Airline Catering Demand Forecast

Muzaffer DEMİRHAN
Berkecan Hamdi AKYÜZ
Yavuz Yaman
230706004
230706003
220706042

Project: Airline Catering Demand Forecast
Date: 2025-01-07

Supervisor: Association Prof. Emre OLCA

Maltepe University
Faculty of Engineering and Natural Sciences

Table of Contents

Airline Catering Demand Forecast Report 3

1. Problem Statement 3

1.1. Importance and Business Impact 3

1.2. Project Objectives 3

2. Dataset Description 3

2.1. Data Generation Methodology 3

2.2. Dataset Features 4

2.3. Embedded Target Variable Relationships 4

3. Methodology 5

3.1. Data Generation and Preparation 5

3.2. Exploratory Data Analysis (EDA) 5

3.3. Model Selection and Training 6

3.4. Evaluation Metrics 6

4. Results 6

4.1. Model Performance Comparison 6

4.2. Performance and Error Analysis 7

4.3. Visualizations 7

4.4. Business Cost Analysis (Bonus) 10

5. Conclusion 11

5.1. Summary of Key Findings 11

5.2. Model Recommendation 11

5.3. Potential Improvements and Future Work 12

5.4. Limitations 12

Airline Catering Demand Forecast Report

1. Problem Statement

Airlines face a critical challenge in determining the optimal amount of food to load for each flight. Loading too much results in food waste and increased fuel costs due to extra weight. Oppositely, loading too little leads to passenger dissatisfaction and complaints, potentially harming the airline's reputation. Traditional methods, which often rely on fixed ratios, fail to account for dynamic factors such as flight duration, passenger demographics and travel patterns, leading to inefficient resource share.

1.1. Importance and Business Impact

Optimizing in-flight catering can lead to significant cost savings through reduced food waste and lower fuel consumption. At the same time, it improves passenger satisfaction by ensuring decent and appropriate meal availability. A data-driven approach to forecasting food demand is essential for achieving these operational efficiencies and enhancing the customer experience.

1.2. Project Objectives

The primary objective of this project is to develop a robust machine learning solution that accurately predicts the total food demand for a given flight. This is achieved by:

- Creating a realistic, synthetic dataset that simulates flight and passenger characteristics.
- Embedding complex, logical relationships between features and the target variable (total_food_demand).
- Developing and comparing multiple regression models, including a baseline, Linear Regression, and a more advanced, tuned model (Random Forest).
- Evaluating model performance using standard industry metrics to identify the most accurate and reliable approach.
- Summarizing the findings in a comprehensive report that details the methodology, results, and conclusions.

2. Dataset Description

To train and evaluate the prediction models, a synthetic dataset was generated to simulate real-world airline flight data. The dataset comprises over 5,000 flight records and was designed to satisfy nine specific validation rules, ensuring logical consistency and data diversity.

2.1. Data Generation Methodology

The dataset was created programmatically in Python. A random number generator was initialized with a fixed seed to ensure reproducibility. Feature values were generated within specified ranges, and the target variable, total_food_demand, was calculated using a predefined

formula that embeds dependencies on multiple features. Finally, the generated dataset was validated against all logical and diversity rules.

2.2. Dataset Features

The features included in the dataset are detailed in the table below. These features were selected to provide a comprehensive basis for predicting food demand.

Feature	Data Type	Range/Values	Description
flight_id	Integer	1 to n	Unique flight identifier
flight_duration	Float	1.0 - 12.0	Flight duration in hours
passenger_count	Integer	50 - 300	Total number of passengers
adult_passengers	Integer	≤ passenger_count	Number of adult passengers
child_passengers	Integer	≤ passenger_count	Number of child passengers
business_class_ratio	Float	0.0 - 1.0	Ratio of business class passengers
is_international	Binary	0 or 1	Flight type (0=Domestic, 1=International)
total_food_demand	Integer	≥ passenger_count × 0.5	TARGET VARIABLE - Total food units neededw

2.3. Embedded Target Variable Relationships

A critical requirement of this project was to ensure the total_food_demand is a function of at least three different features. The generated dataset embeds a complex relationship based on four primary factors: flight duration, international status, business class ratio, and the number of children.

