COMPARING ML/DL ALGORITHMS IN REAL ESTATE STOCK PRICES ANALYSIS AND PREDICTION

1st Ngo Thuy Yen Nhi
dept. name of organization (of Aff.)
name of organization (of Aff.)
Ho Chi Minh City, Viet Nam
e-mail: 21521230@gm.uit.edu.vn

2nd Nguyen Duong
dept. name of organization (of Aff.)
name of organization (of Aff.)
Ho Chi Minh City, Viet Nam
e-mail: 21521990@gm.uit.edu.vn

3rd Pham Duy Khanh
dept. name of organization (of Aff.)
name of organization (of Aff.)
Ho Chi Minh City, Viet Nam
e-mail: 21522211@gm.uit.edu.vn

4th Nguyen Ngoc Ha My
dept. name of organization (of Aff.)
name of organization (of Aff.)
Ho Chi Minh City, Viet Nam
e-mail: 21522351@gm.uit.edu.vn

5th Le Thuan Hieu
dept. name of organization (of Aff.)
name of organization (of Aff.)
Ho Chi Minh City, Viet Nam
e-mail: 21522072@gm.uit.edu.vn

***Abstract*—Machine learning and deep learning techniques is renowned for playing a pivotal role in the realm of Business Analytics nowadays by leveraging time-series datasets with the aspiration to provide businesses with analytics-driven insights that assist in facilitating decision-making, improve performance, and optimize processes. Due to the rapid and diverse development of algorithms, there aries a critical need to scrutinize the efficacy of them within specific domains. This paper addresses this imperative by utilizing and evaluating five distinct algorithms - FEDformer, TBATS, AR-MOS, Kalman Filter, and ResNet towards three different time-series datasets from three distinguished real estate companies in Viet Nam. Within the context of this study, the objective is to analyze and forecast the trajectory their stock prices based on their historical daily stock price records. The study delves into an extensive comparative analysis of these algorithms' capabilities and outlines a detailed methodology that illuminates the intricacies of each algorithm's application and the rationale behind their selection. Subsequently, the study applies the findings obtained from the completed analyses to elucidate the strengths and limitations of each approach. Ultimately, through systematic rigorous evaluation and comparison, this paper aims to enhance the understanding of the suitability and competently of these algorithms for stock prices analysis and prediction within the real estate domain in Viet Nam.**

***Index Terms*—Analytic, AR-MOS, Comparison, Efficacy, FEDformer, Kalman Filter, Market capitalization, Prediction, Real estate, ResNet, Stock prices, TBATS, Time-series.**

I. Introduction

In the contemporary era shaped by rapid societal and technological advancements, a typical modern business or industrial organization is constantly in demand of a cutting-edge and robust analytical tools for effective decision-making and risk management. Therefore, there arises a pressing need for advanced analytical tools and methodologies to navigate and extract insights from vast volumes of data.

In general, the world, and specifically Viet Nam, the real estate sector has experienced rapid growth and transformation in recent years, driven by urbanization, economic development, and foreign investment. Amidst this dynamic landscape, accurate estate analysis and forecasting play a pivotal role in facilitating informed decision-making and strategic planning for real estate companies. Stock prices, as a key indicator of a company's value and performance, serves as a crucial metric for investors, analysts, and industry stakeholders alike. Accurate prediction of market capitalization not only aids investors in decisions making but also assists real estate companies in strategizing their operations and investments effectively.

Given the importance need for robust analysis and predictive models within the real estate domain in Viet Nam, Machine learning and Deep learning techniques have emerged as powerful tools in this regard, offering the potential to uncover patterns, trends, and predictive signals within time-series datasets. Against this backdrop, this study investigates the capabilities of various machine learning and deep learning algorithms in analyzing and forecasting stock prices for three prominent real estate companies in Vietnam (DXG, DIG and NVL). Ultimately, the objective is to elucidate the efficacy and comparative performance between the five methodologies of ML/DL algorithms - FEDformer, TBATS, AR-MOS, Kalman Filter, and ResNet by leveraging three different time-series datasets based on historical daily stock price records of those three Vietnamese prominent real estate companies.

