Name of the student: **Sunay Chhajed**

Roll No: **19816876**

Degree for which submitted: **MS**

Department: **Economic Sciences**

Thesis title: **Application of Large Language Models in Forecasting Stock Prices**

Thesis Supervisor: **Dr. Abhinava Tripathi**

Month and year of thesis submission: **May, 2024**

Abstract:

In this article, we explore the research-related use cases of Chat-GPT and Gemini platforms. This research is novel and provides a robust estimation of stock prices with ChatGPT and Gemini. More specifically, we examine Security price prediction as follows.

First, this paper examines the potential of Large Language Models (LLMs) in forecasting stock prices. Utilizing various sets of time-series data spanning over a one-year period from diverse sectors, LLMs are given instructions, and then this pre-trained LLM forecasting model is used to predict the future prices of stocks in sectors such as Sensex, Gold, Nifty 50, and Bank Nifty. A novel model is introduced to assess the out-of-sample forecasting accuracy of these LLMs. We also compare the accuracy across both platforms, that is, Chat-GPT and Gemini, with our novel method. This paper ultimately aims to provide a robust analysis of the abilities of generative AI models in forecasting financial data. The offered insights into the comparative performance of

ChatGPT and Gemini can facilitate future applications and continue to add to the ongoing discussion of LLMs in stock price prediction.

Table of Contents:

LIST OF FIGURES

LIST OF TABLES

1.1 Introduction

The field of Artificial Intelligence (AI) is growing at a fast rate, encompassing significant advancements in several industries like healthcare, robotics, manufacturing, etc [1-3]. A very recent and profound development is the use of AI in generative language models. These large language models (built on transformer models) have significant applications in the financial domains [4-5].

An important milestone in LLM development was InstructGPT, a framework that allowed for instruction fine-tuning of a pre-trained language model based on reinforcement learning from human feedback(RLHF) [6-7]. At present, we have ChatGPT as a successor of InstructGPT. ChatGPT is equipped with several advancements that lead to great performance in several NLP tasks, such as reasoning and generalised text generation. These NLP capabilities have invited several applications in diverse domains such as education, healthcare, human-machine interaction, medicine and scientific research.

Google's Gemini and Open AI's ChatGPT are at the forefront, and these multimodal LLMs can process and understand information given in text, code, image, etc. These can generate high-quality and accurate human-like text to cater to the needs of users[8-10]. We assess their effectiveness and performance in predicting stock prices. Moreover, with the advancement of GPU and the increase in computational power, large-scale deep networks with huge amounts of parameters are trained so that more information can be learned to deal with general downstream tasks with better performance [11].

The LLM models discussed above have already been studied and judged on their performance and ability to offer an edge in accuracy and depth on specific challenges in financial data interpretations [9]. We added to the existing strand by using these LLMs to predict stock prices.

The idea behind this was to understand whether this state-of-the-art generative technology can be used to predict these stock prices accurately. We have done a comparative analysis of Gemini and ChatGPT with reference to a novel forecasting model, in our case, the ARIMA model.

We trained our own ARIMA model on time-series data to predict the stock prices to the utmost accuracy possible. Eventually, we compare the values predicted by ChatGPT and Gemini, given time-series data as input, to our forecasted values of the ARIMA model. This comparative scenario will provide us with insights into the adoption of generative AI. This research highlights a thorough evaluation of ChatGPT and Gemini and encourages further exploration of their role in empowering financial data analysis and predictions.

2.1 Literature Review

Human intervention has always been an integrated part of finance related sectors[2,12-13]. However, financial data and reports that are reviewed by finance experts (Tripathi et al., 2019; Tripathi, Vipul, et al., 2020b; Tripathi & Pandey, 2021), which mostly lead to errors[14-16]. Human based traditional methods prove to be increasingly cumbersome with the growth of data[15,17-19]. These problems are solved efficiently by Generative AI, which can easily detect patterns and provide insights with the use of tools that can interpret the financial data to finally focus on making decisions that are crucial for short as well as long term goals.

However, financial forecasting (Tripathi, Dixit, et al., 2020; Tripathi et al., 2020, 2021b) is regarded as one of the most challenging tasks due to the complexities in real time stock market [20-21]. This surges a demand for reliable models for prediction to maximize profits and minimize risks [22]. This motivation drove researchers to develop hybrid approaches to enhance stock price predictions by combining the strengths of different models including the use of Artificial Neural Networks (ANNs) [20]. But language is a major part of communication and is essential for human machine interaction. Being trained on hundreds of billions of parameters, LLMs provide significant breakthroughs [23-25]. LLMs are built on transformer architecture, which, on recent advancements in computation power, becomes capable of handling human level performance in various contexts [26-28]. Additionally, a new framework called Self Correlated Reinforcement Learning with the Local-Global model (SCRL-LG) harnesses the power of LLMs for multi-modal stock return prediction [29].

OpenAI's ChatGPT is considerably renowned for its exceptional conversational prowess and its ability to produce structured and well informed text over diverse tasks [30-32]. Whereas Google's Gemini, even being new in commercial use, shows promise with its versatility and adaptability across various applications. [33-36]. The true essence of any model, however, lies in its efficiency, performance and adaptability in the dynamic financial landscape [37-41]. Efficiency plays a crucial role in indicating the model's ability to perform tasks with minimal resources, such as minimizing computational costs and processing time, especially with the growing data volumes. Emphasizing accuracy and reliability while adapting seamlessly transitioning between tasks and evolving while learning shows the performance of these LLMs. Although the primary strength of these LLMs lies in recognizing patterns and multiple nuances, context specific evaluation is important to garner their full potential.

ChatGPT is widely known for its natural conversational skills, while Google's Gemini has the upper hand in its precision and factual accuracy. ChatGPT excels in understanding context and is capable of offering explanations and creative content generation through its ability to learn and grow by prompt interactions. Google's robust knowledge base and search technology enable it to process data in image, audio or visual format [9].

In this paper, we utilize LLMs such as ChatGPT and Gemini, as well as the traditional ARIMA based models to forecast stock prices. This helps us to compare the strengths and limitations of LLMs and the ARIMA models based on short and long term stock price predictions.

Statistical models govern the basis of the ARIMA models (Tripathi, 2021; Tripathi et al., 2021a; Tripathi, Vipul, et al., 2020a; Tripathi & Dixit, 2023). In general, predictions can be made using statistical and Artificial Intelligence(AI) techniques [20]. Several time series forecasting studies using ANNs techniques lag behind the ARIMA models, mainly due to the efficiency and robustness of ARIMA models that ace short term predictions [42-44]. It has been widely used in the finance and economics related fields. There are several other related models such as the traditional regression analysis, exponential smoothing and the generalized autoregressive conditional heteroskedasticity (GARCH). ARIMA model has been used extensively in various real life applications, such as forecasting [45-50]. These are widely used in financial forecasting [48,50]. The models are capable of generating efficient short term forecasts. In [51], the superior performance of ARIMA models over complex structural models on short term stock price predictions. Building these models is deeply discussed in [52].

3.1 Data

A time series is a collection of data points that are organized by a time stamp and are often used in forecasting future instances. Examples of time series data include stock prices, churn rate, temperature measurements, etc. In this study, we have used the time series data for four key securities which are: Bank Nifty (in INR), Nifty 50 (in INR), Sensex (in INR) and Gold (in USD). These four securities are major stock indices in the Indian market, hence giving us a diverse representation of financial markets. The data for these securities is readily available and is relevant for a broader audience as it captures some critical components of the Indian financial ecosystem.

Data was sourced from Yahoo Finance, which is a widely recognized platform for obtaining historical financial data. This obtained data was stored in Excel files for pre-processing, where we used R, a widely known statistical computation language, to clean the data of any Nan values or empty cells. This was done to ensure data quality and consistency throughout the research. The data was further cleaned to remove the unwasted columns, and relevant fields like ‘Date’ and ‘Adjusted Close’ values were taken as final input.

The LLMs models used here, ChatGPT and Gemini, take data in text format and hence, clear instructions were given to ensure the LLMs read the data in a tabular structure. The processed data is then used to task the LLMs with predicting stock prices for three different time horizons: 20 days, 30 days, and 40 days. Input data spans over a period of one year and can be visualized in Figure 1 - 4.

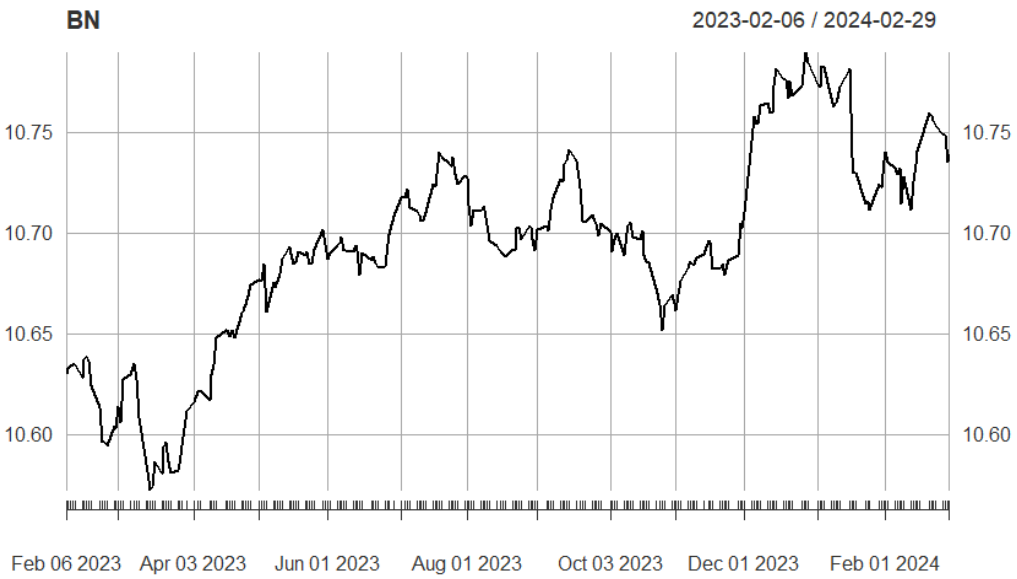

Figure 1 Represents the stock price data of Bank Nifty

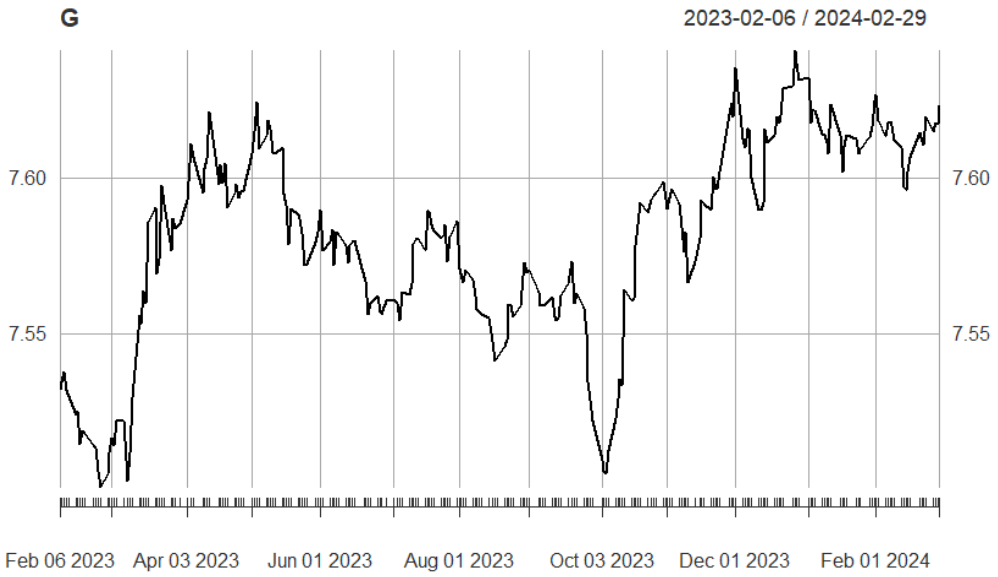

Figure 2: Represents the stock price data of Gold

Figure 3: Represents the stock price data of Nifty 50

Figure 4: Represents the stock price data of Sensex

Thus, this study's fundamental approach involves the use of consistent and pre-processed financial data to forecast stock prices.

