***Review Article***

Time Series Analysis Using Diverse Algorithms and ARIMA Model

Smit Borasaniya#1a, Ishan Pansuriya#1b, Ketul Patel#1b, Dr. Gaurav Kumar Ameta#2

*Department of Computer Science - Bachelors in Technology Program, Indus University, Rancharda, Ahmedabad, Gujarat, INDIA.*

*Professor, Department of Computer Science, Indus University, Rancharda, Ahmedabad, Gujarat, INDIA.*

1 [smit.borasaniya@gmail.com](mailto:smit.borasaniya@gmail.com), 2 [ishanpansuriya@gmail.com](mailto:ishanpansuriya@gmail.com), 3 [ketul.patel3000@gmail.com](mailto:ketul.patel3000@gmail.com)

***Abstract: The different methods of learning implemented by different algorithms gives a unique perspective for various insights that can be gathered form a dataset. The commonly used forecasting methods are ANN, RNN, LSTM, SVM, and ARIMA analysis. ARIMA Model method is a way to integrate technical analysis for making cognizant financial decision. It can be used to predict short as well as long term stock prices or any other values that are tightly bounded with time, such as climate change or any statistical analysis. The stock market in particular has its biggest challenge of predicting the stock prices. As stock price data represents a financial time series data which becomes more difficult to predict due to its characteristics and dynamic nature. It can also be merged with other models and algorithms for more veracity to the insights produced. The following review paper consists of an array of algorithms generally used in time series analysis and an elaboration on how different parameters can help gain better intuition using ARIMA model for the intrinsic properties of datasets such as RMSE, AIC, SBC, variance of error and maximum likelihood values. Various types of datasets can have different sets of parameters being considered at the same time to improve its efficiency. This opens up numerous possibilities of implementation of the algorithm.***

***Keywords: (AIC) Akaike information criterion, Artificial Neural Networks (ANN), Auto Regressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), Maximum likelihood, Recurrent Neural Networks (RNN), Root Mean Squared Error (RMSE), (SBC) Schwarz’s Bayesian Criterion and (SVM) Support Vector Machine and Variance of error.***

1. **INTRODUCTION**

Forecasting is the system of gathering predictions for the future based totally on historical and present information and the study of trends. The forecasting procedure provides with a fast and austere way to generate the forecasts for many time series in a single step. Forecasting uses an extrapolative method(s), where the forecasts for a series are only the function of time and past values of the series, not of any other additional variables. Due to the sequential nature, stochastic approach is widely used to draw inferences about the characteristics of the time series data.[10] Time-series analysis is a statistical method of analyzing data from repeated observations on a single unit or individual at regular intervals over a large number of observations. [9] Time series forecasting is a process to predict the future with the help of history data. [10] Stock market is a booming sector of today’s economy; people are investing in stocks for a good return on investment. Prediction is a comparable, however extra accepted time period. [[1](https://scholar.google.com/citations?view_op=view_citation&hl=it&user=vb9EOUMAAAAJ&citation_for_view=vb9EOUMAAAAJ:HeT0ZceujKMC)] With the need for more veracity in the trends of values for the stock prices, the trend forecasting becomes more necessary and essentials for stakeholders. [2]

The beauty of the ARIMA model lies in its simplicity and can be applied only to stationary time series [p, d, q]. Therefore, our data from the time series, which are seasonal and nonstationary, are transformed into a stationary series for the application of the ARIMA model. [11] The model can be developed using sophisticated statistical techniques. [12] Another concept of autoregressive model known as Periodic Autoregressive (PAR) model and Laplace Autoregressive (LAR) has also been frequently used recently in various environmental, hydrological and meteorological studies. [13] Another statistical model is based on the seasonal analysis of the ARIMA time series (SARIMA) and is further improved by incorporating short-term solar radiation predictions derived from the NWP (numerical weather prediction) model. [14] Deep Learning (DL) techniques have recently developed as powerful machine learning tools that can do shape identification, regression analysis, and prediction.

### ORGANIZATION OF PAPER

A review on the methods of time series analysis, used in different types of industries based on different factors and criteria’s and how it can be used to predict future values on the basis of small or large amount of historical data is conducted. Basically, a comparison is made between the usage of ARIMA model in banking stock market data to predict 200 observations and in demand analysis in food industry to provide the backing support for the robustness of the applications of ARIMA model in the field of predicting future time series values.

