# Comparative Analysis of ARIMA and LSTM Models for Agricultural Product Price Forecasting

Chenxi Xia *

College of Business, Southern University of Science and Technology, Shenzhen, 518055, China

* Corresponding author: 12010419@mail.sustect.edu.cn

**Abstract.** The fluctuations in vegetable prices can have an impact on the economy. Machine learning can identify price trend changes. This study investigates the performance of ARIMA and LSTM models in predicting price trends for agricultural products, focusing on greens and lotus roots. The objective was to ascertain the superior model in terms of accurately reflecting market oscillations—a critical aspect for stakeholders in the agricultural sector. The investigation contrasted the ARIMA model's adeptness at detecting linear tendencies against the LSTM's capacity to decode intricate nonlinear dynamics. Empirical assessment employing metrics such as RMSE, MAE, and SMAPE disclosed a consistent supremacy of the ARIMA model over the LSTM in both datasets. This was particularly evident within the lotus root forecasts, where the discrepancy in error metrics for LSTM was remarkably pronounced. The outcomes indicate that notwithstanding the sophisticated structure of LSTM, ARIMA models maintain their status as robust and precise tools for agricultural time series forecasting. This is especially pertinent in contexts demanding computational efficiency and model interpretability. This investigation evaluates the efficacy of ARIMA and LSTM models in forecasting time-series data for agricultural commodities.

**Keywords:** Price prediction, ARIMA, LSTM.

## 1. Introduction

Predictive modeling of vegetable prices using machine learning techniques is of paramount importance. The cost of vegetables plays a pivotal role in determining both consumer spending and farmers' revenue; significant fluctuations can lead to considerable economic repercussions. Through the analysis of extensive historical datasets, machine learning can discern price trends and patterns, subsequently forecasting impending price shifts. Notably, machine learning techniques are adept at interpreting nonlinear patterns, thereby offering a more nuanced understanding of the intricate variations associated with vegetable pricing. Such predictions can facilitate more informed planning and decision-making processes for farmers, consumers, and governing bodies. This is imperative not only for the robustness of rural economies but also for a country's sustained economic advancement.

Recent academic research has seen noteworthy progress in this area. Assis and Remali [1] examined various time-series methods for cocoa bean pricing, emphasizing the ARIMA model's effectiveness. Adanacioglu and Yercan [2], studying tomato prices in Turkey, adjusted for their pronounced seasonality and applied a seasonal ARIMA model. On a different note, Ge and Wu [3] employed a multivariate linear regression approach to predict corn prices. Minghua et al. [4] utilized a back-propagation neural network for predicting agricultural prices, highlighting its superiority over traditional statistical approaches. Wang et al. [5] turned their attention to garlic pricing, focusing on its non-linear nature, using a hybrid ARIMA-SVM model, demonstrating the hybrid's increased prediction precision over individual ARIMA and SVM models. Yin et al. [6] presented an innovative STL-Attention-based LSTM model. They combined the Loess (STL) preprocessing with an LSTM-based attention mechanism. Departing from conventional regression methods, Panigrahi and Behera [7] introduced a hybrid ETS-ANN model, splitting the time series into two components. These studies collectively validate the applicability of machine learning techniques in the domain of price forecasting.

This study conducts a comparative analysis of ARIMA and LSTM models to predict the prices of agricultural commodities, leveraging historical data from the Kalimati Market. It aims to ascertain

which model more precisely mirrors the market trends for greens and lotus roots, ultimately facilitating stakeholders in executing informed economic decisions.

## 2. Material and Method

### 2.1. Dataset

Kalimati Tarkari is a place for selling vegetables and fruits in Kathmandu Valley. The price of Fruits and Vegetables from July 2020 to June 2023 is grabbed from the website of Kalimati Fruits and Vegetables Market Development Board (https://kalimatimarket.gov.np/). Fig 1 depicts the average price trajectory of two vegetables—greens and lotus root—at the Kalimati Fruits and Vegetable Market, spanning from July 2020 to June 2023.