4.1 Methodology

ARIMA Model:

The autoregressive integrated moving average model (ARIMA) is a widely used time series model because of its strong short-term predictability and simplicity, and it is extensively used in financial time series forecasting [53]. It involves the concepts of autoregression (AR), differencing (I) and moving average (MA).

The autoregressive (AR) component (p):

The AR component accounts for the relationship between the current observation and a specified number of lagged observations.(1)

The parameter p indicates the number of lagged variables to be incorporated in the model.

The AR(p) model can be expressed as:

$$Yt = \alpha 1 Yt\text{-}1 + \alpha 2 Yt\text{-}2 + \ldots \alpha p Yt\text{-}p + \varepsilon t$$

ε represents the random component (random shock) [61], α1, α2, …, in this equation are the autoregressive coefficients. and Yt is the time series at time t.

The integrated (I) component (d):

The I component involves differencing of the time series to make it stationary. Stationary time series have statistical properties that remain over time.

A first order difference d = 1 indicates that the difference between two successive values of Y is constant. An integrated process is defined by equation (2)

$$Yt = Yt\text{-}1 + \varepsilon t$$

where the random perturbation εt is a white noise.

The moving average (MA) component (q):

The MA component accounts for the relationship between the current observation and a random term based on past forecast errors.

The order q indicates the number of previous periods to be incorporated in the current value.

A moving average is defined by equation (3)

$$Yt = u + \varepsilon t + \theta 1\varepsilon t\text{-}1 + \theta 2\varepsilon t\text{-}2 + \ldots + \theta q\varepsilon t\text{-}q$$

Where u is the mean of the series, θ1, θ2, …, θq are the moving average coefficients, ε error term associated.

Many researchers acknowledge that the estimation of parameters does require a large number of observations. Hence, there are certain limitations associated with using the ARIMA model [54].

The ARIMA model is denoted as ARIMA(p, d, q).

The ARIMA model utilizes the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the sample data as fundamental tools to identify the order of the ARIMA model, as proposed by Box and Jenkins [55]. The approach is also termed the Box-Jenkins

methodology, which consists of a sequence of steps for recognising, estimating and evaluating ARIMA models with time series data [52].

When comparing different forecasting models, statistical measures known as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) play a crucial role. While fitting ARIMA models with different sets of parameters (different values of p, d and q), AIC and BIC values are computed for each model variant to identify the best fitting model by identification of goodness of fit and complexity of model or number of parameters. The model showing the lowest values for AIC and BIC values is considered the best fit for forecasting while maintaining accuracy and not making the model overly complex.

In Figure 5, The outlined flowchart shows a sequence of steps for creating an ARIMA model facilitating comprehensive time series analysis. The imported data from the Excel file undergoes logarithmic transformation to set the initial stage for a rigorous time series analysis. The original dataset was plotted to check the data's stationarity by transforming the series by differencing methods.

After ensuring the stationarity of data, the autocorrelation function (ACF) and the partial autocorrelation function (PACF) are examined for different series to identify the potential autoregressive (AR) and moving average (MA) orders of the ARIMA model. Statistical tests like Ljung-Box and Box-Pierce tests are then used to evaluate autocorrelation within the data.

This preliminary analysis is then continued with model selection based on AIC and BIC criteria. Changing the various parameters (p, d and q), different ARIMA models are constructed, and best fit model is selected with the conduction of out of sample forecasting. Multiple error metrics are calculated to utilize the performance of different ARIMA models. A thorough examination of the

data and analysis of different model selection and evaluation is carefully considered which leads to a robust forecast for time series data.

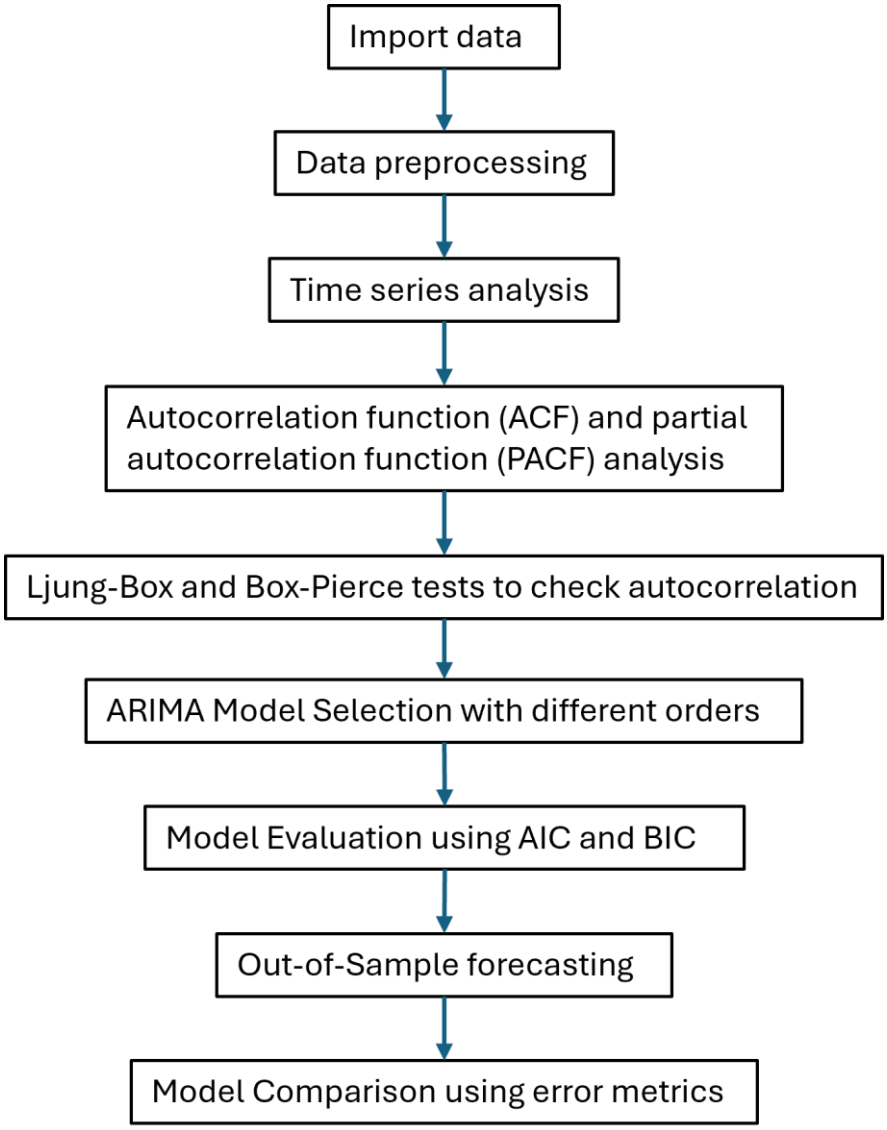

Figure 5: The flowchart of ARIMA Model methodology

An important part of our research comprised a basic set of steps to give instructions to the LLMs; this term is coined as prompt engineering. Prompt engineering is basically how, through a basic of instructions human-machine interaction is made possible where a chain of thoughts is

conveyed to the machine by entering instructions in textual format. This improves the ability of LLMs to understand and perform complex reasoning.[56] Chain-of-thought (CoT) prompting is also a recent technique incorporated while instructing LLMs is a step-by-step question -> answer -> reason -> evolve process, which successes the overall learning and then answering process of generating relevant information. [57] The prompt used by us clearly outlined a clear and specific task for both ChatGPT and Gemini by setting a scenario where the model is given the role of a proficient forecaster.

The objective of predicting the stock price values for different time horizons was clearly specified, and well-defined instructions on how to read the dataset were provided. We explicitly needed to mention not to provide us the approach or the code of forecasted data because if not, then both, ChatGPT and Gemini provided Python or R code to forecast and not the forecasted values directly. This clearly helped us to define the objective and clearly lay out the path for LLMs to follow to give us the desired outcome.

5.1 Analysis and Results:

We began by converting the input data into a time-series dataset by built-in functions of R. Our initial step aimed to check the stationarity of the series. We plotted the differencing order series to see whether our series was in the required format. To further confirm this, we conducted Box-Ljung and Box-Pierce tests for autocorrelation on both the zero-order and first-order differenced series. The p-values obtained indicated that the first-order difference series was stationary.

First-order difference series	Box-Pierce test	Box-Ljung test
Nifty 50	p-value = 0.254	p-value = 0.2512
Bank Nifty	p-value = 0.2729	p-value = 0.27
Sensex	p-value = 0.3528	p-value = 0.35
Gold	p-value = 0.3918	p-value = 0.389

Table 1: Containing p-values for Box-Pierce and Box-Ljung test

Further, we generated the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for the time series to understand the level of significant lag values.

