Stock Price Analysis and Prediction for Vietnam Stock Market using Machine Learning

Ta Nguyen Thanh
Institute for Artificial Intelligence
UET-VNU
23020437@vnu.edu.vn

October 24, 2025

Contents

1 Introduction

The Vietnamese stock market is one of Southeast Asia's most dynamic capital markets, offering significant opportunities alongside inherent risks due to its characteristic volatility. Predicting stock movements is a major challenge, as prices are influenced by a complex mix of economic, political, and social factors that traditional analysis often struggles to model effectively.

This report tackles this challenge by applying machine learning techniques to forecast stock price movements. Our objective is to develop a predictive model from historical data that can serve as a quantitative tool for investors, aiding in more informed, data-driven decision-making. The following sections will detail our methodology, model development, and a thorough performance evaluation, aiming to demonstrate the value of computational methods in navigating Vietnam's evolving market.

2 Data collecting

Code and data for this project can be found in this GitHub Repository. The notebooks for crawling and analysis are presented in folder **notebook**. All the data of Vietnam Stock Market is gathered from the VnStock API. We've collected information for 1721 companies, with fields such as symbol, exchange, time, organisation name - short name, product-group id. For each company, we call the API to get the stock data from January 1st 2012 up until October 15th 2025. In case a company joined the stock market later than January 1st 2012, then its stock data will be collected since its own first date. After crawling data and removing companies whose data is corrupted, there are 1688 companies left. Each company's data is written to a CSV file and comprises of 5 fields:

- time (the date when the stock is on the market)
- open (price at the start of the trading day)
- close (price at the end of the trading day)
- high (highest price recorded throughout the trading day)
- low (lowest price recorded throughout the trading day)
- volume (number of trades throughout the trading day)

These CSVs strictly follow the OHLCV market data format. After data is collected, they are sent to HDFS system before being processed by Spark.

3 System Architecture

The analysis presented in this report was conducted within a distributed computing environment, built using Docker to simulate a big data ecosystem. The entire setup is orchestrated through a docker-compose file, ensuring a reproducible and isolated environment for data processing and analysis.

The system is comprised of two main clusters:

Hadoop Cluster: At the core, a Hadoop Distributed File System (HDFS) is used for robust, distributed storage of the stock market data. This cluster includes a central namenode for managing the file system namespace and four datanode services for storing data blocks across multiple nodes. Resource management is handled by Hadoop YARN, with a dedicated resourcemanager and nodemanager. A historyserver is also deployed to provide a web UI for monitoring completed application logs.

Spark Cluster: Apache Spark serves as the primary engine for large-scale data processing. The cluster is configured in standalone mode, consisting of a spark-master node that coordinates tasks and spark-worker nodes that perform the computations.

All services, including the Hadoop and Spark components, are connected on a unified Docker bridge network (sparknet), allowing for seamless inter-container communication. Data analysis and model training were performed interactively using a Jupyter pyspark-notebook, which connects directly to the Spark master to execute distributed jobs on the cluster. This containerized architecture provides a scalable and efficient platform for big data analysis.

4 Data Analysis

To gain a deeper understanding of the market's behavior, we conducted an analysis of historical data from five major, influential stocks on the Vietnamese stock market: Vingroup (VIC), Vinhomes (VHM), BIDV (BID), VPBank (VPB), and Hoa Phat Group (HPG). This selection represents key sectors of the economy. The analysis was primarily focused on key financial metrics derived from daily trading data, including the open, high, low, and closing prices, which form the basis for most technical analyses.

