Stock Price Prediction Using Gradient Boosting and ARIMA

1. Introduction

Stock price prediction is a challenging problem due to the inherent volatility of financial markets. Various modelling techniques, including machine learning and statistical methods, have been explored to predict stock prices. In this study, we compare two models:

1. **Gradient Boosting Regression (GBR)** – A machine learning technique that captures complex patterns in stock prices.
2. **Autoregressive Integrated Moving Average (ARIMA)** – A statistical time series model designed for trend-based forecasting.

The goal is to evaluate the performance of these models and determine which one provides the most accurate predictions. The study also discusses the implications of the findings on trading strategies.

---

2. Methodology

2.1 Data Collection

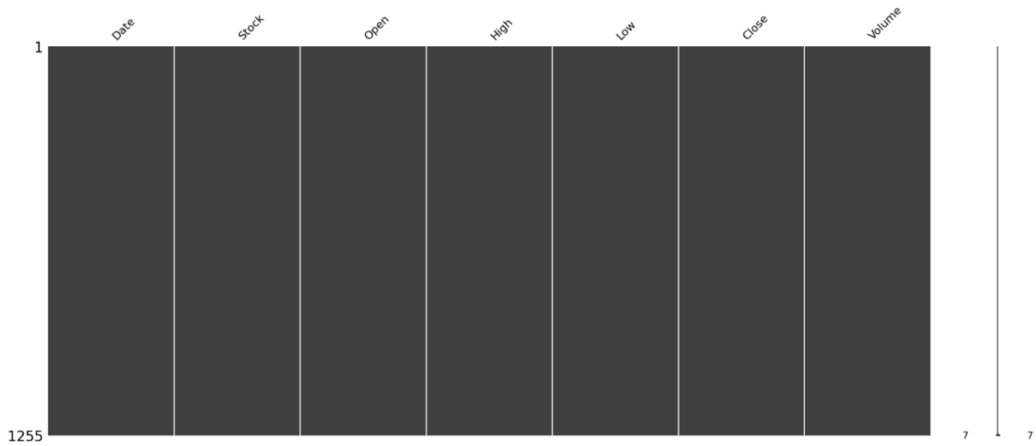
The dataset consists of historical stock prices with the following features:

- **Date, Open, High, Low, Close, Volume**
- **Moving Averages:** SMA_7, SMA_30, SMA_50
- **Lag Features:** Close_Lag_1, Close_Lag_3, Close_Lag_7
- **Daily Return:** The percentage change in stock price from the previous day

The dataset was split into:

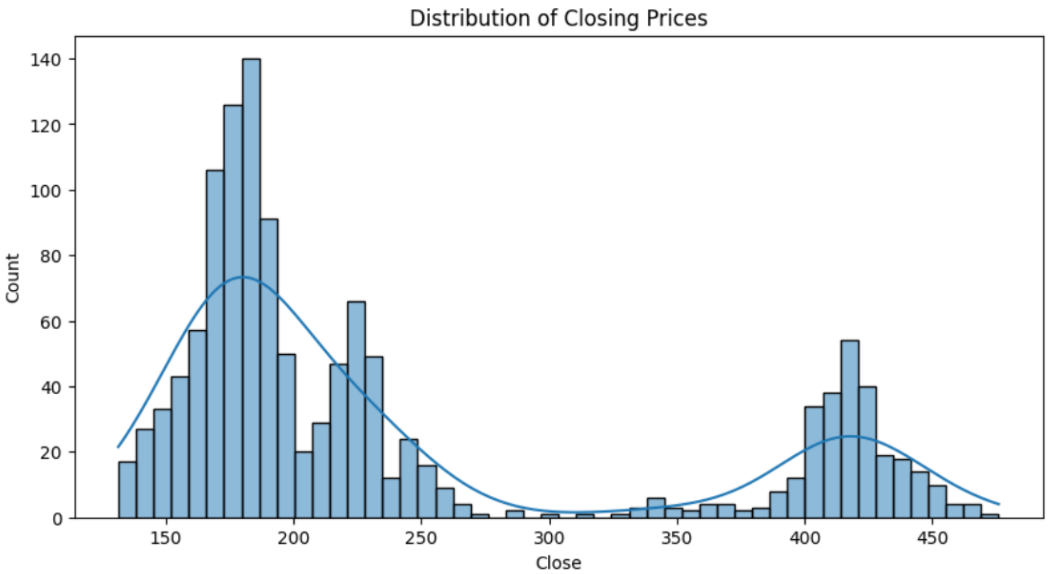
- **Training Set:** 80% of the data
- **Testing Set:** 20% of the data

