Stock Market Prediction of Vietnamese Companies Using Some Machine Learning Models

Vinh Quang Quach
Faculty of Information System
University of Information Technology
Thu Duc City, Vietnam
20521811@gm.uit.edu.vn

Sy Le Van
Faculty of Information System
University of Information Technology
Thu Duc City, Vietnam
20521854@gm.uit.edu.vn

Thinh Xuan Nguyen
Faculty of Information System
University of Information Technology
Thu Duc City, Vietnam
20521967@gm.uit.edu.vn

Abstract— The stock market allows buyers and sellers to interact and transact shares that represent their ownership of a business. The creation of trustworthy prediction models for the equities market enables investors to make better choices. This research aims to predict the future stock prices of some large businesses (Vingroup, Vietcombank, FPT) in Vietnam. In addition. Machine Learning consists of making computer tasks using human intelligence is currently the top trending technique. [5] It is now a potent analytical tool for managing investments effectively in the financial markets. A novel method that can assist investors in making better investment and management decisions to achieve improved performance of their securities investments has been made possible by the widespread use of ML in the financial sector. [11] In this research, we first review the shares prices of the above business in recent years of the Vietnamese stock markets. After that, we use data and combine some machine learning algorithms, and price stock prediction in the future.

Keywords—stock, linear regression, non-linear regression, ARIMA, Prophet, LSTM.

1. Introduction

Investors, publicly traded firms, and governments are all clearly interested in forecasting stock price changes. The question of whether the market can be forecast has been up for dispute. According to the Random Walk Theory (Malkiel, 1973), prices are established arbitrarily, making it impossible to outperform the market. But with AI advancements, it has been empirically demonstrated that stock price movement is predictable. [9]

The Stock Market is a highly complex system, where huge chunks and volumes of information. Data is generated instantaneously and constantly changes in small proportions with different factors and diversity. The ownership of companies is divided among the shareholders, who are the real owners of a company. The shareholders are the owners of the company's shares that are offered in publicly listed companies. The stocks symbolize the company's ownership in parts and are the stakeholders in the company's profits and losses.

Since the stock market is primarily dynamic, nonlinear, complex, nonparametric, and chaotic in character, stock market prediction is regarded as a tough task of the financial time series prediction process. The stock market is additionally influenced by a variety of macroeconomic factors, including political developments, corporate policies, general economic conditions, investor expectations, institutional investment preferences, movement in other stock markets, investor psychology, etc. [10] [12] [13]

Regarding the stock markets Vietnamese, Vietnam's stock market fluctuates at the same time as the stock market in the world. The change in Vietnam's stock market in recent years is the renewal of current laws and policies related to control, evaluation, and transparency in financial statements and information. The Vietnamese stock market is expected to make a remarkably strong recovery in 2022, fueled by expectations of a period of strong growth after vaccines are widely distributed and the economy fully reopens. Vingroup, Vietcombank, and Vietnam Airlines are large enterprises with outstanding activities on the Vietnamese stock market**.**

1.1 Vietcombank (VCB)

Vietcombank is one of the major banks in Vietnam, which was officially established and put into operation on April 1, 1963, as a state-owned commercial bank. In recent years, Vietcombank's profit has increased rapidly, thereby increasing the value of VCB shares outstanding in the market.

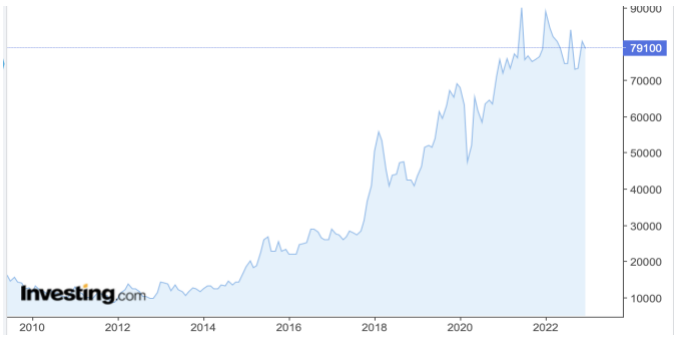

Figure 1: Vietcombank stock market from 2010 to 2022

1.2 Vingroup (VIC)

The company's main business fields include real estate trading, office rental services, housing, machinery, construction equipment, hotel business, entertainment services, entertainment…Vingroup is currently ranked second in terms of capitalization on the stock exchange with approximately VND 370,000 billion, after Vietcombank. Due to the impact of COVID-19, the stock price of this company has also fluctuated sharply in the last seven years.

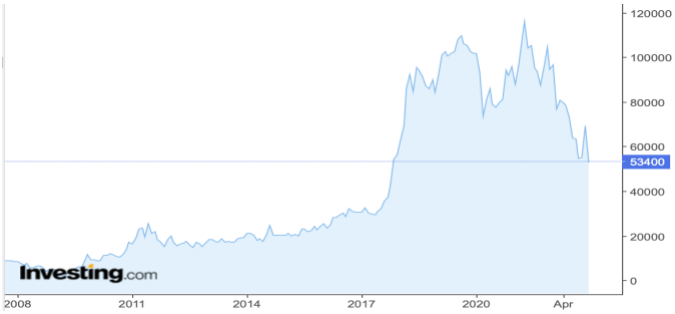

Figure 2: Vingroup stock market from 2010 to 2022

1.3 FPT Telecom (FPT)

FPT Telecom Joint Stock Company (referred to as FPT Telecom for short) is one of the leading telecom and Internet service providers in the region. FPT Telecom is a member of FPT Corporation - one of the largest information technology service companies in Vietnam with the main business of providing information technology products and services.

