UNIVERSITY OF INFORMATION TECHNOLOGY

FACULTY OF INFORMATION SYSTEMS

Prediction Stock Price Using Statistical Models

- NGUYEN NGOC HUYEN - 20521423
- NGUYEN THI NGOC THAO - 20521935
- PHAM HA MINH QUYEN - 20521824

PREDICTION STOCK PRICE USING STATISTICAL MODELS

Nguyen N. Huyen
STAT3013.N11.CTTT-EN
University of Information Technology
20521423@gm.uit.edu.vn

Nguyen T. N. Thao
STAT3013.N11.CTTT-EN
University of Information Technology
20521935@gm.uit.edu.vn

Pham H. M. Quyen
STAT3013.N11.CTTT-EN
University of Information Technology
20521824@gm.uit.edu.vn

***Abstract*— Stocks are simply understood as valuable papers confirming ownership of stocks of business. Many companies decide to issue stocks to serve their growth plan. Stocks represent the ownership of the business by each shareholder, so the business is not obliged to return such capital contribution to the owners of its corporate stocks. The stock market has established itself as an effective capital mobilization channel for the economy. In this paper, we investigate the predictive stock price of some Corporation in Vietnam.**

I. INTRODUCTION

Stocks are securities that confirm the owner's lawful rights and interests to part of the stock capital of the issuing organization. In 1602, the Dutch East India Company issued its first shares through the Amsterdam Stock Exchange, and it was the first company to issue stock and bonds.

The purpose of stock valuation is to determine the true value of a stock at a given point in time, find out the potential of the stock and to make relevant investment decisions. For businesses, stock valuation is one of the important steps of a joint stock company when it wants to offer shares, raise capital and increase its influence in the market. For investors, stock valuation helps investors know which stocks are worth buying and have the greatest potential for return. An easy way to do that is to gauge how much the stock is worth. Then, we will proceed to buy the stock if the share price is lower than the value we value. Or sell the stock (if the investor owns the stock) if the share price is already higher than the valuation to make a profit.

There are many algorithms and techniques that help us predict prices. In this paper, we will use Linear Regression, Non-Linear Regression, ARIMA, LSTM, and Prophet models to predict the stock prices of some corporations in the next 30 days. From there, evaluate and compare the above 5 models.

II. RELATED WORK

Our work builds on prior research and research as a basis to conduct supervised machine learning prediction on the price of the stock. The system is designed for predictions of the future prices of stocks for the next 5 years using Facebook Prophet which can be used for better investments [1]. Manoj S Hegde et al. [2] investigated that The Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) capable of solving in volute linear problems, and also there is a discussion about the usage of RNN (Recurrent Neural Networks) to predict the share prices. Forecasting has been dominated by linear methods for many decades. Linear methods are easy to develop and implement and they are also relatively simple to understand and interpret.

There are some researches use Linear Regression to predict stock market. [3] used Linear Regression to predict stock market trading volume and draw conclusions from that we can define marketing and customer satisfaction guidelines that are pertinent to all clientele and their preferences, finally resulting in higher output and earnings. [4] Other research show that result of Linear Regression can be improved by transformation process to normalize the input common data

Nonlinear Granger causality tests to examine the dynamic relation between daily aggregate stock price and trading volume. It proves useful information on whether knowledge of past stock price movements in trading volume, and vice versa[5]. Nonlinearities are sufficiently important that they can be detected with nontrivial power using various tests for nonlinearities [6].

In Addition, [7] a group of authors studied the performance of the ARIMA time model in predicting stock prices from many different companies. They draw conclusions, compared to other sectors, the accuracy of forecasts made using the ARIMA model for the banking and automotive industries is lower. Therefore, a better model is required for predicting the stock prices of the companies in the aforementioned industry.

III. METHODS

A. Data Collection

We use stock price data available on investing website, which is a platform that provides price list data of stock tickers in real time, we have downloaded historical price data of stocks of some businesses in Vietnam from December 1, 2017, *to*

December 1, 2022, and save it as a csv file. Each line includes date, open, price, high, low, vol, change.