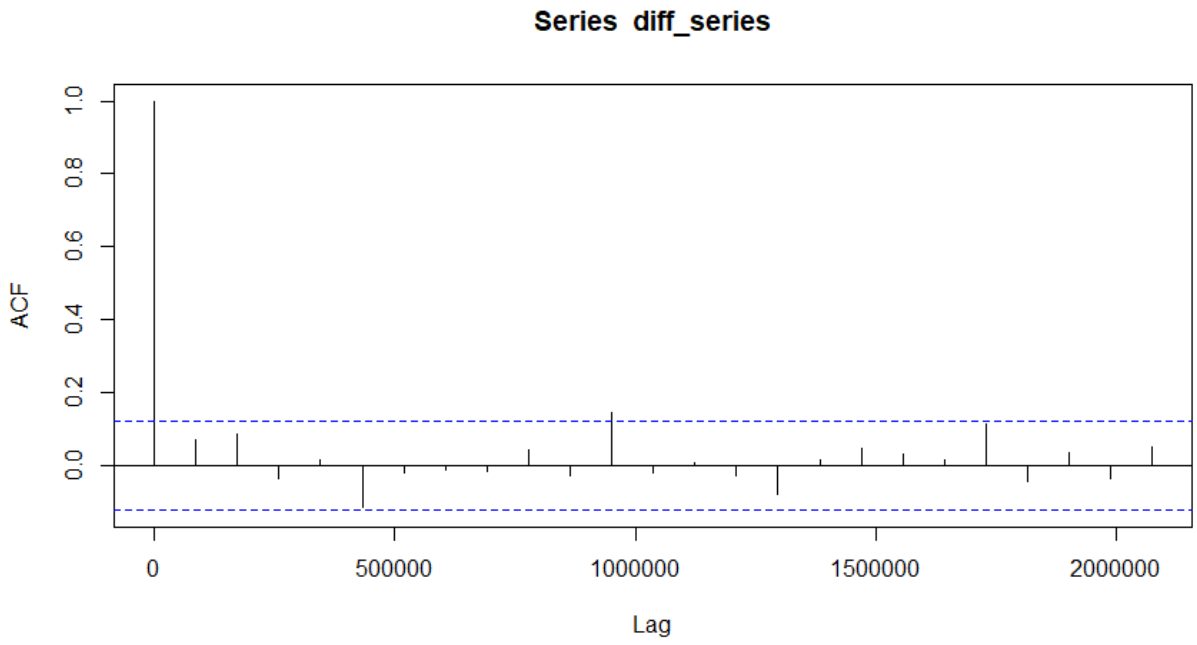

Figure 6: ACF for Nifty 50

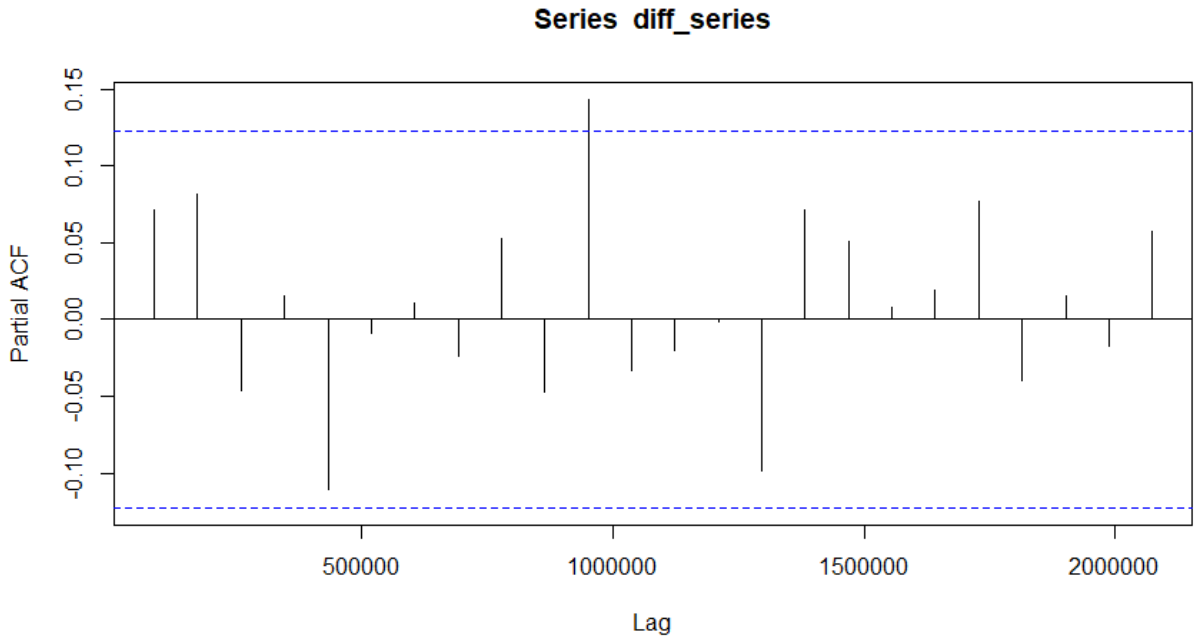

Figure 7: PACF for Nifty 50

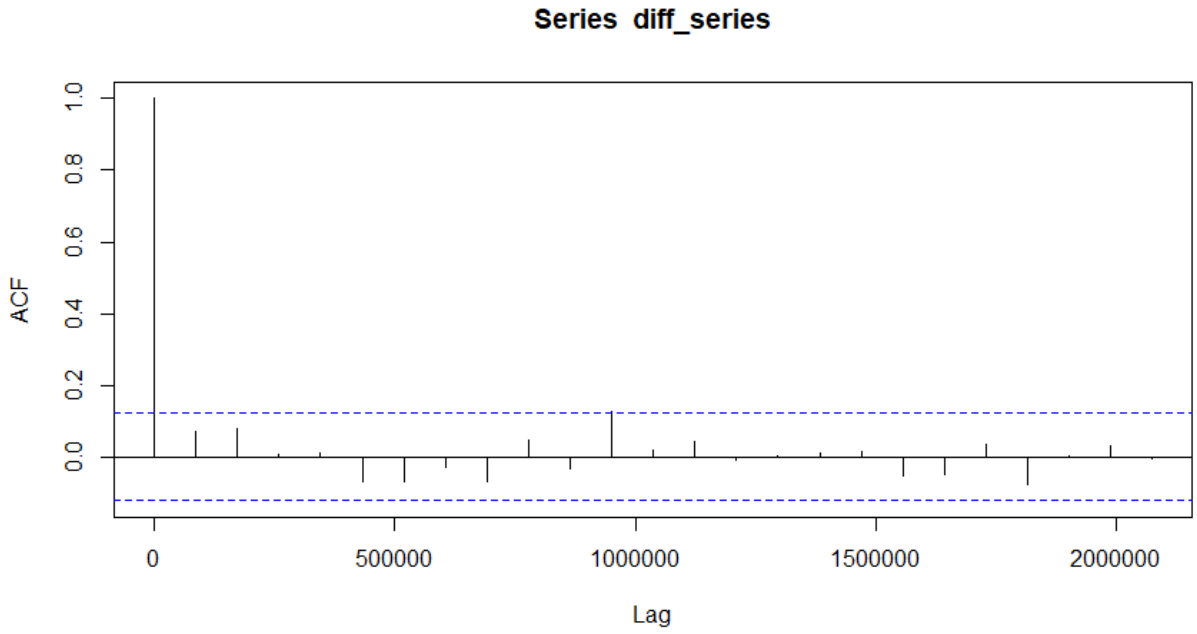

Figure 8: ACF for Bank Nifty

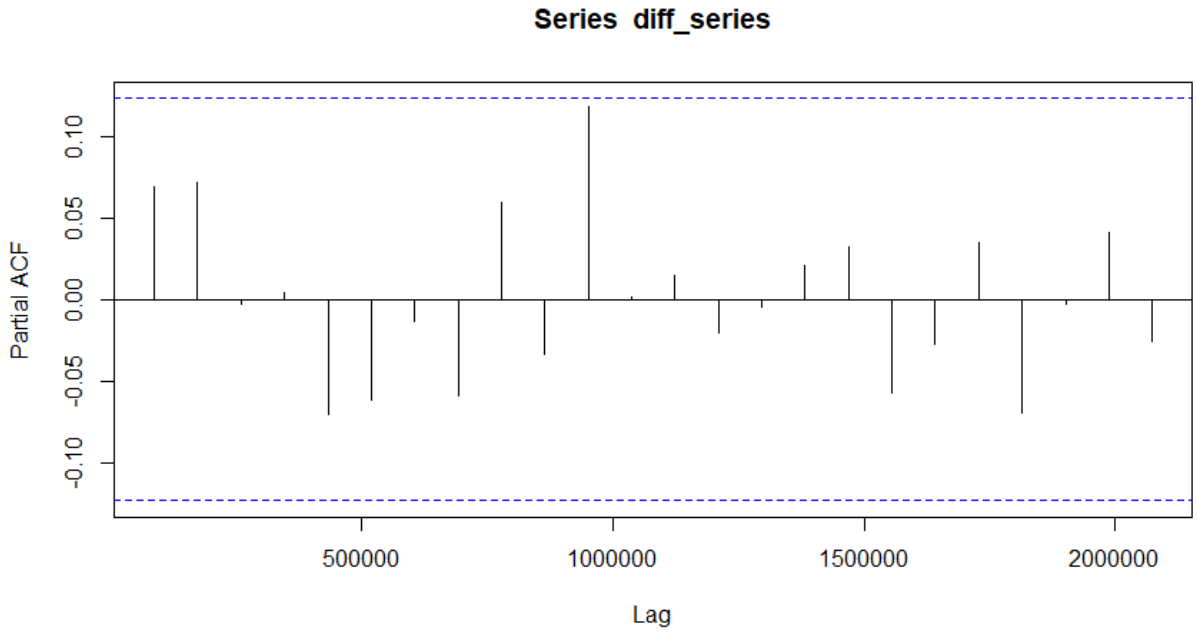

Figure 9: PACF for Bank Nifty

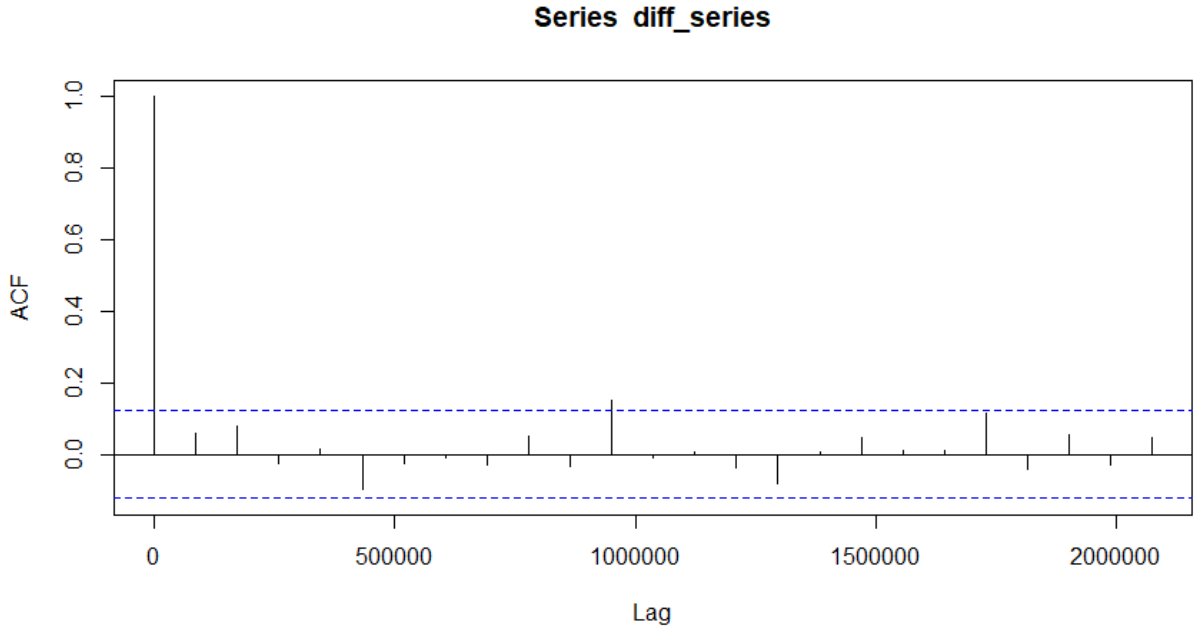

Figure 10: ACF for Sensex

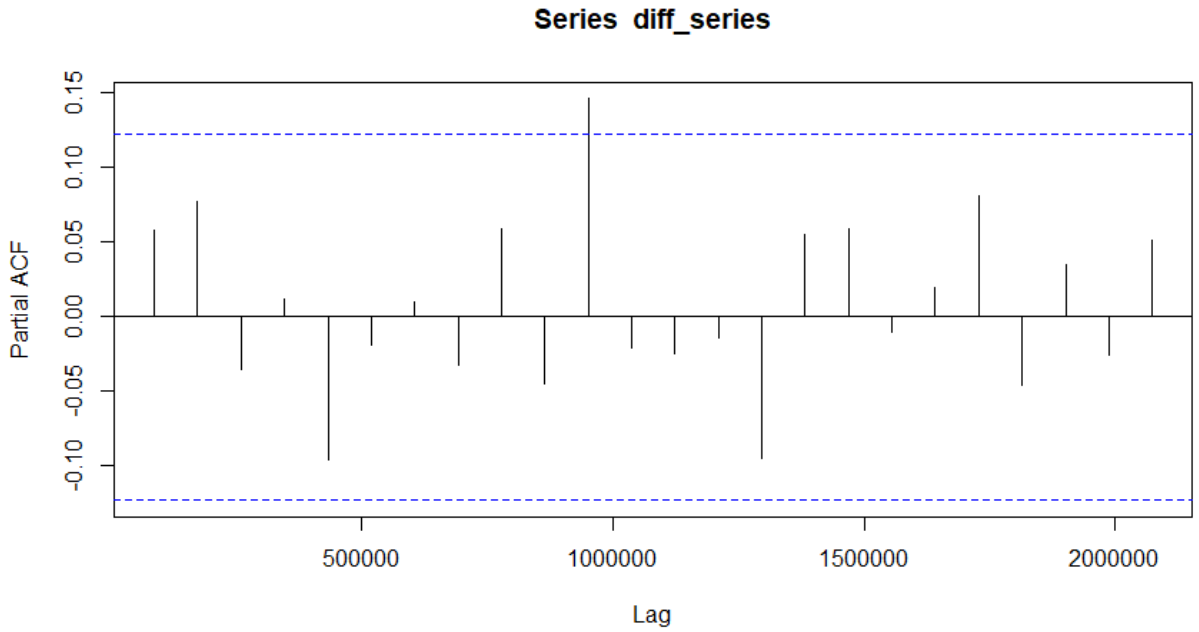

Figure 11: PACF for Sensex

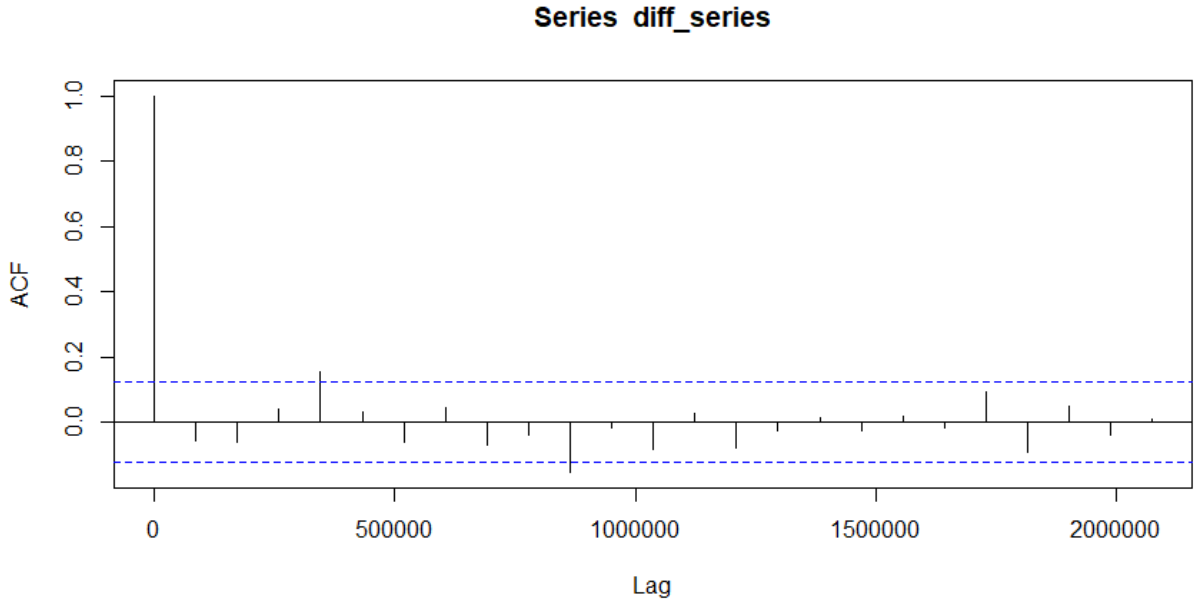

Figure 12: ACF for Gold

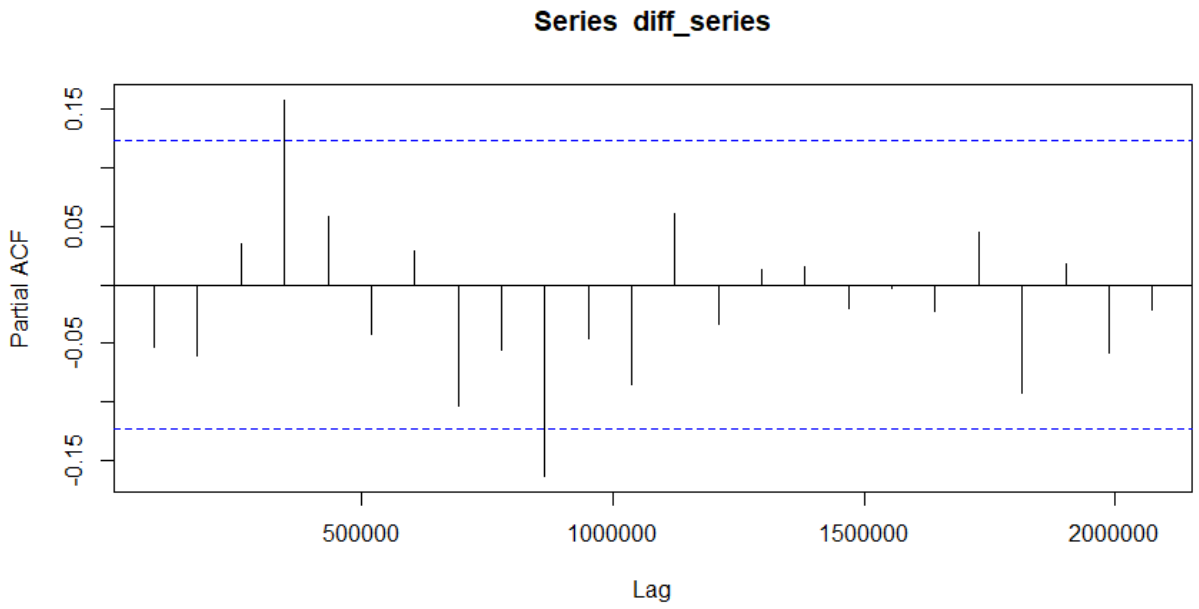

Figure 13: PACF for Gold

With different combinations of p, d and q parameter values, we proceeded to fit various ARIMA models and for each model we checked AIC and BIC values. Then, forecasts for three different time horizons: 20 days, 30 days, and 40 days were calculated and simultaneously, forecasts for the same time horizons were conducted by ChatGPT and Gemini to include these models in comparative analysis. The results for all these forecasted horizons by our seven models are visually represented which can be seen in Figures 14-25.

