Forecasting gold prices using time series analysis

1st Nguyen Nhat Thuong
STAT3013.N11.CTTT
University of Information Technology
Ho Chi Minh city
20522000@gm.uit.edu.vn

2nd Le Quang Hoa
STAT3013.N11.CTTT
University of Information Technology
Ho Chi Minh city
20521331@gm.uit.edu.vn

3rd Kieu Xuan Dieu Huong
STAT3013.N11.CTTT
University of Information Technology
Ho Chi Minh city
20521381@gm.uit.edu.vn

Abstract*—The world in general and Vietnam in particular, gold price fluctuations have a significant impact on many financial activities of Vietnam and the world. The development of a reliable predictive model can provide insights into the volatility, behaviours, and dynamics of the gold price, and can ultimately provide an opportunity to make substantial profits. Currently, in Vietnam, gold price forecasting models are still limited. To contribute to solving the above problem, we study and build a gold price forecasting model, in this study we use 5 models to forecast gold prices based on past prices, the data set includes: include: "date" and "price" daily for the past 20 years. We conduct research and comparison to come up with the best model among Non-Linear Regression, Linear Regression, ARIMA, FBProphet, and LSTM models.*

Keywords—Time series Forecasting, Linear Regression, Non-linear Regression, Arima, FBProphet, LSTM, Forecasting gold prices.

I. Introduction

Currently, with the world economic situation in general and Vietnam being complicated and unpredictable, the World Bank has also issued a warning of a global recession [1], in Vietnam Prime Minister Pham Minh Chinh at the opening session of the 4th session (the XV National Assembly) also commented, "The world situation in 2023 may change rapidly, complicatedly and unpredictably. slowing down; the risk of economic recession and the risks of finance, currency, public debt, energy security, food, information increase" [2] with the above situation that inflation may increase, If the local currency can depreciate, the stock market may decline, gold is one of the choices for financial protection, as well as the best return for our investment. But the big problem here is how to buy gold at a favorable and reasonable price. In that context, although there are five models to predict the gold price, there are still limitations of its own. To contribute to solving the above problem, especially to predict more accurately the gold price in the future, we study and build a gold price prediction model using time series analysis.

II. Related works

1. Time Series Analysis and Forecasting of Gold Price using ARIMA and LSTM Model; Research Dhruvi Sarvaiya and Disha Ramchandani, Department of Computer Engineering; Thadomal Shahani Engineering College, Mumbai, India; In this paper the authors have made use of various ARIMA models of permutations of p, d, q values to conclude that ARIMA model of order (1,1,2) as was deemed fit by the Augmented Dickey-Fuller test. and use LSTM model to improve accuracy by RNN with four layers of interaction structure [3].

2. Implementation multiple linear regression in neural network predict gold price; Research Musli Yanto, Sigit Sanjaya, Yulasmi, Dodi Guswandi 4 and Syafri Arlis 5; 1,4,5 Department of Informatics Engineering, Faculty of Computer Science, Universitas Putra Indonesia YPTK, Indonesia, 2,3 Department of Management, Faculty of Economic and Business, Universitas Putra Indonesia YPTK, Indonesia; In this paper using multiple linear regression method and Artificial Natural Network Parameters are taken from training data and variables are extracted from dataset using correlation [4].

3. Application of ARIMA Model to Forecast Gold Price in Vietnam; Ho Thanh Tri, Phan Dao, Nguyen Van Ninh and Juraj Sipko; In this paper using the model. ARIMA with four models, ARIMA(1,1,1), ARIMA(1,1,5), ARIMA(5,1,1), ARIMA(5,1,5) and result the forecast ARIMA (5,1,5) model is excellent in this case with the forecast error is about 3.46% [5].

4. Forecasting Gold Prices Using Multiple Linear Regression Method; Research Z. Ismail 1, A. Yahya 2

and A. Shabri 3; 1, 3 Department of Mathematics, Faculty of Science, 2 Department of Basic Education, Faculty of Education; University Technology Malaysia, 81310 Skudai, Johor Malaysia; In this paper using Linear Regression model to study gold price from London with some factors affecting gold price and factors used as independent variable [6].

III. Methodology

Research method is very important, so before the research we proceed to build a framework diagram of the research method.

