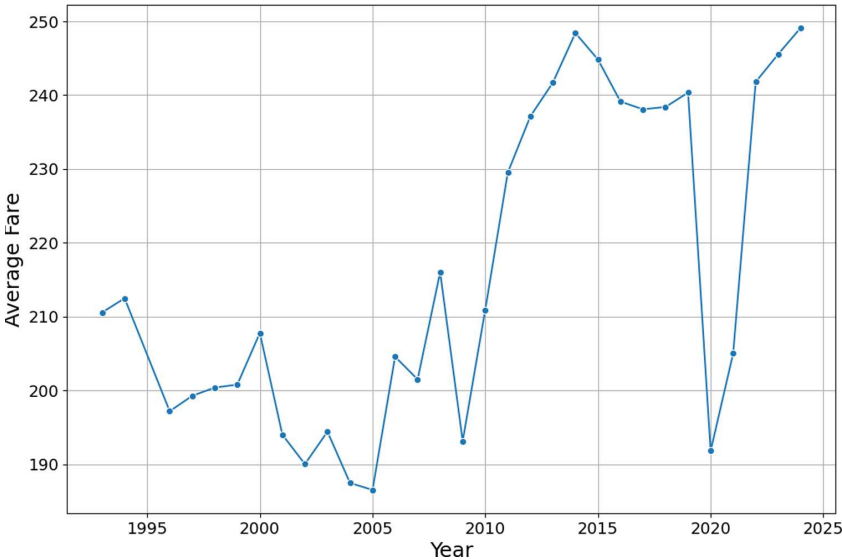


Figure 5
Hexbin plot of airfare vs. flight distance

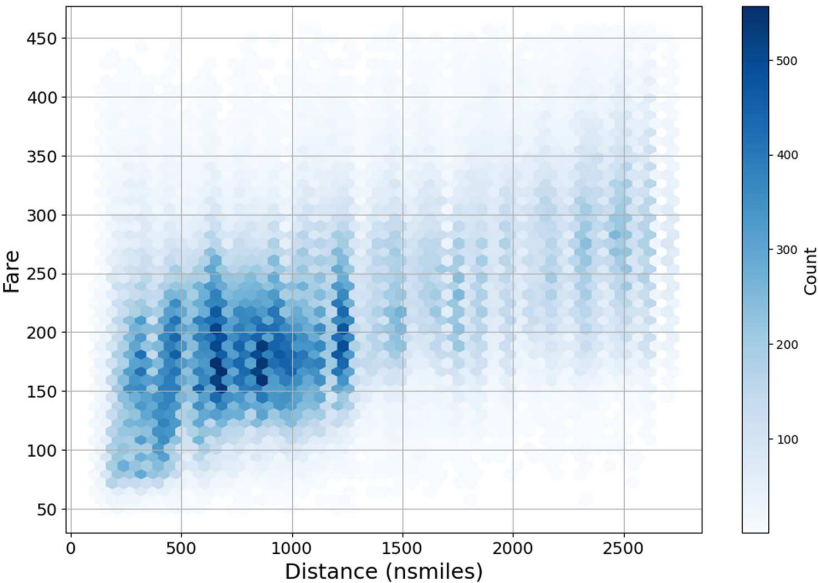
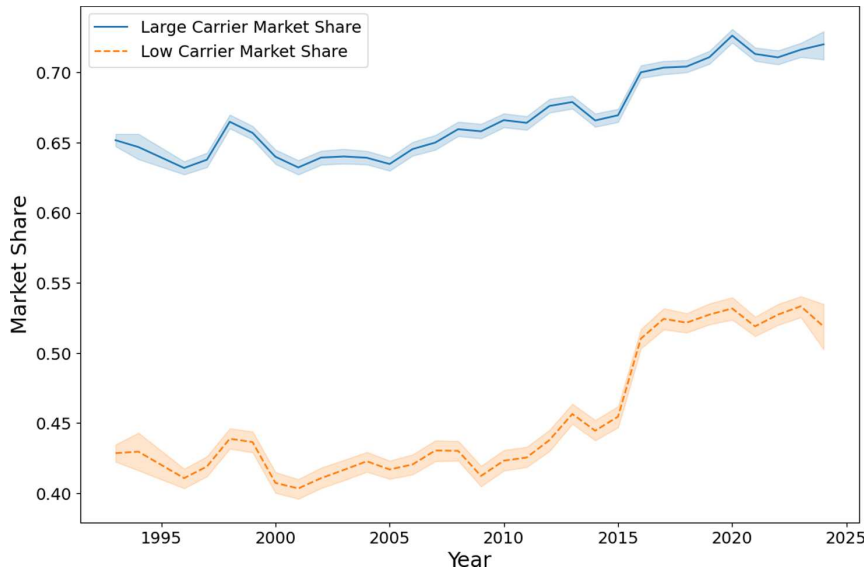


