UNIVERSITY OF INFORMATION TECHNOLOGY

FACULTY OF INFORMATION SYSTEMS

Using Time Series Analysis To Forecast Crude Oil Price

- PHẠM THỊ THUỲ DƯƠNG - 20521221
- BÙI QUANG ANH - 18520441
- BÙI THU HÀ - 20521266

Using Time Series Analysis To Forecast Crude Oil Price

1st Pham Thi Thuy Duong
STAT3013.N11.CTTT
University of Information Technology
Ho Chi Minh city
20521221@gm.uit.edu.vn

2nd Bui Quang Anh
STAT3013.N11.CTTT
University of Information Technology
Ho Chi Minh city
18520441@gm.uit.edu.vn

3rd Bui Thu Ha
STAT3013.N11.CTTT
University of Information Technology
Ho Chi Minh city
20521266@gm.uit.edu.vn

***Abstract* —** With the development of time series forecasting, recently there have been many scientific research papers on the fields of research and finance using this method. In this study, we use Linear Regression, Non-Linear Regression, ARIMA, SARIMA, Prophet and LSTM models to compare and forecast the volatility of crude oil close prices from 2016 to 2022.

***Keywords* —** *Crude oil prices, Linear Regression, Non-Linear Regression, Prophet, ARIMA, SARIMA, LSTM models*

I. INTRODUCTION

Crude oil is one of the largest energy supplies in the world, which plays a significant role in current life. Due to its great importance in the world economy, forecasting the price of crude oil is also an equally important factor to contribute to improving the decisions of governments. Moreover, because of many impact factors such as wars, epidemics… the world economy is also greatly affected, leading to crude oil also affected. Therefore, recently, research models of crude oil price prediction have become more and more popular and especially interesting.

II. RELATED WORK

Based on previous research on the methods of computational models for us as a basis to make predictions about crude oil prices.

Adalat Muradov, Yadulla Hansanli, Nazim Hajiyev used Trend, ARIMA, Holt, ARCH/GARCH forecasting models to forecast oil prices in the period 2018-2020. The results show that the rise of 1% increase in the average price of Brent and WTI oil in the world market leads to an increase in the price of AzerLight oil by about 0.95% per barrel. [1]

Yanhui Chen, Kaijian He, Geoffrey K.F. Tso using the deep learning model to capture the unknown complex nonlinear characteristics of the crude oil price movement. They further propose a new hybrid crude oil price forecasting model based on the deep learning model. Using the proposed model, major crude oil price movement is analyzed and modeled. The performance of the proposed model is evaluated using the price data in the WTI crude oil markets.

They concluded that DBN based model tracks the actual returns more closely than the LSTM model. LSTM produces forecasts with significantly larger fluctuation range, with very large value of transient events. In the fast changing crude oil data, the LSTM model may not adapt to changes fast enough to incorporate the new changes available. The market is mostly dominated with shorter term memory behavior. [2]

W. Ahmad, M. Aamir, U. Khalil, M. Ishaq, N. Iqbal, M. Khan, they proposed a new hybrid method that combines the median ensemble empirical mode decomposition and group method of data handling (MEEMD-GMDH) to reduce mood splitting problems and forecast crude oil price.

In Predictive Performance of Single Models, the GMDH model attains the smallest value (better performance) on the metrics (RMSE, MAE, and MAPE)and indicated that the ARIMA model performed worst because the classical econometric and time series method does not perform well for nonlinear time series. The forecast rating performance (RMSE) for both crude oil markets are shown in order of GMDH -ANN model attained second- The ARIMA model takes the third place in terms of predictive performance.

In Predictive Performance of Hybrid Models, The MEEMD-based model is statistically superior to the ANN and ARIMA models, and their p-values are less than < 0.01 for both markets, which indicates the superiority of the MEEMD-GMDH model. [3]

Öznur Öztunç Kaymak and Yiğit Kaymak have improved a model to give more accurate forecasts of crude oil prices and compare with ANNs (artificial neural network), SVM (support vector machine) models.