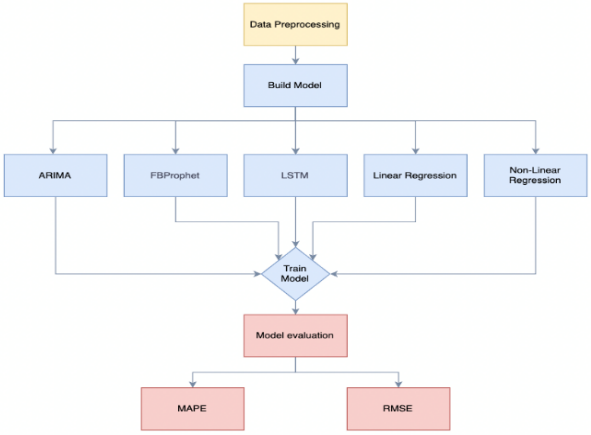

Figure 1: Research framework.

In the framework we divide into 3 phases: the first stage is preprocessing the data, the second stage is building the model and training the model, and the third stage is evaluating the model.

A. Data processing

As mentioned above, our article will provide a time series forecast based on how accurately we can predict the price of gold. For this purpose, we requested a dataset that provides gold price data, the dataset we chose was selected from Kaggle, which is one of the reputable large communities and platforms for datasets. Our initial raw data consists of 19 columns of data, and then we first process and select 2 columns, it includes 'Date' column Y-M-D and column 'Prices' (VND), after processing for data management, the dataset consists of 5289 lines and the data is entered by date from 01/01/2002 to 04/08/2022.

Figure 2: Gold price data chart after processing.

B. Modeling

1) Non-linear Regression

Nonlinear time-series analysis is a group of techniques used to gather dynamical data regarding a data set's progression of values. This framework relies critically on the concept of reconstruction of the state space of the system from which the data are sampled, and the information regarding their temporal behaviours is accessible through a single variable's time series. The aim of this method is to forecast the future direction of the time series, and in some cases, to reconstruct the motion equations. However, in practice, there are several challenges that limit the effectiveness of this strategy, such as whether the signal completely and precisely captures the dynamics and whether it contains noise. Additionally, the numerical algorithms that we use to instantiate these ideas are not perfect; they involve approximations, scale parameters, and finite-precision arithmetic, among other things. Nevertheless, nonlinear time series analysis has been used for thousands of real data sets from a variety of systems.
A simple non-linear regression model is expressed as follows:

$$y = f(x, \beta) + \varepsilon \quad (1)$$

Where:
X is vector of P predictors
β is vector of k parameters
F is the known regression function
ε is the error

2) Linear Regression

Linear regression is a statistical method for determining the value of a dependent variable from an independent variable. Linear regression is also a measure of the connection between two variables. A dependent variable is predicted using this modeling approach based on one or more independent factors. Of all statistical methods, linear regression analysis is the one that is most frequently utilized.
a. Descriptive - It aids in evaluating the degree of correlation between the outcome (the dependent variable) and the predictor factors.
b. Adjustment - It adjusts the influence of variables or confounders.
c. Predictors - It aids in the estimation of significant risk variables that have an impact on the dependent variable.
d. Extent of prediction: - It aids in determining how much a change in the independent variable of one "unit" might impact the dependent variable.
e. Prediction - It assists in quantifying new cases.
Following the formation of the predictive network pattern, the topic will turn to linear regression, which is likewise a relatively straightforward technique for examining the relationship between a predictor variable and a response variable.

$$y = \beta_0 + \beta_1 + e \quad (2)$$

To create multiple linear regression equations, the variables were extracted from the dataset using correlation, and the parameters were taken from the training data. The number r, also known as the linear correlation coefficient, quantifies the direction and intensity of a relationship between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (3)$$

3) ARIMA

ARIMA, or Autoregressive Integrated Moving Average is known as a time series forecasting model, was created by George Box and Gwilyn Jenkins in 1970. This model is widely used in the finance field. Therefore, to predict the golden price in future ARIMA is one of our choices. The ARIMA model was created by combining the Moving Average (MA) and Auto-Regressive (AR) models, both of which predict future values using lagged data [7] .
Auto Regressive (AR only) model is is a stationary process which Y_t depends only on its own lags..

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \epsilon_t \quad (4)$$

Where:
α: is the intercept term.
Y_{t-1}: is the lag1 of the series.
β_1 : is the coefficient of lag1.

