iJRASET
International Journal For Research in
Applied Science and Engineering Technology

Volume: 12 Issue: V Month of publication: May 2024

DOI: https://doi.org/10.22214/ijraset.2024.61757

Unravelling Stock Market Patterns: Analysis and Predictive Modelling Using Time Series and Deep Learning

Snigdha Iyengar
Christ Deemed to be University, India

Abstract: Within the dynamic landscape of the financial markets, where fortunes can shift with the wind, the precision of future prediction holds immense value. This research embarks on a quest to enhance the accuracy of stock market forecasts, venturing into the domain of advanced time series models. We shall closely examine three prominent methodologies: the venerable ARIMA, the multifaceted VARMA, and the potent deep learning architecture known as LSTM. Through a rigorous investigation of their strengths and limitations, we shall subject these models to both univariate and multivariate data extracted from the Indian stock market. This comparative analysis aims to glean invaluable insights into their predictive capabilities, ultimately paving the path for the development of even more sophisticated forecasting tools. By wielding these instruments with confidence, investors can navigate the intricate dynamics of the market with a newfound certainty and optimize their decision-making for greater returns. Our endeavor transcends mere numerical analysis; it seeks to illuminate the complex dance of the market, unravel its hidden patterns, and ultimately empower investors to gain the upper hand in this challenging yet potentially rewarding financial sphere.

I. INTRODUCTION

Time series forecasting is an important problem that has been widely studied, and several linear and nonlinear models have been proposed to improve the prediction accuracy. The autoregressive integrated moving average (ARIMA) model is one of the most popular and important models.[3] Considering not only univariate time series data but multivariate data as well in stock market volatility, A basic vector model is the VARMA model.

Although VARMA models are flexible in their representation of several types of time series models, such as vector autoregressive (VAR) and vector moving average (VMA) models, their major limitation is that linear correlation is assumed in the structure and nonlinear characteristics cannot be captured. However, real-world problems are always complex, and hence, building a VARMA model for a multivariate time series requires some attention[4].

Apart from the statistical time series models for univariate and multivariate series, the combination of statistics and learning models have polished several machine learning algorithms, such as a critical neural networks, gradient boosted regression trees, support vector machines and, random forecast. These algorithms can reveal complex patterns characterized by non-linearity as well as some relations that are difficult to detect with linear algorithms.. A large number of studies is currently active on the subject of machine learning methods used in finance, some studies used tree-based models to predict portfolio returns .others used deep learning in the production of future values of financial assets.[2]

The dynamic, complex, evolutionary and chaotic nature of the market clearly demonstrates the limitations of classical statistical methods for forecasting time series of stock prices, and requires more powerful methods to complete the task. In particular, when dealing with market trends, we need methods that can work with a large amount of "noisy" and non-linear data. Given the disadvantages imposed by statistical methods, machine learning methods such as artificial neural networks (ANNs) in combination with heuristic algorithms will be used as an alternative.[7].

Various machine learning techniques have been proposed over the decades for predicting stock market prices. An Artificial Neural Network (ANN) is introduced, incorporating Long Short-Term Memory (LSTM) to enhance trend forecasting in stock market data. LSTM contributes to maintaining both short-term and long-term memory in conjunction with the temporal aspects of the data. The objective is to elevate the effectiveness of trend forecasting in stock market data, facilitating more informed trading decisions.[1][2].

II. ARIMA MODEL

An ARIMA model is a vibrant univariate forecasting method to project the future values of a time series. the Quantitative forecasting models make use of the data available to make predictions into future.

The model basically sums up the interesting patterns in the data and presents a statistical association between the past and current values of the variable. Likewise, we can say, that quantitative forecasting models are used to extrapolate the past and present behavior into future. Some examples of the Quantitative models include the regression analysis models, smoothing models and the time series models.[4]

Autoregressive Integrated Moving Average, is a commonly employed time series forecasting model within the realms of statistics and econometrics. This model is crafted to discern various aspects of time series data, such as trends and seasonality.

1) *Autoregressive (AR) Component:* This part entails modeling the connection between a current observation and past observations, known as lags.
   The "p" parameter signifies the number of lag observations incorporated into the model.
2) *Integrated (I) Component:* The integrated component involves differencing the time series data to achieve stationarity. Stationary data, characterized by a consistent mean and variance, facilitates easier modelling.
   The "d" parameter denotes the order of differencing necessary for attaining stationarity.
3) *Moving Average (MA) Component:* The moving average component models the association between a current observation and a residual error from a moving average model applied to lagged observations.The "q" parameter represents the order of the moving average.

By combining these components, ARIMA is represented as ARIMA(p, d, q). The objective is to identify the optimal values for these parameters, creating an effective model for forecasting future values within the time series. ARIMA equation can be written as:

$$Yt = \phi1\ yt\text{-}1 + et - \theta1\ et\text{-}1$$

III. VARMA MODEL

A *vector autoregression moving average (VARMA) model* is a multivariate time series model containing a system of *n* equations of *n* distinct, stationary response variables as linear functions of lagged responses and other terms. VARMA models are also characterized by their degree *p,q*; each equation in a VARMA(*p,q*) model contains *p,q* lags of all variables in the system.

.

VARMA(p,q) equation:

$$Y_t= c+ {}_{p}\sum_{j=1} \Phi jyt{-}j+ {}_{q}\sum_{k=1} \Theta k\varepsilon t{-}k{+}\varepsilon t$$

Where,

- y_t is the *n*-by-1 vector of distinct response time series variables at time *t*.
- *c* is an *n*-by-1 vector of constant offsets in each equation.
- Φ_j is an *n*-by-*n* matrix of AR coefficients, where $j = 1,...,p$ and Φ_p is not a matrix containing only zeros.
- ε_t is an *n*-by-1 vector of random Gaussian innovations, each with a mean of 0 and collectively an *n*-by-*n* covariance matrix Σ. For $t \neq s$, ε_t and ε_s are independent.
- Θ_k is an *n*-by-*n* matrix of MA coefficients, where $k = 1,...,q$ and Θ_q is not a matrix containing only zeros.
- Φ_0 and Θ_0 are the AR and MA structural coefficients, respectively.

