Machine Learning for Agricultural Price Prediction: A Case of Coffee Commodity in Vietnam Market

Conference Paper · December 2023

DOI: 10.1109/BCD57833.2023.10466313

9 authors, including:

Thong Doan
University of Information Technology
1 PUBLICATION **7** CITATIONS
SEE PROFILE

Nam Phương
Van Lang University
81 PUBLICATIONS **823** CITATIONS
SEE PROFILE

Thanh Ngoc Nguyen
RMIT Vietnam
25 PUBLICATIONS **116** CITATIONS
SEE PROFILE

Linh Duc Tran
RMIT Vietnam
16 PUBLICATIONS **79** CITATIONS
SEE PROFILE

Machine Learning for Agricultural Price Prediction: A Case of Coffee Commodity in Vietnam Market

Thuan Nguyen Le Ngoc[1], Dieu Tin Lam[1], Trang Nguyen Hai Minh[1], Thong Chanh Doan[1], Nam Phuong Nguyen[1]
Hien Manh Nguyen[2], Thanh Ngoc Nguyen[1], Linh Duc Tran[1], Nhat-Quang Tran[1*]
[1] School of Science, Engineering and Technology, RMIT University, Ho Chi Minh City, Vietnam
[2] Phuonghai Technology Science JSC
* E-mail: quang.tran26@rmit.edu.vn

Abstract – Predicting agricultural commodity prices is crucial for farmers, governments, and related stakeholders to make important decisions about crop planning and food security. This is especially important in economies that rely greatly on agriculture, such as Vietnam and other developing countries, where the instability and inadequate predictability of agricultural commodity prices may lead to overproduction, "good season - devaluation" phenomenon, and significant economic losses. Aiming to address this issue, this paper presents the use of machine learning models to predict the price of coffee, a major agricultural commodity in Vietnam. Various machine learning techniques, namely LSTM (Long Short-Term Memory), ARIMA (Autoregressive Integrated Moving Average), SARIMA (Seasonal ARIMA), GRU (Gated Recurrent Unit), SVM (Support Vector Machine), and RF (Random Forest), are trained and evaluated using different data, including historical prices and additional data, such as fuel prices and weather data. The results of numerical experiments demonstrated that using additional data can significantly improve prediction performance.

Keywords—Price Prediction, ARIMA, SARIMA, LSTM, GRU, SVM, Random Forest

I. INTRODUCTION

According to the World Bank, "agricultural development is one of the most powerful tools to end extreme poverty, boost shared prosperity, and feed a projected 10 billion people by 2050. Growth in the agriculture sector is two to four times more effective in raising incomes among the poorest compared to other sectors" [1]. Overall, agriculture accounts for 4% of global GDP and more than 25% in some least developing countries. With a turnover of over 53 billion USD annually, Vietnam ranks among the top 15 countries worldwide for agricultural exports, selling goods to more than 200 countries and territories [2].

The prediction of agricultural commodity prices is crucial for farmers, businesses and governments in Vietnam and developing countries, where agriculture plays an important role in the economy. However, due to a lack of research and investment on this topic, such countries may face a serious issue: the price of agricultural commodities can plunge sharply at the time of bumper crops, badly damaging farmers and the national economy. According to VOV, a Vietnamese government broadcaster, the country's agriculture pricing is unstable and hugely dependent on unmanageable factors such as exporting [3]

Fluctuations in crop prices can have a significant impact on farmers' incomes, agricultural investments, and overall market dynamics. Reliable price forecasts empower farmers to make informed decisions, enhancing their productivity, profitability, and livelihoods. Therefore, developing and utilizing robust crop price prediction models and tools are essential for supporting the resilience and prosperity of the agricultural sector, ensuring the well-being of farmers, and sustaining the broader economy.

Crop price prediction is also relevant to traders, policymakers, and other stakeholders within the agricultural value chain. Traders rely on price forecasts to plan their procurement and distribution strategies, ensuring a smooth flow of agricultural commodities. Policymakers utilize price predictions to design and implement effective agricultural policies, support farmers, and maintain market stability. Accurate crop price forecasts enable stakeholders to anticipate market trends, manage inventories, and develop strategies that promote sustainable agricultural growth and equitable market outcomes [4].