The rest of the paper is organized as follows: Section III has a discussion of the methodologies used for time series analysis including six types of algorithms (ANN, RNN, LSTM, SVM, ARIMA) which help to make the quality prediction of the historical data. Section IV includes the investigation of the factors weighed in for long-term prediction of values and short-term predictions of values through ARIMA models, considering two particular researches for banking stock market data (A) and forecasting demand of food (B). Session V concludes the paper and discusses future directions.

### METHODS USED FOR TIME SERIES ANALYSIS

Time-series forecasting is extensively used for non-stationary data. Non-stationary data are called the data whose statistical properties e.g., the mean and standard deviation are not constant over time but instead, these metrics vary over time. These non-stationary input data (used as input to these models) are usually called time-series. Some examples of time-series include the temperature values over time, stock price over time, price of a house over time etc. So, the input is a signal (time-series) that is defined by observations taken sequentially in time.

### *ANN*

Artificial Neural Networks (ANN), commonly simply called Neural Networks (NN), are computer systems that are inspired by the biological neural networks that make up the brains of animals. An ANN is based on a collection of connected units or nodes called artificial neurons that loosely model the neurons in a biological brain. Each connection, like synapses in a biological brain, can carry a signal to other neurons. receives a signal, processes it and can signal neurons connected to it. The "signal" in a connection is a real number, and the output of each neuron is calculated by a nonlinear function of the sum of its inputs. The connections are called edges. Neurons and edges generally have a weight that adapts as learning progresses. The weighting increases or decreases the signal strength of a connection. Neurons can have a threshold so that a signal is only sent when the added signal exceeds that threshold. are added in layers. Different layers can perform different transformations on their inputs. Signals travel from the first layer (the input layer) to the last layer (the output layer), possibly after traversing the layers several times.

Traditional statistical models that include exponential smoothing, moving average, and ARIMA make its prediction linear. Support Vector Machines (SVM) and Artificial Neural Networks (ANN) are widely used today to estimate stock price movements. Artificial neural networks are widely used to solve many problems because of their versatility. ANN can be viewed as a computational or mathematical model that is inspired by the functional or structural properties of biological neural networks. These neural networks are designed to extract patterns from noisy data. A large sample of data known as the training phase then feeds the data into the network that was not included in the training phase, this phase is known as the validation or prediction phase. The only reason for this procedure is to predict new results. Algorithms provide an accuracy of 99.9% using Levenberg-Marquardt, scaled conjugate gradient and Bayesian regularization.

Achieving moderately correct forecasts of a statistic may be a vital however difficult task. ARIMA and ANN are two wide standard and effective prediction models. ARIMA assumes linear information generation’s function, whereas ANN is best suited for nonlinearly generated time series. But, it's virtually not possible to determine the precise nature of a series and a real-world time series most frequently contains each linear still as skew correlation structures. [3]

Input Layers

Hidden Layers

Output Layers

**Figure 1:** Simple architecture of a neural network

The Figure 1 shows the basic architecture of a neural network with input layers where input is given, hidden layers where the basic computing is performed and the output.

Yj is the output of node j, f (.) is the transfer function, wij the connection weight between node j and node i in the lower layer and Xij is the input signal from the node i in the lower layer to node j.

### *RNN*

RNN is a class of ANN where connections are established based on directed graphs along a temporal sequence. The first layer is normally a feed ahead neural community observed with the aid of using recurrent neural community layer in which a few statistics it had within side the preceding time-step is remembered with the aid of using a reminiscence function. Forward propagation is carried out on this case. It saves statistics required for its future use. If the prediction is wrong, the getting to know price is hired to make small changes. Hence, making it progressively boom closer to making the proper prediction in the course of the backpropagation. Recurrent Neural Networks may provide better predictions than the neural networks used in any study, such as, LSTM (Long Short-Term Memory) [4]

Both finite impulse and infinite impulse recurrent networks can have additional stored states, and the storage can be under direct control by the neural network. The storage can also be replaced by another network or graph if that incorporates time delays or has feedback loops. Such controlled states are referred to as gated state or gated memory, and are part of [long short-term memory](https://en.wikipedia.org/wiki/Long_short-term_memory) networks (LSTMs) and [gated recurrent units](https://en.wikipedia.org/wiki/Gated_recurrent_unit). It is also known as Feedback Neural Network (FNN).

