

Stock Price Prediction for IT, Banking and Health Care using ARIMA Model

submitted by

Nitish Kumar Mandal

Class Roll No: 002210503052

Exam Roll: MCA244051

Master of Computer Application

under the supervision of

Prof. (Dr.) Mahantapas Kundu

in the

Department of Computer Science and Engineering

Faculty of Engineering and Technology

Jadavpur University

2024

Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF RECOMMENDATION

To Whom It May Concern

I hereby recommend that the project report entitled “Stock Price Prediction for IT, Banking and Health Care using ARIMA Model” has been carried out by Nitish Kumar Mandal (Reg. No.: 163644 of 2022-23, Roll No.: 002210503052, Exam Roll: MCA244051), under my guidance and supervision and be accepted in partial fulfilment of the requirement for the degree of Master of Computer Application in department of Computer Science and Engineering, Jadavpur University during the academic year 2023-2024.

Prof. (Dr.) Mahantapas Kundu
Project Supervisor
Dept. of Computer Science and Engineering
Jadavpur University

Prof. (Dr.) Nandini Mukhopadhyay
Head of the Department
Dept. of Computer Science & Engineering
Jadavpur University

Prof. Dipak Laha
Dean, Faculty Council
of Engineering & Technology
Jadavpur University, Kolkata-032

Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF APPROVAL

This is to certify that the project report entitled “**Stock Price Prediction for IT, Banking and Health Care using ARIMA Model**” is a bona fide record of work carried out by Nitish Kumar Mandal, in fulfillment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University. It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approves the project report only for the purpose for which it has been submitted.

Signature of Examiner 1

Date :

Signature of Examiner 2

Date :

Faculty of Engineering and Technology
Jadavpur University

**DECLARATION OF COMPLIANCE WITH
ACADEMIC ETHICS**

I hereby declare that this project report entitled " Stock Price Prediction for IT, Banking and Health Care using ARIMA Model " is an outcome of my genuine and independent work, as part of my Master of Computer Application program.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, all materials and results that are not original to this work have been fully cited and referenced.

Nitish Kumar Mandal

Registration No.: 163644 of 2022-23

Examination Roll No.: MCA244051

Thesis Title :- Stock Price Prediction for IT, Banking and Health Care using ARIMA Model.

Signature with Date

ACKNOWLEDGEMENTS

With my most sincere and gratitude, I would like to thank Prof. (Dr.) Mahantapas Kundu, Department of Computer Science & Engineering, my supervisor, for his overwhelming support throughout the duration of the project. His motivation always gave me the required inputs and momentum to continue with my work, without which the project work would not have taken its current shape. His valuable suggestions and numerous discussions have always inspired new ways of thinking. I feel deeply honored that I got this opportunity to work under him.

I would like to express my sincere thanks to all my teachers for providing sound knowledge base and cooperation.

I would like to thank all the faculty members of the Department of Computer Science & Engineering of Jadavpur University for their continuous support.

Last, but not the least, I would like to thank my batch mates for staying by my side when I needed them.

Nitish Kumar Mandal

Class Roll No.: 002210503052

Examination Roll No.: MCA244051

Department of Computer Science & Engineering

Jadavpur University

Contents

Sl. No	Topic	Page No.
	Abstract	01
Chapter 1	Introduction	02 - 05
1.1	Stock Price Analysis	02
1.2	Stock Price Prediction	02 - 03
1.3	Applications	03
1.4	Motivation	04
1.5	Scope of the Present Report	04
1.6	Organization of Project Work	04 - 05
Chapter 2	Related Works	05 - 07
Chapter 3	Methodology	08 - 10
3.1	ARIMA Model	08 - 10
3.1.1	Testing and Ensuring Stationarity	09 - 10
3.1.2	Identification of p and q	10
3.1.3	Estimation and Forecasting	10
Chapter 4	Experimental Results	11 - 39
4.1	Dataset Description	11
4.2	IT Sector	11 - 18
4.2.1	Dataset	11 - 12
4.2.2	Data Preprocessing and Visualization	12
4.2.3	Testing and Ensuring Stationarity	13 - 15
4.2.4	Identification of p and q	15 - 16
4.2.5	Dataset Split into Train and Test	16 - 17
4.2.6	Model Training	17 - 18
4.2.7	Summary	18
4.3	Banking Sector	19 - 26
4.3.1	Dataset	19
4.3.2	Data Preprocessing and Visualization	19 - 20
4.3.3	Testing and Ensuring Stationarity	20 - 22
4.3.4	Identification of p and q	23 - 24
4.3.5	Dataset Split into Train and Test	25
4.3.6	Model Training	25 - 26
4.3.7	Summary	26
4.4	Health Care Sector	27
4.4.1	Dataset	27
4.4.2	Data Preprocessing and Visualization	27 - 28
4.4.3	Testing and Ensuring Stationarity	28 - 30
4.4.4	Identification of p and q	31 - 32
4.4.5	Dataset Split into Train and Test	33
4.4.6	Model Training	33 - 34

4.4.7	Summary	34
4.5	MAPE	35
4.6	Results on Data from IT Sector	35 - 36
4.7	Results on Data from Banking Sector	37 - 38
4.8	Results on Data from Health Care Sector	38 - 39
Chapter 5	Conclusion	40
	References	41 - 42

ABSTRACT

Predicting stock prices is crucial for making smart financial decisions, and aiding investors in understanding and navigating its effectiveness and limitations the ever-changing financial markets. This research explores using the Autoregressive Integrated Moving Average (ARIMA) model to predict stock prices, assessing its effectiveness and limitations. The study explains the ARIMA model's theory, highlighting its ability to identify patterns over time, making it useful for predicting future prices. Historical stock price data from various sectors, including IT, Banking, and Health Care, is used to apply the ARIMA model, and its performance in making accurate predictions is evaluated.

The methodology involves several key steps, including data preprocessing, the ADF test, parameter estimation, model selection, and model evaluation. Data preprocessing involves cleaning and transforming raw stock price data to ensure consistency and reliability in the analysis. The ADF test helps determine whether a time series is stationary and can indicate if differencing is needed. Parameter estimation is done by evaluating the PACF and ACF plots. Model selection entails identifying the most appropriate ARIMA specifications, considering factors such as stationarity, autocorrelation, and partial autocorrelation in the data. Model evaluation is conducted using Mean Absolute Percentage Error (MAPE) to assess the accuracy and robustness of the ARIMA model's predictions.

The analysis of the ARIMA model encompasses a comprehensive examination of its predictive capabilities across different stocks and time periods. This research evaluates the ARIMA model's performance in predicting stock prices from various sectors, including IT, Banking, and Health Care. By analyzing historical stock price data, the study assesses the model's accuracy and reliability in forecasting future stock prices. The results provide insights into the strengths and limitations of the ARIMA model in different market conditions, offering valuable information for investors and analysts in making informed decisions based on forecasted stock price trends.

The findings contribute to the existing knowledge on stock price prediction methodologies, offering valuable insights and practical guidance for investors, financial analysts, and researchers. This research advances understanding of stock price prediction using the ARIMA model, paving the way for future research endeavors.

Chapter 1

Introduction

1.1 Stock Price Analysis

The advancements in stock price prediction have gained significant importance among expert analysts and investors. The stock market prediction for analysing the trends is complicated due to intrinsic noisy environments and large volatility with respect to the market trends. The complexities of the stock prices adapt certain factors that involve quarterly earnings' reports, market news, and varying changing behaviours. The traders depend on various technical indicators that are based on the stocks, which are collected on a daily basis. Even though these indicators are used to analyse the stock returns, it is complicated to forecast daily and weekly trends in the market [1]. The accurate prediction of stock trends is interesting and a complex task in the changing industrial world. Several aspects, which affect the behaviour of stock trends, are economic and non-economic, which are taken into consideration. Thus, predicting the stock market is considered as a major challenge for increasing production [2]. Traditional techniques reveal that the stock market earnings are predicted from previous stock returns and other financial variables and macroeconomics. The prediction of stock market revenues directed the investors towards examining the causes of predictability. The forecasting of stock trends is a difficult process as it is influenced by several aspects, which involve trader's expectations, financial circumstances, administrative events, and certain aspects related to the market trends. Moreover, the list of stock prices is usually dynamic, complicated, noisy, nonparametric, and nonlinear by nature [3]. The forecasting of financial time series becomes an issue due to certain complex features, like volatility, irregularities, noise, and changing trends [4].

1.2 Stock Price Prediction

Due to rapid improvement in Computing Performance, many Stock Prediction Techniques are come into picture and methods for fining information from data are gaining more interest [5]. Working with large volumes of data has become increasingly common in modern times. Through the application of diverse techniques, data mining can uncover patterns within stock prices, facilitating the prediction of future stock values.

For a while now, researchers have been really interested in guessing what stock prices will be in the future. Some people think it's impossible to predict stock prices accurately, but many studies disagree with this idea. They say that if you create and fine-tune predictive models the right way, they can actually predict future stock prices very accurately and reliably. They've found that how you pick the information to use in

your model, the techniques you use, and how you make the model better all impact how accurate it is. In this context, researchers suggested a way to predict stock prices by breaking down time series data. Machine learning and deep learning are now widely used to predict stock prices and understand how they change over time. One method for predicting stock prices involves breaking down the time series data into different components. A Time-Series decomposition-based approach was also used to conduct predictive analysis of the FMCG, realty, large-cap, mid-cap, banking, and other sectors of India [6-8]. Researchers use some models to predict stock price, those are: One time series model (Holt-Winters Exponential Smoothing), one econometric model (ARIMA), two machine learning models (Random Forest and MARS), and two deep learning-based models (simple RNN and LSTM).[11]

1.3 Applications

In this section, we will delve into practical applications of time series analysis, focusing particularly on its relevance in predicting stock prices. Time series analysis plays a crucial role in accurately forecasting future trends and patterns, empowering investors to make informed decisions in the stock market. Let's examine some examples:

- **Stock Price Prediction:** Time series analysis enables analysts to study historical stock price and identify patterns and trends. By leveraging technique such as ARIMA (Autoregressive Integrated Moving Average) investors can forecast future stock prices with a reasonable degree of accuracy.
- **Market Trend Analysis:** Time series analysis helps in analyzing market trends over time, including identifying bullish (rising) or bearish (falling) market trends. This information is crucial for investors to adjust their investment strategies accordingly.
- **Risk Management:** By analyzing historical stock price data using time series techniques, investors can assess the risk associated with different investment options. This allows them to make informed decisions to mitigate potential losses.
- **Portfolio Optimization:** Time series analysis aids in optimizing investment portfolios by identifying correlations and relationships between different stocks and assets. Investors can use this information to diversify their portfolios effectively and maximize returns while minimizing risk.

1.4 Motivation

Accurate prediction of stock prices is crucial for investors, financial institutions, and the economy as a whole. It aids in making informed decisions, managing risks, and maximizing returns. However, forecasting stock prices is challenging due to the inherent complexity and volatility of financial markets, as well as the presence of various sources of uncertainty and noise. To address these challenges, time series analysis, particularly using the ARIMA model, has emerged as a powerful tool for modeling and predicting stock prices. ARIMA models can capture the temporal dependencies and patterns in stock price data, incorporating factors such as historical trends, seasonality, and market dynamics. Additionally, advancements in machine learning have paved the way for the development of sophisticated models capable of handling large-scale datasets and achieving high prediction accuracy. By applying the ARIMA model and comparing it with other traditional forecasting techniques, such as linear regression and decision trees, we aim to identify the most effective approach for predicting stock prices. Our study provides valuable insights into the performance of different models in stock price prediction, informing future research and practical applications in financial markets.

1.5 Scope of Present Work

The aim of this study is to develop ARIMA model for predicting stock prices in three different sectors: Information Technology (IT), Banking, and Healthcare. Specifically, the models under comparison will include the ARIMA model. The focus will be on utilizing historical stock price data to forecast future stock prices accurately and assessing the performance of the developed models. Additionally, the research will examine that how a model can accurately predict different sector stocks. which can offer valuable insights for portfolio management and investment strategies.

1.6 Organization of Project Work

In this thesis, we present a study on time series analysis for stock price prediction. The goal of this work is to develop accurate prediction model that can help investors and financial institutions to make informed decisions, managing risks, and maximizing returns. This work can be divided into following chapters.

- **Chapter 1: Introduction**

This chapter provides an overview of the research problem, objectives, and motivation behind the study. It also outlines the scope of the research, and presents a brief overview of the methodologies and approaches used.