A. Deep learning networks (neural network)

LSTM and RNN:

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) known for capturing past data to make future predictions. In an Artificial Neural Network (ANN) with one hidden layer, the input layer nodes connect to the hidden layer through synapses with weights as decision makers for signals. Learning involves adjusting these weights, and after completion, the ANN will have optimal weights. The hidden layer applies a sigmoid or tanh activation function on the weighted sum from the input layer. The output layer, obtained after applying the SoftMax function, minimizes error between training and test data.

For predicting future values based on past sequences, RNNs use earlier stages to learn and forecast trends. However, RNNs struggle with long-term memory. LSTM, with its "memory line" and gates, addresses this limitation, allowing the retention of information from earlier sequences for more accurate forecasting. [9]

The ability of memorizing sequence of data makes the LSTM a special kind of RNNs. Every LSTM node most be consisting of a set of cells responsible of storing passed data streams, the upper line in each cell links the models as transport line handing over data from the past to the present ones, the independency of cells helps the model dispose filter of add values of a cell to another. In the end the sigmoidal neural network layer composing the gates drive the cell to an optimal value by disposing or letting data pass through. Each sigmoid layer has a binary value (0 or 1) with 0 “let nothing pass through”; and 1 “let everything pass through.” The goal here is to control the state of each cell, the gates are controlled as follow: - Forget Gate outputs a number between 0 and 1, where 1 illustration “completely keep this”; whereas, 0 indicates “completely ignore this.” - Memory Gate chooses which new data will be stored in the cell. First, a sigmoid layer “input door layer” chooses which values will be changed. Next, a *tanh* layer makes a vector of new candidate values that could be added to the state. **-** Output Gate decides what will be the output of each cell. The output value will be based on the cell state along with the filtered and freshest added data. [9].

IV. METHODOLOGY

A. Data Collection

The data that is focused here is of Indian stock Infosys ltd.

the dataset is divided into 2 parts, one for univariate analysis and other one for multivariate analysis.

Infosys ltd stock closing price is taken under consideration from 1/05/2023 to 17/11/2023.the data is concentrated on hourly basis for 5 months having total of 250+ rows.

For multivariate analysis purpose, the same stock Infosys Ltd is taken with increased time period of 1/01/2008 to 30/11/2008 for daily basis consisting of 4000 rows.

B. Feature Selection

For Multivariate Analysis purpose for time series,The features selected for prediction and fitting are open values, close values, high values, low values for Infosys Stock and the stocks affecting the Indian stocks that are Sensex, Nifty50 are selected.

C. Procedure

For univariate time series data, Visualizing the time series data to identify trends, seasonality, and other patterns. Converting the data into stationarity and applying statistical tests before fitting.

Applying the Autoregressive Integrated Moving Average (ARIMA) model to the univariate time series (closing prices) with suitable parameters (p,q).

Assessing the performance of the ARIMA model using appropriate evaluation metrics (e.g., Mean Absolute Error, Mean Squared Error).

For multivariate time series data, Visualizing the time series data to identify trends, seasonality, and other patterns. Converting the data into stationarity and applying statistical tests before fitting.Fitting a Vector Autoregressive Moving Average (VARMA) model to capture the interdependencies among variables with suitable parameters(p,q).

D. Comparison with LSTM

Implementing the Long Short-Term Memory (LSTM) model for both univariate and multivariate time series prediction.

Utilizing the historical closing prices, opening prices, low prices, high prices, Sensex close values, and Nifty closing values.

V. ANALYSIS

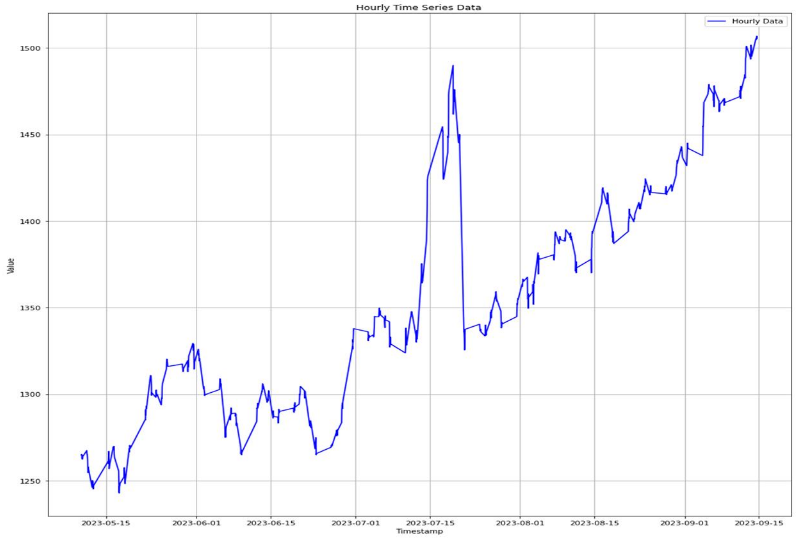
A. Univariate Time Series (ARIMA)

A time series needs to be lacking trend and seasonality, in order to be stationary. Such type of time series are characterized by having a constant variance and constant mean over a given period of time. The ``trend and seasonality" component may affect a time series at different instants.

