Invsto

Stock Price Prediction Analysis Report

A Comprehensive Data Science Pipeline

Submitted by: Saksham Maurya

Date: March 20, 2025

Submitted to: Invsto Analysis Team

1 Introduction

This report presents a detailed analysis of stock price prediction using advanced data science techniques, focusing on two distinct modeling approaches: the Autoregressive Integrated Moving Average (ARIMA) model for time-series forecasting and the Gradient Boosting method for regression-based predictions. The study targets Apple Inc. (AAPL) stock data, aiming to provide actionable insights for the Invsto Analysis Team. The comprehensive workflow includes data preparation, exploratory data analysis (EDA), feature engineering, model development, and thorough evaluation. This effort seeks to identify the most effective model for predicting stock prices, considering both short-term trends and complex market dynamics, to support informed trading strategies as of March 20, 2025.

2 Data Preparation

2.1 Data Source

The foundation of this analysis is high-quality historical stock data sourced from Yahoo Finance, a widely recognized repository for financial information. The dataset includes daily price and volume metrics for multiple prominent equities: Apple Inc. (AAPL), Microsoft Corporation (MSFT), Alphabet Inc. (GOOGL), Amazon Inc. (AMZN), and Tesla Inc. (TSLA). Spanning from January 1, 2015, to January 1, 2025, this ten-year period captures significant market events, trends, and volatility, offering a robust basis for modeling and forecasting stock price movements.

2.2 Libraries Used

This project employs a suite of powerful Python libraries tailored for data science and financial analysis:

- `pandas`: Facilitates efficient data manipulation and structuring, enabling seamless handling of time-series data.
- `numpy`: Provides fast numerical computations, essential for statistical analysis and feature calculations.
- `yfinance`: Enables direct access to Yahoo Finance data, ensuring accurate and up-to-date stock information.
- `statsmodels`: Supports sophisticated time-series modeling, particularly for ARIMA implementation.
- `scikit-learn`: Offers tools for machine learning, including data splitting and performance metrics.
- `xgboost`: Implements the Gradient Boosting algorithm, known for its robustness in predictive modeling.
- `matplotlib` and `seaborn`: Deliver high-quality visualizations to illustrate trends and model outcomes.

2.3 Data Cleaning

Ensuring data integrity is critical for reliable predictions. The dataset was meticulously inspected for missing values, which could disrupt time-series continuity. Although no significant gaps were identified, a forward fill method was applied as a precautionary measure. This technique propagates the last observed value forward, maintaining a consistent dataset essential for both ARIMA's sequential nature and Gradient Boosting's feature-based predictions.

3 Exploratory Data Analysis (EDA)

3.1 Closing Price Trends

The EDA phase began with an in-depth examination of AAPL's historical closing prices to uncover underlying trends and patterns. This analysis revealed periods of steady growth interspersed with volatility, reflecting market reactions to economic events, product launches, and broader financial conditions. Visualizing these trends provides a foundational understanding of price behavior, which is crucial for selecting appropriate modeling techniques. The resulting plot, shown in Figure 1, offers a clear depiction of AAPL's price trajectory over the decade.