- **Chapter 2: Related Work**

This chapter reviews the existing work on stock price analysis and stock price prediction. It discusses the various models and techniques used in these fields and highlights the strengths and limitations of each approach. The chapter also examines the recent developments in these areas and discusses the research gaps and opportunities for future research.

- **Chapter 3: Data Collection, Preprocessing, Result and Analysis**

This chapter describes the data collection process, including the sources and types of data used in the study. It also discusses the steps taken to preprocess and clean the data, This chapter also presents the results and analysis of the study, including the performance of the developed models in predicting stock price.

- **Chapter 4: Conclusion and Future work**

This chapter outlines the significant findings and contributions of the study, addressing both its limitations and potential avenues for future research. Additionally, it offers recommendations tailored to investors and financial institutions to enhance their stock analysis methodologies and provide insights for informed decision-making regarding buying or selling actions.

Chapter 2

Related Works

This chapter presents a thorough literature review on time series analysis in the context of stock price prediction. To evaluate the model's effectiveness in predicting stock prices accurately, it is essential to analyze various stocks and assess how well the model performs across different datasets. For predicting stock prices, it is essential to work with the daily closing price of a stock. The current state of stock prediction research papers can be divided into three groups.

Holt-Winters Exponential Smoothing deals with univariate data. This method takes into account the weighted averages of prior data, where the weights are decreasing exponentially as the observations grow older. Single exponential smoothing captures single parameter *alpha*, which captures the level of the time series. Double Exponential smoothing, which is also called the Holts Exponential smoothing, captures level and

trend. In Holts Winter Exponential Smoothing, along with level and trend, seasonality in the data is also captured. Here a parameter called *gamma* is added, along with *alpha* and *beta*, to control the influence of the seasonal component. The capturing of the three parameters results in a higher accuracy value in forecasting [11]. This model works well in producing short time forecasts, but it has shortcomings, including the normalization of seasonal indices, the choice of the starting values, and the choice of smoothing parameters [10]. Among three sectors Holt-Winters Exponential Smoothing model performed best in the case of the health sector (on SUN PHARMA data) and performed poorly for the banking sector (on ICICI) data [11].

ARIMA comprises of three terms Autoregression (AR) which means the forecasted value depends on linear combinations of own past observed values. Integrated (I) means differencing of the time series to make it stationary and Moving Average (MA) which means that the number of lagged forecast errors should be used in the ARIMA Model. ARIMA models are known to be robust and efficient in financial time series forecasting especially short-term prediction than even the most popular ANNs technique. But the only disadvantage associated with the method is the series needs to be converted to stationary before applying ARIMA. Moreover, the assumption of the ARIMA model is constant variance, but financial time series exhibit changes in volatility, and this feature doesn't come under the ARIMA assumption [12]. ARIMA performed well for both the IT (Infosys) and banking (ICICI) sector and performed poorly for the healthcare sector (SUN PHARMA) [11].

MARS and Random Forest are the machine learning techniques also used for stock price prediction. MARS (Multivariate Adaptive Regression Splines) is a powerful machine learning technique that combines the flexibility of regression with the adaptability of decision trees. MARS is particularly effective for modeling complex relationships in data without requiring prior assumptions about the underlying structure. By partitioning the input space into smaller regions and fitting simple linear models within each region, MARS can capture nonlinearities and interactions between variables. This makes it suitable for tasks such as regression, classification, and prediction. Among the three industries, the model performed best in the case of the banking Sector (on ICICI data) and performed poorly in the case of the health Sector (on SUN PHARMA data) [11].

Random Forest is a powerful and versatile machine learning algorithm used for both classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or the mean prediction (regression) of the individual trees. This ensemble method harnesses the collective wisdom of multiple decision trees to improve prediction accuracy and generalization while mitigating overfitting. Random Forest is renowned for its robustness, scalability, and ability to handle high-dimensional data with ease. Among the three industries, the model performed best on the health sector (SUN PHARMA data) and performed poorly on the Banking Sector (ICICI data) based on RMSE/Test Mean ratio [11].

For deep learning, two models – Simple RNN and Stacked LSTM – were used for model building. For RNN or LSTM, we predict values based on past records as well. Here the sequencing is important. RNN works on the elements stored in the short-term memory. These elements help in the sequence prediction. It has at least one feedback connection, which helps in looping activations. For model building, in the first case, we go for the simple RNN model. It will have a sequential of RNN layers followed by a dropout layer. Finally, a dense layer is used, which is a deeply connected neural network layer. However, simple RNN has some shortcomings. It cannot remember memories that are too old. Also, it faces a vanishing gradient problem, i.e., the gradient almost has no effect during backpropagation due to the presence of so many layers. LSTM handles this issue well. LSTMs are capable of learning long-term dependencies. It contains a simple RNN cell, cell state (used for long-term memory), and three gates – forget, input and output. Forget gate basically determines which data to keep and which one to remove from memory. Among the three industries, the LSTM model also performed best on the IT stocks and performed poorly on Banking stocks. The prediction was made on the truncated values using Simple RNN and Stacked LSTM [11].

Chapter 3

Methodology

3.1 ARIMA Model

Stock market forecasting has always been a subject of great interest for investors, analyst and traders seeking to make informed investment decisions. One of the most popular methods in time series analysis is the Autoregressive Integrated Moving average (ARIMA) model. This model can able to capture trends, seasonality, and stationary behavior, the ARIMA has proven to be a powerful tool for forecasting stock prices [13].

The ARIMA model is a handy tool for analyzing and predicting sequential data. It combines three important elements:

- Autoregressive (AR) for looking at past observations
- Differencing (I or Integrated) for handling changes in data patterns
- Moving average (MA) for considering error terms

ARIMA is also known as the Box-Jenkins approach. Box and Jenkins claimed that non-stationary data can be made stationary by differencing the series. A Pure AR (Auto-Regressive only) Model is a model which relies only on its own lags. Hence, we can also conclude that it is a function of the lags of Y_t

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \epsilon_t$$

where, Y_{t-1} is the lag1 of the series. β_1 is the coefficient of lag1 and is the term of intercept that is calculated by the model.

Similarly, a Pure MA (Moving Average only) model is a model where Y_t relies only on the lagged predicted errors.

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \dots + \phi_q \epsilon_{t-q}$$

Where, the error terms are the AR models errors of the corresponding lags. The errors ϵ_t and ϵ_{t-1} are the errors from the equations given below:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_0 Y_0 + \epsilon_t$$

$$Y_{t-1} = \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \dots + \beta_0 Y_0 + \epsilon_{t-1}$$

Thus, we have concluded Auto-Regressive (AR) and Moving Average (MA) models, respectively.

Let us now understand the equation of an ARIMA Model.

An ARIMA model is a model where the series of time was subtracted at least once in order to make it stationary, and we combine the Auto-Regressive (AR) and the Moving Average (MA) terms. Hence, we got the following equation:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \dots + \phi_q \epsilon_{t-q}$$

Forecasted Y_t = Constant + Linear Combination Lags of Y (up to p lags) + Linear Combination of Lagged Predicted Errors (up to q lags)[15].

3.1.1 Testing and Ensuring Stationarity

To model a time series with the Box-Jenkins approach, the series has to be stationary. A stationary time series means a time series without trend, one having a constant mean and variance over time, which makes it easy for predicting values.

- **Testing for stationarity** - We test for stationarity using the Augmented Dickey-Fuller unit root test. The p-value resulting from the ADF test has to be less than 0.05 or 5% for a time series to be stationary. If the p-value is greater than 0.05 or 5%, you conclude that the time series has a unit root which means that it is a non-stationary process [13].
- **Differencing (d)** – To convert a non-stationary process to a stationary process, we apply the differencing method. Differencing a time series means finding the differences between consecutive values of a time series data. The differenced values form a new time series dataset which can be tested to uncover new correlations or other interesting statistical properties[16].

We can apply the differencing method consecutively more than once, giving rise to the "first order differencing", "second order differencing", etc. This how each order differencing works:

First-order differencing ($d=1$): This involves subtracting each observation from its preceding observation. It eliminates a linear trend from the data.

Second-order differencing ($d=2$): If the first-differenced series is still non-stationary, a second-order differencing can be applied to remove any remaining trend or seasonality.

Hence, the purpose of a higher order differencing is to make the time series data completely stationary.

We apply the appropriate differencing order (d) to make a time series stationary before we can proceed to the next step.

3.1.2 Identification of p and q

Since 'd' or differencing has been applied in the last step, in this step, we identify the appropriate 'p' order of Autoregressive (AR) model and 'q' order of Moving average (MA) model processes by using the Autocorrelation function (ACF) and Partial Autocorrelation function (PACF)[15].

Identifying the p order of AR model : For AR models, the ACF will dampen exponentially and the PACF will be used to identify the order (p) of the AR model. If we have one significant spike at lag 1 on the PACF, then we have an AR model of the

order 1, i.e. AR(1). If we have significant spikes at lag 1, 2, and 3 on the PACF, then we have an AR model of the order 3, i.e. AR(3).

Identifying the q order of MA model: For MA models, the PACF will dampen exponentially and the ACF plot will be used to identify the order of the MA process. If we have one significant spike at lag 1 on the ACF, then we have an MA model of the order 1, i.e. MA(1). If we have significant spikes at lag 1, 2, and 3 on the ACF, then we have an MA model of the order 3, i.e. MA(3).

3.1.3 Estimation and Forecasting

Once we have determined the parameters (p , d , q) we estimate the accuracy of the ARIMA model on a training data set and then use the fitted model to forecast the values of the test data set using a forecasting function.

In the end, we cross-check whether our forecasted values are in line with the actual values.

Chapter 4

Experimental Results

4.1 Dataset Description

The use of an appropriate dataset is crucial to demonstrate the effectiveness of any predictive model. To evaluate the accuracy of our model, we utilized datasets from three different sectors: TCS (IT sector), HDFC Bank (Banking sector), and CIPLA (Healthcare sector). By analyzing these diverse stocks, we can comprehensively assess the model's accuracy across various industries. Through this experiment, we aim to determine how accurately the ARIMA model can predict stock prices across these three sectors.

We collected the datasets from Yahoo Finance. For each of the three stocks, we selected data from the period spanning 03/01/2005 to 28/12/2023. During the experiment, 90% of the data was used for training the model, while 10% was used for testing and prediction.

4.2 IT sector

4.2.1 Dataset

For analyzing a dataset of the IT sector, we chose TCS, one of the oldest IT companies in India. The dataset contains the following columns: 'Date' for the corresponding date, 'Open' for the opening price of the stock on that day, 'High' for the highest price of the stock on that day, 'Low' for the lowest price of the stock on that day, 'Close' for the closing price of the stock on that day, and 'Volume' for the total trading volume of the stock on that day.

	Date	Open	High	Low	Close	Volume
0	2005-01-03	166.932495	169.587494	166.932495	169.157501	4734168.0
1	2005-01-04	168.750000	169.250000	166.645004	167.657501	4979104.0
2	2005-01-05	167.524994	167.524994	160.037506	165.625000	9604320.0
3	2005-01-06	165.000000	168.250000	161.562500	165.057495	8006888.0
4	2005-01-07	166.250000	168.125000	164.082504	167.494995	7623136.0
5	2005-01-10	168.582504	169.925003	165.250000	165.932495	7521568.0
6	2005-01-11	166.250000	166.912506	162.625000	164.220001	7986376.0
7	2005-01-12	165.744995	170.994995	160.837494	162.257507	20720032.0
8	2005-01-13	163.750000	167.337494	159.375000	161.612503	27757952.0
9	2005-01-14	163.087494	164.737503	159.625000	162.050003	15757776.0

Table 1 TCS Stock data from 2005 to 2023

4.2.2 Data Preprocessing and Visualization

Pre-processing contributes in transforming the data into an efficient input format that will be fed to the model. For this dataset at first, we check if there is any missing value is present or not, and we see there are lots of missing value present in the dataset. Deleting all rows having invalid or missing values can result in information loss, which would result in imprecise output [14]. After that, we convert the format of the 'Date' column from string to datetime to prevent future errors when iterating over dates.



Figure 1 TCS Stock price data from 2005 to 2024

4.2.3 Testing and Ensuring Stationarity

ADF test: The Augmented Dickey-Fuller (ADF) test's null hypothesis is that the time series is not stationary. Thus, if the ADF test's p-value is less than the significance level (0.05), then we will reject the null hypothesis and infer that the time series is definitely stationary. As we can observe, the p-value is more significant than the level of significance. Therefore, we can difference the series and check the plot of ADF as shown below.