The results show that their model outperforms other forecasting models, especially ANN and SVM (the model is known to produce better results than traditional methods of predicting oil prices). In terms of both RMSE value and MAE value, the proposed model gives better results than other models. [4]

Quanying Lu, Shaolong Sun, Hongbo Duan & Shouyang Wang proposed a variable selection and machine learning framework that combines variable selection (BMA) and forecasting methods (LSTM) to forecast oil prices and compare forecast performance. Its with other new and major variable selection methods (elastic-net and branch and blade Lasso). Moreover, compared with other popular benchmark forecasting methods (RW, ARMA, MLP, RBFNN, GRNN, ENN, WNN, ELM) and the results obtained are as follows.

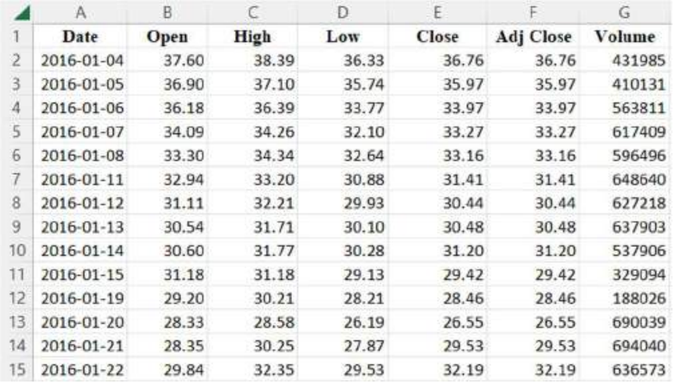
Among variable selection machine learning integration models, the BMA-LSTM integration model performed the best, followed by Spike and Slab LASSO-LSTM and GLMNET-LSTM.

Selection method-variable machine learning is the best direction prediction performance, and it can also be seen that ARIMA's direction predicts performance is worst. [5]

Kian Zhang and Min Hong, in their study, constructed an LSTM (short for Long Short-Term Memory neural network) model to conduct the forecasting crude oil price based on data from February 1986 to May 2021; an ANN (short for Artificial Neural Network) model and a typical ARIMA (short for Autoregressive Integrated Moving Average) model are taken as the comparable models. The results show that, first, the LSTM model has strong generalization ability, with stable applicability in forecasting crude oil prices with different timescales. Second, as compared to other models, the LSTM model generally has higher forecasting accuracy for crude oil prices with different timescales. Third, an LSTM model-derived shorter forecast price timescale corresponds to a lower forecasting accuracy. [6]

III. DATA AND METHODOLOGY

A. Data sources



	A	B	C	D	E	F	G
1	**Date**	**Open**	**High**	**Low**	**Close**	**Adj Close**	**Volume**
2	2016-01-04	37.60	38.39	36.33	36.76	36.76	431985
3	2016-01-05	36.90	37.10	35.74	35.97	35.97	410131
4	2016-01-06	36.18	36.39	33.77	33.97	33.97	563811
5	2016-01-07	34.09	34.26	32.10	33.27	33.27	617409
6	2016-01-08	33.30	34.34	32.64	33.16	33.16	596496
7	2016-01-11	32.94	33.20	30.88	31.41	31.41	648640
8	2016-01-12	31.11	32.21	29.93	30.44	30.44	627218
9	2016-01-13	30.54	31.71	30.10	30.48	30.48	637903
10	2016-01-14	30.60	31.77	30.28	31.20	31.20	537906
11	2016-01-15	31.18	31.18	29.13	29.42	29.42	329094
12	2016-01-19	29.20	30.21	28.21	28.46	28.46	188026
13	2016-01-20	28.33	28.58	26.19	26.55	26.55	690039
14	2016-01-21	28.35	30.25	27.87	29.53	29.53	694040
15	2016-01-22	29.84	32.35	29.53	32.19	32.19	636573

Figure 1 – The data of Crude Oil

The data is about the crude oil price from 4/1/2016 to 12/12/2022.

Source data: Yahoo! Finance.

In this research, we will predict the close price of 30 days later (13/12/2022 → 11/1/2023).

B. Linear Regression

Linear regression is a data analysis technique that predicts the value of a variable based on the value of another variable.