The primary goal of this paper is twofold as it embarks on a comprehensive examination of the five algorithms methodologies:

Firstly, to evaluate the performance and effectiveness of the five selected algorithms in analyzing and forecasting daily

stock price data for the specified real estate companies. Secondly, to conduct a comparative analysis of their efficacy and performance, highlighting the strengths as well as limitations of each approach, and examine the suitability and competency of these algorithms in the context of real estate market analysis.

The structure of this paper is organized to provide a comprehensive understanding of these objectives to outline the suitability and competency of these algorithms. Following this introduction, the subsequent section – Related Work, reviews existing literature and research relevant to ML/DL applications in real estate analysis. Then, the Material section provides the dataset descriptions, data selection criteria, and descriptive statistics employed in this study. Subsequently, in the Methodology section, the study presents the criteria for algorithm selection, and the evaluation metrics utilized in this investigation. In the following section, the Results section presents the findings derived from the application of each algorithm and the evaluation methods deployed. Finally, the Conclusion section summarizes the key findings, discusses their implications, and outlines avenues for future considerations.

II. Related Works

Stock prices can be influenced due to various factors such as company's status, market sentiment, or economic indicators, etc. In recent years, a plethora of studies had been dedicated for analysing and predicting stock prices, employing a variety of ML/DL algorithms, and statistical models. The methodologies encompass statiscal and analytical approaches achieved through a diverse selection of those ML/DL techniques and hybrid forecasting methods.

Therefore, this section focuses on critically examining existing literatures and research papers pertinent to the comparative performance of various ML/DL algorithms as well as statistical models' applications in stock price analysis and prediction. The aim is to contextualize recent studies, identifying key trends, methodologies, and findings in the field and provide an overview of the ML/DL algorithms used in this work comparing with others in different scenarios. This will assist further investigation and provide suggested selection of methodologies applied in the study to evaluate and determine the efficacy of each ML/DL algorithms and models utilized in this research.

A research article released in 2023 by Agus Tri Haryono, Riyanarto Sarno, and Kelly Rossa Sungkono suggested that the architecture of FEDformer, another transformer of time-series, outperformed its counterparts such as AutoFormer, Informer, and Reformer in predicting stock prices. Additionally, the performances of AutoFormer, Informer, and Reformer were also stated to be more advanced compared to ARIMA and LSTM models. For stock prices forecasting, the evaluation methods deployed including R2, RMSE, MSE, and MAPE with the time lag variants of 5, 10, 20, 100, and 200 for a specified stock issuer. The outcomes revealed that FEDformer possessed a superior efficacy with an RMSE score 83.08% lower than Transformer and 84.83% lower than Informer. Futhermore, the MAPE score of FEDformer was also 85.17% lower than Transformer and lower by 96.66% compared to Informer. Without a doubt, FEDformer emerges as a dominant selection among other transformer models in resolving limitations on trend capture in long-term time series forecasting. Another study conducted in 2022 by Sasha S. Yamada and Ogulcan E. Örsel issued an efficacy comparison between a linear Kalman filter and different varieties of LSTM to forecast future stock prices based on historical stock prices data spanning on a 10-year period. The research found that the accuracy of each model is significantly influenced by the volatility of the stock being forecasted. With the training dataset consisting of 7.5 years of historical stock prices, the testing set used used to predict next-day stock prices consists of 2.5 years of prices. By applying RMSE and MAE evaluation method, the findings showed that Kalman Filter demonstrated an RMSE score of 62.67% and MAE score of 87.65% higher than average RMSE and MAE score of different LSTM variances. As a result, for low-volatility stocks, a linear Kalman filter can predict next-day prices with very reasonable accuracy. However, this error increases significantly for greater volatile stocks, making LSTM architectures a much more suitable choice for forecasting in such scenarios.

In a blog published in 2021 by Nadeem, the efficacy of TBATS models in stock prices time-series forecasting was proposed, compared with SARIMA and SARIMAX with 2 Fourier Terms. By utilizing MAE evaluation method, the study indicated the dominance of TBATS with an MAE score of 46.6% and 1.19% lower than the other 2 models. Although TBATS still possesses a few drawbacks as it can slow the computation down, this model is preferable when dealing with time-varying seasonality and is user-friendly for handling data with multiple seasonal patterns.