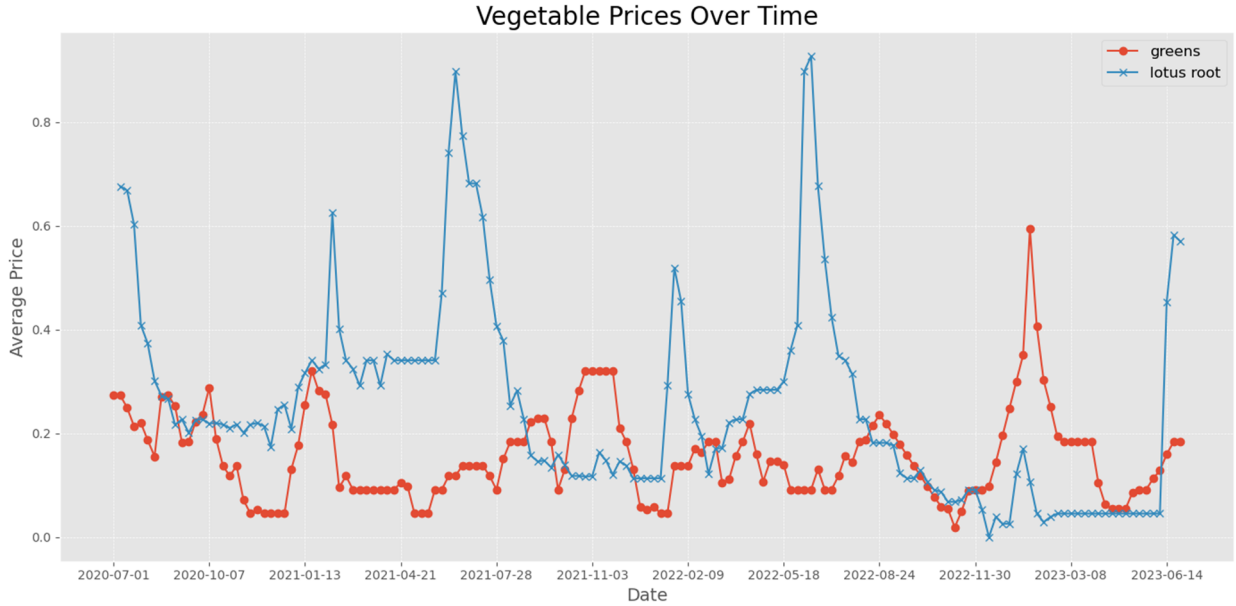


**Figure 1.** Kalimati Fruits and Vegetable Market

### 2.2. Autoregressive Integrated Moving Average Model

The ARIMA model intricately intertwines three distinct components: the Autoregressive model (AR), the Integrated process (denoted as 'I' for differencing), and the Moving Average model (MA).

The foundational concept of the ARIMA model is predicated on the utilization of a dataset's historical information for the prognostication of future trends. A datum at a specific time is influenced not only by the historical data from a preceding period but also by stochastic events during that time. This implies that the ARIMA model postulates that data fluctuate around a grand temporal trend, which is shaped by historical data, while the fluctuations are shaped by random events within a given time frame, with the overarching trend itself possibly being unstable.

In essence, the ARIMA model endeavors to discern the latent temporal patterns within the data through the exploitation of autocorrelation and differencing. These extracted patterns are then employed to forecast future observations. Specifically:

(1) The AR component addresses the autoregressive aspect of the time series, considering the influence of observations from previous periods on the current value.

(2) The I component aims to achieve stationarity in non-stationary time series through methods such as first or second differencing, thus mitigating trends and seasonal effects.

(3) The MA component deals with the moving average portion of the series, taking into account the impact of past forecast errors on the present value.

These Mathematical formulas can be expressed below.

$$AR: Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \xi_t \quad (1)$$

$$MA: Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \quad (2)$$

Where $Yt$ is the time series data we are considering. $\varphi_1$ to $\varphi_p$ are the parameters of the AR model, which describe the relationship between the current value and the past $p$ time points. $\theta_1$ to $\theta_q$ are the parameters of the MA model that describe the relationship between the current value and the error at $q$ time points in the past. $\epsilon_t$ is the error term at $t$ time. $c$ is a constant term.