The formula for a simple linear regression is: [7]

$$y = \beta_0 + \beta_1 X + \epsilon$$

- y is the predicted value of the dependent variable (y) for any given value of the independent variable (x).
- β_0 is the intercept, the predicted value of y when the x is 0.
- β_1 is the regression coefficient – how much we expect y to change as x increases.
- X is the independent variable (the variable we expect is influencing y).
- ϵ is the error of the estimate, or how much variation there is in our estimate of the regression coefficient.

C. Non-Linear Regression

Non-Linear regression is a form of regression in which the dependent or criterion variables are modeled as a non-linear function of model parameters and one or more independent variables.

A simple nonlinear regression model is expressed as follows: [8]

$$Y = f(X, \beta) + \epsilon$$

X: is a vector of P predictors
β: is a vector of k parameters
F(-): is the known regression function
ϵ : is the error term

D. Prophet

Prophet is a method for predicting time series data that uses an additive model to suit non-linear trends with seasonality that occurs annually, monthly, daily, and on weekends as well as during holidays. Strongly seasonal time series and multiple seasons of historical data are ideal for it. Prophet typically manages outliers well and is robust to missing data and changes in the trend.

Prophet can be considered a nonlinear regression model of the form: [9]

$$y_t = g(t) + s(t) + h(t) + \varepsilon_t$$

Where $g(t)$ describes a piecewise-linear trend (or "growth term"), $s(t)$ describes the various seasonal patterns, $h(t)$ captures the holiday effects, and ε_t is a white noise error term.

Prophet is open-source software released by Facebook's Core Data Science team. It is available for download on *CRAN* and *PyPI.*

- So, Prophet is the Facebook' open-source tool for making time series predictions.
- Prophet decomposes time series data into trend, seasonality, and holiday effect.

- **Trend** models non periodic changes in the time series data.
- **Seasonality** is caused due to the periodic changes like daily, weekly, or yearly seasonality.
- **Holiday effect** which occurs on irregular schedules over a day or a period of days.
- **Error terms** are what is not explained by the model.

Prophet also imposes the strict condition that the input columns must be named as **ds (the time column)** and **y (the metric column)**.

E. ARIMA

Autoregressive Integrated Moving Average (ARIMA) models is widely used in demand forecasting. ARIMA models are generally denoted as ARIMA (p,d,q) where p is the order of autoregressive model, d is the degree of differencing, and q is the order of moving-average model. ARIMA models use differencing to convert a non-stationary time series into a stationary one, and then predict future values from historical data. These models use “auto” correlations and moving averages over residual errors in the data to forecast future values. [10]

So how does ARIMA model work?

- AutoRegressive - AR(p) is a regression model with lagged values of y, until p-th time in the past, as predictors. Here, p = the number of lagged observations in the model, ε is white noise at time t, c is a constant and φs are parameters.

$$\hat{y}_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

- Integrated I(d) - The difference is taken d times until the original series becomes stationary. A stationary time series is one whose properties do not depend on the time at which the series is observed.

$By_t = y_{t-1}$ where B is called a backshift operator

Thus, a first order difference is written as

$$y'_t = y_t - y_{t-1} = (1 - B)y_t$$

In general, a d th-order difference can be written as

$$y'_t = (1 - B)^d y_t$$

- Moving average MA(q) - A moving average model uses a regression-like model on past forecast errors. Here, ε is white noise at time t, c is a constant, and θs are parameters

$$\hat{y}_t = C + \theta_1 \in_{t-1} + \theta_2 \in_{t-2} + \cdots + \theta_q \in_{t-q}$$

F. SARIMA

SARIMA stands for Seasonal-ARIMA, and it includes seasonality contribution to the forecast. The importance of seasonality is quite evident and ARIMA fails to encapsulate that information implicitly.

The Autoregressive (AR), Integrated (I), and Moving Average (MA) parts of the model remain as that of ARIMA. The addition of Seasonality adds robustness to the SARIMA model. It’s represented as:

$$SARIMA\ (p,d,q)(P,D,Q)_m$$

where m is the number of observations per year. We use the uppercase notation for the seasonal parts of the model, and lowercase notation for the non-seasonal parts of the model. [11]

G. LSTM

Long short-term memory (LSTM) was first mentioned in 1997 by Sepp Hoch Reiter and Jürgen Schmid Huber. By using constant error carousel (CEC) units, LSTM treats the exploding and vanishing gradient problems. The preliminary version of the LSTM block incorporated cells, input, and output gates. It is famous for its design, which is a specific type of recurrent neural network, utilizing many different real-world applications that the standard version doesn’t. An RNN’s biggest problem is that it only holds the previous state information, causing the vanishing gradient problem. This problem has been solved by LSTM.