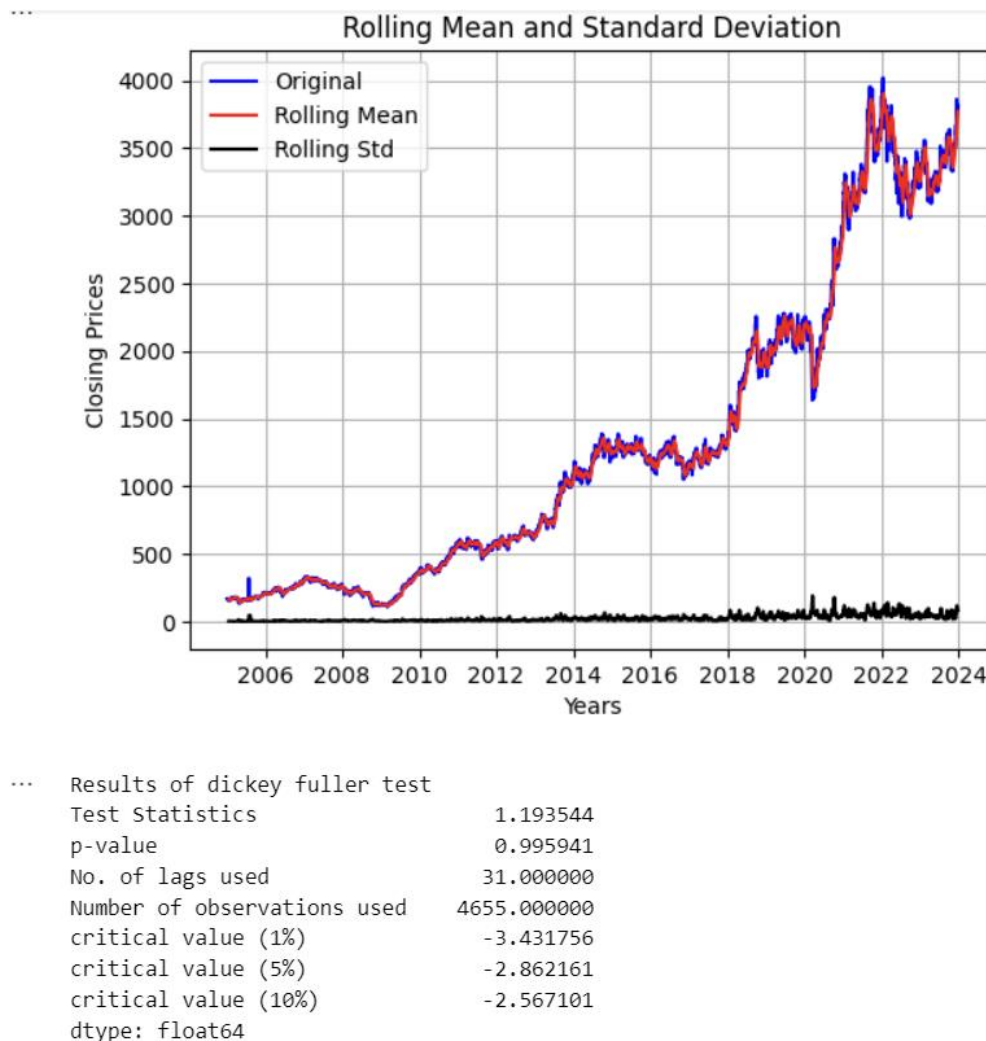


Figure 2 Augmented Dickey-Fuller test

In the above figure we perform ADF test and get some data about the time series data. Those are:

Test Statistics: This is the calculated value from the Augmented Dickey-Fuller test. A higher value indicates that the series is more likely to have a unit root.

p-value: The p-value measures the probability that the observed data would occur if the null hypothesis were true.

No. of lags used: This is the number of lagged difference terms used in the test. Lags are included to account for autocorrelation in the residuals.

Number of observations used: This indicates the number of data points used in the test after adjusting for the lags.

Critical values: In the context of the Augmented Dickey-Fuller (ADF) test, critical values are thresholds used to determine whether to reject the null hypothesis that a unit root is present in the time series. The critical values depend on the chosen significance levels, which are commonly set at 1%, 5%, and 10%.

Critical value (1%) : The test statistic value at the 1% significance level. If the test statistic is less than this critical value, we reject the null hypothesis at the 1% level.

Critical value (5%) : The test statistic value at the 5% significance level. If the test statistic is less than this critical value, we reject the null hypothesis at the 5% level.

Critical value (10%) : The test statistic value at the 10% significance level. If the test statistic is less than this critical value, we reject the null hypothesis at the 10% level.

In Fig 2 ADF test shows that p-value is higher than the significance level (0.05) & test statistics value is greater than critical values so, the series is non-stationary. To convert a non-stationary process to a stationary process, we apply the differencing method.

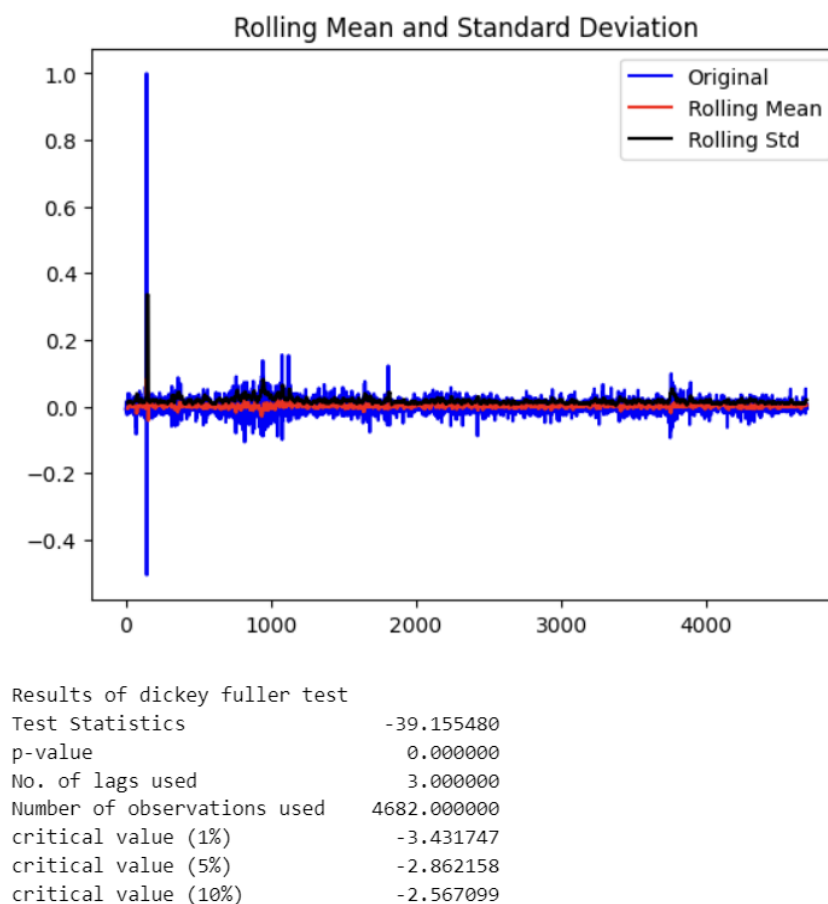


Figure 3 Augmented Dickey-Fuller test with first order difference

In Fig 3 ADF test shows that p-value is lower than the significance level (0.05) & test statistics value is less than critical values so, the series is now stationary. So, the order of d is 1.

4.2.4 Identification of p and q

Now we have to identify the order (p) of the AR model.

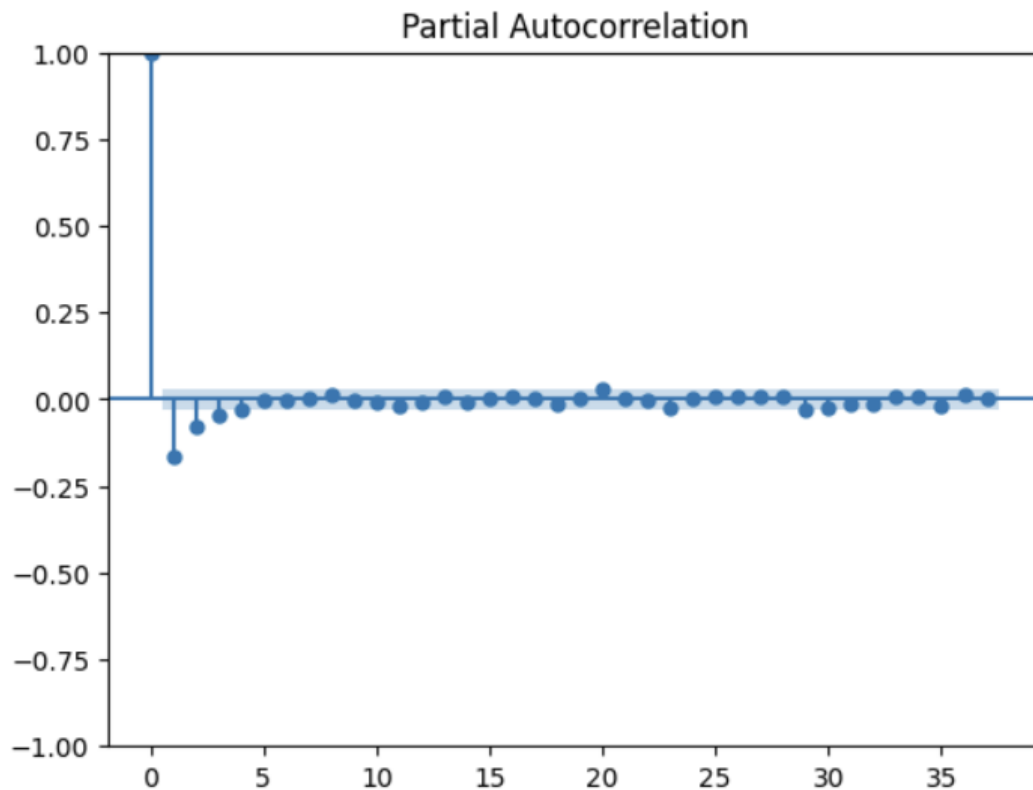


Figure 4 PACF plot

Here in Fig 4 the PACF plot shows that significant spikes at lag 1, 2, and 3 on the PACF, then we have an AR model of the order 3, i.e. AR (3).

Now we have to identify the order (q) of the MA model. For MA models, the PACF will dampen exponentially and the ACF plot will be used to identify the order of the MA process. If we have one significant spike at lag 1 on the ACF, then we have an MA model of the order 1, i.e. MA (1). If we have significant spikes at lag 1, 2, and 3 on the ACF, then we have an MA model of the order 3, i.e. MA (3).

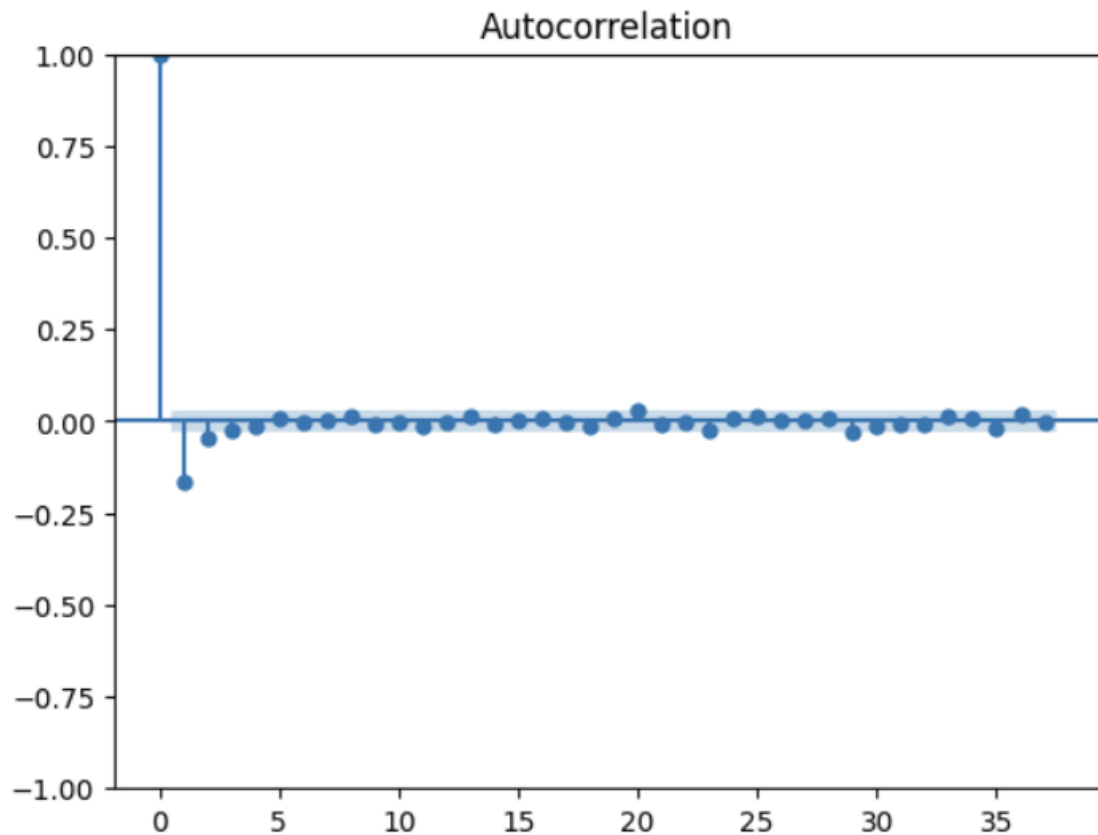


Figure 5 ACF plot

Here in Fig 5 the ACF plot shows that significant spikes at lag 1, 2 on the ACF, then we have an MA model of the order 2, i.e. MA (2). Now we have the values of p, d, q that is (3,1,2).

