MACHINE LEARNING MODEL FOR ANALYZING CLIMATIC AND ECONOMIC INFLUENCES ON VEGETABLE PRICES: FORECASTING CARROT PRICES IN DAMBULLA MARKET

G.G.M.P.KUMARA

Submitted in partial fulfillment of the requirements of the degree of

B.Sc. Honours in Statistics

Department of Statistics and Computer Science

University of Peradeniya

ABSTRACT

This is where you write your abstract ...All the pages in the Project Report/Thesis must be computer printed only on one side of the page using Times New Roman (size 12) font with 1.5-line spacing.

DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

. . . .
G.G.M.P.Kumara
November 22, 2025

. . . .
Prof.Amalka Pinidiyaarachchi
November 22, 2025

ACKNOWLEDGEMENT

And I would like to acknowledge ...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

All the pages in the Project Report/Thesis must be computer printed only on one side of the page using Times New Roman (size 12) font with 1.5-line spacing. However, the following components of the project report/thesis should have single-line spacing: declaration, abstract, acknowledgement, table of contents, list of tables, list of figures, list of abbreviations, table titles, figure captions and references. Each reference must be separated by a single-line spacing. Margins on each page must be maintained as follows: left hand, 40 mm; right hand, 15 mm; top and bottom, 25 mm.

All the pages in the Project Report/Thesis must be computer printed only on one side of the page using Times New Roman (size 12) font with 1.5-line spacing. However, the following components of the project report/thesis should have single-line spacing: declaration, abstract, acknowledgement, table of contents, list of tables, list of figures, list of abbreviations, table titles, figure captions and references. Each reference must be separated by a single-line spacing. Margins on each page must be maintained as follows: left hand, 40 mm; right hand, 15 mm; top and bottom, 25 mm.

All the pages in the Project Report/Thesis must be computer printed only on one side of the page using Times New Roman (size 12) font with 1.5-line spacing. However, the following components of the project report/thesis should have single-line spacing: declaration, abstract, acknowledgement, table of contents, list of tables, list of figures, list of abbreviations, table titles, figure captions and references. Each reference must be separated by a single-line spacing. Margins on each page must be maintained as follows: left hand, 40 mm; right hand, 15 mm; top and bottom, 25 mm.

All the pages in the Project Report/Thesis must be computer printed only on one side of the page using Times New Roman (size 12) font with 1.5-line spacing. However, the following components of the project report/thesis should have single-line spacing: declaration, abstract, acknowledgement, table of contents, list of tables, list of figures, list of abbreviations, table titles, figure captions and references. Each reference must be separated by a single-line spacing. Margins on each page must be maintained as follows: left hand, 40 mm; right hand, 15 mm; top and bottom, 25 mm.

(?)

1.2 Outline of Thesis

CHAPTER 2
METHODOLOGY

This research employs a systematic multi-phase approach to develop and evaluate machine learning models for forecasting carrot prices in the Dambulla market. The methodology consists of five main phases: data collection, data preprocessing, model selection, model training, and model evaluation. The best-performing model is subsequently integrated into a Retrieval-Augmented Generation (RAG) based AI agent to provide intelligent price forecasting and analytical insights.

2.1 Data Collection

Historical data was collected from multiple authoritative sources to capture the various factors influencing carrot prices in the Dambulla market. Market price data, including daily wholesale carrot prices and trading information, was obtained from the Central Bank of Sri Lanka database. Weather and climate data, particularly precipitation measurements from major carrot-growing regions across Sri Lanka, was sourced from the Copernicus Climate Data Store, which provides reliable meteorological observations. Fuel price data, including diesel and petrol prices that affect transportation costs, was collected from the Ceylon Petroleum Corporation (ceypetco.gov.lk) official website. Additional market indicators such as supply factors from various cultivation regions, demand levels, and trading activity information were compiled from agricultural market reports. The dataset spans from January 2020 to July 2025, comprising daily observations to ensure sufficient granularity for time series analysis.

2.2 Data Preprocessing

Data preprocessing was conducted in two main steps to ensure data quality and prepare the dataset for model training.