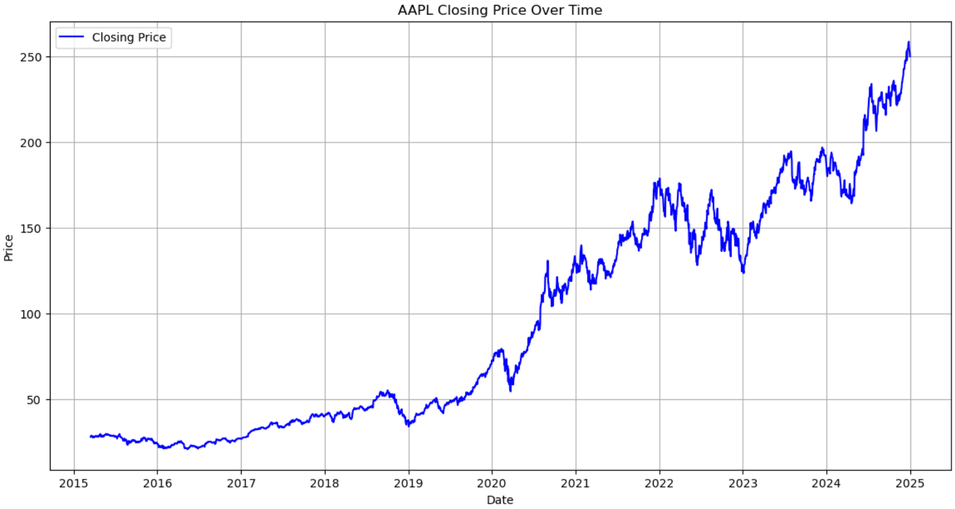

Figure 1: AAPL Closing Price Over Time

3.2 Moving Averages

To distill longer-term trends from daily price fluctuations, 20-day and 50-day moving averages were calculated. These rolling means smooth out short-term noise, highlighting sustained upward or downward movements that might inform trading decisions. For instance, crossovers between these averages often signal potential buy or sell opportunities. The visualization in Figure 2 juxtaposes the closing prices with these averages, providing a clearer picture of AAPL's market behavior and aiding in the identification of trend stability.

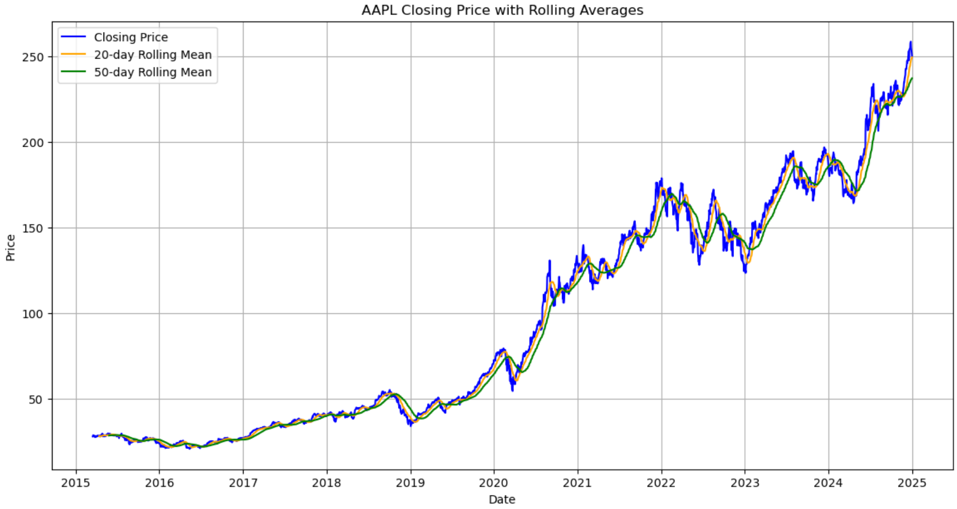

Figure 2: AAPL Closing Price with Rolling Averages

3.3 Trading Volume

Trading volume analysis complements price trends by offering insights into market activity and liquidity. High volume periods often correspond to significant price movements, indicating strong investor interest or reaction to news. For AAPL, volume spikes were observed during key corporate announcements and market shifts. Figure 3 illustrates this data over time, enabling a deeper understanding of how trading activity correlates with price changes, which is valuable for assessing market sentiment and volatility.