4.1 Stock Daily Mean Price

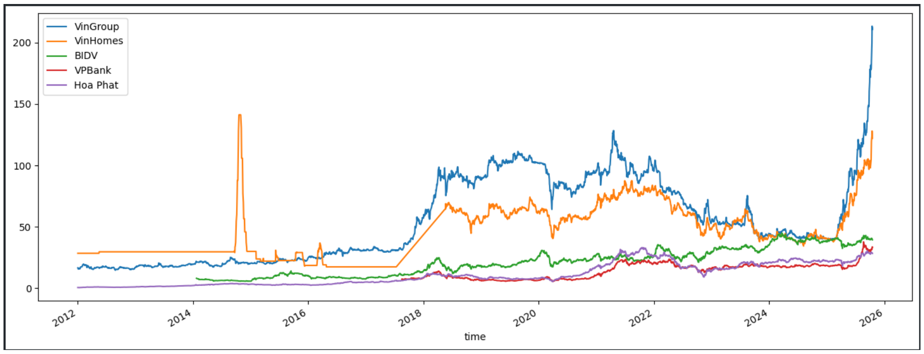

Figure 1: Stock Daily Mean Price

Our initial step involved calculating the daily mean price for each stock, a technique used to smooth out intraday fluctuations and reveal clearer long-term trends. As depicted in the line graph in Figure 1, we can infer that, in overall terms, the value of these selected stocks is increasing steadily over the observed period. Notably, VIC and VHM, which are the two stocks with the highest capitalization on the market, show significant growth. Despite experiencing a mild decrease between 2022 and 2024, they still managed to outgrow the other stocks and show a pronounced surge in their values during the latter half of 2025.

4.2 Volatility from Daily Returns

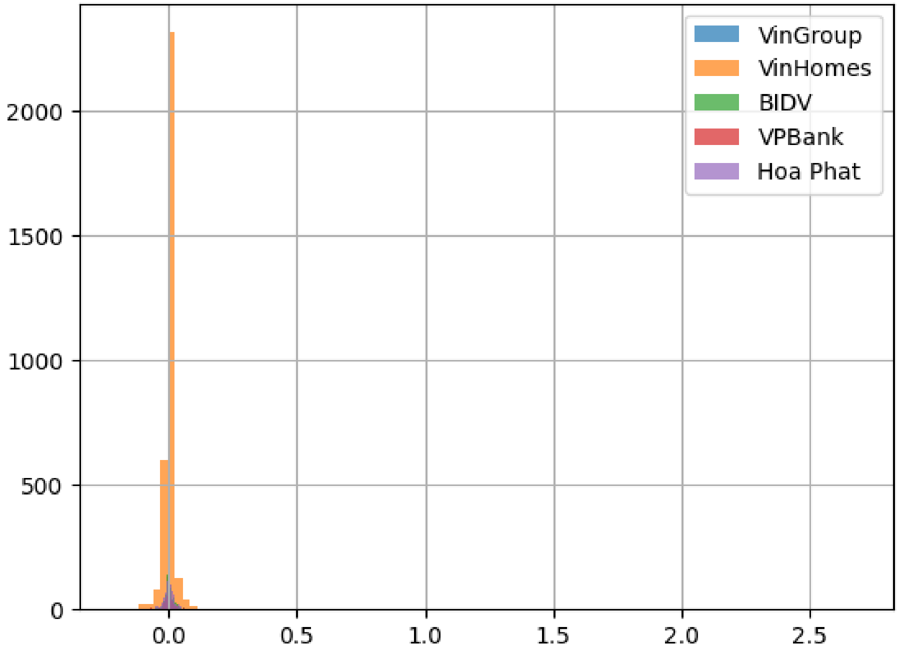

Figure 2: Volatility from Daily Returns

Building on the price data, we derived the daily percentage return. This metric is essential as it measures the day-to-day volatility and performance of each stock, independent of its absolute price. To obtain a quantitative measure of this volatility, the standard deviation of these daily returns was then calculated. This statistical analysis revealed that VHM exhibited the highest volatility among the selected stocks during the observed period. A visualization of this volatility distribution is displayed in the histogram in Figure 2.