If the difference is ignored for the time being (i.e., d=0). Formally, the formula of the ARIMA model can be expressed as:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (3)$$

By integrating these three elements, the ARIMA model is adept at capturing trends and variations in data, as well as handling data characterized by transient, sudden changes, or a high degree of noise. Consequently, the ARIMA model exhibits commendable performance across numerous time series forecasting challenges.

$$MA: Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \quad (4)$$

### 2.3. LSTM

Long Short-Term Memory (LSTM) networks are a specialized category of Recurrent Neural Networks (RNNs), which have seen widespread application across diverse areas including speech recognition, language processing, machine translation, image captioning, and text detection. The unique feature of RNNs is their ability to incorporate information from prior sequences to address current tasks. However, RNNs struggle with maintaining relevance across longer sequences due to long-term dependency issues, causing difficulty in context linkage over extended gaps. To address these challenges, including the vanishing gradient issue, Hochreiter and colleagues [8] developed LSTM networks. An LSTM network is composed of various gates—input, forget, and output—and a cell state, all operating within a single layer. The architecture of an LSTM is depicted in a given illustration in Fig 2. In particular, the forget gate's functionality is described by a specific equation. This gate processes both the previous hidden state $h_{t-1}$ and the current input $x_t$, applies matrix multiplication with its trainable weight matrix W, and adds a bias term $b$. The output of this gate, $f$ , is then passed through a sigmoid function, which outputs a value ranging between 0 and 1. Consequently, a result close to 1 indicates the retention of more information from the prior cell state $C_{t-1}$, while a result near 0 implies the omission of information from $C_{t-1}$.

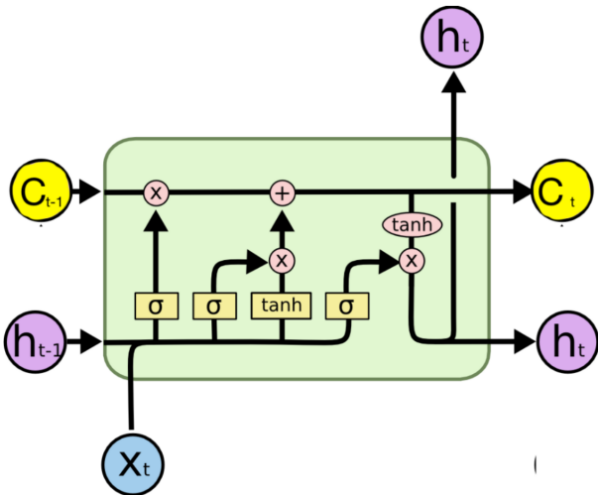


**Figure 2.** LSTM structure

## 2.4. model Validation

### 2.4.1. Extending window splitting

The Expanding Window Splitter is a method employed for time series model validation, ensuring temporally coherent data division (Fig 3). With this approach, the length of the training sequence incrementally increases over consecutive folds, maintaining a complete historical sequence. Conversely, the test sequence remains constant in length. This entails dividing the time series into contiguous blocks, where each, after its initial testing phase, is used for model retraining. Notably, the first block exclusively serves training purposes, preserving the model's temporal integrity.