Moving Average (MA only) model is a stationary process which is depends on the lagged forecast errors.

$$Y_t = \alpha + \epsilon_t + \emptyset_1 \epsilon_{t-1} + \emptyset_2 \epsilon_{t-2} + \cdots + \emptyset_q \epsilon_{t-q} \quad (5)$$

Where:
α: is the intercept term.
ϵ_t : is the error.

The errors ϵ_t and ϵ_{t-1} are from the equation:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_0 Y_0 + \epsilon_t \quad (6)$$

$$Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} + \ldots + \beta_0 Y_0 + \epsilon_t \quad (7)$$

From (4) and (5), we have ARIMA with order p, d, q can be estimated by the following formula:

$$Y_T = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} \epsilon_t + \emptyset_1 \epsilon_{t-1} + \emptyset_2 \epsilon_{t-2} + \cdots + \emptyset_q \epsilon_{t-q}$$

4) FBProphet

Facebook Prophet is an open-source program created by the company that analyzes time series data using a decomposable additive model. It takes holidays into account and often fits nonlinear data with seasonality on an annual, monthly, and daily basis. Prophet fits a variety of linear and nonlinear time functions as components, combining them into the following equation [9]:

$$y(t) = g(t) + s(t) + h(t) + e(t) \quad (8)$$

where:
y(t): predictions (forecast).
g(t): trend alludes to changes throughout a significant stretch of time
s(t): seasonality weekly, daily, yearly.
h(t): holidays
e(t): error term represents any surprising changes not obliged by the model

The Facebook prophet has multiple trends, seasonality, change points, and holiday parameters which needs to be tuned for better results. The goal of Prophet is to tease out the signal in a data set and forecast that signal into the future. The approach to getting at that signal of Prophet is breaking down the signal into three pieces trend, season, and holiday.
The logistic growth model:

$$g(t) = \frac{C}{1 - e^{-k(t-m)}} \quad (9)$$

where:
C: carry capacity
k: growth rate
m: an offset parameter

The linear growth model:

$$y = \begin{cases} \beta_0 + \beta_1 x & x \leq c \\ \beta_0 - \beta_2 x + (\beta_1 + \beta_2) x & x > c \end{cases} \quad (10)$$

5) LSTM

Long short-term memory (LSTM) is an improved network from the RNN model to overcome the vanishing gradient. The structure of the LSTM network has four layers that interact with each other [10].

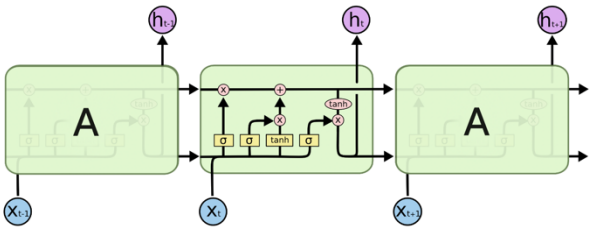

Figure 3: Construction of a node in an LSTM network.

And the above process is done by the following (11) – (16) equations [11]:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (11)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (12)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (13)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (14)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (15)$$

$$h_t = o_t * \tanh(C_t) \quad (16)$$

In addition, the LSTM is capable of removing or adding information necessary for the cell state, which is regulated by the gates. The fact that the LSTM has good memorization makes it possible to predict time series and avoid the problem of remote dependencies. But that also makes the LSTM model complicated and difficult to understand as well as the speed will be much slower than RNN.

C. Algorithm

1) ARIMA

Stage 1: Raw Data
The data will be download on Kaggle in file csv and uploaded to google drive and use google Colab to connect.

Stage 2: Data split
The forecasting model is trained using the daily gold price data from January 2002 till March 2018 and tested using data from March 2018 till June 2022. The model will be split in 90- 10, 80-20 and 70-30.

Stage 3: Check for stationary
To ARIMA we must check the stationary of the data by ADF test (Augmented Dickey-Fuller) test. The test is for the null hypothesis that there is a unit root (non-stationary). If the test is a non-stationary, we will use differencing to make a non-stationary time series to be stationary.

ADF Statistic: -0.370760
p-value: 0.914868

In addition, we use decomposition to see the components of the dataset include trend, seasonal and residual.