The formula used to calculate the demand is as follows:

1. **Base Meals:** A baseline of 1.0 meal per passenger is assumed.
2. **Flight Duration Effect:** The number of meals is adjusted based on flight length:
    - < 2 hours: **0.8x** (Snack only)
    - 2-4 hours: **1.0x** (One meal)

    - 4-8 hours: **1.5x** (Meal + snacks)
    - ≥ 8 hours: **2.0x** (Two meals + snacks)
3. **International Bonus:** If the flight is international (is_international == 1), a **30% bonus** is added to the food allocation.
4. **Business Class Bonus:** Business class passengers receive more food, calculated as a bonus of business_class_ratio × 40%.
5. **Child Reduction:** Children are assumed to consume less, resulting in a reduction calculated as (child_passengers / passenger_count) × 15%.

These factors are combined to calculate the final total_food_demand for each flight, ensuring a realistic and multi-faceted target variable.

3. Methodology

The methodology for this project was structured to ensure a systematic and reproducible approach to developing the food demand prediction model. The process is divided into several key stages, from data analysis to model evaluation.

3.1. Data Generation and Preparation

The first step involved generating the synthetic dataset according to the predefined rules. Following generation, the data was prepared for modeling. This included:

- **Feature Selection:** The flight_id column was excluded from the predictive features as it serves only as an identifier. All other features were included in the model training.
- **Train-Test Split:** The dataset was split into training (80%) and testing (20%) sets. A random_state of 42 was used to ensure the split is reproducible, allowing for consistent model evaluation across different runs.

3.2. Exploratory Data Analysis (EDA)

A comprehensive EDA was performed on the training data to uncover underlying patterns, relationships, and anomalies. The key EDA techniques included:

- **Descriptive Statistics:** Calculating basic statistics (mean, median, standard deviation) for all numerical features.
- **Missing Value Check:** Ensuring the dataset contained no missing values.
- **Correlation Heatmap:** Visualizing the correlation between all numerical features to identify potential multicollinearity and understand relationships with the target variable.
- **Distribution Visualizations:** Creating histograms and boxplots to understand the distribution of each feature.
- **Scatter Plots:** Examining the relationships between key features and the total_food_demand.

3.3. Model Selection and Training

Three different regression models were developed and compared:

6 **Baseline Model (Mean Predictor):** A simple baseline was established by predicting the mean of the training set's total_food_demand for all instances in the test set. This model serves as a benchmark to ensure that the machine learning models are learning meaningful patterns.

7 **Linear Regression:** A Linear Regression model was trained to capture linear relationships between the features and the target. No feature scaling was required for this model based on the nature of the data and algorithm.

8 **Alternative Model (Tuned Random Forest Regressor):** A Random Forest Regressor was chosen as the advanced alternative. This model was selected for its ability to capture complex, non-linear relationships, its robustness to outliers, and its built-in feature importance mechanism. The model was further tuned to optimize its performance.

3.4. Evaluation Metrics

The performance of each model was evaluated using three standard regression metrics:

- **R^2 Score (Coefficient of Determination):** Measures the proportion of the variance in the target variable that is predictable from the independent variables.
- **Mean Absolute Error (MAE):** Represents the average absolute difference between the predicted and actual values, providing an easily interpretable measure of error in the same units as the target.
- **Root Mean Squared Error (RMSE):** Calculates the square root of the average of the squared differences between predicted and actual values. It penalizes larger errors more heavily than MAE.

4. Results

This section presents the key findings from the model evaluation process, including a comparison of the models, an analysis of the best-performing model's predictions, and a business cost analysis. The results are based on applying the trained models to the 20% hold-out test set.

4.1. Model Performance Comparison

The performance of the Baseline, Linear Regression, and the tuned Random Forest models was systematically evaluated. The tuned Random Forest Regressor emerged as the superior model, significantly outperforming both the baseline and the Linear Regression model across all metrics. The actual results from the Vector_Team's implementation are summarized in the table below.