Figure 6
Market share of large vs. low-cost carriers over time



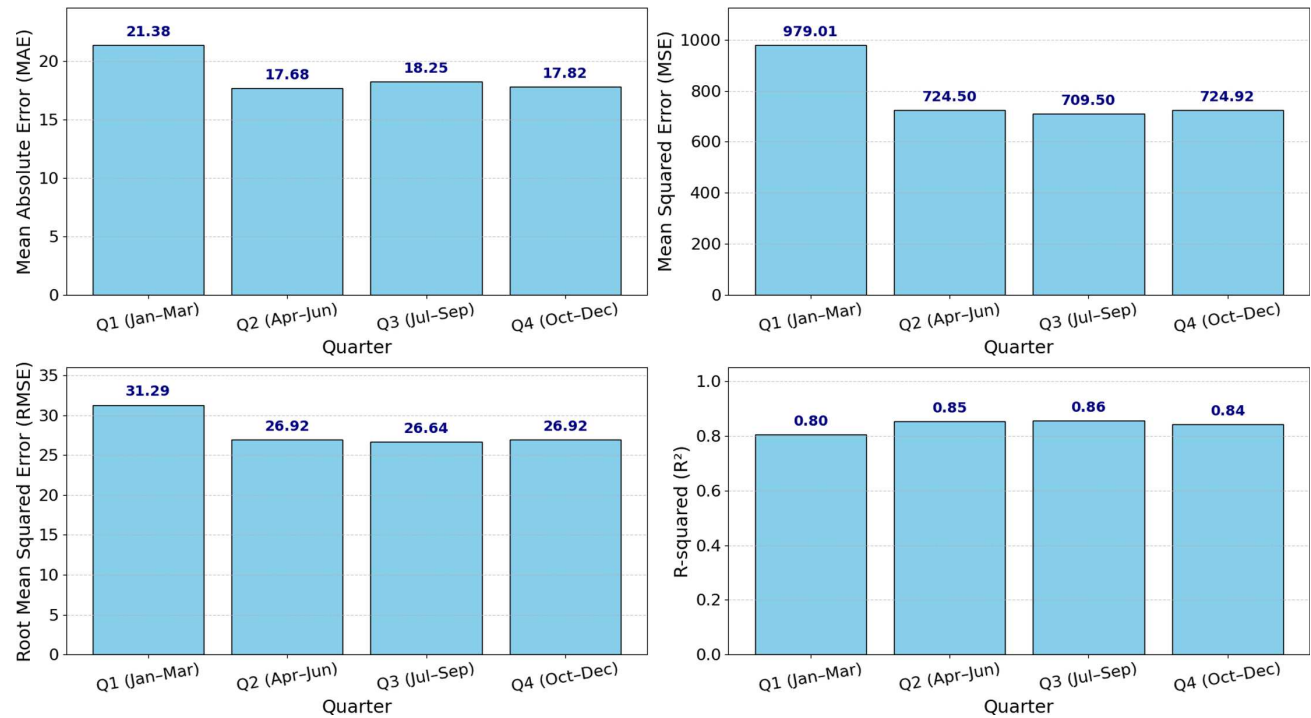
4.3. Seasonal performance and error analysis

From the analysis, it has also been observed that there is variation between quarters. The first quarter (Q1) shows slightly higher error values with MAE = 21.38, RMSE = 31.29, and MSE = 979.01. The remaining quarters (Q2–Q4) displayed a consistent and consistent performance with MAE ranging between 17.7 and 18.3 and RMSE between 26.6 and 26.9. Likewise, the R^2 values ranged from approximately 0.80 in Q1 and were between 0.84 and 0.86 in Q2–Q4. So, the findings suggest that the first

quarter volatility is in line with the industry trend. Hence, the unstable pricing pattern has resulted from post-holiday demand slowdowns, fare promotions, and uneven capacity adjustments.

In addition to the increased error values observed in Q1, the quarterly trend in Figure 7 shows that imperial demand instability does affect model performance. In comparison to Q1, Q4 has been observed to be more sensitive to sudden market fluctuations. This has been caused by year-end toggle behavior and post-holiday demands shift. Comparable prediction errors were observed in Q2 and Q3, indicating similar predictive performance

Figure 7
Model performance metrics by quarter



during these quarters. This suggests the fact that fares during this period followed more predictable operational and seasonal pattern. The significance of seasonality as a practical and statistical factor influencing airfare variation is reinforced by this multi-perspective assessment.

Table 1 summarizes quarterly performance metrics, which indicate that Q1 produces consistently higher error values, while Q2 and Q3 demonstrate greater predictive stability. Quarter 4 maintains moderate error levels as it is likely to be influenced by year-end travel behavior.

