Iosr Journal Of Economics And Finance (Iosr-Jef)
E-Issn: 2321-5933,P-Issn: 2321-5925.Volume 15, Issue 2 Ser. 1 (Mar. – Apr. 2024),Pp 15-21
www.Iosrjournals.Org

Forecasting Stock Index Using A Hybrid ARIMA-SVM Model

Dhawani Nilesh Shah[1], Manishkumar Thaker[2]
[1]*research Scholar, Department Of Statistics, Gujarat University, Ahmedabad, Gujarat, India*
[2]*associate Professor & Hod, Statistics Department, M.G.Science Institute, Ahmedabad, Gujarat, India*

Abstract:
The stock market has high volatility; therefore, it is challenging to forecast future prices. This paper presents an innovative hybrid method for combining sophisticated machine learning models (SVM) with traditional statistical models, such as time series models based on autoregressive integrated moving average (ARIMA). The objective of this paper is to overcome the limitations of the ARIMA model in financial time series forecasting by combining it with the SVM model. The proposed combination method hybridised both ARIMA and SVM models to capitalise on the unique strengths of ARIMA and SVM models in linear and nonlinear modelling respectively. The daily closing price of the Nifty 50 index is used as the dataset for the period from January 1, 2018, to December 15, 2023. When compared with the individual models, the results show that the hybrid model is accurate. The hybrid model's performance was assessed using two metrics: mean square error (MSE) and root mean square error (RMSE).

Key Word*: ARIMA, Hybrid Models, SVM, Time Series*

Date of Submission: 02-03-2024 Date of acceptance: 12-03-2024

I. Introduction

Economic time series such as exchange rates and stock indexes are non-linear, dynamic and non-stationary in nature. It is influenced by various factors like geopolitical events, economic conditions and government policies. As a result, predicting the trend of the financial time series and making predictions for the future period is a crucial endeavour for both academicians as well as market participants [1]. Several studies have studied the price fluctuations in the stock market dynamics using a wide range of linear as well as non-linear time series models. Traditional time series models such as autoregressive integrated moving average (ARIMA), probabilistic models, regression models and vector autoregression models among others have been used in many earlier studies to forecast stock price time series. The approximation of traditional time series methods to model stock price time series is not always satisfactory since these time series exhibit high non-linearity. With the increase in computational ability, now the focus has been shifted to the use of artificial intelligence and machine learning techniques such as Artificial neural network (ANN), support vector machine (SVM) and random forest (RF). The main aim of machine learning methods is to identify complex patterns and make accurate predictions, whereas traditional statistical models are not accurate enough for large datasets [2].

ANN is a very effective tool for time series modelling and forecasting, it has certain limitations in learning the pattern if the data is noisy and has complex dimensions. It also has the issue of overfitting, which results in a loss of generalizability [3]. SVM and RF have also been used successfully to forecast a variety of time series. SVM and RF are two examples of more recent, simpler models that have been used to produce better results than traditional methods. Furthermore, ANN models might tend to fall into a local optimal solution, whereas SVM and RF offer solutions that might be the global optimum. Therefore, overfitting is unlikely to happen when using RF and SVM. For nonlinear time series, machine learning models such as SVM are utilised [4].

In emerging economies like India, the effectiveness of hybrid models—the combination of linear and nonlinear models such as ARIMA-GARCH, ARIMA-ANN, and ARIMA-SVM is crucial. In this paper, we build a hybrid model, ARIMA-SVM, our evaluation compares this hybrid model with independent models like SVM and ARIMA. Empirical results indicate that hybrid ARIMA-SVM outperforms both independent models in terms of MSE and RMSE. Overall, this confirms the superiority of hybrid models over independent models. However, no published work has been found for forecasting the closing price of the Nifty 50 index of the Indian Stock market with the ARIMA-SVM model.

The remaining paper is divided into six sections. Section 2 provides a literature review; Section 3 discusses data and research methodology. Furthermore, the results and findings from the application of the

hybrid model employed are detailed in Section 4. The paper ends with a conclusion and way forward in Section 5.