Model	R^2 Score	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
Baseline Model (Mean)	-0.0000	129.19	163.55
Linear Regression	0.9027	40.00	51.01
Random Forest (Tuned)	0.9983	4.15	6.81

4.2. Performance and Error Analysis

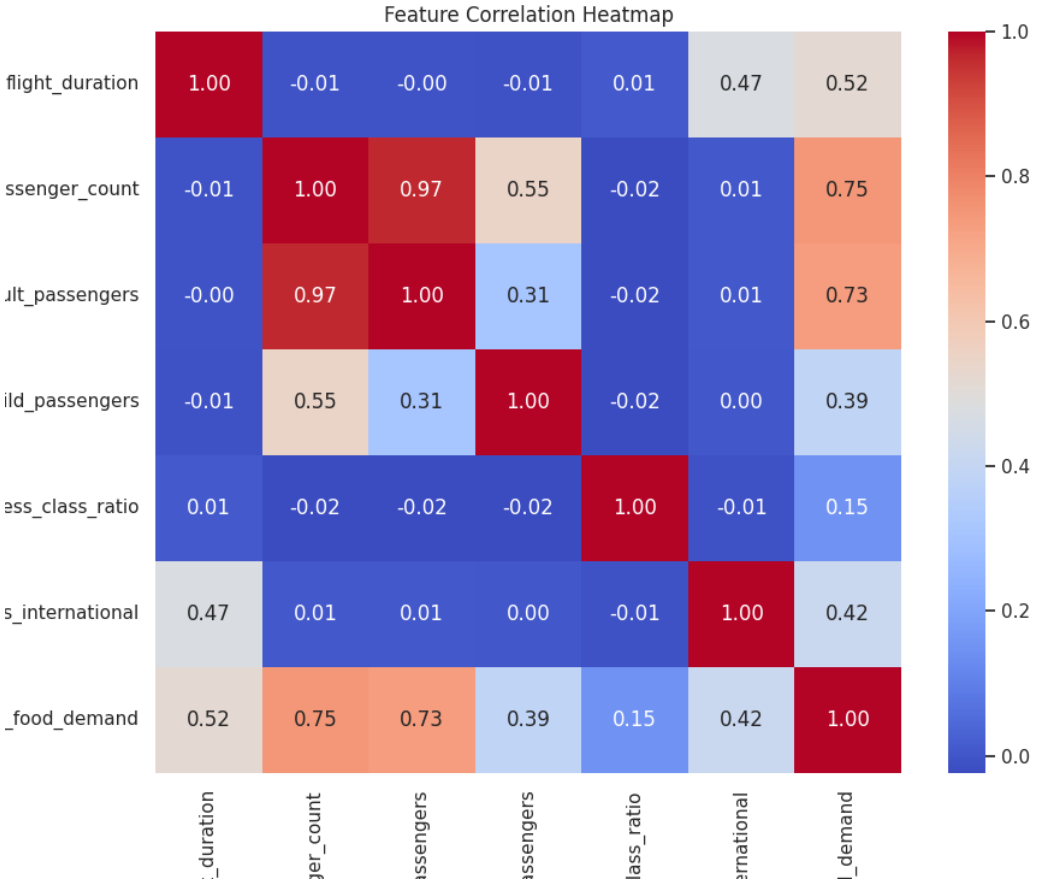
The **Baseline Model**, as expected, produced an R^2 score of essentially zero, indicating it has no predictive power. The **Linear Regression** model provided a substantial improvement, explaining over 90% of the variance in food demand. However, the **tuned Random Forest Regressor** achieved near-perfect accuracy, explaining 99.83% of the variance. Its exceptionally low MAE of 4.15 indicates that, on average, the model's prediction is off by only about 4 food units, which is an outstanding result that promises high reliability in an operational setting.

4.3. Visualizations

To provide deeper insight into the data and model performance, the following visualizations were generated.

