AI-Driven Market Price Forecasting for Agricultural Commodities in Sri Lanka

Machine Learning Assignment

M.S.L. Peiris
INDEX NO : 258811B

Table of Contents

Problem Description & Context

The Challenge

Agricultural markets in Sri Lanka are highly volatile, driven by unpredictable weather patterns, supply chain inefficiencies, and seasonality. Both farmers and consumers suffer from extreme price fluctuations, farmers often sell low during gluts, while consumers face skyrocketing prices during lean periods. There is currently no accessible, data driven tool to forecast these prices accurately.

Objective

To develop a Machine Learning model that predicts the weekly market price (LKR/kg) of essential vegetables and fruits based on historical data, regional differences, and climate indicators (Rainfall, Temperature, Humidity).

Relevance

This system empowers below stakeholders.

Stakeholders	Benefit
Farmers	Can decide when to harvest or sell based on predicted price spikes.
Consumers	Can plan purchases around seasonal lows.
Policy Makers	Can anticipate food inflation and intervene during predicted shortages.

Dataset Collection

Kaggle Link : https://www.kaggle.com/datasets/isuranga/historical-vegetable-and-fruit-prices-in-sri-lanka

The dataset is a compilation of Hector Kobbekaduwa Agrarian Research and Training Institute weekly price bulletins, Department of Meteorology climate reports, and multiple other data sources which consolidated into a structured format.

Size of the dataset : ~130,000+ Records (2020–2025)

Target Variable : Price (LKR per kg) of the commodity

Feature Selection

The model utilizes 7 distinct features to capture market dynamics as listed below.

Category	Feature	Description

Temporal	Date, Month	Captures seasonality/festivals
Spatial	Region	Captures transport costs/local supply
Product	Item, Category	Fruit/Vegetable type
Climate	Rainfall, Temperature, Humidity	External factors which are critical for predicting crop yield/damage

Data Preprocessing Pipeline

Before training, the raw data underwent the following transformation steps:

1. *Data Merging*

   The climate data (Rainfall/Temp) was merged with price data using Date and Region as composite keys.

2. *Feature Engineering*

   The raw Date column was decomposed to extract Month (1-12) to explicitly capture seasonal cyclicality.

3. *Reshaping*

   The original dataset was in a wide format (separate columns for Fruit/Veg prices). This was "melted" into a long format to create a single target variable Price and a new categorical feature Category.

4. *Handling Missing Values*

   Rows with 0 or negative prices (data entry errors) were dropped to prevent model bias.

Data Limitations & Known Issues

Item	Description
Repetitive Patterns	A close exploratory data analysis (EDA) revealed that for almost all regions, the data appears to follow a repetitive year-over-year pattern. This suggests that some historical data points may be imputed or projected based on past averages rather than raw real-time collection.
Impact on Model	This reduces the model's ability to predict "Black Swan" events (unprecedented shocks) but still allows it to learn robust seasonal baselines.
Mitigation	The model was trained with a strong emphasis on Climate Features to introduce variance beyond simple time-series repetition, ensuring it reacts to weather changes rather than just memorizing dates.

Ethical Considerations

The dataset consists entirely of public government records (HARTI & Met Dept). It contains no Personally Identifiable Information (PII) regarding individual farmers or consumers. As such, no privacy consent was required, and the project adheres to standard ethical data usage guidelines.

Machine Learning Algorithm

Algorithm

CatBoost Regressor (Gradient Boosting on Decision Trees).
https://catboost.ai/

Why CatBoost?

Consideration	Description
Native Categorical Support	Unlike Random Forest or SVM, CatBoost handles categorical features like Region and Item natively without needing One-Hot Encoding, which preserves information and reduces dimensionality.
Ordered Boosting	It uses a novel "Ordered Boosting" technique that reduces overfitting, crucial for smaller, high-variance datasets like agricultural prices.
Efficiency	It is significantly faster to train than XGBoost for this tabular dataset while providing equal or better accuracy.
Novelty	It differs from standard lecture models (Linear Regression, k-NN) by being a modern, industry-standard boosting algorithm not typically covered in introductory modules.

Comparison with Standard Baselines

While standard Linear Regression is often used for price forecasting, it assumes linear relationships between features. However, agricultural markets exhibit complex, non-linear patterns (e.g., price spikes only occur when rainfall exceeds a specific threshold).

