FORECASTING THE STOCK PRICES OF VIETNAMESE REAL ESTATE COMPANIES USING STATISTICAL MODELS AND MACHINE LEARNING ALGORITHMS

TRAN THI KIM ANH[1], PHI QUANG THANH[2], AND LE MINH CHANH[3]

[1]Faculty of Information Systems, University of Information Technology, (e-mail: 21520596@gm.uit.edu.vn)
[2]Faculty of Information Systems, University of Information Technology, (e-mail: 21521449@gm.uit.edu.vn)
[3]Faculty of Information Systems, University of Information Technology, (e-mail: 21521882@gm.uit.edu.vn)

ABSTRACT The stock market serves as a cornerstone of Vietnam's finance, providing a platform for investors to trade securities such as stocks, bonds, and derivatives. Developing predictive models for stock prices will aid investors in making decisions more efficiently. In this research, we utilize statistical and machine learning algorithms, as well as deep learning such as Linear Regression, Support Vector Machine (SVM), ARIMA, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Recurrent neural network (RNN), Holt-Winters, TimesNet to forecast the stock prices of three prominent real estate companies: Vinhomes (JSC), Novaland (NVL), and Nam Long Corp (NLG). By leveraging a diverse array of methodologies, we aim to gain insights into the behavior of these stocks and enhance investors' ability to make well-informed decisions in the dynamic real estate market.

INDEX TERMS **Keywords** - Linear Regression, SVM, ARIMA, LSTM, GRU, RNN, Holt-Winters, TimesNet.

I. INTRODUCTION

Vietnam's stock market holds a pivotal position in the country's financial landscape, serving as a key platform for investors to engage in the trading of various securities, including stocks, bonds, and derivatives. With its dynamic nature and significant impact on the economy, the stock market plays a crucial role in facilitating capital mobilization, fostering business growth, and contributing to overall economic development.

In recent years, the adoption of predictive modeling techniques has gained traction within the Vietnamese stock market ecosystem. These predictive models, leveraging statistical and machine learning algorithms, enable investors to forecast stock prices with greater accuracy and efficiency. By harnessing the power of data-driven insights, investors can make more informed decisions, mitigate risks, and seize lucrative investment opportunities.

In this research, we focus on predicting the stock prices of three prominent companies in the Vietnamese real estate sector: Vinhomes, Novaland, and Nam Long Corp. Through the application of various predictive algorithms, including but not limited to Linear Regression, Support Vector Machine (SVM), ARIMA, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Recurrent neural network (RNN), Holt-winters, and TimesNet, we aim to provide valuable insights into the future behavior of these stocks. By examining historical data and market trends, we seek to enhance investors' understanding of the dynamics influencing stock prices in the Vietnamese real estate market.

By leveraging these predictive models, investors can gain a competitive edge in the stock market, optimize their investment strategies, and navigate the complexities of the Vietnamese real estate sector with confidence and precision. Through this research endeavor, we strive to contribute to the advancement of predictive modeling techniques within Vietnam's stock market, ultimately empowering investors to make informed decisions and achieve their financial objectives.

II. RELATED WORKS

Ghosalkar and Dhage (2018) [1] presented a study on real estate value prediction using linear regression at the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). Their research aimed to assess the efficacy of linear regression in forecasting real estate values, contributing to advancements in real estate valuation methodologies.

Lin, Guo, and Hu (2013) [2] introduced an SVM-based

approach for predicting stock market trends at the 2013 International Joint Conference on Neural Networks (IJCNN). Their research highlights the utilization of Support Vector Machines (SVM) in analyzing financial data for trend prediction.

Ariyo, Adewumi, and Ayo (2014) [3] presented a study on stock price prediction utilizing the ARIMA (AutoRegressive Integrated Moving Average) model at the 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. Their research focused on applying time series analysis techniques to forecast stock prices, contributing to advancements in financial modeling methodologies.

Sunny, Maswood, and Alharbi (2020) [4] introduced a deep learning-based approach for stock price prediction using LSTM (Long Short-Term Memory) and Bi-Directional LSTM models at the 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES). Their research aimed to leverage advanced neural network architectures to analyze stock market data and forecast price movements, potentially offering enhanced predictive capabilities in financial markets.