In a study conducted by Bowen Song and Heng Liu in 2018 titled "Stock Price Trend Prediction Model Based on Deep Residual Network and Stock Price Graph", ResNet model was introduced for stock prices prediction and exhibited the average accuracy of 0.40, which surpassed the stochastic indicator of 0.33 when using the stock price graph as input. The paper compared the ResNet model with three selected widely used financial time-series prediction models: SVM, DNN, and CNN. Overall, ResNet and CNN variants demonstrated the highest accuracy and showcased the most stablility in the classifier evaluation index. The deep learning models, such as ResNet, CNN, and DNN, were found to be superior to SVM due to their ability to capture the hidden dynamics of stocks with their complex network structure. Additionally, CNN and ResNet filter features through convolution operations, optimize the learning process, and mitigate overfitting compared to DNN, thereby improving accuracy.

Collectively, these studies contribute to the increasing collection of literature on ML/DL applications in real estate stock price prediction, exhibiting the versatility and effectiveness of various models across different market conditions and time frames. By synthesizing key findings and methodologies, this section provides a comprehensive overview of the state-of-the-

art techniques and avenues for future research in the field.

III. Materials

A. Dataset

In this article, we scrutinize the stock prices data of three following prominent real estate companies in Viet Nam:

- Dat Xanh Group (DXG),
- Development Investment Construction JSC (DIG),
- No Va Land Investment Group Corporation (NVL).

Each dataset presents a record of each company's stock prices, thoroughly documented spanning from January 1, 2018, to March 22, 2024, with a daily time frame on trading activities. These datasets are sourced from Investing.com, a reputable financial platform renowned for providing comprehensive trading information for companies or financial organization.
The datasets feature the following array of variables, providing crucial insights and materials for the study. The variables include:

- Date: Presenting the trading date in the format DD/MM/YYYY.
- Price: Denoting the closing price of the stock on the respective trading day.
- Open: Signifying the opening price of the stock on the respective trading day.
- High: Presenting the highest recorded of the stock price within the trading day.
- Low: Indicating the lowest recorded of the stock price during the trading period.
- Vol: Evaluating the trading volume for each day, represented by the number of shares traded (expressed in millions of shares).
- Change (%): Determining the percentage change in the stock's value compared to the preceding trading day.

All data referring to stock-related metrics are denominated in Vietnamese Dong (VND), ensuring consistency and coherence throughout the analysis.

B. Descriptive Statistics

TABLE I
DXG, DIG, NVL's Descriptive Statistics

Descriptive Statistic	Company		
	DIG	***DXG***	***NVL***
Count	1552	1552	1552
Mean	20575.539	17846.990	41771.515
Std	16677.794	7144.933	23474.855
Min	6591.800	6739.100	10250
25%	8852.700	12814.250	28928
50%	14799.700	17000	34248
75%	25800	20586.350	60244.250
Max	98196.700	46750	92366

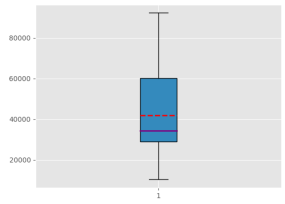

Fig. 1. NVL stock price's boxplot

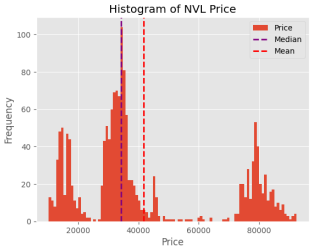

Fig. 2. NVL stock price's histogram

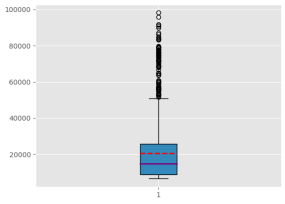

Fig. 3. DIG stock price's boxplot

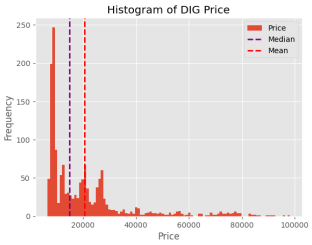

Fig. 4. DIG stock price's histogram

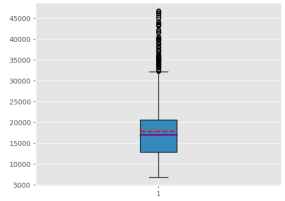

Fig. 5. DXG stock price's boxplot

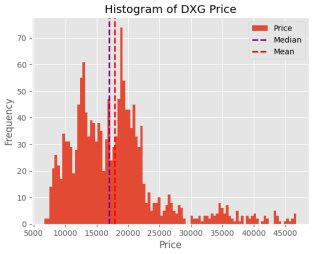

Fig. 6. DXG stock price's histogram

IV. Methodology

A. Linear Regression

Linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data which are called linear models. Linear regression model follows a very particular form, a regression model is linear when all terms in the model have a constant or a parameter multiplied by an independent variable. And by adding the terms together, the equation is formed:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