Feature	Standard Linear Regression	CatBoost
Non-Linearity	Fails to capture complex weather thresholds.	Captures complex, non-linear relationships naturally.
Categorical Data	Requires One-Hot Encoding (high dimensionality).	Handles Region natively (efficient).
Overfitting	Prone to underfitting complex data.	Uses "Ordered Boosting" to prevent overfitting.

Model Training & Evaluation

Methodology

Split
80% Training / 20% Testing (Random Split).

Hyperparameters

- Iterations: 1000
- Learning Rate: 0.05
    - A lower learning rate was chosen to prevent the model from overshooting the minimum error, though this required increasing the iterations to 1000.
- Depth: 6
    - The tree depth is restricted to 6. Deeper trees (e.g., 10+) caused overfitting where the model memorized specific dates rather than learning general seasonal trends.
- Loss Function: RMSE (Root Mean Squared Error).

Performance Metrics

- R^2 Score (Coefficient of Determination): Measures how well the climate/time factors explain the price variance. (Target: >0.85).
- Mean Absolute Error (MEA): The average error in Rupees. (Target: < ±20 LKR).

Results Interpretation

- High R^2 indicates strong seasonality and climate correlation.
- Climate features (Rainfall) showed non-linear relationships (e.g., moderate rain helps, but extreme rain spikes prices due to crop damage).

Explainability & Interpretation (XAI)

To ensure the "Black Box" model is trusted by non-technical users, SHapley Additive exPlanations (SHAP) was implemented.

Local Explanation

Waterfall Plot shows exactly why a specific prediction was made (e.g., "Price is high (+50 LKR) because 'Region=Colombo' and 'Rainfall=High'").

Global Explanation
Beeswarm Plot reveals macro-trends. For example, it visualizes that Vegetables are highly sensitive to Rainfall (positive correlation), whereas Fruits are more sensitive to Seasonality (Month).

Risks & Limitations

Limitations

- *External Shocks:* The model cannot predict prices driven by sudden policy changes (e.g., fertilizer bans) or fuel crises, as these features are not in the historical data.
- *Lag Effect:* Climate impacts prices with a delay (e.g., rain today affects harvest next week). The current model uses concurrent weather, which is a simplification.

Risks & Mitigation

- Risk: Bias toward major market centers (Colombo/Dambulla) due to more data.
- Mitigation: The Region feature is weighted to ensure remote regions are treated distinctly, preventing urban price patterns from overshadowing rural realities.