II. Literature Review

In the last two decades, several studies have used a wide range of linear and non-linear models for forecasting volatile time series. The most popular and traditional time series model is the linear Box-Jenkins or ARIMA model [6]. The ARIMA approach is both simple and yields accurate results and is widely used, however, one of the major limitations of the ARIMA model is the pre-assumed linear form. The real-time datasets are highly volatile and considered non-linear where the mean and variance of the time series change over time. Thus, identifying the non-linear relationship and the application of non-linear models may help in improving the forecasting accuracy.

Several authors have found that non-linear models outperform traditional models like ARIMA. Khashei and Bijari (2010) discussed the effectiveness of the ANN model, (p, d, q), for time series forecasting. Their study integrates neural networks with ARIMA components, improving forecasting accuracy and providing a robust approach for modelling and predicting time series data [7].

While ANN can be a very successful tool for time series, it also has limitations in learning the patterns. Thissen et al. (2003) in his paper discussed the application of SVM in time series forecasting, showcasing their utility in modelling complex data patterns. The study addresses the efficacy of support vector machines for forecasting, emphasizing their potential in handling intricate time-dependent datasets. The research contributes insights into leveraging machine learning techniques for improved time-series predictions [8]. SVM has better generalisation capability than conventional techniques and there is no increase in the number of parameters [9].

Multiple attempts have been made to build a simple and accurate model. Recently, the RF method has been used in forecasting developed by Breiman (2001) and exhibits outstanding performance [10]. Kane et al. (2014) conducted a comparative analysis between ARIMA and Random Forest models for predicting avian influenza H5N1 outbreaks. Their study provides insights into the performance of these time series models, contributing valuable information for enhancing the accuracy of infectious disease outbreak predictions [11]. Zhang (2003) noted the prevalent use of autoregressive terms as inputs in neural networks, despite the persistent non-linear patterns in ARIMA model residuals. To address this, he developed a hybrid ARIMA-ANN model using ARIMA residuals as inputs, comparing its performance with independent ARIMA and ANN models. Results from real datasets indicated the hybrid model significantly enhanced forecasting accuracy, demonstrating its effectiveness in capturing non-linear patterns in time series data.

In the context of India, many previous studies have adopted AR, ARIMA, ANN, SVM, etc. for forecasting purposes, but there is a lack of hybrid models for forecasting the daily closing price of NIFTY50. Also, no published work was found for the same. As mentioned in Section 1, our study focuses on building a hybrid ARIMA-SVM model.

III. Data And Methodology

We use the closing price of Nifty 50-one of the index of the Indian stock market for the period January 1, 2018, to December 15, 2023. The data were obtained from the website of the National Stock Exchange of India (https://www.nseindia.com/).

Forecasting Methodology

In this section, we explain the non-linear model (SVM), the linear model (ARIMA) and the hybrid models.

ARIMA Methodology

The Box-Jenkins methodology (1976) is an approach for developing and designing ARIMA models for forecasting [12]. These models capture the sequential relationship within time series, where autoregressive (AR) terms depict the connections among dependent variables and moving average (MA) terms depict the reliance of dependent variables on preceding error terms [5]. Primarily, it is employed for stationary data, where the mean, variance and autocorrelation remain constant over time. In the case of non-stationary data, it can be employed post-transformation through methods like differencing and logarithm. The ARIMA model comprises autoregressive (AR) components denoted by p, differencing is denoted by d, and moving average (MA) is denoted by q. Historical data decomposes through an AR process that preserves past occurrences, an integrated (I) process that stationarise data and an MA process that captures errors. The p in the AR component represents a linear relationship that dependent variables share with their lagged values. The resultant model with the combination of AR (p) and MA (q) is expressed as:

$$y_t = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \varepsilon_t y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

Where ε_t is white noise $\sim N(\mu, \sigma^2)$, $\{\emptyset = 1, \dots, p\}$ and $\{\theta = 1, \dots, q\}$ are the coefficients of AR(p) and MA(q) components, respectively.