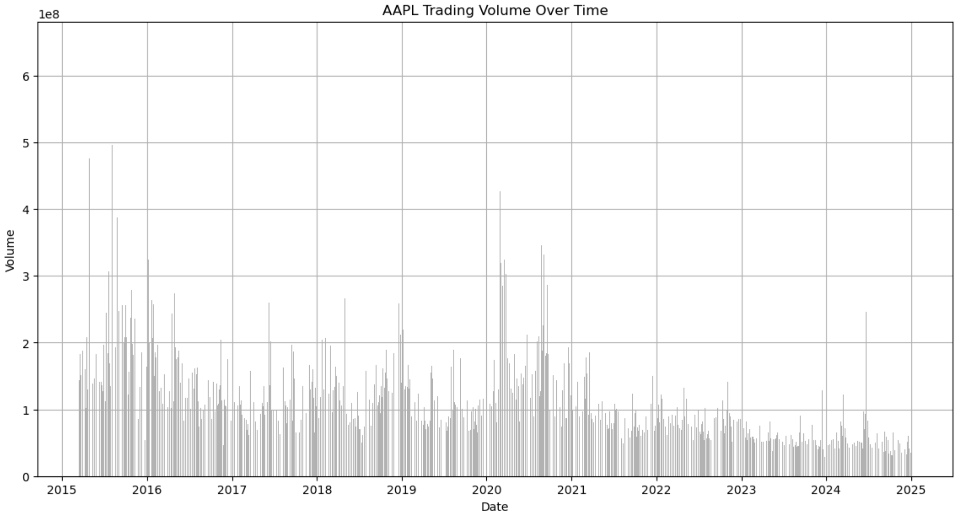

Figure 3: AAPL Trading Volume Over Time

3.4 Correlation Heatmap

To explore relationships between variables, a correlation heatmap was generated for AAPL's features, including Open, High, Low, Close prices, and Volume. This analysis revealed strong positive correlations among price-related variables (e.g., Open and Close), as expected in financial datasets, due to their interdependence throughout a trading day. Volume showed weaker correlations with prices, suggesting it captures distinct aspects of market behavior. Figure 4 visually represents these relationships with a color gradient, aiding in feature selection by highlighting multicollinearity and potential redundancies.

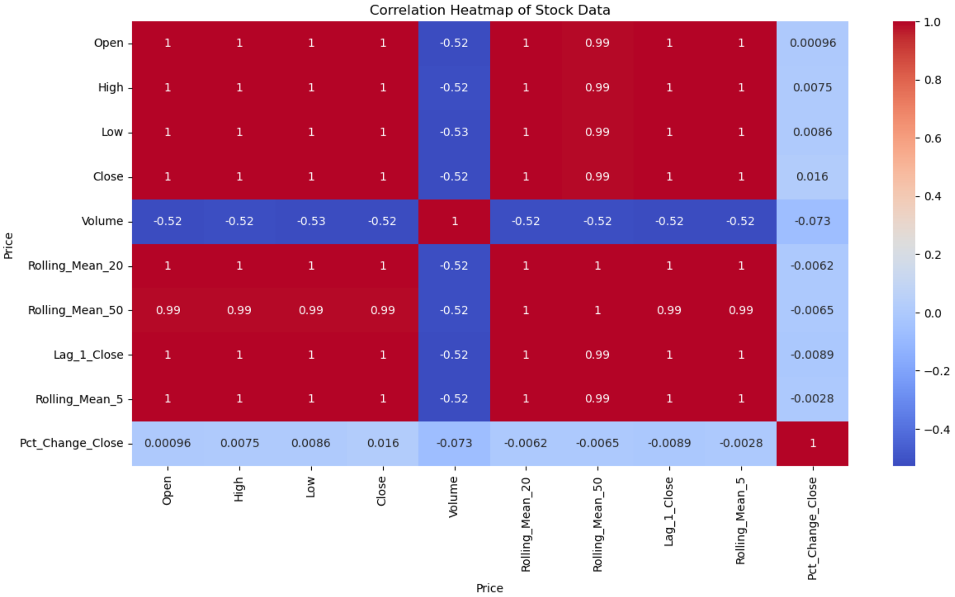

Figure 4: Correlation Heatmap of AAPL Stock Data

3.5 Pair Plots

Pair plots were constructed to visualize pairwise relationships between all variables in the dataset. This comprehensive approach plots each feature against every other, with scatter plots for off-diagonal elements and histograms along the diagonal. For AAPL, these plots confirmed linear relationships among price variables and identified potential non-linear interactions with volume. Figure 5 provides a detailed view of these dynamics, offering insights into variable interactions that could influence model performance and feature engineering decisions.