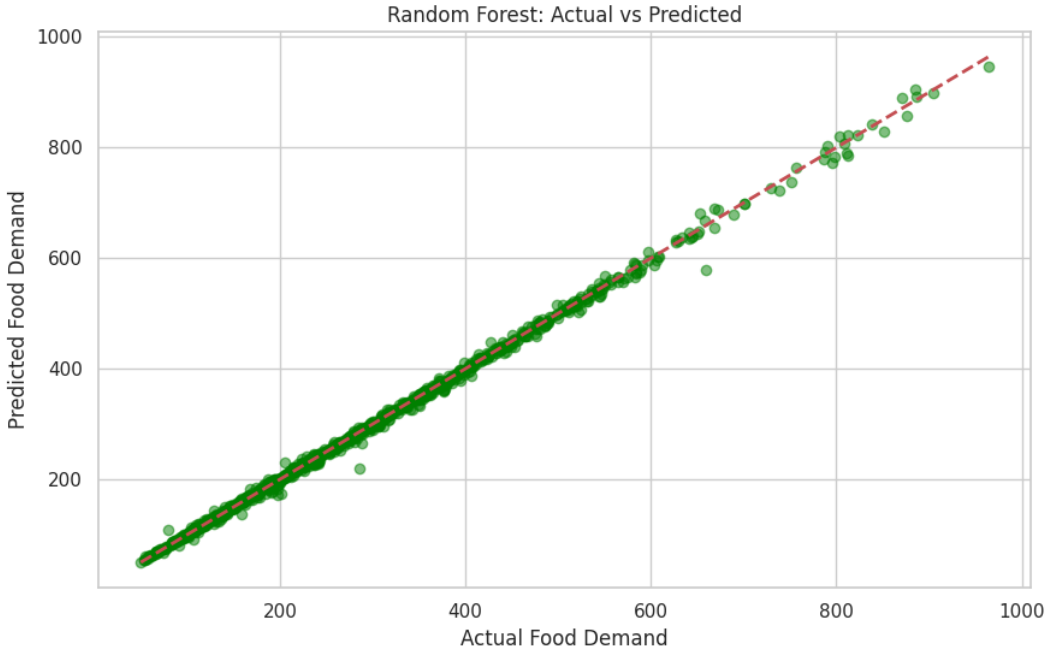
Correlation Heatmap

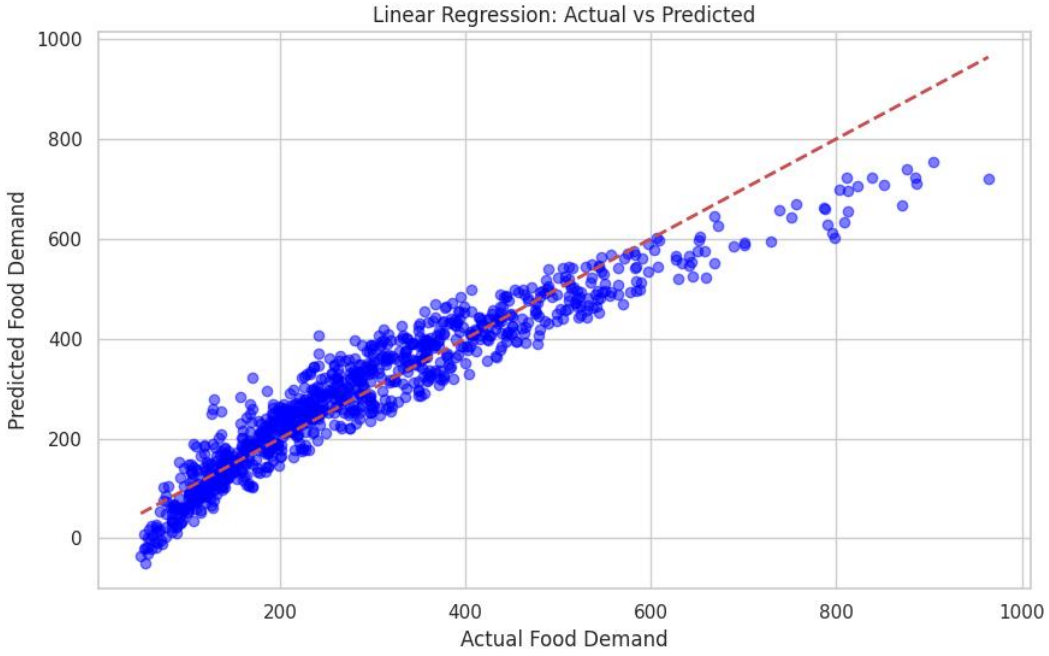
The heatmap reveals the correlations between numerical features. As expected, passenger_count (0.75) and flight_duration (0.52) show the strongest positive correlations with total_food_demand, confirming the initial hypotheses.