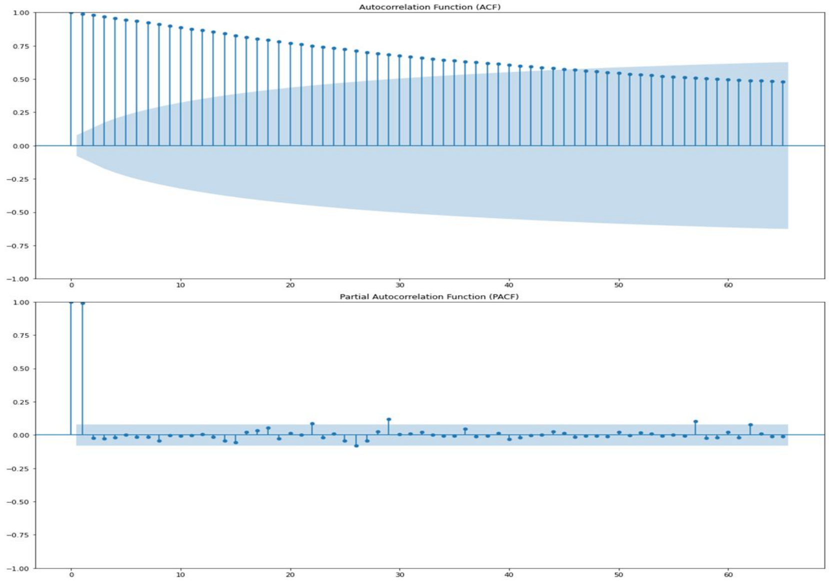
Visualization of univariate time series data that is Infosys LTD closing price stock for the time period of May-September 2023(hourly).

1) Statistical Analysis:

Data visualization for checking the presence of trend or seasonality, and using Augmented Dicky Fuller(ADF) test for checking stationarity. The p-value>0.05 implies the data being non-stationary and visualizations show trend in the time series data.



(Infosys Stock Data for may-september)



(ACF-PACF plots for non-stationary time series)

.

B. Stationary Time Series and Modelling:

Conversion of non-stationary time series data into stationary time series data using differencing technique and checking the data using ADF test. The p-value for ADF test is p-value<0.05 implying the stationarity of the data.

In the next step we go for forecasting the series. We use the ``ARIMA (p,d,q) model'' for predicting the next values in the time series. We use the auto.arima () function in to get the results. Auto.arima() function chooses the best parameters of ``ARIMA(p,d,q)'' to get the forecasted series. The auto.arima() function uses a `trace' that justifies why the parameters (p,d,q) chosen are best suited for the ``ARIMA(p,d,q) model''.

Where

P: defines the lags in PACF plot

q: defines the no. of lags in ACF plot

d: differencing

Since, the time series is already differenced and converted into stationarity, d=0 for our model fitting and forecasting. Using the auto modeling the results for suitable order are : (2,0,2)

lags	p	d	q	AIC
15	1	0		4223.782
20	2	0	2	4218.887
30	3	0	4	4222.796
40	4	0	4	4228.638
50	3	0	4	4222.796
65	2	0	4	4222.344
75	2	0	3	4224.51

The ARIMA model with lowest AIC value has order (2,0,2) which is same before differencing (2,1,2).
The RMSE value for ARIMA(2,0,2) model's predicted next 6 hours is **8.55675.**

C. Univariate Time Series (LSTM):

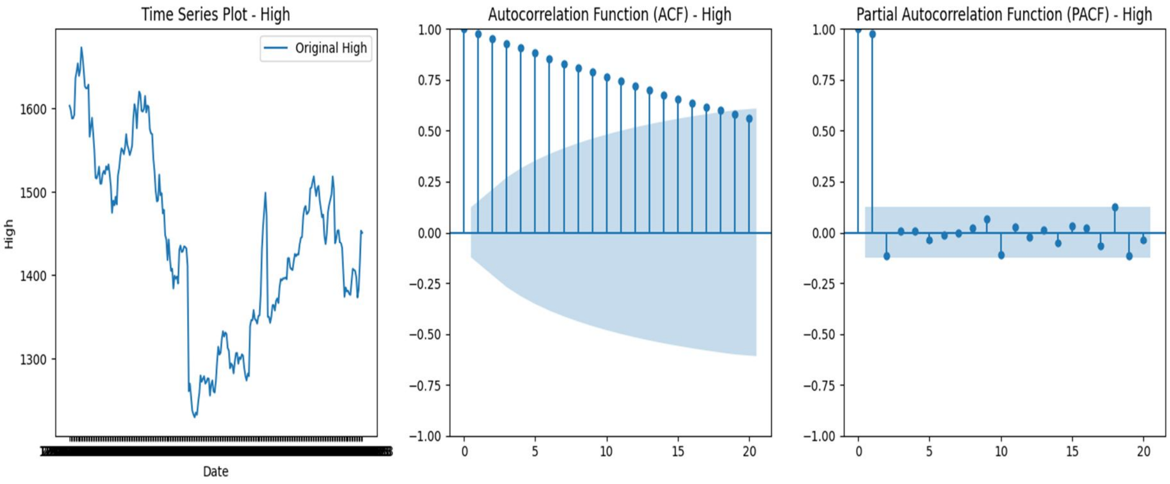
We use the same data of Infosys ltd May-September) on hourly basis for fitting LSTM model. LSTM was applied on the same data with Univariate time series values with window size as 10.
To build our model we are going to use the LSTM, our model uses 80% of data for training and the other 20% of data for testing. For training we use mean squared error to optimize our model. Also, We used 50 epochs for training and batch-size as 32 .
LSTM predicted the same next 6 hours value for the Infosys stock closing price for 15th September 2023.
The RMSE value for predicted v/s actual for LSTM is **4.0987.**

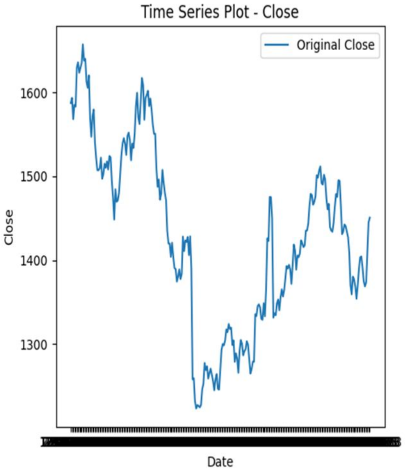
D. Multivariate Time Series (VARMA):

The data used for this model was Infosys Ltd's stock value, where the features considered were high, low, open, close and Sensex closing value.
The time period taken for this data was daily value from 1/11/2022 to 17/11/2023.
The data consisted of 250 rows and 5 columns.
The suitable multivariate time series model was fitted for the prediction of upcoming values.
The dataset was divided into training, testing and validation data and the best model is fit with respect to best p , q values.