2.2.1 Data Visualization

Exploratory data analysis was performed to understand the underlying patterns, trends, and relationships within the dataset. This phase involved creating several visualizations using Python libraries, specifically matplotlib and seaborn. The following plots were generated:

Time Series Visualizations:

- Historical carrot price trends over the entire study period to identify seasonality, cyclical patterns, and overall price movements
- Precipitation patterns from Nuwara Eliya and other major growing regions to observe rainfall distribution over time

Relationship Analysis:

- Scatter plots and correlation analyses between carrot prices and external factors including precipitation levels, fuel prices, supply factors, demand indicators, and trading activity levels
- Comparative visualizations of carrot prices during different market conditions (high demand vs. low demand, high supply vs. low supply)

Distribution Analysis:

- Histograms and box plots to examine price distribution and identify potential outliers
- Statistical summaries of key variables to understand data characteristics

These visualizations provided crucial insights into data patterns and informed subsequent modeling decisions, particularly regarding feature selection and the appropriateness of different modeling approaches.

2.2.2 Missing Value Treatment

To address missing values in the dataset, forward filling (also known as last observation carried forward) was employed. This method propagates the last observed value forward to fill subsequent missing entries, which is appropriate for time series data where values typically change gradually over time. Forward filling was chosen over backward filling or interpolation to avoid introducing future information into historical records, thereby preventing data leakage that could artificially inflate model performance. This approach ensures temporal integrity in the dataset, maintaining the chronological causality essential for accurate time series forecasting.

2.3 Model Selection

Three distinct machine learning approaches were selected for carrot price forecasting based on their theoretical foundations and demonstrated effectiveness in agricultural price prediction literature.

2.3.1 ARIMA (AutoRegressive Integrated Moving Average)

ARIMA models were selected as they are widely used in agricultural commodity price forecasting and have shown consistent performance across numerous studies. ARIMA is particularly well-suited for univariate time series data that exhibits stationary behavior after appropriate differencing. The model captures autocorrelation patterns in historical price data, making it effective for short to medium-term forecasting. Its interpretability and established track record in economic forecasting make it a valuable baseline model for comparison.

2.3.2 LSTM (Long Short-Term Memory)

LSTM neural networks were chosen due to their superior capability in capturing long-term dependencies and complex non-linear patterns in time series data. Unlike traditional statistical methods, LSTMs can effectively model the sequential nature of time series while handling multiple input features simultaneously. Recent literature demonstrates that LSTM models consistently achieve high prediction accuracy in vegetable price forecasting, often outperforming traditional methods when dealing with volatile and non-stationary price data. The model's ability to learn from both univariate and multivariate inputs makes it particularly suitable for this research.

2.3.3 Random Forest

Random Forest, an ensemble learning method, was selected as a comparative model due to its robust performance in handling high-dimensional data with complex non-linear relationships. The model has been successfully applied in agricultural price prediction studies and offers advantages including resistance to overfitting, ability to capture feature interactions, and relatively low sensitivity to hyperparameter tuning. Its inclusion provides a non-sequential modeling perspective that serves as an important benchmark against time series-specific approaches.

2.4 Model Training

Model training was conducted using Python programming language with scikit-learn, TensorFlow, and statsmodels libraries. The dataset was divided into training, validation, and test sets using a temporal split to maintain chronological order and prevent data leakage.

2.4.1 ARIMA Model Training

Two ARIMA configurations were trained to evaluate the impact of external factors:

Univariate ARIMA: This baseline model was trained using only historical carrot price data. The model parameters (p, d, q) representing the autoregressive order, degree of differencing, and moving average order respectively, were determined through systematic grid search and validated using the Akaike Information Criterion (AIC). Stationarity of the price series was tested using the Augmented Dickey-Fuller test prior to model fitting.

Multivariate ARIMA (ARIMAX): An extended ARIMA model incorporating exogenous variables was trained to assess whether external factors improve forecasting accuracy. This model included weather conditions, fuel prices, supply factors, and market indicators as additional input features alongside historical prices.

2.4.2 LSTM Model Training

LSTM models were implemented using TensorFlow/Keras framework with appropriate architecture design:

Univariate LSTM: A baseline LSTM network was trained using only the sequence of historical carrot prices. The model architecture included multiple LSTM layers with dropout regularization to prevent overfitting. Training was conducted using backpropagation through time with early stopping based on validation loss to avoid overtraining.

Multivariate LSTM: To leverage the full spectrum of available data, a multivariate LSTM model was developed incorporating external factors alongside price history. Prior to training this model, statistical filter methods were applied to identify the most relevant features for price prediction. This feature selection process reduced dimensionality and improved model efficiency while maintaining predictive power. The filtered feature set included key variables from precipitation data, fuel prices, supply factors, demand indicators, and temporal features. The multivariate LSTM architecture was designed to process multiple input features simultaneously across the temporal sequence, enabling the model to learn complex interactions between different factors affecting carrot prices.