LSTM was constructed to avoid the issue of long-term dependencies. Remembering information for a long duration is basically its default behavior. All RNNs have repeating chain modules of the neural network. This repeating module is a simple structure, such as the tanh layer in an RNN. LSTM also has an identical chaining structure instead of having a single neural network layer. [12]

Equation of LSTM: [13]

- Equation for LSTM forget gate:

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right)$$

where $x^{(t)}$ is the current input vector and $h^{(t)}$ is the current hidden layer vector containing the outputs of all the LSTM cells, and b^f, u^f and W^f are respectively biases, input weights and recurrent weights for forget gates.

- Equation for LSTM internal state update:

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} = g_i^{(t)} \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right)$$

where b, u and W respectively denote the biases, input weights and recurrent weights into the LSTM cell

- Equation for LSTM input:

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right)$$

- Equation for LSTM output:

Output $h_i^{(t)}$ of the LSTM cell can be shut off via the output gate $q_i^{(t)}$ which also uses a sigmoid unit for gating

$$h_i^{(t)} = \tan h\left(s_i^{(t)}\right) q_i^{(t)}$$

$$q_i^{(t)} = \sigma\left(b_i^0 + \sum_j U_{i,j}^0 x_j^{(t)} + \sum_j W_{i,j}^0 h_j^{(t-1)}\right)$$

b^0, U^0 and W^0 are biases, input weights and recurrent weights.

IV. MODEL SETTING

A. Data analysis

1. Close Price

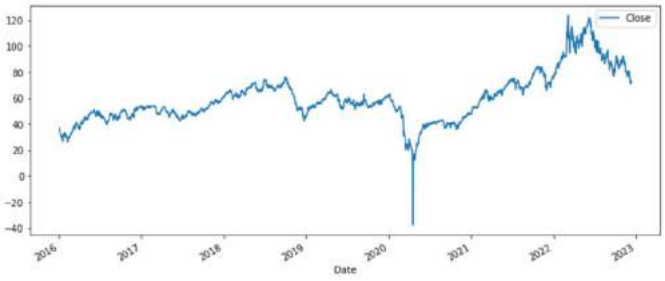


Figure 2 – Graph of close price

In the Close plot, we can easily see the movement of the Close price from 2016 to 2022.