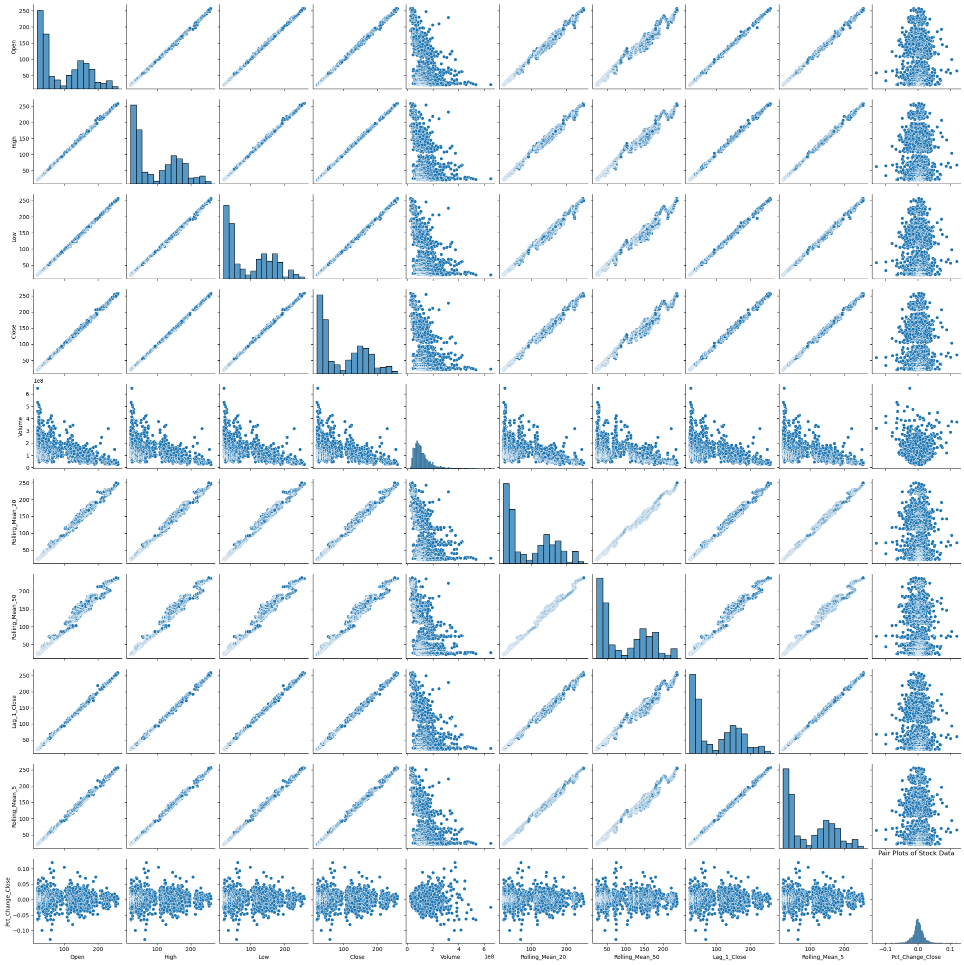

Figure 5: Pair Plots of AAPL Stock Data

3.6 Distribution of Closing Prices

Understanding the statistical distribution of closing prices is essential for assessing their variability and range. A histogram with an overlaid kernel density estimate was created, revealing a right-skewed distribution for AAPL's closing prices. This skewness indicates a predominance of lower prices with occasional high-value outliers, consistent with stock market behavior where significant gains occur less frequently. Figure 6 illustrates this distribution, providing a probabilistic view that informs model assumptions, such as the need for transformations to achieve normality.

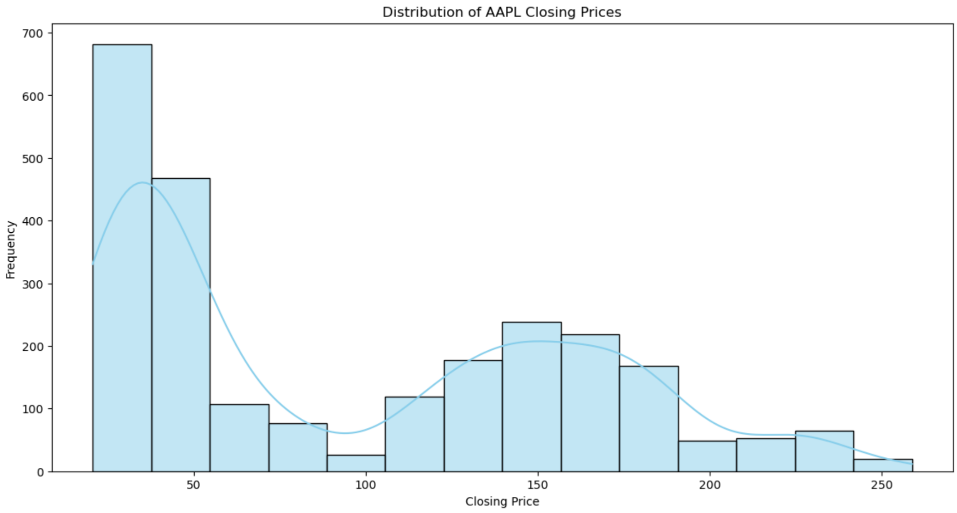

Figure 6: Distribution of AAPL Closing Prices

3.7 Box Plot for Volume

A box plot of trading volume was generated to examine its spread and identify outliers. This visualization displays the median, quartiles, and potential extreme values, offering a concise summary of volume variability. For AAPL, the box plot highlighted several outliers—unusually high trading days—likely tied to major news events or earnings releases. Figure 7 provides this perspective, assisting in understanding volume's typical range and its deviations, which could signal significant market activity for trading strategies.