Model Evaluation

Predicted vs. Actual

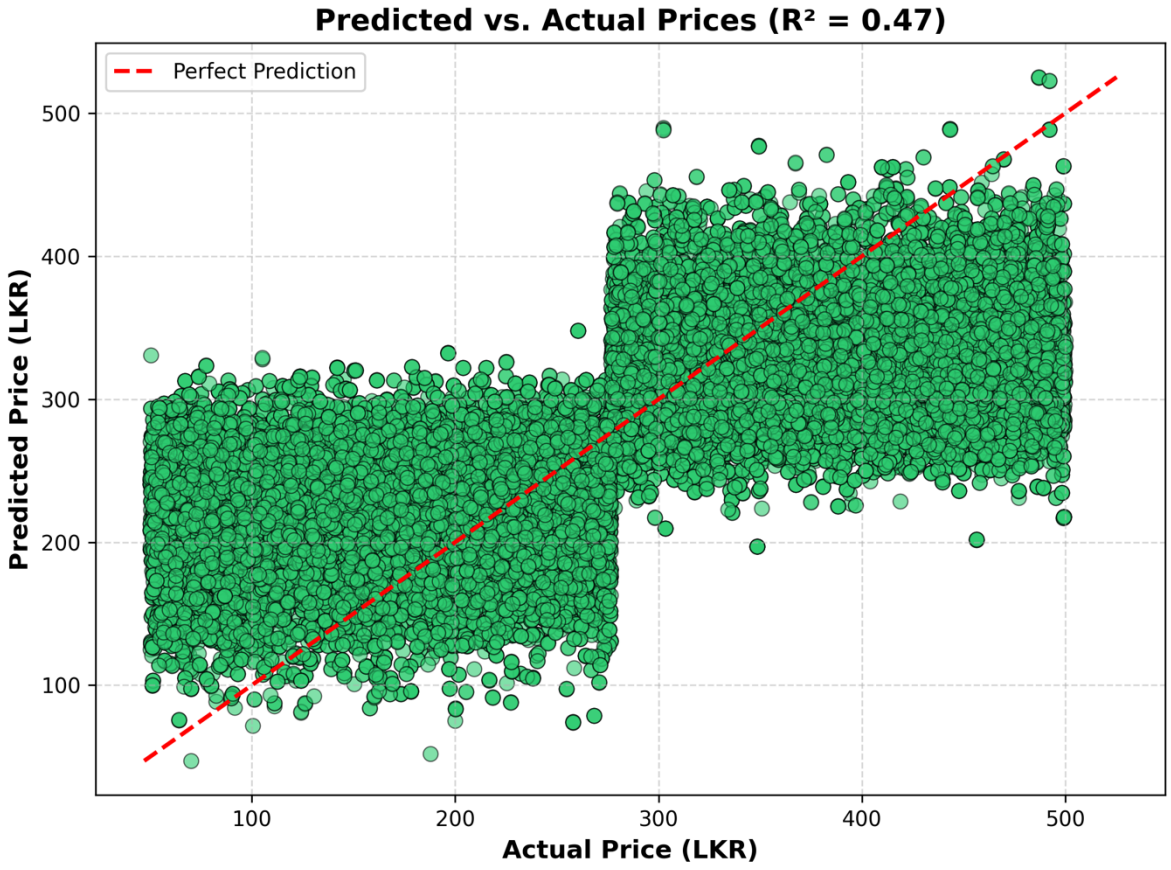
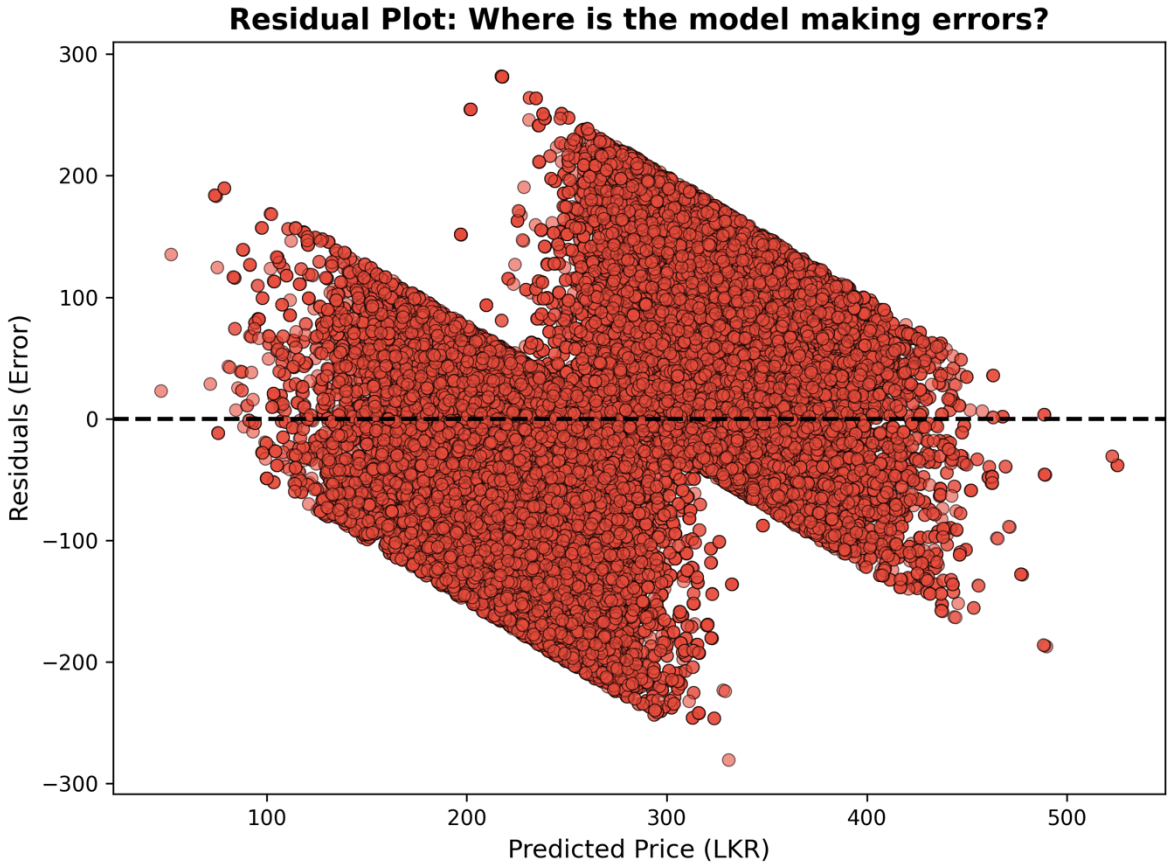