B. Linear Regression

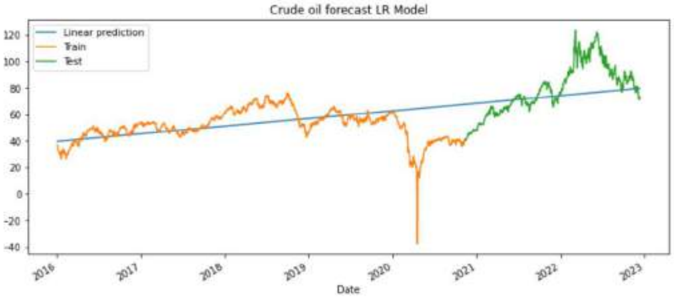


Figure 3 – Result of model Linear Regression 7-3

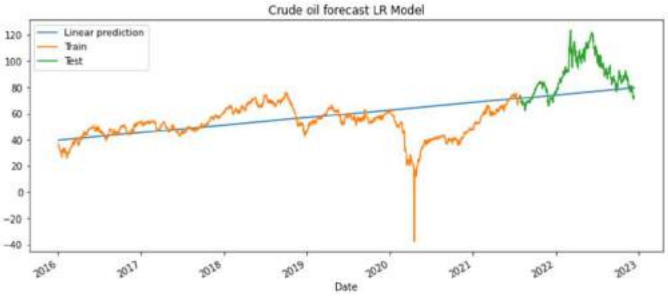


Figure 4 – Result of model Linear Regression 8-2

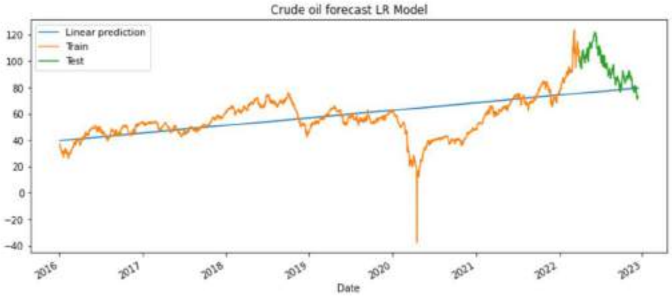


Figure 5 – Result of model Linear Regression 9-1

In 3 models train test 7-3, 8-2 and 9-1, it can be seen that the forecast of the Linear Regression model is steadily increasing.

C. Non-Linear Regression

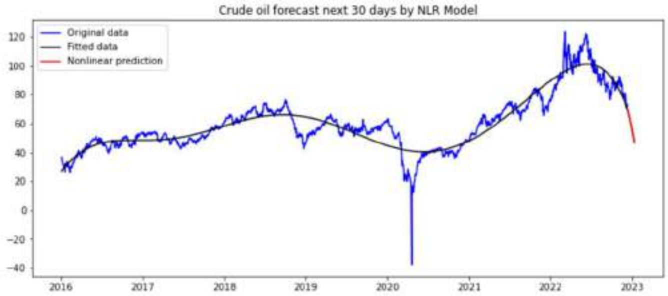


Figure 6 – Result of model Non-linear Regression

In the Non-Linear Regression model, it can be seen that the prediction of this model for the next 30 days is decreasing.

D. Prophet

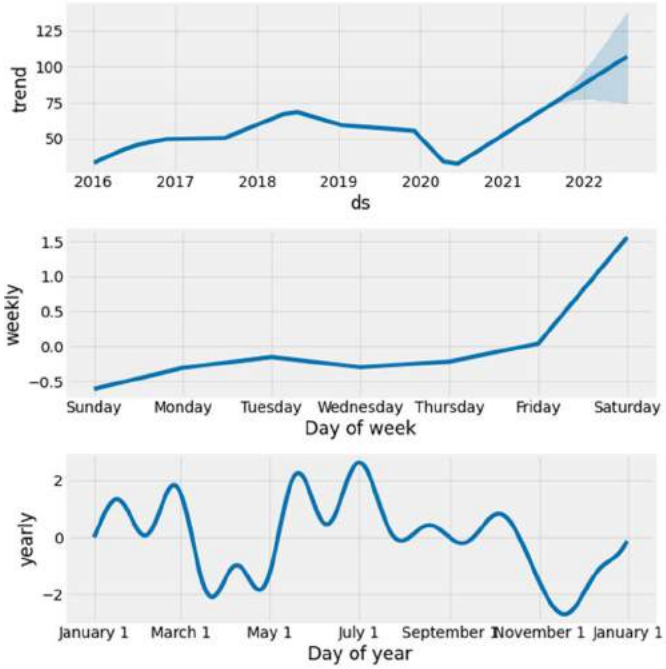
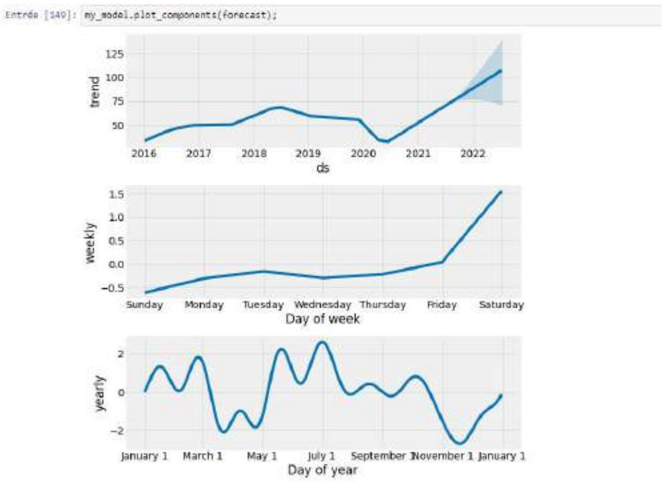