Support Vector Machines

The support vector machine (SVM) was proposed by vapnik [13]. Unlike other neural networks, SVM aims to minimize an upper bound of the generalisation error rather than the empirical error, based on the structured risk minimization principle. Also, a collection of high dimensional linear functions is applied by SVMs model to create a regression function. The SVM regression function is expressed as follows:

$$y = w\emptyset(x) + b; \tag{2}$$

Where $\emptyset(x)$ is the high dimensional feature space which is non-linear mapped from the input space x. The coefficients w and b are estimated by minimising:

$$R_{SVM_s} = C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i y_i) + \frac{1}{2}\|w\|^2 \tag{3}$$

$$L_\varepsilon(d_i y_i) = \begin{cases} |d - y| - \varepsilon, |d - y| \geq \varepsilon \\ otherwise \end{cases} \tag{4}$$

Where,
C and ε are prescribed parameters, E is the tube size of SVM, d_i is the actual closing price in i[th] period
$C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i y_i)$ is the empirical risk measured by ε intensive loss function
$\frac{1}{2}\|w\|^2$ is regularised term
C is the regularisation constant that evaluates the trade-off between the empirical risk and the flatness of the model
ς and ς^* are the slack variables representing the distance from actual values to the corresponding boundary values of ε tube.

The above equation 3 can be transformed to the following constrained formation:

Minimise:
$$R(w, \varsigma, \varsigma^*) = \frac{1}{2}ww^T + C^*(\sum_{i=1}^{N}(\varsigma_i + \varsigma_i^*)) \tag{5}$$
Subject to:
$$w\phi(x_i) + b_i - d_i \leq \varepsilon + \varsigma_i^* \tag{6}$$
$$d_i - w\phi(x_i) - b_i \leq \varepsilon + \varsigma_i \tag{7}$$
$$\varsigma_i, \varsigma_i^* \geq 0, \qquad i = 1,2, \dots, N$$

Equation 5 can be solved by Langragian multiplier and maximising dual function of equation 5.

A kernel function $K(x_i, x_j)$is a function that measures the similarity between two vectors x_i and x_j in a high-dimensional space $\phi(x_i)$ and $\phi(x_j)$.It does this by computing the dot product of $\phi(x_i)$ and $\phi(x_j)$, i.e., $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$. Different kernel functions can produce different inner products, which can be used to build machines with various kinds of non-linear decision boundaries in the input space. The best model is the one that minimises the error estimate among the different kernel functions.

Hybrid Methodology

The volatility in the stock market prices makes it difficult to predict accurately, to address this issue, a hybrid model is build combining both linear and non-linear modelling techniques. The ARIMA and SVM models have different capabilities to capture characteristics in linear or non-linear domains. Therefore, a hybrid model composed of ARIMA and SVM components can model both linear and nonlinear patterns with improved forecasting performance.
The hybrid model (Y_t) can be represented as follows:

$$Y_t = L_t + N_t \tag{8}$$

Where L_t is the linear part and N_t is the nonlinear part of the hybrid model. Both L_t and N_t are estimated from the data set.

L_t is the forecast value of the ARIMA model at time t. Let e_t be the residual at time t obtained from ARIMA model then,

$$e_t = Y_t - \hat{L}_t \tag{9}$$

The residuals are modelled by SVM and can be represented as follows:

$$e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-n}) + \Delta t \tag{10}$$

Where f is a nonlinear function modelled by SVM and Δt is the random error. Therefore, the combined forecast is,

$$\hat{Y}_t = \hat{L}_t + \hat{N}_t \tag{11}$$

Notably, $\hat{N}_t$ is the forecast value of equation 10.

IV. Results And Findings

The ARIMA, SVM, and ARIMA-SVM models are estimated for the given dataset. The model estimation selection process is then followed by an empirical evaluation and the relative performance of the models is measured by statistical measures.