4.3 Correlation Matrix

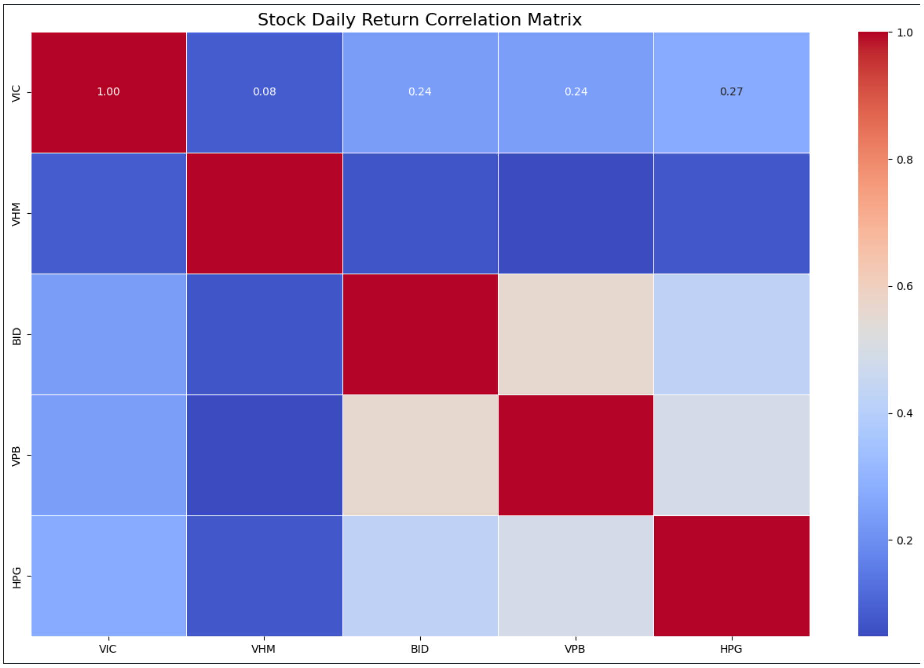

Figure 3: Heatmap for Stock Correlation

Finally, to examine how these stocks move in relation to one another, a correlation matrix of their daily returns was generated. This matrix, presented as a heatmap in Figure 3, provides insight into potential portfolio diversification. The results showed a particularly strong positive correlation between the two banking stocks, BIDV and VPBank. This suggests that their prices tend to move in the same direction, likely driven by shared sector-specific factors. Conversely, other stocks, such as those in real estate and industry, showed weaker correlations with each other and with the banking sector, indicating a degree of market diversification.

4.4 Traded Value and Volume

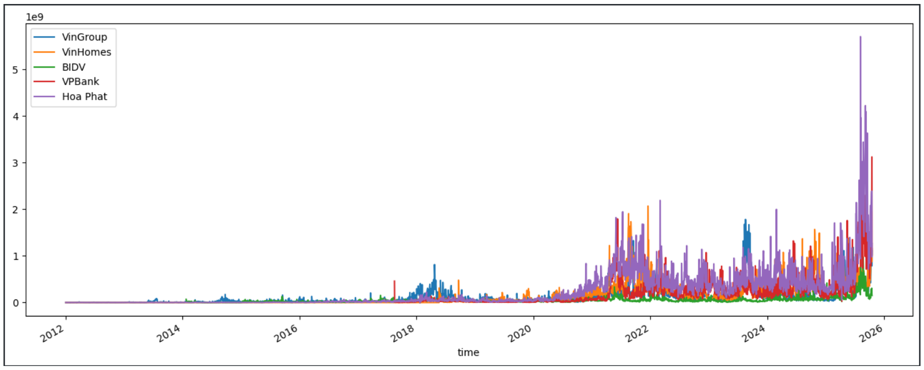

Figure 4: Traded Value

To gain a more granular view of market activity and liquidity, we analyzed the daily traded value, which is calculated by multiplying the closing price by the daily volume. This metric, visualized in Figure 4, represents the total monetary value of shares exchanged daily and serves as a strong indicator of market interest and capital flow. The plot highlights periods of significant market engagement, with VIC and VHM again demonstrating exceptionally high traded values, confirming their dominant role in attracting market capital. Furthermore, the raw trading volume for each stock was plotted individually in Figure 5 and Figure 6. This analysis allows for a clearer view of the ebb and flow of trading activity over time, revealing spikes in volume that often correspond to significant price movements or major news events for each specific company.