Figure 7 – Result of model Prophet 7-3



```
Entrée [149]: my_model.plot_components(forecast);
```

Figure 8 – Result of model Prophet 8-2

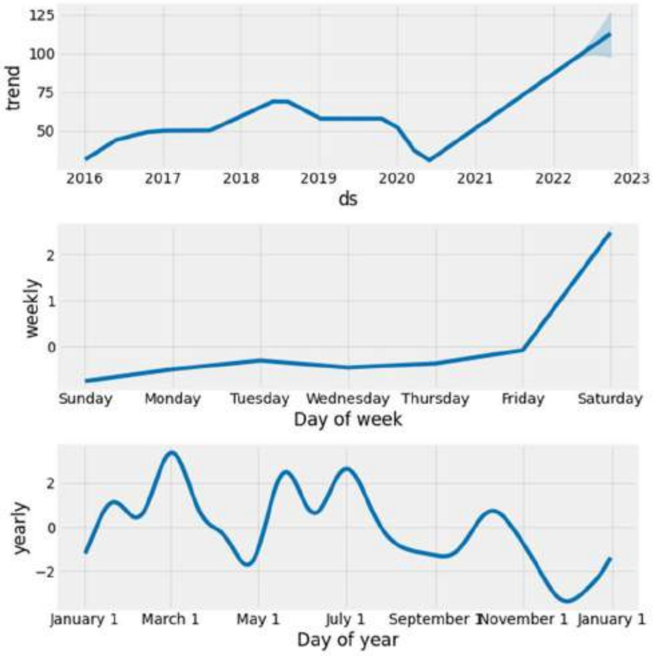


Figure 9 – Result of model Prophet 9-1

- Close price is not linear, lowest in between 2020 and 2021 and will be highest by 2023. According to the trend model we can see the price will increase linear from the middle of 2021 onwards.
- About the week, we can see Close prices will increase sharply at the weekend and the lowest at the beginning of the week specifically on the 2nd day.
- According to the year, we can see that Close price increases and decreases irregularly.

E. ARIMA

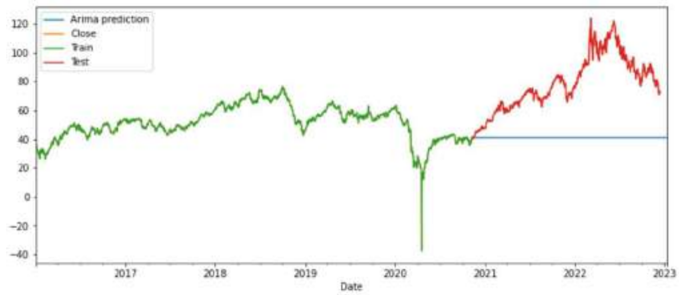


Figure 10 – Result of model ARIMA 7-3

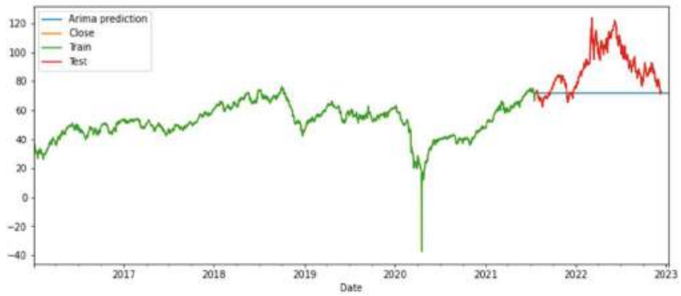


Figure 11 – Result of model ARIMA 8-2

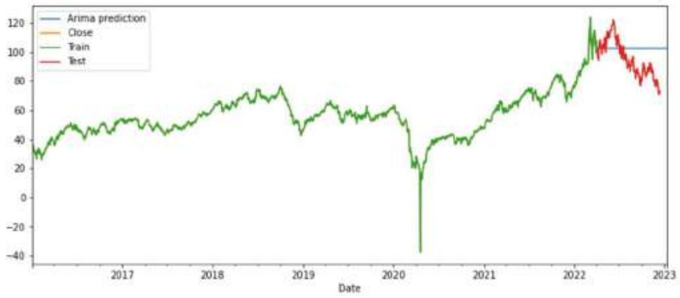


Figure 12 – Result of model ARIMA 9-1

In 3 models train test 7-3, 8-2 and 9-1, it can be seen that the forecast of the ARIMA model is neither increasing nor decreasing.

