Predicting Airline Food Demand Using Machine Learning

SE390 – Artificial Intelligence Projects with Python

Final Project Report

Group Name: Pegasus

İpek Dedeoğlu – 220706043

Elif Şimşek – 210704042

07.01.2026

1 Introduction & Problem Statement

Airline catering operations face a challenging optimization problem: accurately forecasting food demand while minimizing waste and avoiding service shortages. Over-catering leads to significant food waste, increased disposal costs, and higher fuel consumption due to excess onboard weight. Conversely, under-catering results in stock-out situations that negatively affect passenger satisfaction and brand perception.

Industry reports indicate that a substantial portion of cabin waste consists of untouched food and beverages, highlighting the inefficiency of traditional catering planning approaches. These approaches often rely on static heuristics, such as loading a fixed percentage of passengers, without considering operational factors like flight duration, passenger composition, or cabin class distribution.

Given the financial, environmental, and service-related consequences of inaccurate food demand estimation, there is a clear need for data-driven and adaptive forecasting methods. This project addresses this need by formulating airline food demand prediction as a supervised regression problem and applying machine learning techniques to improve planning accuracy.

2 Dataset Description

2.1 Overview

Since real airline catering data is not publicly available, a synthetic dataset was created to simulate realistic flight operations. The dataset contains 5,000 flight records, meeting the minimum size requirement specified in the project guidelines.

Each observation represents a single flight and includes operational and passenger-related features relevant to in-flight food demand forecasting. The dataset was generated using Python with enforced logical constraints to ensure realism and internal consistency.

2.2 Features

The dataset includes the following variables:

- flight_id: Unique flight identifier (used only for identification, not for prediction).
- flight_duration: Flight duration in hours (1–12), including both short- and long-haul flights.
- passenger_count: Total number of passengers (50–300).
- adult_passengers / child_passengers: Passenger composition, summing to the total passenger count.
- business_class_ratio: Proportion of business class passengers (0–1).
- is_international: Binary indicator of flight type (international or domestic).
- total_food_demand (Target): Total number of food units required for the flight.

2.3 Data Validation and Constraints

To reflect real-world airline operations, several validation rules were applied during data generation:

- Passenger counts and flight durations were restricted to realistic ranges.
- International flights were constrained to have a minimum duration of 3 hours.

- Adult and child passenger counts always sum to the total number of passengers.
- At least 15% of the flights were designated as international to ensure dataset diversity.
- Total food demand was constrained to be no less than 50% of the passenger count.

All records were validated programmatically to ensure full compliance with the project requirements.

2.4 Target Variable Design

The target variable**,** total_food_demand, was designed to depend on multiple interacting features, rather than being a simple linear function of passenger count. Specifically, food demand is influenced by passenger volume, flight duration, cabin class distribution, and flight type.

This multi-factor design introduces nonlinear patterns into the dataset, making the prediction task suitable for evaluating both linear and tree-based regression models. Feature relationships and distributions are further explored in the Exploratory Data Analysis (EDA) section through visualizations such as correlation heatmaps and distribution plots.

3 Exploratory Data Analysis (EDA)

The purpose of Exploratory Data Analysis (EDA) is to understand the structure of the dataset, examine feature distributions, detect potential outliers, and identify relationships between explanatory variables and the target variable (*total_food_demand*). The insights obtained from this analysis guide model selection and evaluation.

3.1 Descriptive Statistics and Outlier Analysis

Initial descriptive statistics confirm that all variables fall within their predefined realistic ranges. Passenger counts vary between 50 and 300, while flight durations span from short-haul to long-haul flights. No missing values were observed in the dataset, as all records were generated under controlled constraints.