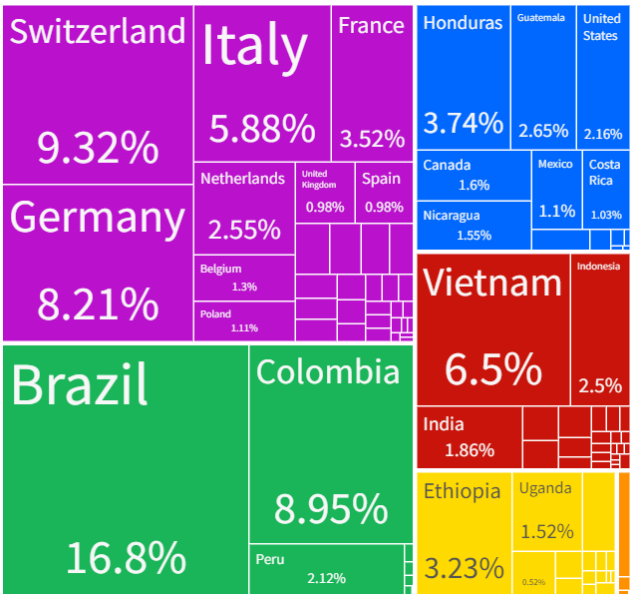

Fig. 1. Vietnam among the top coffee exporters [5].

By developing and evaluating effective agricultural price prediction models, this paper aims to contribute to improving the planning and decision-making process in agriculture and the sustainable development of the agricultural sector in Vietnam and other developing countries. Particularly, this paper presents the use of machine learning models to predict the price of coffee, one of Vietnam's most important agricultural commodities as illustrated in **Fig. 1**. Various different machine learning techniques, namely **LSTM** (Long short-term memory), **ARIMA** (Autoregressive Integrated Moving Average), **SARIMA** (Seasonal ARIMA), **GRU** (Gated Recurrent Unit), **SVM** (Support Vector Machine), and **RF** (Random Forest), will be trained and evaluated using different data, including historical prices and additional data, such as fuel prices and weather data (precipitation and temperature).

II. LITERATURE REVIEW

The agricultural sector in Vietnam is large, growing, and makes significant contribution to the country's economy. In 2020, agriculture employed 36.23% of the workforce and constituted 12.66% of the Vietnam GDP [6, 7] . Despite its economic significance, predicting crop prices using machine learning is in its early stages in Vietnam. There is little research on the use of machine learning for this purpose, and the historical price data available in the national database [8] has not yet been widely used in machine learning applications.

Historical price data are time series, so **ARIMA** (Autoregressive Integrated Moving Average) and **SARIMA** (Seasonal ARIMA) models [9] can be considered. Notably, in some cases, the seasonal element in SARIMA can be neglected, and using a simpler ARIMA model may improve the robustness of the time series analysis [10, 11]. In addition to ARIMA, other time series models have also been experimented with, such as **LSTM** (Long Short-Term Memory) [12] and **GRU** (Gated Recurrent Unit) [13]. These models collectively present a comprehensive approach to predicting agricultural prices, encompassing a range of methodologies from traditional time series analysis to sophisticated deep learning architectures.

Furthermore, in [14], the authors explored the predictive modeling of crop prices using **RF** (Random Forest) [15], **SVM** (Support Vector Machine) [16], and gradient-boosting algorithms. Their study highlighted the potential of machine learning in accurately forecasting agricultural prices, particularly when incorporating weather data as features. Additionally, Nishant et al. [17] focused on utilizing basic parameters easily understandable by farmers, such as state, district, season, and region, to enhance prediction accuracy. They experimented with techniques such as Kernel Ridge, Lasso, Elastic Net, and Stacking Regression. Their study emphasized the creation of a meta-model by aggregating predictions from base models, which showed a significant improvement in prediction accuracy through Stacked Regression. This methodology informs our approach to model stacking for improved prediction performance.

III. PROPOSED SOLUTION

In this study, we focus on forecasting coffee prices in Vietnam market using a variety of machine learning models mentioned above: **ARIMA**, **SARIMA**, **LSTM**, **GRU**, **SVM**, and **RF** (Random Forest). These models will be trained and evaluated using diverse datasets, encompassing not only historical crop prices but also additional relevant variables such as fuel prices, precipitation, and temperature data. The objective is to predict coffee prices with a lead time of one month in advance.