In those figures below, this is all data our group has been collected from 1/12/2017 to 1/12/2022 (figure 4, figure 5, figure 6):

Then we will be using the Matplotlib library in Python order to visualize the data of MSN, SSI, VND before predicting (figure 1, figure 2, figure 3):

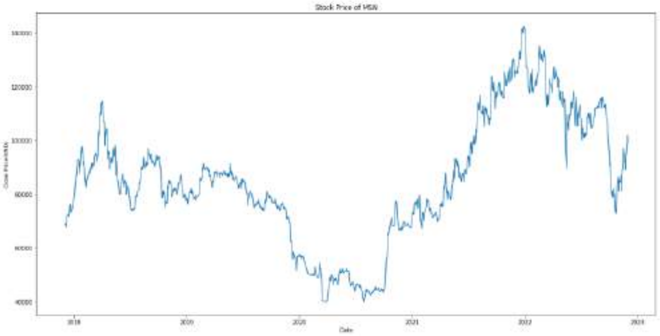

Figure 1: Movement's stock of MSN

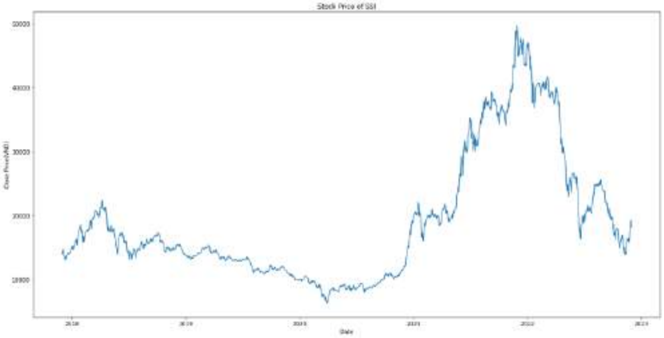

Figure 2: Movement stock of SSI

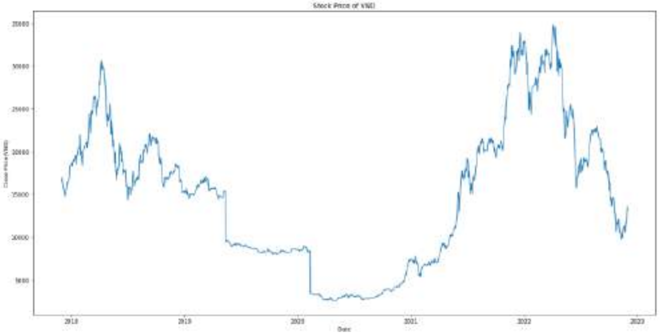

Figure 3: Movement stock of VND

In this study, we will calculate and figure out the impacts of the following two important factor are selected: Date and Price. Then we will use those value to predict the value of the date after to consider that is correct or not, that is shown through the 5 models below: Linear Regression, Non-Linear Regression, LSTM, ARIMA and Prophet.

B. *Learning Algorithms*

Linear Regression Linear Regression (LR) is the mostused statical technique relationship between two variables with straight line based on a line best fit. In order to predict the dependent variable using an independent variable while helping to minimize the squared error, linear regression seeks to find the line of best fit.

$$\hat{y} = \mathrm{b_o} + \mathrm{b_1}x$$

Where:
$\hat{y}$: the predicted dependent variable
$\mathrm{b_o}$: the slope of the line
$\mathrm{b_1}$: the intercept of the line
x: the independent variable also the variable we are using to predict y

Non-Linear Regression Non-Linear Regression (NLN) is a statistical tool that shows the relationship between a dependent variable and two or more independent variables. This is a valuable tool for predicting modeling is a statistical tool that relates the two variables in a non-linear curved relationship.

General form:

$$\mathrm{Y_i} = f(\mathrm{X_i}\,, y) + \hat{\mathrm{I}}_\mathrm{i}$$

Each $\mathrm{Y_i}$ observation is set to be the sum of the mean response $f(\mathrm{X_i}\,, y)$ and a random error term $\hat{\mathrm{I}}_\mathrm{i}$

ARIMA ARIMA (Autoregressive Integrated Moving Average) is a general class of statistical models for time series analysis forecasting. ARIMA uses a time series past value and/or forecast errors to predicts its future values. [8]

ARIMA model assumption – stationary: the time series has its statistical properties remain constant across time.

Three components/parameters: AR + I + MA (p, d, q)

AR (Auto regression) : the time series is linearly regressed on its own past values.

- p: the number of past values included in the AR model
  $y_\mathrm{t} = c + q_1\mathrm{y_{t\text{-}1}} + q_2\mathrm{x_{t\text{-}2}} + \cdots + q_\mathrm{p}\mathrm{y_{t\text{-}p}} + \hat{\mathrm{I}}_\mathrm{t}$ [9]

I (Integrated): if not stationary, the time series can be differenced to become stationary, i.e., compute the differences between consecutive observations.