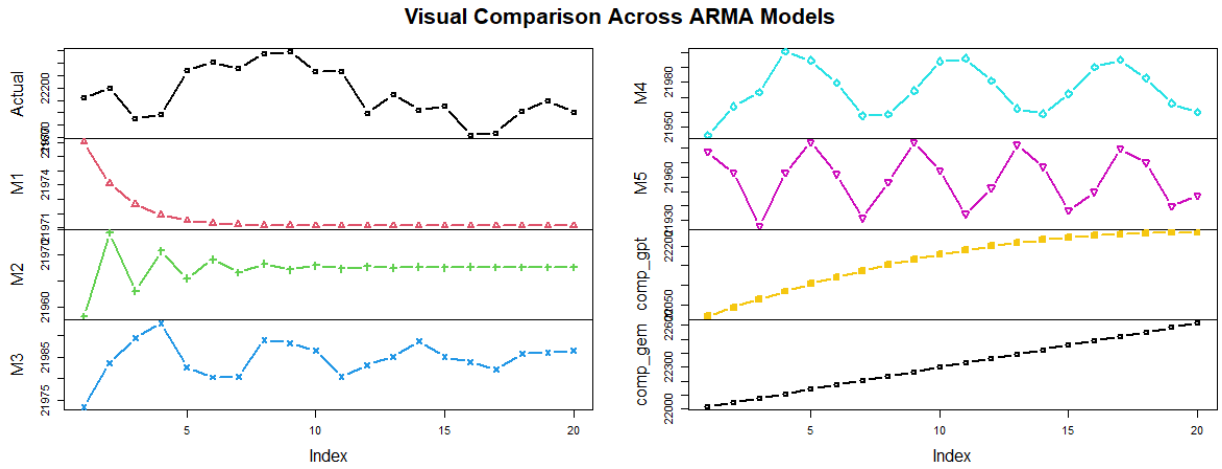

Figure 14: Nifty 50 ARIMA model comparison for 20 days horizon

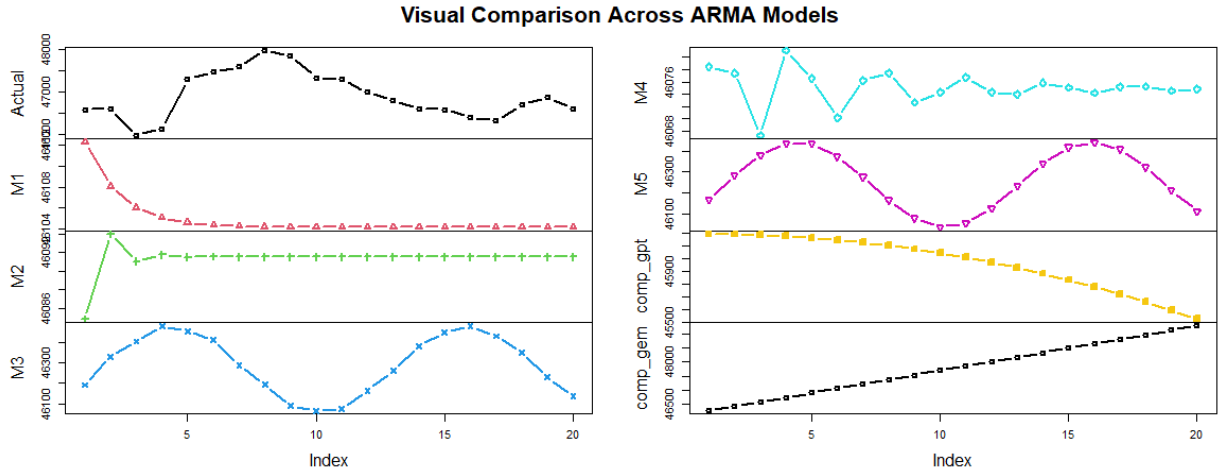

Figure 15: Bank Nifty ARIMA model comparison for 20 days horizon

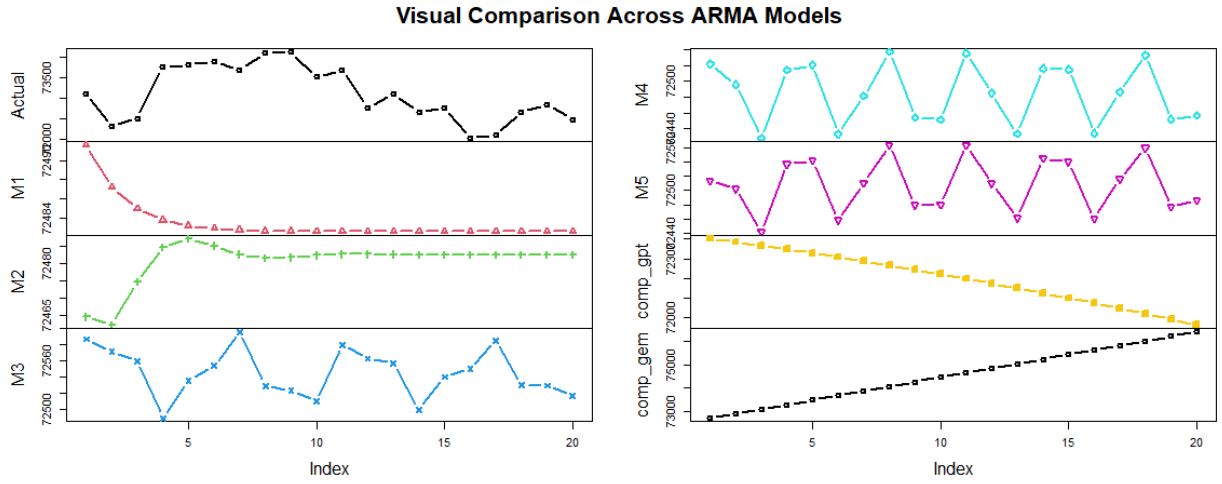

Figure 16: Sensex ARIMA model comparison for 20 days horizon

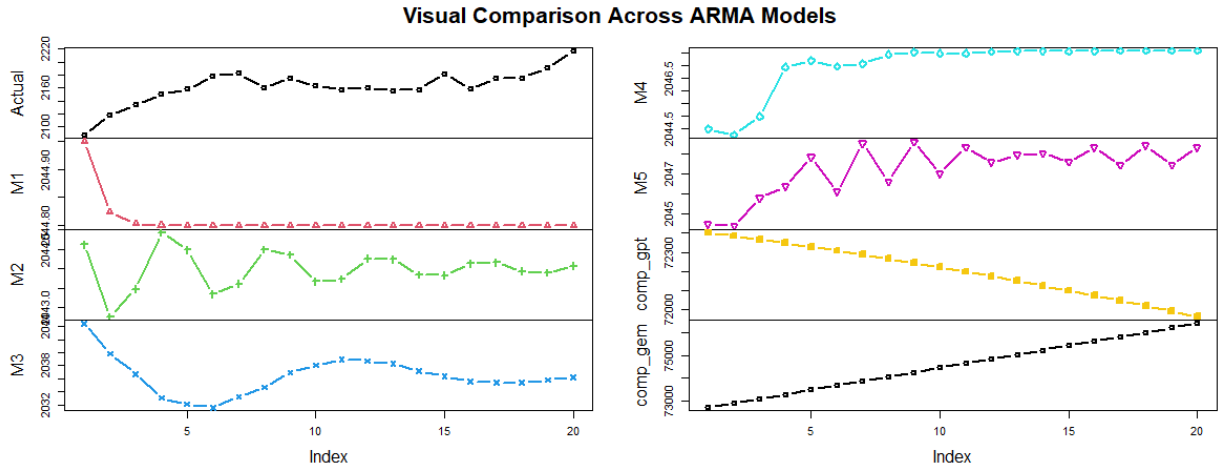

Figure 17: Gold ARIMA model comparison for 20 days horizon

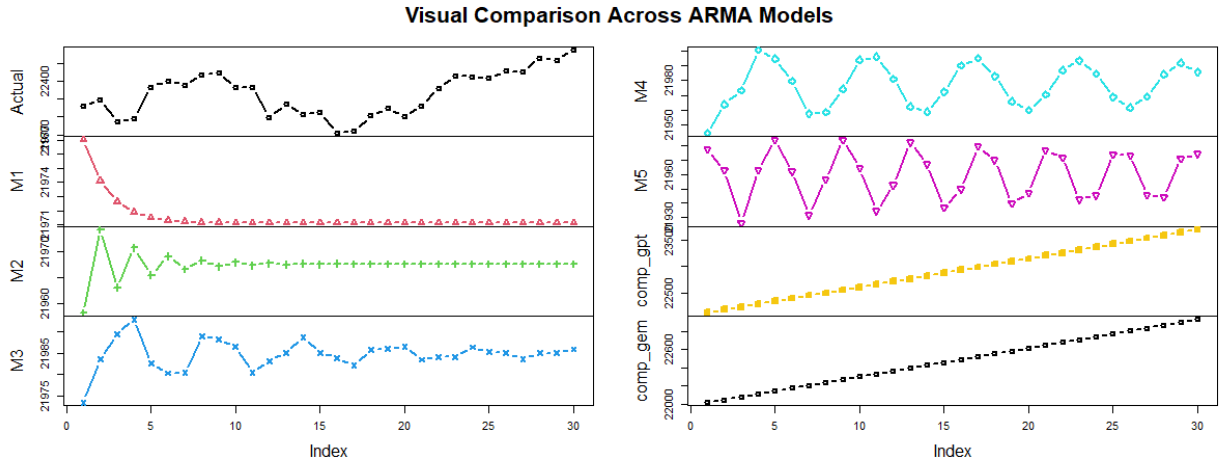

Figure 18: Nifty 50 ARIMA model comparison for 30 days horizon

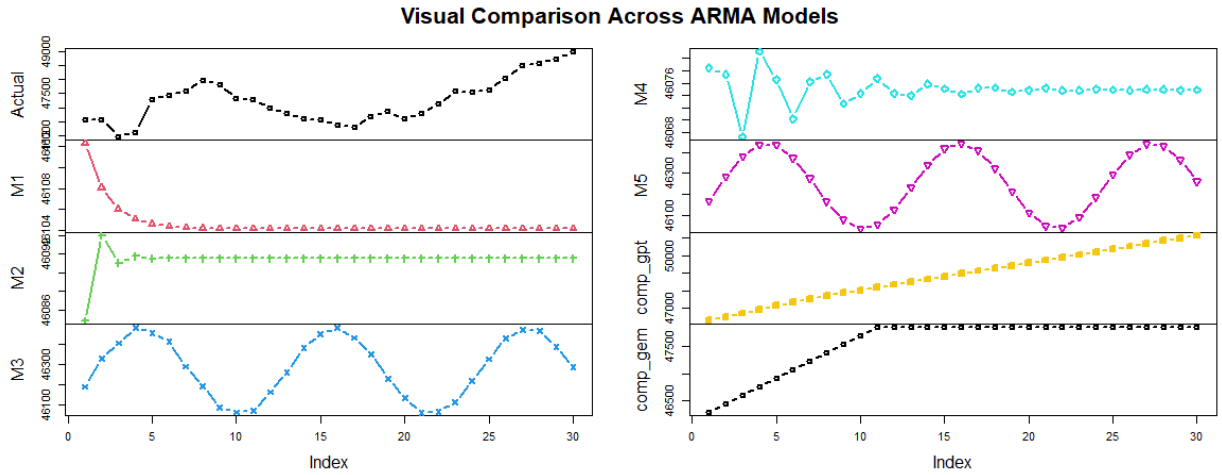

Figure 19: Bank Nifty ARIMA model comparison for 30 days horizon

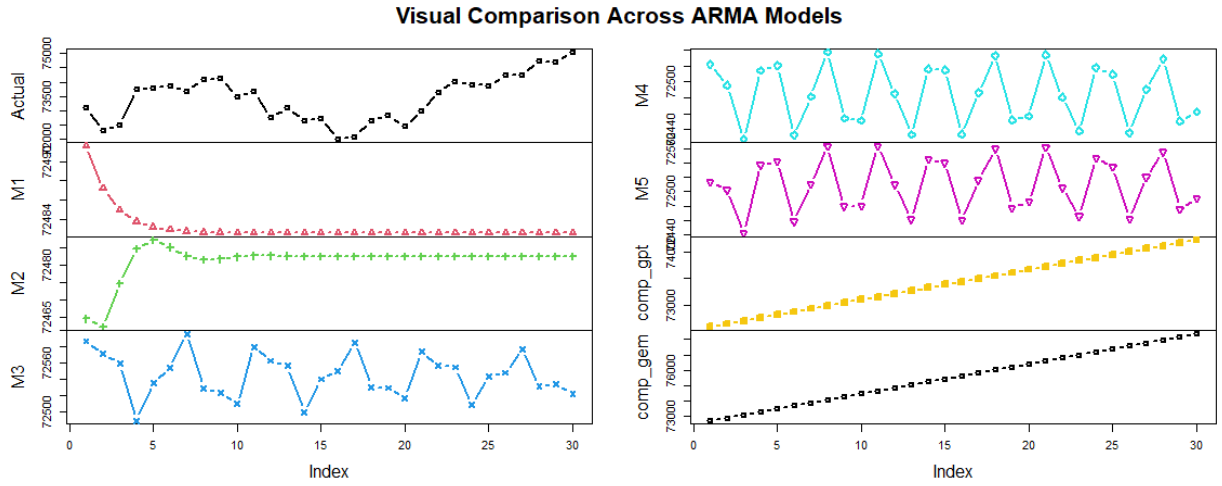

Figure 20: Sensex ARIMA model comparison for 30 days horizon

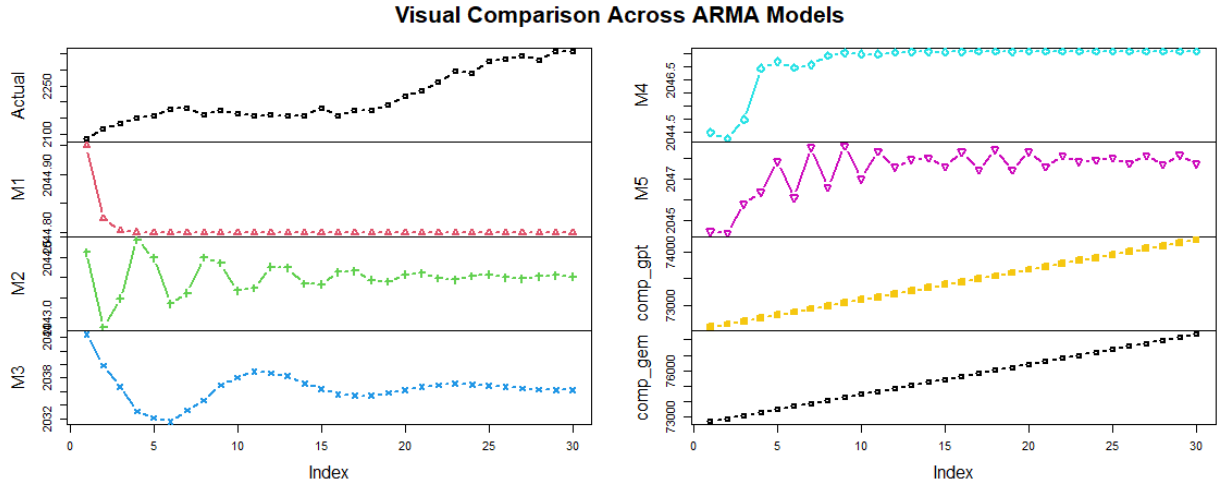

Figure 21: Gold ARIMA model comparison for 30 days horizon

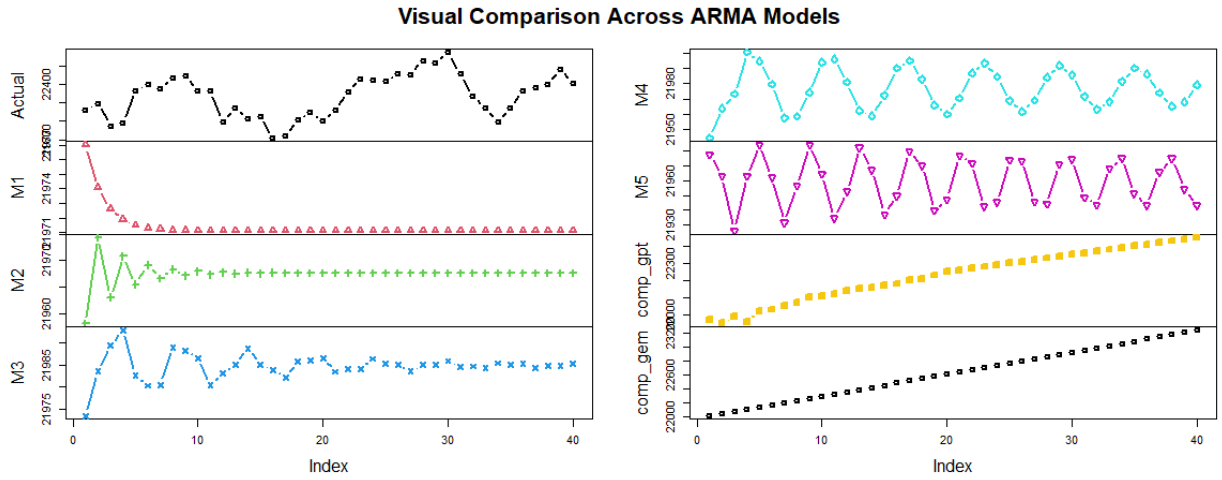

Figure 22: Nifty 50 ARIMA model comparison for 40 days horizon