2.2 Data Processing

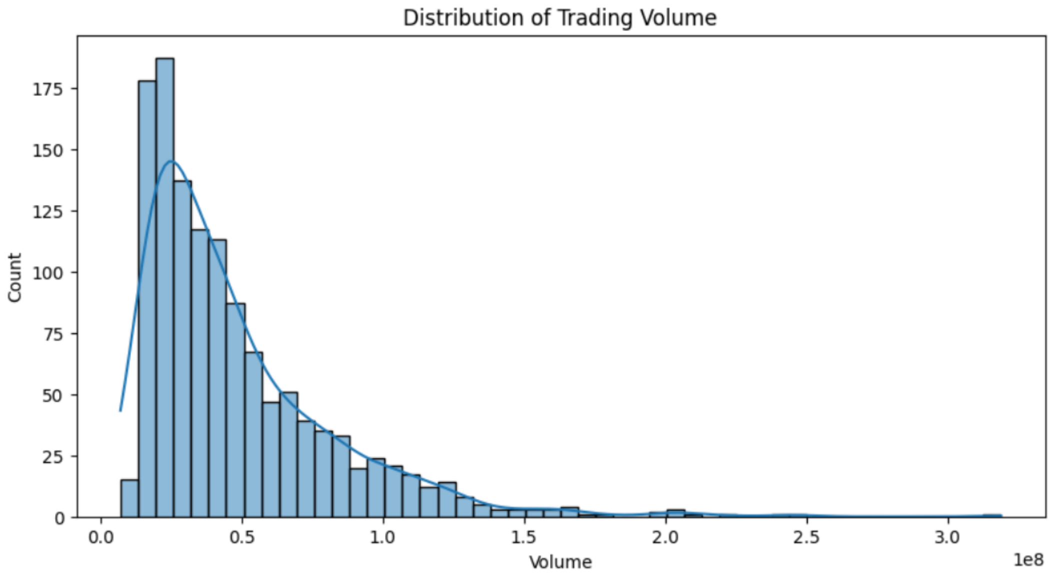


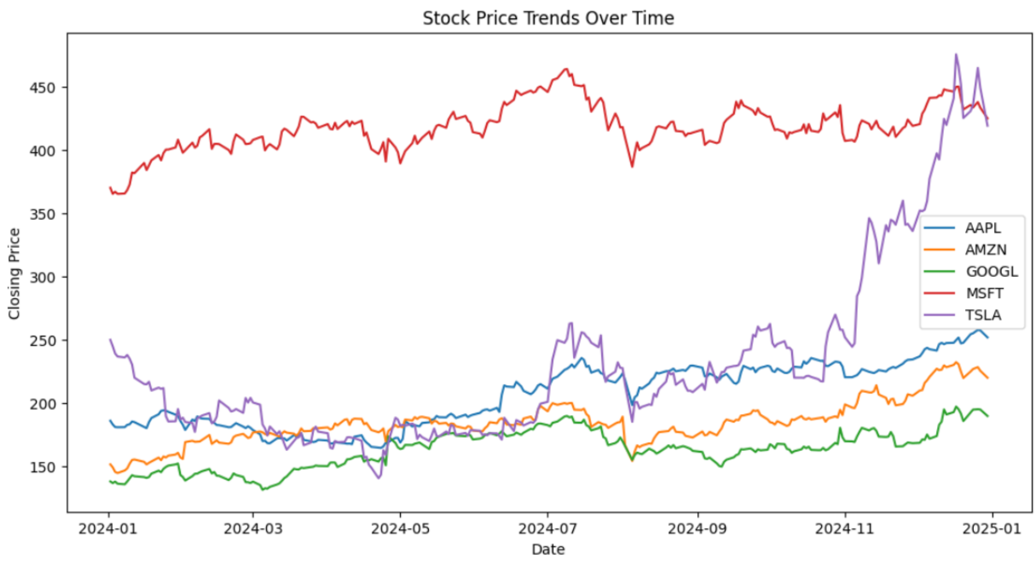
- Missing values were handled using interpolation techniques.
- Feature engineering was performed by adding lag variables and moving averages to enhance predictive performance.
- The dataset was normalized to ensure better convergence during model training.