1) Statistical Analysis:

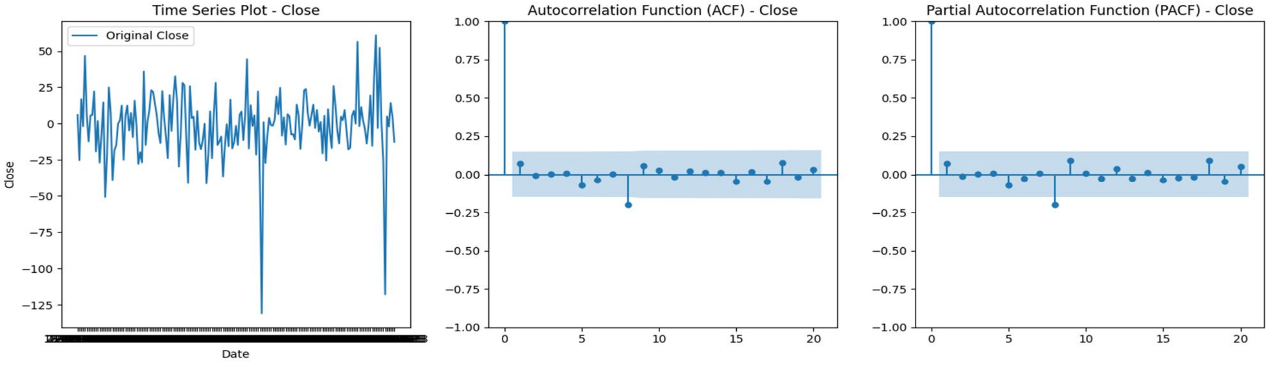
Data visualization for checking the presence of trend or seasonality, and using Augmented Dicky-Fuller (ADF) test for checking stationarity. The p-value>0.05 implies the data being non-stationary and visualizations show trend in the time series data.





Visually checking the trend present in the data and in ACF,PACF plots. The ADF test results also in p-value>0.05 resulting in non-stationary data.

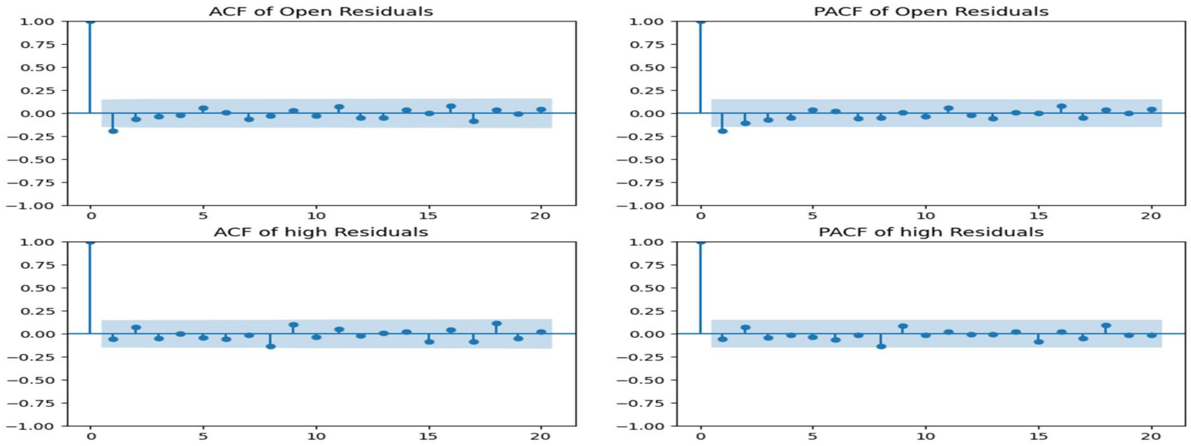
Converting the non-stationary data into stationary data by using differencing and checking the p-value for ADF test resulted in stationary data as shown in the plots.

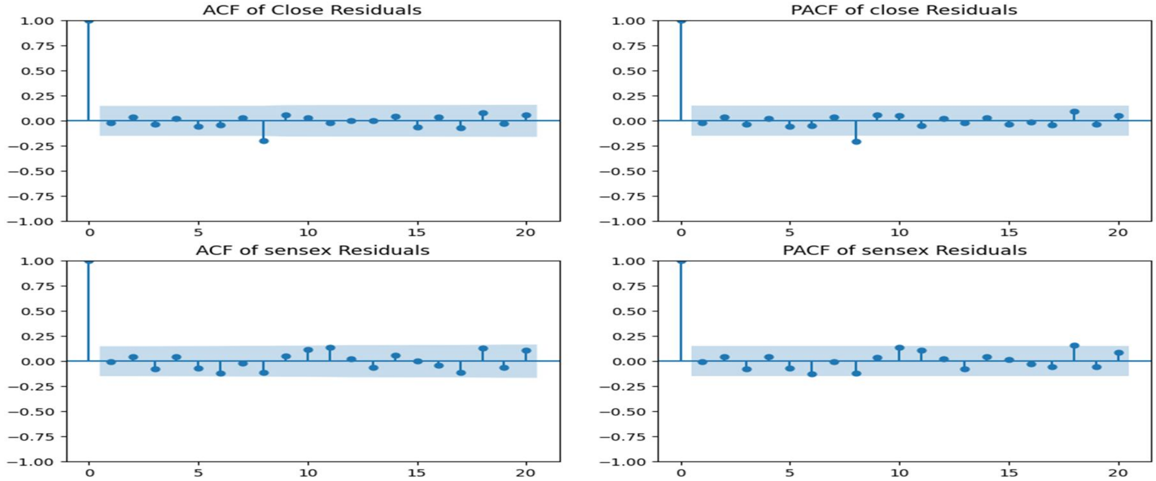


By looking at the plots, it is clear the time series data for all the features are stationary.

2) Residual Analysis:

Residual Analysis was conducted, as per time series assumption. The residuals were again tested using ADF test; with p-value<0.05.Resulting in stationary observations and uncorrelated variables. The residuals were visually displayed for the checking of stationarity with respect to ACF, PACF plots. The assumption of residuals are fulfilled which makes our model efficient and ready for forecasting.

