Stock Returns Prediction: Comparison between ARIMA and Support Vector Regression

Tao He [1,*]

[1] School of Public Administration, Nanjing University of Finance & Economics, Nanjing, China

Abstract. One of the key subjects for academic scholars and professional investors is the forecasting of stock returns and prices. As advancements in computing technologies have progressed, the machine learning approach has garnered significant attention. This study compares and evaluates two forecasting models, the SVR model, and the ARIMA model. The analysis is conducted using data from the NYSE Composite Index, the S&P 500 Index, the Dow Jones Industrial Average Index, and the Nasdaq Composite Index, spanning from January 1, 2012, to September 1, 2023, as the experimental time series. Initially, before applying the ARIMA model, the stationarity of the time series data is verified through ACF and PACF analyses. Grid search and cross-validation methods are then used to pinpoint three main feature factors. The output of both models is then evaluated using metrics like MAE, MSE, and RMSE before prediction values are produced. The SVR model performs better than other models, which suggests that it could be a good option for future stock return forecasting projects based on empirical evidence.

1 Introduction

The anticipation of forecasting stock prices and returns has ignited the enthusiasm of both researchers and investors. While the Efficient Market Hypothesis (EMH) posits that predicting stock market behavior is an insurmountable challenge, advancements in computer science technology and machine learning algorithms have bolstered the prospect of achieving more precise predictions of stock prices and returns [1]. Machine learning models have exhibited outstanding capabilities in capturing non-linear correlations, adjusting to dynamic market dynamics, and effectively managing intricate datasets [2]. Recent years have seen a surge in studies applying machine learning techniques, including Gradient Boost, Random Forest, SVR, as well as deep learning models like neural networks, and LSTM [3-9]. Hybrid models are used widely as well [10]. These studies have highlighted the potential and advantages of machine learning in stock price prediction, meeting the demand for more reliable forecasting methods. However, research still faces challenges, including the need for comparative studies with traditional statistical models and addressing interpretability issues in machine learning models.

To anticipate stock prices, this research compares and contrasts the ARIMA and SVR models. The four primary stock market indices for the period from January 1, 2012, through September 1, 2023 are the Nasdaq Composite Index, S&P 500, Dow Jones Industrial Average, and NYSE Composite Index. The following succinct summary represents the main contributions of this study.

The ACF and Partial PACF are used to first evaluate the stationarity of the time series dataset. Then, a thorough strategy using grid search and cross-validation is applied to locate and utilize the dataset's most relevant feature elements, improving the SVR model's capacity for prediction. Critical metrics like MAE, MSE, and RMSE are computed to systematically assess the performance of models that both produce anticipated values.

The rest of this essay is organized as follows. In section 2, the research's data and methodology are detailed. The experimental findings are laid out and explained in section 3. Section 4 gives the reseach's conclusion.

2 Data and Methodologies

2.1 Data

The four basic prices—highest price, lowest price, opening price—are used to gauge stock market performance. The closing price is one of these variables that is thought to be particularly good at capturing daily market activity and mood. Logarithmic

[*] Corresponding author: taohe@ldy.edu.rs

returns, which are usually computed using daily closing prices, are widely used to assess the performance of particular stock indices. For this study, Yahoo Finance was used to get historical logarithmic daily returns for the NYSE Composite Index, Nasdaq Composite Index, S&P 500, and Dow Jones Industrial Average, four well-known stock indices. The experimental dataset spans the dates of September 1, 2023, and January 1, 2012, respectively.

Fig. 1 displays the distribution of logarithmic returns for four stock indices. Table 1 provides the statistical features of these returns. Over the entire time series, each index has been observed 2934 times. The standard deviation (std) values in Table 1 are all relatively low, hovering around 0.01. This indicates that the logarithmic returns for these indices exhibit limited dispersion from their respective means. Lower standard deviations imply a reduced level of risk and a more predictable pattern of returns. A notable aspect is the presence of asymmetry in the data. While the minimum values (min) are slightly negative for all indices, the maximum values (max) are positive. This asymmetry implies that, over the analyzed time series, the indices experienced more significant positive returns compared to negative returns. This skewness in returns may be of interest to investors and portfolio managers seeking to capitalize on positive market movements. The quartile values (25%, 50%, and 75%) offer insights into the spread and distribution of returns. Across all four indices, the median (50%) is close to zero, indicating that half of the observations fall above and half below the mean. The quartiles (25% and 75%) reveal that the majority of returns are clustered in the range between -0.01 and 0.01. This suggests that most observations fall within a relatively narrow band of returns, reinforcing the concept of limited dispersion. The mean values remain close to zero, indicative of a stable average performance over the entire time series.