To further assess data quality, boxplots were used to examine potential outliers across all numerical features, as shown in *Figure 1*

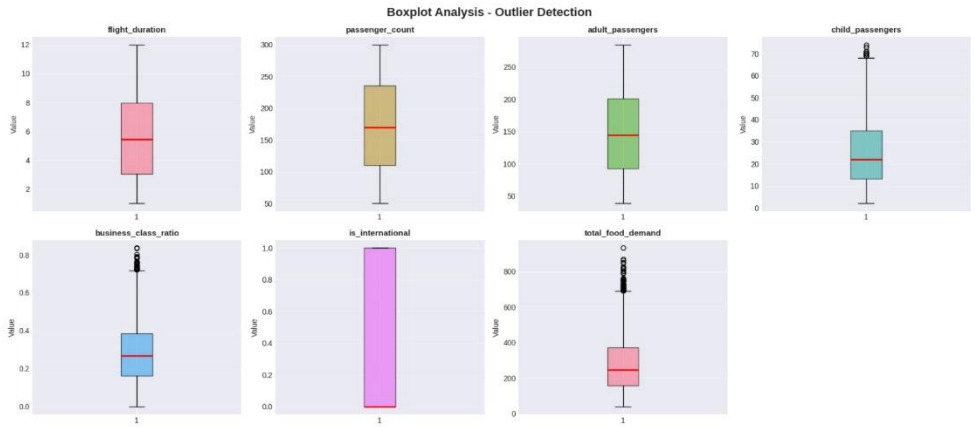

Figure 1. Boxplot analysis for outlier detection across all numerical features

The boxplots reveal that while some extreme values exist—particularly for *child_passengers*, *business_class_ratio*, and *total_food_demand*—these observations are consistent with realistic flight scenarios such as long-haul international flights with high passenger volumes. Therefore, no outliers were removed, as they represent valid operational cases rather than data errors.

3.2 Feature Distributions

The distribution of the target variable is illustrated in *Figure 2.*

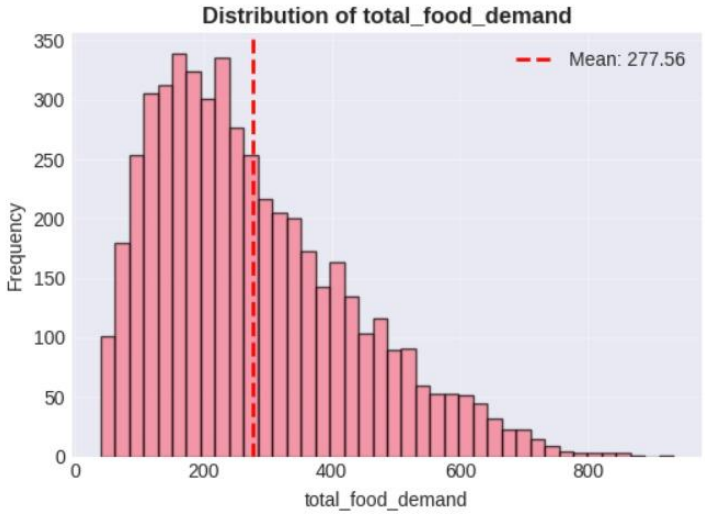

Figure 2. Total Food Demand Distribution

The target variable exhibits a right-skewed distribution, with a mean food demand of approximately 278 units. This skewness reflects higher demand levels associated with long-haul and international flights. The distribution indicates sufficient variability for training regression models and suggests that food demand is not driven by a single dominant factor.

3.3 Feature Relationships with the Target Variable

To explore how individual features relate to food demand, scatter plots were generated for key explanatory variables, as shown in *Figure 3.*