F. SARIMA

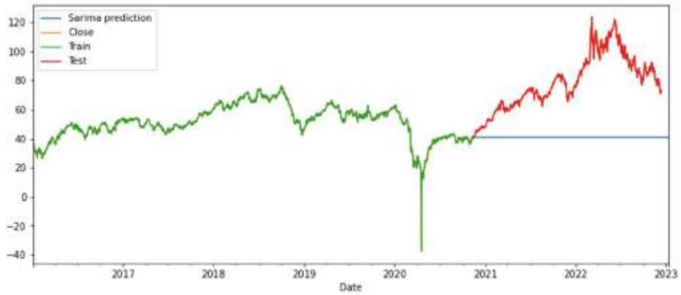


Figure 13 – The result of model SARIMA 7-3

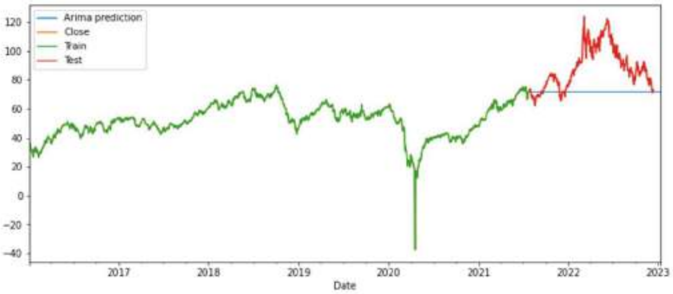


Figure 14 – Result of model SARIMA 8-2

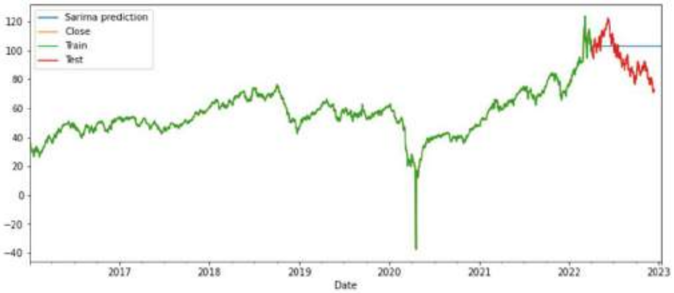


Figure 15 – Result of model SARIMA 9-1

The same as model ARIMA, when we add seasonal forecast, 3 train test models 7-3, 8-2 and 9-1 also gave the same results, neither increasing nor decreasing.

