

Flight Price Prediction Using Comparative Data Analytics and Regression Models

Dev Srivastava
M.Sc Computer Science
University of Warwick, United Kingdom

Abstract—Flight ticket prices can differ widely even for similar routes, often for reasons that are not obvious to travelers. Factors such as airline choice, journey duration, number of stops, and travel date all play a role, but they do not fully explain how prices are set. This report explores how effectively flight prices can be predicted using only the structured information that is visible to users at booking time.

The analysis follows a complete data analytics pipeline, beginning with cleaning and standardizing raw flight data and ending with a comparison of several regression models. Duration values were converted into a consistent numerical format, temporal features were extracted from travel dates, and categorical variables were encoded for modeling. A baseline predictor, Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression were evaluated using Mean Absolute Error, Root Mean Squared Error, and the coefficient of determination on a held out test set.

The results show clear differences between model classes. Linear models struggle to capture the complexity of pricing behavior while ensemble based methods perform substantially better. Further inspection of residuals indicates that prediction errors increase for higher priced tickets suggesting the presence of unobserved pricing factors. Overall, the study highlights both what can and cannot be learned about flight pricing from observable itinerary features alone.

I. INTRODUCTION

Flight pricing is something most travelers encounter yet it often appears unpredictable. Two tickets for the same route can differ significantly in price depending on the airline, the number of stops, the length of the journey, or the date of travel. These differences are driven by various factors such as demand patterns, competition amongst airlines, their revenue management strategies, consumer demand patterns and operational costs. However, many of these factors are not directly visible to customers which makes understanding and predicting flight prices a challenging analytical problem.

From the perspective of data analytics, this creates an important limitation. Users and third party booking platforms only have access to structured information shown during a flight search, such as airline name, route, travel duration, number of stops, and calendar details. Internal airline information, including seat availability or real time demand is not available. As a result, the key question is not whether prices can be predicted perfectly but how much of the observed price variation can be explained using this limited but realistic set of features.

This project takes a comparative approach to address this question. Rather than relying on a single predictive model sev-

eral regression techniques are evaluated ranging from simple baselines to more flexible non linear methods. Simpler models provide transparency and interpretability whereas complex models are used to capture interactions and non linear effects that may exist in the data. This allows performance differences to be analysed in a structured and meaningful way.

The focus of the coursework is not purely on maximising predictive accuracy. But an equal emphasis is placed on data preparation, evaluation methodology, and interpretation of results. Diagnostic plots and feature importance analysis are used to understand where models perform well and where they struggle, particularly for high priced tickets.

The study is guided by the following research questions:

- **RQ1:** To what extent can structured flight attributes alone explain observed airfare variation?
- **RQ2:** Do non-linear ensemble models provide meaningful analytical advantages over linear baselines in modelling airfare dynamics?
- **RQ3:** Which feature groups dominate price formation in this dataset?

These questions are operationalised through the following hypotheses:

- **H1:** Linear models underperform due to their limited ability to capture non-linear interactions between route structure and temporal features.
- **H2:** Ensemble methods outperform single learners by reducing bias and/or variance through aggregation or sequential error correction.
- **H3:** Duration and calendar-based features contribute substantial explanatory power, beyond airline identity indicators alone.

The report is further segregated in the following manner:

- **Section II** - Reviews related work on flight pricing and regression modelling.
- **Section III**- Describes the dataset and its analytical relevance.
- **Section IV**- Details the data cleaning and preprocessing steps.
- **Section V**- Outlines the modelling methodology and evaluation metrics.
- **Section VI**- Results
- **Sections VII and VIII**- Discussion and Limitations
- **Section IX**- Conclusion & Future Works

II. RELATED WORK

Flight price forecasting has been studied using different statistical models and related machine learning methods, ranging from linear regression to ensemble approaches. From an aviation economics viewpoint, airfare dynamics reflect a mixture of cost drivers, competitive interactions, and demand segmentation, which produces non-linear behavior at the itinerary level [1]. In real world scenarios, consumer-facing indicators frequently rely on structured itinerary attributes and historical observations rather than the airline in-house reality such as remaining inventory or revenue management constraints.