4.2.5 Dataset Split into Train and Test

Now we have to split the dataset into train and test. The length of the dataset is 4687. Now we split the data into 9:1 ratio. As we use 90% data as train data to train the model and 10% of the data we use to compare the predicted values with test data.



Figure 6 Train and test data split plot

We have the values of p , d , q and also split the dataset into train and test. Now we have to train the ARIMA model with the values p , d , q .

4.2.6 Model Training

First of all, we pass the train dataset and the values of p , d , q in the ARIMA model. The forecasting process iterates through the test data points. For each iteration, the model is fit to the training data and used to generate a forecast. The predicted value is printed and appended to a list of model predictions. Subsequently, the actual observed value from the test set is appended to the training data to ensure that the model is continually updated with the most recent observations, thereby refining its predictive accuracy. This iterative process enhances the model's ability to adapt to new data points, making it a robust approach for time series forecasting in dynamic environments.

```

SARIMAX Results
=====
Dep. Variable:          y      No. Observations:          4686
Model:                ARIMA(3, 1, 2)  Log Likelihood          -21715.305
Date:                Sun, 12 May 2024  AIC                    43442.610
Time:                17:07:46    BIC                    43481.323
Sample:                0      HQIC                    43456.223
                        - 4686
Covariance Type:        opg
=====
               coef      std err          z      P>|z|      [0.025      0.975]
-----
ar.L1          -0.0082      0.009      -0.941      0.347      -0.025      0.009
ar.L2          -0.9922      0.003     -382.771      0.000      -0.997     -0.987
ar.L3          -0.0196      0.008      -2.392      0.017      -0.036     -0.004
ma.L1          -0.0028      0.004      -0.756      0.450      -0.010      0.005
ma.L2           0.9859      0.004     278.078      0.000      0.979      0.993
sigma2        619.3003      5.034     123.015      0.000     609.433     629.167
=====
Ljung-Box (L1) (Q):                0.08  Jarque-Bera (JB):                25275.92
Prob(Q):                          0.77  Prob(JB):                  0.00
Heteroskedasticity (H):            22.53  Skew:                   -0.22
Prob(H) (two-sided):              0.00  Kurtosis:                14.37
=====

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

```

Figure 7 model summary

4.2.7 Summary

Model Parameters: The summary will list the ARIMA model parameters (p, d, q), which represent the number of lag observations (p), the degree of differencing (d), and the size of the moving average window (q). These parameters define the structure of the ARIMA model.

Coefficient Significance: The coefficients for the AR, I, and MA terms, along with their standard errors and p-values, are provided. Significant p-values (typically < 0.05) indicate that the corresponding term is a meaningful contributor to the model.

Model Fit Statistics: Key statistics such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are displayed. Lower values of AIC and BIC suggest a better-fitting model.

Residuals Analysis: The summary often includes diagnostics for the residuals, such as autocorrelation plots and the Ljung-Box test. Ideally, residuals should resemble white noise, indicating that the model has adequately captured the patterns in the data. The ideal Ljung-Box test result is greater than 0.05, indicating that the model is suitable. Here the value is 0.08 so, the model is suitable.

4.3 Banking Sector

4.3.1 Dataset

For analyzing a dataset of the Banking sector, we chose HDFC, one of the oldest Banks in India. The dataset contains the following columns: 'Date' for the corresponding date, 'Open' for the opening price of the stock on that day, 'High' for the highest price of the stock on that day, 'Low' for the lowest price of the stock on that day, 'Close' for the closing price of the stock on that day, and 'Volume' for the total trading volume of the stock on that day.

	Date	Open	High	Low	Close	Volume
0	2005-01-03	51.605000	52.779999	51.605000	52.525002	1295580.0
1	2005-01-04	52.505001	53.099998	50.500000	51.945000	2238060.0
2	2005-01-05	52.000000	52.000000	48.700001	50.005001	3987870.0
3	2005-01-06	50.005001	50.980000	47.900002	48.514999	4829510.0
4	2005-01-07	48.799999	49.375000	48.450001	49.025002	2649000.0
5	2005-01-10	50.000000	50.799999	48.525002	48.974998	6612870.0
6	2005-01-11	49.000000	49.174999	48.000000	48.340000	2391110.0
7	2005-01-12	48.340000	48.689999	47.509998	48.095001	1418790.0
8	2005-01-13	48.000000	50.799999	48.000000	49.889999	5395920.0
9	2005-01-14	49.900002	52.200001	49.224998	51.494999	7951920.0

Table 2 HDFC Stock data from 2005 to 2023

4.3.2 Data Preprocessing and Visualization

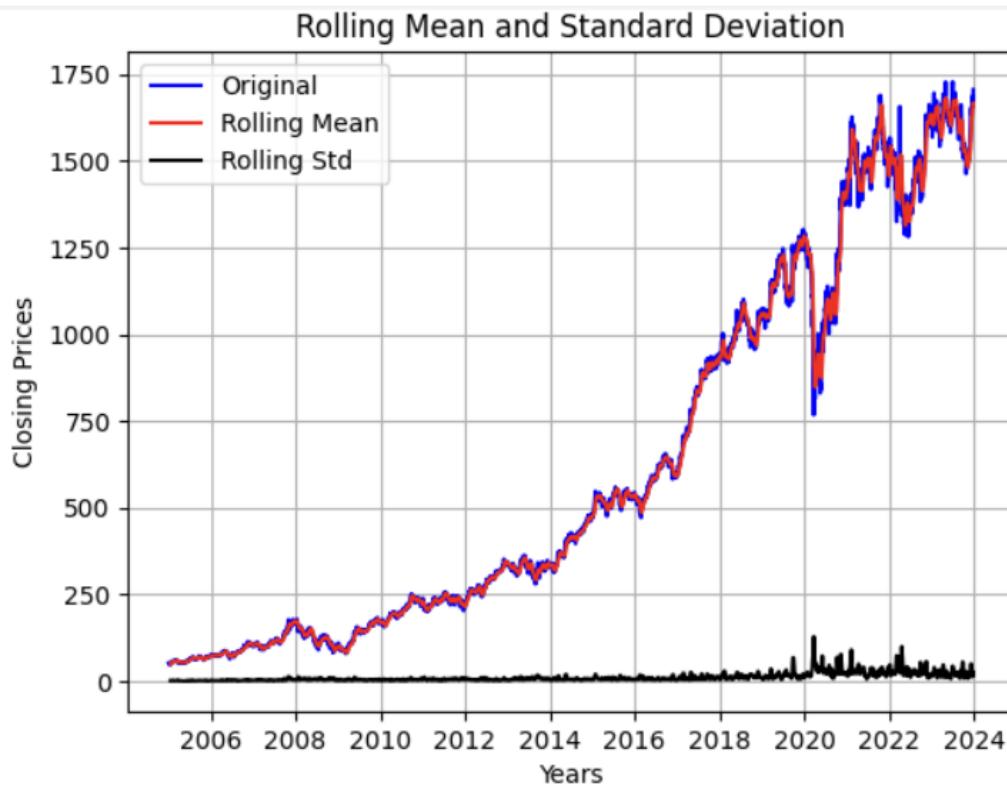
Pre-processing contributes in transforming the data into an efficient input format that will be fed to the model. For this dataset at first, we check if there is any missing value is present or not, and we see there are lots of missing value present in the dataset. Deleting all rows having invalid or missing values can result in information loss, which would result in imprecise output [14]. After that, we convert the format of the 'Date' column from string to datetime to prevent future errors when iterating over dates.



Figure 8 HDFC Stock price data from 2005 to 2024

4.3.3 Testing and Ensuring Stationarity

ADF test: The Augmented Dickey-Fuller (ADF) test's null hypothesis is that the time series is not stationary. Thus, if the ADF test's p-value is less than the significance level (0.05) or the test statistics value is less than critical values, then we will reject the null hypothesis and infer that the time series is definitely stationary. As we can observe, the p-value is more significant than the level of significance. Therefore, we can difference the series and check the plot of autocorrelation as shown below.



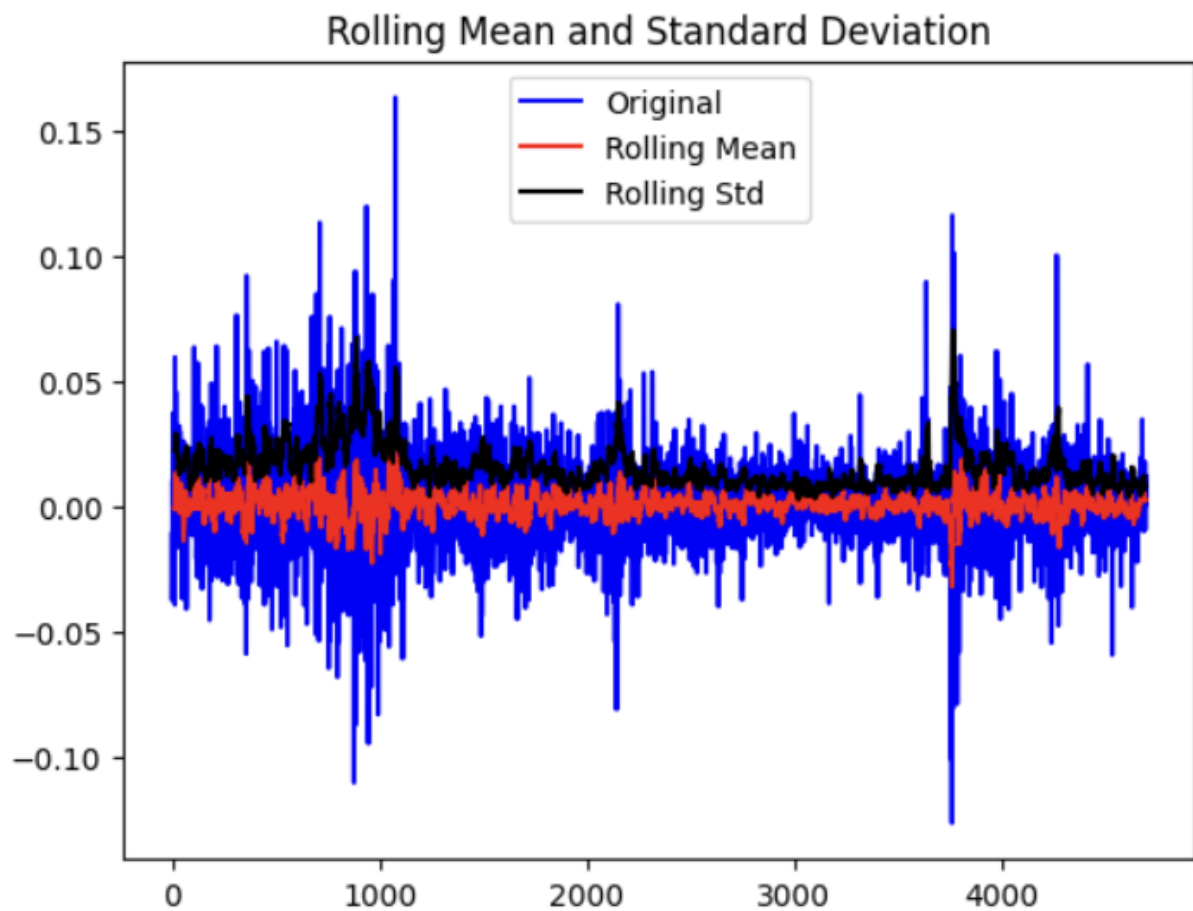
```

... Results of dickey fuller test
Test Statistics          0.460552
p-value                 0.983623
No. of lags used        31.000000
Number of observations used 4655.000000
critical value (1%)     -3.431756
critical value (5%)     -2.862161
critical value (10%)    -2.567101
dtype: float64

```

Figure 9 Augmented Dickey-Fuller test

In Fig 9, the ADF test shows that the p-value is higher than the significance level (0.05) and the test statistic value is greater than the critical values, so the series is non-stationary. To convert a non-stationary process to a stationary process, we apply the differencing method. We subtract the current day's stock closing price from the next day's stock closing price until the end of the dataset and store this in a list called 'Returns'. After that, we perform another ADF test with the 'Returns' list and check the values.



```
Results of dickey fuller test
Test Statistics                -30.850918
p-value                       0.000000
No. of lags used              5.000000
Number of observations used    4680.000000
critical value (1%)           -3.431748
critical value (5%)           -2.862158
critical value (10%)          -2.567099
dtype: float64
```

Figure 10 Augmented Dickey-Fuller test with first order difference

In Fig 10 ADF test shows that p-value is lower than the significance level (0.05) & test statistics value is less than critical values so, the series is now stationary. So, the order of d is 1.