Figure 3: FPT Telecom stock market from 2010 to 2022

2. Related Work

In recent years, many techniques have been applied to analyze the factors that affect stock prices and related sectors. In the study of Amey Bhadamkar and Sonali Bhattacharya. They using Machine Learning Algorithms, Time Series Forecasting, Sentiment Methods to analysis the unique relationship between Elon Musk's Tweets and Tesla's stock value. They discovered that in the short run, the number of tweets Elon Musk wrote and his interaction marginally corresponds with the stock price of Tesla by examining the frequency of replies or tweets and the stock value varying over time. In addition, their research showed they discovered that Tesla's closing price changes had a direct,parallel association with Musk's engagement after evaluating all of the data.These findings are helpful to the general public since they reveal that Elon Musk's Twitter involvement is a good predictor of stock price rise**.**

Mohammad Almasarweh and S. AL Wadi used ARIMA for predicting banking stock market. They collected around 200 observations for implementing the forecast and the result appeared that ARIMA(1,1,2) was the best one with Root Mean Square Error was 1.4%.

In a study of ARIMA-intervention, Jeffrey E Jarrett and Eric Kyper used ARIMA model with intervention to forecast and analyze Chinese stock price. They fetched the database from PACAP-CCER China Database developed by the Pacific-Basin Capital Markets (PACAP) Research Center at the University of Rhode Island (USA) and the SINOFIN Information Service Inc, affiliated with the China Center for Economic Research (CCER) of Peking University (China). The findings show that China was impacted by the global financial crisis, which also had an impact on its stock market and manufacturing sector.

Adil Moghar and Mhamed Hamiche used LSTM to predict two stock markets in the New York Stock Exchange (GOOGL and NKE) fetched from yahoo finance. They tried different number of epochs and length of dataset and the results is varied. They concluded that with less data and more epochs, their testing results become less volatile.

Beijia Jin, Shuning Gao, Zheng Tao used a hydrid combination of ARIMA and Prophet in Google Stock Price Prediction during the COVID-19 pandemic. Their research came into conclusion that the error of the ARIMA model is less than Prophet model, which resulted in ARIMA had more accuracy than Prophet through the RMSE calculation, where ARIMA was 0.01842 and Prophet was 47.56. Ultimately, it appears that ARIMA performs better than Prophet when it comes to forecasting during pandemic.

3. Methodology

3.1. Data Collection

We extracted dataset about stock market price of Vietcombank, Vingroup and FPT Telecom available on Investing, which contains all the stock trading information of all companies in the world including Vietnam. Each data set contains:

- Open: the first price at which a stock trades upon the opening of the exchange on a trading day.
- Price (also know as Close Price): the last price at which a stock trades upon the end of the exchange on a trading day.

- High: the maximum price of stock.
- Low: the minimum price of stock.
- Volume: the number of shares traded in a particular stock.

The original datasets contain over 5 thousands of stock price since the establishment day. We reduced the range of data from 2015 to 2022 and gathered around 1700 rows of each parameters . After that, we started cleaning and formatting the dataset for an ease of use. By changing the decimal format into "#.#", which originally was "#,#" and removed unnecessary rows, we obtained a clean and numerical dataset.

3.2 Algorithm

Linear Regression Linear Regression (LR) is a predictive model that develops a straight line between a dependent variable and multiple independent variables. Linear fit is achieved by minimizing the mean squared error between predicted and actual output. [1] The formulation of Simple Linear Regression can be described as below:

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

where,
y is a dependent variable
β_0 is the intercept
β_1 is the regression coefficient
X is independent variable
ϵ is the error of the estimate

Non-linear Regression Non-linear regression is a type of regression analysis where observational data are represented by a function that depends on one or more independent variables and is a nonlinear combination of the model parameters. In order to fit the data, a method of sequential approximations is used. The statistical form of Non-linear Regression can be describe below: [2]

$$y \sim f(x, \beta)$$

where,

y is dependent variable

x is independent variable

β is the vector of parameter

Auto Regressive Integrated Moving Average Auto Regressive Integrated Moving Average is one of the most popular statistical analysis model when it comes to forecast and predict. Time series data are used by ARIMA to either comprehend the data set better or to forecast future patterns based on historical data. ARIMA consists of three components:

- Autoregression (AR): a model in which a changing variable regresses on its own lag, or prior, values.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

- Integrated (I): model looks for differences between static data and previous values. The goal is to obtain stationary data that is not affected by seasonality. [4]

  $$By_t = y_{t-1}$$

  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): incorporates the dependency between an observation and a residual error resulting from the application of a moving average model to lagged observations.