Earlier work on this topic had explored predictive modeling for fare analysis using machine learning and web-derived observations, emphasizing the complexity introduced by random dynamics and heterogeneous market conditions [2]. Linear regression remains attractive due to simplicity and interpretability however, its additive structure may not be able to function when interactions dominate. Decision trees can capture non-linear splits and interactions automatically but may exhibit high variance and overfitting when used as single learners [3]. Random Forests reduce variance by averaging many uncorrelated trees, improving generalization while retaining flexibility [4]. Gradient boosting methods build additive models sequentially, often achieving strong predictive performance by reducing bias through iterative fitting of residual structure [5].

Methodologically, standard learning texts emphasize the importance of cross-validation, careful evaluation, and bias-variance trade-offs [5], [6]. In regression, MAE and RMSE are widely used; RMSE penalizes large errors more strongly, while MAE provides an interpretable mean absolute deviation [7]. These principles motivate the evaluation design in this project, particularly the inclusion of cross-validation RMSE for stability and residual diagnostics for interpretation of the error-structure.

III. DATASET DESCRIPTION

The dataset contains structured records of flight options with associated prices. Each row corresponds to a flight itinerary, and the target variable is **Price**. Explanatory variables include categorical and numerical information:

- **Carrier information:** airline name (categorical).
- **Route information:** source city and destination city (categorical).
- **Trip structure:** number of stops (discrete numerical).
- **Timing features:** departure/arrival time fields and travel date components (day and month).
- **Duration:** originally stored in mixed textual formats (e.g., hours and minutes), later converted to a single numeric representation in minutes.

A key contextual point is representativeness: although this dataset does not include internal airline inventory levels or competitor pricing in real time, it closely reflects the type of structured attributes available at query time to consumers and third-party aggregators. From a decision-support perspective,

TABLE I
SUMMARY OF DATASET ATTRIBUTES

Attribute	Type	Description
Airline	Categorical	Operating airline
Source	Categorical	Departure city
Destination	Categorical	Arrival city
Total_Stops	Numerical	Number of intermediate stops
Duration (min)	Numerical	Total travel time in minutes
Day / Month	Numerical	Calendar features
Price	Numerical	Target variable

these are the variables a user can actually observe when choosing between itineraries (route, stops, airline, and travel date). Thus, the dataset is suitable for modelling *observable* determinants of fare differences, while any remaining error can be interpreted as the contribution of latent factors (e.g., demand shocks, special events, last-minute booking effects, or strategic pricing).

A. Summary of Dataset Attributes

Table I provides a concise overview of the key variables used in the analysis, including their data types and analytical roles. This summary serves as a reference point for the subsequent exploratory and modeling stages. Together, these attributes form the basis for feature engineering and regression analysis developed in the following sections.

B. Exploratory Data Analysis Summary

Before modeling, a brief exploratory analysis was used to understand price range, distributional shape, and potential outliers. In typical flight-price datasets, prices are right-skewed: many observations cluster in a lower-mid fare band, with fewer examples of very high prices. This matters analytically for two reasons. First, evaluation metrics such as RMSE can be dominated by a small number of extreme fares, since squared error penalizes large deviations more heavily [7]. Second, model residuals may display heteroscedasticity (non-constant variance), where errors become larger at higher price levels; this often signals missing explanatory variables or non-linear regime changes in the target distribution.

EDA also motivates feature engineering choices. Duration tends to correlate with price, but the relationship is not purely linear: stops, airline identity, route competitiveness, and timing can modify the marginal effect of duration. Similarly, calendar variables (day and month) can capture seasonal demand patterns and day-level fluctuations. These insights justified (i) converting duration to a consistent numerical feature and (ii) including temporal decomposition features to capture seasonal variability.