Table 1
Quarterly performance table

Quarter	MAE	RMSE	R^2
Q1	21.38	31.29	0.8
Q2	17.68	26.91	0.85
Q3	18.25	26.64	0.86
Q4	17.81	26.92	0.84

4.4. Statistical validation of seasonal differences

In order to statistically prove the existence of seasonal differences, a one-way ANOVA was run to test the average airfare (USD) across quarters. The test showed $F = 46.19$, $p = 7.89 \times 10^{-30}$ ($p < 0.001$), confirming that the differences in mean airfare across quarters were statistically significant. A post hoc Tukey's HSD test was then applied to determine pairwise differences, finding that Q1 was significantly different compared to Q3 and Q4 (reject = True), Q2 was significantly different compared to Q4 (reject=True), and no significant difference was found between Q3 and Q4 (reject = False). Tukey's HSD indicated that there is also a significant difference between Q2 and Q3 (reject = True), demonstrating a shift in the levels of fares in the middle of the year and not in the effects of Q1. These findings support the quarter variable as an important explanatory variable in the predictive model, which can explain the asymmetry observed in the seasonal error

behavior and yield better interpretability in terms of variations in the fare over time.

4.5. Model diagnostics

These findings are also supported by model diagnostics. The actual versus predicted scatter plot showed that many points fell along the 45-degree reference line, indicating good correspondence between the predicted and observed values, as shown in Figure 8. This expansion at higher fares will be more variable at long distances or high-demand routes, which is characteristic of airline pricing. There were some slight underestimates in the high fare range, which is because of the conservative bias seen in the ensemble models due to the averaging of predictions.

4.6. Residual diagnostics

Two additional diagnostic tests were conducted to evaluate model robustness and also to confirm the accuracy in terms of prediction related to the use of random forest. The two diagnostic checks conducted are (i) residual distribution analysis and (ii) residuals-versus-predicted scatter evaluation. The plot illustrates that the model does not suffer from major statistical weakness and generalizes well.

It has also been observed from Figure 9 that it resulted in a bell shape, which means there is an absence of symmetric over- or under-prediction. The mean residual value (≈ -0.63) is effectively zero, demonstrating that prediction errors are unbiased across the dataset. The relatively narrow spread of the residuals ($SD \approx 28.0$) reflects concentrated error boundaries. This shows that it is consistent with the overall RMSE as reported earlier. This distribution shows that the model does not consistently depart from actual fare values and retains steady error behavior.

Figure 10 presents the residual-versus-predicted plot. It shows a dense symmetrical cloud of points around the zero-residual axis. It shows no visible curvature or directional pattern. Hence, it can be said that the model does not exhibit structural bias across low, mid, or high fare ranges. It also shows that there is little expansion of residual variance caused by higher