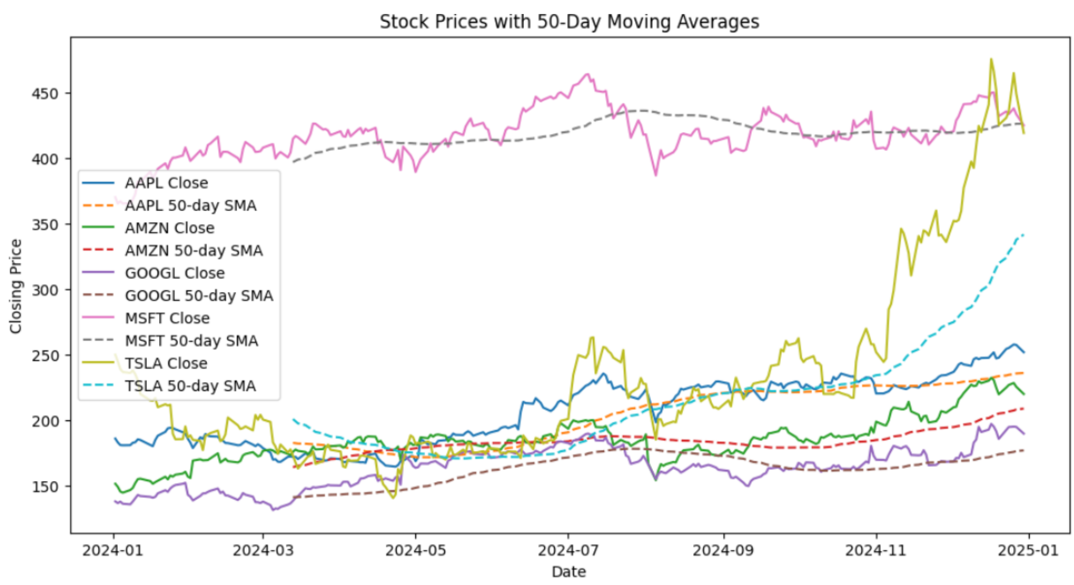
2.3 Exploratory Data Analysis (EDA)