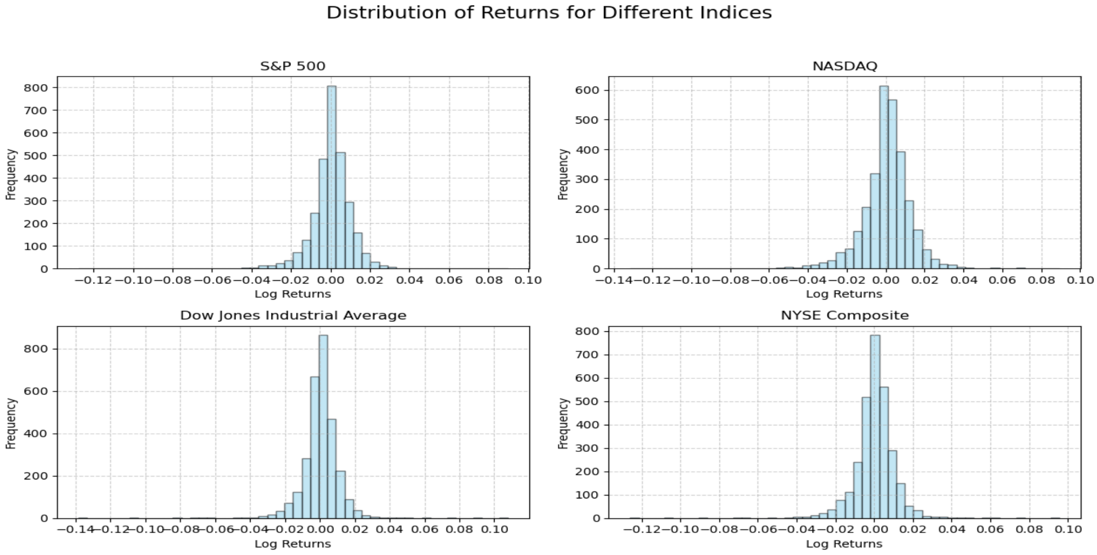

Fig. 1. Visualization of Time Series (Photo/Picture credit: Original)

Table 1. Statistical Description of Time Series

	S&P 500	NASDAQ	Dow Jones Industrial Average	NYSE Composite
count	2934.00	2934.00	2934.00	2934.00
mean	0.00	0.00	0.00	0.00
std	0.01	0.01	0.01	0.01
min	-0.13	-0.13	-0.14	-0.13
25%	-0.00	-0.00	-0.00	-0.00
50%	0.00	0.00	0.00	0.00
75%	0.01	0.01	0.01	0.01
max	0.09	0.09	0.11	0.10

2.2 ARIMA

The Box-Jenkins technique, which was first presented by Box and Jenkins in 1970, is another name for the ARIMA model. This model represents a conventional linear time series forecasting approach renowned for its effectiveness in short-term prediction. The model is comprised of three essential components, which collectively constitute its structure.

Autoregressive (AR) Component: The autoregressive (AR) component depicts the autocorrelation between the present observation and previous observations at various time delays. It shows that a time series' present value is connected to its earlier values.

Integrated (I) Component: To render the time series stationary, the data must be differentiated as part of the integrated (I) component. Differentiating eliminates the impacts of trends or seasonality by deducting the current value from the prior value.

Moving Average (MA) Component: The MA component shows how the white noise error terms for the prior and present data relate to one another. It implies that random noise from the past affects the current value of a time series.

ARIMA(p, d, q) is the representation of the ARIMA model, where:

The "p" parameter indicates the order of the Autoregressive (AR) component, denoting how many past values are considered.

The "d" parameter represents the order of the differencing procedures used to render the time series stationary.

The "q" parameter signifies the order of the Moving Average (MA) component, indicating how many prior error terms are incorporated into the model.