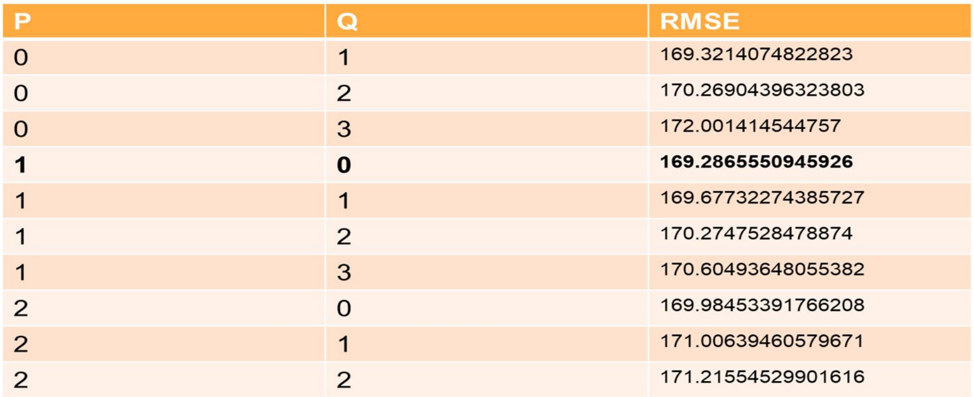




3) VARMA model fitting and forecasting:

The multivariate time-series VARMA model was fitted to forecast upcoming values of all the features which have high correlation among each other.

The best order (p , q) for the data with suitable p , q parameters is VARMA(1,0) which is VAR(1) .



P	Q	RMSE
0	1	169.3214074822823
0	2	170.26904396323803
0	3	172.001414544757
1	**0**	**169.2865550945926**
1	1	169.67732274385727
1	2	170.2747528478874
1	3	170.60493648055382
2	0	169.98453391766208
2	1	171.00639460579671
2	2	171.21554529901616

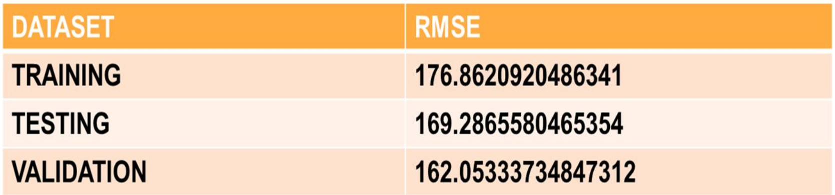
For the VAR(1,0) model fitting, the data was spit into training testing and validation data with training as 70%,testing as 25% and validation as 5% of the data.

After fitting the data with VARMA(1,0) model the following results were obtained:

The equation of the VARMA (1,0) model obtained as :

$$A_t = -0.0326A_{t-1} - .0565B_{t-1} - 0.2525C_{t-1} + 0.9505D_{t-1} - 0.0022F_{t-1} + \varepsilon_1$$
$$B_t = 0.0326A_{t-1} - 0.3776B_{t-1} - 0.1883C_{t-1} + 0.7575D_{t-1} - 0.0022F_{t-1} + \varepsilon_2$$
$$C_t = -0.0939A_{t-1} - 0.1506B_{t-1} - 0.6433C_{t-1} + 0.7833D_{t-1} - 0.0019F_{t-1} + \varepsilon_3$$
$$D_t = -0.2344A_{t-1} - 0.3449B_{t-1} - 0.0628C_{t-1} + 0.0073D_{t-1} - 0.0005F_{t-1} + \varepsilon_4$$
$$F_t = -1.46379A_{t-1} + 2.1152B_{t-1} + 0.4991C_{t-1} + 0.1432D_{t-1} + 0.0463F_{t-1} + \varepsilon_5$$

The RMSE values for the model is as :



DATASET	RMSE
TRAINING	176.8620920486341
TESTING	169.2865580465354
VALIDATION	162.05333734847312

E. Multivariate Time Series (LSTM):

To compare our results, we used LSTM model for fitting and prediction of the same data set.

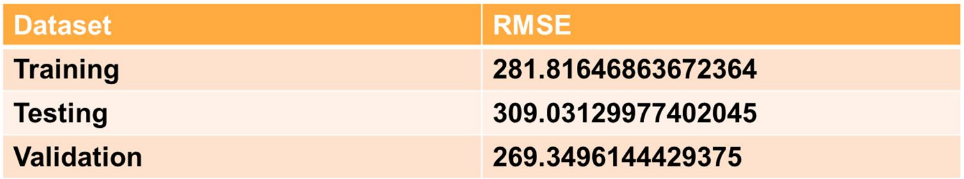
The data used for this model was Infosys Ltd's stock value, where the features considered were high, low, open, close and Sensex closing value. The time period taken for this data was daily value from 1/11/2022 to 17/11/2023.The data consisted of 250 rows and 5 columns same as for VAR model.

The data was again divided into 3 parts, the data was spit into training testing and validation data with training as 70%, testing as 25% and validation as 5% of the data

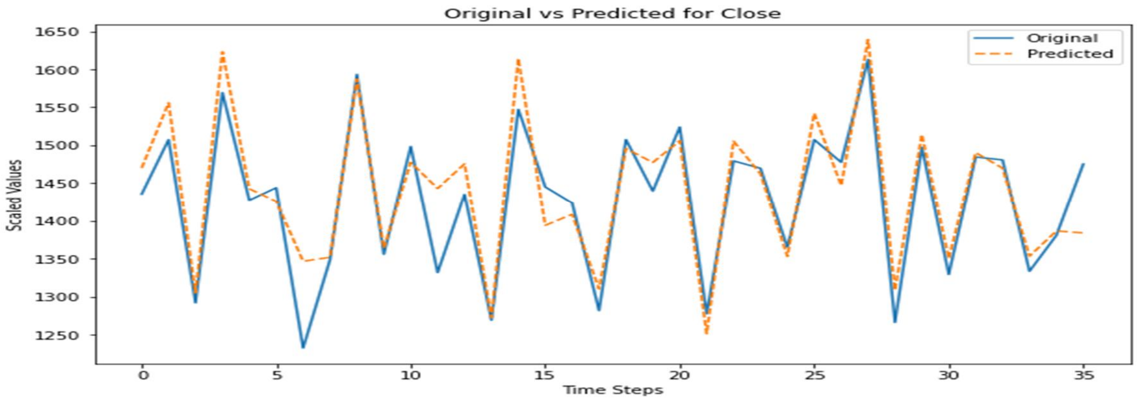
The LSTM model created and fitted with 50 units and a single hidden layer and batch size of 32.