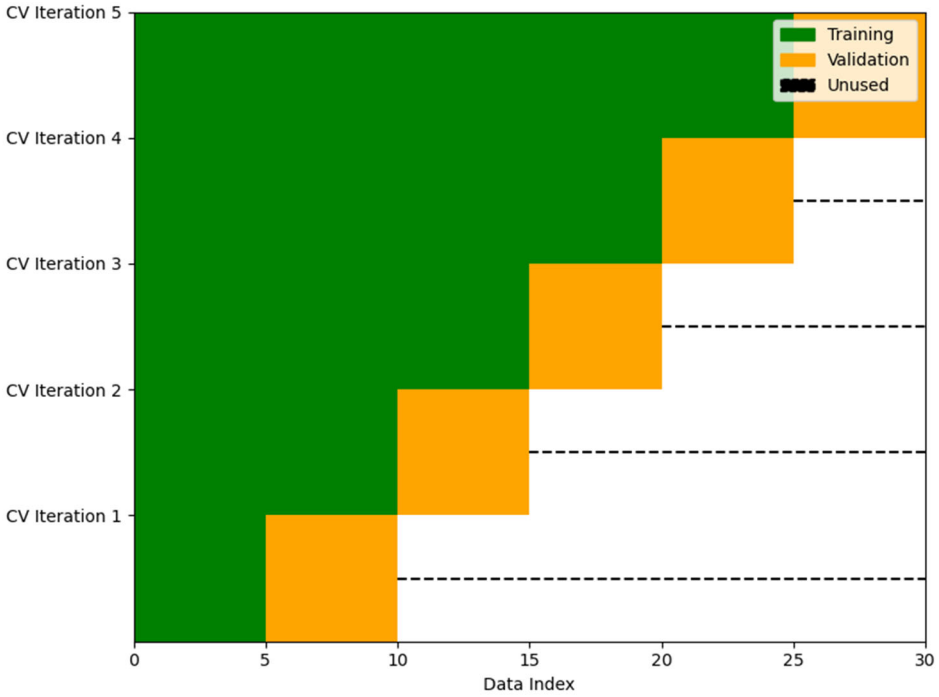


**Figure 3.** Extending window splitting

## 2.5. Model Evaluation

### 2.5.1. Root Mean Squared Error

RMSE stands as a predominant metric in assessing the precision of predictive models, primarily in regression analyses. This metric quantifies the average squared discrepancies between the forecasted and actual outcomes, subsequently extracting the square root of this mean. The formula can be expressed below.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{y}_i)^2} \tag{5}$$

### 2.5.2. Symmetric Mean Absolute Percentage Error

SMAPE serves as a measure to evaluate the veracity of forecasts, especially within the realm of time series predictions. Unlike RMSE, SMAPE computes the percentage deviation between anticipated and observed results. This renders it optimal for juxtaposing models of varying magnitudes. The symmetric nature of SMAPE ensures an impartial assessment of both over-predictions and under-predictions. Expressed in percentage terms, a diminished SMAPE value signifies superior model efficacy. The formula can be expressed below.

$$SMAPE = \frac{100\%}{n}\Sigma_{t=1}^{n}\frac{|y_i - \hat{y}_i|}{(|\hat{y}_i| + |y_i|)/2} \tag{6}$$

### 2.5.3. Mean Absolute Error (MAE)

MAE offers a lucid and discernible metric for gauging a predictive model's prowess. It ascertains the mean absolute disparity between projected and actual readings. MAE elucidates the magnitude by which a model's forecasts deviate from factual values, disregarding the error's direction. The formula can be expressed below.

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |(y_i - \hat{y}_i)| \quad (7)$$

## 3. Results

### 3.1. Results for ARIMA Prediction

Fig 4 presents a comparative line graph illustrating the actual versus predicted average prices of greens on the test set using the ARIMA model. As shown in Fig 4, the forecast analysis of the 'greens' time series predominantly displays a close approximation between predicted and actual values, illustrating the model's aptitude for capturing the overarching data trends accurately. However, an aberration is observed on or around January 18, 2023, where an unforeseen peak in the actual values emerges, possibly signaling an atypical incident or extreme market volatility that eluded the model's predictive scope. The model's predictions align with actual data once more following this outlier, underscoring its resilience and ability to recalibrate its forecasting accuracy in the face of anomalies.