2.3 SVR

Support Vector Regression (SVR), a type of a prediction method frequently employed, is a derivative of the Support Vector Machine (SVM). In essence, SVM finds a hyperplane that divides all data points into discrete categories. If a data point resides above this separating hyperplane, it is classified as positive, while those below it are classified as negative. This traditional SVM model is primarily employed for predicting whether a stock will exhibit an upward or downward trend in the future. However, for more precise forecasting, an extended version of SVM known as SVR is employed to address regression-related challenges. SVR's core ideas are similar to those of SVM, but it has a distinct goal. Finding a function or curve that reduces the gap between training data points and the function while maintaining a predetermined tolerance level (epsilon) is the objective of SVR.

3 Results

3.1 ARIMA

Python is used as the experimental software for this study. Model verification, parameter estimation, and diagnostic assessment are the three fundamental processes in the creation of an ARIMA model. The model has checked the stationarity of the stock data in Fig. 2. The moving average(q) and autoregressive(p) parameters are then searched using the grid search to get the AIC and BIC with the lowest values.

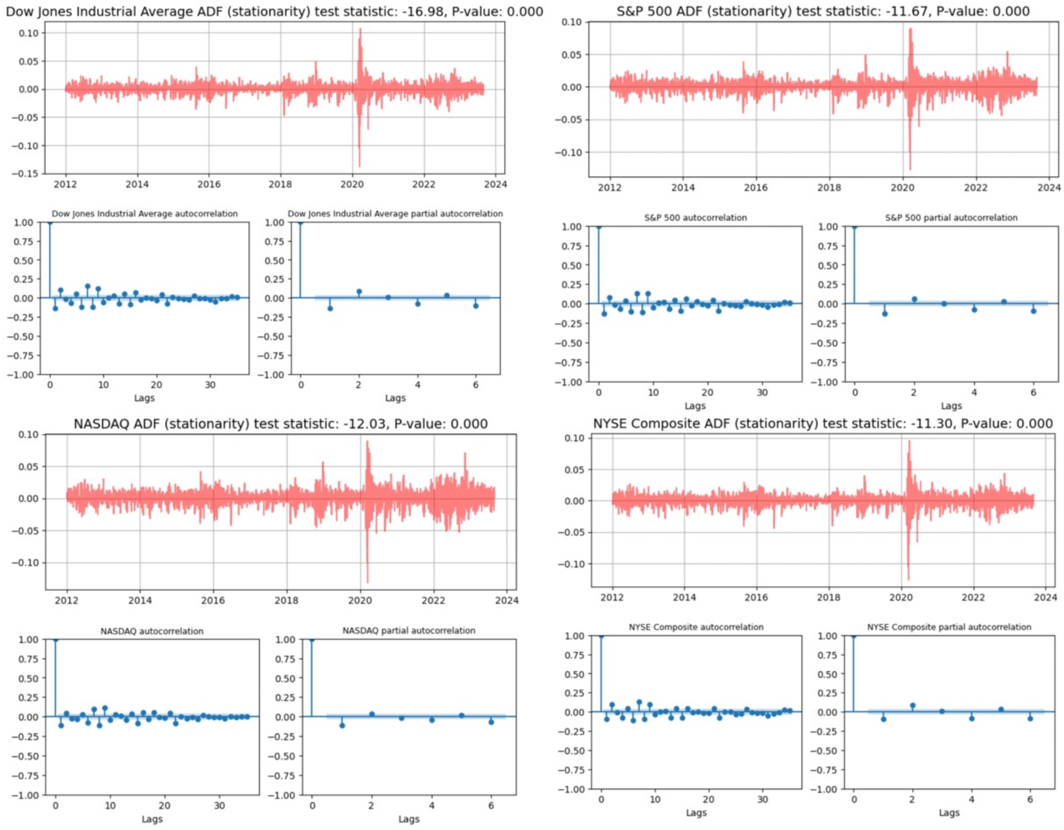

Fig. 2. ADF and PADF results (Photo/Picture credit: Original)

The threshold is set as 0.05. According to the results of the evaluation criteria, the P-value is nearest to zero and the time series showed stationarity.

When the model is set as SARIMAX (0,0,2), AIC and BIC two are the smallest. The degree of accuracy of the anticipated returns compared to the actual returns is graphically depicted in Fig. 3. Both the anticipated values and the actual values exhibit a similar fluctuating pattern. Though not as great as real values, anticipated values do fluctuate to a lesser level.