4.3.4 Identification of p and q

Now we have to identify the order (p) of the AR model.

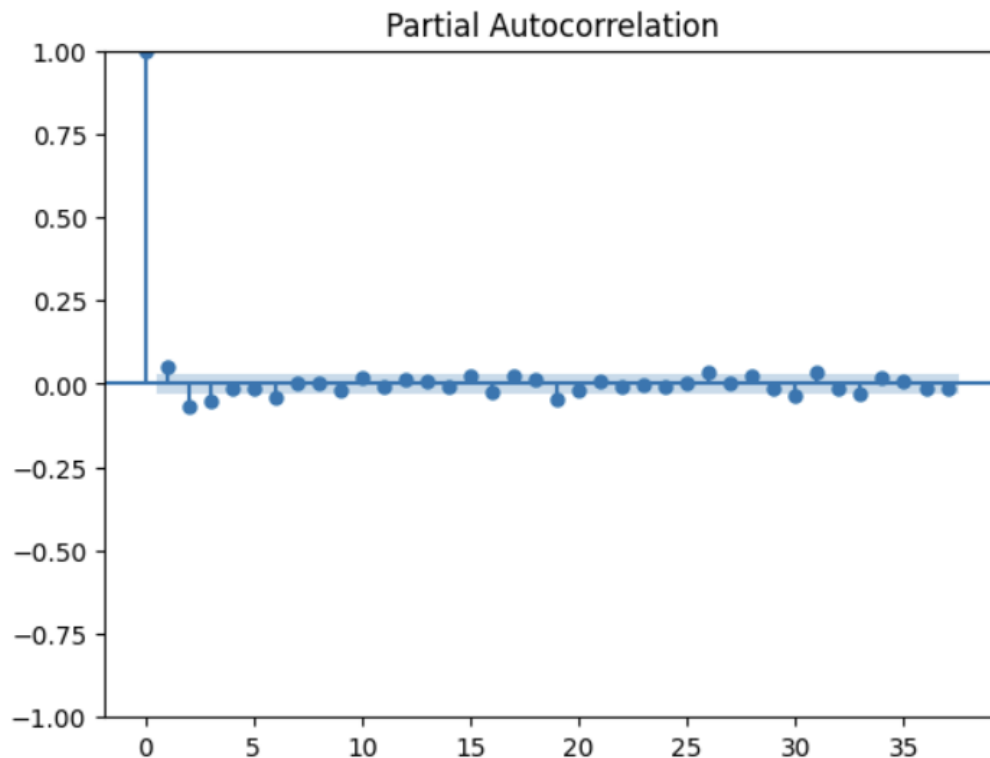


Figure 11 PACF plot

Here in Fig 11 the PACF plot shows that significant spikes at lag 2, and 3 on the PACF, then we have an AR model of the order 2, i.e. AR (2).

Now we have to identify the order (q) of the MA model. For MA models, the PACF will dampen exponentially and the ACF plot will be used to identify the order of the MA process. If we have one significant spike at lag 1 on the ACF, then we have an MA model of the order 1, i.e. MA (1). If we have significant spikes at lag 1, 2, and 3 on the ACF, then we have an MA model of the order 3, i.e. MA (3).

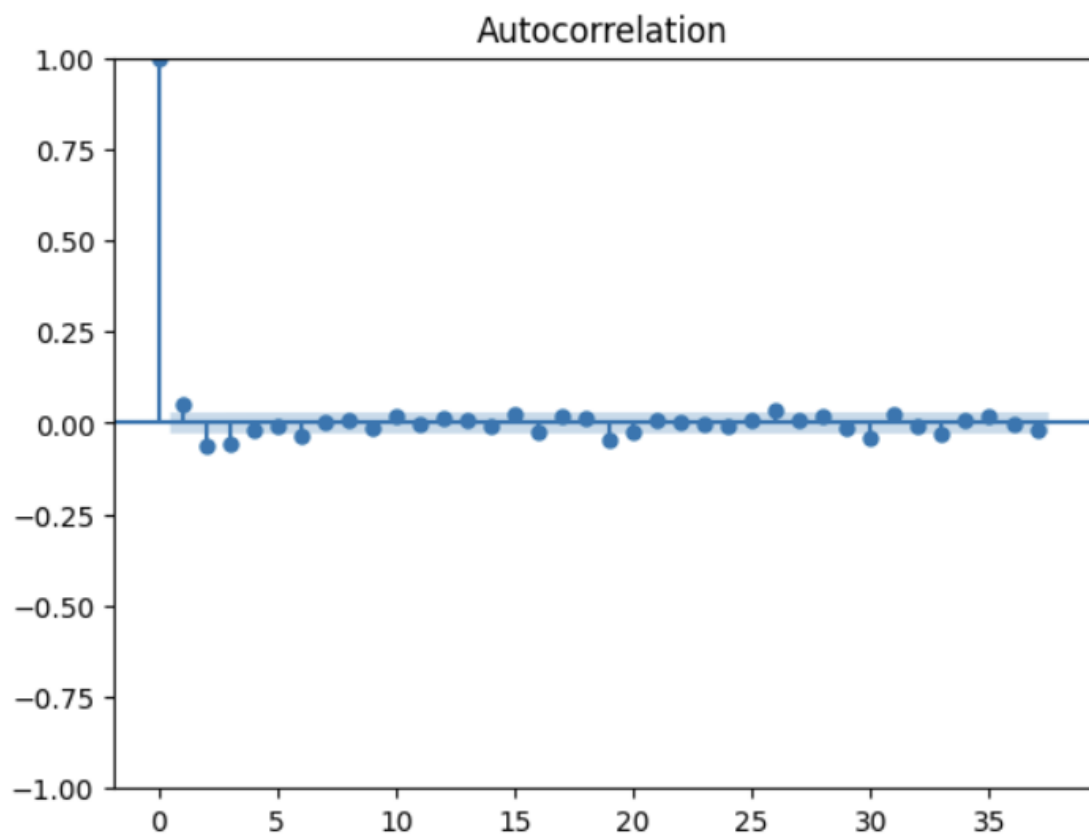


Figure 12 ACF plot

Here in Fig 12 the ACF plot shows that significant spikes at lag 2,3 on the ACF, then we have an MA model of the order 2, i.e. MA (2). Now we have the values of p, d, q that is (2,1,2).

4.3.5 Dataset Split into Train and Test

Now we have to split the dataset into train and test. The length of the dataset is 4687. Now we split the data into 9:1 ratio. As we use 90% data as train data to train the model and 10% of the data we use to compare the predicted values with test data.



Figure 13 Train and test data split plot

We have the values of p , d , q and also split the dataset into train and test. Now we have to train the ARIMA model with the values p , d , q .

4.3.6 Model Training

First of all, we pass the train dataset and the values of p , d , q in the ARIMA model. The forecasting process iterates through the test data points. For each iteration, the model is fit to the training data and used to generate a forecast. The predicted value is printed and appended to a list of model predictions. Subsequently, the actual observed value from the test set is appended to the training data to ensure that the model is continually updated with the most recent observations, thereby refining its predictive accuracy. This iterative process enhances the model's ability to adapt to new data points, making it a robust approach for time series forecasting in dynamic environments.

```

=====
SARIMAX Results
=====
Dep. Variable:          y      No. Observations:          4686
Model:                 ARIMA(2, 1, 2)  Log Likelihood          -18164.381
Date:                 Mon, 20 May 2024  AIC                   36338.761
Time:                 12:54:57    BIC                   36371.022
Sample:              0      HQIC                   36350.105
                        - 4686
Covariance Type:      opg
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
ar.L1         -0.4828      0.012     -40.942      0.000     -0.506     -0.460
ar.L2         -0.9331      0.010     -90.874      0.000     -0.953     -0.913
ma.L1          0.5168      0.013      40.810      0.000      0.492      0.542
ma.L2          0.9200      0.011      82.385      0.000      0.898      0.942
sigma2       136.5230      0.910     150.018      0.000     134.739     138.307
=====
Ljung-Box (L1) (Q):          0.53  Jarque-Bera (JB):          74582.94
Prob(Q):                   0.47  Prob(JB):              0.00
Heteroskedasticity (H):      42.01  Skew:              0.11
Prob(H) (two-sided):         0.00  Kurtosis:          22.55
=====

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

```

Figure 14 model summary

4.3.7 Summary

Model Parameters: The summary will list the ARIMA model parameters (p, d, q), which represent the number of lag observations (p), the degree of differencing (d), and the size of the moving average window (q). These parameters define the structure of the ARIMA model.

Coefficient Significance: The coefficients for the AR, I, and MA terms, along with their standard errors and p-values, are provided. Significant p-values (typically < 0.05) indicate that the corresponding term is a meaningful contributor to the model.

Model Fit Statistics: Key statistics such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are displayed. Lower values of AIC and BIC suggest a better-fitting model.

Residuals Analysis: The summary often includes diagnostics for the residuals, such as autocorrelation plots and the Ljung-Box test. Ideally, residuals should resemble white noise, indicating that the model has adequately captured the patterns in the data. The ideal Ljung-Box test result is greater than 0.05, indicating that the model is suitable. Here the value is 0.53 so, the model is suitable.

4.4 Health Care Sector

4.4.1 CIPLA Dataset

For analyzing a dataset of the Health Care sector, we chose CIPLA, one of the oldest Health Care companies in India. The dataset contains the following columns: 'Date' for the corresponding date, 'Open' for the opening price of the stock on that day, 'High' for the highest price of the stock on that day, 'Low' for the lowest price of the stock on that day, 'Close' for the closing price of the stock on that day, and 'Volume' for the total trading volume of the stock on that day.