A. Training Data

Details of all datasets, which will be used for training, are outlined below:

- ***Coffee historical price data:*** were obtained from the thitruongnongsan.gov.vn website [8], an official platform of the Vietnamese government. We used Robusta coffee data, as this type of coffee is more popular in Vietnam. The dataset covers daily prices from January 2019 to December 2022.
- ***Fuel price data:*** were obtained from the autofun.vn website [18]. There are several different types of fuel, such as Ron 95, diesel, and train fuel. However, according to CoffeeConcept [19], most coffee beans are transported by container-trucks throughout the country, with some companies using ships as an alternative. Both vehicles run on diesel, so we use only diesel prices as the fuel data for analysis.
- ***Precipitation & Temperature data:*** weather data were retrieved from open-meteo.com [20]. Many natural conditions can affect crop yield, such as wind speed, light, and humidity. However, according to National Geographic [21], precipitation and temperature are the most crucial factors. Precipitation is the amount of condensation from clouds, including rain or snow. It is the amount of water plants receive from the sky apart from watering. Robusta coffee grows best at 24°C to 30°C and will die below 10°C [22]. Additionally, total precipitation needs to be between 2000 and 2500 mm/year. Too much or too little rain can negatively affect coffee crops.

Initially, it was found that the raw data files had several issues, including missing dates and values, incorrect timestamp format, and inconsistent units. Therefore, to ensure uniformity, we standardized all files to use the same units and timestamp format. The data are then preprocessed by filling in missing values with the nearest available values and removing outliers. As a result, **Fig. 2** shows time series graphs of all features.

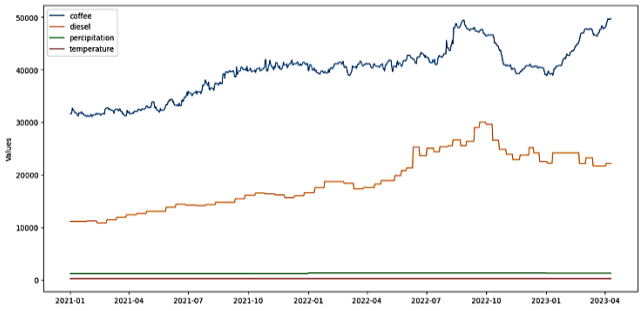

Fig. 2. Time series graphs of training data features

Table I shows a snippet of the final dataset in data frame format. All prices are in Vietnamese Dong (VND) currency. Since the transportation and storage time of coffee beans can take up to several months, each row in **Table I** includes the diesel prices of the three previous months (Fuel-m3, Fuel-m2, and Fuel-m1). For precipitation, it is usually measured by the total amount annually, thus we simply summed all precipitation values within a year to obtain the total amount annually. Furthermore, for the temperature, we calculated the mean temperature value for each year. Note that we also tried to use the total number of days with temperatures from 24°C to 30°C, but this method did not give good results compared to the mean temperature. Additionally, according to [19], Robusta coffee is harvested from October to December annually and takes 8-9 months to grow before it can be harvested again. Therefore, we use the weather data (precipitation and temperature) of the previous year for training purposes.

Table I. A snippet of final dataset in data frame format

	yyyy-mm-dd	Coffee price	Fuel-m3	Fuel-m2	Fuel-m1	Precipitation (mm)	Temperature
0	2021-01-01	31667.5	11430.0	11210.0	11120.0	1174.199994	24.937842
1	2021-01-02	31667.5	11430.0	11210.0	11120.0	1174.199994	24.937842
2	2021-01-03	31667.5	11430.0	11210.0	11120.0	1174.199994	24.937842
3	2021-01-04	32717.0	11430.0	11210.0	11120.0	1174.199994	24.937842
4	2021-01-05	32417.0	11430.0	11210.0	11120.0	1174.199994	24.937842
...	...	...	...	...	...	...	...
864	2023-05-15	55566.5	20140.0	20500.0	22520.0	1259.800013	24.170057
865	2023-05-16	56066.5	20140.0	20500.0	21560.0	1259.800013	24.170057
866	2023-05-17	56666.5	20140.0	20500.0	21560.0	1259.800013	24.170057
867	2023-05-18	57000.0	20140.0	20500.0	21560.0	1259.800013	24.170057
868	2023-05-19	57500.0	20140.0	20500.0	21560.0	1259.800013	24.170057

869 rows × 7 columns

B. Training Models

As discussed above, six machine learning models will be trained and evaluated, namely ARIMA, SARIMA, LSTM, GRU, SVM, and Random Forest (RF). The first four models, ARIMA, SARIMA, LSTM, and GRU, are times series models, hence we train them using only the coffee historical price data. For ARIMA and SARIMA, we split the time series into 70% training data and 30% test data. For other models, we use the coffee historical prices from the past 60 days for training and predict the price of the 30th day after the training period. Random Forest and SVM are capable of handling multivariate data. Therefore, we can feed both historical prices and other features to these models for training. All models use the same training and test dataset to ensure consistency and comparable results.