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

Each of the AR, I, and MA components are included in the model as a parameter. Hence we get ARIMA(p, d, q) as a complete function. Below is a formula of ARIMA, which is also combined by three components of ARIMA.

$$Y_t = \beta_1 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p}$$

Long Short Term Memory The Long Short-Term Memory (LSTM) Network is an extended version of the RNN proposed by Sepp Hochreiter and Jurgen in 1997. LSTM is designed to solve the Gradient Vanishing problem in RNNs caused by long-term dependencies. An LSTM network can consist of many interconnected LSTM memory cells, and the specific architecture of each cell is shown in the figure. The LSTM architecture consists of one cell state C_t and three gates - one forget gate f_t, one input gate i_t, and one output gate o_t. [5]

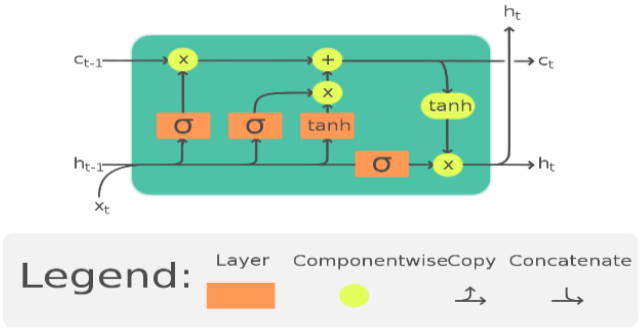

Figure 4: LSTM architecture

Forget gate is where the information in the previous timestamp will be decided to be kept or forgotten. The equation of forget can be written as:

$$f_t = \sigma(x_t * U_f + H_{t-1} * W_f)$$

Figure 5: Forget gate equation

where,

X_t: input to the current timestamp

U_f: weight associated with the input

H_{t-1}: The hidden state of the previous timestamp

W_f: It is the weight matrix associated with hidden state

Input gate is where the importance of the new information carried by the input get quantified. The equation of input gate can be written as:

$$i_t = \sigma(x_t * U_i + H_{t-1} * W_i)$$

Figure 6: Input gate equation

where,

X_t: Input at the current timestamp t

U_i: weight matrix of input

H_{t-1}: A hidden state at the previous timestamp

W_i: Weight matrix of input associated with hidden state

Output gate is responsible for deciding the output of each cell. The equation of output is shown below, which is similar to the forget gate and input gate: [5]

$$(_oW * _{t-1}H + _oU * _tx)\sigma = _to$$

Figure 7: output gate equation

Prophet When it comes to forecasting time series that has multiple seasonality, Prophet is one of the most suitable algorithms for this. Prophet is an open-source algorithm developed by Facebook's data science research team. By fitting curved trends with yearly, monthly, and daily seasonality as well as holiday impacts. [7] Prophet is used to forecast time series data based on an additive model without facing drawbacks of other algorithm. [8] Below is the function of Prophet which is the sum of three different functions plus the error.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

Figure 8: Prophet function

where, [3]

$y(t)$ is the function of the additive regression model

$g(t)$ is the function of trends

$s(t)$ is the function of seasonality

$h(t)$ is the function of holiday effect

$\varepsilon(t)$ is the function of error

4. Result

For a simple model like Linear Regression and Non-linear Regression, we gave all the data to the train set since both models don't have to use the train test splitting technique. As for machine learning models like ARIMA, LSTM and Prophet. Splitting data into train set and test set is one of the most essential step. We choose the rate of train-test for these three models to be 7-3, 8-2 and 9-1, where,

7-3: 70% of dataset is train data and 30% of data is test data
8-2: 80% of dataset is train data and 20% of data is test data
9-1: 90% of dataset is train data and 10% of data is test data

Then, we calculate two values, which are Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). MAPE shows the average difference between predicted values and actual values, RMSE shows the accuracy of forecasting result. The result is shown below:

Company	Model	Train-Test	MAPE (%)	RMSE
VCB	Linear	10-0	10.00	6.17
	Non-linear	10-0	8.47	5.14
	ARIMA	7-3	12.81	11
		8-2	11.16	9
		9-1	11.14	9
	LSTM	7-3	4.04	4
		8-2	2.93	3
		9-1	3.09	3
	Prophet	7-3	5.03	4
		8-2	5.15	3
		9-1	5.08	3
VIC	Linear	10-0	32.17	23.43
	Non-linear	10-0	8.84	7.77
	ARIMA	7-3	17.00	17
		8-2	26.99	23
		9-1	21.28	16
	LSTM	7-3	4.28	5
		8-2	6.11	5
		9-1	4.11	3
	Prophet	7-3	4.62	5
		8-2	4.60	5
		9-1	4.51	5
FPT	Linear	10-0	29.20	10.68
	Non-linear	10-0	8.05	4.08

	ARIMA	7-3	43.17	36
		8-2	8.06	9
		9-1	7.63	7
	LSTM	7-3	8.02	7
		8-2	6.57	6
		9-1	5.73	5
	Prophet	7-3	5.16	3
		8-2	5.02	3
		9-1	4.82	3

Table 1: Statistical table of MAPE and RMSE of each model for three datasets

In addition, we also plotted the graph for each of the train test split and model to see how much the predicted value is changed compared to the actual value visually.