The daily closing price of Nifty 50 is used in this study to examine the performance of the proposed model. The time series data has undergone a log transformation to address the skewness and enhance the normality in the variable distribution. Figure 1 shows the closing price of Nifty 50 from January 01, 2018, to December 15, 2023. This plot clearly shows there is an upward trend. Data from January 01, 2018 to October 06, 2023 is used as a training data set and closing prices from October 07,2023 to December 15,2023 is used as testing data.

Table 1: Descriptive Statistics for Train and Test Series

	Training data set	**Testing data set**
Mean	9.45	9.83
Std deviation	0.2	0.05
Minimum	8.93	9.73
Maximum	9.82	9.96

Table 1 summarises the train and test dataset. Two indices, MSE(mean square error) and RMSE (root mean square error) are used as measures of accuracy. The indices are shown as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{12}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{13}$$

Where *n* is the number of forecasting periods, y_i is the actual closing price at period *i* and $\hat{y}_i$ is the forecasted price at period *t*.

In this study, auto.arima function of the pmdarima library in Python is used to build the model, optimal order and optimal seasonal order is selected based on the determined criteria (AIC values). The table given below shows the top 5 models with the least AIC values.

Table 2: AIC Values for Different ARIMA Models

Model	**AIC**
ARIMA (5,1,4)	-6981.38
ARIMA (4,1,5)	-6981.62
ARIMA (5,1,5)	-6970.3
ARIMA (4,1,4)	-6963.27
ARIMA (5,1,4)	-6982.42

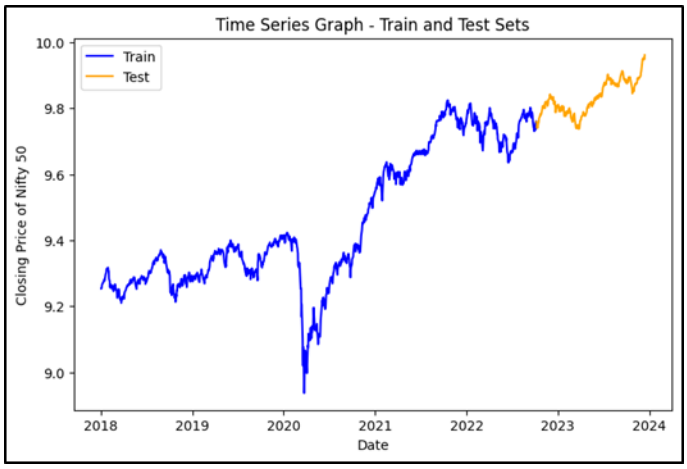

Fig 1: Time Series Plot of Closing Price of Nifty 50

The most appropriate ARIMA model is ARIMA (5,1,4), the figure 2 shows the comparison of test data and forecasted values of the daily closing price using the 'predict' function. The ARIMA model exhibits subpar forecasting performance and has demonstrated low accuracy; it has yielded a linear function.