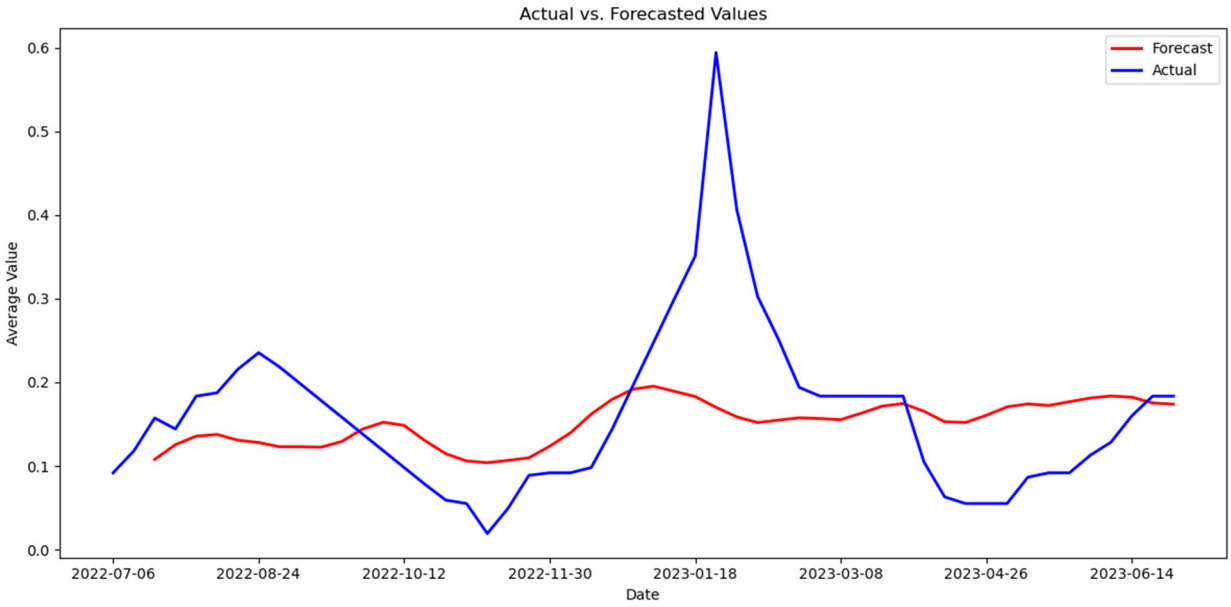


**Figure 4.** ARIMA Prediction of greens average price over time

Fig 5 presents a comparative line graph illustrating the actual versus predicted average prices of lotus root on the test set using the ARIMA model. In the case of the 'lotus root' time series, the initial predictive output diverges significantly from the actual values, particularly around July 6, 2022. This divergence may reflect an initial calibration phase where the model adjusts to the patterns of volatility within the data. Thereafter, the predictions adhere closely to the actual value trajectory. Yet, the model does not consistently capture minor fluctuations, especially in the period following January 18, 2023. From March 8, 2023, the model fails to project an uptick in actual values, culminating in a notable predictive discrepancy as actual values ascend sharply towards the dataset's end.

In both instances, the ARIMA model appears adept at capturing the general trend of the price data, as evidenced by the proximity of forecasted to actual values at the majority of temporal points. Despite the model's ability to track the overall trajectory of the data, its performance is less satisfactory when confronting abrupt changes, such as spikes or sudden trend shifts. There seems to be a lag in the ARIMA model's responsiveness to sudden changes in actual values. This lag is

particularly noticeable in its reaction to the spike in 'greens' prices and the final ascending trend in 'lotus root' prices. While the model demonstrates robustness in handling stable or moderately volatile data, its predictive accuracy diminishes when faced with severe fluctuations and outliers.