For Vietcombank,

Figure 9: Predicted result of Linear Regression

Figure 10: Predicted result of Non-linear Regression

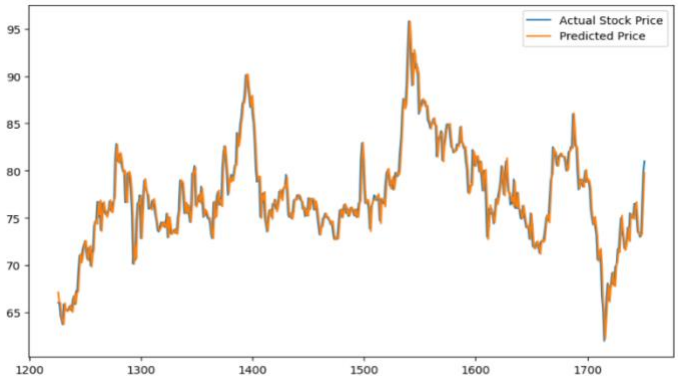

Figure 11: Predicted result of ARIMA with train-test rate 7-3

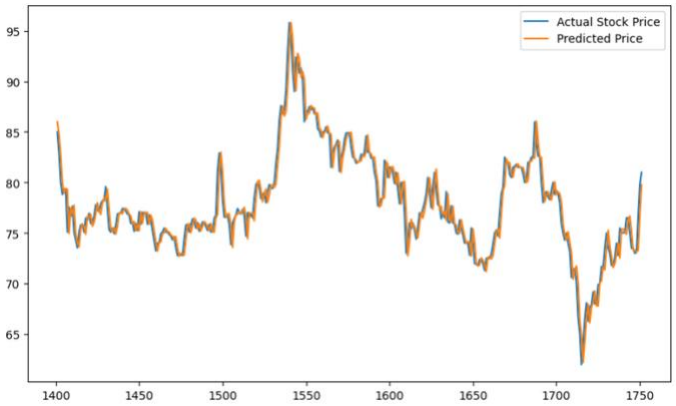

Figure 12: Predicted result of ARIMA with train-test rate 8-2

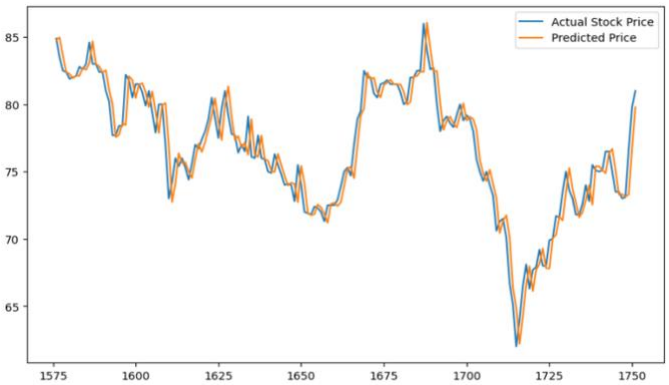

Figure 13: Predicted result of ARIMA with train-test rate 9-1

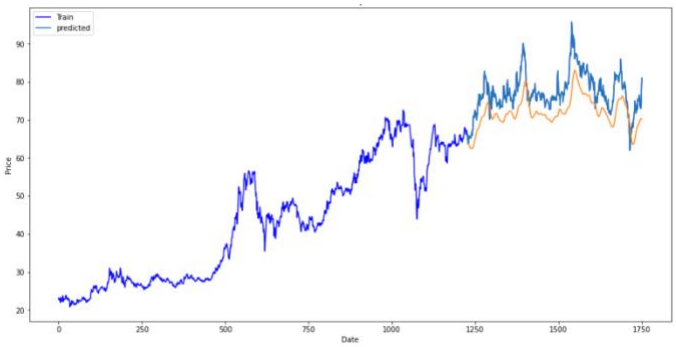

Figure 14: Predicted result of LSTM with train-test rate 7-3

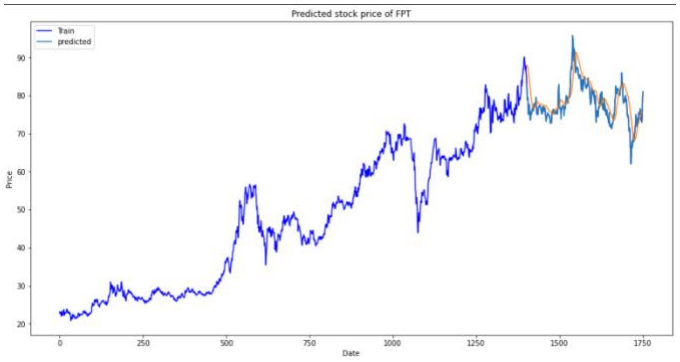

Figure 15: Predicted result of LSTM with train-test rate 8-2

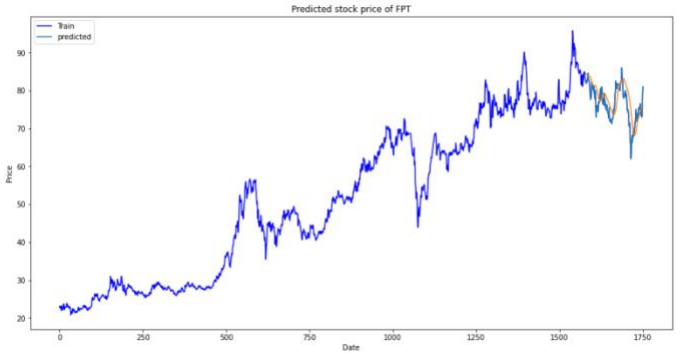

Figure 16: Predicted result of LSTM with train-test rate 9-1

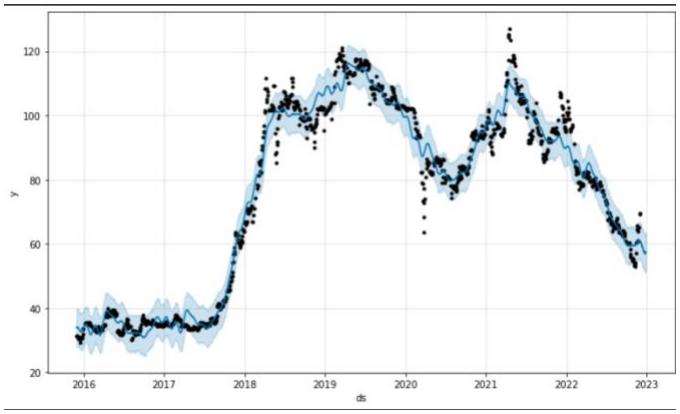

Figure 17: Predicted result of Prophet with train-test rate 7-3

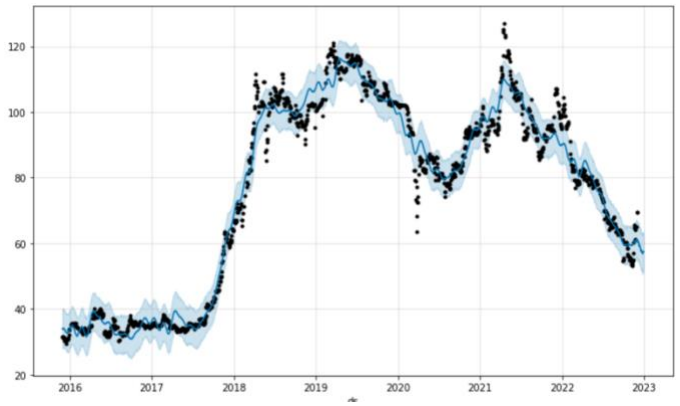

Figure 18: Predicted result of Prophet with train-test rate 8-2

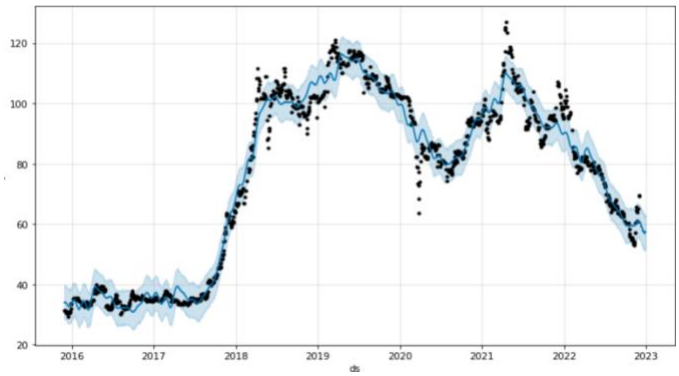

Figure 19: Predicted result of Prophet with train-test rate 9-1