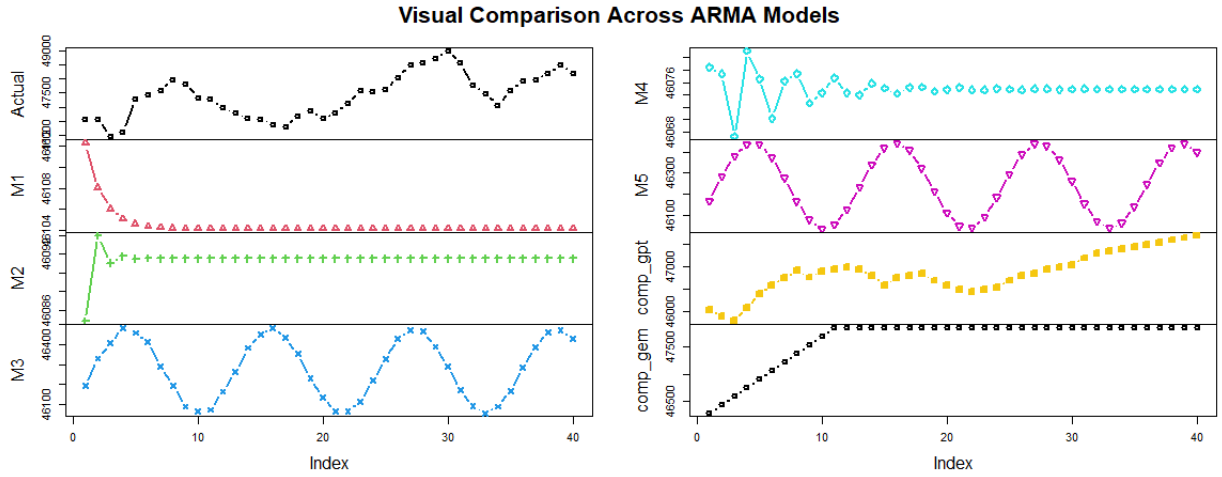

Figure 23: Bank Nifty ARIMA model comparison for 40 days horizon

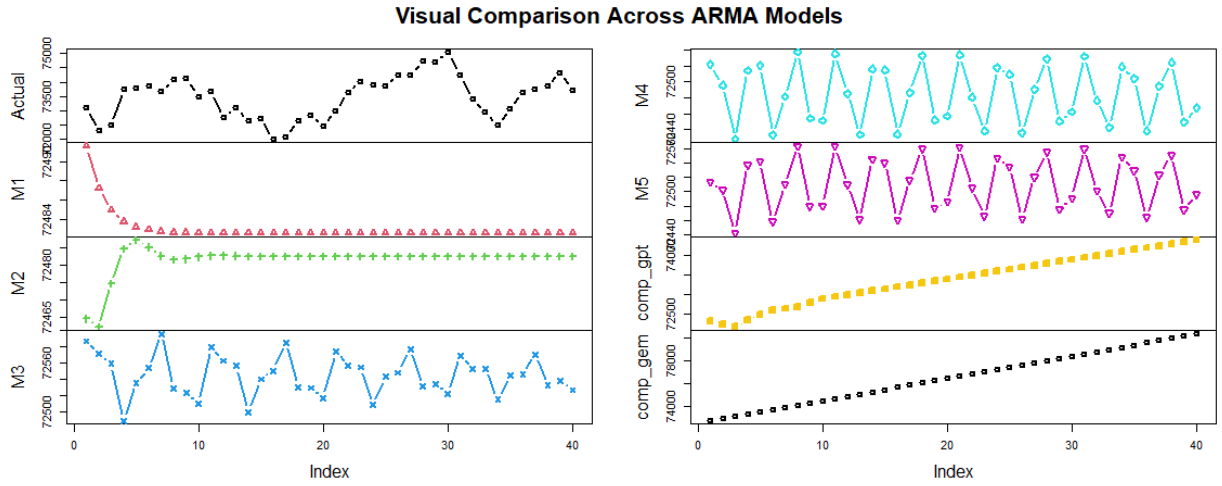

Figure 24: Sensex ARIMA model comparison for 40 days horizon

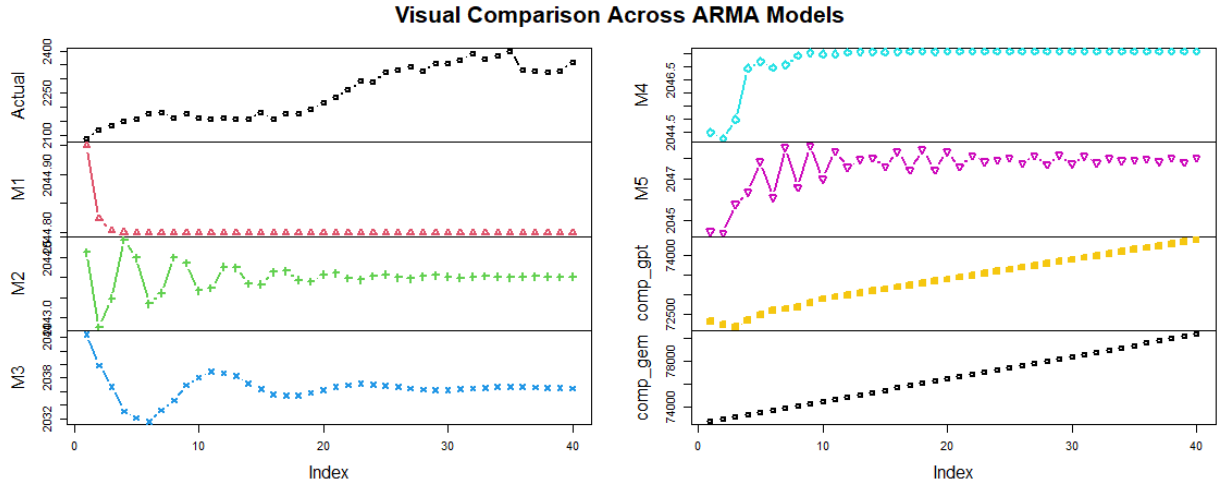

Figure 25: Gold ARIMA model comparison for 40 days horizon

Error metrics were calculated for all the forecasted time series of our seven parameters. These six different error metrics, namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Root Mean Squared Log Error (RMSLE), Mean Absolute Percent Error (MAPE) and Symmetric Mean Absolute Percent Error (SMAPE), were scaled and then an average of all these errors was

calculated which was termed as ‘Complex’ error. These error values were compared and ranked to identify the best-fitting model for all the time horizons for all four securities. In tables 2-14, we show the different results of errors obtained and rank them based on their best fit for predicting stock prices.