IV. EXPERIMENT RESULTS

After training, the performance of each machine learning model is evaluated using four key measures, namely Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE). As previously mentioned, ARIMA, SARIMA, LSTM, and GRU, being time series models, are trained using only coffee price data. Their performance metrics are presented in **Table II**. Among these time series models, **SARIMA** provides the best performance with an **RMSE** value of **1086.39** and a **MAE** value of **791.45**. This is a good performance, given that the coffee prices are typically from about **30000 to 50000** VND as shown in **Fig. 2**.

For traditional machine learning models, Random Forest (RF) and SVM, diverse combinations of features are employed for training. Their performance results are summarized in **Table III**. Notably, it is found that the RF model consistently outperforms SVM across all feature combinations. The most effective model emerges as **RF**, trained with **coffee price**, **fuel price**, and **precipitation**, presenting an **RMSE** of **637.34** and a **MAE** of **379.19**, which is even significantly better than the performance of time series models evaluated above.

In general, it is demonstrated that incorporating additional relevant variables such as fuel prices, precipitation, and temperature data can improve performance of machine learning models in predicting crop prices. Altogether, the results in **Table II** and **Table III** show that the traditional **RM** and **SVM** models can significantly outperform the time series models, especially in some appropriate feature combinations.

Table II. Performance of time series models

Features	Models			
	LSTM	GRU	SARIMA	ARIMA
Coffee price	RMSE: 2458.79 (đồng) MAE: 1929.19 (đồng) MAPE: 0.04 MASE: 11.07	RMSE: 2514.72 (đồng) MAE: 1999.03 (đồng) MAPE: 0.05 MASE: 11.47	RMSE: 1086.39 (đồng) MAE: 791.45 (đồng) MAPE: 0.02 MASE: 19.75	RMSE: 3428.60 (đồng) MAE: 3166.56 (đồng) MAPE: 0.07 MASE: 16.25

Table III. Performance of Random Forest (RF) and SVM trained with different data combinations.

Features	Models	
	RF	SVM
Coffee price	RMSE: 1063.05 (đồng) MAE: 642.84 (đồng) MAPE: 0.02 MASE: 0.15	RMSE: 1464.27 (đồng) MAE: 1236.29 (đồng) MAPE: 0.03 MASE: 0.29
Coffee price & Fuel price	RMSE: 639.96 (đồng) MAE: 379.93 (đồng) MAPE: 0.01 MASE: 0.09	RMSE: 1263.34 (đồng) MAE: 1080.30 (đồng) MAPE: 0.03 MASE: 0.26
Coffee price & Precipitation	RMSE: 1042.87 (đồng) MAE: 623.63 (đồng) MAPE: 0.01 MASE: 0.15	RMSE: 1270.77 (đồng) MAE: 1113.89 (đồng) MAPE: 0.03 MASE: 0.26
Coffee price & Temperature	RMSE: 986.71 (đồng) MAE: 575.06 (đồng) MAPE: 0.01 MASE: 0.14	RMSE: 1338.40 (đồng) MAE: 1169.54 (đồng) MAPE: 0.03 MASE: 0.28
Coffee price & Fuel price & Precipitation	RMSE: 637.34 (đồng) MAE: 379.19 (đồng) MAPE: 0.01 MASE: 0.09	RMSE: 1225.70 (đồng) MAE: 1060.13 (đồng) MAPE: 0.03 MASE: 0.25
Coffee price & Precipitation & Temperature	RMSE: 995.88 (đồng) MAE: 581.15 (đồng) MAPE: 0.01 MASE: 0.14	RMSE: 1190.78 (đồng) MAE: 1030.62 (đồng) MAPE: 0.03 MASE: 0.24
All features	RMSE: 639.63 (đồng) MAE: 382.54 (đồng) MAPE: 0.01 MASE: 0.09	RMSE: 1182.69 (đồng) MAE: 1012.15 (đồng) MAPE: 0.03 MASE: 0.24

This can be explained by two factors: first, the lack of additional features, namely fuel data and weather, when training the time series models; and second, the modest amount of data available, since deep learning models, such as

LSTM and **GRU**, require a great deal of input data to be effective.