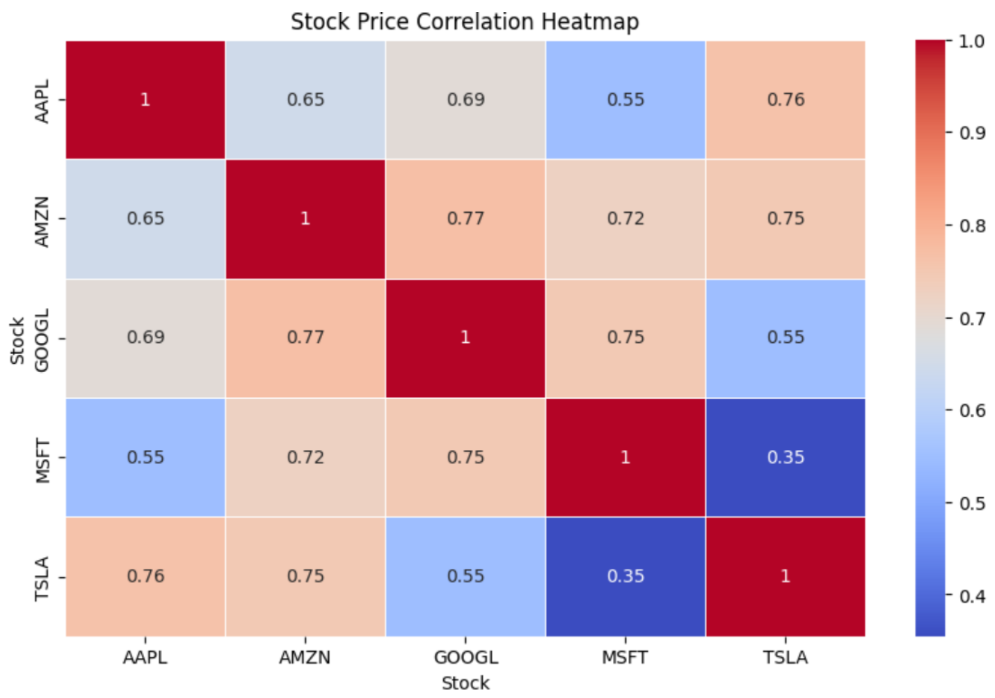


- The stock prices exhibited seasonality and upward trends over time.
- Moving averages provided insights into trend reversals.
- Daily returns indicated volatility, which is crucial for predicting price movements.


Distribution of Trading Volume
Count
Volume
1e8


Stock Price Trends Over Time
Closing Price
Date
AAPL
AMZN
GOOGL
MSFT
TSLA


Stock Prices with 50-Day Moving Averages
Closing Price
Date
AAPL Close
AAPL 50-day SMA
AMZN Close
AMZN 50-day SMA
GOOGL Close
GOOGL 50-day SMA
MSFT Close
MSFT 50-day SMA
TSLA Close
TSLA 50-day SMA



---

3. Model Development and Results

3.1 Gradient Boosting Regression (GBR)

Gradient Boosting is an ensemble learning method that builds multiple weak decision trees and improves their performance iteratively. It is particularly effective for capturing **non-linear dependencies** in stock price data.

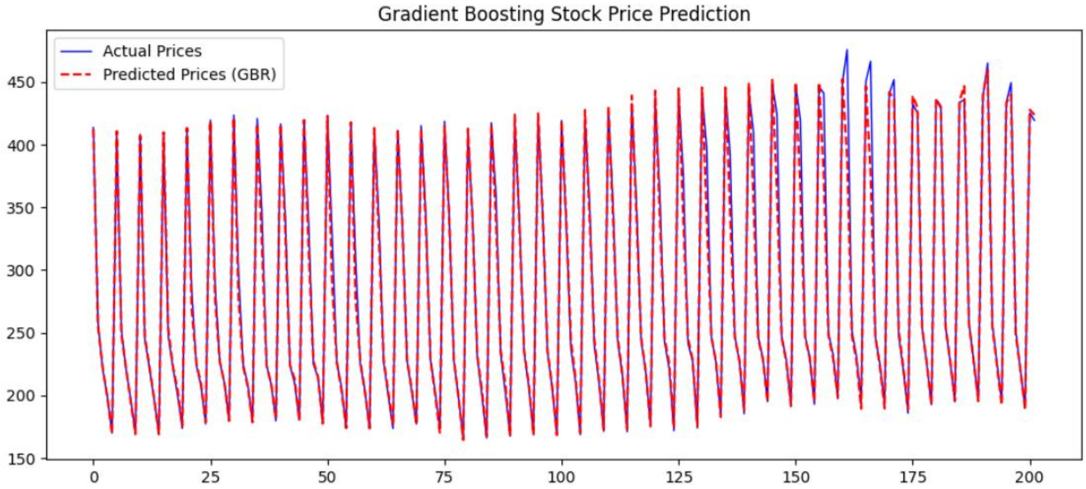
Results:

- **Mean Absolute Error (MAE):** 5.3895
- **Root Mean Squared Error (RMSE):** 11.7057
- **Mean Absolute Percentage Error (MAPE):** 1.62%
- **Model Accuracy:** 98.38%

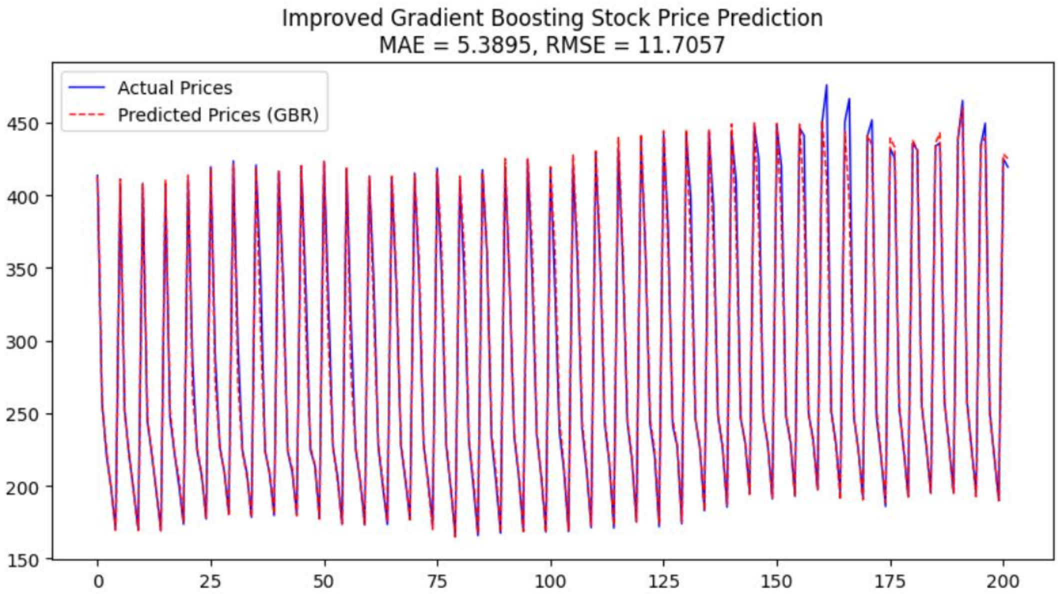
Visualization: ***(Leave space for GBR prediction graph)***

- The predicted values (red dashed line) closely follow the actual prices (blue line).
- GBR effectively captures daily price fluctuations, making it suitable for short-term price forecasting.