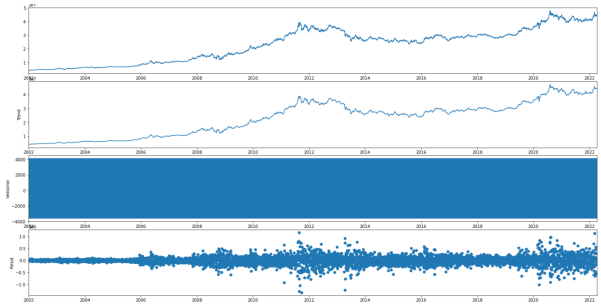

Figure 4: The component of the data.

Stage 4: Determine ARIMA models parameters p, d, q
In stage 3, we find that our data is a non-stationary so that we use auto_arima to find (p,d,q) for the best model without using PACF plot and ACF plot.

Stage 5: Fit the model
After deciding the parameters of p,d,q we fit the model in Python.

ARMA Model Results

Dep. Variable:	Prices	No. Observations:	4760
Model:	ARMA(6, 2)	Log Likelihood	-65849.718
Method:	css-mle	S.D. of innovations	246109.654
Date:	Tue, 03 Jan 2023	AIC	131719.436
Time:	14:55:45	BIC	131784.116
Sample:	01-01-2002 - 03-30-2020	HQIC	131742.161

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.061e+07	nan	nan	nan	nan	nan
ar.L1.Prices	-0.8234	nan	nan	nan	nan	nan
ar.L2.Prices	0.9396	2.38e-05	3.94e+04	0.000	0.940	0.940
ar.L3.Prices	0.8837	3.07e-05	2.88e+04	0.000	0.884	0.884
ar.L4.Prices	-0.0183	0.000	-60.784	0.000	-0.019	-0.018
ar.L5.Prices	0.0112	nan	nan	nan	nan	nan
ar.L6.Prices	0.0071	nan	nan	nan	nan	nan
ma.L1.Prices	1.8326	0.007	278.826	0.000	1.820	1.846
ma.L2.Prices	0.8957	0.007	136.771	0.000	0.883	0.909

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0000	-0.0000j	1.0000	-0.0000
AR.2	-1.0099	-0.2782j	1.0475	-0.4572
AR.3	-1.0099	+0.2782j	1.0475	0.4572
AR.4	2.4155	-4.2433j	4.8826	-0.1676
AR.5	2.4155	+4.2433j	4.8826	0.1676
AR.6	-5.3978	-0.0000j	5.3978	-0.5000
MA.1	-1.0230	-0.2643j	1.0566	-0.4598
MA.2	-1.0230	+0.2643j	1.0566	0.4598

Figure 5: ARIMA model summary.

2) FBProphet

Stage 1: Raw data process
Because in Prophet is built with two main features (ds has a date format, timestamping, represents a quantitative value and y representing a measure that we predict) we must rename these columns and change the format of ds column.

Stage 2: Data split
The forecasting model is trained using the daily gold price data from January 2002 till March 2018 and tested using data from March 2018 till June 2022 (90-10)

Stage 3: Create and fit the model
The model was then fitted using a fit( ) method.In the table ds, as we know, is the time series data. yhat is the prediction, yhat_lower and yhat_upper are the uncertainty levels (it basically means the prediction and actual values can vary within the bounds of the uncertainty levels). Next is trend, which depicts the long-term expansion, contraction, or stagnation of the data;

trend lower and trend upper indicate the degrees of uncertainty.

Stage 4: Data visualize

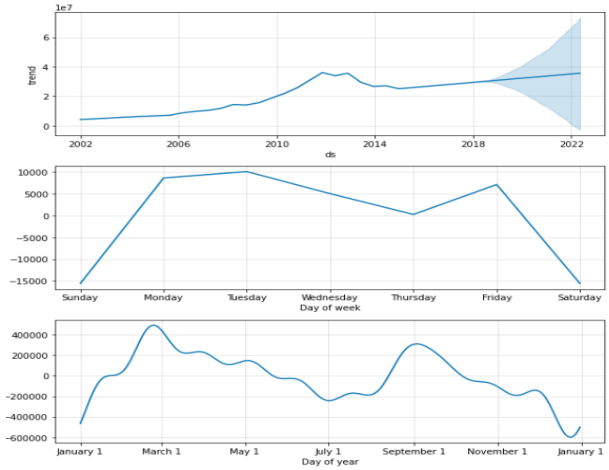

Figure 6: The trends and seasonality (in a year, week, day) of the time series data.