For Vingroup,

Figure 20: Predicted result of Linear Regression

Figure 21: Predicted result of Non-linear Regression

Figure 22: Predicted result of ARIMA with train-test rate 7-3

Figure 23: Predicted result of ARIMA with train-test rate 8-2

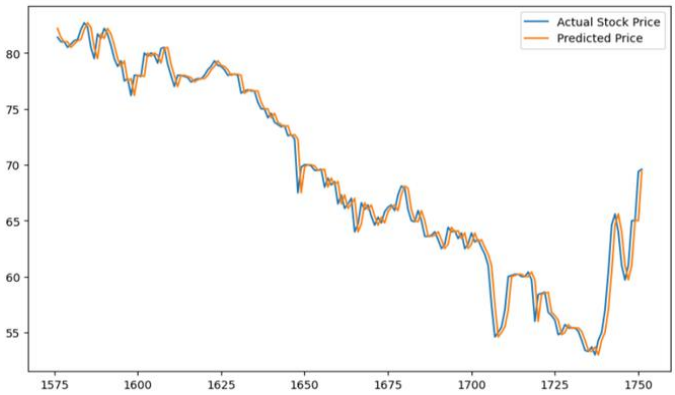

Figure 24: Predicted result of ARIMA with train-test rate 9-1

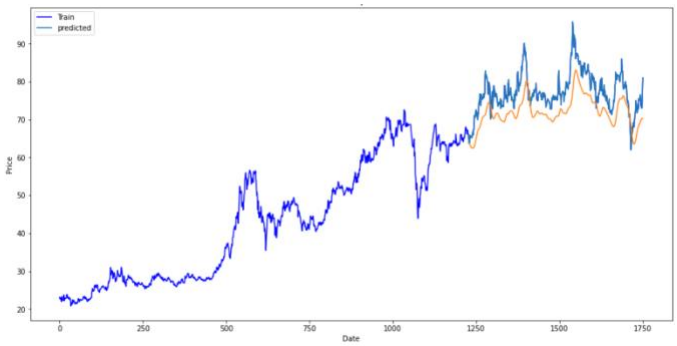

Figure 25: Predicted result of LSTM with train-test rate 7-3

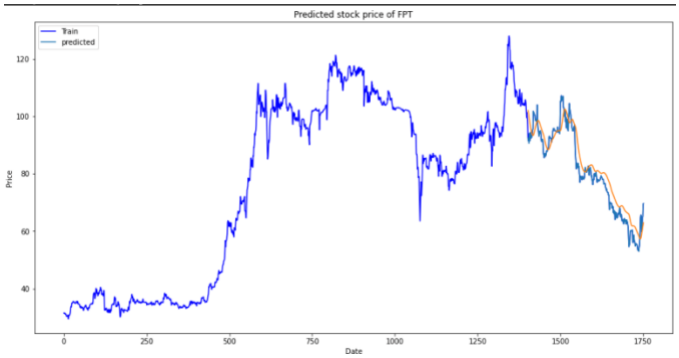

Figure 26: Predicted result of LSTM with train-test rate 8-2

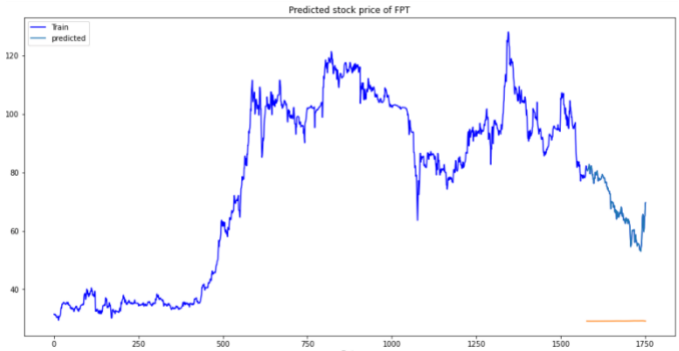

Figure 27: Predicted result of LSTM with train-test rate 9-1

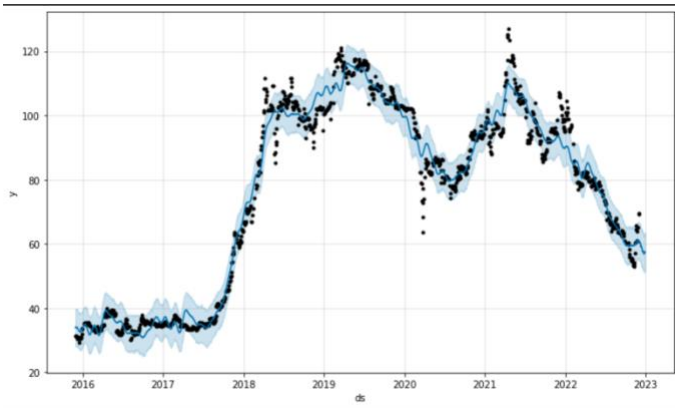

Figure 28: Predicted result of Prophet with train-test rate 7-3

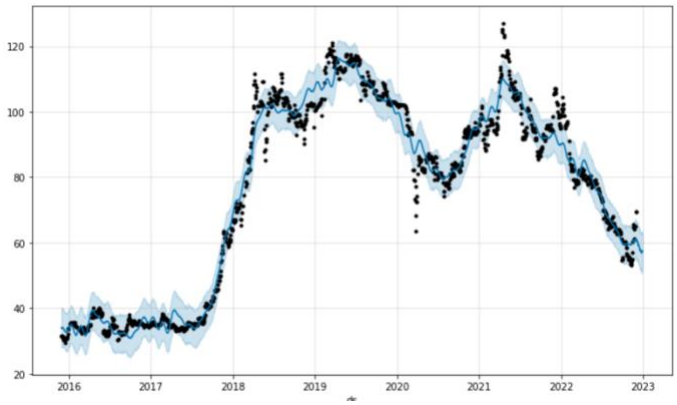

Figure 29: Predicted result of Prophet with train-test rate 8-2

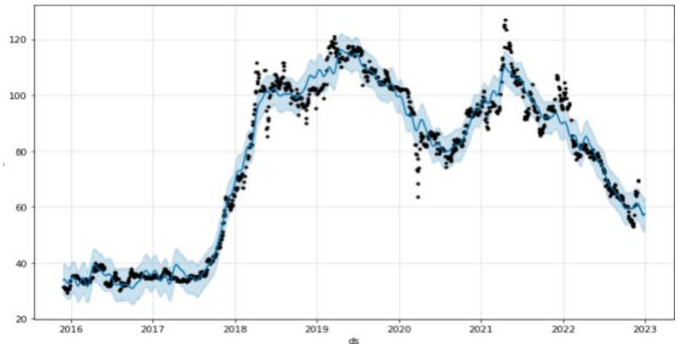

Figure 30: Predicted result of Prophet with train-test rate 9-1

For FPT Telecom,

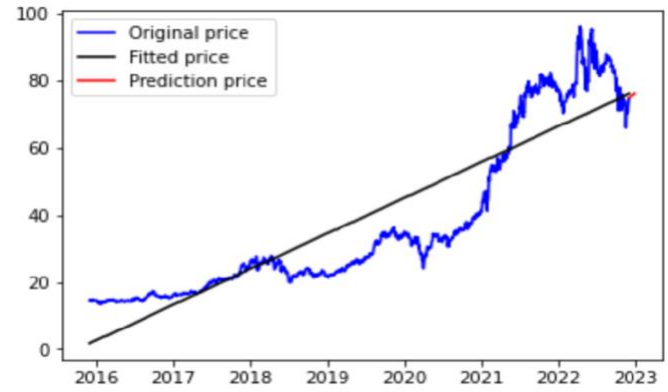

Figure 31: Predicted result of Linear Regression