Model Performance: MAE = 5.9869, RMSE = 14.7300

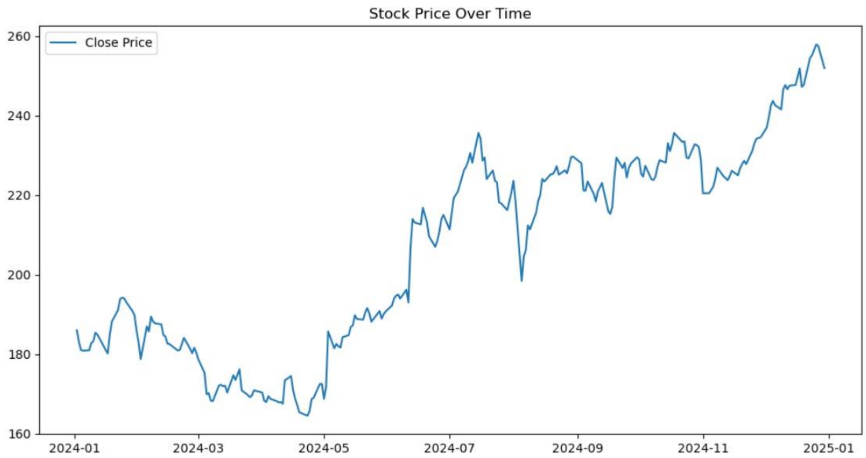


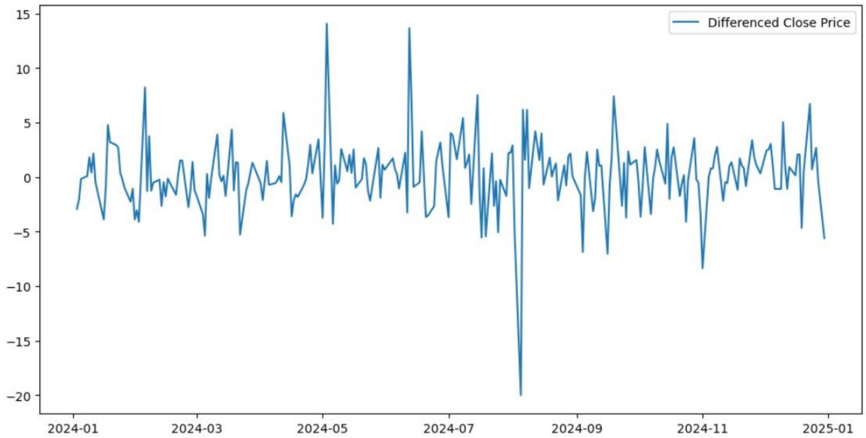
Improved Model Performance: MAE = 5.3895, RMSE = 11.7057



3.2 ARIMA Model

ARIMA is a traditional time series model that forecasts future values based on past trends. It is particularly useful for capturing long-term market trends but struggles with short-term fluctuations.






```
ADF Statistic: -15.075841060181713
p-value: 8.593227809057714e-28
The time series is stationary.
```

```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=1298.595, Time=0.40 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=1294.306, Time=0.03 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=1295.972, Time=0.07 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=1295.930, Time=0.10 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=1294.002, Time=0.03 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=inf, Time=0.70 sec

Best model:  ARIMA(0,1,0)(0,0,0)[0]
Total fit time: 1.343 seconds
```