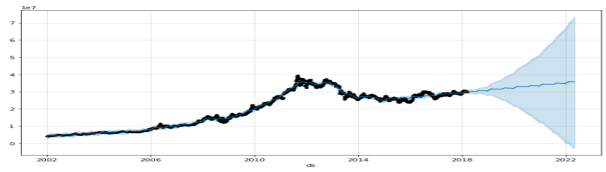

Figure 7: Compare actual and project value with train data of 70% and test data of 30%

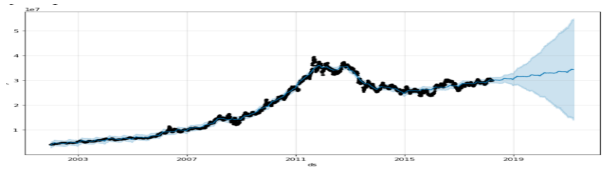

Figure 8: Compare actual and project value with train data of 80% and test data of 20%

Figure 9: Compare actual and project value with train data of 90% and test data of 10%

3) LSTM

Different neural network types can be created by combining various elements, such as network topology, training process, etc. We have taken into account long short-term memory and recurrent neural networks for this experiment.
In this section, we'll go over our system's methodology. Our system is divided into the following stages:

Stage 1: Raw Data

During this period, I downloaded the data from Kaggle and processed the data truncation with Excel then uploaded to google drive and connect it to google colab.

Stage 2: Data Processing
Format the date column structure and proceed to DataFrame data.

Stage 3: Split Dataset
In this step we will proceed to divide the data into parts to train and test, we choose 3 rates first 70% train 30% test, second 80% train 20% test, finally 90% train and 10% test.

Stage 4: Model Building
This stage involves feeding data into a neural network that has been trained to forecast biases and random weights. A sequential input layer, two LSTM layers, two Dropout layers for rote learning, and a dense output layer with linear activation functions are all included in our LSTM model. The following is the code for the neural network used by Keras:

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 15, 64)	16896
dropout (Dropout)	(None, 15, 64)	0
lstm_1 (LSTM)	(None, 15, 64)	33024
dropout_1 (Dropout)	(None, 15, 64)	0
lstm_2 (LSTM)	(None, 32)	12416
dropout_2 (Dropout)	(None, 32)	0
dense (Dense)	(None, 25)	825
dropout_3 (Dropout)	(None, 25)	0
dense_1 (Dense)	(None, 1)	26

Figure 10: Summary LSTM model

Stage 5: Model Learning
In this step, we proceed to machine learning with epochs of 100 and save only the best model with the lowest loss index.

Stage 6: Finally we proceed to plot the 3 cases

Figure 11: Compare actual and project value with train data of 70% and test data of 30% with MAPE of 5.77%

Figure 12: Compare actual and project value with train data of 80% and test data of 20% with MAPE of 4.83%

Figure 13: Compare actual and project value with train data of 90% and test data of 10% with MAPE of 5.52%

D. Metrics

We utilize the Root Mean Square Error to assess the system's effectiveness (RMSE). The RMSE value is used to reduce error or the discrepancy between the desired and obtained output values. The root mean square error (RMSE) is the sum of all square errors.

$$RMSE = \sqrt{\frac{1}{n}\sum(\hat{Y}_i - Y_i)^2} \qquad (17)$$

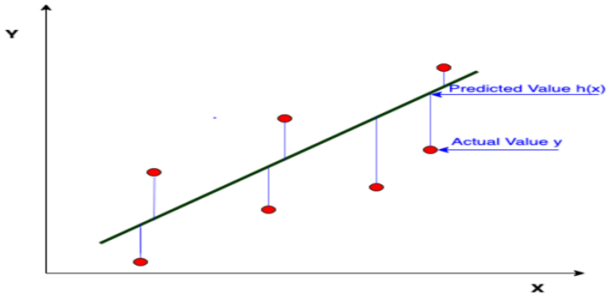

Figure 14. RMSE value calculate

The RMSE error metric is widely used and is a great all-purpose error metric for numerical forecasts.RMSE amplifies and harshly penalizes big errors in comparison to the analogous Mean Absolute Error.
The mean absolute percentage error, or MAPE, is expressed as a percentage. The accuracy of a forecasting system is statistically measured by the mean absolute percentage error (MAPE).

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (18)$$

MAPE-value	Accuracy of forecast
Less than 10%	Highly Accurate Forecast
11% to 20%	Good Forecast
21% to 50%	Reasonable Forecast
More than 51%	Inaccurate Forecast

Figure 15: Interpretation of MAPE Results for Forecasting Accuracy.