Figure 32: Predicted result of Non-linear Regression

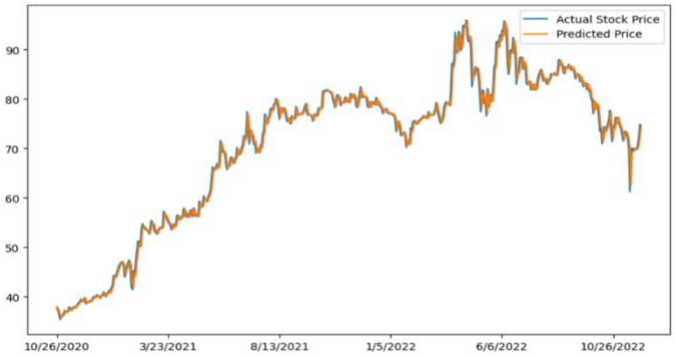

Figure 33: Predicted result of ARIMA with train-test rate 7-3

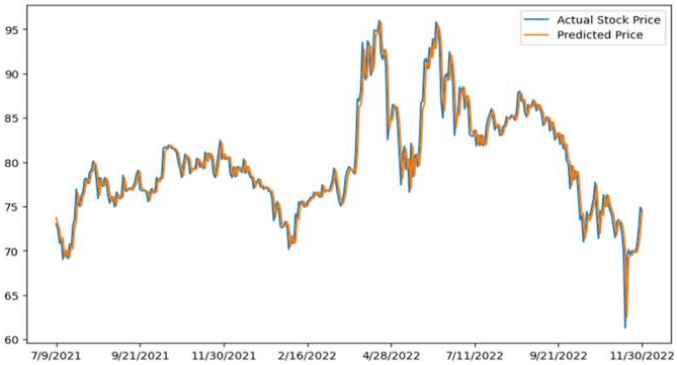

Figure 34: Predicted result of ARIMA with train-test rate 8-2

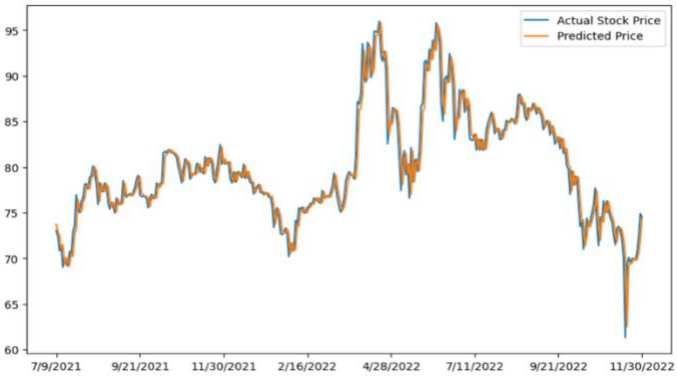

Figure 35: Predicted result of ARIMA with train-test rate 9-1

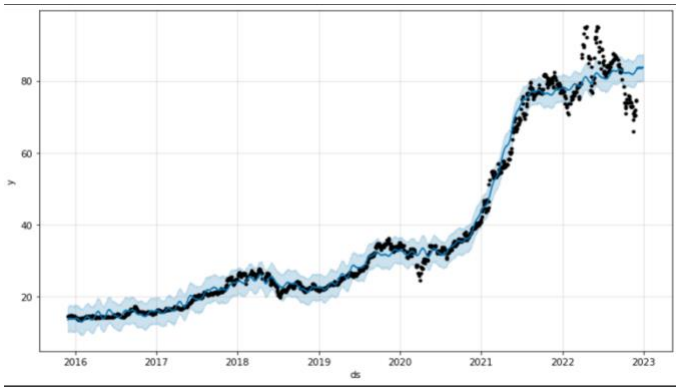

Figure 36: Predicted result of Prophet with train-test rate 7-3

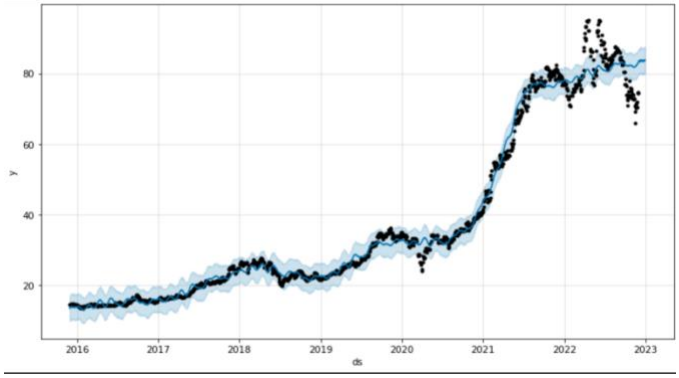

Figure 37: Predicted result of Prophet with train-test rate 8-2

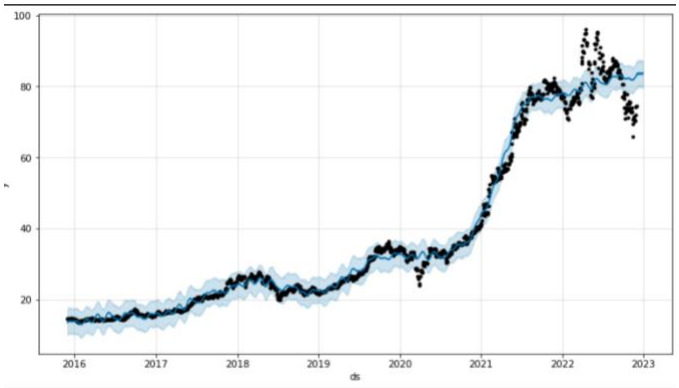

Figure 38: Predicted result of Prophet with train-test rate 9-1

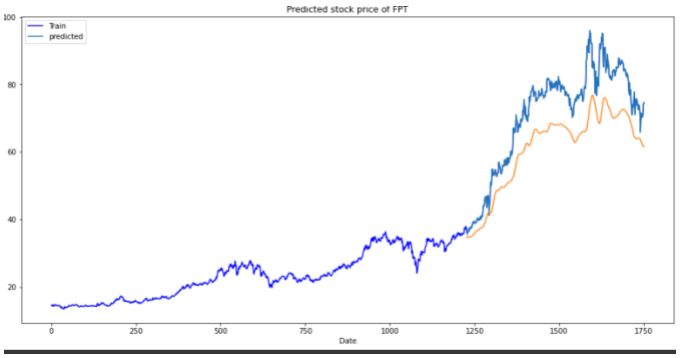

Figure 39: Predicted result of LSTM with train-test rate 7-3

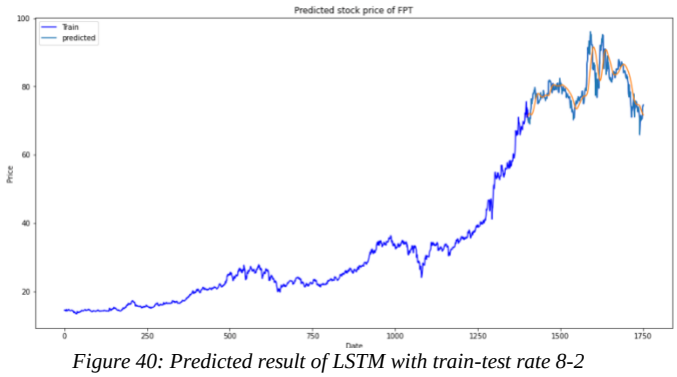

Figure 40: Predicted result of LSTM with train-test rate 8-2

Figure 41: Predicted result of LSTM with train-test rate 9-1

5. Discussion