### *LSTM*

Neural networks used in Deep Learning consists of different layers connected to each other and work on the structure and functions of the human brain. It learns from huge volumes of data and used complex algorithms to train a neural net. The recurrent neural network works on the principle of saving the output of a layer and feeding this back to the input in order to predict the output of the layer.

Units are enforced to learn very long sequences. This is a more general version of the gated recurrent system. LSTM is more benign than other deep learning methods like RNN or traditional feed forward because LSTMs tackle the evanescent gradient issue. Unlike contemporary model for prediction which uses feed forward neural systems, LSTM uses input associations i.e. Not only does the procedure focus on closing day value for stock market data but also all the data points arrangements throughout the day. Which requires a model which incorporates cross-approval which is achieved by training of the model using the pre-partitioned information. The motivation of tuning the trends of stock, is to explicitly amend the calculation so that it can educate to feature data and calibrate itself.

The LSTM module is composed of a cell, a data door, a front door and a door with a view. The cell collects values over arbitrary time intervals, and the three inputs manipulate the development of records inside and out of the cell. Thus, the predominant benefit of the LSTM is each LSTM unit collects statistics for both, an extended or quick period of time (ergo the name) without explicitly using the activation function inside the recurrent components. This lets in LSTMs to take care of the evaporating slope issue – as the value positioned away withinside the reminiscence cell isn't always iteratively adjusted; the inclination does not disappear while it is modelled by the LSTM model. The paper suggests that the algorithm is able to prove the with minimum loss rate of 0.0024 and if the epoch batch rates are increased then training will be more efficient. [5]

In the Figure 2, the general system architecture of LSTM is demonstrated.

Raw Data

Feature Extraction

Training Data

Prediction

Testing Data

Trained Data Result

**Figure 2:** System Architecture of LSTM [5]

The Figure 2 shows the system architecture of LSTM where the main components/features of the data are selected form the raw data and split into training and testing sets. The model is then trained and analyzed to gain insights on the predicted values of stock prices.

### *SVM*

Support Vector Machines are efficient supervised learning algorithms applicable for both classification and regression. It is a discriminative classifier that is formally defined by a separating hyperplane.

In classification problems there are a set number of outputs that a feature set can be labeled as, whereas the output can take on continuous values in regression problems. [6] In Saahil’s paper the problem of stock price forecasting as a classification problem. The feature set of a stock’s recent price volatility and momentum, along with the index’s recent volatility and momentum, are used to predict whether or not the stock’s price m days in the future will be higher (+1) or lower (−1) than the current day’s price. Specifically, it is a binary classification problem. [6]

There are no assumptions made in the dataset and all the numeric problems can be dealt with SVM. The linear separability of the data plays a significant role in deciding the degree of tolerance in SVM. The penalty term that is passed as hyperparameter in SVM when it comes to linearly separable and nonlinear solutions is called 'C', which is called the degree of tolerance. The decision limit depends on a small margin and fewer support vectors. Therefore, the black box method, has a tendency towards overfitting and the very strict calculation. It is a useful method that can be carried out even if its high stability is not impaired. [8]

### *ARIMA*

An ARIMA model could be a category of statistical models for associate analyzing and prognostication statistic data. It expressly caters to a collection of ordinary structures in time series data, and as such provides a straightforward however, powerful methodology for creating skillful time series forecasts.

ARIMA is a form that stands for Autoregressive Integrated Moving Average. It's a generalization of the less complicated Autoregressive Moving Average and adds the notion of integration. This acronym is descriptive, capturing the key aspects of the model itself.

Briefly, they are:

• AR: Autoregression. A model that uses the dependent relationship between Associate in Nursing observation and a few varieties of lagged observations. The auto regression is given by:

• I: Integrated. the utilization of differencing of raw observations (e.g., subtracting an observation from an observation at the previous time step) so as to create the statistic stationary.

• MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations. The moving average is given by:

Each of those additives are explicitly designated withinside the version as a parameter. A widespread notation is used of ARIMA (p, d, q) in which the parameters are substituted with integer values to fast suggest the unique ARIMA version being used. The parameters of the ARIMA model are defined as follows:

• p: Lag order or the number of lagging observations included in the model.

• d: Degree of differencing or the number of times that the raw observations are differenced.

• q: Order of moving average or the size of the moving average window.