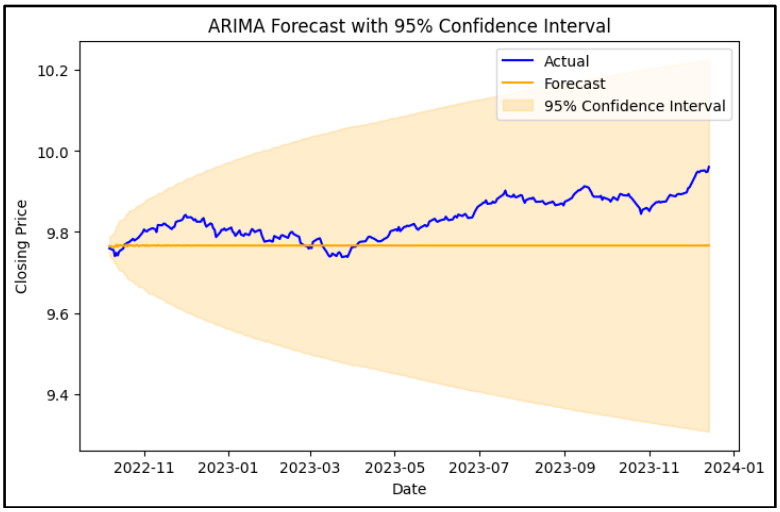

Fig 2: Actual and Forecast Results by ARIMA Model

For the SVM model, we propose a Support vector regression model, here we use the Gaussian radial basis function as the kernel function. To determine the parameters of the SVM, we employed grid search method that systematically explored the range of hyperparameters. The grid search focused on tuning hyperparameters — radial basis function (σ), epsilon-insensitive loss function(ε) and cost parameter (C). Grid search method enables to select the best model, without causing overfitting or underfitting of the training data. By applying this optimised SVR model on training data, we assess the performance using metrics such as mean square error (MSE) and root mean square error (RMSE).

The suitable parameter for the model are C=1, epsilon = 0.01 and gamma =0.001. The figure 3 shows the comparison of actual values and forecasted values of the test data for daily closing price using SVR method.

The Hybrid model was built by taking residuals of the ARIMA model as the input for the SVM model. This study used the value of C and σ experimented between the range of 1 and 100, and further Radial basis kernel was used. The grid search method provided with the optimal model with C = 10, epsilon= 0.01 and gamma = 0.01. The resulting Hybrid model is labelled as Hybrid ARIMA-SVM, the figure 4 provides a visual representation of the comparison between actual and forecasted closing price of Nifty 50 using Hybrid model.

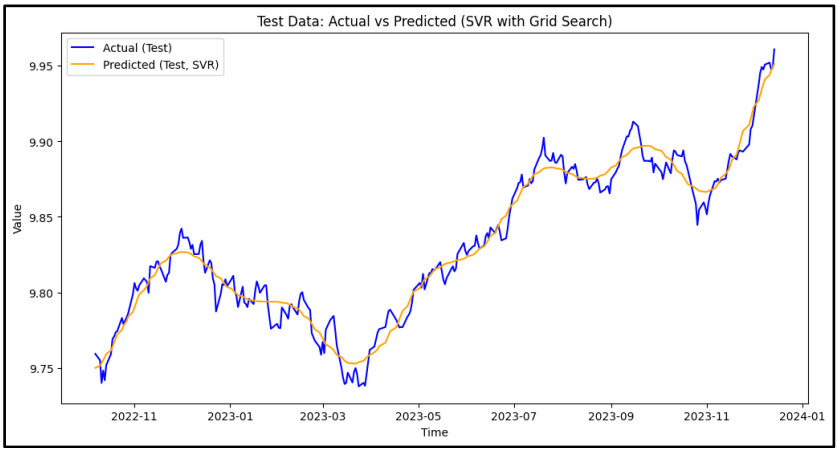

Fig 3: Actual and Forecast Results by SVM Method

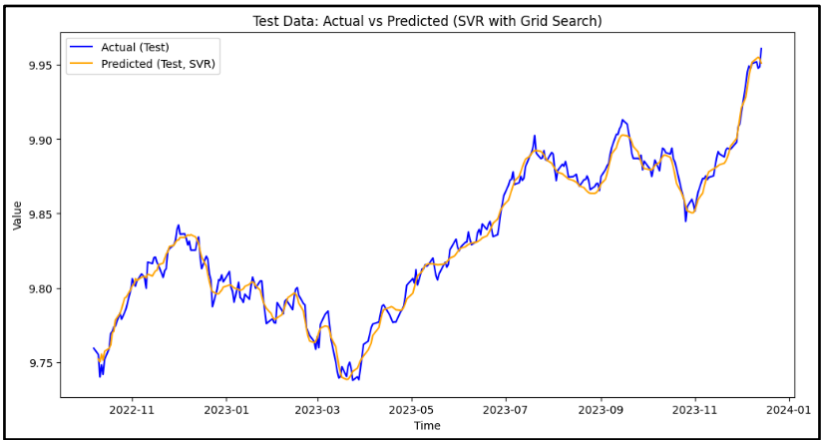

Fig 4: Actual and Forecast Results by Additive Hybrid Model

The table given below displays the forecasting performance results, including MSE and RMSE, obtained from ARIMA, SVM and hybrid model. The ordering of the models is based on their performance on the dataset. Notably, the hybrid models exhibit lower errors compared to the other models, as evident from the table. This observation implies that neither the ARIMA nor the SVM model alone can fully capture all the patterns present in the data.