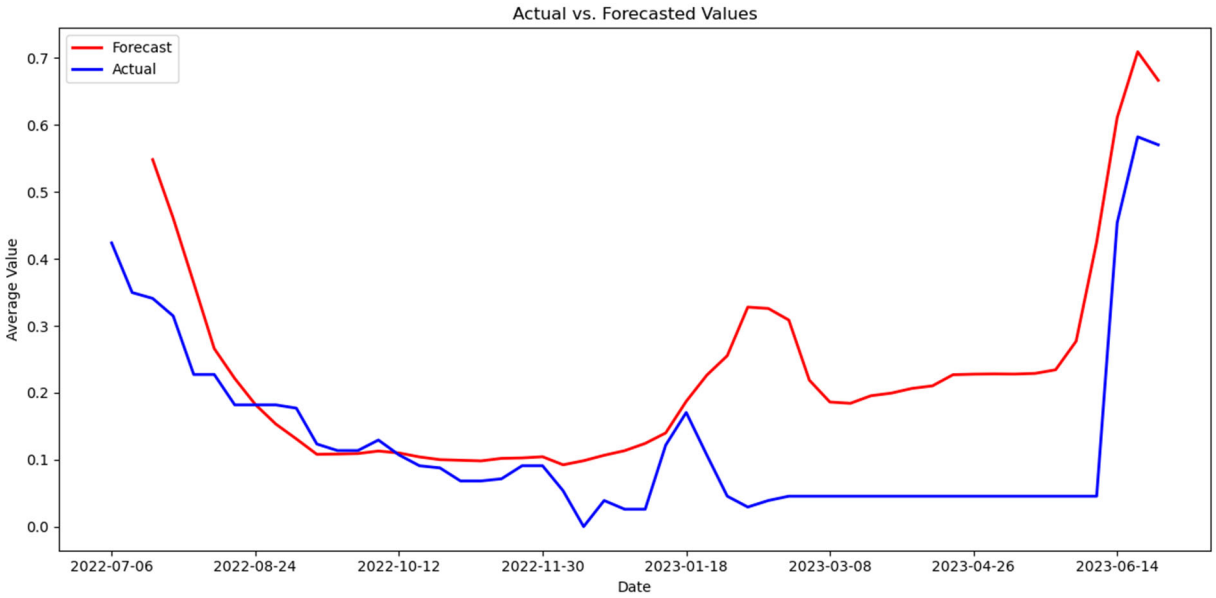


**Figure 5.** ARIMA Prediction of lotus root average price over time

### 3.2. Results for LSTM Prediction

The Long Short-Term Memory (LSTM) model exhibits a consistent deviation between predicted and actual values across the majority of time points under observation based on Fig 6. Notably around January 2023, the forecasted values displayed a peak whereas the actual data remained relatively stable, indicating a period of suboptimal predictive performance by the model. Moreover, a closer alignment between predicted and actual data at the beginning and end segments of the time series suggests the model's enhanced forecasting accuracy at the margins of the sequence. Overall, while the model offers some reference value in predicting this price series, it demonstrates a lower precision in capturing fluctuations during certain intervals.

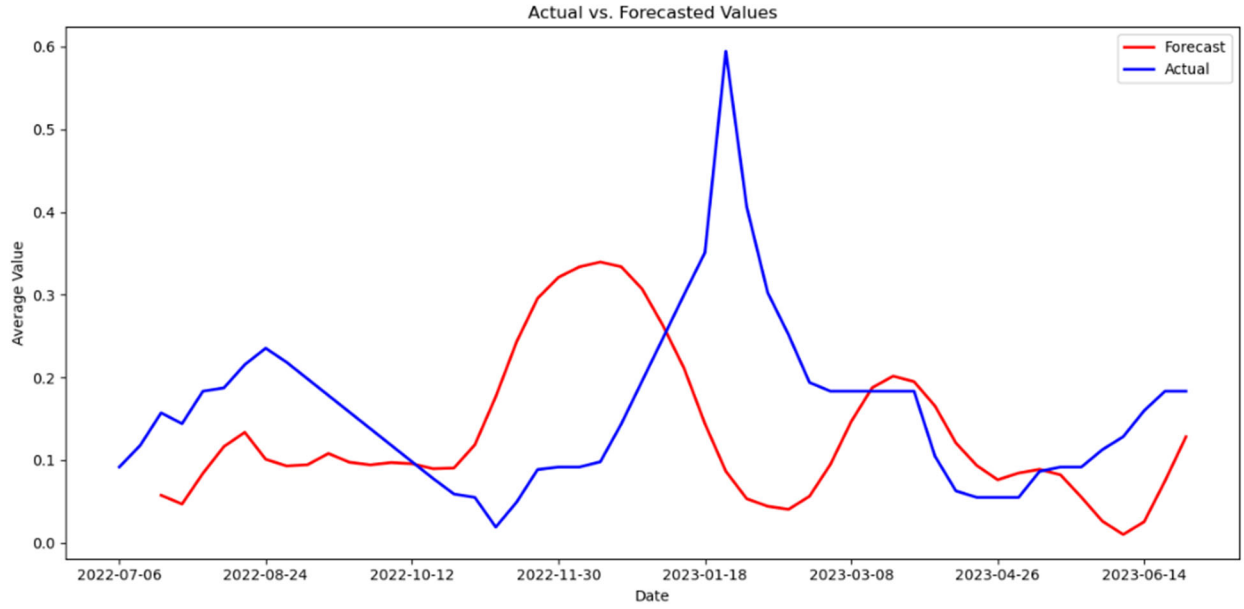


**Figure 6.** LSTM Prediction of greens average price over time

Throughout the time series based on Fig 7, the predicted trajectory mirrors the actual values. The LSTM model adeptly captures the pricing trends, particularly from July to August 2022, where the

predictive and actual curves rise and fall in near synchrony. However, it is important to note that the model failed to capture a significant peak in mid-January 2023, which may indicate limitations in the model's ability to anticipate sudden or irregular market movements. Additionally, towards the end of the sequence, the model appears to have overestimated the uptrend in pricing.