2.4.3 Random Forest Model Training

A Random Forest regression model was trained using the same multivariate feature set employed in the multivariate LSTM model. The ensemble model consisted of multiple decision trees trained on bootstrapped samples of the dataset. Unlike the sequential models (ARIMA and LSTM), Random Forest treats each time point as an independent observation, providing a different modeling perspective. This approach serves as a benchmark to evaluate whether the temporal dependencies captured by sequential models provide significant advantages over ensemble methods in this forecasting task.

2.5 Model Evaluation

All trained models were evaluated using standard time series forecasting metrics to ensure comprehensive performance assessment. The evaluation was conducted on a held-out test set that was not used during model training or hyperparameter tuning. Model performance was compared across all approaches to identify the most accurate forecasting method for carrot prices in the Dambulla market. The best-performing model was selected based on prediction accuracy, stability across different market conditions, and practical applicability for operational use.

2.6 AI Agent Development Using RAG Architecture

Following model evaluation, the predictions from the best-performing model were integrated into an intelligent AI agent built using Retrieval-Augmented Generation (RAG) architecture. This agent combines the strengths of information retrieval systems with large language models to provide accurate, context-aware responses to user queries about carrot price forecasts.

2.6.1 RAG System Architecture

The RAG-based AI agent consists of three main components:

Vector Database: Prediction results from the best-performing model, including forecasted prices, actual prices, prediction dates, and relevant contextual information (such as market conditions, holidays, and supply factors), were stored in a vector database implemented using ChromaDB. Each prediction record was converted into a rich text representation and embedded into high-dimensional vector space using the BAAI/bge-base-en-v1.5 sentence transformer model. This embedding process enables efficient semantic search and retrieval of relevant price information based on user queries.

Retrieval Component: When a user submits a query, the system converts the query into the same vector space and performs similarity search to retrieve the most relevant prediction records from the database. This retrieval mechanism ensures that the AI agent has access to specific, factual information about price forecasts rather than relying solely on the language model's parametric knowledge.

Language Model: The retrieved information is provided as context to the Qwen 2.5-7B-Instruct large language model, which generates natural language responses to user queries. The model was configured with 4-bit quantization to enable efficient inference on standard computing hardware while maintaining response quality. The language model synthesizes the retrieved prediction data with its general knowledge about agricultural markets to produce coherent, informative answers.

2.6.2 Query Types and Response Generation

The AI agent is designed to handle two distinct categories of user queries:

Date-Specific Queries: When users request price information for specific dates (e.g., "What is the carrot price on 2024-06-15?"), the system retrieves the corresponding prediction record and provides a factual response including the predicted price, actual price (if available), prediction accuracy, and relevant contextual factors such as whether the date falls on a holiday or weekend.

Analytical Queries: For questions requiring analysis and reasoning (e.g., "Why did carrot prices increase from 2024-04-02 to 2024-04-08?"), the system retrieves relevant prediction records for the specified date range, calculates statistical measures (price change, volatility, trend direction), identifies contextual factors (holidays, supply disruptions, weather events), and uses the language model to generate an analytical explanation. The response combines quantitative analysis with qualitative reasoning to provide comprehensive insights into price movements.

2.6.3 Implementation Details

The AI agent was implemented in Python using the LangChain framework for orchestrating the RAG pipeline, ChromaDB for vector storage and retrieval, and the Transformers library for interfacing with the Qwen language model. The system employs prompt engineering techniques to ensure that the language model generates responses strictly based on retrieved information rather than hallucinating facts. Response generation is constrained to maintain factual accuracy, with the model instructed to acknowledge when information is unavailable rather than providing speculative answers.

This RAG-based approach ensures that the AI agent provides reliable, traceable price forecasts grounded in actual model predictions while leveraging the natural language understanding and generation capabilities of large language models to deliver user-friendly, contextually rich responses.

2.7 Summary of Methodology

This comprehensive methodology integrates traditional statistical methods (ARIMA), modern deep learning techniques (LSTM), and ensemble approaches (Random Forest) to develop robust carrot price forecasting models. The multi-phase approach ensures rigorous data preparation, systematic model comparison, and practical deployment through an intelligent AI agent. By combining quantitative forecasting with qualitative analysis capabilities, the research delivers both accurate predictions and interpretable insights valuable for stakeholders in the agricultural supply chain.

References.bib