Another notable point is that the **RF** model trained with **coffee price** and **fuel price** achieved very close performance (RMSE: 639.96, MAE: 379.93), compared to the best **RF** model which also incorporated the **precipitation** data. This strongly indicated that fuel price is an important factor that affects coffee prices, while weather data has a minor effect. This is consistent with the time series graphs shown in **Fig. 2**, where diesel and coffee prices seem to have similar trends.

The minor effect of weather data could be due to the fact that the tropical weather of Vietnam may be relatively stable in terms of precipitation and temperature over years. However, with climate change, this stability may not persist in the future. At that time, weather data may contribute more to the prediction task, as crop yield can be severely affected by climate change.

V. CONCLUSION

In this study, we used various machine learning models trained with diverse combinations of features to predict coffee price in Vietnam. Our numerical experiments revealed that the **Random Forest (RF) model** trained with **additional fuel data** outperformed other models and demonstrated the best performance. This suggests that other ensemble learning methods, such as AdaBoost, Gradient Boosting, and stacking models, are also promising for crop price prediction. In addition, the integration of a broader range of relevant features, such as crop yield, agricultural and policy news, market trends, geopolitical events, and consumer behaviors, can unveil complex patterns influencing prices and further improve model performance. These factors could be considered for future research on this topic.

REFERENCES

[1] "Agriculture and Food." The World Bank. https://www.worldbank.org/en/topic/agriculture/overview (accessed Oct, 2023).

[2] C. Khoi, "Vietnam exporting agricultural products to nearly 200 countries and territories," in *VnEconomy*, ed, 2023.

[3] N. Trang, "Domestic agricultural product prices increase and decrease unstable because they still depend heavily on exports," in *VOV*, ed, 2022.

[4] T. Vinh, "Market forecasting technology helps eliminate the worry of "good season - devaluation"," in *Vietnamnet*, ed, 2022.

[5] Exporters of Coffee (2021) [Online] Available: https://oec.world/en/profile/hs/coffee

[6] "Vietnam - Employment In Agriculture (% Of Total Employment)." https://tradingeconomics.com/vietnam/employment-in-agriculture-percent-of-total-employment-wb-data.html (accessed Oct, 2023).

[7] "Vietnam: GDP share of agriculture." www.theglobaleconomy.com/Vietnam/Share_of_agriculture/ (accessed Oct, 2023).

[8] "Agricultural product market information." http://thitruongnongsan.gov.vn/vn/default.aspx (accessed Oct, 2023).

[9] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[10] P. D. Khanh. "Lesson 19 - ARIMA model in time series." https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html (accessed Oct, 2023).

[11] R. Wanyama. "Time Series Forecasting for Predicting Store Sales: A Comprehensive Guide." Medium. https://medium.com/@rasmowanyama/title-time-series-forecasting-for-store-sales-a-comprehensive-guide-33346108c2fe (accessed Oct, 2023).

[12] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37-45, 2012.

[13] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[14] B. Panigrahi, K. C. R. Kathala, and M. Sujatha, "A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models," *Procedia Computer Science*, vol. 218, pp. 2684-2693, 2023.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.

[16] S. Suthaharan and S. Suthaharan, "Support vector machine," *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp. 207-235, 2016.

[17] P. S. Nishant, P. S. Venkat, B. L. Avinash, and B. Jabber, "Crop yield prediction based on Indian agriculture using machine learning," in *2020 International Conference for Emerging Technology (INCET)*, 2020: IEEE, pp. 1-4.

[18] "Gasoline Prices Today in Vietnam - Gasoline Prices RON 92, RON 95, Diesel, Kerosene." AutoFun. https://www.autofun.vn/dung-cu/gia-xang-dau (accessed Oct, 2023).

[19] "Storing and Transporting Coffee Beans." CoffeeConcept. https://ddkaffee.com/blogs/kien-thuc-ca-phe/luu-tru-va-van-chuyen-ca-phe-nhan (accessed Oct, 2023).

[20] Historical Weather API [Online] Available: https://open-meteo.com/en/docs/historical-weather-api

[21] "Climate and Crop Growth - How does the weather affect plant growth?" National Geographic. https://media.nationalgeographic.org/assets/activity/assets/climate-crop-growth-1.pdf (accessed Oct, 2023).

[22] "Optimal weather conditions for coffee growing." WeatherPlus. https://weatherplus.vn/dieu-kien-thoi-tiet-toi-uu-cho-trong-ca-phe (accessed Oct, 2023).