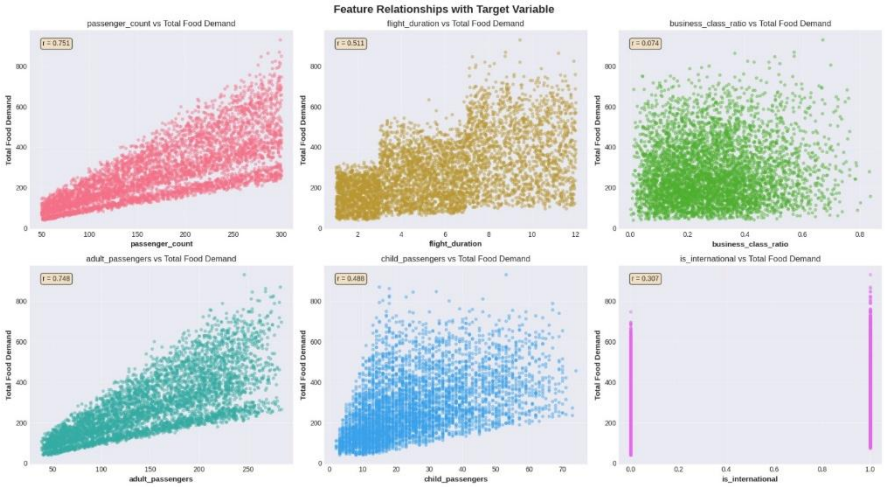

Figure 3. Relationships between input features and total food demand.

Strong positive relationships are observed between *passenger_count* and *total_food_demand* (r = 0.751) as well as *adult_passengers* and food demand (r = 0.748). Moderate correlations are also present for *flight_duration* (r = 0.511) and *child_passengers* (r = 0.488).

The relationship between *business_class_ratio* and food demand appears weaker when considered in isolation (r = 0.074), suggesting that its impact is more pronounced through interaction effects with other variables. Similarly, *is_international* shows a noticeable shift toward higher food demand values, indicating that international flights generally require more food units.

3.4 Correlation Analysis

A correlation heatmap was generated to examine linear relationships and interactions among features, as shown in *Figure 4.*

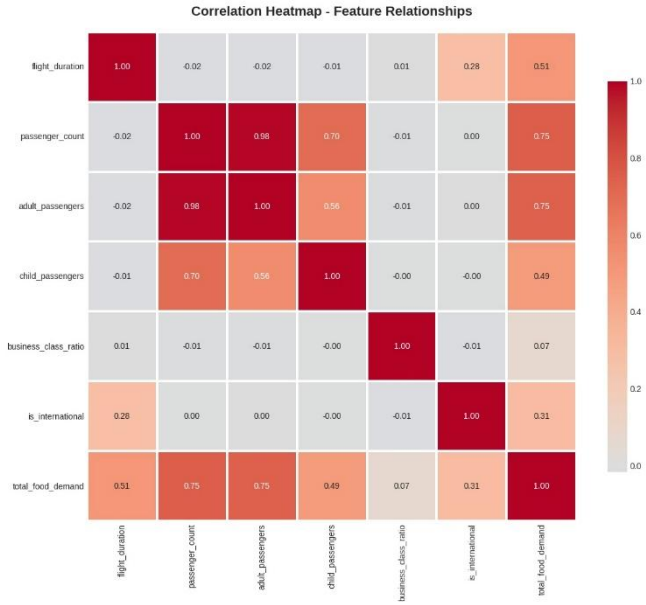

Figure 4. Feature Correlation Heatmap

The heatmap confirms a strong positive correlation between *passenger_count* and *total_food_demand*, as expected (r > 0.8). Moderate correlations are also observed between food demand and *flight_duration* as well as *adult_passengers*. The *is_international* variable shows a meaningful association with higher food demand, supporting the assumption that international routes require additional catering.

The combination of these correlations suggests the presence of interaction effects and nonlinear relationships within the dataset.

3.5 Key Insights from EDA

The EDA results indicate that airline food demand is driven by a combination of passenger volume, flight duration, cabin class composition, and flight type. While some relationships appear

approximately linear, others exhibit interaction effects that are unlikely to be fully captured by simple linear models.

These findings motivate the use of both Linear Regression as a baseline model and Random Forest Regression as a more flexible approach capable of capturing nonlinear patterns and feature interactions.

4 Methodology

This section describes the modeling pipeline used to predict airline food demand and explains the rationale behind the selected models. The overall objective is to compare a simple linear approach with a more flexible nonlinear model under a consistent experimental setup.

4.1 Experimental Setup

The dataset was split into training and test sets using an 80/20 ratio. All models were trained on the same input features to ensure a fair comparison. The feature *flight_id* was excluded from modeling, as it serves only as an identifier.