The no. of epochs run for the model were 100.

For the LSTM sequential model the results obtained are:



Dataset	RMSE
Training	281.81646863672364
Testing	309.03129977402045
Validation	269.3496144429375

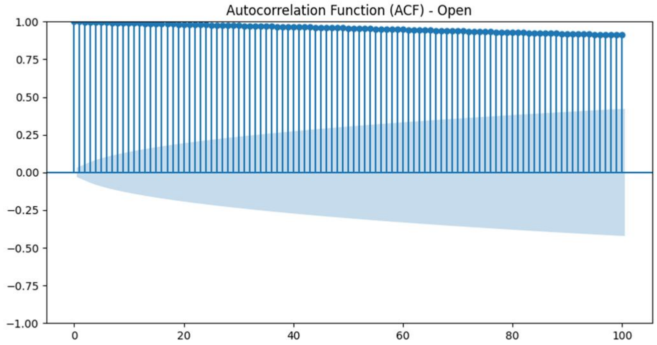


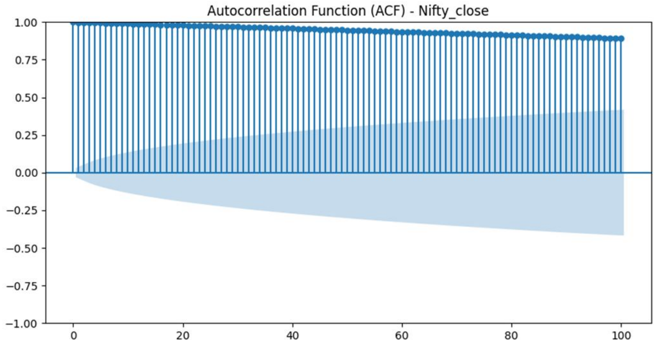
The second analysis is also done on Multivariate Analysis. The dataset chosen is Infosys LTD stock data, with features as open, close, high, low values and closing values of Sensex and Nifty50 to understand the impact of these stocks on Infosys Stock price.

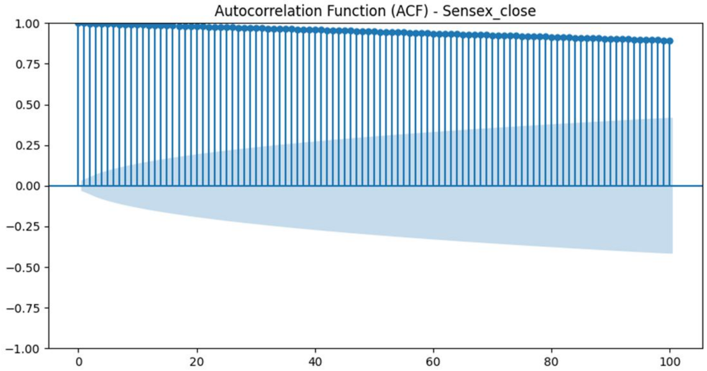
The dataset was taken for a span of 15 years on daily basis from 1/01/2008 to 30/11/2023.The data consisted of 4000 rows and 6 columns.

1) Statistical Analysis:

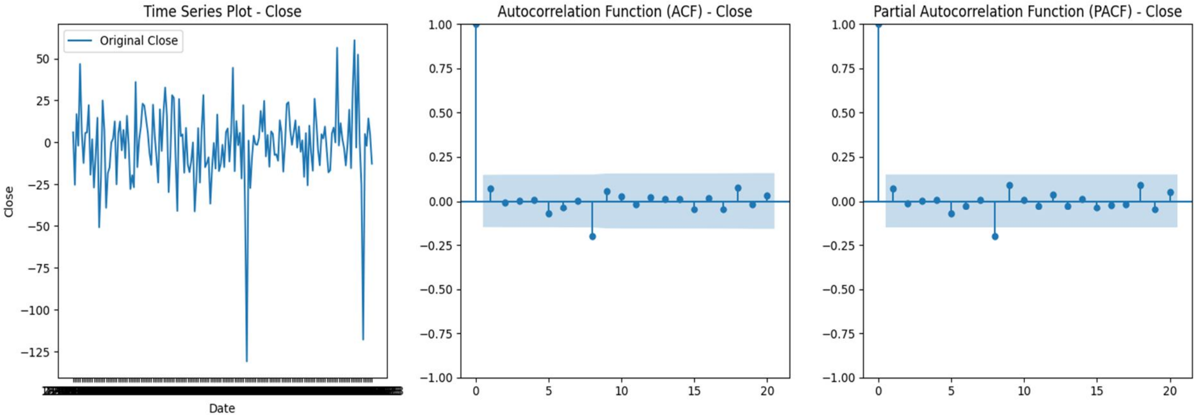
Data visualization for checking the presence of trend or seasonality, and using Augmented Dicky-Fuller (ADF) test for checking stationarity. The p-value>0.05 implies the data being non-stationary and visualizations show trend in the time series data. ACF plot shows trend present in the series and need of conversion to non-stationary data.