Jaiswal and Singh (2022) [5] proposed a hybrid Convolutional Recurrent (CNN-GRU) model for stock price prediction, leveraging both CNN and GRU architectures. Their research aimed to combine the strengths of CNN for feature extraction and GRU for capturing sequential dependencies, potentially improving the accuracy of stock price forecasts.

Syavasya and Muddana (2021) [6] developed a machine learning-based time series prediction method using Holt-Winters Exponential Smoothing with Multiplicative Seasonality. Their research aimed to improve forecasting accuracy by incorporating advanced machine learning techniques into traditional time series analysis.

III. MATERIALS

A. DATASET

The reference datasets used are sourced as follows: The historical stock price data of Vinhomes (VHM), No Va Land Investment Group Corp (NVL) and Nam Long Investment Corp (NLG). The datasets are obtained from the investing.com website, and the data is available within the time range from March 1, 2019, to March 1, 2024. Because the project goal is to predict closing prices, we'll only analyze data from the "Close" column (in VND). The dataset contains the following columns:

- Date: Represents the date when the financial data was recorded.
- Price (also known as the Close Price): Refers to the price of the stock at the end of exchange.
- Open: Illustrate the opening price of the stock at the beginning of the trading day.
- High: Represents the highest price reached by the stock during the trading day.
- Low: Indicates the lowest price reached by the stock during the trading day.
- Vol.: Stands for volume, which represents the number of shares traded during the trading day.
- Change: Reflects the percentage change in the price of the stock compared to the previous trading day.

B. DESCRIPTIVE STATISTICS

For this project, we will use Python programming language to visualize data in figures.

TABLE 1. VHM, NVL, NLG's Descriptive Statistics

	VHM	NVL	NLG
Count	1252	1252	1252
Mean	62,066	43,756	31,100
Std	11,878	25,614	10,816
Min	38,450	10,250	14,414
25%	53,900	23,350	21,509
50%	61,768	34,213	30,132
75%	71,569	76,000	37,200
Max	88,722	92,366	63,723

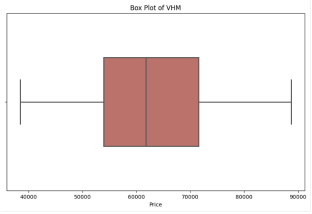

FIGURE 1. VHM stock price's boxplot

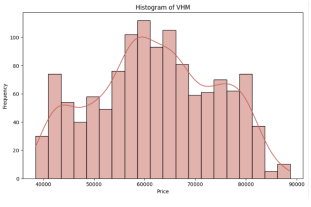

FIGURE 2. VHM stock price's histogram

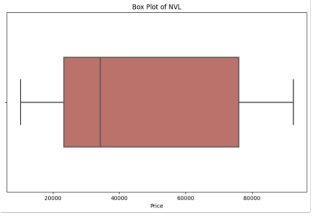

FIGURE 3. NVL stock price's boxplot

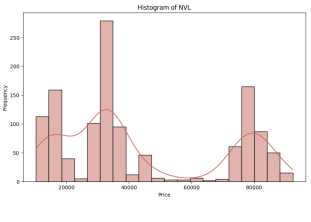

FIGURE 4. NVL stock price's histogram

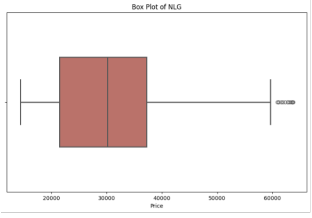

FIGURE 5. NLG stock price's boxplot

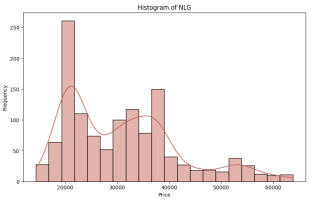

FIGURE 6. NLG stock price's histogram

Based on the data:

- VHM has the highest average at 62,066, followed by NVL at 43,756, and NLG at 31,100.
- NVL has the highest variability with a standard deviation of 25,614.

- NVL also has the widest range of values, from 10,250 to 92,366.
- NLG has the lowest average and variability.

Overall, NVL stands out for its high variability and wide range of values, while VHM consistently maintains higher averages. NLG consistently has lower values compared to the other two groups.