Figure 8
Actual versus predicted fare

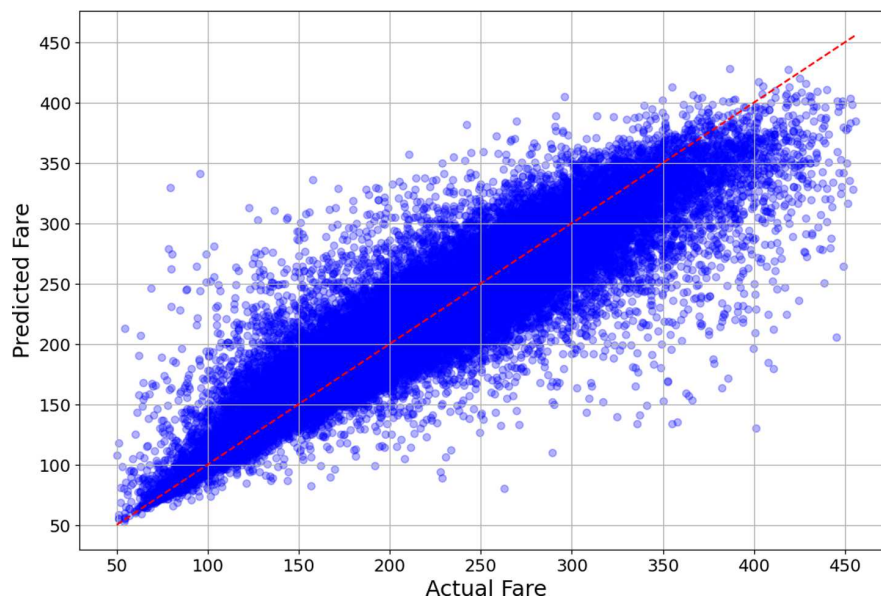


Figure 9
Residual distribution curve

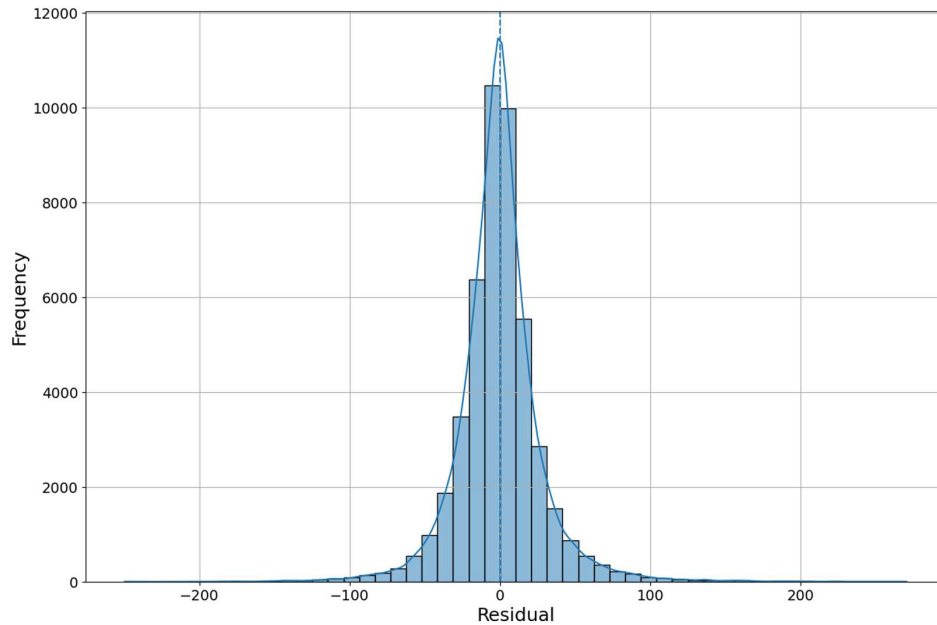
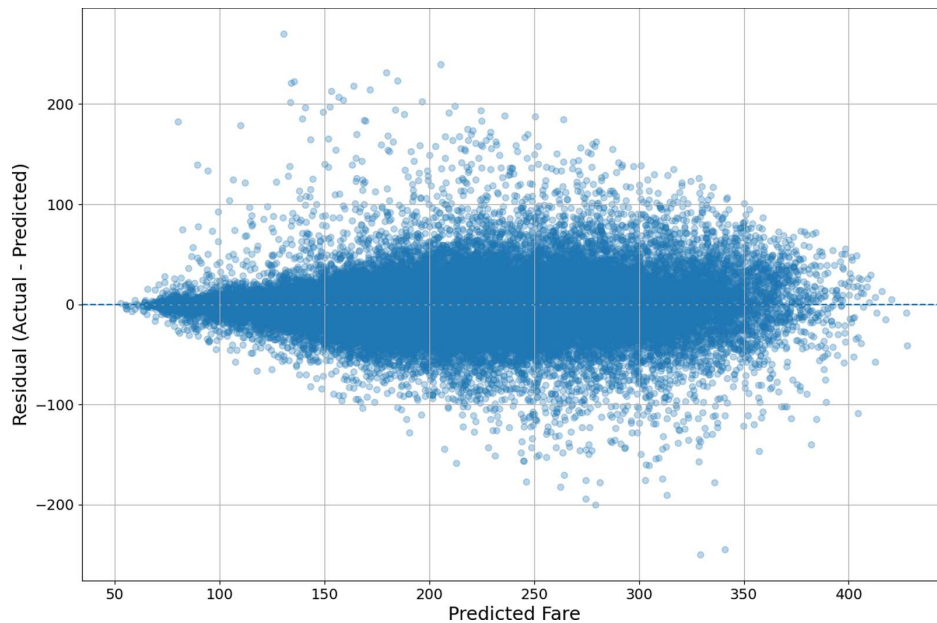


Figure 10
Residuals vs. predicted scatter plot



predicted fares. This indicates that the model is more stable. This demonstrates that the implemented model's predicted errors are well-behaved and constant across fare levels, as supported by the balanced vertical dispersion and the lack of significant heteroscedasticity.