A linear regression model is created with the necessary number and kind of terms, and the statistics is prepared using a degree of differencing to make it stationary, that is, to remove fashion and seasonal systems that have a negative impact on the regression model. A charge of zero can be used for a parameter, indicating that the parameter is no longer in use. In this way, the ARIMA version can be set up to behave like an ARMA version, or even a simple AR, I, or MA version. When using an ARIMA version for a temporal collection, the underlying system that generated the observations must also be an ARIMA system. This could also be beneficial appear obvious however, it allows to encourage the desire to confirm the model's assumptions within the raw observations and within the residual forecasting errors from the model. Primarily ARIMA model is denoted as:

The Seasonal Autoregressive Integrated Moving Average (SARIMA) version of ARIMA is used for seasonal time collection forecasting. ARIMA and its distinctive versions are primarily based totally at the famous Box-Jenkins precept and so those also are extensively referred to as the Box-Jenkins models. [1]

### APPLICATION OF ARIMA

### *ARIMA MODEL IN PREDICTING BANKING STOCK MARKET DATA: RMSE*

### For predicting, the ARIMA model was used to banking stock market data in this article. The results are obtained using the MINTAB software. In the period 1993 to 2017, 2000 observations were gathered for each variable from associated databases. Because several ARIMA models can be created for one column of data using different values of p, d, and q, RMSE is chosen as a criterion for finding the fitting ARIMA model. As a result, the fitted ARIMA model has a lower RMSE. [7]

They have concluded on these observations:

1. The values of p, d, and q are between 0 and 2 solely because these values cannot be negative, and they should not be greater than 2 otherwise the parameter estimation will be useless.
2. The RMSE is set between 4.00 and 5.00 depending on the dataset. As a result, after utilizing the program to construct the dataset, ARIMA (1,1,2) was found to be the best with an RMSE of 1.4.
3. In some circumstances, the ARIMA model is not fitted, indicating that the dataset cannot be estimated, and this should be discarded.

Banking time series data and financial time series in general is

very hard in decomposition and forecasting because the data are non-stationary and non-linear with high heteroscedasticity.[7]

### *FORECASTING OF DEMAND USING ARIMA MODEL: 4 CRITERIA’S (AIC, SBC, STANDARD ERROR, AND MAXIMUM LIKELIHOOD)*

### Based on the Box–Jenkins approach, the study was carried out in three parts: identification, estimation, and verification. [8] The paper used four parameters namely AIC, SBC, variance of error and maximum likelihood. The data taken is for food company and prediction for its future demand is made. They used SPSS time series module and a fast maximum likelihood estimation algorithm.

### AIC stands for [Akaike information criterion](https://en.wikipedia.org/wiki/Akaike_information_criterion" \o "Akaike information criterion) and SBC for Schwarz’s Bayesian Criterion. AIC is calculated by

### where L is the likelihood of the data and k=d. SBC or BIC is calculated by

### where AIC is [Akaike information criterion](https://en.wikipedia.org/wiki/Akaike_information_criterion), as calculated before and k=d. In both the cases, p is the order of the AR part and q is the order of the MA part.

### The aforementioned algorithms were used to find values of the four parameters for different models like (1,1,1), (1,0,1) etc. On which they have based their predictions. The ARIMA model (1, 0, 1) is selected because all the coefficients are significantly optimized than any other models taken into consideration. The IBM SPSS Forecasting was used to then obtain the prediction based on the best parameters selected which had the best optimized values for the 4 parameters from the list of all permuted values P, D, Q i.e. (1, 0, 1) was selected.

### CONCLUSION AND FUTURE DIRECTIONS

1. ***CONCLUSION***

In “ARIMA Model in Predicting Banking Stock Market Data” they have made short term forecast on banking stock market data and collected 200 observations, the best model was selected with the criteria of MSE for short term prediction. Other forecast horizons for stock market data, such as industrial data, can be investigated in the future. Creating new models that combine qualitative and quantitative methodologies to generate accurate forecasts and improve forecast accuracy in the future. The method for short term prediction is suitable for cases of the high-technology market especially for the banks since it gives a significant indicator for the future but is inefficient for not so tightly bounded time series data. There are many factors for different types of datasets to be considered which if taken wrongly may result a varied unrelated misleading output/prediction. This strategy can be used and is appropriate for high-tech market scenarios, particularly for banks, because it provides a substantial indicator for the future. The approach was designed for short-term forecasting and is not suitable for long-term forecasting.

The forecasting made in “Forecasting of demand using ARIMA model” paper is dependent on four criteria’s namely SBC, AIC, standard error, and maximum likelihood. This helps in predicting values for a longer time period (January 2016 to October 2016 i.e., 10 months). The criteria’s make the algorithm more feasible for forecasting future demands and reliable guidelines. Thus, enabling businesses to make profound production and purchasing planning, eliminating the high risk of losses.