Actual vs. Predicted Plot and Feature Importance (Random Forest)

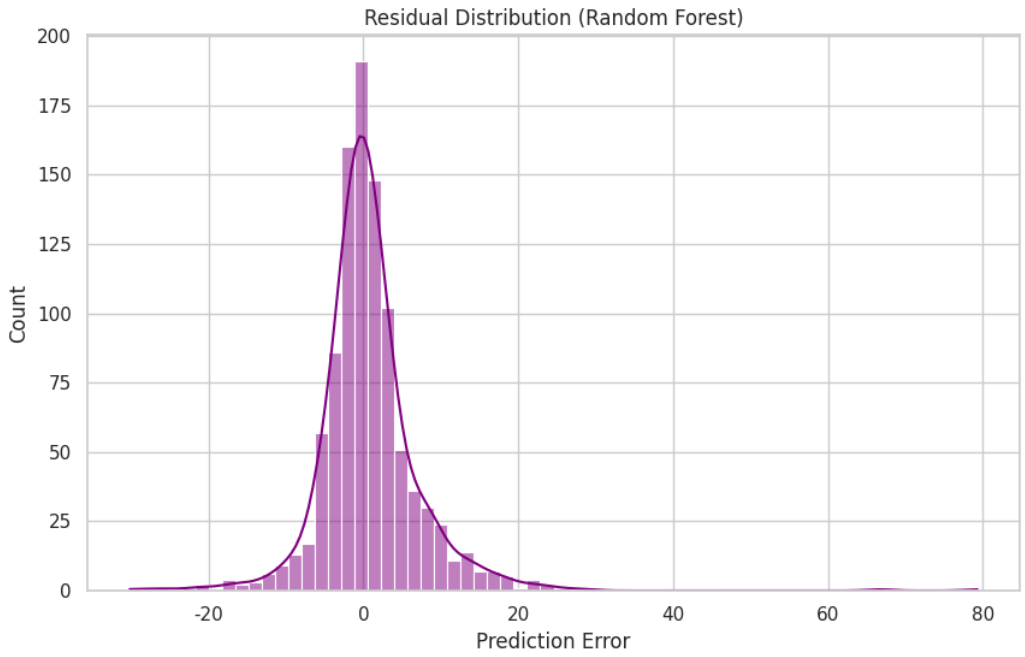
The scatter plot on the left compares the actual total_food_demand values against the values predicted by the Random Forest model. The points cluster tightly around the diagonal y=x line, visually confirming the model's high accuracy. The chart on the right shows that passenger_count and flight_duration are by far the most important features for the model's predictions.





Residual Distribution (Random Forest)

