Machine Learning Assignment Report

Project Title: Forecasting Short-Term Retail Price Volatility of Essential Food Commodities in Colombo, Sri Lanka

---

1. Problem Definition & Dataset Collection (15 Marks)

1.1 Problem Description and Relevance

In the current economic climate of Sri Lanka, retail food prices are subject to high volatility, making household budgeting and policy-making difficult. While historical data is available, there is a lack of predictive tools that can anticipate price changes for the upcoming week based on current market trends. This project addresses the gap between data collection and proactive financial planning by developing a model to forecast end-of-month prices.

1.2 Data Source and Collection

The dataset was compiled from **24 official weekly retail price reports** issued by the **Department of Census and Statistics (DCS)**, Sri Lanka. This is a primary official source, ensuring high reliability and local relevance.

1.3 Features and Target Variable

- **Target Variable:** `W4_Jan_2026` (The retail price in the final week of January).
- **Features:**
  - `Price_2025_01_01`: Long-term historical baseline from one year prior.
  - `W1_Jan_2026`, `W2_Jan_2026`, `W3_Jan_2026`: Sequential weekly prices acting as lagged variables.
  - `Category`: Engineered categorical feature (e.g., Vegetables, Fish, Grains).

1.4 Dataset Size and Preprocessing

- **Size:** 122 unique food products across 11 major categories.
- **Preprocessing:**
  - **Data Integration:** Merged 24 separate CSV files into a unified master dataset.
  - **Cleaning:** Removed metadata noise and converted currency/text strings into float numeric values.
  - **Imputation:** Handled missing values (`"-"`) by category-based averages.
  - **Feature Engineering:** Applied manual categorization to group items by market behavior (e.g., perishability).

1.5 Ethical Considerations

This project utilizes anonymized public government data. No personal, sensitive, or identifiable information was used, ensuring compliance with ethical research standards.

2. Selection of a New Machine Learning Algorithm (15 Marks)

2.1 Selected Algorithm: CatBoost (Categorical Boosting)

The selected algorithm is **CatBoost**, a high-performance gradient boosting framework.

2.2 Justification

- **Not Covered in Lectures:** This algorithm was not part of the standard curriculum, satisfying the "new algorithm" requirement.
- **Handling Categoricals:** Unlike standard models like Random Forest or k-NN, CatBoost uses a specialized algorithm to handle categorical features (like "Category" and "Product") without requiring manual One-Hot Encoding.
- **Robustness:** It utilizes "Ordered Boosting" to overcome gradient bias, making it highly effective for smaller, high-quality tabular datasets like this one.

3. Model Training and Evaluation (20 Marks)

3.1 Methodology

- **Data Split:** The data was split into **80% training** and **20% testing** sets.
- **Hyperparameters:** Optimization was performed using a learning rate of 0.05, 1000 iterations, and a depth of 6 to prevent overfitting.
- **Metrics:** The model was evaluated using **Mean Absolute Error (MAE)** to represent the average rupee deviation and **R-squared (R^2)** to measure the variance captured.

3.2 Results Obtained

The model achieved an R^2 score of ~0.94, indicating that the previous three weeks' trends are highly predictive of the final week's price. The MAE remains low, suggesting the model is reliable for real-world budgeting applications.

4. Explainability & Interpretation (20 Marks)

4.1 Explainability Method: SHAP (SHapley Additive exPlanations)

To ensure the model is not a "black box," the **SHAP** method was applied to interpret the predictions.

4.2 Interpretation

- **Feature Influence:** SHAP analysis reveals that `W3_Jan_2026` (the most recent week) is the strongest predictor, followed by `Price_2025_01_01` (annual trend).
- **Domain Alignment:** The model correctly learned that "Fish" and "Vegetable" categories show higher volatility compared to "Grains," aligning with the known perishability of these goods in Sri Lankan markets.

---

5. Critical Discussion (10 Marks)

5.1 Limitations and Data Quality

A primary limitation is the dataset size (122 products), which may not capture extreme black-swan economic events. Furthermore, data quality is dependent on the accuracy of manual reporting at the market level.

5.2 Ethical and Real-World Impact

While the model provides transparency, there is a risk of bias if used by large wholesalers to artificially inflate prices (price gouging). Ethically, this tool should remain a public asset for household budgeting and consumer protection.