IV. DATA CLEANING AND PREPROCESSING

Raw datasets commonly contain missing values, inconsistent formatting, and mixed data types. Since supervised learning models assume consistent numeric representations,

preprocessing is very important for both performance and interpretability.

A. Handling Missing Values

Initial inspection identified a small number of missing or null entries across certain columns. Two standard approaches are: (i) imputation (e.g., mean/median for numerical, most-frequent for categorical), or (ii) removal of incomplete rows. Because airfare pricing depends on interacting variables, naive imputation can create unrealistic combinations (e.g., imputed airline-route pairs) that can change relationships. Therefore, rows with missing values were removed to preserve data integrity and ensure models were trained on complete feature vectors.

B. Removal of Redundant Attributes

Some attributes were redundant or uninformative for modelling (e.g., identifiers or columns duplicating information represented elsewhere). Removing them reduces noise, improves training efficiency, and simplifies interpretation. For linear models, this also reduces multicollinearity risk.

C. Standardisation of Temporal and Duration Formats

Temporal variables were decomposed into components to represent seasonality and day-level effects (e.g., extracting month and day). Duration values were originally stored in textual formats. These were converted into a single numerical feature that is total minutes. This standardization ensures the model treats duration as a continuous variable with a consistent scale and reduces parsing related noise.

D. Categorical Encoding

Categorical variables such as airline, source, and destination cannot be used directly by many regression models. One-hot encoding was used to transform each category into binary indicator columns. While one-hot encoding increases dimensionality, it enables both linear and tree based models to use categorical information effectively.

E. Train-Test Split and Leakage Prevention

A standard 80:20 split was used to create training and test partitions. All model training, validation, and hyperparameter decisions were performed using only the training portion, with final reporting on the held out test set. This prevents information leakage and ensures reported performance reflects generalization. The target variable was retained on its original monetary scale, so MAE and RMSE remain interpretable in the same units as the original prices.

F. Final Dataset Preparation

After cleaning and encoding, the final modeling dataset consisted entirely of numerical features. The feature set includes continuous variables (e.g., duration in minutes), discrete variables (e.g., number of stops), temporal components (day and month), and one-hot encoded categorical indicators (airline and route fields). This uniform representation is a requirement for standard scikit-learn regression models and enables fair comparison across model classes.

V. METHODOLOGY

A. Feature Engineering

Feature engineering transforms raw attributes into representations better aligned with predictive structure. The main engineered features used in this project are:

- **Duration in minutes:** conversion from mixed formats to a single numeric feature.
- **Calendar decomposition:** month and day extraction to capture seasonal/daily pricing effects.
- **Stop count representation:** numeric encoding of number of stops.
- **One-hot encoding:** airline, source, and destination indicators.

The motivation is to enable both linear and non-linear models to learn useful relationships while keeping the pipeline reproducible and interpretable [6].

B. Models Considered

Models were chosen to form a controlled complexity ladder, enabling analytical comparison of assumptions and error behaviour.

1) *Baseline Mean Predictor:* A baseline predictor provides a reference point by predicting the mean training price for all test instances. While simplistic, it is useful to confirm that models are learning structure rather than merely matching a global average.

2) *Linear Regression:* Linear Regression models the target as a weighted sum of features. It is fast and interpretable but limited when relationships are non-linear or interaction-driven unless interactions are explicitly engineered [6].

3) *Decision Tree Regressor:* Decision trees capture non-linear splits and interactions automatically. However, they are prone to overfitting when deep and can exhibit high variance across different train-test splits [3].

4) *Random Forest Regressor:* Random Forests reduce variance through bagging and feature randomness. Each tree learns a slightly different structure, and the final prediction averages across trees, typically improving generalization [4].

5) *Gradient Boosting Regressor:* Gradient boosting fits an additive sequence of weak learners that iteratively correct previous errors. It can be highly accurate by reducing bias and modeling structured residual patterns, but may require careful tuning to avoid overfitting [5].