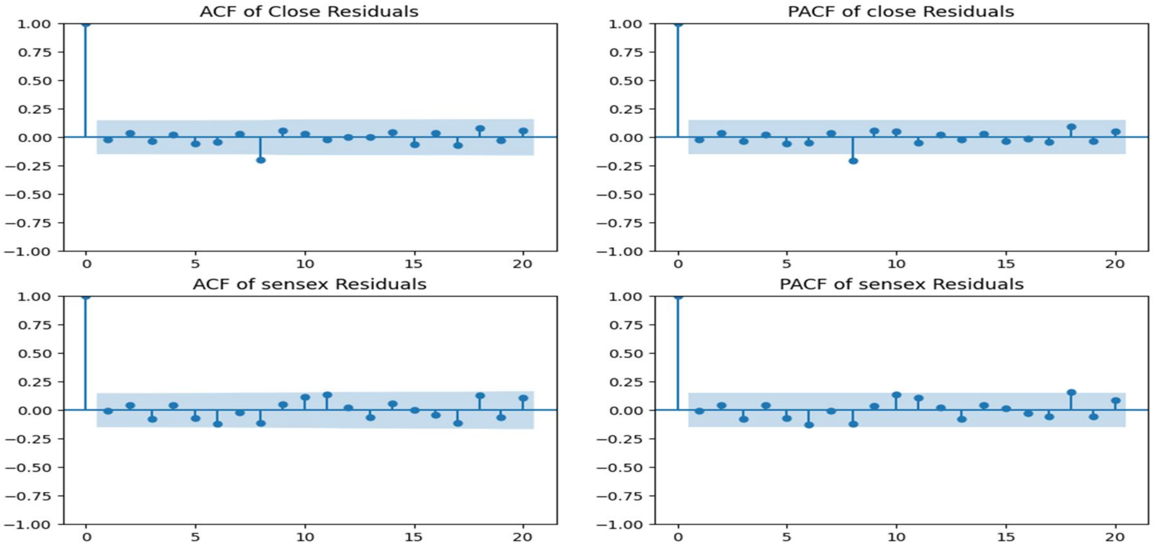
Converting the non-stationary data into stationary data by using differencing and checking the p-value for ADF test resulted in stationary data as shown in the plots.



By looking at the plots, it is clear the time series data for all the features are stationary.

2) Residual Analysis:

Residual Analysis was conducted, as per time series assumption. The residuals were again tested using ADF test; with p-value<0.05.Resulting in stationary observations and uncorrelated variables. The residuals were visually displayed for the checking of stationarity with respect to ACF, PACF plots. The assumption of residuals are fulfilled which makes our model efficient and ready for forecasting.



3) VARMA model fitting and forecasting:

The multivariate time-series VARMA model was fitted to forecast upcoming values of all the features which have high correlation among each other.

The best order (p,q) for the data with suitable p , q parameters is VARMA(0,2) which is VMA(2) .

P	Q	RMSE
1	1	116.4190866047861
1	2	116.41829815382188
2	0	116.42064417195844
2	1	116.4234194210862
2	1	116.4234194210862
0	2	116.40762353826513
1	0	116.4221534325657

For the VARMA (0,2) model fitting, the data was split into training testing and validation data with training as 70%,testing as 25% and validation as 5% of the data.

After fitting the data with VARMA (0,2) model the following results were obtained:

The equation of the VARMA (0,2) model obtained as :

The RMSE values for the model is as:

DATASET	RMSE
training	116.40762353826513
testing	165.1540209329463
validation	167.68623209975104

F. LSTM Analysis

To compare our results, we used LSTM model for fitting and prediction of the same data set. The dataset chosen is Infosys LTD stock data, with features as open, close, high, low values and closing values of Sensex and Nifty50 to understand the impact of these stocks on Infosys Stock price.

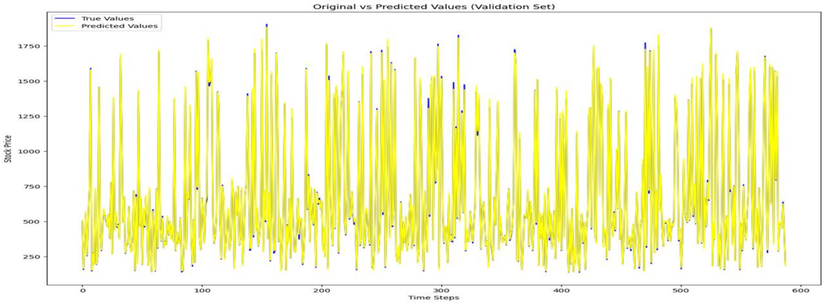
The dataset was taken for a span of 15 years on daily basis from 1/01/2008 to 30/11/2023.The data consisted of 4000 rows and 6 columns.

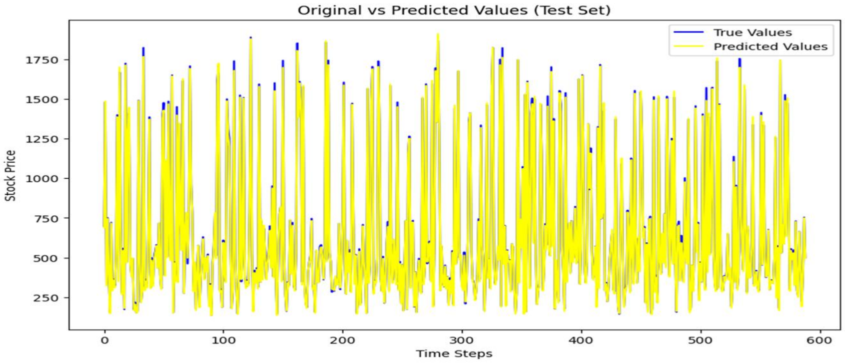
The LSTM model is fitted for this data using hyperparameter tuning and selecting best parameters of LSTM for best fit and lowest RMSE values.The parameters considered for cross validation and hyper-parameter tuning are no. of epochs, no. of neurons,batch size and sequence for fitting the model.The data is transformed using standardization and LSTM model is fit using these parameters.

The results obtained for this model (sequence=8) is:

units	epochs	Batch_size	RMSE_test
50	50	32	247.8991326155294
50	50	64	296.7788849332451
100	50	32	252.76396278277406
100	50	64	281.49965945766996
50	100	32	258.5460915140095
50	100	64	281.23587958785924
100	**100**	**32**	**207.85170009377197**
100	100	64	268.88421135045945

The best Model for fitting and forecasting with lowest RMSE for number of sequence is LSTM(units=100,epochs=100,batch_size=32)





For the same hyper-parameters of LSTM with sequence number=10,The best suited LSTM model is LSTM(units=100,epochs=100,batch_size=32).