- d: the number of times the time series is differenced
  $\nabla y_\mathrm{t} = \mathrm{y_t} - \mathrm{y_{t\text{-}1}}$ [10]

MA (Moving average): the time series is 'regressed' on the past forecast errors.

- q: the number of past forecast errors included in the MA model

$$y_\mathrm{t} = c + \hat{\mathrm{I}}_\mathrm{t} + \mathrm{q_1}\hat{\mathrm{I}}_\mathrm{t\text{-}1} + \mathrm{q_2}\hat{\mathrm{I}}_\mathrm{t\text{-}2} + \cdots + \mathrm{q_q}\hat{\mathrm{I}}_\mathrm{t\text{-}q}$$

ARIMA(p,d,q) full equation:

$$\nabla y_\mathrm{t} = c + j_1\,\tilde{\mathrm{N}}y_{\mathrm{t}-1} + \cdots + j_\mathrm{p}\tilde{\mathrm{N}}y_{\mathrm{t}-\mathrm{p}} + \hat{\mathrm{I}}_\mathrm{t} + q_1\hat{\mathrm{I}}_\mathrm{t\text{-}1} \ldots + q_\mathrm{q}\hat{\mathrm{I}}_\mathrm{t\text{-}q} \quad [11]$$

LSTM Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) used in the field of deep learning. Unlike standard feedforward neural networks (FNNs), LSTMs contain feedback connections. Networks not only process single data points (such as images), but also entire data series (such as speech or video).

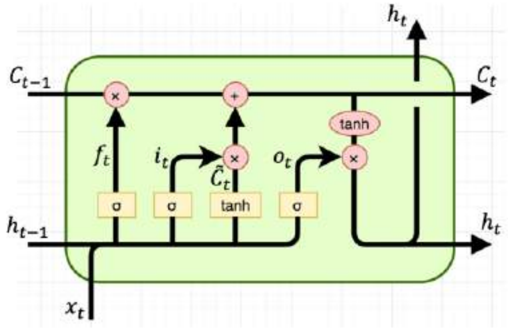

Figure 4: LSTM model

Forget gate: $f_t = \sigma(U_f * x_t + W_f * h_{t-1} + b_f)$

Input gate: $i_t = \sigma(U_i * x_t + W_i * h_{t-1} + b_i)$

Output gate: $o_t = \sigma(U_o * x_t + W_o * h_{t-1} + b_o)$

Prophet Prophet is open-source software released by Facebook's Core Data Science team. Prophet is intended to forecast univariate time series datasets using an additive model that fits non-linear trends with yearly, weekly, and daily seasonality. While compared to other models, this model requires very little calculation time.

$$y(t) = g(t) + h(t) + s(t) + e.t$$

Where:
y(t): Additive Regressive Model
g(t): Trend factor
h(t): Holiday component
s(t): Seasonality component
e.t: Error Term

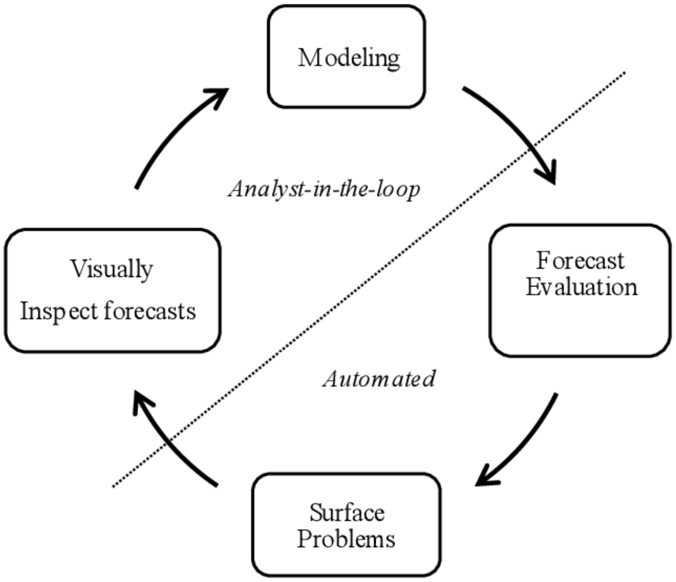

Figure 5: Prophet model.