C. Hyperparameter Strategy

To strengthen methodological robustness, tree-based models were evaluated with reasonable hyperparameter settings affecting model capacity and regularization. Typical parameters include:

D. Evaluation Metrics

Three standard regression metrics were used.

TABLE II
TYPICAL HYPERPARAMETERS CONSIDERED FOR TREE-BASED MODELS

Model	Key Hyperparameters	Purpose
Decision Tree	max_depth, min_samples_split, min_samples_leaf	control overfitting
Random Forest	n_estimators, max_depth, max_features, min_samples_leaf	variance reduction
Gradient Boosting	n_estimators, learning_rate, max_depth, subsample	regularisation + bias reduction

1) *Mean Absolute Error (MAE)*: MAE measures the average absolute deviation:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|.$$

It is robust and easy to interpret in the same units as the target.

2) *Root Mean Squared Error (RMSE)*: RMSE is:

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

It penalizes larger errors more strongly and is sensitive to outliers [7].

3) R^2 : R^2 measures explained variance relative to a mean predictor. Higher values indicate better fit; however, for real-world pricing, R^2 should be interpreted alongside MAE/RMSE since it can mask regime-specific errors.

E. Cross-Validation for Stability

Cross-validation RMSE was computed to reduce dependence on a single train-test split. This is important because model performance can vary depending on how the data is partitioned, especially when the target distribution is skewed or contains rare high price cases. The bias variance framework explains why flexible models can overfit without proper regularization and why ensembles can improve stability by reducing variance and/or bias [5], [8].

VI. RESULTS

A. Model Performance Comparison

Table III summaries test set performance across models. Ensemble methods substantially reduce prediction error compared with the baseline and linear model, supporting the view that non linear structure exists in the feature set.

B. Metric Comparison Across Models

Figures 1–6 provide visual comparisons across models. The baseline performs worst confirming that a constant price estimate is insufficient. Linear Regression improves substantially but remains limited by its linear hypothesis. Decision Tree reduces error further indicating that non linear splits capture

TABLE III
MODEL PERFORMANCE COMPARISON (TEST SET AND CROSS-VALIDATION)

Model	MAE	RMSE	R^2	CV RMSE
Baseline Mean Predictor	3676.40	4643.71	–	–
Linear Regression	1975.10	2854.20	0.62	2828.05
Decision Tree	1217.21	2233.49	0.77	2280.10
Random Forest	1154.24	1975.17	0.82	2055.28
Gradient Boosting	1334.61	1906.70	0.83	–

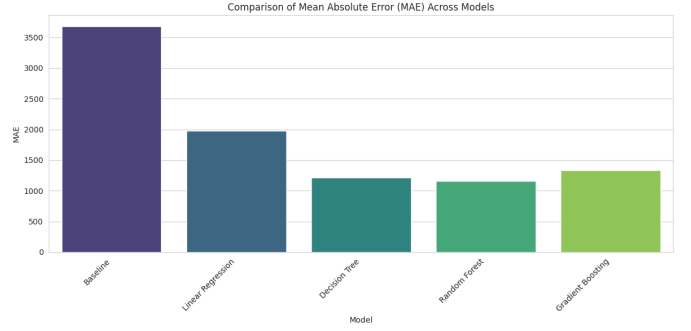


Fig. 1. Comparison of MAE across models (view 1).

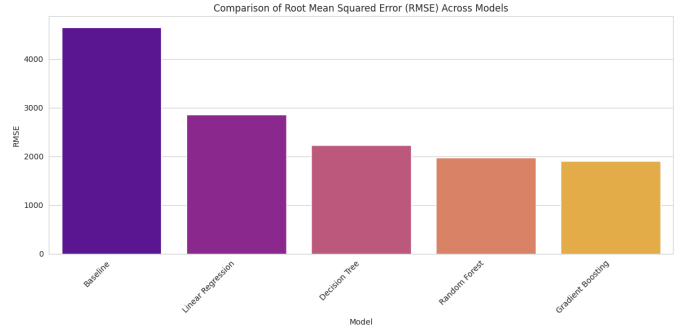


Fig. 2. Comparison of RMSE across models (view 1).