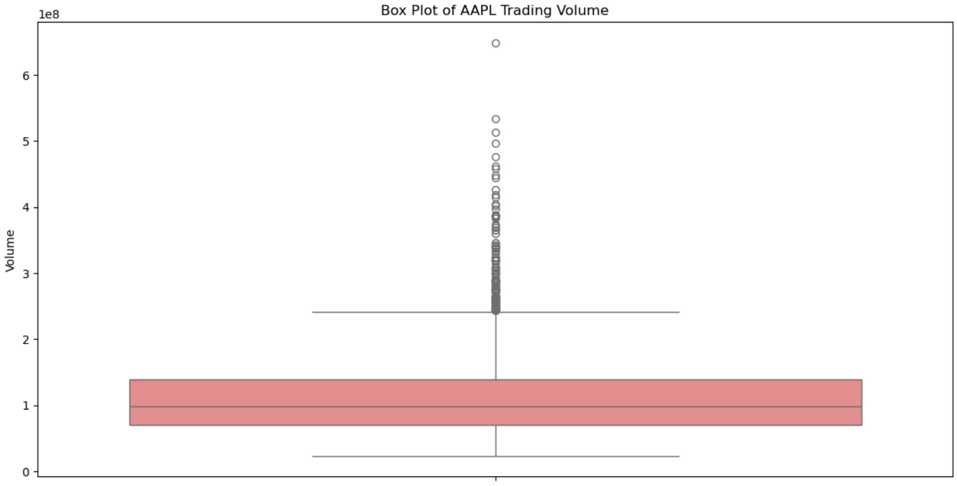

Figure 7: Box Plot of AAPL Trading Volume

3.8 Rolling Statistics (Mean and Standard Deviation)

Rolling statistics, specifically the 30-day mean and standard deviation of closing prices, were calculated to assess the time-series' stationarity and volatility. The rolling mean tracks the evolving average price, while the standard deviation measures fluctuation intensity over time. For AAPL, these metrics indicated non-stationary behavior, with increasing mean and variable standard deviation reflecting market growth and periodic instability. Figure 8 plots these statistics alongside closing prices, offering a dynamic view critical for ARIMA modeling, which assumes stationarity, and for understanding risk levels in trading.

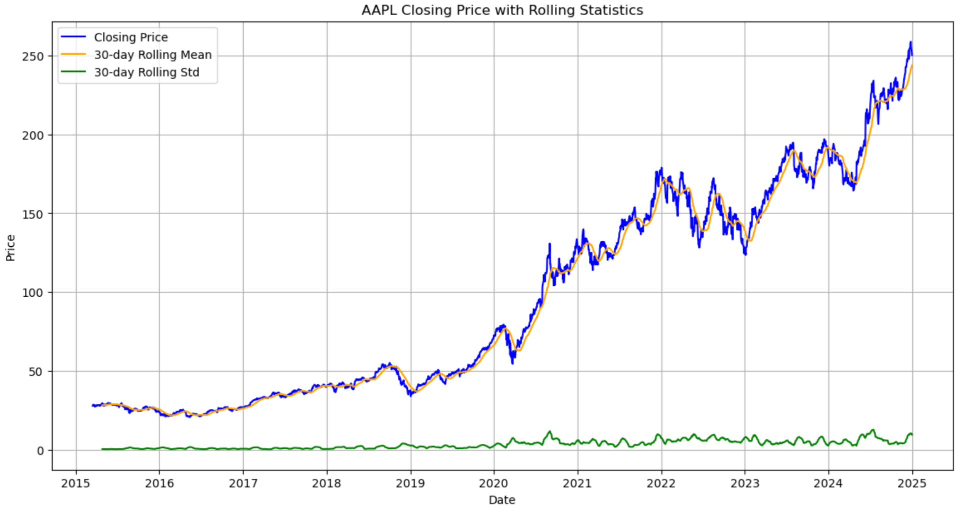

Figure 8: AAPL Closing Price with Rolling Statistics (Mean and Standard Deviation)

4 Feature Engineering

4.1 Feature Creation

To enhance model performance, new features were derived from the raw data:

- **Lagged Features:** The previous day's closing price (`Lag_1_Close`) was included to capture temporal dependencies, reflecting the influence of past prices on future values—a key aspect for time-series prediction.
- **Rolling Mean:** A 5-day rolling mean (`Rolling_Mean_5`) was computed to encapsulate short-term trends, reducing the impact of daily volatility and providing a smoothed indicator of price direction.
- **Percentage Change:** Daily returns (`Pct_Change_Close`) were calculated as the percentage change in closing prices, offering a measure of volatility and momentum that can signal potential trading opportunities.