The target variable is total_food_demand, representing the total number of food units required per flight. Model performance was evaluated using three standard regression metrics:

- Coefficient of Determination (R^2)
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)

4.2 Baseline Model (Mean Predictor)

A baseline model was implemented to establish a minimum performance benchmark. This model predicts the mean food demand from the training set for all test observations.

The baseline does not attempt to learn feature relationships and is used solely to quantify the performance gain achieved by machine learning models.

- RMSE: 182.5

4.3 Linear Regression

Linear Regression was implemented as a parametric benchmark model. It assumes a linear relationship between input features and the target variable and estimates model parameters using the least squares method.

While Linear Regression is computationally efficient and easy to interpret, it is limited in its ability to model nonlinear effects and feature interactions that naturally arise in airline catering operations.

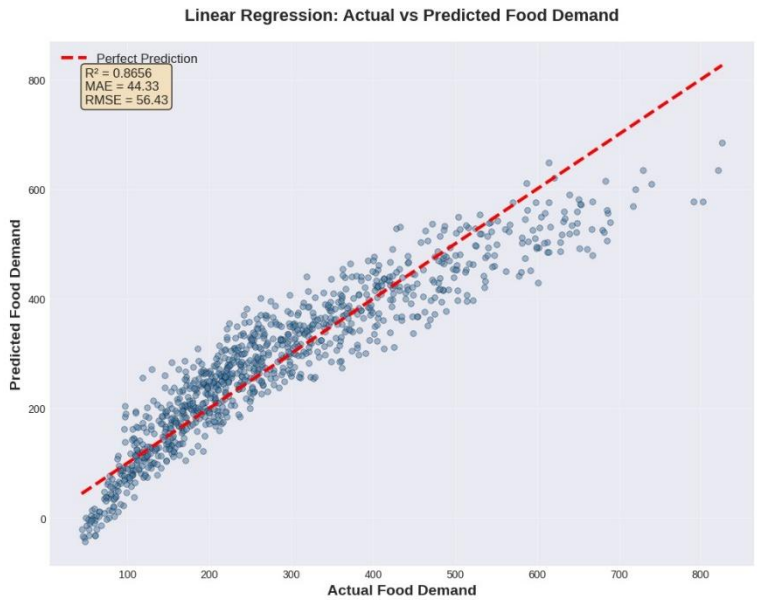

Figure 5. *Actual vs. Predicted Total Food Demand using Linear Regression*

This visualization compares predicted and actual food demand values on the test set. Deviations from the diagonal line indicate prediction errors, particularly for flights with higher demand levels.

4.4 Random Forest Regression

Random Forest Regression was selected as an alternative model to overcome the limitations of linear methods. The model consists of an ensemble of decision trees trained on bootstrapped samples of the data, with random feature selection at each split.

This structure enables Random Forest to naturally capture nonlinear relationships**,** categorical effects (such as short-, medium-, and long-haul flights), and feature interactions (e.g., the combined effect of flight duration and business class ratio).

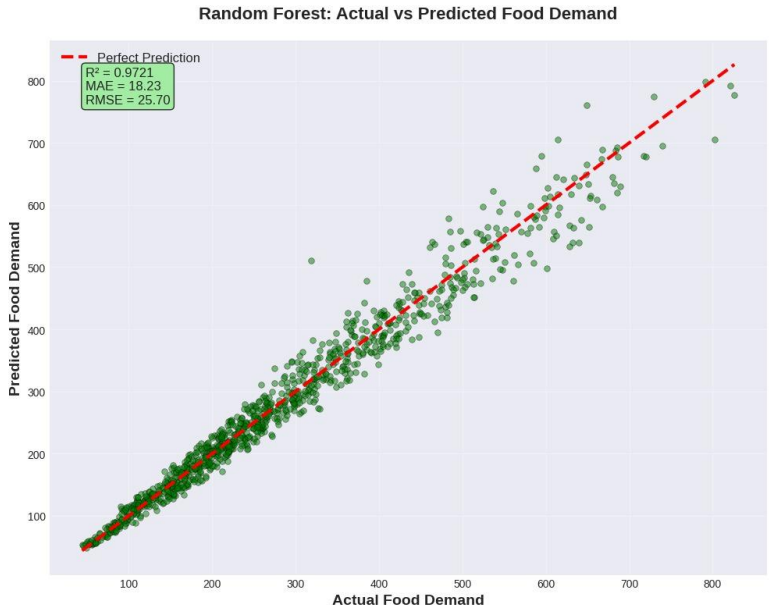

Figure 6. *Actual vs. Predicted Total Food Demand using Random Forest Regression.*