G. *LSTM*

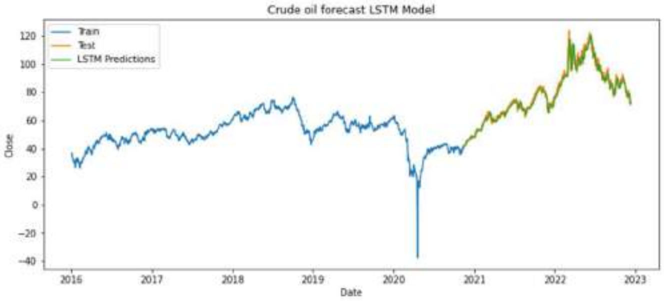


Figure 16 – Result of model LSTM 7-3

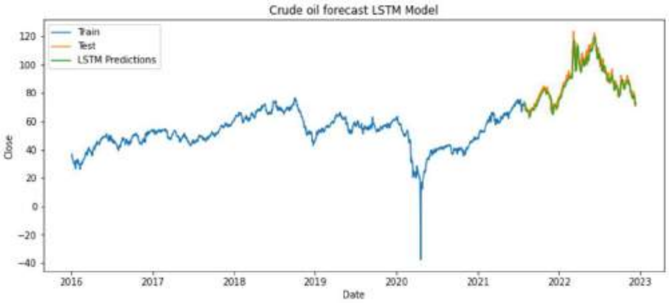


Figure 17 – Result of model LSTM 8-2

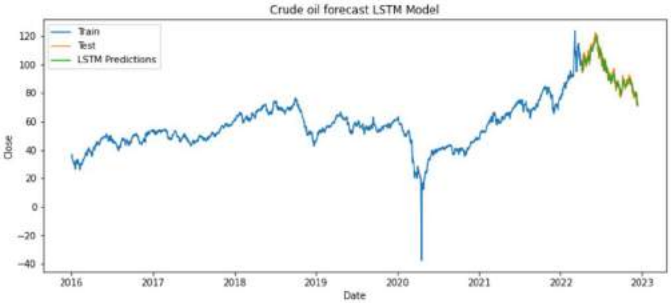


Figure 18 – Result of model LSTM 9-1

In the LSTM model, we can see that the prediction line in all 3 train tests does not move much. This means that the prediction of this model is fit and has high accuracy.

V. CONCLUSION

Model	Train-test	RMSE	MAPE (%)
Linear Regression	7-3	32.06	53.8
	8-2	38.4	74
	9-1	43.62	85.29
Non-Linear Regression		25.84	31.5
ARIMA	7-3	42.38	92.79
	8-2	22.21	24.79
	9-1	14.1	11.77
SARIMA	7-3	42.27	92.24
	8-2	22.24	24.84
	9-1	14.43	11.93
Prophet	7-3	12	0.09
	8-2	14	0.7
	9-1	18	0.89
LSTM	7-3	3.51	3.02
	8-2	4.14	3.5
	9-1	3.33	2.89

RMSE (root mean square error): Variance or standard deviation of the forecast series compared with reality.

MAPE (mean absolute percentage error): Average percentage of absolute error. This indicator shows how much the forecast value deviates from the actual value.

Based on the comparison table, we can in turn found that:

- The model with the highest accuracy is Prophet 7-3 (0.09%)
- The model with the lowest accuracy is ARIMA 7-3 (92.24%).

But overall, the model with the lowest MAPE ratio is Linear Regression.

- Train test 7-3: 53.8%
- Train test 8-2: 74%
- Train test 9-1: 85.29%

Therefore, we can also conclude that Linear Regression model is also not good to predict the future Close price of Crude Oil.

At the same time, we can also see that the percentage of MAPE of Prophet and LSTM is very low, so we should use these 2 models to predict for the next days.

REFERENCES

[1] A. MURADOV, Y. HASANLI, N. HAJIYEV, "CRUDE OIL PRICE FORECASTING TECHNIQUES IN THE WORLD MARKET", 2018.

[2] Y. Chen, K. He, G.K.F. Tso, "Forecasting Crude Oil Prices: A Deep Learning based Model", 2017.

[3] W. Ahmad, M. Aamir, U. Khalil, M. Ishaq, N. Iqbal, M. Khan, "A New Approach for Forecasting Crude Oil Prices Using Median .

[4] Öznur Öztunç Kaymak, Yiğit Kaymak, "Prediction of crude oil prices in COVID-19 outbreak using real data", March 11, 2022.

[5] Q.Lu, S.Sun, H.Duan, S.Wang, "Analysis and forecasting of crude oil price based on the variable selection-LSTM integrated model", September 24, 2021.

[6] K.Zhang, M.Hong, "Forecasting crude oil price using LSTM neural networks", May 9, 2022.

[7] Rebecca Bevans, "Simple Linear Regression | An Easy Introduction & Examples", November 15, 2022.

[8] CFI Team, "Nonlinear Regression", December 20, 2022.

[9] R.J.Hyndman, G.Athanasopoulos, "Forecasting: Principles and Practice (3rd ed) – 12.2 Prophet model", December 10, 2022.

[10] Neha Bora, "Understanding ARIMA Models for Machine Learning", November 9, 2021.

[11] A.Bajaj, "ARIMA & SARIMA: Real-World Time Series Forecasting", November 11, 2022.

[12] B.V.Vishwas, Ashish Patel, "Hands-on Time Series Analysis with Python", 2020, p.223.

[13] S.Srihari, "Long-Short Term Memory and Other Gated RNN", p.23-25.