	Date	Open	High	Low	Close	Volume
0	2005-01-03	128.000000	128.800003	126.400002	127.500000	413807.0
1	2005-01-04	128.660004	128.800003	126.199997	127.059998	474117.0
2	2005-01-05	127.000000	127.019997	114.800003	120.459999	769930.0
3	2005-01-06	122.000000	122.040001	116.000000	117.400002	1237895.0
4	2005-01-07	117.239998	118.779999	116.019997	116.639999	1145887.0
5	2005-01-10	116.000000	119.000000	111.800003	112.680000	956135.0
6	2005-01-11	112.760002	113.599998	106.019997	107.160004	1658220.0
7	2005-01-12	108.800003	109.199997	98.199997	101.019997	2455915.0
8	2005-01-13	102.800003	107.559998	102.400002	105.739998	2382132.0
9	2005-01-14	106.400002	108.739998	105.260002	106.279999	1249067.0

Table 3 CIPLA Stock data from 2005 to 2023

4.4.2 Data Preprocessing and Visualization

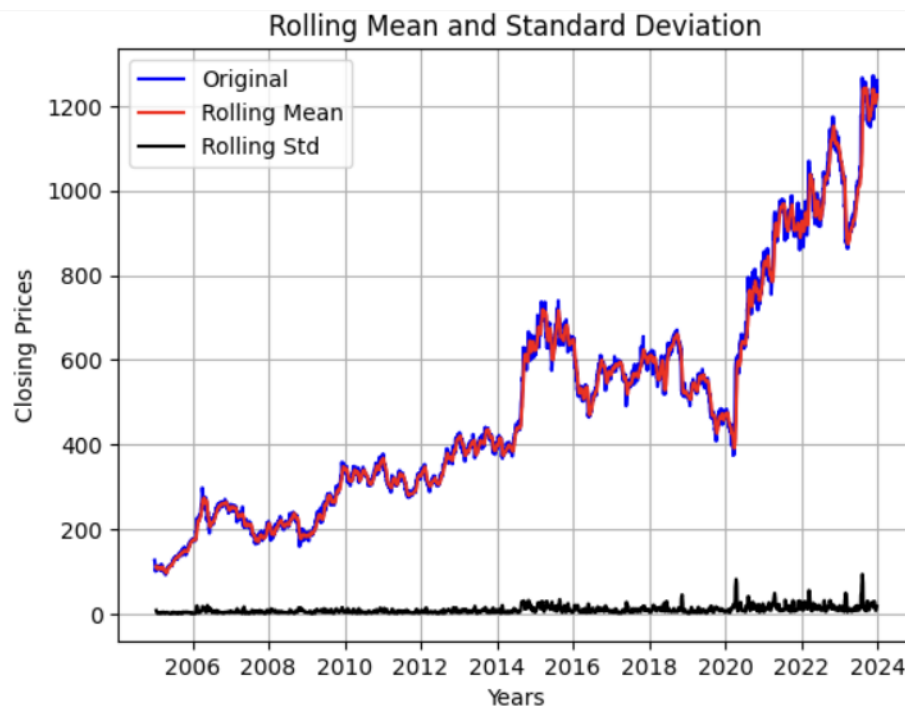
Pre-processing contributes in transforming the data into an efficient input format that will be fed to the model. The different pre-processing methods used in this research work are elaborated in this section. For this dataset at first, we check if there is any missing value is present or not, and we see there are lots of missing value present in the dataset. Deleting all rows having invalid or missing values can result in information loss, which would result in imprecise output [14]. After that, we convert the format of the 'Date' column from string to datetime format to prevent future errors when iterating over dates.



Figure 15 CIPLA Stock price data from 2005 to 2024

4.4.3 Testing and Ensuring Stationarity

ADF test: The Augmented Dickey-Fuller (ADF) test's null hypothesis is that the time series is not stationary. Thus, if the ADF test's p-value is less than the significance level (0.05), then we will reject the null hypothesis and infer that the time series is definitely stationary. As we can observe, the p-value is more significant than the level of significance. Therefore, we can difference the series and check the plot of autocorrelation as shown below.



```
Results of dickey fuller test
Test Statistics           0.333903
p-value                  0.978848
No. of lags used         1.000000
Number of observations used 4658.000000
critical value (1%)      -3.431755
critical value (5%)      -2.862161
critical value (10%)     -2.567100
dtype: float64
```

Figure 16 Augmented Dickey-Fuller test

In the above figure we perform ADF test and get some data about the time series data. Those are:

Test Statistics: This is the calculated value of the test statistic from the Dickey-Fuller test. A higher value indicates that the series is more likely to have a unit root.

p-value: The p-value measures the probability that the observed data would occur if the null hypothesis were true.

No. of lags used: This is the number of lagged difference terms used in the test. Lags are included to account for autocorrelation in the residuals.

Number of observations used: This indicates the number of data points used in the test after adjusting for the lags.

Critical values: In the context of the Augmented Dickey-Fuller (ADF) test, critical values are thresholds used to determine whether to reject the null hypothesis that a unit root is present in the time series. The critical values depend on the chosen significance levels, which are commonly set at 1%, 5%, and 10%.

Critical value (1%) : The test statistic value at the 1% significance level. If the test statistic is less than this critical value, we reject the null hypothesis at the 1% level.

Critical value (5%) : The test statistic value at the 5% significance level. If the test statistic is less than this critical value, we reject the null hypothesis at the 5% level.

Critical value (10%) : The test statistic value at the 10% significance level. If the test statistic is less than this critical value, we reject the null hypothesis at the 10% level.

In Fig 16 ADF test shows that p-value is higher than the significance level (0.05) & test statistics value is greater than critical values so, the series is non-stationary. To convert a non-stationary process to a stationary process, we apply the differencing method.

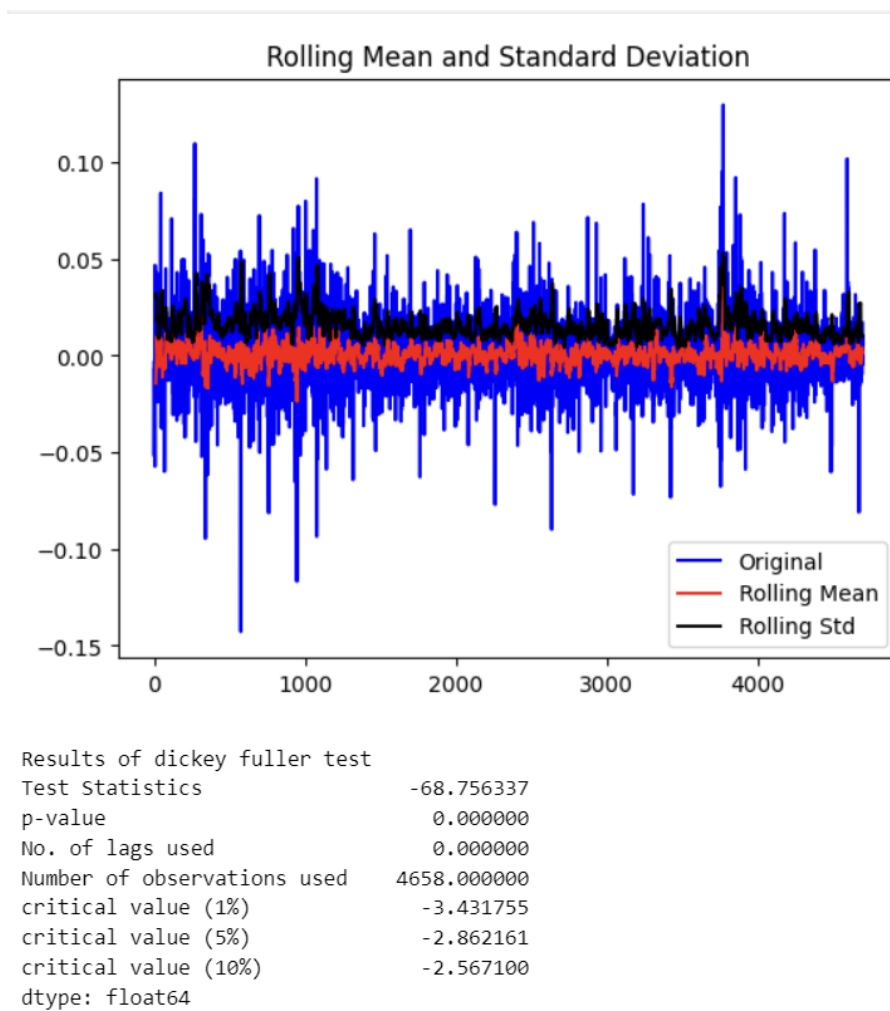


Figure 17 Augmented Dickey-Fuller test with first order difference

In Fig 17 ADF test shows that p-value is lower than the significance level (0.05) & test statistics value is less than critical values so, the series is now stationary. So, the order of d is 1.

4.4.4 Identification of p and q

Now we have to identify the order (p) of the AR model.

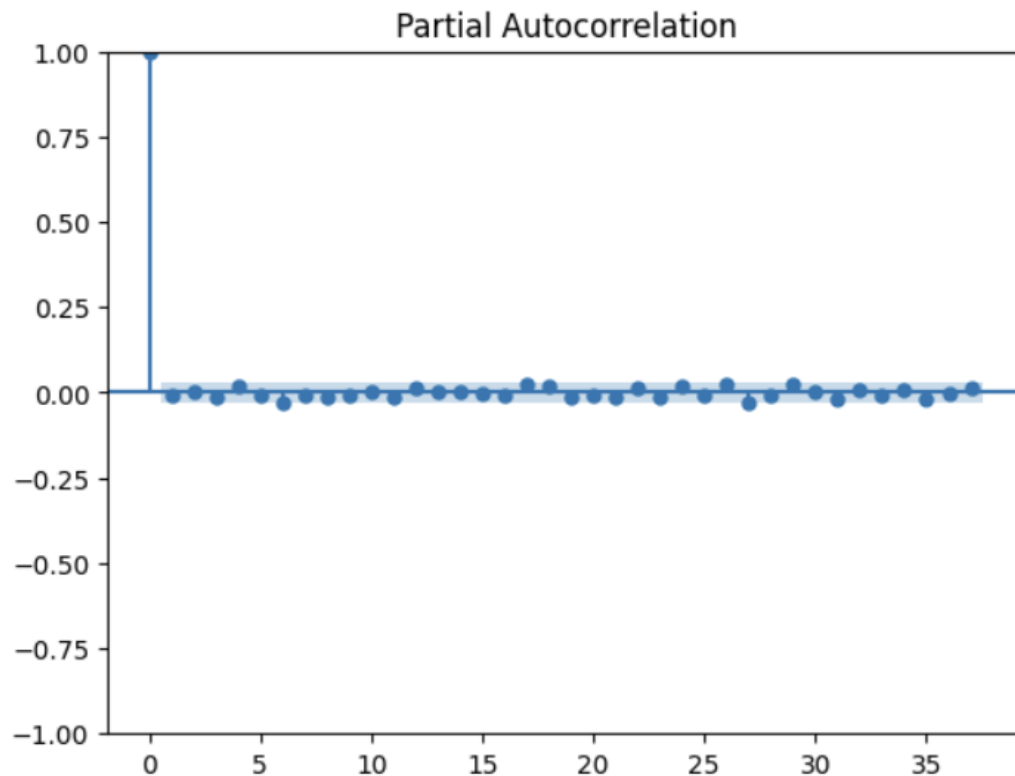


Figure 18 PACF plot

Here in Fig 18 the PACF plot shows that significant spikes at lag 1 on the PACF, then we have an AR model of the order 1, i.e. AR (1).

Now we have to identify the order (q) of the MA model. For MA models, the PACF will dampen exponentially and the ACF plot will be used to identify the order of the MA process. If we have one significant spike at lag 1 on the ACF, then we have an MA model of the order 1, i.e. MA (1). If we have significant spikes at lag 1, 2, and 3 on the ACF, then we have an MA model of the order 3, i.e. MA (3).