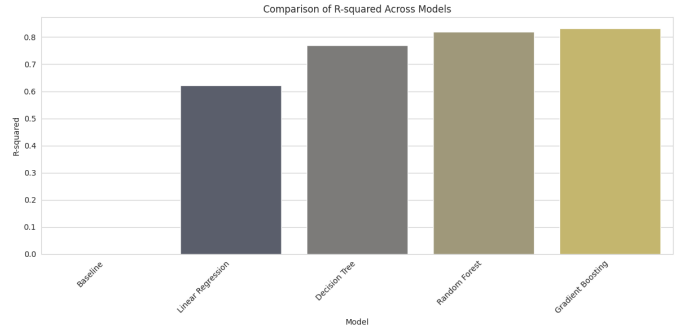


Fig. 3. Comparison of R^2 across models (view 1).

important structure. Random Forest and Gradient Boosting achieve the lowest RMSE and highest R^2 , supporting the conclusion that ensemble approaches better model complex fare interactions.

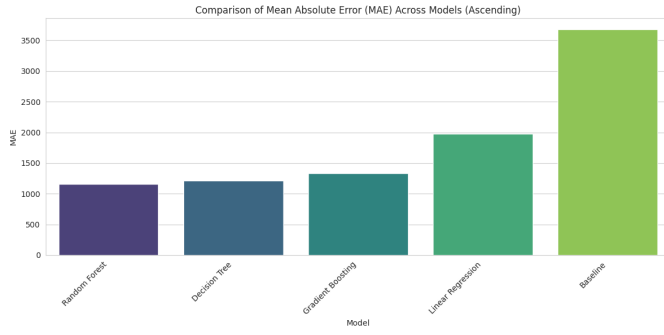


Fig. 4. Comparison of MAE across models (view 2).

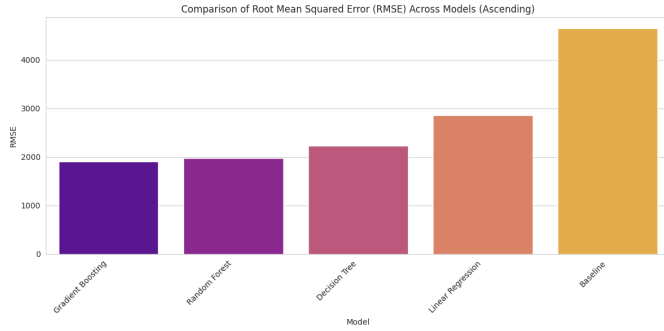


Fig. 5. Comparison of RMSE across models (view 2).

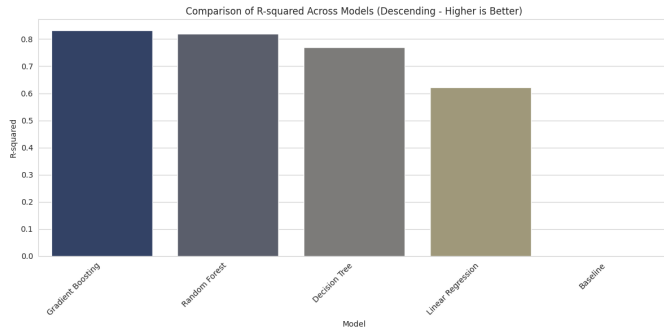


Fig. 6. Comparison of R^2 across models (view 2).

C. Linear Regression Diagnostics

Figure 7 plots actual versus predicted values for the Linear Regression model. A perfect predictor would lie on the diagonal line. Predictions align reasonably in the mid price region but diverge more for higher fares, which is consistent with linear underfitting for extreme prices and suggests that high price tickets may be driven by additional interactions or missing explanatory signals.