In this study of Stock Price Prediction, we used different machine learning models to analyze and forecast the Price column of three different companies in Vietnam. Linear Regression is the model that has the worst performance so far as it can only portray the linear relationship between Price and Date. As for the best performance model, none of these above had made any prediction on three datasets. LSTM performs well in FPT Telecom dataset but poorly in Vingroup, resulting in a low MAPE in the train-test rate of 7-3 and 8-2. ARIMA is best at Vietcombank and Prophet is also Vietcombank. We think that these models are the average choices for forecasting these three dataset as there are better ones.

6. Reference

[1] Alex Greaves, Benjamin Au, Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin, December 8, 2015, pp. 4-5.

[2] https://en.wikipedia.org/wiki/Nonlinear_regression

[3] Beijia Jin, Shuning Gao, Zheng Tao, ARIMA and Facebook Prophet Model in Google Stock Price Prediction, October 21, 2022, pp. 62.

[4] https://www.capitalone.com/tech/machine-learning/understanding-arima-models/

[5] Adil MOGHAR ,Mhamed HAMICHE, Stock Market Prediction Using LSTM Recurrent Neural Network, pp. 1170

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] https://github.com/facebook/prophet

[8] https://towardsdatascience.com/time-series-analysis-with-facebook-prophet-how-it-works-and-how-to-use-it-f15ecf2c0e3a

[9] Xiao Ding, Yue Zhang, Ting Liu, Junwen Duan, Using Structured Events to Predict Stock Price Movement: An Empirical Investigation, pp. 1415

[10] Yakup Kar, Melek Acar Boyacioglu, Ömer Kaan Baykan, Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange, pp. 1

[11] Mehtabhorn Obthong, Nongnuch Tantisantiwong, Watthanasak Jeamwatthanachai, Gary Wills1, A Survey on Machine Learning for Stock Price Prediction: Algorithms and Techniques**,** pp. 1

[12] Masoud, Najeb MH. (2017) "The impact of stock market performance upon economic growth." *International Journal of Economics and Financial Issues* 3 (4)**,** pp. 788–798

[13] Murkute, Amod, and Tanuja Sarode. (2015) "Forecasting market price of stock using artificial neural network." *International Journal of Computer Applications* 124 (12), pp. 11-15