1. ***FUTURE DIRECTION***

Only linear predictions are cultivated in ARIMA modeling alone. It requires combination of other forecasting methods like ANN or RNN to support more robust predictability. ARIMA model can be used with a wide variety of algorithms mentioned before to predict values for climate change or seasonal data for prices of stock commodities.

Accounting for an ARIMA model which is feasible for both long term and short-term predictions, a pipeline should be set up where primary predictions are filtered out by RMSE values to gain short term forecast and then find the optimal value for AIC, SBC, variance of error and maximum likelihood to make a long-term prediction. This will return a set of predictions that is acceptable throughout a wide range of long and short values of time.

Testing it with a neural network technique and compare it to ARIMA's results to see if the ANN's power in the food industry can be confirmed. In addition, creating an ARIMA-radial basis function (RBF) combination can help achieve high accuracy.

### REFERENCES

[1] A survey on ARIMA forecasting using time series model ZA Farhath, B Arputhamary, L Arockiam – 2016 Cited by 36 Related articles https://www.ijcsmc.com/docs/papers/August2016/V5I8201626.pdf

[2] Meesad, P., & Rasel, R. I. (2013). Predicting stock market price using support vector regression. 2013 International Conference on Informatics, Electronics and Vision (ICIEV). doi:10.1109/iciev.2013.6572570 https://www.researchgate.net/publication/326326237\_Predicting\_close\_stock\_prices

[3] Ina Khandelwal, Ratnadip Adhikari, Ghanshyam Verma, Time Series Forecasting Using Hybrid ARIMA and ANN Models Based on DWT Decomposition, Procedia Computer Science, Volume 48, 2015, Pages 173-179, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2015.04.167>.

[4] Selvamuthu, D., Kumar, V. & Mishra, A. Indian stock market prediction using artificial neural networks on tick data. Financ Innov 5, 16 (2019). <https://doi.org/10.1186/s40854-019-0131-7>

[5] Pramod, & Pm, Mallikarjuna. (2021). Stock Price Prediction Using LSTM. Test Engineering and Management. 83. 5246-5251.

[6] Madge, Saahil and Swati Bhatt. “Predicting Stock Price Direction using Support Vector Machines.” (2015). (Predicting Stock Price Direction using Support Vector Machines Saahil Madge [saahil\_madge.pdf (princeton.edu)](https://www.cs.princeton.edu/sites/default/files/uploads/saahil_madge.pdf))

[7] ARIMA Model in Predicting Banking Stock Market Data. Mohammad Almasarweh1 & S. AL Wadi2. Modern Applied Science; Vol. 12, No. 11; 2018 ISSN 1913-1844 E-ISSN 1913-1852 Published by Canadian Center of Science and Education

[8] Fattah, Jamal & Ezzine, Latifa & Aman, Zineb & Moussami, Haj & Lachhab, Abdeslam. (2018). Forecasting of demand using ARIMA model. International Journal of Engineering Business Management. 10. 184797901880867. 10.1177/1847979018808673.

[9] Velicer, Wayne & Fava, Joseph. (2003). Time Series Analysis. 10.1002/0471264385.wei0223.

[10] Athiyarath, S., Paul, M. & Krishnaswamy, S. A Comparative Study and Analysis of Time Series Forecasting Techniques. SN COMPUT. SCI. 1, 175 (2020). <https://doi.org/10.1007/s42979-020-00180-5>

[11] Fara, L., Diaconu, A., Craciunescu, D., & Fara, S. (2021). Forecasting of Energy Production for Photovoltaic Systems Based on ARIMA and ANN Advanced Models. International Journal of Photoenergy.

[12] H. X. Zhao and F. Magoulès, “A review on the prediction of building energy consumption,” Renewable and Sustainable Energy Reviews, vol. 16, no. 6, pp. 3586–3592, 2012

[13] Sales, J. D., dealer J. W., Yevjevich V. And Lane, W. L., “Applied Modeling of Hydrologic Time series”, Water Resources Publication, LLC, 1997

[14] S. Vagropoulos, G. Christos Chouliaras, E. G. Kardakos, and A. G. Bakirtzis, “Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting,” in 2016 IEEE International Energy Conference (ENERGYCON), pp. 1–6, Leuven, Belgium, 2016.