IV. Result

In this evaluation, five distinct models were assessed, to be specific LSTM, ARIMA, LN, NLN, FBProphet for the time series examination on 7-3, 8-2, 9-1 training and testing. RMSE and MAPE scores were used to assess model execution
In the table 1 show the forecasting metrics results for examining test data of the dataset. From the outcomes, one can without much of a stretch see that LSTM has better execution compared to different models for all evaluation metrics. LSTM has performed well in all train-test set. The RMSE values are lower than ARIMA and FBProphet. The best MAPE score is 5.83% which is for LSTM by the data at the rate 8-2 train test. Best RMSE is 2149828.17 for 8-2 train-test split by LSTM.
To assess overall model performance. The deep learning-based model LSTM outperform ARIMA, and FBProphet in term of mean absolute percentage error. At the same time LSTM has the lowest error in root mean squared error metrics compared with other models.

Model	Train-test	RMSE	MAPE (%)
Linear Regression	**10-0**	4777274.13	16.65
Non-Linear Regression	**10-0**	4545507.81	21.78
ARIMA	**7-3**	8727545.1	16.27
	8-2	8984086.99	18.05
	9-1	4149475.21	8.73
FBProphet	**7-3**	14211069.53	38.73
	8-2	5900284.21	12.19
	9-1	2509956.14	5.16
LSTM	**7-3**	2452811.30	5.77
	8-2	2149828.17	4.83
	9-1	2438809	5.52

V. Conclusion

Global commerce, raw material costs, stalled shipping, plummeting stock markets, a high unemployment rate, and declining currencies are just a few of the effects of the economic boom that might render an intermediate bank worthless. Stocks influenced by dividends are frequently held by well-known corporations that dominate their respective markets. In a recession, gold and dividend-paying stocks have historically been regarded as the finest investments. This study used ARIMA, FBProphet, LSTM, LR, and NLR models to analyse and predict the price of gold. In terms of MAPE and RMSE assessment metrics, LSTM outperforms all other models for most of the train test set. The trend analysis demonstrates that ARIMA and FBProphet have significant MAPE errors because they were unable to accurately anticipate lower values. The techniques used produced positive outcomes. However, it limits the scope of our analysis to the model's suitability, which may also be enhanced by spending more time examining hyper-optimization strategies.

References

[1] Overview of the world's economic situation for q3 and all of 2022.

[2] Assessing the risk of world economic recession in 2023, the Prime Minister proposed solutions to cope.

[3] Time Series Analysis and Forecasting of Gold Price using ARIMA and LSTM Model; Dhruvi Sarvaiya - Disha Ramchandani; Ijraset International Journal For Research in Applied Science and Engineering Technology, 2022.

[4] Implementation multiple linear regression in neural network predict gold price; Musli Yanto, Sigit Sanjaya, Yulasmi, Dodi Guswandi, Syafri Arlis; Indonesian Journal of Electrical Engineering and Computer Science, 2021.

[5] Application of ARIMA Model to Forecast Gold Price in Vietnam; Ho Thanh Tri, Phan Dao, Nguyen Van Ninh and Juraj Sipko; 11th International Days of Statistics and Economics, 2017.

[6] Forecasting Gold Prices Using Multiple Linear Regression Method; Z. Ismail, A. Yahya and A. Shabri; American Journal of Applied Sciences 6, 2009.

[7] Using ARIMA model to analyse and predict bitcoin price; Yang Si; BCP Business & Management 12, 2022.

[8] On Forecasting Nigeria's GDP: A Comparative Performance of Regression with ARIMA Errors and ARIMA Method; Christogonus Ifeanyichukwu Ugoh, Udochukwu Victor Echebiri, Gabriel Olawale Temisan, Johnpaul Kenechukwu Iwuchukwu, Emwinloghosa Kenneth Guobadia; International Journal of Mathematics and Statistics Studies 9, 2022.

[9] Evaluation of Time Series Forecasting Models for Estimation of PM2.5 Levels in Air; Satvik Garg, Himanshu Jindal; 6th I2CT 2021.

[10] Understanding LSTM Networks.

[11] A CNN–LSTM model for gold price time-series forecasting; Ioannis E. Livieris, Emmanuel, Panagiotis Pintelas; SI 6, 2020