Table 4: Performance Comparison for the Daily Closing Price of Nifty 50 Stock Index

Model	MSE	RMSE
Hybrid ARIMA-SVM	7.8	2.7
SVM	15.3	3.8
ARIMA	95.3	9.7

V. Discussion

For more than fifty years, the autoregressive integrated moving average (ARIMA) model has been prevalent in various fields of time series forecasting. Recently, artificial neural networks (ANN) have shown the capability in capturing nonlinear data patterns. This study is motivated by the observation that various forecasting models may enhance each other in accurately representing data sets. Consequently, a hybrid model combining ARIMA and support vector machines (SVM) is proposed. The presented model is believed to greatly improve the prediction performance of the single ARIMA model or the single SVMs model in forecasting closing price of Nifty 50 stock index. However, future research should address some problems of structured selection of optimal parameters of the hybrid model and also building Multiplicative Hybrid Model.

References

[1]. Hanssens, D. M., Parsons, L. J., & Schultz, R. L. (2003). Market Response Models: Econometric And Time Series Analysis (Vol. 2). Springer Science & Business Media.

[2]. Cerqueira, V., Torgo, L., & Soares, C. (2019). Machine Learning Vs Statistical Methods For Time Series Forecasting: Size Matters. Arxiv Preprint Arxiv:1909.13316.

[3]. Zhang, G., Patuwo, B.E. And Hu., M.Y. (1998) 'Forecasting With Artificial Neural Networks: The State Of The Art', International Journal Of Forecasting, Vol. 14, No. 1, Pp.35–62.
[4]. Falát, L., Pančíková, L., & Hlinková, M. (2015, July). Prediction Model For High-Volatile Time Series Based On Svm Regression Approach. In 2015 International Conference On Information And Digital Technologies (Pp. 77-83). Ieee.
[5]. Time Series Forecasting Of Daily Gold Prices In Ahmedabad: An Evaluation Of Arima, Ann, And Hybrid Arima-Ann Approaches, International Journal Of Emerging Technologies And Innovative Research (Www.Jetir.Org | Ugc And Issn Approved), Issn:2349-5162, Vol.10, Issue 12, Page No. Ppb342-B348, December-2023.
[6]. Singh, A., & Mishra, G. C. (2015). Application Of Box-Jenkins Method And Artificial Neural Network Procedure For Time Series Forecasting Of Prices. Statistics In Transition New Series, 16(1), 83-96.
[7]. Khashei, M., & Bijari, M. (2010). An Artificial Neural Network (P, D, Q) Model For Timeseries Forecasting. Expert Systems With Applications, 37(1), 479-489.
[8]. Thissen, U. V. B. R., Van Brakel, R., De Weijer, A. P., Melssen, W. J., & Buydens, L. M. C. (2003). Using Support Vector Machines For Time Series Prediction. Chemometrics And Intelligent Laboratory Systems, 69(1-2), 35-49.
[9]. Sivapragasam, C., & Liong, S. Y. (2005). Flow Categorization Model For Improving Forecasting. Hydrology Research, 36(1), 37-48.
[10]. Breiman, L. (2001). Statistical Modeling: The Two Cultures (With Comments And A Rejoinder By The Author). Statistical Science, 16(3), 199-231.
[11]. Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014). Comparison Of Arima And Random Forest Time Series Models For Prediction Of Avian Influenza H5n1 Outbreaks. Bmc Bioinformatics, 15(1), 1-9.
[12]. Box, G.E.P. And Jenkins, G.M. (1976) Time Series Analysis: Forecasting And Control. Revised Edition, Holden Day, San Francisco
[13]. Vapnik V. The Nature Of Statistic Learning Theory. New York: Springer; 1995.