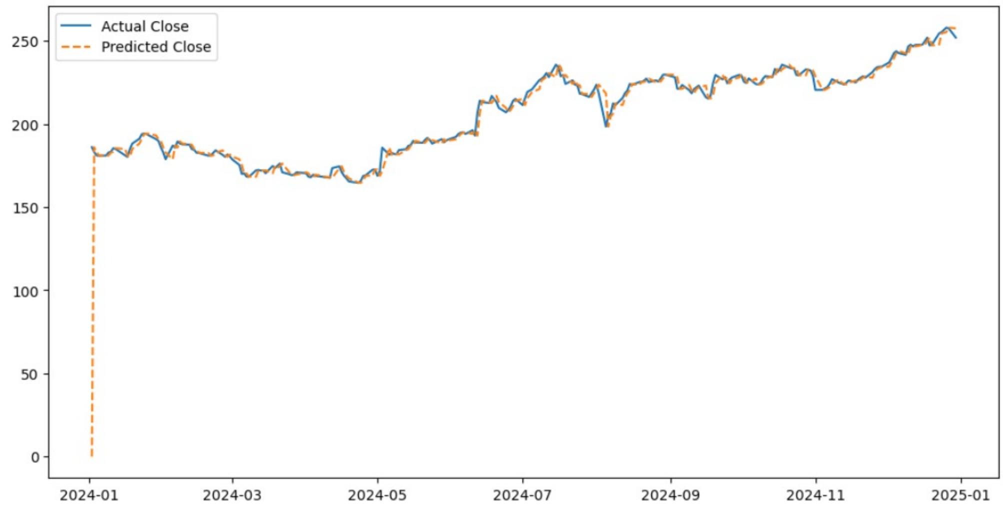
SARIMAX Results

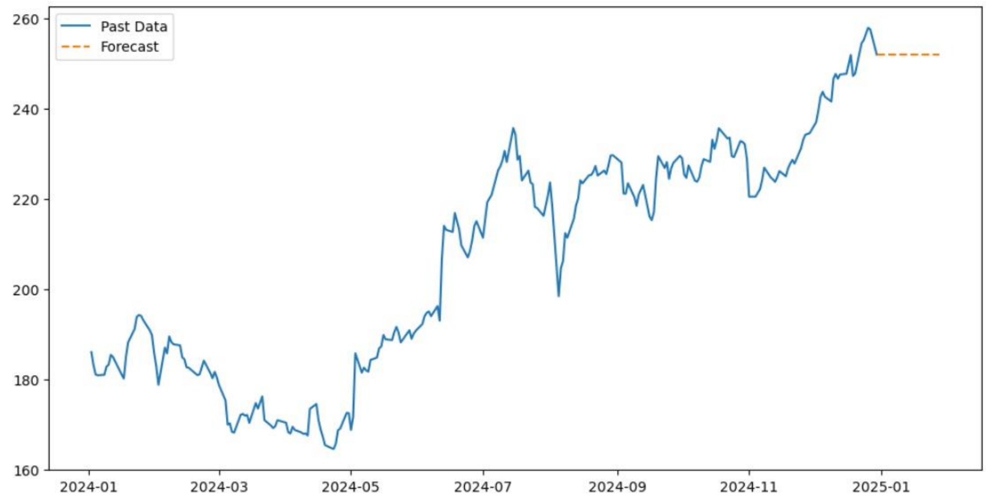
Dep. Variable:	y	**No. Observations:**	251
Model:	SARIMAX(0, 1, 0)	**Log Likelihood**	-646.001
Date:	Tue, 04 Mar 2025	**AIC**	1294.002
Time:	23:49:59	**BIC**	1297.524
Sample:	0	**HQIC**	1295.420
	- 251		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
sigma2	10.2793	0.419	24.555	0.000	9.459	11.100

Ljung-Box (L1) (Q):	0.33	**Jarque-Bera (JB):**	656.58
Prob(Q):	0.56	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.37	**Skew:**	-0.43
Prob(H) (two-sided):	0.15	**Kurtosis:**	10.89