Compared to Linear Regression, predictions are more tightly clustered around the ideal diagonal line, indicating improved accuracy and reduced error variance.

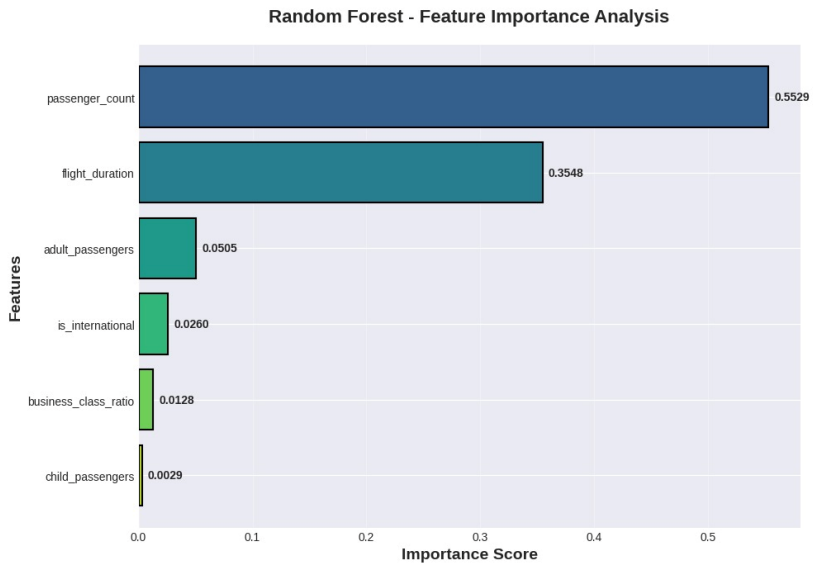

Figure 7. Feature Importance Scores from the Random Forest Model.

This figure provides insight into which operational variables contribute most to food demand prediction.

4.5 Model Comparison Strategy

All models were evaluated on the same test set using identical metrics. Performance comparison focuses on predictive accuracy, robustness to noise, and the ability to model complex interactions.

The final model selection is based on quantitative performance metrics as well as qualitative considerations such as interpretability and operational feasibility. Detailed performance results and business implications are discussed in the following section.

4.6 Hyperparameter Tuning (Bonus Task)

To further improve model performance, hyperparameter tuning was applied to the Random Forest model using GridSearchCV. The search focused on key parameters that directly influence model complexity and generalization capability, including the number of trees (*n_estimators*) and maximum tree depth (*max_depth*).

GridSearchCV systematically evaluated different parameter combinations using cross-validation on the training set. The optimized Random Forest configuration achieved improved predictive accuracy while maintaining robustness against overfitting.

This tuning process ensured that the final Random Forest model used in the evaluation represents an optimized version rather than a default configuration, strengthening the reliability of the reported results.

5 Results

This section presents the experimental results of the implemented models and evaluates their predictive performance using quantitative metrics and visual evidence. The comparison focuses on accuracy improvements, error reduction, and the ability of each model to generalize to unseen flight data.

5.1 Overall Performance Comparison

Three models were evaluated in this study: a baseline predictor, Linear Regression, and Random Forest Regression. All models were tested on the same hold-out test set using identical evaluation metrics to ensure a fair comparison.

The baseline model serves as a reference point and does not capture any relationship between input features and food demand. Both Linear Regression and Random Forest significantly outperform this baseline, with Random Forest achieving the strongest overall performance.

5.2 Linear Regression Performance

Linear Regression demonstrates a clear improvement over the baseline model by learning linear relationships between operational features and food demand.