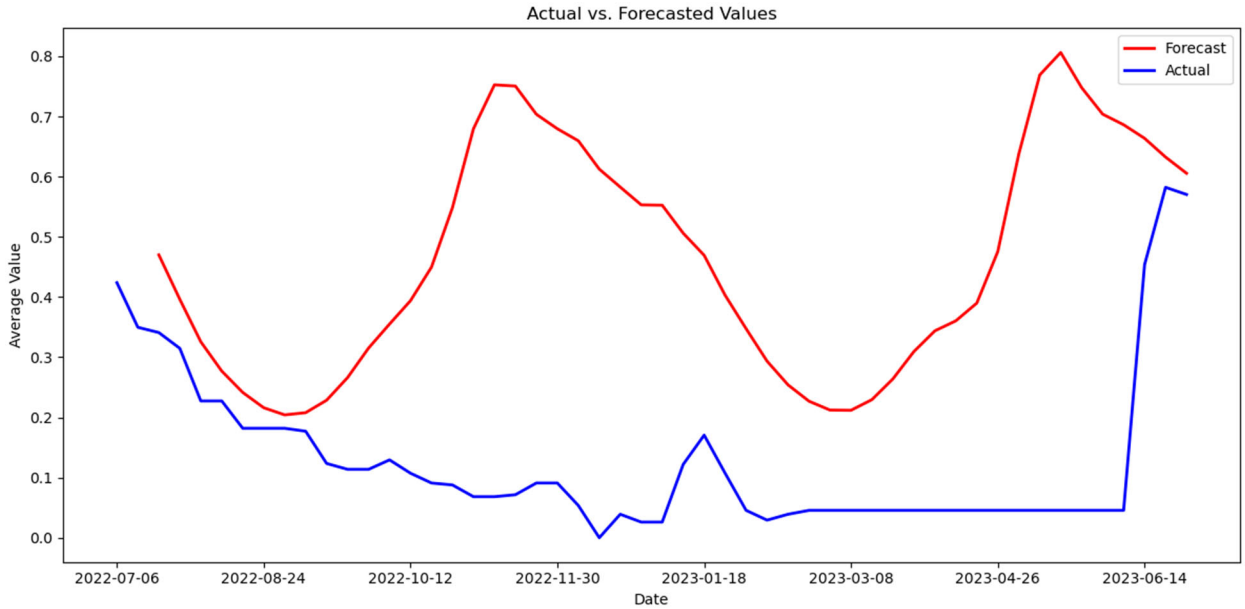


**Figure 7.** LSTM Prediction of lotus root average price over time

### 3.3. Evaluation

As shown in Table 1, for the "greens" dataset, the ARIMA model demonstrated considerable predictive accuracy, with a low Root Mean Square Error (RMSE) of 0.0989 and Mean Absolute Error (MAE) of 0.0686, suggesting a proximate alignment of forecasted values with actual ones in most instances. However, a relatively high Symmetric Mean Absolute Percentage Error (SMAPE) of 47.07% indicates potential substantial relative errors at certain data points, particularly where actual price fluctuations were minimal. In contrast, on the "lotus root" dataset, the ARIMA model's prediction error notably increased, with RMSE rising to 0.1693 and MAE to 0.1205, while a sharp escalation in SMAPE to 92.96% revealed frequent and significant deviations between predicted and actual values, denoting poor predictive performance on this dataset. In essence, while the ARIMA model achieves an acceptable level of precision on the "greens" dataset, its efficacy substantially diminishes on the "lotus root" dataset.

When deploying the LSTM model for forecasting "greens" and "lotus root", the performance was suboptimal for both datasets. For "greens," the LSTM model exhibited a higher RMSE of 0.1536 and MAE of 0.1075 than the ARIMA model, with SMAPE at 70.83% also indicating significant predictive discrepancies. The performance was even more deficient for "lotus root," with markedly increased RMSE and MAE values of 0.4106 and 0.3335, respectively, and SMAPE soaring to 114.85%, which signifies extensive predictive errors and, in certain cases, a forecasted trend contradictory to actual data trends. These assessment outcomes suggest that the LSTM model may require further tuning and optimization to handle these specific time series datasets.

**Table 1.** Evaluation of ARIMA and LSTM

| | RMSE | SMAPE | MAE |
|---|---|---|---|
| ARIMA (greens) | 0.0989 | 47.07% | 0.0686 |
| ARIMA (lotus root) | 0.1693 | 92.96% | 0.1205 |
| LSTM (greens) | 0.1536 | 70.83% | 0.1075 |
| LSTM (lotus root) | 0.4106 | 114.85% | 0.3335 |

Upon comprehensive evaluation of the metrics, it is discernible that the ARIMA model, on the whole, surpasses the LSTM model in forecasting accuracy for both products. This disparity is particularly pronounced in the "lotus root" forecasts, where the error indices of the LSTM model are significantly greater than those of the ARIMA model, suggesting difficulties encountered by LSTM in addressing specific features of this dataset. However, it is crucial to acknowledge that the elevated values of SMAPE might amplify the perceived inaccuracies in the model's predictions, especially when the data points are proximal to zero.