IV. METHODOLOGY

A. LINEAR REGRESSION

Simple linear regression describes the relationship between one variable's magnitude and that of another—for instance, as X increases, Y might also increase, or it could decrease. The difference is that while correlation measures the strength of an association between two variables, regression quantifies the nature of the relationship. [7]

A simple linear regression model has the form:

$$Y = \beta_0 + \beta_1 X_1$$

Where:

- Y is the dependent variable (Target Variable).
- X_1, X_2 are the independent (explanatory) variables.
- β_0 is the intercept term.
- β_1 is the regression coefficient for the independent variable.

When there are multiple predictors, the equation is extended to accommodate them: Multiple Linear Regression. Instead of a line, we now have a linear model—the relationship between each coefficient and its variable (feature) is linear.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Where:

- Y is the dependent variable (Target Variable).
- $X_1, X_2, \ldots, X_k$ are the independent (explanatory) variables.
- β_0 is the intercept term.
- $\beta_1, ..., \beta_k$ are the regression coefficients for the independent variables.
- ε is the error term.

B. ARIMA

The Autoregressive Integrated Moving Average (ARIMA) [8] model utilizes time-series data and statistical analysis to interpret the data and forecast future values. ARIMA aims to understand data patterns by analyzing its past values and employs linear regression to make predictions.

The ARIMA model is typically denoted with the parameters (p, d, q), which can be assigned different values to modify the model and apply it in different ways.

Some of the limitations of the model are its dependency on data collection and the manual trial-and-error process required to determine parameter values that fit best.

Meaning of each component in the Arima model:

- **AutoRegression (AR)**: refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.

The form of AR is:

$$Y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \varepsilon_t$$

Where:

- Y_t is the current value.
- α_0 is the constant term.
- p is the number of orders.
- $\alpha_1, ..., \alpha_p$ are the auto-regression coefficient.
- ε_t is the error term.

- **Integrated (I)**: represents the differencing of raw observations to allow the time series to become stationary.
- **Moving Average (MA)**: incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

The form of MA is:

$$Y_t = \beta_0 + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \cdots + \beta_q \varepsilon_{t-q} + \varepsilon_t$$

Where:

- Y_t is current observed value.
- $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$ are forecast error.
- β_0 is the intercept term.
- $\beta_1, ..., \beta_q$ mean values of Y_t and moving average coefficients.
- ε_t random forecasting error of the current period. The expected mean value is 0.
- q is the number of past errors used in the moving average.

C. HOLT-WINTERS

The Holt-Winters model is a time series forecasting method developed by Charles Holt and Peter Winters in 1960. It is a linear model used to forecast values in a time series with trends and seasonality that vary over time.

Holt-Winters is a model of time series behavior. Forecasting always requires a model, and Holt-Winters is a way to model three aspects of the time series: a typical value (average), a slope (trend) over time, and a cyclical repeating pattern (seasonality). [9]

The formula for the basic version of the Holt-Winters model (simple exponential smoothing) is:

$$F(t+1) = \alpha Y(t) + (1-\alpha)(F(t) + T(t))$$

Where:

- $Y(t)$ is the value at time t.
- $F(t)$ is the forecasted value at time t.
- $T(t)$ is the trend value at time t.
- α is the model's smoothing constant ranging from 0 to 1, determining the importance of past values for the current forecast.

The formula for the enhanced version of the Holt-Winters model (Holt's linear exponential smoothing) is:

$$F(t+1) = \alpha Y(t) + (1-\alpha)(F(t) + T(t))$$
$$T(t+1) = \beta * (F(t+1) - F(t)) + (1-\beta) * T(t)$$

Where:

- β is the model's coefficient for trend, ranging from 0 to 1, determining the importance of trend changes for the forecast.

The formula for the enhanced version of the Holt-Winters model with seasonality adds a seasonal factor to the enhanced model's formula.

$$F(t+1) = \alpha(Y(t) - S(t-m)) + (1-\alpha)(F(t) + T(t))$$
$$T(t+1) = \beta * (F(t+1) - F(t)) + (1-\beta) * T(t)$$
$$S(t+1) = \gamma(Y(t) - F(t+1)) + (1-\gamma) * S(t-m+1)$$

Where:

- $S(t-m)$ is the seasonal value at time $t-m$, with m being the number of repetitions in the seasonal cycle.
- γ is the model's coefficient for seasonality, ranging from 0 to 1, determining the importance of seasonal changes for the forecast.