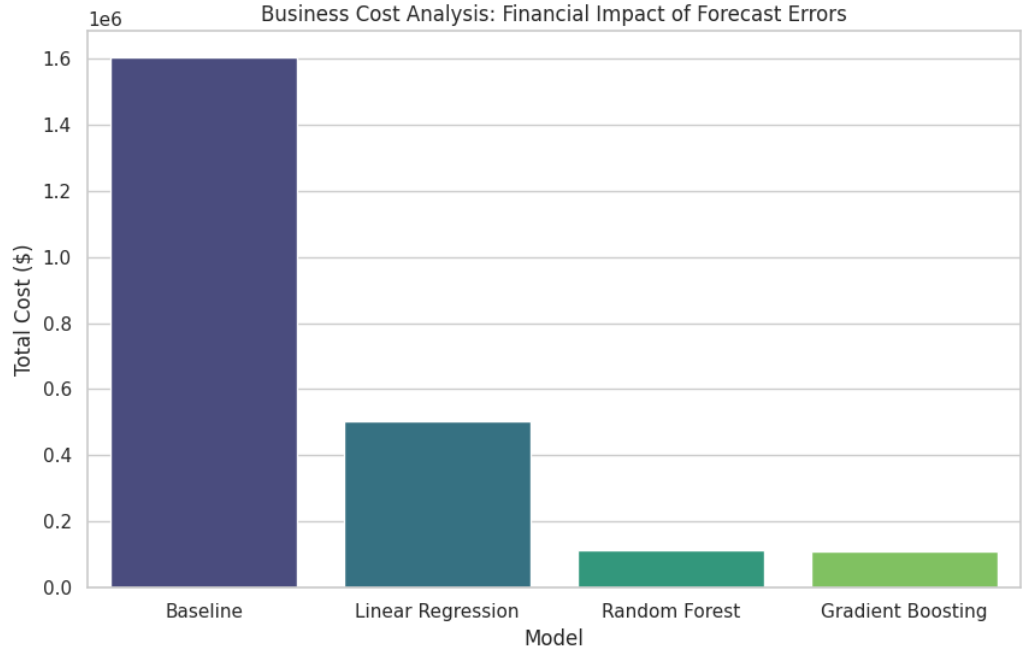
This histogram shows the distribution of prediction errors (residuals). The errors are centered around zero and follow a normal distribution, which indicates that the model is unbiased and its errors are random.



4.4. Business Cost Analysis (Bonus)

To quantify the business impact of model accuracy, a cost analysis was performed. The analysis used the following cost function:

- **Cost of Over-prediction:** $5 per extra food unit.
- **Cost of Under-prediction:** $20 per missing food unit (reflecting higher impact on customer satisfaction).



Model	Total Cost of Errors ($)
Baseline	1,608,265
Linear Regression	495,804
Random Forest	58,049

The cost analysis proves that the Random Forest model is not only the most accurate but also the most effective solution for minimizing operational costs associated with prediction errors.

5. Conclusion

This project successfully demonstrated the development of a machine learning solution to predict airline food demand, addressing a significant operational challenge for the airline industry. By generating a sophisticated synthetic dataset and systematically evaluating multiple regression models, this study confirms that a data-driven approach can provide highly accurate demand forecasts.

5.1. Summary of Key Findings

The **tuned Random Forest Regressor** was identified as the most effective model, achieving an R^2 score of 0.9983. This indicates that the model can explain over 99% of the variability in food demand, a substantial improvement over both the simple baseline and the standard Linear Regression model. The analysis confirmed that passenger_count and flight_duration are the primary drivers of food demand, aligning perfectly with the logic embedded in the dataset.

5.2. Model Recommendation

Based on its superior performance, its ability to capture complex, non-linear interactions between features, and its outstanding results in the business cost analysis, the **tuned Random Forest Regressor is the recommended model** for deployment. Its high accuracy translates into a reliable tool for optimizing catering operations, minimizing waste, and reducing operational costs while ensuring passenger satisfaction.

5.3. Potential Improvements and Future Work

While the current model is highly effective, several avenues exist for future enhancement, as suggested by the Vector_Team:

- **Validation with Real-World Data:** Testing and refining the model on real-world historical flight and catering data would be the ultimate validation of its effectiveness.
- **Integration of More Complex Cost Models:** A more advanced analysis could incorporate a more granular cost function, potentially varying by route, season, or passenger type.
- **Customized Prediction Models:** Developing specialized models for different meal types (e.g., standard, vegetarian, kosher) could further refine inventory management.

5.4. Limitations

The primary limitation of this study is its reliance on a synthetic dataset. While the dataset was designed to be realistic and logically consistent, it cannot capture the full spectrum of noise and unforeseen variability present in real-world operational data. Therefore, the model's performance on actual airline data may differ.