Complex Error Values												
	Nifty 50 _20	Nifty 50 _30	Nifty 50 _40	Bank Nifty_20	Bank Nifty_30	Bank Nifty_40	Sensex _20	Sensex _30	Sensex _40	Gold_20	Gold_30	Gold_40
M1	-0.12	-0.13	0.03	-0.12	0.04	0.30	-0.16	-0.07	-0.11	-0.20	-0.19	-0.17
M2	-0.09	-0.12	0.05	-0.09	0.05	0.32	-0.16	-0.07	-0.11	-0.20	-0.18	-0.17
M3	-0.52	-0.42	-0.10	-0.89	-0.37	0.28	-0.50	-0.28	-0.34	-0.59	-0.59	-0.59
M4	-0.10	-0.14	0.02	-0.06	0.07	0.33	-0.17	-0.07	-0.11	-0.20	-0.19	-0.18
M5	-0.02	-0.10	0.09	-0.31	-0.11	0.16	-0.19	-0.09	-0.12	-0.20	-0.19	-0.18
ChatGPT	-0.30	0.92	-0.73	0.19	0.77	-0.54	0.03	-0.41	-0.28	1.84	1.84	1.83
Gemini	0.88	-0.20	0.65	0.82	-0.59	-0.59	0.90	0.88	0.92	1.88	1.88	1.88

Table 2: Shows the complex error metric values

Complex Error Ranking												
	Nifty 50 _20	Nifty 50 _30	Nifty 50 _40	Bank Nifty_20	Bank Nifty_30	Bank Nifty_40	Sensex _20	Sensex _30	Sensex _40	Gold_20	Gold_30	Gold_40
M1	3	4	4	3	4	5	4	5	5	4	4	4
M2	5	5	5	4	5	6	5	6	6	5	5	5
M3	1	1	2	1	2	4	1	2	1	1	1	1
M4	4	3	3	5	6	7	3	4	4	3	3	3
M5	6	6	6	2	3	3	2	3	3	2	2	2
ChatGPT	2	7	1	6	7	2	6	1	2	6	6	6
Gemini	7	2	7	7	1	1	7	7	7	7	7	7

Table 3: Shows the complex error ranking based on their values

Nifty 50 Error Rank - pred_20						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	4	2	4	3	2	3
M2	5	4	5	5	4	5
M3	2	1	2	1	1	1
M4	3	5	3	4	5	4
M5	6	6	6	6	6	6
nifty_gpt	1	3	1	2	3	2
nifty_gem	7	7	7	7	7	7

Table 4: Shows the rank of Error Metrics of Nifty 50 for 20 days forecasting

Bank Nifty Error Rank - pred_20						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	3	3	3	3	3	3
M2	4	4	4	4	4	4
M3	1	1	1	1	1	1
M4	5	5	5	5	5	5
M5	2	2	2	2	2	2
bank_nifty	6	6	6	6	6	6
bank_nifty	7	7	7	7	7	7

Table 5: Shows the rank of Error Metrics of Bank Nifty for 20 days forecasting

Sensex Error Rank - pred_20						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	4	4	4	4	4	4
M2	5	5	5	5	5	5
M3	1	1	1	1	1	1
M4	3	3	3	3	3	3
M5	2	2	2	2	2	2
sensex_gp	6	6	6	6	6	6
sensex_ge	7	7	7	7	7	7

Table 6: Shows the rank of Error Metrics of Sensex for 20 days forecasting

Gold Error Rank - pred_20						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	3	3	3	3	3	4
M2	4	4	4	4	4	5
M3	5	5	5	5	5	1
M4	2	2	2	2	2	3
M5	1	1	1	1	1	2
gold_gpt	6	6	6	6	6	6
gold_gem	7	7	7	7	7	7

Table 7: Shows the rank of Error Metrics of Gold for 20 days forecasting

Nifty 50 Error Rank - pred_30						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	4	4	4	4	4	4
M2	5	5	5	5	5	5
M3	2	2	2	2	2	1
M4	3	3	3	3	3	3
M5	6	6	6	6	6	6
nifty_gpt	7	7	7	7	7	7
nifty_gem	1	1	1	1	1	2

Table 8: Shows the rank of Error Metrics of Nifty 50 for 30 days forecasting

Bank Nifty Error Rank - pred_30						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	4	4	4	4	4	4
M2	5	5	5	5	5	5
M3	2	2	2	2	2	2
M4	6	6	6	6	6	6
M5	3	3	3	3	3	3
bank_nifty	7	7	7	7	7	7
bank_nifty	1	1	1	1	1	1

Table 9: Shows the rank of Error Metrics of Bank Nifty for 30 days forecasting

Sensex Error Rank - pred_30						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	5	5	5	5	5	5
M2	6	6	6	6	6	6
M3	2	2	2	2	2	2
M4	4	4	4	4	4	4
M5	3	3	3	3	3	3
sensex_gp	1	1	1	1	1	1
sensex_ge	7	7	7	7	7	7

Table 10: Shows the rank of Error Metrics of Sensex for 30 days forecasting

Gold Error Rank - pred_30						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	3	3	3	3	3	4
M2	4	4	4	4	4	5
M3	5	5	5	5	5	1
M4	2	2	2	2	2	3
M5	1	1	1	1	1	2
gold_gpt	6	6	6	6	6	6
gold_gem	7	7	7	7	7	7

Table 11: Shows the rank of Error Metrics of Gold for 30 days forecasting

Nifty 50 Error Rank - pred_40						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	4	4	4	4	4	4
M2	5	5	5	5	5	5
M3	2	2	2	2	2	2
M4	3	3	3	3	3	3
M5	6	6	6	6	6	6
nifty_gpt	1	1	1	1	1	1
nifty_gem	7	7	7	7	7	7

Table 12: Shows the rank of Error Metrics of Nifty 50 for 40 days forecasting

Bank Nifty Error Rank - pred_40						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	5	5	5	5	5	5
M2	6	6	6	6	6	6
M3	3	3	3	3	3	4
M4	7	7	7	7	7	7
M5	4	4	4	4	4	3
bank_nifty	2	2	2	2	2	2
bank_nifty	1	1	1	1	1	1

Table 13: Shows the rank of Error Metrics of Bank Nifty for 40 days forecasting

Sensex Error Rank - pred_40						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	5	5	5	5	5	5
M2	6	6	6	6	6	6
M3	2	2	2	2	2	1
M4	4	4	4	4	4	4
M5	3	3	3	3	3	3
sensex_gp	1	1	1	1	1	2
sensex_ge	7	7	7	7	7	7

Table 14: Shows the rank of Error Metrics of Sensex for 40 days forecasting

Gold Error Rank - pred_40						
	RMSE	MAE	RMSLE	MAPE	SMAPE	Complex
M1	3	3	3	3	3	4
M2	4	4	4	4	4	5
M3	5	5	5	5	5	1
M4	2	2	2	2	2	3
M5	1	1	1	1	1	2
gold_gpt	6	6	6	6	6	6
gold_gem	7	7	7	7	7	7

Table 15: Shows the rank of Error Metrics of Gold for 40 days forecasting

As you can see from the Tables 2-5, for the 20-day horizon for all four securities, ARIMA models are doing better in forecasting (also as evident in ranks). In the case of stocks (Nifty 50, Bank Nifty, Sensex) model M3 (3, 1, 3) is doing better and in the case of Gold M5 (5, 1, 3) is doing better.

From Tables 6-9 of the 30-day horizon, in case of stocks atleast one of the LLMs are always performing better, as you can observe in case of Nifty 50 and Bank Nifty Gemini is doing better and in case of Sensex, ChatGPT is doing better. But in case of Gold, LLMs have the worst performance among all models and M5(5, 1, 3) is performing best.

From tables 10-13 of the 40-day horizon, in the case of stocks again one of the LLMs is always performing better, as can be seen in the case of the Nifty 50 and Sensex, ChatGPT is doing better, and the case of Bank Nifty, Gemini is performing better. But in case of Gold again, ChatGPT and Gemini are performing worst, and model M5(5, 1, 3) is the best performing model. In Gold, we can observe the same rankings across all three horizons for all the computed models. Also, the ARIMA model is performing better, which in this is M5(5, 1, 3) and LLMs are performing worst, for all the models and in the case of stocks, the second-best model is always an ARIMA model.

6.1 Conclusion

The research evaluated the performance in forecasting abilities of large language models (LLMs) like ChatGPT and Gemini in a comparative analysis with our novel ARIMA models. We spanned our analysis over four major securities (Nifty 50, Bank Nifty, Sensex, Gold) and three time horizons (20, 30 and 40 days).

For shorter time forecasts (20 days), the ARIMA model showed superior accuracy especially in predicting stock (for Nifty 50, Bank Nifty and Sensex) and gold prices, while extending the horizon to 30 or 40 days, either of the LLMs, ChatGPT or Gemini, outperformed ARIMA models for stock predictions. Notably, Gemini performed better in longer time horizons while ChatGPT showed competitive results.

For Gold price predictions, across all horizons, ARIMA models consistently outperformed ChatGPT and Gemini. This highlights the possibility of further exploring the reliability of LLMs in forecasting prices of commodities and simultaneously, further fine-tuning the models to more accurately predict the stock prices. The ARIMA models showed consistent results across all tested scenarios.

Our study showed that ARIMA model are better for short-term forecasting for commodities like gold while LLMs have a better performance in longer-term stock price predictions. Further exploration of why LLMs performed better for certain data types and different time horizons might add to the ongoing discussion to the use of LLMs in the field of finance and provide a broader view of how LLMs can be made mode effective in financial forecasting.

References:

[1] A. Bahrammirzaee, “A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems,” Neural Computing and Applications, vol. 19, no. 8, pp. 1165–1195, 2010.

[2] Y. Hilpisch, Artificial Intelligence in Finance. O’Reilly Media, 2020.

[3] J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik, “Artificial intelligence and machine

learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis," Journal of Behavioral and Experimental Finance, vol. 32, p. 100577, 2021.

[4] F. K¨onigstorfer and S. Thalmann, "Applications of artificial intelligence in commercial banks–a research agenda for behavioral finance," Journal of behavioral and experimental finance, vol. 27, p. 100352, 2020.

[5] P. Giudici, "Fintech risk management: A research challenge for artificial intelligence in finance," Frontiers in Artificial Intelligence, vol. 1, p. 1, 2018.

[6] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray et al., "Training language models to follow instructions with human feedback," Advances in neural information processing systems, vol. 35, pp. 27 730–27 744, 2022.

[7] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," Advances in neural information processing systems, vol. 30, 2017.

[8] M. Dowling and B. Lucey, "Chatgpt for (finance) research: The bananarama conjecture," Finance Research Letters, vol. 53, p. 103662, 2023.

[9] N. Rane, "Role and challenges of chatgpt and similar generative artificial intelligence in finance and accounting," Available at SSRN 4603206, 2023.