D. SUPPORT VECTOR MACHINE - SVM

Support Vector Regression (SVR) is a type of machine learning algorithm used for regression analysis. The goal of SVR is to find a function that approximates the relationship between the input variables and a continuous target variable, while minimizing the prediction error. [10]

Support Vector Regression (SVR) uses the same principle as SVM, but for regression problems. [10]

SVR can handle non-linear relationships between the input variables and the target variable by using a kernel function to map the data to a higher-dimensional space. This makes it a powerful tool for regression tasks where there may be complex relationships between the input variables and the target variable. [10]

TABLE 2. Kernels used in the Support Vector Machine

Kernel	Equation
Linear	$x^T z$
Polynomial	$\left(r + \gamma x^T z\right)^d$
RBF	$\exp\left(-\gamma \frac{\|x-z\|_2^2}{2\sigma^2}\right), \gamma > 0$
Sigmoid	$\tanh\left(\gamma x^T z + r\right)$

V. RESULT

A. EVALUATION METHODS

B. VHM DATASET

C. NVL DATASET

D. NLG DATASET

VI. CONCLUSION

A. SUMMARY

B. FUTURE CONSIDERATIONS

ACKNOWLEDGMENT

First and foremost, we would like to express our sincere gratitude to **Assoc. Prof. Dr. Nguyen Dinh Thuan** and **Mr. Nguyen Minh Nhut** for their exceptional guidance, expertise, and invaluable feedback throughout the research process. Their mentorship and unwavering support have been instrumental in shaping the direction and quality of this study. Their profound knowledge, critical insights, and attention to detail have significantly contributed to the success of this research.

This research would not have been possible without the support and contributions of our mentors. We would like to extend our heartfelt thanks to everyone involved for their invaluable assistance, encouragement, and belief in our research. Thank you all for your invaluable assistance and encouragement.

REFERENCES

[1] Ghosalkar, N. N., and Dhage, S. N. (2018). Real Estate Value Prediction Using Linear Regression. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). doi:10.1109/iccubea.2018.8697639.

[2] Lin, Y., Guo, H., and Hu, J. (2013). An SVM-based approach for stock market trend prediction. The 2013 International Joint Conference on Neural Networks (IJCNN). doi:10.1109/ijcnn.2013.6706743.

[3] Ariyo, A. A., Adewumi, A. O., and Ayo, C. K. (2014). Stock Price Prediction Using the ARIMA Model. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. doi:10.1109/uksim.2014.67.

[4] Istiake Sunny, M. A., Maswood, M. M. S., and Alharbi, A. G. (2020). Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model. 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES). doi:10.1109/niles50944.2020.9257950.

[5] Jaiswal, R., and Singh, B. (2022). A Hybrid Convolutional Recurrent (CNN-GRU) Model for Stock Price Prediction. IEEE, doi:10.1109/CSNT54456.2022.9787651.

[6] C. Syavasya and A. L. Muddana, "Machine learning based Time series prediction using Holt-Winters Exponential Smoothing with Multiplicative Seasonality," IEEE, doi:10.1109/ICEECCOT52851.2021.9708006.

[7] P. Bruce, A. Bruce, and P. Gedeck, Practical Statistics for Data Scientists: 50+ Essential Concepts Using r and Python. Sebastopol, CA: O'Reilly Media, 2020.

[8] "Autoregressive Integrated moving average (ARIMA)," Corporate Finance Institute, https://corporatefinanceinstitute.com/resources/data-science/autoregressive-integrated-moving-average-arima/ (accessed May 5, 2024).

[9] "Autoregressive Integrated moving average (ARIMA)," Corporate Finance Institute, https://corporatefinanceinstitute.com/resources/data-science/autoregressive-integrated-moving-average-arima/ (accessed May 5, 2024).

[10] A. Sethi, "Support vector regression tutorial for machine learning," Analytics Vidhya, https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/ (accessed May 15, 2024).