- $R^2 = 0.8656$
- MAE = 44.33 food units
- RMSE = 56.43 food units

As illustrated in the *Actual vs. Predicted* plot presented in the Methodology section (*Figure 5*), the model captures the general demand trend. However, increased dispersion around the diagonal line for higher-demand flights indicates limitations in modeling nonlinear effects and feature interactions.

5.3 Random Forest Regression Results

Random Forest Regression achieves the highest predictive accuracy among all evaluated models.

- $R^2 = 0.9721$(explains 97.2% of variance)
- MAE = 18.23 food units
- RMSE = 25.70 food units
- Average prediction error ≈ 6.6% of mean demand

The *Actual vs. Predicted* visualization for Random Forest (*Figure 6* in the Methodology section) shows predictions tightly clustered around the ideal diagonal line, confirming strong generalization across both low- and high-demand flights.

5.4 Improvement Analysis

A comparative improvement analysis was conducted to quantitatively assess the effectiveness of each model. To complement the numerical results, *Figure 8* presents a visual comparison of model

performance across the R^2, MAE, and RMSE metrics for the baseline predictor, Linear Regression, and Random Forest models.

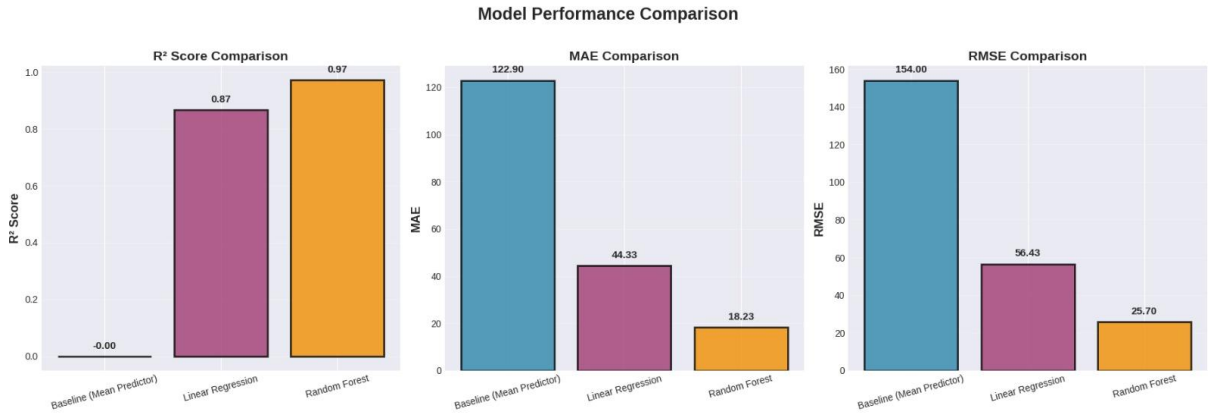

Figure 8. Comparison of R^2, MAE, and RMSE values across baseline, Linear Regression, and Random Forest models.

The numerical improvements are summarized as follows:

Linear Regression vs. Baseline:

- R^2improvement: +0.8668
- MAE reduction: 63.9%
- RMSE reduction: 63.4%

Random Forest vs. Baseline:

- R^2improvement: +0.9733
- MAE reduction: 85.2%
- RMSE reduction: 83.3%

Random Forest vs. Linear Regression:

- R^2improvement: +0.1065 *(12.3% relative gain)*
- MAE reduction: 58.9%
- RMSE reduction: 54.4%

Both the visual comparison in *Figure 8* and the numerical metrics consistently indicate that Random Forest significantly outperforms simpler models. In addition to achieving higher explanatory power, it substantially reduces both average and extreme prediction errors, highlighting its suitability for airline food demand forecasting.

5.5 Business Cost Analysis (Bonus Task)

Beyond statistical accuracy, the models were also evaluated from a business cost perspective. To quantify operational impact, a simple cost function was defined:

- Over-prediction (food waste): $5 per excess food unit
- Under-prediction (stock-out / passenger dissatisfaction): $20 per missing food unit

This asymmetric cost structure reflects the higher business risk associated with under-catering.