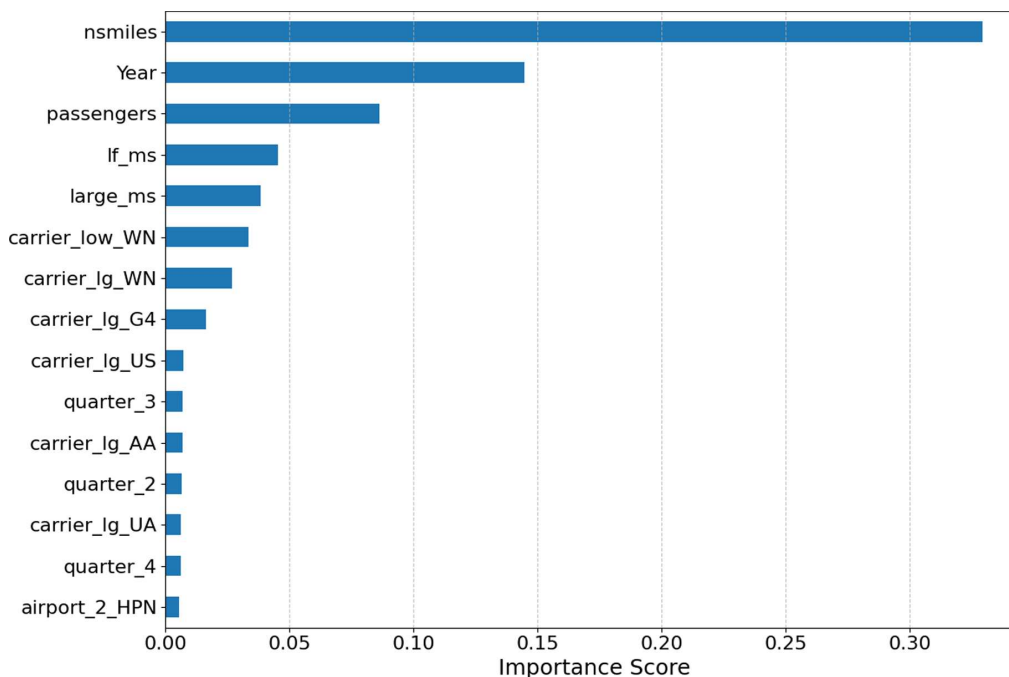
Finally, it can be said that the use of the random forest model did generate unbiased, consistent, and statistically reliable predictions. This has also been confirmed by the residual histogram and scattered diagnostic plot. This plot does provide an extra layer of validation beyond traditional performance measurement. Hence, it also increases confidence in the model's applicability related to prediction application.

4.7. Feature importance analysis

It has been confirmed that route distance, load factor, and seasonal indicators (*i.e.*, quarter) have emerged as the most significant predictors of fare determination, according to the feature relevance analysis. These results highlight the fact that operational and temporal factors are the main causes of airfare variability, which supports both the exploratory data analysis and earlier research.

The plot in Figure 11 shows that route distance (nsmiles), year, and passenger volume contribute most strongly to fare predictions, followed by load factor (lf_ms) and carrier/seasonal

Figure 11
Top 15 feature importances identified by the random forest model



attributes. The distribution highlights the dominance of operational factors over categorical carrier identifiers.

4.8. Prediction evaluation

The prediction framework in this study also proved its practical utility by generating plausible scenario-based forecasts of future airfare trends. For instance, if a user inputs a future setup with the year, quarter, route, and carrier bits, the model then outputs a forecast airfare of 258.25 USD. This shows that the model yields estimates that are both interpretable and highly plausible. The robustness of this predictive framework is underpinned by several key factors.

First, the model showed decent generalization, with an R^2 of 0.838 on the test data. Further, the residual checks didn't show any big systematic skew, which means that the predictive mistake level stays steady across different fare bands. Finally, the feature importance analysis revealed that the model used relied on meaningful operational and time-related variables, rather than random or spurious patterns. Even quarterly error measures stayed stable, which means that the model seems to capture temporal changes well, with the only expected increase in volatility showing up in Q1. Because of these capabilities, the framework offers actionable insights for the decision-making process by airlines, travel agencies, and policy planners. Overall, it can be said that the random forest model delivered strong predictive power and high interpretability, clearly highlighting how operational and seasonal factors shape airfares. Moreover, pairing statistical validation (*i.e.*, ANOVA and Tukey's HSD) with ML helped not only to confirm significant quarterly variations but also to support the integration of statistical inference with ML.

5. Discussion and Conclusion

The results of the empirical study prove that the Random Forest Regressor provides a powerful and explainable model of

airline fare dynamics. As covered earlier, linear regression will struggle to capture nonlinear fare dynamics; hence, the use of a random forest model will demonstrate improved predictive capability due to its ensemble structure and ability to model complex feature interactions. The use of a random forest model makes it suitable for airfare prediction, keeping a note that pricing behavior is influenced by temporal and other multiple independent operational factors. The overall R^2 of 0.838 and RMSE of 28.04 indicate that the model has the predictive strength, which is in line with or greater than other research studies as covered in Reference [21], where it is reported an R^2 value ranging from 0.78 to 0.82. The overall finding gives a clear indication that the use of random forest is suitable for heterogeneous transportation data containing continuous and categorical predictors. The combination of seasonal factors (quarter) with the operational factors, which include distance and load factor, greatly increased the model accuracy, which is also supported by the analysis of the feature importance.