IV. Results

With ARIMA model included three parameters p, d, and q. They take non-negative integer values, indicating which specific ARIMA model is used. With LSTM, model is built through six layers (1 input layer, 3 hidden layers, 1 output layer), unit = 64 for each hidden layer, learning rate = 0.0001, batch size = 32, epoch = 100. Next, about model Prophet, the main parameters for Prophet models: growth (The form of the trend: "linear", or "logistic"). The changepoint (maximum number of trends changepoints allowed when modeling the trend). Changepoint range (The range affects how close the changepoints can go to the end of the time series). The larger the value, the more flexible the trend. Prior Scale (Controls flexibility of: Changepoints, Seasonality, Holidays). The log10_trans (converts priors to a scale from 0.001 to 100, which effectively weights lower values more heavily than larger values).

Stock	Model	Train - Test	MAPE	RMSE
MSN	ARIMA	9 – 1	13.68%	20086.54
		8 - 2	12.33%	19660.59
		7 -3	30.70%	29502.10
	LSTM	9 - 1	5.17%	6576.81
		8 – 2	**5.03%**	**6994.01**
		7 - 3	6.83%	9251.22
	Prophet	9 – 1	9.08%	11425.58
		8 – 2	8.13%	9183.69

		7 – 3	9.23%	9342.98
	Linear	10 – 0	25.29%	21803.14
	Non-Linear	10 – 0	6.80%	6812.59
VND	ARIMA	9 - 1	30.62%	8585.72
		8 – 2	26.47 %	10409.17
		7 - 3	53.88%	9788.04
	LSTM	9 - 1	7.64%	1772.65
		8 - 2	7.43%	2158.46
		7 - 3	**7.30%**	**2054.68**
	Prophet	9 - 1	14.64%	3554.32
		8 - 2	16.15%	3041.33
		7 - 3	14.94%	1632.90
	Linear	10 – 0	97.37%	8365.98
	Non-Linear	10 - 0	15.52%	2361.88
SSI	ARIMA	9 - 1	20.75%	6102.71
		8 – 2	41.54%	22879.02
		7 – 3	31.43%	9973.13
	LSTM	9 – 1	8.09%	2094.29
		8 - 2	6.75%	2239.98
		7 – 3	**6.33%**	**2396.93**
	Prophet	9 – 1	11.07%	4360.94
		8 – 2	10.94%	3875.02
		7 - 3	10.55%	2835.99
	Linear	10 – 0	43.65%	8514.69
	Non-Linear	10 - 0	8.31%	2424.81

Table 1 Evaluation of 5 models with the close stock price from 3 companies.

Values give best prediction stock price of MSN company in each model:

With ARIMA model with the train-test ratio respectively 8-2 have values MAPE = 12.33% and RMSE = 20086.54

With LSTM model with the train-test ratio respectively 8-2 have values MAPE = 5.03% and RMSE = 6994.01

With Prophet model with the train-test ratio respectively 8-2 have values MAPE = 8.13% and RMSE = 9183.69

Values give best prediction stock price of SSI company in each model:

With ARIMA model with the train-test ratio respectively 9-1 have values MAPE = 20.75% and RMSE = 6102.71

With LSTM model with the train-test ratio respectively 7-3 have values MAPE = 6.33% and RMSE = 2396.93

With Prophet model with the train-test ratio respectively 7-3 have values MAPE = 10.55% and RMSE = 2835.99

Values give best prediction stock price of VND company in each model:

With ARIMA model with the train-test ratio respectively 8-2 have values MAPE = 26.47 % and RMSE = 10409.17

With LSTM model with the train-test ratio respectively 7-3 have values MAPE = 7.30% and RMSE = 2054.68

With Prophet model with the train-test ratio respectively 9-1 have values MAPE = 14.64% and RMSE = 3554.32

From these values above, shown that:

For company MSN, the first 3 - 5 days there is an average deviation between 30 predictions and 30 actual days of about 26%, then on the 7th day the average deviation decreases to 25% and after 15 days it drops to 20%, but in the last 3 days, the deviation increased sharply to 95%. The total mean deviation after 30 days is 29%.

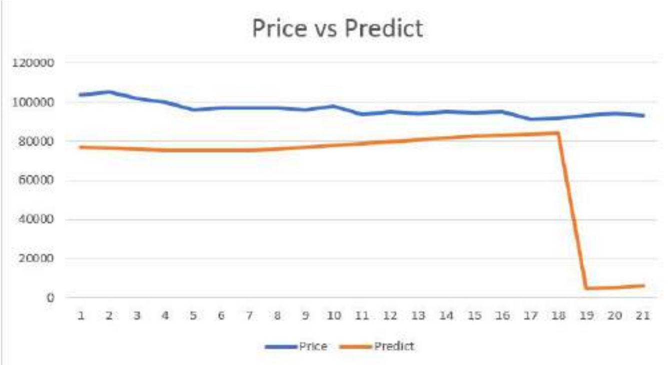

Figure 6 Actual and predicted variance of stock price MNS

With SSI company, the first day has a 13% difference between the predicted value and the actual value, the first 3 days, the difference between the first 3, 5, 7, 15 days is 15%, respectively, 16%, 16% and 19%. The predicted results increase gradually but not too high (about 1-3%).