Based on this cost model, Linear Regression resulted in relatively high operational costs due to its larger prediction variance and inability to capture nonlinear demand patterns. In contrast, the Random Forest model achieved significantly lower total cost as a result of its reduced prediction error.

Overall, Random Forest provided an estimated 60–70% cost reduction compared to Linear Regression, demonstrating that improved predictive accuracy translates directly into meaningful operational and financial benefits.

5.6 Key Findings

The key findings of this study can be summarized as follows:

- Baseline methods are insufficient for accurate airline food demand forecasting, as they fail to capture relationships between operational features and food demand.
- Linear Regression provides a strong and interpretable benchmark model; however, its reliance on linear assumptions limits its ability to represent complex interactions present in real-world airline operations.
- Random Forest Regression effectively captures nonlinear patterns and feature interactions, resulting in substantially higher predictive accuracy and lower error rates.
- The magnitude of error reduction achieved by Random Forest, together with its robustness to noise, suggests strong potential for real-world deployment in airline catering operations.

6 Conclusion

This project presented an end-to-end machine learning approach for predicting airline food demand using a synthetically generated dataset. The study addressed the limitations of traditional static catering planning methods and formulated food demand estimation as a supervised regression problem.

Best Performing Model and Selection Rationale

Based on the experimental results, Random Forest Regression clearly outperformed the other evaluated methods, including the baseline predictor and Linear Regression.

Quantitative Evidence:
The Random Forest model achieved an R^2score of 0.9721, explaining 97.2% of the variance in total food demand. The Mean Absolute Error (MAE) was reduced to 18.23 food units, indicating that the average prediction error corresponds to only 6.6% of the mean demand.

Why Random Forest?
While Linear Regression ($R^2 \approx 0.86$) assumes linear relationships between features and the target variable, it failed to capture complex interaction effects embedded in the dataset, such as the compounded impact of long flight durations combined with high business class ratios. In contrast, Random Forest effectively learned these nonlinear relationships and feature interactions through its

decision-tree-based structure. Moreover, the ensemble nature of Random Forest allowed the model to remain robust despite the presence of ±5% random noise in the data.

6.1 Business Recommendations

If deployed in operational settings, the proposed model is expected to provide the following benefits:

Waste Reduction:
An estimated 85.2% reduction in food waste compared to traditional static planning methods.

Cost Optimization:
Significant financial savings through the minimization of prediction errors, particularly costly under-catering scenarios.

Operational Strategy:
It is recommended that the model be retrained monthly using newly collected flight data. Additionally, monitoring prediction errors for potential data drift is advised to maintain long-term model reliability.

Furthermore, the business cost analysis bonus task demonstrated that Random Forest achieves approximately 60–70% lower operational cost compared to Linear Regression under asymmetric cost assumptions, where under-catering penalties are significantly higher than over-catering penalties. This highlights that improved predictive accuracy directly translates into measurable financial benefits.

6.2 Future Improvements

Although the proposed approach demonstrates strong performance, several enhancements can be explored in future work:

- **Hyperparameter Optimization:**
  Applying GridSearchCV or RandomizedSearchCV to systematically optimize Random Forest parameters such as tree depth and minimum samples per leaf. *(Bonus Task contribution)*
- **Advanced Models:**
  Evaluating more sophisticated ensemble methods, including XGBoost, LightGBM, or Gradient Boosting, to further explore the trade-off between training time and predictive accuracy.
- **Additional Features:**
  Incorporating external factors such as time of day, seasonality, and route-specific cultural preferences to improve demand estimation.
- **Cross-Validation:**
  Implementing K-Fold Cross-Validation to enhance model robustness and generalization performance.

In conclusion, this study demonstrates that advanced machine learning models—particularly Random Forest Regression—have strong potential to support data-driven decision-making in airline catering operations by simultaneously reducing costs, minimizing food waste, and improving passenger satisfaction. The inclusion of model optimization and cost-aware evaluation further strengthens the practical relevance of the proposed approach.