Figure 8 plots residuals versus predicted prices. Ideally, residuals are randomly scattered around zero with roughly constant variance. Instead, residual spread increases with predicted price suggesting heteroscedasticity. This phenomenon is common in cost prediction tasks and supports the use of non linear models that can better capture conditional patterns and interactions.

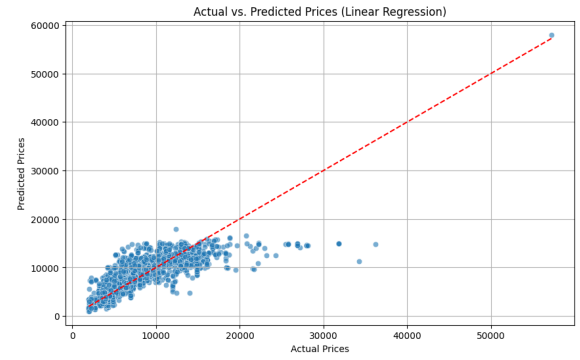


Fig. 7. Actual vs. predicted prices (Linear Regression).

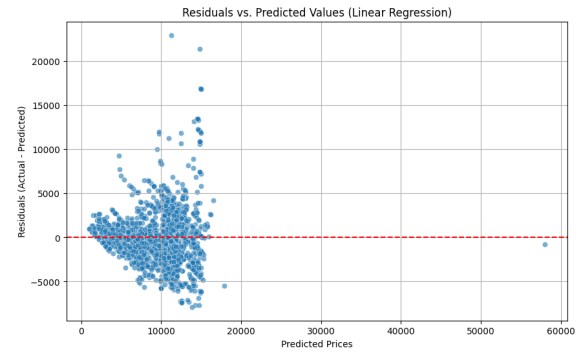


Fig. 8. Residuals vs. predicted values (Linear Regression).

D. Feature Importance Analysis

Interpretability is important in applied analytics especially when results inform downstream decisions. Tree based models enable post-hoc analysis via feature importance scores. Figures 9 and 10 show that `duration_in_min` is the most influential feature, indicating that longer travel time is associated with higher prices. This aligns with intuitive expectations: longer flights can imply greater operational cost, more complex routing, and different demand patterns.

Temporal variables (day and month) also appear as important predictors. This is consistent with seasonality and day-level effects, where certain travel days and months experience higher demand and therefore higher observed prices. Airline and route indicators also contribute reflecting heterogeneous pricing policies across carriers and city pairs.

Importance should be interpreted carefully they quantify how a feature reduces impurity or error within the tree ensemble not causal influence. Nonetheless, they provide valuable analytics insight into which attributes the model relies on most and can guide future feature collection.

E. Gradient Boosting Diagnostics

Figure 11 shows the actual versus predicted plot for Gradient Boosting. Compared to Linear Regression, the scatter lies closer to the diagonal for many points indicating improved fit. Boosting is effective at capturing structured non linear interactions by sequentially correcting residual errors [5].

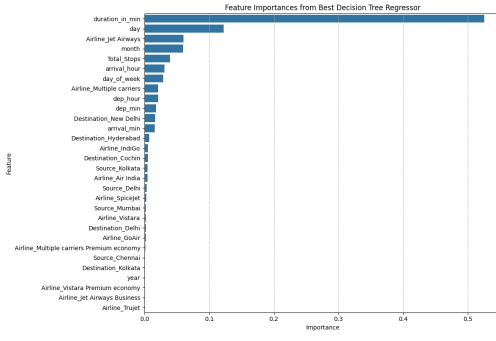


Fig. 9. Feature importances from the best Decision Tree Regressor.