With the help of quarterly performance analysis, it can be seen that Q1 yielded the largest error (MAE = 21.38; RMSE = 31.29) compared to the relatively consistent predictions in Q2–Q4. Such a situation is in accordance with the known patterns in airline pricing literature since Q1 is traditionally marked by the decline in post-holiday demand, promotional pricing campaigns, and weather-related conflicts. Thus, the slight decrease in the accuracy of the model in this timeframe is because of the nature of the market and not because of a flaw in the model. On the other hand, the steady outcomes of the model in the range of Q2–Q4 indicate that the model has a good generalization capacity when the demand is moderate and peak. These seasonal patterns are statistically supported by the results of the ANOVA and the Tukey's HSD test.

The large variations between quarters are especially visible between Q1 and Q3/Q4 and between Q2 and Q4. This finding supports not only the inclusion of temporal features but also the approach to how we can interpret the ML outputs. Mixing statistical inference with predictive modeling bridges the traditional

gap between data-driven practice and theory-driven reasoning, representing a significant methodological advancement.

In line with recent methodological reviews, this study takes a more balanced approach by integrating statistical validation. This improves both reliability and explanatory value [22].

The feature importance analysis gave substantial insights into the factors that drive airfare. Variables like route distance, load factor, and carrier type have been the most influential predictors, which is consistent with operational realities and market dynamics in shaping pricing behavior. The fact that the quarter variable stays prominent reinforces that time-related factors and seasonal fluctuations matter a lot for how fares get formed, and it also suggests the need to model seasonality explicitly in order to capture airfare volatility properly. These interpretable outcomes are useful in practice because it can help airlines operator in understanding on how operational conditions and seasonal demand cycles can affect fare adjustments. This ultimately supports more informed pricing strategies and capacity planning decisions as an approach that matches recent elasticity-based analyses in airline markets [17].

The finding in this paper demonstrates the practical applicability of the used model (*i.e.*, random forest). For example, the scenario-based prediction for Q3 2027 on the JFK–LAX route showed a predicted fare of \$258.25, which illustrates the fact that the framework has the ability to generate interpretable and forward-looking outputs. These outputs can support the decision-making process for airlines and travel agencies, as well as for policy planners. Though this framework is presented as a proof of concept, it has the potential to scale and demonstrate the practical relevance of the model in the real world. This also suggests that the synthesis of ML with statistical hypothesis testing helps in both predictive accuracy and interpretability. This leads to a meaningful methodological contribution to previous studies. The stability and robustness of the predictive framework are further supported by this residual diagnostic analysis.

This study, though having strengths in many aspects, has some limitations. For example, the current model relies on aggregated quarterly data, which may limit or hide intra-quarter fluctuations and temporary shocks such as holidays or sudden demand spikes. As we know, the forecasts rely on historical trends, so it could be assumed that future market behavior will broadly follow the same seasonal rhythm. The main reason for using this framework in this paper is its adaptability and robustness. This study could be extended to include macroeconomic variables, such as inflation, fuel prices, and GDP, as well as other event-related factors like pandemics and major travel disruptions.

For our future work, we will expand this framework by benchmarking it against additional baseline models (*i.e.*, boosting techniques, tree-based approaches, and linear methods). The framework will further validate predictive performance under one unified evaluation protocol. Furthermore, upcoming research will explore advanced deep learning architectures such as Long Short-Term Memory networks and Transformer-based models to effectively capture sequential interactions and nonlinear temporal dynamics. Integrating these models will enable real-time prediction within the dynamic pricing framework. This in turn will enhance revenue management systems by leveraging them more effectively.

In this study, the research integrates statistical validation, seasonal analysis, and ML prediction to have an innovative, explainable, and data-driven airfare prediction framework. This study shows both prediction accuracy and analytical value with the use of U.S. Airline Flight Routes and Fares (1993–2024)

data. The Random Forest Regressor not only demonstrated strong predictive performance compared to literature baselines, but its feature-level transparency also shows the most influential operational and time-dependent drivers of fare variability. The statistical method (*i.e.*, ANOVA and Tukey's HSD testing) provided further confirmation related to the significance of seasonality, showing how data-driven learning can be tied to statistical accuracy, even when dealing with complex real-world conditions in practice. The proposed scenario-based prediction will help the framework expand into other domains too, for example, in forecasting future fares under different market situations.

The model used in this paper does form a scalable platform for future development of interpretable AI-based airfare prediction, although it is subject to limitations related to data granularity and macroeconomic uncertainty.