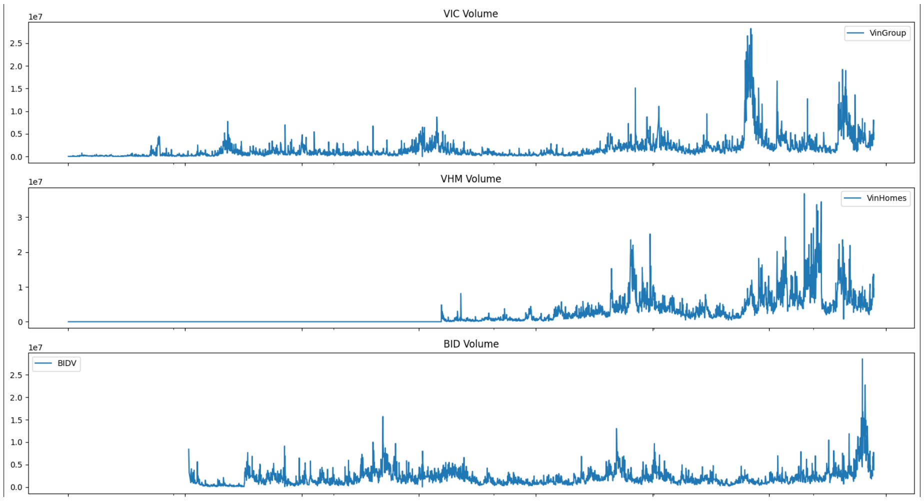

Figure 5: Traded Volume for 3 Biggest Companies

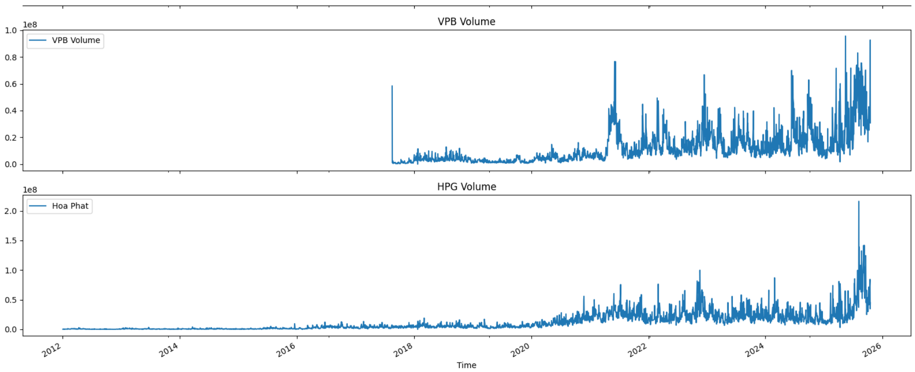

Figure 6: Traded Volume for smaller 2 Companies

4.5 Moving Average Analysis

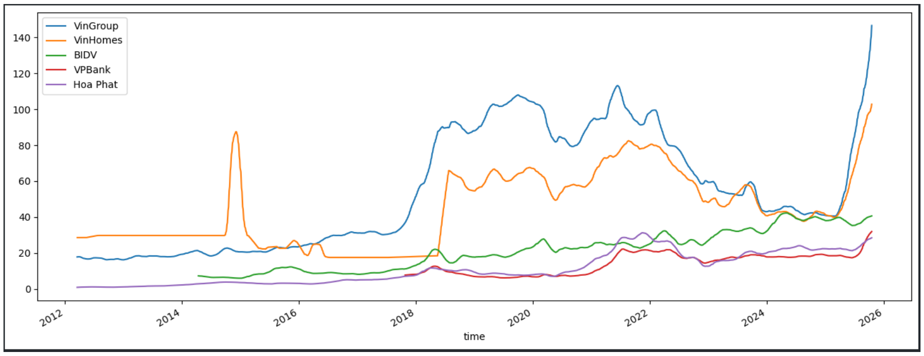

Figure 7: 50-day Moving Average

Finally, to smooth out price data and better identify underlying trends, we calculated two key Simple Moving Averages (SMA) for each stock's closing price. The 50-day SMA, plotted in Figure 7, is a common short-term indicator used to gauge recent price momentum. It reacts more quickly to price changes, helping to identify the current, immediate trend. In contrast, the 200-day SMA, shown in Figure 8, is a standard long-term trend indicator. By averaging the price over a much longer period, it provides a clearer view of the established, underlying market direction and is often used by investors to determine the primary market cycle (bullish or bearish).