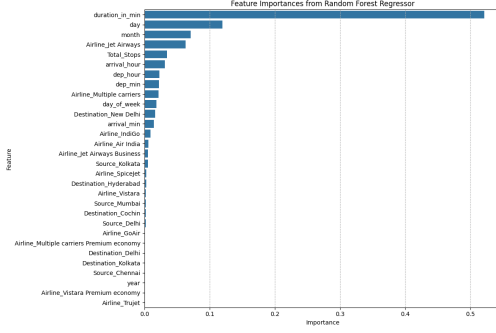


Fig. 10. Feature importances from the Random Forest Regressor.

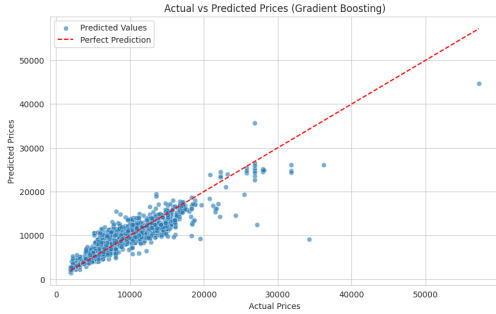


Fig. 11. Actual vs. predicted prices (Gradient Boosting).

However, some deviations remain for extreme fares, which is expected due to limited feature coverage and potential outliers.

VII. DISCUSSION

A. Answering the Research Questions

RQ1 (explainability from structured features): The improvement from the baseline to Linear Regression indicates that structured itinerary features contain meaningful predictive information. Calendar components, stops, duration, and categorical route/airline indicators jointly explain a substantial portion of price variability ($R^2 \approx 0.62$ for Linear Regression). However, the presence of notable residual variance, especially for high fares suggests that structured attributes alone cannot fully explain pricing outcomes consistent with real world revenue management where latent demand and inventory signals are influential.

RQ2 (benefit of non-linear ensembles): The results show a clear performance progression across model complexity. Decision Trees improve over linear models by capturing interactions but are vulnerable to overfitting and instability [3]. Random Forest reduces variance through averaging and achieves strong generalization, consistent with theory and prior evidence [4]. Gradient Boosting achieves the highest R^2 and lowest RMSE in this comparison. This supports the claim that modeling airfare requires non linear representational capacity and interaction capture validating **H1** and **H2**.

RQ3 (dominant feature groups): Feature importance indicates that duration dominates followed by temporal variables and selected airline/route indicators. This supports **H3**: pricing is strongly related to itinerary length and time-of-travel effects while airline identity and destination contribute additional differentiation.

B. Why Gradient Boosting Can Outperform Random Forest in This Setting

Although Random Forest achieves a lower MAE than Gradient Boosting in the reported comparison, Gradient Boosting attains stronger R^2 and lower RMSE. This pattern can occur when boosting reduces large errors for harder cases (thereby improving RMSE and explained variance) by fitting residual structure sequentially, while Random Forest may produce robust median-like predictions that minimize typical absolute deviations.

Analytically, this dataset is medium dimensional after one-hot encoding and contains structured interactions. Boosting can be particularly effective in such settings because it incrementally improves fit on regions where the current ensemble underperforms. In contrast, Random Forest primarily reduces variance via averaging it may smooth predictions in a way that helps MAE but can leave structured residual patterns uncorrected. The diagnostic plots provide supporting evidence that linear residuals show heteroscedasticity and interaction driven error growth motivating models that learn conditional structure rather than global additive patterns.

C. Error Patterns, Heteroscedasticity, and Outliers

Residual diagnostics indicate that error variance increases for higher predicted prices (heteroscedasticity). Several plausible explanations align with domain intuition and prior research on airfare dynamics [1], [2]:

- **Missing demand-side signals:** last minute booking urgency, holiday spikes, or sudden demand shifts are not explicitly encoded.
- **Inventory and dynamic pricing:** internal seat availability and airline revenue management policies can cause abrupt price changes not predictable from itinerary structure alone.
- **Data imbalance:** fewer high fare samples limit the model's ability to learn extreme regimes, increasing error for rare cases.

These factors explain why even the best-performing model still exhibits deviations for extreme fares.