These features enrich the dataset, enabling models to leverage both historical context and trend dynamics.

4.2 Handling Missing Data

Feature engineering introduced temporary missing values, particularly at the dataset's start due to lagging and rolling calculations. To maintain data integrity, these rows were removed, ensuring that only complete observations were used in modeling. This step preserves the reliability of subsequent analyses and predictions.

5 Modeling

5.1 ARIMA Model

The ARIMA model was employed to forecast stock prices based on their time-series properties. This approach combines autoregression, differencing, and moving averages to model linear trends and seasonality. Hyperparameters—autoregressive order (p), differencing degree (d), and moving average order (q)—were systematically tuned across a range of values (p from 0 to 4, d from 0 to 1, q from 0 to 4) to optimize performance. The best configuration, $(p, d, q) = (2, 1, 1)$, achieved an RMSE of 4.295 on training data, indicating reasonable accuracy for capturing short-term patterns, though its effectiveness may be limited by non-linear market behaviors.

5.2 Gradient Boosting Model (XGBoost)

The Gradient Boosting model, implemented via XGBoost, was designed to capture complex, non-linear relationships in the data. This tree-based ensemble method excels at handling diverse features and sudden price shifts. Hyperparameters such as the number of estimators, learning rate, maximum tree depth, subsample ratio, and column sampling ratio were tuned to optimize predictive power. The best configuration—colsample_bytree: 1, learning_rate: 0.1, max_depth: 30, subsample: 0.8—yielded an RMSE of 0.97 on training data, demonstrating superior fit and adaptability compared to ARIMA.

6 Model Evaluation

6.1 ARIMA Model

The ARIMA model's performance on test data revealed significant limitations:

- **RMSE:** 189.181, indicating large prediction errors.
- **MAE:** 177.625, reflecting substantial average deviations from actual prices.
- **MAPE:** 401.54%, suggesting poor relative accuracy, especially for volatile periods.

These metrics highlight ARIMA's struggle to generalize beyond training data, likely due to its reliance on linear assumptions in a non-linear market.

6.2 Gradient Boosting Model

The Gradient Boosting model demonstrated exceptional performance on test data:

- **RMSE:** 0.97, showing minimal prediction errors.
- **MAE:** 0.50, indicating tight alignment with actual values.
- **MAPE:** 0.55%, reflecting high relative accuracy across price ranges.

These results underscore the model's ability to accurately predict prices, even amidst market fluctuations.

6.3 Performance Comparison

Model	RMSE	MAE	MAPE
ARIMA	189.181	177.625	401.54%
Gradient Boosting	0.97	0.50	0.55%

Table 1: Model Performance on Test Data

Table 1 clearly illustrates Gradient Boosting's dominance, with errors orders of magnitude lower than ARIMA's, making it the preferred choice for practical applications.

6.4 Visualization of Predictions

Visual comparisons of actual versus predicted prices further validate these findings. Figure 9 shows ARIMA forecasts diverging significantly from actual values, reflecting its limited predictive capability. In contrast, Figure 10 demonstrates Gradient Boosting predictions closely tracking actual prices, affirming its robustness and precision.