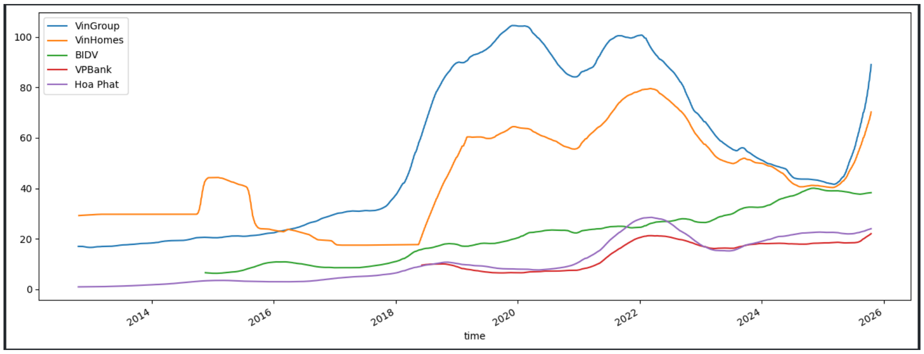

Figure 8: 200-day Moving Average

5 Stock Price prediction with Machine Learning

For the predictive modeling portion of this report, we transitioned from broad market analysis to a specific case study: forecasting the stock price of Vinamilk (VNM). The primary objective was to build a functional model capable of predicting future price movements based purely on its own historical time-series data. To achieve this, we employed a Long Short-Term Memory (LSTM) neural network. This advanced model is a special type of recurrent neural network (RNN) that is exceptionally well-suited for learning long-term dependencies and patterns in sequential data, making it a powerful choice for financial time-series like stock prices.

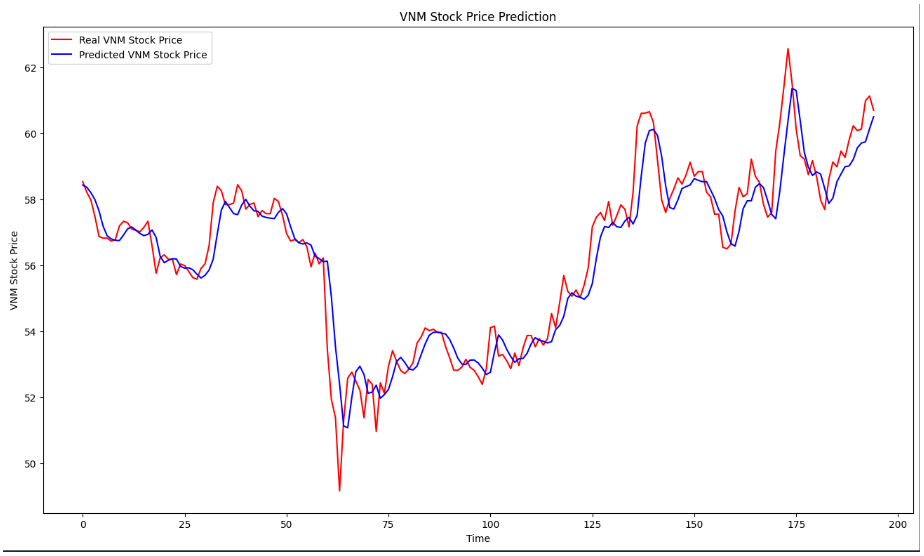

Figure 9: Result of Stock Price Prediction

To properly train and validate the model, the historical data for VNM was carefully divided into two distinct sets: a training set, which contained all available data up to the end of 2024, and a test set, which consisted of data exclusively from 2025. The LSTM model was then trained on the historical training data. It learned to identify patterns from a sequence of the past 60 days of mean prices to predict the price for the single next day.

After the training phase was complete, the model's predictive performance was rigorously evaluated by using it to forecast stock prices for the unseen 2025 test period. The final step in this evaluation involved plotting the model's predicted stock prices directly against the actual, real prices observed in 2025. This visualization, which can be seen in Figure 9, provides a clear and direct comparison of the model's accuracy. It effectively demonstrates the model's ability to capture the complex trends and fluctuations of the VNM stock price throughout the test year.