[10] Y. Cao and J. Zhai, "Bridging the gap–the impact of chatgpt on financial research," Journal of Chinese Economic and Business Studies, vol. 21, no. 2, pp. 177–191, 2023.

[11] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, "A brief overview of chatgpt: The history, status quo and potential future development," IEEE/CAA Journal of Automatica Sinica, vol. 10, no. 5, pp. 1122–1136, 2023

[12] M. S. Khan and H. Umer, "Chatgpt in finance: Applications, challenges, and solutions," Heliyon, vol. 10, no. 2, 2024.

[13] K. Wenzlaff and S. Spaeth, "Smarter than humans? validating how openai's chatgpt model explains crowdfunding, alternative finance and community finance," Validating how OpenAI's

ChatGPT Model Explains Crowdfunding, Alternative Finance and Community Finance (December 22, 2022), 2022.

[14] G. Fatouros, J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis, "Transforming sentiment analysis in the financial domain with chatgpt," Machine Learning with Applications, vol. 14, p. 100508, 2023.

[15] A. W. Lo, M. Singh, J. Musumeci, Z. Nagy, G. Giese, X. Wang, A. Mirabelli, N. Keywork, A. Turetsky, B. Griffiths et al., "From eliza to chatgpt: The evolution of natural language processing and financial applications," The Journal of Portfolio Management, vol. 49, no. 7, pp. 201–235, 2023.

[16] D. Krause, "Large language models and generative ai in finance: An analysis of chatgpt, bard, and bing ai," Bard, and Bing AI (July 15, 2023), 2023.

[17] X. Zhang and Q. Yang, "Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters," in Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 4435–4439.

[18] L. Loukas, I. Stogiannidis, P. Malakasiotis, and S. Vassos, "Breaking the bank with chatgpt: Few-shot text classification for finance," arXiv preprint arXiv:2308.14634, 2023.

[19] S. Itoh and K. Okada, "The power of large language models: A chatgpt-driven textual analysis of fundamental data," Available at SSRN 4546163, 2023.

[20] J.-J. Wang, J.-Z. Wang, Z.-G. Zhang, and S.-P. Guo, "Stock index forecasting based on a hybrid model," Omega, vol. 40, no. 6, pp. 758–766, 2012.

[21] L.-Y. Wei, "A hybrid model based on anfis and adaptive expectation genetic algorithm to forecast taiex," Economic Modelling, vol. 33, pp. 893–899, 2013.

[22] G. S. Atsalakis, E. M. Dimitrakakis, and C. D. Zopounidis, "Elliott wave theory and neuro-fuzzy systems, in stock market prediction: The wasp system," Expert Systems with Applications, vol. 38, no. 8, pp. 9196–9206, 2011.

[23] A. Chernyavskiy, D. Ilvovsky, and P. Nakov, "Transformers:"the end of history" for natural

language processing?" in Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21. Springer, 2021, pp. 677–693.

[24] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," Advances in neural information processing systems, vol. 32, 2019.

[25] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu et al., "Towards a human-like open-domain chatbot," arXiv preprint arXiv:2001.09977, 2020.

26] B. A. y Arcas, "Do large language models understand us?" Daedalus, vol. 151, no. 2, pp. 183–197, 2022.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.

[28] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[29] Y. Ding, S. Jia, T. Ma, B. Mao, X. Zhou, L. Li, and D. Han, "Integrating stock features and global information via large language models for enhanced stock return prediction," arXiv preprint arXiv:2310.05627, 2023.

[30] A. S. George and A. H. George, "A review of chatgpt ai's impact on several business sectors," Partners Universal International Innovation Journal, vol. 1, no. 1, pp. 9–23, 2023.

[31] Ö. Aydin and E. Karaarslan, "Is chatgpt leading generative ai? what is beyond expectations?" Academic Platform Journal of Engineering and Smart Systems, vol. 11, no. 3, pp. 118–134, 2023.

[32] L. A. Smales, "Classification of rba monetary policy announcements using chatgpt," Finance

Research Letters, vol. 58, p. 104514, 2023.

[33] P. Perera and M. Lankathilake, “Preparing to revolutionize education with the multi-model genai tool google gemini? a journey towards effective policy making,” J. Adv. Educ. Philos, vol. 7, pp. 246–253, 2023.

[34] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, and M. N. Halgamuge, “From google gemini to openai q*(q-star): A survey of reshaping the generative artificial intelligence (ai) research landscape,” arXiv preprint arXiv:2312.10868, 2023.

[35] H. R. Saeidnia, “Welcome to the gemini era: Google deepmind and the information industry,” Library Hi Tech News, no. ahead-of-print, 2023.

[36] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth et al., “Gemini: a family of highly capable multimodal models,” arXiv preprint arXiv:2312.11805, 2023.

[37] G.-G. Lee, E. Latif, L. Shi, and X. Zhai, “Gemini pro defeated by GPT-4V: Evidence from education,” arXiv preprint arXiv:2401.08660, 2023.

[38] C. Fu et al., “A challenger to gpt-4v? early explorations of gemini in visual expertise,” arXiv preprint arXiv:2312.12436, 2023.

[39] Y. Wang and Y. Zhao, “Gemini in reasoning: Unveiling commonsense in multimodal large language models,” arXiv preprint arXiv:2312.17661, 2023.

[40] M. Masalkhi et al., “Google deepmind’s gemini ai versus chatgpt: A comparative analysis in ophthalmology,” Eye, pp. 1–6, 2024

[41] Z. Qi, Y. Fang, M. Zhang, Z. Sun, T. Wu, Z. Liu, D. Lin, J. Wang, and H. Zhao, “Gemini vs gpt-4v: A preliminary comparison and combination of vision-language models through qualitative cases,” ArXiv, vol. abs/2312.15011, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266550760

[42] M.M. Mohamed, “Forecasting stock exchange movements using neural networks: empirical evidence from Kuwait”, Expert Systems with Applications, vol. 27, no. 9, pp.6302–6309, 2010

[43] L.C. Kyungjoo, Y. Sehwan and J. John, "Neural Network Model vs. SARIMA Model In Forecasting Korean Stock Price Index (KOSPI), Issues in Information System, vol. 8 no. 2, pp. 372-378, 2007.

[44] N. Merh, V. P. Saxena, and K. R. Pardasani, "A comparison between hybrid approaches of ann and arima for indian stock trend forecasting," Business Intelligence Journal, vol. 3, no. 2, pp. 23–43, 2010.

[45] C. Javier, E. Rosario, J. Francisco, and J. C. Antonio, "Arima models to predict next electricity price," IEEE Transactions on Power Systems, vol. 18, no. 3, pp. 1014–1020, 2003.

[46] R. Nochai and T. Nochai, "Arima model for forecasting oil palm price," in Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and applications. Academia Penang, 2006, pp. 13–15.

[46a] Tripathi, A., Dixit, A., & Vipul. (2021b). Liquidity commonality in the cryptocurrency market. *Applied Economics*, *0*(0), 1–15. https://doi.org/10.1080/00036846.2021.1982128

[47] M. Khashei, M. Bijari, and G. A. R. Ardali, "Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (anns)," Neurocomputing, vol. 72, no. 4-6, pp. 956–967, 2009.

[48] C.-M. Lee and C.-N. Ko, "Short-term load forecasting using lifting scheme and arima models,"Expert Systems with Applications, vol. 38, no. 5, pp. 5902–5911, 2011.

[49] M. Khashei, M. Bijari, and G. A. R. Ardali, "Hybridization of autoregressive integrated moving average (arima) with probabilistic neural networks (pnns)," Computers & Industrial Engineer-ing, vol. 63, no. 1, pp. 37–45, 2012.

[50] C.-C. Wang, "A comparison study between fuzzy time series model and arima model for fore-casting taiwan export," Expert Systems with Applications, vol. 38, no. 8, pp. 9296–9304, 2011.

[50a] Tripathi, A., & Pandey, A. (2021). Information dissemination across global markets during the spread of COVID-19 pandemic. *International Review of Economics & Finance*, *74*, 103–115. https://doi.org/10.1016/j.iref.2021.02.004

[51] A. Meyler, G. Kenny, and T. Quinn, "Forecasting irish inflation using arima models," 1998.

[52] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock price prediction using the arima model,"

in 2014 UKSim-AMSS 16th international conference on computer modelling and simulation. IEEE, 2014, pp. 106–112.

[52a] Tripathi, A., Dixit, A., & Vipul. (2021a). Information content of order imbalance in an order-driven market: Indian Evidence. *Finance Research Letters*, *41*, 101863. https://doi.org/10.1016/j.frl.2020.101863

[53] L. Zhihao, W. Junpei, Z. Xiaoliang, and N. Huijun, "Research on covid-19 epidemic based on arima model," in Journal of Physics: Conference Series, vol. 2012, no. 1. IOP Publishing, 2021, p. 012063.

[53a] Tripathi, A., Vipul, & Dixit, A. (2020). Limit order books: A systematic review of literature. *Qualitative Research in Financial Markets*, *12*(4), 505–541. https://doi.org/10.1108/QRFM-07-2019-0080

[53b] Tripathi, A., Vipul, & Dixit, A. (2020a). Liquidity commonality beyond best prices: Indian evidence. *Journal of Asset Management*, *21*(4), 355–373. https://doi.org/10.1057/s41260-020-00164-3

[53c] Tripathi, A., Vipul, V., & Dixit, A. (2020b). Adaptive market hypothesis and investor sentiments: Global evidence. *Managerial Finance*.

[54] J. Fattah, L. Ezzine, Z. Aman, H. El Moussami, and A. Lachhab, "Forecasting of demand using arima model," International Journal of Engineering Business Management, vol. 10, p. 1847979018808673, 2018.

[54a] Tripathi, A., & Dixit, A. (2023). Global Component of Sentiment in Futures Markets: Evidence from Covid-19 Pandemic. *American Business Review*, *26*(2), 4.

[55] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time series analysis: forecasting and control. John Wiley & Sons, 2015.

[56] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou et al., "Chain-of-thought prompting elicits reasoning in large language models," Advances in neural information processing systems, vol. 35, pp. 24 824–24 837, 2022.

[56a] Tripathi, A. (2021). The Arrival of Information and Price Adjustment Across Extreme Quantiles: Global Evidence. *IIM Kozhikode Society & Management Review*, *10*(1), 7–19. https://doi.org/10.1177/2277975220937994

[56b] Tripathi, A., Dixit, A., & Vipul. (2019). Liquidity of financial markets: A review. *Studies in Economics and Finance*. https://doi.org/10.1108/SEF-10-2018-0319

[56c] Tripathi, A., Dixit, A., & Vipul. (2020). Liquidity commonality in extreme quantiles: Indian evidence. *Finance Research Letters*, 101448. https://doi.org/10.1016/j.frl.2020.101448

[57] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," Advances in neural information processing systems, vol. 35, pp. 22 199–22 213, 2022.