Overall, this study contributes to the literature on combining ML accuracy with traditional interpretability, offering a novel pathway applicable to both research and industry. The suggested framework can therefore be viewed as a practical and methodological transition point between econometric modeling and predictive data science. This framework aims to assist evidence-based decision-making within the aviation industry in a more data-intensive future. Further, the inclusion of a structured prediction workflow also highlights the potential for event-driven forecasting in airline pricing environments, thereby supporting continuous monitoring and informed decision-making.

Acknowledgments

This study was supported by the guidance of academic colleagues, researchers, and students at York St John University, whose insights contributed meaningfully to the development and refinement of the work. The authors also acknowledge the contributions of industry practitioners and analysts whose perspectives helped contextualize the practical relevance of airfare prediction and interpretability.

Funding Support

This research is supported and funded by York St. John University, Research School under QR Project QR24-25-46, AY24-25.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new datasets were generated for the study. The analysis was conducted using publicly available data from the open-source repository at [US Airline Flight Routes and Fares 1993–2024](https://www.bls.gov/data/airfare). The source code supporting the findings of this study is available in RaYDaR at <https://doi.org/10.25421/yorksj.32509494>.

Author Contribution Statement

Swathi Ganesan: Conceptualization, Methodology, Software, Validation, Data curation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Amjad Alam:** Software, Validation, Funding acquisition. **Sangita Pokhrel:** Validation, Data curation. **Ganapathy Kumar:** Data curation, Supervision, Project administration. **Shiv Kunwar:** Software, Resources, Writing – original draft. **Minel Somasiri:** Formal analysis, Resources, Writing – original draft. **Nalinda Somasiri:** Formal analysis, Supervision, Project administration.