6 Predicting Stock Price using Spark MLlib

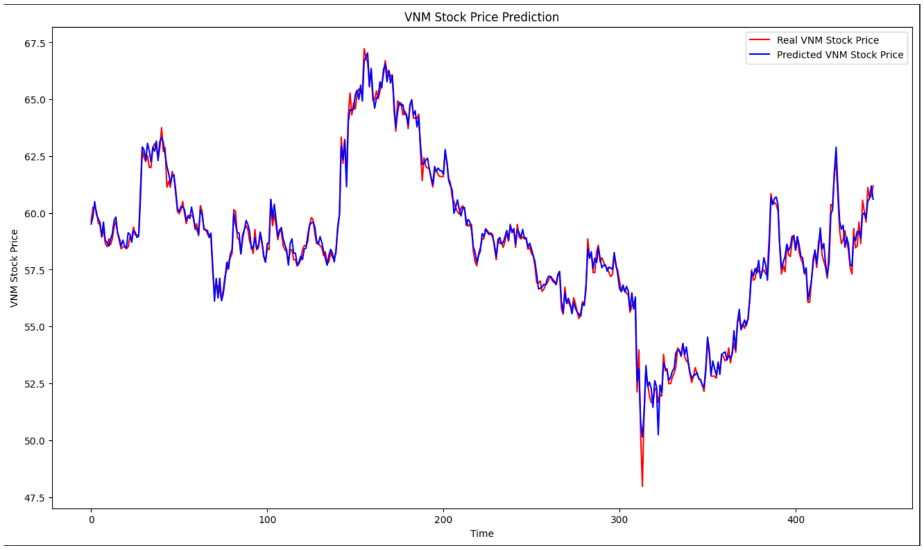

Figure 10: Result of Stock Price Prediction with Spark MLlib

In addition to the LSTM model, we explored an alternative approach using Spark's MLlib library to build a Linear Regression model. This second model differed in its objective, features, and evaluation period.

Instead of predicting the mean price, this model was designed to forecast the closing price. It was trained using a different set of features:

- MA30: A 30-day moving average of the closing price, to provide trend context.
- Open, High, Low: The daily OHL prices.
- Volume: The daily trading volume.

For this model, the data was split differently: the training set included all data prior to 2024, and the test set comprised all data from January 1, 2024, to the present.

The model was trained on the pre-2024 data and evaluated on the 2024-2025 test data. Performance was measured using the Root Mean Squared Error (RMSE), which resulted in a test RMSE of approximately 0.306. This low RMSE value indicates a strong fit, suggesting that the selected features (MA30, OHL, and volume) are highly predictive of the closing price. The comparison between the actual and predicted prices for this model is visualized in Figure 10.

7 Conclusion

This report provides a dual analysis of the Vietnamese stock market, combining statistical data analysis with a machine learning-based prediction model. The data analysis revealed distinct behaviors among key sectors, with a strong correlation observed between banking institutions, while real estate and industrial stocks showed more independent movement. This highlights the potential for portfolio diversification within the market.

The LSTM model developed to predict Vinamilk's (VNM) stock price demonstrated a promising ability to forecast price trends. By learning from historical data, the model was able to generate predictions for 2025 that captured the general direction of the stock's movement. While no model can predict market behavior with perfect accuracy, the results indicate that machine learning techniques, specifically LSTM networks, are valuable tools for forecasting in volatile markets.

For investors, these findings underscore the importance of data-driven analysis. The correlation study can inform diversification strategies, while the predictive model serves as a practical example of how advanced analytics can supplement traditional investment decision-making. Future work could expand upon this by incorporating a wider range of stocks, integrating macroeconomic data, and exploring other machine learning architectures to further enhance predictive accuracy.

References

[1] vnstocks.com, *VNStock documentation*, https://vnstocks.com/docs/vnstock

[2] thviet79, *Stock-Price*, GitHub repository. https://github.com/thviet79/Stock-Price