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Close   No. Observations:                  251
Model:                 ARIMA(0, 1, 0)   Log Likelihood                -646.001
Date:                Tue, 04 Mar 2025   AIC                           1294.002
Time:                        23:50:27   BIC                           1297.524
Sample:                             0   HQIC                          1295.420
                                - 251
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2        10.2793      0.419     24.555      0.000       9.459      11.100
===================================================================================
Ljung-Box (L1) (Q):                   0.33   Jarque-Bera (JB):               656.58
Prob(Q):                              0.56   Prob(JB):                         0.00
Heteroskedasticity (H):               1.37   Skew:                            -0.43
Prob(H) (two-sided):                  0.15   Kurtosis:                        10.89
===================================================================================
```





Results:

- **Mean Absolute Error (MAE):** 11.3675
- **Root Mean Squared Error (RMSE):** 3.2061
- **Mean Absolute Percentage Error (MAPE):** 4.85%

- ARIMA forecasts a stable trend but does not react well to sudden market movements.
- The predicted values (orange dashed line) remain relatively flat compared to actual stock movements.

---

4. Model Comparison

Model	MAE	RMSE	MAPE	Accuracy
Gradient Boosting (GBR)	**5.3895**	**11.7057**	**1.62%**	**98.38%**
ARIMA	11.3675	**3.2061**	4.85%	N/A

Key Insights:

- **GBR outperforms ARIMA in overall accuracy (98.38%).**
- **GBR has a lower MAE (5.38 vs. 11.36), meaning its predictions are closer to actual values.**
- **ARIMA has a lower RMSE (3.20), indicating a smoother forecast but lacks responsiveness to price fluctuations.**
- **MAPE for GBR (1.62%) is significantly lower than ARIMA's (4.85%), proving that GBR is more reliable for short-term forecasting.**

---

5. Implications for Trading Strategies

5.1 Trading with Gradient Boosting (GBR)

- **Best suited for short-term trading** due to its ability to capture price fluctuations.
- Can be used in **intraday trading** and **swing trading strategies**.
- **Potential strategy:** Use GBR predictions to **set stop-loss and take-profit levels** for trade execution.

5.2 Trading with ARIMA

- More suitable for **long-term investors** looking for general market trends.
- Works well in **momentum trading** by identifying uptrends and downtrends.
- **Potential strategy:** Use ARIMA forecasts for **trend-following investment strategies**.

5.3 Recommendation

- **For active traders:** GBR is the better choice due to its higher accuracy and responsiveness.
- **For long-term investors:** ARIMA can provide useful insights but lacks precision for short-term movements.

A **hybrid approach** combining both models can provide a more robust trading system.

---

6. Conclusion

This study demonstrates that **machine learning models like Gradient Boosting outperform traditional time series models like ARIMA in stock price prediction**. The key findings are:

- **GBR achieves 98.38% accuracy, making it highly reliable for stock price forecasting.**
- **ARIMA provides a stable long-term forecast but fails to capture short-term market fluctuations.**
- **GBR is recommended for short-term trading, while ARIMA is better suited for long-term trend analysis.**
- **Combining both models in a hybrid strategy may enhance trading decisions.**

Final Takeaway:

Machine learning techniques such as **Gradient Boosting** offer superior performance for stock price prediction and can be effectively used in trading strategies. **Future work** could involve incorporating deep learning models like LSTMs or reinforcement learning for even better predictions.

7. Summary and Future Work

7.1 Summary

This study compared two stock price prediction models: **Gradient Boosting Regression (GBR) and ARIMA**. Key findings include:

- **GBR achieved higher accuracy (98.38%) and lower error rates, making it ideal for short-term forecasting.**
- **ARIMA provided a smooth forecast but struggled with short-term fluctuations.**

- **A hybrid approach combining GBR's short-term accuracy with ARIMA's long-term trend analysis can improve trading strategies.**

Overall, machine learning models like **GBR outperform traditional statistical methods like ARIMA for stock price prediction** due to their ability to capture complex market patterns.

7.2 Future Work

To further improve prediction accuracy and robustness, the following enhancements can be explored:

1. Incorporating Deep Learning Models

- **Long Short-Term Memory (LSTM)** networks can capture sequential dependencies in stock price data.
- **Transformer-based models** like Temporal Fusion Transformers (TFT) can enhance time-series forecasting.

2. Hybrid Modeling

- Combine **GBR, ARIMA, and LSTM** to leverage the strengths of each model.
- Use **ensemble methods** to blend multiple forecasts for improved accuracy.

3. Feature Engineering Improvements

- Introduce **macro-economic indicators** (inflation, interest rates, etc.) to improve forecasting.
- Incorporate **sentiment analysis** from financial news and social media.

4. Reinforcement Learning for Trading Strategies

- Develop **automated trading bots** that use **Q-learning** or **Deep Reinforcement Learning (DRL)** to optimize trade execution.

5. Real-time Prediction System

- Deploy models in a **real-time environment** using **streaming data** for live trading signals.
- Implement a **dashboard** for monitoring stock predictions dynamically.

Final Thought

Stock price prediction remains an evolving field, and **leveraging machine learning with domain expertise** can lead to powerful trading strategies. **Integrating deep learning, alternative data sources, and reinforcement learning** could be the next step toward building an **intelligent stock forecasting system**.