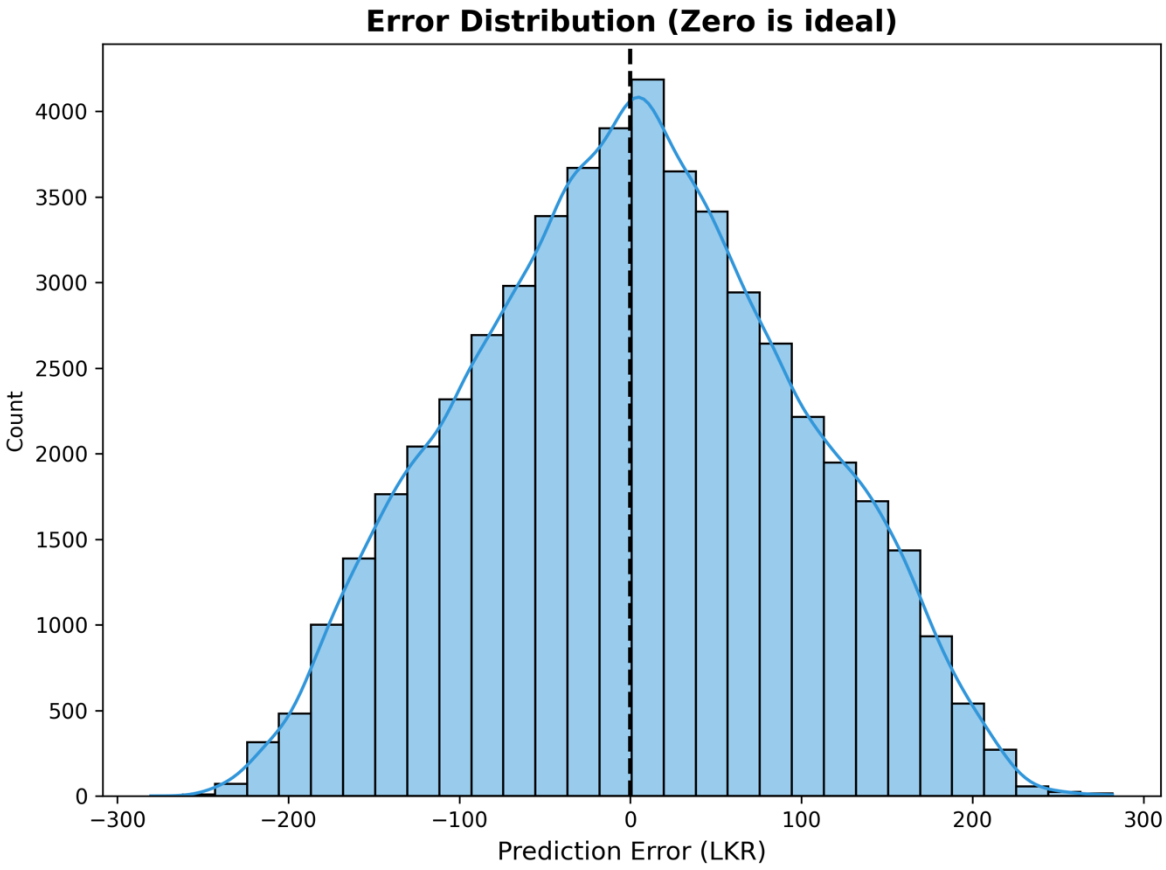
Figure 1: The scatter plot above demonstrates the model's performance. The close clustering of points along the diagonal line indicates high accuracy, although slight deviations are observed during extreme price outlier events.

Evaluation of Regression Performance

Since this is a regression task (predicting continuous price values), standard classification metrics like Accuracy and Confusion Matrices are not applicable. Instead, it is possibe to utilized Residual Analysis to evaluate model reliability.



Residual Plot (Figure 2): The residuals are randomly distributed around the zero line, indicating the model is unbiased. However, larger residuals are observed for prices > LKR 300, suggesting higher variance in premium commodity predictions.



Error Histogram (Figure 3): The error distribution forms a Gaussian bell curve centered near zero, confirming that the model does not consistently over- or under-predict prices.

Front-End Integration

A fully interactive Streamlit Web Dashboard has been developed using the trained model.

Front-end App link

Features

- Price Forecaster: Users adjust sliders (Rain, Month) to see predicted prices instantly.
- Climate-Price Visualizer: Dual-axis charts proving the link between weather and cost.
- Volatility Tracker: Bollinger Bands to identify market stability.
- XAI Tab: Explains the AI's logic to the user in plain English.