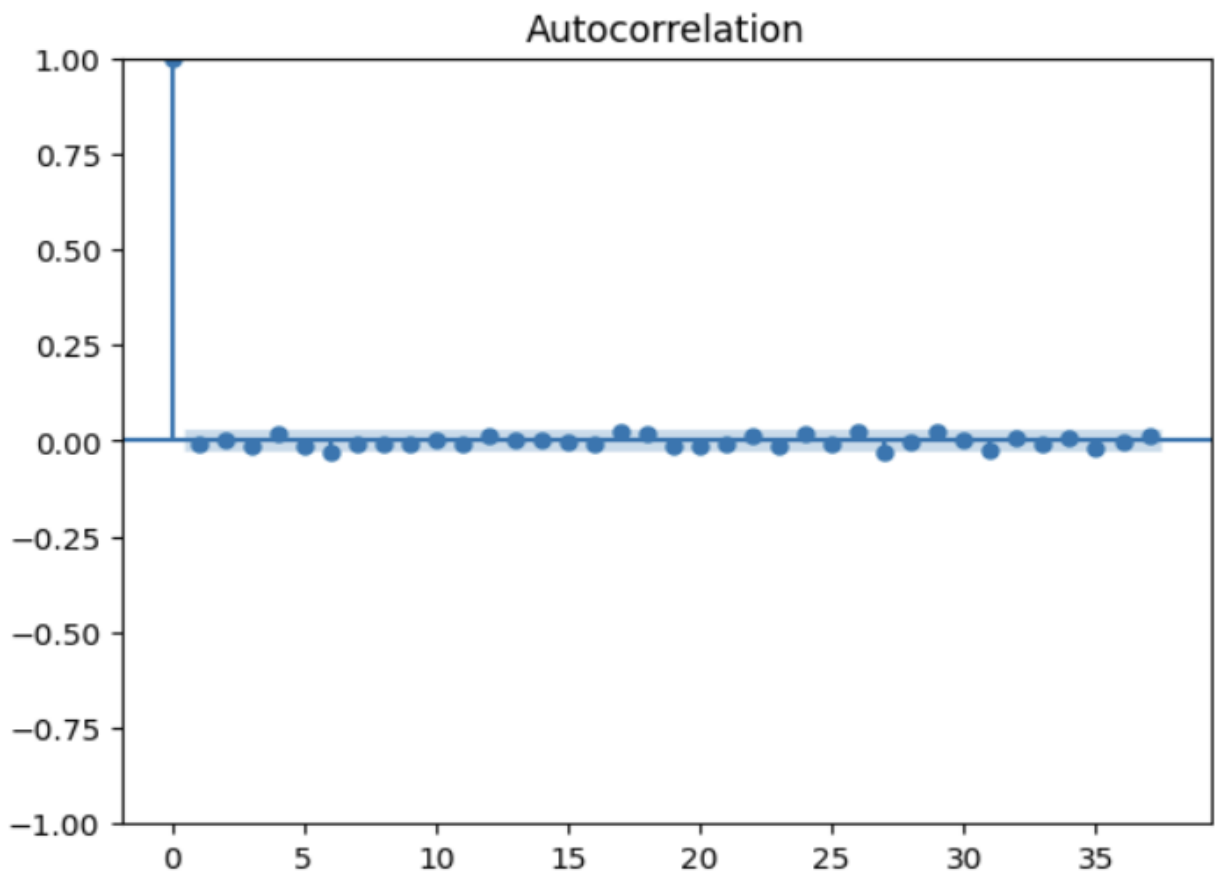


Figure 19 ACF plot

Here in Fig 19 the ACF plot shows that significant spikes at lag 1 on the ACF, then we have an MA model of the order 1, i.e. MA (1). Now we have the values of p, d, q that is (1,1,1).

4.4.5 Dataset Split into Train and Test

Now we have to split the dataset into train and test. The length of the dataset is 4660. Now we split the data into 9:1 ratio. As we use 90% data as train data to train the model and 10% of the data we use to compare the predicted values with test data.



Figure 20 Train and test data split plot

We have the values of p , d , q and also split the dataset into train and test. Now we have to train the ARIMA model with the values p , d , q .

4.4.6 Model Training

First of all, we pass the train dataset and the values of p , d , q in the ARIMA model. The forecasting process iterates through the test data points. For each iteration, the model is fit to the training data and used to generate a forecast. The predicted value is printed and appended to a list of model predictions. Subsequently, the actual observed value from the test set is appended to the training data to ensure that the model is continually updated with the most recent observations, thereby refining its predictive accuracy. This iterative process enhances the model's ability to adapt to new data points, making it a robust approach for time series forecasting in dynamic environments.

```

SARIMAX Results
=====
Dep. Variable:          y      No. Observations:          4659
Model:                 ARIMA(1, 1, 1)  Log Likelihood      -16899.257
Date:                 Mon, 20 May 2024  AIC                33804.514
Time:                 13:07:45    BIC                33823.853
Sample:                0      HQIC                33811.316
                        - 4659
Covariance Type:      opg
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
ar.L1         -0.2103      0.160      -1.318      0.187      -0.523      0.102
ma.L1          0.1539      0.161       0.958      0.338      -0.161      0.469
sigma2         82.9232      0.607     136.640      0.000      81.734      84.113
=====
Ljung-Box (L1) (Q):          0.00  Jarque-Bera (JB):          39225.50
Prob(Q):                    0.99  Prob(JB):              0.00
Heteroskedasticity (H):      7.17  Skew:              0.47
Prob(H) (two-sided):         0.00  Kurtosis:           17.18
=====

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

```

Figure 21 model summary

4.4.7 Summary

Model Parameters: The summary will list the ARIMA model parameters (p, d, q), which represent the number of lag observations (p), the degree of differencing (d), and the size of the moving average window (q). These parameters define the structure of the ARIMA model.

Coefficient Significance: The coefficients for the AR, I, and MA terms, along with their standard errors and p-values, are provided. Significant p-values (typically < 0.05) indicate that the corresponding term is a meaningful contributor to the model.

Model Fit Statistics: Key statistics such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are displayed. Lower values of AIC and BIC suggest a better-fitting model.

Residuals Analysis: The summary often includes diagnostics for the residuals, such as autocorrelation plots and the Ljung-Box test. Ideally, residuals should resemble white noise, indicating that the model has adequately captured the patterns in the data. Here the Ljung-Box test result is 0.00, it indicates autocorrelation exists in the residuals of your time series model, meaning the model may not adequately capture all patterns in the data and might be mis specified or insufficient.

4.5 MAPE

The performance result demonstrates the efficacy of the ARIMA built to forecast the stock price of three different sectors like IT, Bank, and Health Care. The performance of the model was measured using the Mean Absolute Percentage Error (MAPE) as the evaluation metric [17]. Here the general formula for MAPE is:

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|Actual\ value_i - Forecasted\ value_i|}{Actual\ value_i}$$

$N = No. of\ obserations$

4.6 Results on Data from IT Sector

In the IT sector, the closing value of the stock price of TCS was taken. By using the ARIMA model, it predicts the future value of the stock. Then, we compare the predicted values with the test dataset to evaluate the accuracy of our model. The accuracy was measured using MAPE (Mean Absolute Percentage Error).

In the figure below, we only showcase the test data section, where the red line represents the actual price of TCS and the blue dotted line represents the predicted price of the stock. By using this figure, we can visually observe how accurately the model predicts the price and where it makes mistakes.

The table below shows the actual values, predicted values, and the differences between the actual and predicted prices for 10 days.

Date	Actual Price	Predicted Price
07-02-2022	3779	3819.63
08-02-2022	3743.45	3816.40
09-02-2022	3760.55	3786.03
10-02-2022	3770.35	3748.94
11-02-2022	3694.95	3759.39
14-02-2022	3733.75	3765.00
15-02-2022	3817.8	3692.94
16-02-2022	3813.1	3736.58
17-02-2022	3784.2	3821.38
18-02-2022	3793.9	3811.50

Table 4 TCS Actual price and predicted price

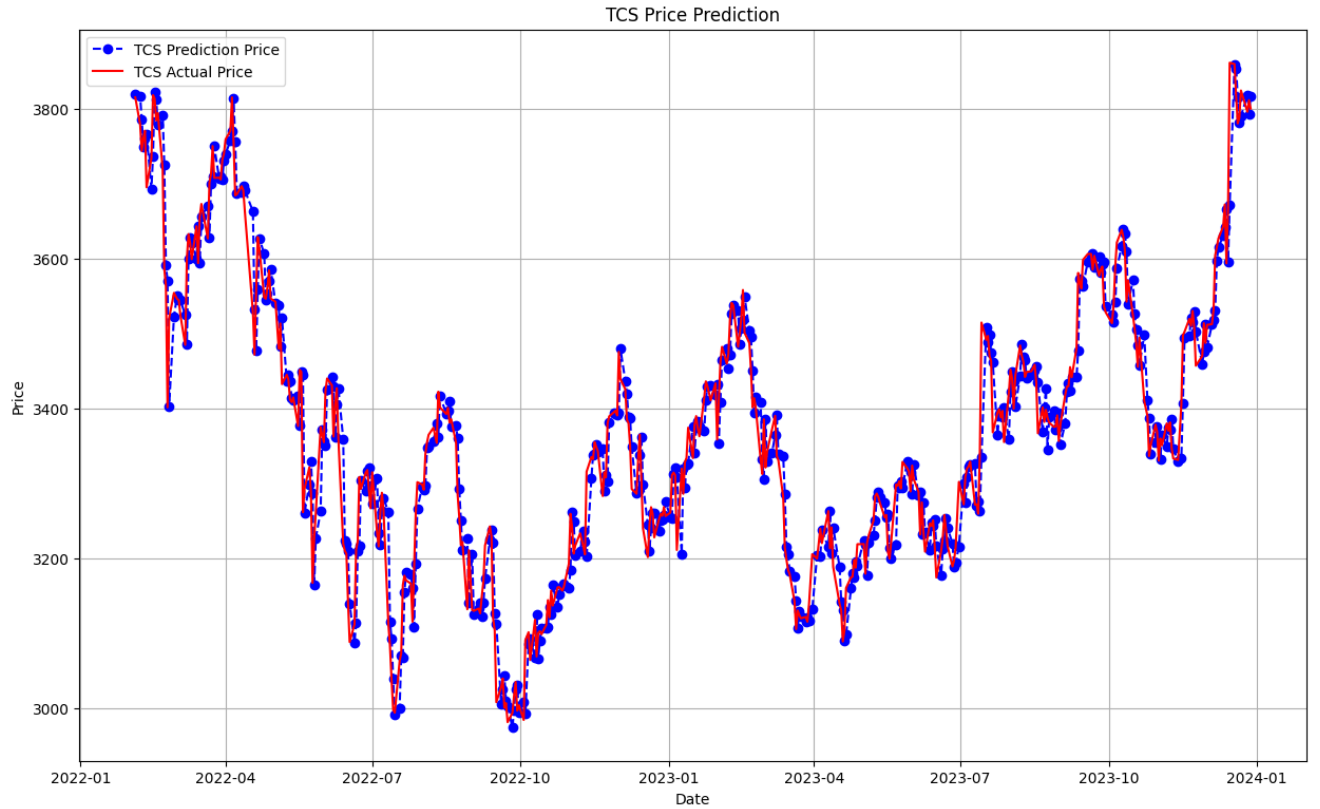


Figure 22 TCS stock actual price and predicted price

Now we have to calculate the accuracy of the model. To Calculate the accuracy, we had used MAPE.

The formula for Mean Absolute Percentage Error (MAPE) is[15]:

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{V(t) - P(t)}{V(t)} * 100 \right|$$

Where:

V(t) = Actual value

P(t) = Forecast value

n = Number of observations

The MAPE of the model is : 6.3%

Now we can say the model accuracy for the IT sector is : 93.68%

4.7 Results on Data from Banking Sector

In the Banking sector, the closing value of the stock price of HDFC was taken. By using the ARIMA model, it predicts the future value of the stock. Then, we compare the predicted values with the test dataset to evaluate the accuracy of our model. The accuracy was measured using MAPE (Mean Absolute Percentage Error).

The table below shows the actual values, predicted values, and the differences between the actual and predicted prices for 10 days.