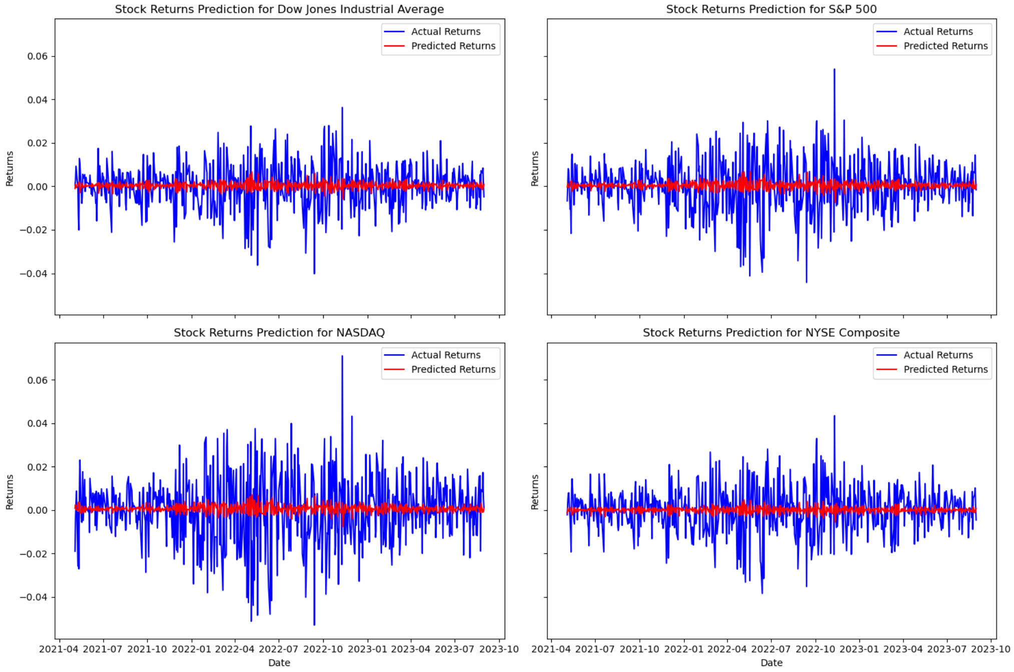

Fig. 3. ARIMA Predicted Result (Photo/Picture credit: Original)

* Corresponding author: taohe@ldy.edu.rs

3.2 SVR

When employing SVR, three approaches are taken into account. The initial step is to divide the data into test and training observations. The second step in feature engineering entails standardizing and lowering the dimensionality. The third step entails parameter definition and model selection using grid search and cross-validation.

Accordingly, the data is split into training sets with approximately 80% business days (2347 observations) with the remaining 20%(587 observations) as test sets, and then relevant features or feature extraction are chosen for model training. These 3 features—C, epsilon, and kernel function—were used to train the model using Principal Component Analysis (PCA). After grid search and cross-validation, the best SVR model selected for prediction has parameters C as 0.1, epsilon as 0.01. The degree of accuracy of the anticipated returns compared to the actual returns is graphically depicted in Fig. 4. The plots demonstrate good forecast accuracy.

