SE 390 01 - Artificial Intelligence Projects with Python

FINAL PROJECT EXAM

Fall 2025-2026

Airline Food Demand Prediction

1. Problem Description

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. Loading too little leads to passenger dissatisfaction and complaints. Traditional methods rely on fixed ratios that fail to account for dynamic factors such as flight duration, passenger demographics, and travel patterns.

Your task is to develop a machine learning solution that predicts the total food demand for a given flight based on various flight and passenger characteristics. This prediction will help airlines optimize their catering operations, reduce waste, and improve customer satisfaction.

Business Impact: Optimizing in-flight catering can lead to significant cost savings and reduced food waste, while simultaneously improving passenger satisfaction.

2. Dataset Requirements

You will create a synthetic dataset that simulates real airline flight data. The dataset must contain at least 5,000 flight records with the following features:

Definition: One "food unit" represents one meal or snack package prepared for a passenger.

Feature	Type	Description
flight_id	Integer	Unique flight identifier
flight_duration	Float	Flight duration in hours (1-12)
passenger_count	Integer	Total number of passengers (50-300)
adult_passengers	Integer	Number of adult passengers
child_passengers	Integer	Number of child passengers
business_class_ratio	Float	Ratio of business class passengers (0-1)
is_international	Binary (0/1)	Whether the flight is international
total_food_demand	**Integer**	**TARGET VARIABLE - Total food units needed**

⚠ **Important:** flight_id is an identifier only. It must NOT be used as a predictive feature in your models.

2.1 Data Validation Rules

Your synthetic dataset must satisfy all of the following logical constraints. Datasets that violate these rules will receive point deductions:

#	Validation Rule
1	adult_passengers + child_passengers == passenger_count
2	0 <= business_class_ratio <= 1
3	1 <= flight_duration <= 12 (realistic flight duration range)
4	if is_international == 1 then flight_duration >= 3
5	50 <= passenger_count <= 300
6	total_food_demand >= passenger_count * 0.5
7	Dataset must contain at least 5,000 rows
8	is_international == 1 for at least 15% of flights (data diversity)
9	flight_duration must include both short (1-3h) and long (8-12h) flights

2.2 Target Variable Requirements

Critical Requirement: Your total_food_demand must depend on at least 3 different features (e.g., passenger_count, flight_duration, and is_international or business_class_ratio). Simply multiplying passenger_count by a constant is not acceptable. Your report must explain what relationships you embedded in the target variable.

3. Project Tasks

Task 1: Exploratory Data Analysis (20 points)

Perform a comprehensive analysis of your dataset to understand the underlying patterns and relationships:

- Display basic statistics using descriptive methods
- Check for missing values and handle them appropriately
- Create a correlation heatmap to identify feature relationships
- Visualize distributions using histograms and boxplots
- Create scatter plots to examine relationships with the target variable

Task 2: Baseline Model (10 points)

Before building any machine learning model, establish a baseline for comparison. The baseline helps you understand whether your ML models are actually learning meaningful patterns.

Baseline Strategy: Use the mean of the training set as the prediction for all test samples (Mean Predictor). Calculate R^2, MAE, and RMSE for this baseline.

Why is this important? If your ML model cannot beat this simple baseline, it indicates that the model has not learned useful patterns from the data.

Task 3: Linear Regression Model (15 points)

Implement a Linear Regression model as your first machine learning approach:

1. Split your data into training (80%) and testing (20%) sets
2. Train a Linear Regression model on the training data
3. Make predictions on the test set
4. Calculate performance metrics: R^2, MAE, RMSE
5. Create a scatter plot comparing actual vs. predicted values

Task 4: Alternative Model of Your Choice (30 points)

Select and implement a second regression model to compare against Linear Regression. You may choose from models such as:

- Random Forest Regressor
- Gradient Boosting Regressor
- XGBoost
- Support Vector Regressor
- Decision Tree Regressor

Required: You must justify your model choice. Explain why you selected this particular algorithm and what advantages it might offer for this problem.

If you choose a tree-based model, include a feature importance analysis showing which features contribute most to the predictions.

Task 5: Model Comparison & Error Analysis (10 points)

Create a comprehensive comparison of all three approaches (Baseline, Linear Regression, and your chosen model):

- Present a comparison table with R^2, MAE, and RMSE for each model
- Discuss which model performs best and explain why
- Analyze the trade-offs between model complexity and performance
- **Include a residual plot or error histogram** showing the distribution of prediction errors for your best model

Task 6: Written Report (15 points)

Submit a separate PDF report that summarizes your work. The report must include:

6. **Problem Statement:** Brief description of the problem and its importance
7. **Dataset Description:** How you created the data and what relationships you embedded (minimum 3 features affecting target)
8. **Methodology:** Your approach to solving the problem
9. **Results:** Key findings and model performance
10. **Conclusion:** Summary and potential improvements

Report Format Requirements

Requirement	Specification
Length	3-5 pages (excluding cover page)
Visuals	Minimum 3 figures (heatmap, actual vs. predicted plot, residual plot or feature importance)
Tables	Minimum 1 table (model performance comparison)
Format	PDF, 11-12pt font, 1.15 line spacing

4. Grading Rubric

Component	Points	Key Criteria
Exploratory Data Analysis	**20**	Thorough analysis, clear visualizations
Baseline Model	**10**	Correct implementation
Linear Regression Model	**15**	Proper implementation, metrics
Alternative Model + Justification	**30**	Model choice justified
Model Comparison & Error Analysis	**10**	Table, residual analysis
Written Report	**15**	Format compliance, clarity
TOTAL	**100**	

Bonus Points (up to +15)

Bonus Task	Points
Hyperparameter tuning (GridSearchCV or RandomizedSearchCV)	**+3**

Business cost analysis: Calculate financial impact using Cost = (OverPredictions × $5) + (UnderPredictions × $20)	**+2**
Implementation of a third model with comparison	**+10**

5. Submission Requirements

5.1 Deliverables

11. **Jupyter Notebook (.ipynb):** Complete code with markdown explanations
12. **Dataset (.csv):** Your generated synthetic dataset
13. **Report (.pdf):** Written summary report

5.2 File Naming Convention

StudentID_Name_Surname.ipynb or GroupName.ipynb

StudentID_Name_Surname_dataset.csv or GroupName_dataset.csv

StudentID_Name_Surname_report.pdf or GroupName_report.pdf

5.3 Important Notes

- **Code Comments:** Include clear comments explaining your code
- **Visualizations:** All plots must have titles and axis labels

⚠ CRITICAL: Plagiarism will result in zero points!!!

Good luck!