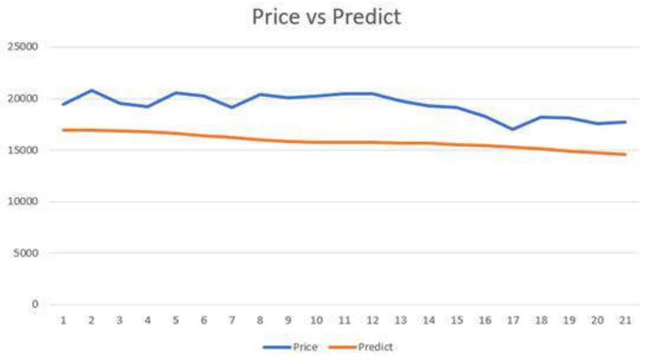

Figure 7 Actual and predicted variance of stock price SSI

For the VND company, the first day deviation is 2% from the actual value, but 3 - 5 days later the deviation between the predicted 30 days and the actual 30 days averages 6% to 8%. Then on 7th day the mean deviation increased to the enjoying level of 10% and after 15 days the flatness still increased sharply to 18%. And the last 3 days, the deviation increased extremely sharply up to 26%. From the results we can see that the level is continuously increasing, and we need to change the model to predict the results to approximate reality.

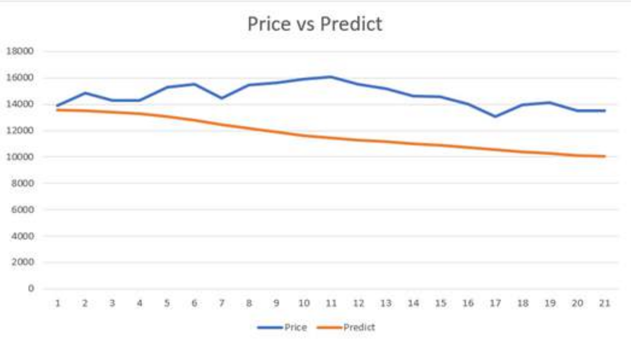

Figure 8 Actual and predicted variance of stock price VND

V. CONCLUSION

In this study, we use different learning algorithms and machine learning algorithms along with 3 types of split train tests to learn each model and their accuracy level. From that we can see, the learning model of Linear Regression and Non-Linear Regression has a rather high error. In addition, the model ARIMA, has similar results even though the prediction results obtained from the test set are only straight lines. Besides, Prophet model gives better MAPE and RSME results than ARIMA. Finally, the LSTM model has the best predictive results in the above models with the lowest MAPE and RSME.

Finally, in the next model studies, we will try to build better models than the current to predict the exact stock price as accurately as possible.

REFERENCES

[1]: Stock Price Prediction using Facebook Prophet - ITM Web of Conferences 44, 03060 (2022)

[2]: M. S. Hegde, G. Krishna and R. Srinath, "An Ensemble Stock Predictor and Recommender System," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1981-1985.

[3] Gharehchopogh, Farhad Soleimanian, Tahmineh Haddadi Bonab, and Seyyed Reza Khaze. "A linear regression approach to prediction of stock market trading volume: a case study." International Journal of Managing Value and Supply Chains 4.3 (2013): 25.

[4] SIEW, Han Lock; NORDIN, Md Jan. Regression techniques for the prediction of stock price trend. In: 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE). IEEE, 2012. p. 1-5.

[5] Granger Causality in the Stock Price- Volume Relation.

[6]: A COMPARISON OF LINEAR AND NONLINEAR UNIVARIATE MODELS FOR FORECASATING MARCROECONOMIC TIME SERIES.

[7] Mondal, Prapanna, Labani Shit, and Saptarsi Goswami. "Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices." International Journal of Computer Science, Engineering and Applications 4.2 (2014): 13.

[8]: Lianne & Justin How to build ARIMA model in Python for time series prediction

[9]: Lianne & Justin How to build ARIMA model in Python for time series prediction

[10]: Lianne & Justin How to build ARIMA model in Python for time series prediction

[11]: Lianne & Justin How to build ARIMA model in Python for time series prediction