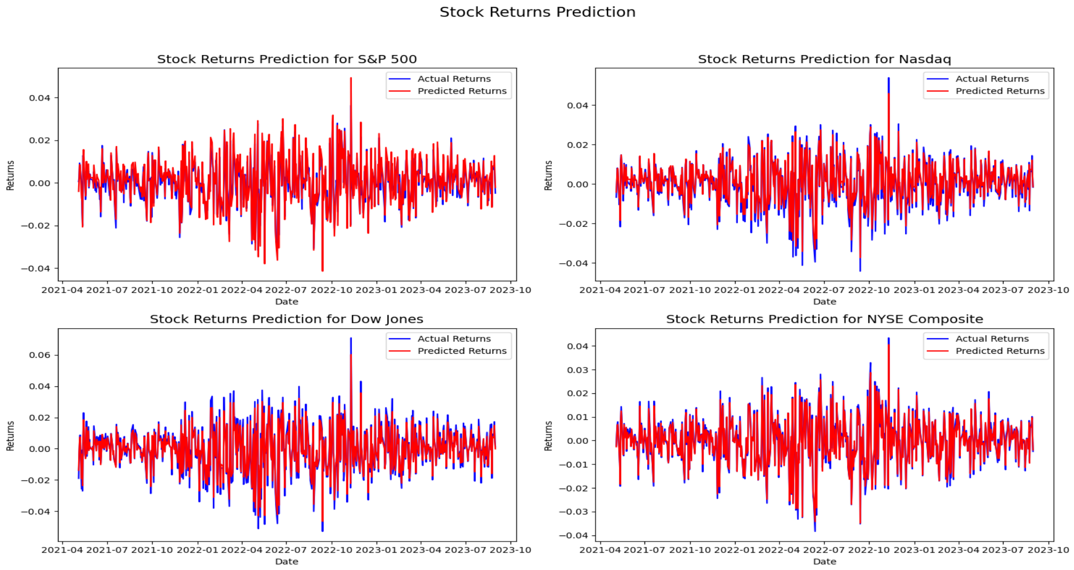

Fig. 4. SVR Predicted Result (Photo/Picture credit: Original)

3.3 Comparison

To quantify the superior performance of SVR, this paper employs several evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). SVR consistently surpasses the ARIMA model with more predicted accuracy across all four stock indexes when comparing the two models, in conclusion. SVR has much better performance as seen by significantly reduced MAE, MSE, and RMSE. This observation holds true across all four stock indexes, demonstrating the robustness and reliability of SVR in forecasting stock prices (See Table 2).

Table 2. Comparison Results

	ARIMA			SVR		
	RMSE	MSE	MAE	RMSE	MSE	MAE
Dow Jones Industrial Average	0.010171	0.000103	0.007645	0.002758	0.000008	0.002133
S&P 500	0.011882	0.000141	0.008936	0.002944	0.000009	0.002323
NASDAQ	0.015607	0.000244	0.011872	0.002158	0.000005	0.001693
NYSE Composite	0.010569	0.000112	0.008043	0.001313	0.000002	0.001029

4 Conclusion

This paper forecasts the S&P 500, Dow Jones Industrial Average, Nasdaq Composite Index, and NYSE Composite (covering the period from January 1, 2012, to September 1, 2023) using ARIMA and SVR models. The superior performance of the SVR model in predicting stock returns highlights the potential of machine learning techniques in financial forecasting. This paper reveals that SVR, as the favored option for predicting stock prices in this comparison analysis, can produce more accurate forecasts compared to classic statistical models like ARIMA due to its capacity to capture complicated non-linear correlations in the data. For additional research, a hybrid model and deep learning strategy should be looked upon. Hybrid models that combine the strengths of different predictive techniques could be a promising area of research. For example, an ensemble model that integrates ARIMA, SVR, and other machine learning algorithms could potentially achieve even higher prediction accuracy. The combination of multiple models might also enhance robustness in the face of varying market conditions. Deep learning, particularly recurrent neural networks (RNNs) and LSTM networks, has shown great promise in time series forecasting. Future research can delve deeper into the application of these neural network architectures to stock return prediction. Additionally, attention mechanisms and advanced variations of LSTM networks could be explored to recognize long-term dependencies and complex patterns in financial data.

References

1. Malkiel, B. G., Fama, E. F. The Journal of Finance, (1970).
2. Banu, M. P. International Journal of Scientific Research in Computer Science Engineering and Information Technology, (2020).
3. Yang, Y., Wu, Y., Wang, P., Jiali, X. E3S Web of Conferences, (2021).
4. Nikou, M., Mansourfar, G., Bagherzadeh, J. Intelligent Systems in Accounting, Finance and Management, (2019).
5. Henrique, B. M., Sobreiro, V. A., Kimura, H. The Journal of Finance and Data Science, (2018).
6. Pansari, R. K., Rasool, A., Wadhvani, R., Dubey, A. Next Generation Systems and Networks, (2023).
7. Pardaz Banu, M. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, (2020).
8. Aditi S., Lavnika M. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, (2023).
9. Mehtab, S., Sen, J. arXiv preprint arXiv:1912.07700, (2019).
10. Göçken, M., Özçalıcı, M., Boru, A., Dosdoğru, A. T. Neural Computing and Applications, (2019).