Date	Actual Price	Predicted Price
07-02-2022	1468.15	1514.28
08-02-2022	1461.84	1523.40
09-02-2022	1497.59	1468.43
10-02-2022	1525.09	1464.03
11-02-2022	1518.84	1500.81
14-02-2022	1473.69	1522.44
15-02-2022	1517.80	1514.43
16-02-2022	1515.75	1473.15
17-02-2022	1506.5	1518.55
18-02-2022	1512.34	1516.78

Table 5 HDFC Actual price and predicted price

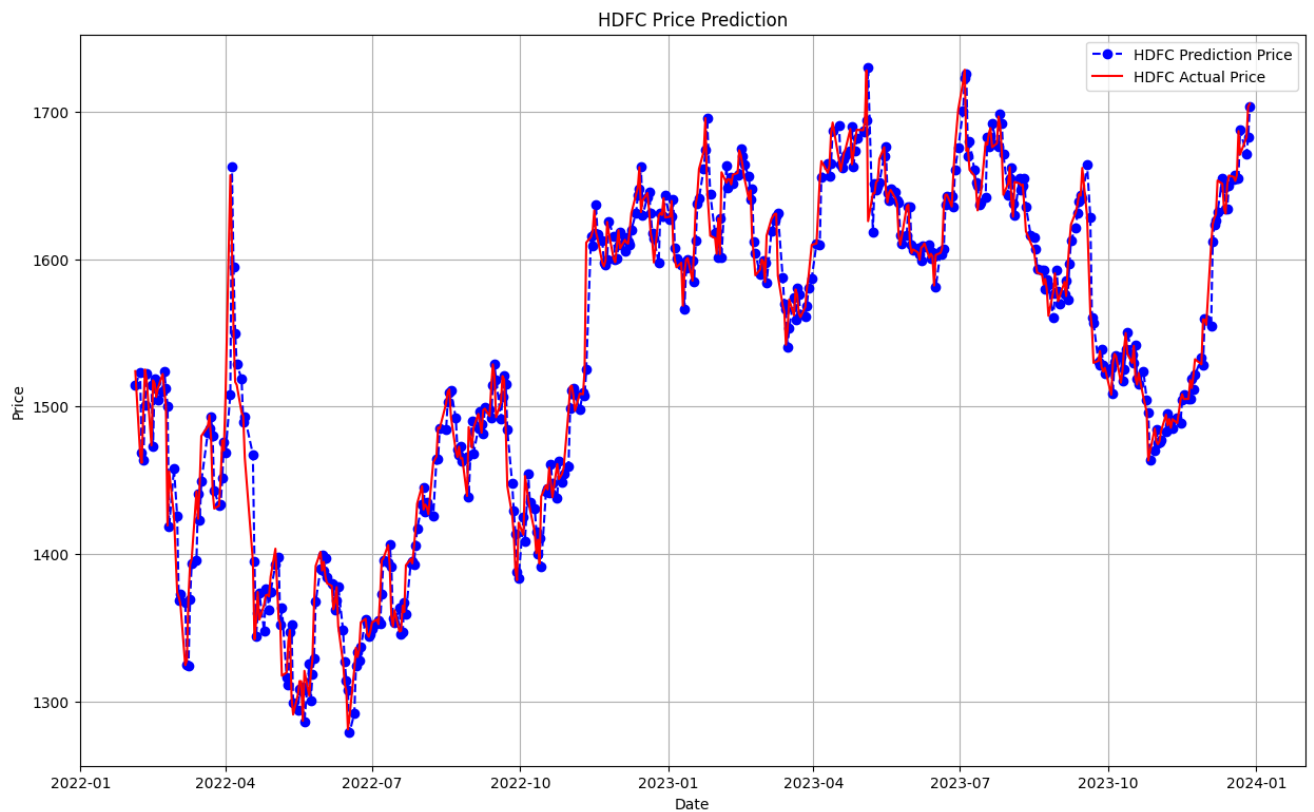


Figure 23 HDFC stock actual price and predicted price

Now we have to calculate the accuracy of the model. To Calculate the accuracy, we had used MAPE.

The MAPE of the model is : 8.3%

Now we can say the model accuracy for the IT sector is : 91.66%

4.8 Results on Data from Health Care Sector

In the Health Care sector, the closing value of the stock price of CIPLA was taken. By using the ARIMA model, it predicts the future value of the stock. Then, we compare the predicted values with the test dataset to evaluate the accuracy of our model. The accuracy was measured using MAPE (Mean Absolute Percentage Error).

In the figure below, we only showcase the test data section, where the red line represents the actual price of CIPLA and the blue dotted line represents the predicted price of the stock. By using this figure, we can visually observe how accurately the model predicts the price and where it makes mistakes. The table below shows the actual values, predicted values, and the differences between the actual and predicted prices for 10 days.

Date	Actual Price	Predicted Price
17-01-2022	910.5	923.29
18-01-2022	902.25	921.33
19-01-2022	895.45	910.83
20-01-2022	882.7	902.62
21-01-2022	866.9	895.78
24-01-2022	891.7	883.21
25-01-2022	904.1	867.54
27-01-2022	927.85	891.11
28-01-2022	932.15	903.51
31-01-2022	944.75	926.94

Table 6 CIPLA Actual price and predicted price

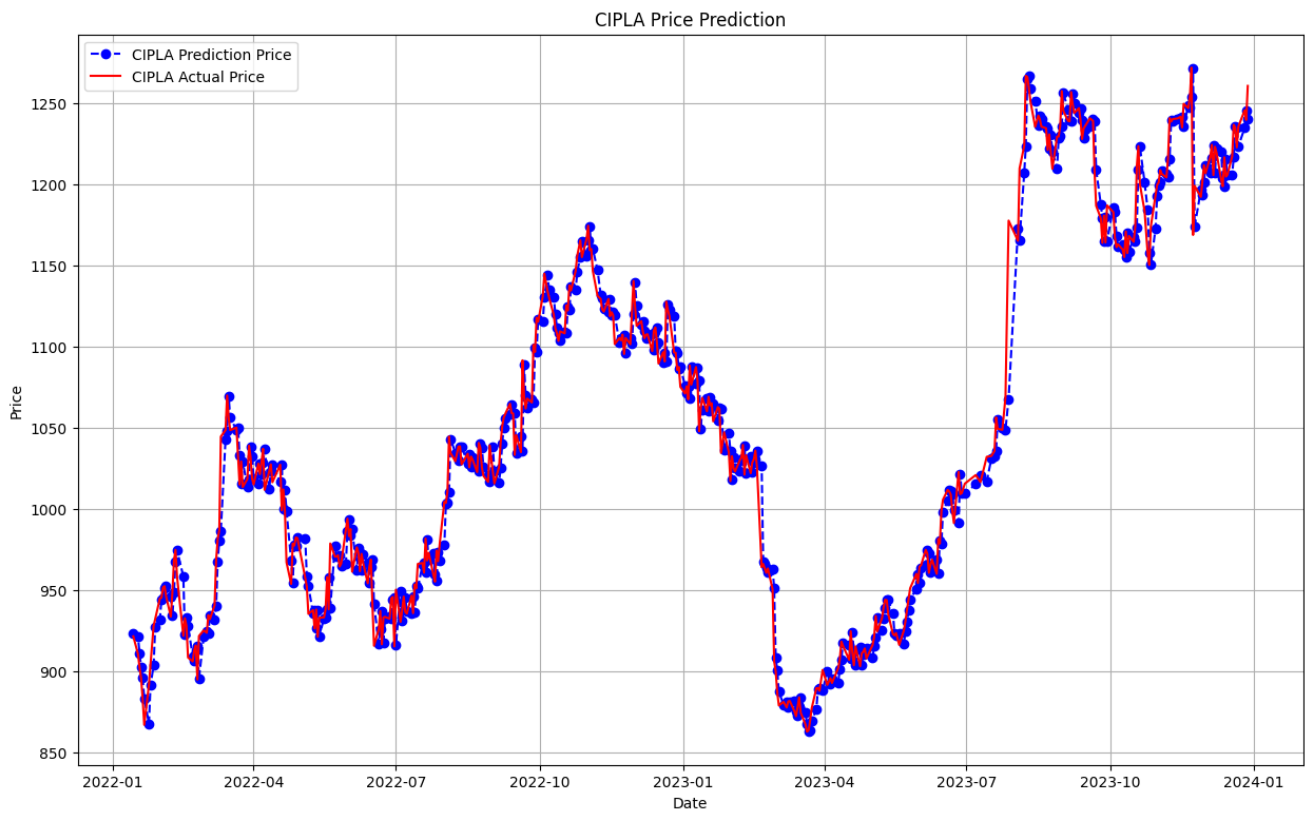


Figure 24 CIPLA stock actual price and predicted price

Now we have to calculate the accuracy of the model. To Calculate the accuracy, we had used MAPE.

The MAPE of the model is : 12.15%

Now we can say the model accuracy for the IT sector is : 87.84%

Conclusion

In conclusion, the ARIMA model proves to be a robust tool for predicting stock prices across various sectors, including IT, Banking, and Healthcare. Our analysis demonstrates that ARIMA effectively captures the underlying patterns in time series data, making it suitable for short-term forecasting. In the IT sector, characterized by rapid innovation and high volatility, ARIMA's adaptability allows for accurate short-term predictions, essential for timely decision-making.

Similarly, in the Banking sector, ARIMA's ability to model seasonality and trends helps in forecasting stock prices influenced by economic cycles and policy changes. For the Healthcare sector, where stock prices are often affected by regulatory developments and breakthroughs in medical research, ARIMA's strength lies in its capacity to incorporate historical data trends to predict future movements.

The accuracy of stock price prediction within the healthcare sector was found to be lower compared to the IT and banking sectors. This discrepancy suggests that the ARIMA model may encounter challenges when applied to healthcare stocks, potentially due to regulatory changes, clinical trial results, healthcare policies, public health crises, drug approvals, and medical breakthroughs, which may not be adequately captured by historical data alone which introduce additional unpredictability compared to the more stable and regulated IT and banking sectors.

The ARIMA model is helpful, but it has some problems. It mainly looks at past data, so it might not be very good at predicting what will happen in crazy or surprising market situations. This can happen during sudden big changes in the market, like financial crises or global sickness outbreaks. In these cases, the ARIMA model might make wrong guesses because it relies too much on old information. Also, it doesn't work well with data that keeps changing a lot and doesn't stay the same over time.

Overall, ARIMA remains a valuable method for investors and analysts seeking to make informed decisions based on historical stock price patterns across these critical sectors. To enhance its predictive power, combining ARIMA with other models or incorporating additional variables, such as macroeconomic indicators or sentiment analysis from news and social media, can provide a more comprehensive forecasting framework. Continuous evaluation and adaptation of the model parameters are essential to maintain its relevance and accuracy in dynamic market conditions.

References

- [1] Jonathan L. Ticknor (2013). “A Bayesian regularized artificial neural network for stock market forecasting”, “Expert Systems with Applications”, Volume 40, Issue 14, 15 October 2013, Pages 5501- 5506.
- [2] Chi-Yuan Yeh, Chi-Wei Huang, Shie-Jue Lee (2011). “A multiple-kernel support vector regression approach for stock market price forecasting”, “Expert Systems with Applications”, Volume 38, Issue 3, March 2011, Pages 2177-2186.
- [3] Melek Acar Boyacioglu (2010), “An Adaptive Network-Based Fuzzy Inference System (ANFIS) for the prediction of stock market return: The case of the Istanbul Stock Exchange”, “Expert Systems with Applications”, Volume 37, Issue 12, December 2010, Pages 7908-7912.
- [4] Ricardo de A. Araújo (2013), “A Morphological-Rank-Linear evolutionary method for stock market prediction”, “Information Sciences”, Volume 237, 10 July 2013, Pages 3-17.
- [5] Kannan K S, Sekar P S, Sathik M M, et al. Financial stock market forecast using data mining techniques[C]//Proceedings of the International Multi conference of Engineers and computer scientists. 2010, 1.
- [6] J. Sen. and T. Datta Chaudhuri, “A predictive analysis of the Indian FMCG sector using time series decomposition-based approach”, *Journ. of Eco. Lib.*, vol. 4, no. 2, pp. 206-226, June 2017.
- [7] J. Sen and T. Datta Chaudhuri, “Understanding the sectors of Indian economy for portfolio choice”, *Int. Journ. of Bus. Forecast. and Mktg. Intel.*, vol. 2, no. 2, pp. 178-222, 2018.
- [8] J. Sen, “A robust analysis and forecasting framework for the Indian mid-cap sector using time series decomposition”, *Journ. of Ins. and Fin. Mgmt*, vol. 3, no. 4, pp. 1- 32, 2017.
- [9] K. Ullah, M. Qasim, “Google stock prices prediction using deep learning”, Proc. of the IEEE 10th Int. Conf. on Sys Engg. and Tech., Nov 9, 2020, pp. 108-113, Shah Alam, Malaysia.
- [10] M. H. Rahman, M. M. Hossain, U. Salma, M. T. F. Khan, “Revenue forecasting using Holt-Winters exponential smoothing”, *Research & Reviews: Journ. of Stats.*, vol 5, no. 3, pp.19-25, 2016.

- [11] Ananda Chatterjee, Hrisav Bhowmick, Jaydip Sen, “Stock Price Prediction Using Time Series, Econometric, Machine Learning, and Deep Learning Models”.
- [12] T. Vantuch and I. Zelinka, “Evolutionary based ARIMA models for stock price forecasting”, Proc. of Interdisci. Symp. on Compl. Sys., pp, 239-247, 2014.
- [13] Milind Paradkar & Chainika Thakar, “Forecasting Stock Prices Using ARIMA Model”
- [14] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago and O. Tabona, "A survey on missing data in machine learning," *Journal of Big data*, vol. 8, no. 140, 2021.
- [15] <https://www.javatpoint.com/arima-model-in-python>
- [16] Rob J Hyndman, George Athanasopoulos, “Forecasting: Principles and Practice”, “2nd Edition”, pp, 139-141, 2018.
- [17] Almasarweh, Mohammad & Alwadi, Sadam. “ARIMA Model in Predicting Banking Stock Market Data”, “Modern Applied Science”, 12. 309.10.5539 /mas. v12n11p309, 2018.