References

- [1] Bala, B., & Behal, S. (2024). A brief survey of data preprocessing in machine learning and deep learning techniques. In *2024 8th International Conference on I-SMAC*, 1755–1762. <https://doi.org/10.1109/I-SMAC61858.2024.10714767>
- [2] Houkan, A., Sahoo, A. K., Gochhayat, S. P., Sahoo, P. K., Liu, H., Khalid, S. G., & Jain, P. (2024). Enhancing security in industrial IoT networks: Machine learning solutions for feature selection and reduction. *IEEE Access*, 12, 160864–160883. <https://doi.org/10.1109/ACCESS.2024.3481459>
- [3] Felahatpisheh, A., Gulati, A., & Valderrama, C. E. (2025). Enhancing machine learning models using a hierarchical two-level genetic algorithm for feature engineering. In *2025 IEEE 11th International Conference on Big Data Computing Service and Machine Learning Applications*, 231–238. <https://doi.org/10.1109/BigDataService65758.2025.00042>
- [4] Khan, R., Shah, A. M., Bareera, Ijaz, A., & Sumeer, A. (2025). Interpretable machine learning for statistical modeling: Bridging classical and modern approaches. *International Journal of Social Sciences Bulletin*, 3(8), 43–50.
- [5] O'Brien Quinn, H., Sedky, Francis, M., J., & Streeton, M. (2024). Literature review of explainable tabular data analysis. *Electronics*, 13(19), 3806. <https://doi.org/10.3390/electronics13193806>
- [6] Sreelekha, M., & Midhunchakkaravarthy. (2024). Intelligent transportation system for sustainable and efficient urban mobility: Machine learning approach for traffic flow prediction. In V. Vimal, I. Perikos, A. Mukherjee, & V. Piuri (Eds.), *Multi-strategy learning environment: Proceedings of ICMSE 2024*, 399–412. Springer. https://doi.org/10.1007/978-981-97-1488-9_30
- [7] Yang, H.-F., & Chen, Y. P.-P. (2019). Hybrid deep learning and empirical mode decomposition model for time series applications. *Expert Systems with Applications*, 120, 128–138. <https://doi.org/10.1016/j.eswa.2018.11.019>
- [8] Guan, Y. (2024). Flight price prediction web-based platform: Leveraging generative AI for real-time airfare forecasting. *Journal of Web Engineering*, 23(2), 299–314. <https://doi.org/10.13052/jwe1540-9589.2325>
- [9] Li, Q., Gao, J., Beck, J. L., Lin, C., Huang, Y., & Li, H. (2024). Probabilistic outlier detection for robust regression modeling of structural response for high-speed railway track monitoring. *Structural Health Monitoring*, 23(2), 1280–1296. <https://doi.org/10.1177/14759217231184584>
- [10] Barsce, J. C., Palombarini, J. A., & Martínez, E. C. (2017). Towards autonomous reinforcement learning: Automatic setting of hyper-parameters using Bayesian optimization. In *2017 XLIII Latin American Computer Conference*, 1–9. <https://doi.org/10.1109/CLEI.2017.8226439>
- [11] Talaat, F. M. (2025). Revolutionizing cardiovascular health: Integrating deep learning techniques for predictive analysis of personal key indicators in heart disease. *Neural Computing and Applications*, 37(1), 1–24. <https://doi.org/10.1007/s00521-024-10453-2>
- [12] Wheeb, A. H., & Khan, M. F. (2025). A survey of several machine learning (ML) algorithms for security solution in Internet of Things (IoT) networks. *Journal of Artificial Intelligence Research & Advances*, 12(1), 1–21.
- [13] Abdella, J. A., Zaki, N. M., Shuaib, K., & Khan, F. (2021). Airline ticket price and demand prediction: A survey. *Journal of King Saud University - Computer and Information Sciences*, 33(4), 375–391. <https://doi.org/10.1016/j.jksuci.2019.02.001>
- [14] Liu, J., Ding, L., Guan, X., Gui, J., & Xu, J. (2020). Comparative analysis of forecasting for air cargo volume: Statistical techniques vs. machine learning. *Journal of Data, Information and Management*, 2(4), 243–255. <https://doi.org/10.1007/s42488-020-00031-1>
- [15] Gürlü, H. E., & Özçalıcı, M. (2025). Optimizing air freight forecasting: The integration of machine learning and empirical mode decomposition. In *2025 9th International Symposium on Innovative Approaches in Smart Technologies* (pp. 1–10). <https://doi.org/10.1109/ISAS66241.2025.11101970>
- [16] Sun, J. (2025). *Machine learning approaches to airfare prediction: A comparative study with hybrid models*. Bachelor's Thesis, LUT University. <https://lutpub.lut.fi/handle/10024/169248>
- [17] Sifat, A. I., & Elahi, M. A. (2025). Elasticity-driven fare optimization in US air travel segments. *Challenges and Issues of Modern Science*, 4(1), 133–138. <https://doi.org/10.15421/cims.4.316>
- [18] Qian, W., Nielsen, T. D., Zhao, Y., Larsen, K. G., & Yu, J. J. (2024). Uncertainty-aware temporal graph convolutional network for traffic speed forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 25(8), 8578–8590. <https://doi.org/10.1109/TITS.2024.3365721>
- [19] Wang, Z., Yin, X. F., Lin, Y. H., Chua, P. C., Li, N., & Fu, X. (2024). Optimizing demand forecasting: A framework with Bayesian optimization embedded reinforcement learning for combined algorithm selection and hyperparameter optimization. In *2024 IEEE Conference on Artificial Intelligence*, 1045–1050. <https://doi.org/10.1109/CAI59869.2024.00190>
- [20] Shi, Y., Wei, P., Feng, K., Feng, D.-C., & Beer, M. (2025). A survey on machine learning approaches for uncertainty quantification of engineering systems. *Machine Learning for Computational Science and Engineering*, 1(1), 11. <https://doi.org/10.1007/s44379-024-00011-x>
- [21] Jaiswal, P., Rasool, A., Wadhvani, R., & Dubey, A. (2024). A study on recent missing value imputation techniques based on machine learning and deep learning. In *2024 IEEE 8th International Conference on Information and Communication Technology*, 1–6. <https://doi.org/10.1109/CICT64037.2024.10899655>
- [22] Bomrah, S., Uddin, M., Upadhyay, U., Komorowski, M., Priya, J., Dhar, E., . . . , & Syed-Abdul, S. (2024). A scoping review of machine learning for sepsis prediction-feature engineering strategies and model performance: A step towards

- explainability. *Critical Care*, 28(1), 180. <https://doi.org/10.1186/s13054-024-04948-6>
- [23] Fernandes, A. A., Koehler, M., Konstantinou, N., Pankin, P., Paton, N. W., & Sakellariou, R. (2023). Data preparation: A technological perspective and review. *SN Computer Science*, 4(4), 425. <https://doi.org/10.1007/s42979-023-01828-8>
- [24] Sreehari, E., & Babu, L. D. (2024). Feature selection techniques for data analysis and decision making in interdisciplinary areas: A systematic review. *IEEE Access*, 12, 188845–188873. <https://doi.org/10.1109/ACCESS.2024.3512878>
- [25] Zhao, H., Gu, M., Qiu, S., Zhao, A., & Deng, W. (2025). Dynamic path planning for space-time optimization cooperative tasks of multiple unmanned aerial vehicles in uncertain environment. *IEEE Transactions on Consumer Electronics*, 71(3), 7673–7682. <https://doi.org/10.1109/TCE.2025.3593383>

How to Cite: Ganesan, S., Alam, A., Pokhrel, S., Kumar, G., Kunwar, S., Somasiri, M., & Somasiri, N. (2026). Interpretable Airfare Prediction Using Random Forests: Seasonal Variation, Statistical Testing, and a Structured Prediction Workflow. *Journal of Data Science and Intelligent Systems*. <https://doi.org/10.47852/bonviewJDSIS62028502>