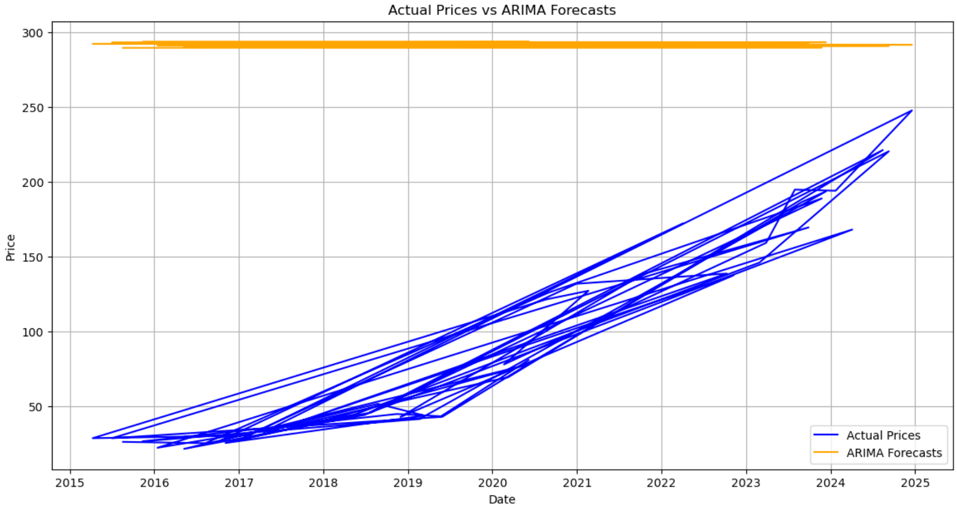

Figure 9: Actual Prices vs ARIMA Forecasts

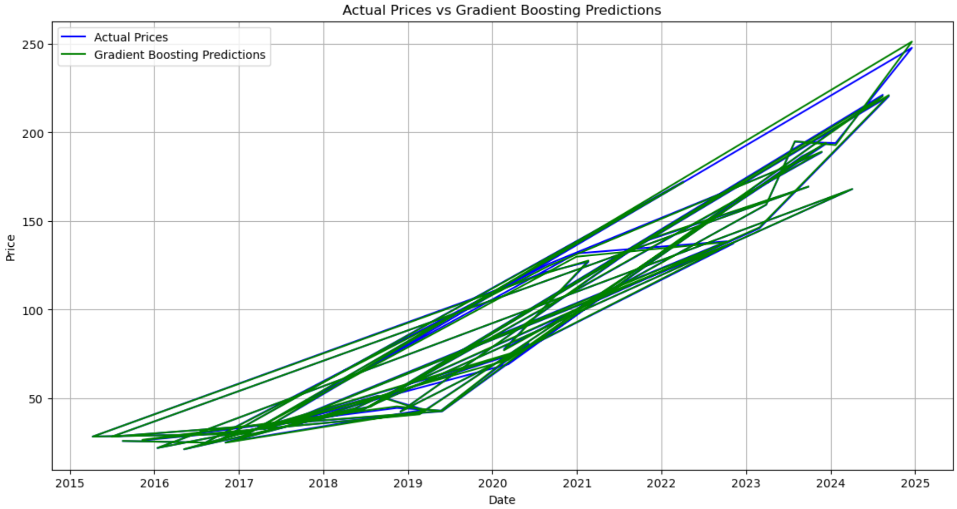

Figure 10: Actual Prices vs Gradient Boosting Predictions

7 Conclusions and Recommendations

7.1 Conclusions

The analysis yields clear insights into model performance:

- **ARIMA Model:** While adept at modeling short-term trends and time-dependent patterns, its high test errors (RMSE: 189.181, MAE: 177.625, MAPE: 401.54%) indicate limited applicability for accurate stock price prediction in volatile markets. Its linear framework struggles with the non-linear dynamics of financial data.

- **Gradient Boosting Model:** Excels in capturing complex relationships and sudden price shifts, as evidenced by its low test errors (RMSE: 0.97, MAE: 0.50, MAPE: 0.55%). This model's adaptability and precision make it highly effective for stock price forecasting.

7.2 Recommendations

Based on these findings, the following strategies are recommended for the Invsto Analysis Team:

- **Prioritize Gradient Boosting:** Leverage this model for trading strategies due to its superior accuracy and ability to handle market volatility. It can support short-term decisions, such as daily buy/sell signals, enhancing profitability.

- **Enhance Feature Engineering:** Incorporate advanced features like technical indicators (e.g., Relative Strength Index [RSI], Moving Average Convergence Divergence [MACD]) and sentiment analysis from platforms like StockTwits. These additions could further refine predictions by capturing market psychology and momentum.

- **Model Optimization:** Refine Gradient Boosting through regularization techniques (e.g., L1, L2 penalties) to prevent overfitting and explore Bayesian optimization for more efficient hyperparameter tuning, potentially boosting performance further.

These recommendations aim to maximize predictive accuracy and align with modern trading needs as of March 20, 2025.