Where:

- Y: dependent variable (affected by other variables)
- X_1, X_2, X_p: independent variables (affecting other variables)
- β_0: regression constant (or intercept). This is the index indicating the value of Y when all X are 0 (no X). When represented on the XY graph, β_0 is the point on the Oy axis where the regression line intersects.

- $\beta_1, \beta_2, \beta_p$: regression coefficients, also known as slope coefficients. This index indicates the level of change in Y caused by the corresponding X.
- ϵ: error term. This index, the larger it is, makes the prediction accuracy of the regression less accurate or more deviated from reality.

B. ARIMA

ARIMA stands for Autoregressive Integrated Moving Average. The ARIMA model is based on the Box and Jenkins method of using three different concepts: autoregressive (AR) model, moving average (MA) model, and integration, together classified as an ARIMA (p, d, q). It is a quantitative forecasting model over time, where the future value of the predictor variable will depend on the movement trend of that object in the past. The model contains three components/parameters: AR + I + MA. AR is denoted as p, where it shows the weighted linear of sum p values based on ARIMA (p, d, q) terminology. The p-value indicates the number of orders. The formula to denote this AR is shown as:

$$\phi_1 = \phi_1 + \phi_2\delta - \delta x_0 + \phi_3\delta - \delta x_1 + \ldots + \delta e_{t-1} = e_t \quad (1)$$

Where p is used to determine the number of orders of past values; t is the time series; ϕ is the coefficient of the AR model; e is the error term with mean zero and variance σ^2.

MA process is denoted by order q in the ARIMA (p, d, q) classification which shows an error value in Equation (4), it also uses the number of orders in the past values, as denoted in Equation (5)

$$x_t = c + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q} + e_t \quad (2)$$

Where t is the time series; θ is the slope coefficient; q is the number of orders needed to identify the past values. To identify how many orders are in the calculation of AR, the parameter of q is used; c is the intercept

Integrated or differentiated versions are denoted as d in ARIMA (p, d, q), which is the number of times the time series got different.

$$I = \Delta x_t = x_t - x_{t-1} \quad (3)$$

Therefore, The ARIMA(p, d, q) can be represented in the following equation:

$$Y_t = c + \varphi_1 Y_{t-1} + \ldots + \varphi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (4)$$

C. ETS

Exponential smoothing method (ETS) is one of the forecasting model analyses of runtime (time series). This method performs a forecasting approach to the grant of a value weighting over a certain constant on a series of observations of the past to predict future values. This smoothing method adds the parameter alpha (α) in this model to reduce the factor of randomness [1]. The exponential term in the weighting is derived from the methods/scales (factor refinement of previous periods that shaped exponential).

Exponential smoothing model variation depending on case data will be foreseen. The following variation of the exponential smoothing model:

- Simple Exponential Smoothing (SES): Simple Exponential Smoothing (SES) is a time series forecasting model calculation method that assumes the observation data pattern is stationary tend to straight lines. The formula of the SES is as follows [2]:

  $$F_t = \alpha D_t + (1 - \alpha) F_{t-1}$$

  Where: D_t is the actual data value in period t F_t is forecasting data value on a period t α is a constant of refinement for the entire data
- Double Exponential Smoothing (DES): Double Exponential Smoothing (DES) or commonly called Holt's Model is a method of forecasting model calculations that assume the observation data pattern has a trend however no seasonal variations have. This model is also called Trend-Adjusted Exponential Smoothing (TAES) due to an adjustment of the trend on the observation data to influence the accuracy of the results. This method is very good for calculating a linear trend that forecasts short and medium term. The formula of the DES is as follows [3]:

  $$F_t = \alpha D_t + (1 - \alpha)(F_{t-1} + T_{t-1})$$

  $$T_t = \beta(F_t - F_{t-1}) + (1 - \beta) T_{t-1}$$

  Where: T_t is the trend in the period t β is a constant refinement for trends
- Triple Exponential Smoothing (TES): Triple Exponential Smoothing (TEST) or commonly called Holt Winters Model is a forecasting model which assumes that the calculation method of observation data has a trend at once seasonal variations. Constant of Gamma (δ) is the parameter that controls the weighting observation data for estimating the existence of seasonal variations. Calculation of the TES is almost the same with DES but there is mining of gamma constant value to smooth the seasonal component, with the following calculation formula:

  $$L_t = \alpha \left(\frac{D_t}{I_{t-L}} \right) + (1 - \alpha)(L_{t-1} + T_{t-1})$$

  $$T_t = \beta(L_t - L_{t-1}) + (1 - \beta) T_{t-1}$$

  $$I_t = \delta \left(\frac{D_t}{L_t} \right) + (1 - \delta) I_{t-L}$$

  Where: I_t is the seasonal index in the period t δ is a constant smoothing seasonal component L is the length of the season

The value of alpha, beta, and gamma in the range 0 to 1. The larger the value of the constant, then the greater the weighting towards granting any observation data.

D. Kalman Filter

Kalman Filter is a Linear-Gaussian State Space Model, a type of time series prediction algorithm, invented by Rudolf Emil Kalman in 1960. [4]

How Kalman filter works:

Initial estimation: Initialize the initial state $\widehat{x}_{0,0}$ and the initial state covariance matrix $P_{0,0}$.

Iterative process of prediction and measurement update:

1) Prediction
   - Prediction of the next state
   $$\widehat{x}_{n+1,n} = F\widehat{x}_{n,n} + Gu_n$$
   - Prediction of the next state uncertainty (error)
   $$P_{n+1,n} = FP_{n,n}F^T + Q$$
2) Measurement update
   - Calculation of the Kalman Gain prediction weight
   $$K_n = P_{n,n-1}H^T(HP_{n,n-1}H^T + R_n)^{-1}$$
   - Update of the state estimate
   $$\widehat{x}_{n,n} = \widehat{x}_{n,n-1} + K_n(Z_n - H\widehat{x}_{n,n-1})$$
   - Update of the estimate uncertainty (error)
   $$P_{n,n} = (1-K_nH)P_{n,n-1}(1-K_nH)^T + K_nR_nK_n^T$$

   Where:
   - x: The state actor
   - F: The state transition matrix
   - G: The control matrix
   - u: The input variable
   - P: The covariance matrix
   - H: The observation matrix
   - H^T: The transpose of the observation matrix
   - K: The Kalman Gain
   - R: The measurement covariance matrix
   - z: The vector measurement

By combining measurement data and prediction models to estimate the state and reduce noise in dynamic systems, Kalman Filter improves the accuracy of the prediction model.

E. TBATS

TBATS models are a sophisticated class of time series models that integrate several techniques to address complex data patterns.

TBATS employs trigonometric functions to model multiple seasonalities simultaneously, such as daily, weekly, and annual cycles. Moreover, a Box-Cox transformation is applied to stabilize the variance, making the data more suitable for modeling.

TBATS models also incorporate ARMA components to manage short-term dynamics and autocorrelations in the residuals, thereby improving forecast accuracy. Additionally, they allow for damping trends, accommodating trends that decrease over time. These features enable TBATS models to effectively handle complex seasonal patterns, nonlinearities, and residual autocorrelations, making them particularly useful for forecasting data with intricate seasonal structures, such as in the case study of forecasting the second wave of the COVID-19 epidemic case study. [5]

V. Result

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] Hanke, J. E. (2008). Business Forecasting (Vol. 8). New Jersey: Pearson Education International.

[2] Hyndman, R. J., Athanasopoulos, G. (2013). Forecasting: Principles and Practice. Melbourne: OTexts

[3] Rim A., GiLang R. A., Sari S. W. Time series forecasting using exponential smoothing to predict the number of website visitor of Sebelas Maret University. (2015b, October 1). IEEE Conference Publication

[4] Alex Becker (www.kalmanfilter.net). (n.d.). Online kalman filter tutorial. https://www.kalmanfilter.net/multiSummary.html

[5] Perone, G. (2021). Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy. the European Journal of Health Economics, 23(6), 917–940. https://doi.org/10.1007/s10198-021-01347-4