## 4. Discussion

In light of the preceding analysis, a detailed discussion has been conducted on the performance of ARIMA and LSTM models in forecasting the prices of "greens" and "lotus root."

Regarding the performance of the ARIMA model, it demonstrates periods of effective alignment with the price trends, underscoring its competence in capturing linear patterns within time-series data. Nonetheless, the model's predictive accuracy diminishes when confronted with volatile price fluctuations and sharp outlier values, with the elevated SMAPE values reflecting this limitation to some extent. This may stem from the inherent linear assumptions of the ARIMA framework, which do not accommodate or anticipate non-linear data fluctuations. To enhance its adaptability to more intricate data patterns, the model necessitates further optimization, potentially including the refinement of parameters such as the autoregressive terms ($p$), degree of differencing ($d$), and moving average components ($q$).

The LSTM model, on the other hand, encounters distinct challenges. Theoretically capable of capturing long-term dependencies within time-series data through its internal memory mechanism, its practical application is frequently contingent on the volume of data, hyperparameter configurations, model complexity, and the extent of training. In the cases examined, the LSTM did not meet expectations, which may be attributed to an inadequately tuned network architecture for the specific data characteristics, or insufficient training data to capture and learn the complex dynamics of pricing.

Ultimately, the selection and evaluation of time-series forecasting models should be dynamic, with regular adjustments and optimizations based on the evolving characteristics and performance of the data. Practically, an ensemble approach, synthesizing predictions from multiple models, may be necessary to achieve more robust and precise outcomes. Moreover, incorporating external data and considering factors such as economic cycles could contribute to further enhancing the predictive capability of the models.

## 5. Conclusion

This article presents a comparative assessment of ARIMA and LSTM models in analyzing the price trends of agricultural products, specifically greens and lotus roots, to evaluate their predictive accuracy concerning market fluctuations. The study reveals a consistent superiority of the ARIMA model across both datasets, with notably lower error metrics for lotus root price predictions, thus demonstrating ARIMA's robustness and precision in agricultural time series forecasting. Despite the LSTM model's structural complexity, its performance did not meet expectations, particularly in addressing price spikes. This suggests that model selection in practical applications should be tailored to the specific characteristics of the dataset, with ongoing adjustments and optimizations to accommodate dynamic changes in data. Future research may explore hybrid approaches that effectively combine ARIMA and LSTM models or incorporate additional external variables to enhance the accuracy of predictive models. Such advancements are vital for precise agricultural price forecasting and bolstering the stability of agricultural markets.

## References

[1] Assis, K., A. Amran, and Y. Remali. Forecasting cocoa bean prices using univariate time series models. Researchers World, 2010, 1 (1): 71.

[2] Adanacioglu, Hakan, and Murat Yercan. An analysis of tomato prices at wholesale level in Turkey: an application of SARIMA model. Custos e@ gronegócio on line, 2012, 8 (4): 52 - 75.

[3] Gu, Y. et al. Blite-Svr: New Forecasting Model for Late Blight on Potato Using Support-Vector Regression. Computers and electronics in agriculture, 2016, 130: 169 - 76.

[4] Wei, M. H. et al. Prediction Model of Agricultural Product's Price Based on the Improved Bp Neural Network. 2012 7th International Conference on Computer Science & Education (ICCSE), IEEE, 2012: 613 - 17.

[5] Wang, B. J. et al. Research on Hybrid Model of Garlic Short-Term Price Forecasting Based on Big Data. Computers, Materials & Continua, 2018, 57.

[6] Yin, H.L. et al. Stl-Attlstm: Vegetable Price Forecasting Using Stl and Attention Mechanism-Based Lstm. Agriculture, 2020, 12 (10): 612.

[7] Panigrahi, S. and Himansu S.B. A Hybrid Ets–Ann Model for Time Series Forecasting. Engineering applications of artificial intelligence, 2017, 66: 49 - 59.

[8] Hochreiter, S., and Jürgen, S. Long short-term memory. Neural computation, 1997, 9.8: 1735 - 1780.