D. Interpretability and Decision-Support Impact

A key analytics goal is not only prediction but also actionable interpretation. The importance ranking suggests that, within this dataset, the duration feature is a primary driver, with calendar effects also contributing meaningfully. In a decision-support setting (e.g., a fare estimator or consumer advisory tool), this implies:

- Duration and routing complexity (stops) should be treated as central signals in fare benchmarking.
- Temporal features can be used to communicate demand periods and expected pricing shifts across the calendar.
- Airline and route indicators capture systematic policy differences, supporting personalised comparisons for specific corridors.

From an interpretability viewpoint Linear Regression provides transparent coefficients but underfits extremes ensembles provide stronger performance but rely on less transparent decision structures. This is a practical trade-off for high-stakes explanations linear baselines may still be valuable for accurate estimation ensembles are preferred.

E. Alternative Analytical Framings

Two alternative framings could strengthen analytics insight in future work:

- **Log-price modelling:** predicting $\log(\text{Price})$ can reduce skewness and heteroscedasticity often improving stability and interpretability of multiplicative effects.
- **Segmented modelling:** splitting fares into regimes (low/medium/high) or stratifying by route could enable models to learn more homogeneous relationships and reduce extreme error growth.

This project retained original prices to keep errors interpretable in monetary units and to maintain direct relevance for decision-support.

VIII. LIMITATIONS

This study is limited by available features and data scope:

- **Missing demand signals:** events, holidays, and booking lead time are not present.
- **Dynamic pricing:** real-time inventory and competitor prices are not included.
- **Potential dataset bias:** certain routes/airlines may dominate the dataset, affecting generalization.
- **Outlier sensitivity:** RMSE is influenced by extreme fares; robust modelling could consider log-transform targets or quantile regression variants.

Despite these limitations, the workflow provides coherent analytics and demonstrates that ensemble regression models can significantly improve fare prediction accuracy using structured attributes.

IX. CONCLUSION AND FUTURE WORK

This report concludes an end-to-end data analytics pipeline for flight price prediction. After systematic preprocessing and feature engineering multiple regression models were evaluated

TABLE IV
ANALYTICAL TAKEAWAYS FROM MODEL COMPARISON

Aspect	Analytical Insight
Best RMSE / R^2	Gradient Boosting captures structured non-linear residual patterns
Lowest MAE	Random Forest provides robust average performance
Key predictor	Duration in minutes dominates importance rankings
Calendar effects	Day and month contribute meaningful seasonal/daily signal
Error behaviour	Larger errors at high prices (heteroscedasticity; outliers)
Interpretability	Linear models are transparent but underfit extremes; ensembles are more accurate

using standard metrics. The comparative results show that non-linear tree-based ensembles outperform linear baselines, suggesting airfare depends on interactions and non-linear effects not captured by simple additive models. Diagnostic plots further reveal heteroscedastic error patterns especially for high fares consistent with missing demand side signals and dynamic pricing effects. Feature importance analysis shows duration and calendar variables dominate model reliance with airline and route indicators providing additional differentiation.

Future work could improve realism and performance by integrating booking lead time, holiday indicators, and dynamic market signals. Methodologically, more systematic hyperparameter optimization and robust evaluation across multiple splits would strengthen conclusions. Interpretability could be enhanced with partial dependence analysis or SHAP style explanations to better characterize interaction effects. Finally, a time aware validation strategy could mirror real-world deployment where models are trained on past data and evaluated on future periods.

REFERENCES

- [1] E. Pels and N. Njegovan, "Airfare pricing and demand analysis," *Journal of Air Transport Management*, vol. 15, no. 1, pp. 1–6, 2009.
- [2] O. Etzioni, R. Turchetta, and C. A. Knoblock, "A predictive model for flight fare analysis," *Artificial Intelligence*, vol. 170, no. 16–17, pp. 1243–1287, 2006.
- [3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1984.
- [4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.
- [6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.
- [7] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)? arguments against avoiding rmse in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.
- [8] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992.