The lowest RMSE value obtained for test data is 196.2530982315928.The RMSE value for the validation dataset is 186.56467958304881.

The best suited LSTM model for the dataset of 15 years is LSTM(units=100,epochs=100,batch_size=32) with number of sequence to be 10.

VI. RESULTS

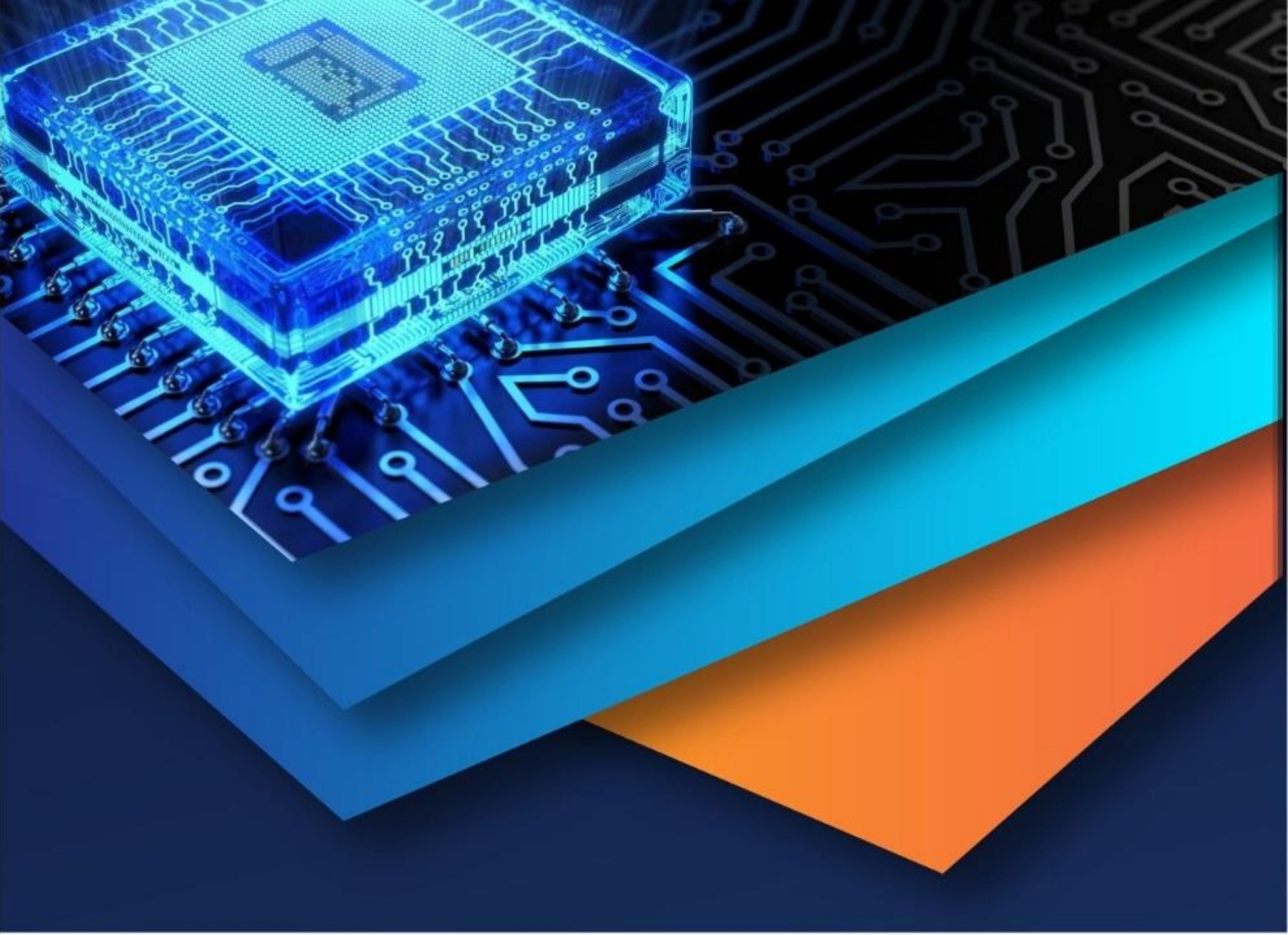
For univariate time series analysis, the ARIMA model, configured as ARIMA(2,0,2), showcased its ability to discern trends and seasonality, yielding a respectable RMSE of 8.55675. However, the LSTM model outperformed ARIMA with a notably lower RMSE of 4.0987, underscoring its proficiency in capturing intricate patterns in stock prices.

In the multivariate realm, VARMA models, particularly VARMA(1,0), demonstrated their effectiveness in capturing dependencies among features, resulting in an RMSE of 162 for the first dataset. In comparison, LSTM, with hyperparameter tuning, exhibited a competitive RMSE of 269. For the second dataset, VARMA achieved an RMSE of 167, while LSTM, with meticulous tuning, recorded an RMSE of 182.

VII. CONCLUSION

In Conclusion, this research undertook a comprehensive exploration of stock market forecasting using univariate and multivariate time series models, with a specific focus on ARIMA, VARMA, and LSTM methodologies. The investigation utilized datasets extracted from the Indian stock market, primarily centered around Infosys Ltd.

The comparative analysis highlighted the strengths and weaknesses of each model. While VARMA models provided valuable insights into the interplay among various features, LSTM's advanced deep learning architecture demonstrated a superior ability to handle non-linear relationships and intricate patterns. The choice between these models would depend on the specific requirements of the forecasting task and the nature of the underlying data. This research sheds light on the dynamic interplay between traditional statistical models and advanced machine learning techniques in the context of stock market forecasting. Investors and analysts can leverage these findings to make informed decisions, navigating the complexities of the financial landscape. As the field continues to evolve, the synergy between statistical and machine learning models will likely play a crucial role in shaping the future of stock market predictions, offering improved insights for optimizing investment strategies.



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY