Stock Price Prediction using Time Series Forecasting

Report by

ALEKHYA CHATTERJEE

Introduction

Problem Statement

Accurately predicting stock prices is crucial for making informed investment decisions. However, stock prices are influenced by numerous factors such as economic conditions, market sentiment, and company performance. Traditional forecasting methods often fail to account for these complexities, leading to unreliable predictions. This project aims to address this challenge by implementing machine learning and time-series forecasting models to predict stock prices.

Objective

The primary goal of this study is to analyze historical stock price data and develop predictive models using **ARIMA (AutoRegressive Integrated Moving Average)** and **XGBoost (Extreme Gradient Boosting)**. The project will:

- Perform exploratory data analysis and feature engineering.
- Apply ARIMA and XGBoost to forecast stock prices.
- Evaluate model performance using **RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R2 SCORE**.
- Compare the effectiveness of statistical and machine learning approaches.
- Provide insights into potential trading strategies based on the findings.

Scope of the Report

This report is structured as follows:

- **Data Collection & Preprocessing**: Gathering historical stock price data and cleaning it for analysis.
- **Feature Engineering**: Creating new features such as lagged returns, moving averages, and volume changes.
- **Model Development**: Implementing ARIMA and XGBoost models for stock price forecasting.
- **Model Evaluation & Comparison**: Analyzing the performance of both models using standard error metrics.
- **Findings & Trading Strategy Implications**: Discussing the results and their impact on investment decisions.

Significance of the Study

The study will help investors, traders, and financial analysts understand the strengths and limitations of statistical and machine learning models in stock price prediction. By comparing ARIMA and XGBoost, this research will provide insights into choosing the most reliable model for short-term and long-term forecasting.

Methodology

1. Data Collection

The dataset for this study was sourced from **yfinance** API, including historical stock price data for selected companies. The data included key financial metrics such as **closing price, trading volume, and other relevant indicators** over a 1 year time period.

2. Data Preprocessing

To ensure high-quality inputs for modeling, the following preprocessing steps were applied:

- **Handling Missing Values:** Missing data points were identified and either imputed using forward-fill techniques or removed if necessary.
- **Feature Engineering:** New features such as **lagged returns, moving averages, and percentage changes** were created to enhance predictive power.
- **Stationarity Check:** The **ADF and KPSS tests** were conducted to verify stationarity, and **first-order differencing** was applied to non-stationary series.
- **Train-Test Split:** The dataset was split into **80% training and 20% testing sets** for model evaluation.

3. ARIMA Model Development

The ARIMA (AutoRegressive Integrated Moving Average) model was used for time-series forecasting:

- **Parameter Selection:** The **ACF and PACF plots and Grid search Results** were analyzed to determine optimal (p, d, q) values.
- **Model Training:** The ARIMA model was trained using the training dataset, and predictions were made on the test set.

4. XGBoost Model Development

A Gradient Boosting approach using **XGBoost** was applied for predictive modeling:

- **Feature Selection:** Relevant features such as **lagged prices, moving averages, and volume changes** were used.

- **Hyperparameter Tuning: Optuna** was used to optimize hyperparameters for improved performance.
- **Model Training:** The model was trained on the training set and evaluated on the test set.

5. Model Evaluation

Both models were assessed using the following metrics:

- **Root Mean Squared Error (RMSE)**
- **Mean Absolute Error (MAE)**
- **R^2 Score (for XGBoost model)**

A **comparative analysis** was conducted to determine which model provided the most accurate and reliable stock price predictions.

Exploratory Data Analysis (EDA)

1. Summary Statistics

To understand the fundamental characteristics of the dataset, we first computed summary statistics for each stock. These statistics include measures such as **mean, median, standard deviation, minimum, and maximum values** for key features like closing price and volume.

Summary Statistics for AAPL:

Price	Close	High	Low	Open	Volume
Ticker	AAPL	AAPL	AAPL	AAPL	AAPL
count	250.000000	250.000000	250.000000	250.000000	2.500000e+02
mean	217.733153	219.735418	215.446490	217.478133	5.538131e+07
std	23.698061	23.843097	23.304666	23.650546	3.036794e+07
min	164.224548	165.617963	163.308874	164.572913	2.323470e+07
25%	209.257889	213.492477	207.966055	209.961835	4.047068e+07
50%	224.504593	226.366475	222.522789	224.464611	4.822680e+07
75%	232.820576	234.570783	229.482695	232.752745	6.041002e+07
max	258.735504	259.814335	257.347047	257.906429	3.186799e+08

Summary Statistics for MSFT:

Price	Close	High	Low	Open	Volume
Ticker	MSFT	MSFT	MSFT	MSFT	MSFT
count	250.000000	250.000000	250.000000	250.000000	2.500000e+02
mean	420.156811	423.673314	416.416669	420.266144	2.057371e+07
std	16.989484	16.613416	16.977067	16.789476	7.500249e+06
min	378.769989	385.220001	376.910004	379.000000	7.164500e+06
25%	410.525772	414.022701	407.205190	410.773451	1.627195e+07
50%	418.478348	422.753401	415.122228	419.120264	1.910730e+07
75%	429.462509	431.753317	425.053223	429.681908	2.285900e+07
max	464.854340	465.639777	461.772294	464.297590	6.426370e+07

Summary Statistics for GOOGL:

Price	Close	High	Low	Open	Volume
Ticker	GOOGL	GOOGL	GOOGL	GOOGL	GOOGL
count	250.000000	250.000000	250.000000	250.000000	2.500000e+02
mean	172.380429	174.204888	170.565623	172.336955	2.714162e+07
std	13.053964	13.208105	12.927557	13.000695	1.074299e+07
min	148.319000	149.664464	146.882294	148.410454	1.024210e+07
25%	162.910454	165.028495	161.722779	163.265981	2.025885e+07
50%	170.720444	172.907862	168.361029	170.328832	2.412560e+07
75%	181.419346	183.521474	180.377883	181.922379	3.118368e+07
max	206.142593	206.811821	202.576693	203.156027	7.037390e+07

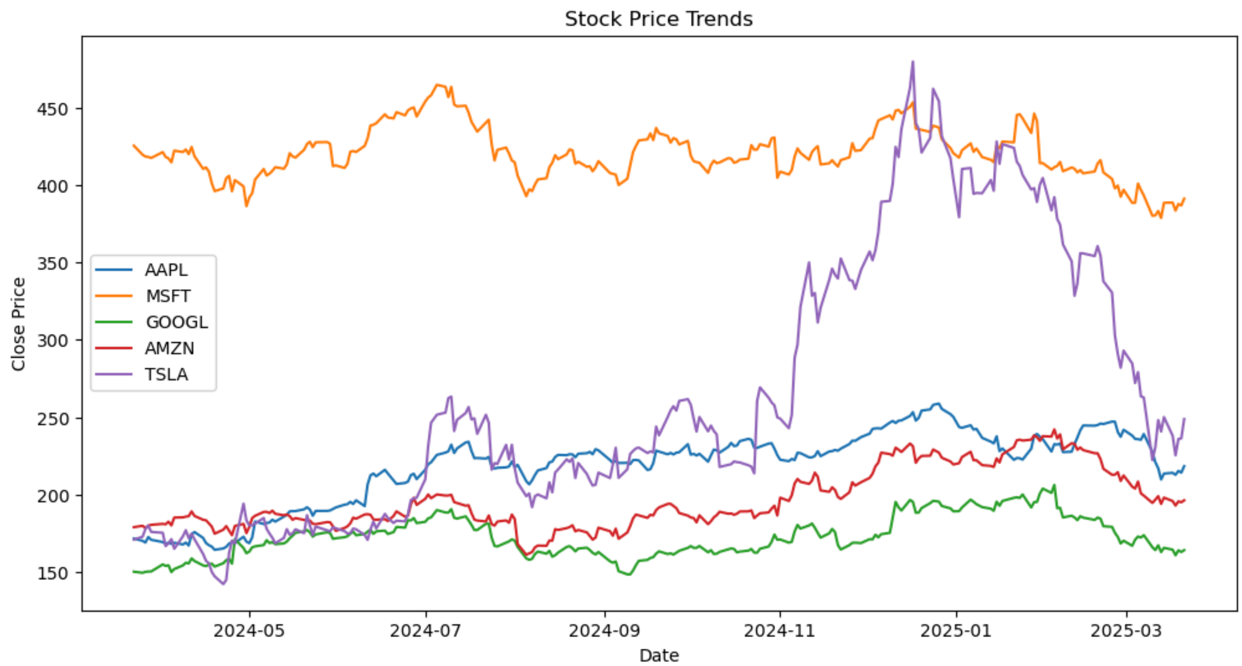
Summary Statistics for AMZN:

Price	Close	High	Low	Open	Volume
Ticker	AMZN	AMZN	AMZN	AMZN	AMZN
count	250.000000	250.000000	250.000000	250.000000	2.500000e+02
mean	196.267120	198.360360	193.954520	196.369400	3.976012e+07
std	19.277490	19.380422	19.089038	19.224122	1.575399e+07
min	161.020004	162.960007	151.610001	154.210007	1.500750e+07
25%	182.592503	184.777496	180.495003	182.782505	2.994568e+07
50%	188.729996	190.525002	186.394997	188.900002	3.628435e+07
75%	208.867504	212.527496	206.745003	209.212494	4.296842e+07
max	242.059998	242.520004	238.029999	239.020004	1.414484e+08

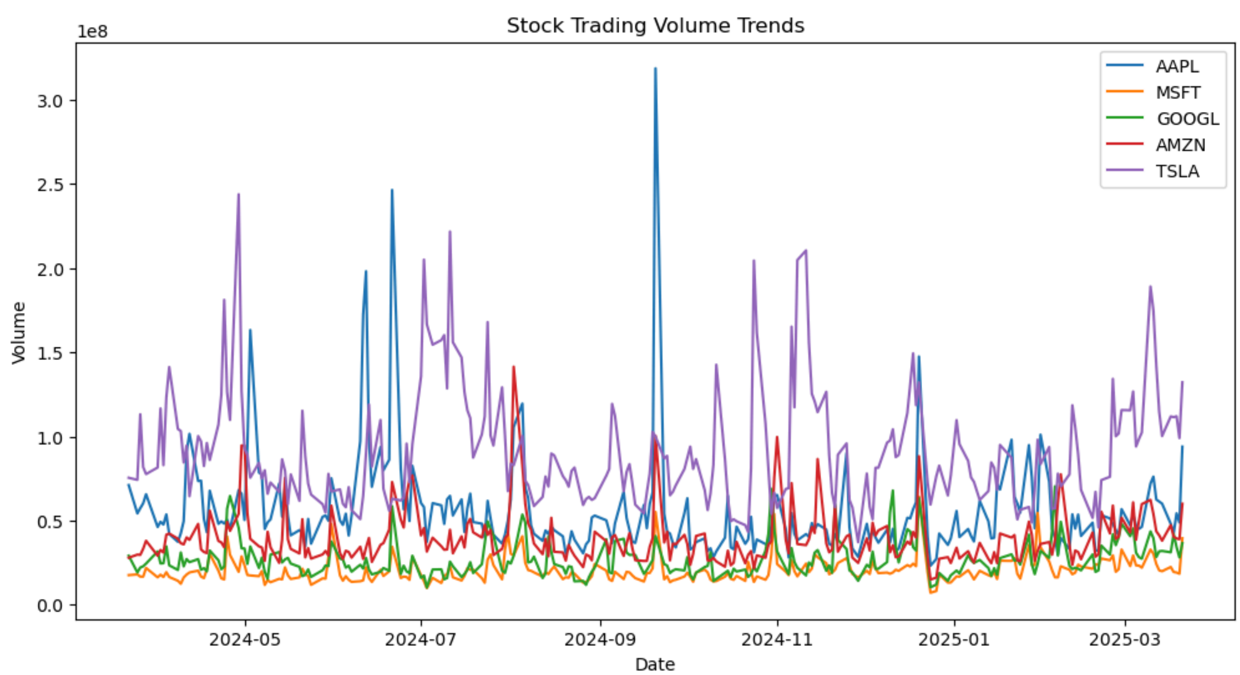
Summary Statistics for TSLA:

Price	Close	High	Low	Open	Volume
Ticker	TSLA	TSLA	TSLA	TSLA	TSLA
count	250.000000	250.000000	250.000000	250.000000	2.500000e+02
mean	261.740920	267.992280	255.597000	262.099400	9.190228e+07
std	85.426252	88.109607	82.783298	85.955553	3.460384e+07
min	142.050003	144.440002	138.800003	140.559998	3.716760e+07
25%	187.372505	190.982498	182.414997	186.544994	6.745688e+07
50%	238.129997	244.115005	232.235001	234.775002	8.291510e+07
75%	335.605011	346.259995	328.487495	340.082512	1.075146e+08
max	479.859985	488.540009	457.510010	475.899994	2.438697e+08

2. Stock Price Trends Over Time

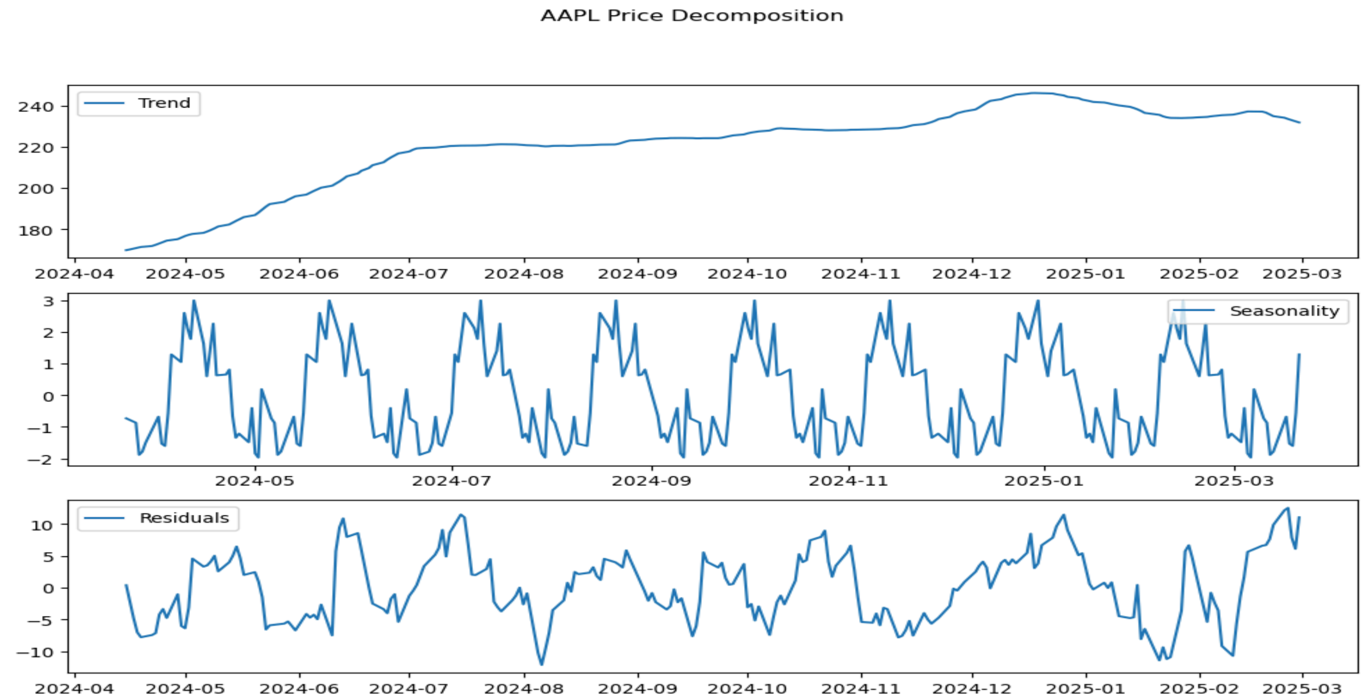


3. Trading Volume Trends Over Time

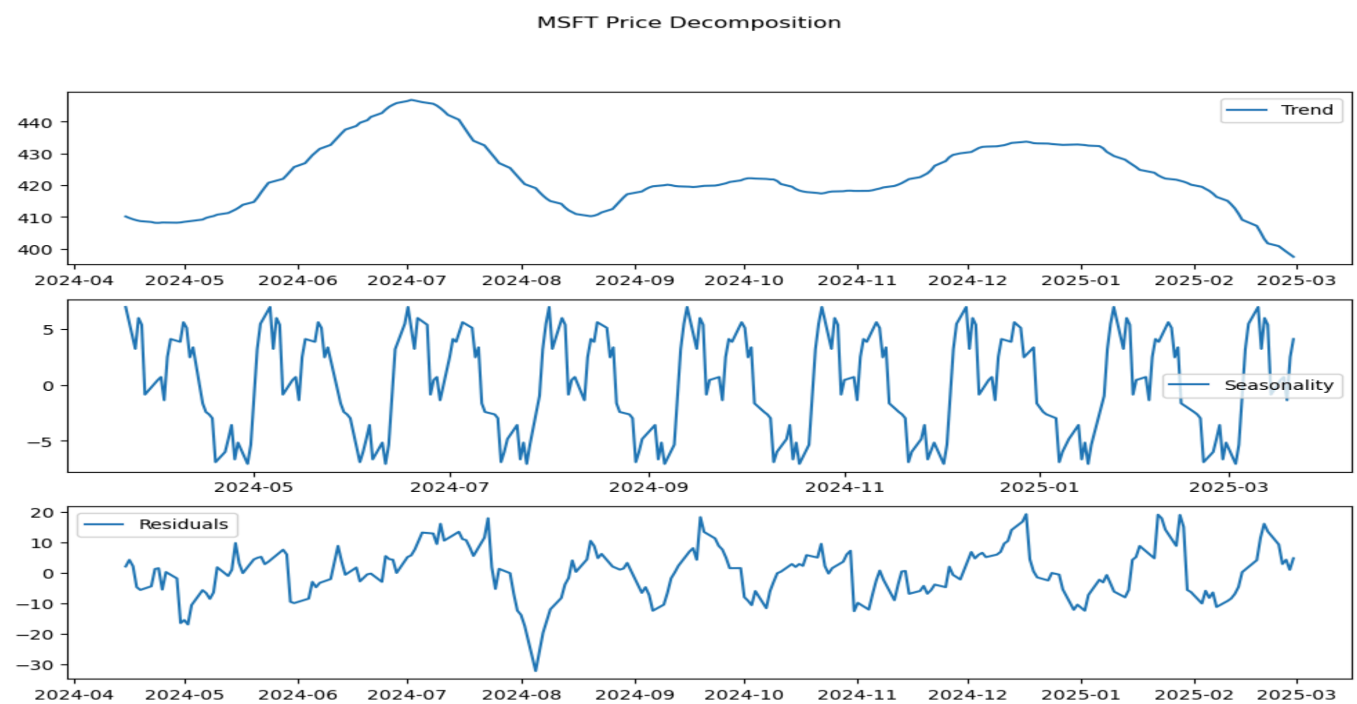


4. Seasonality & Trend Analysis

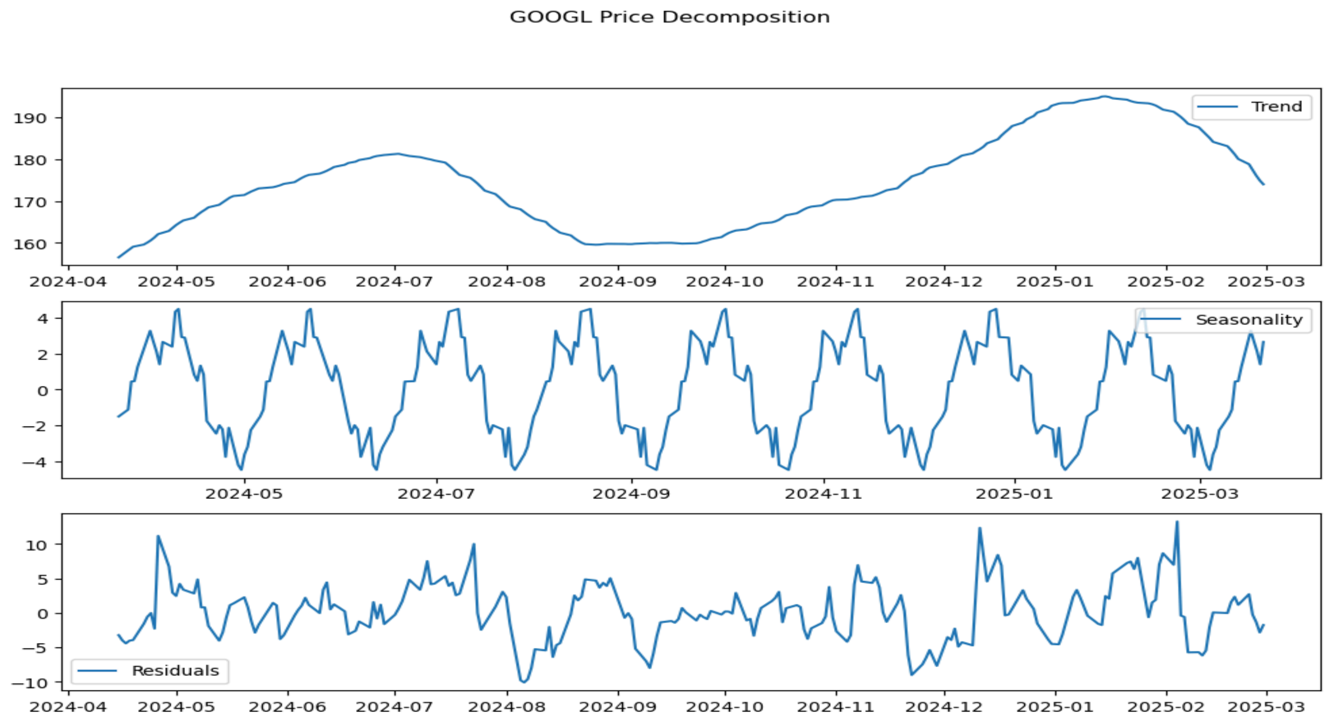
AAPL



MFST

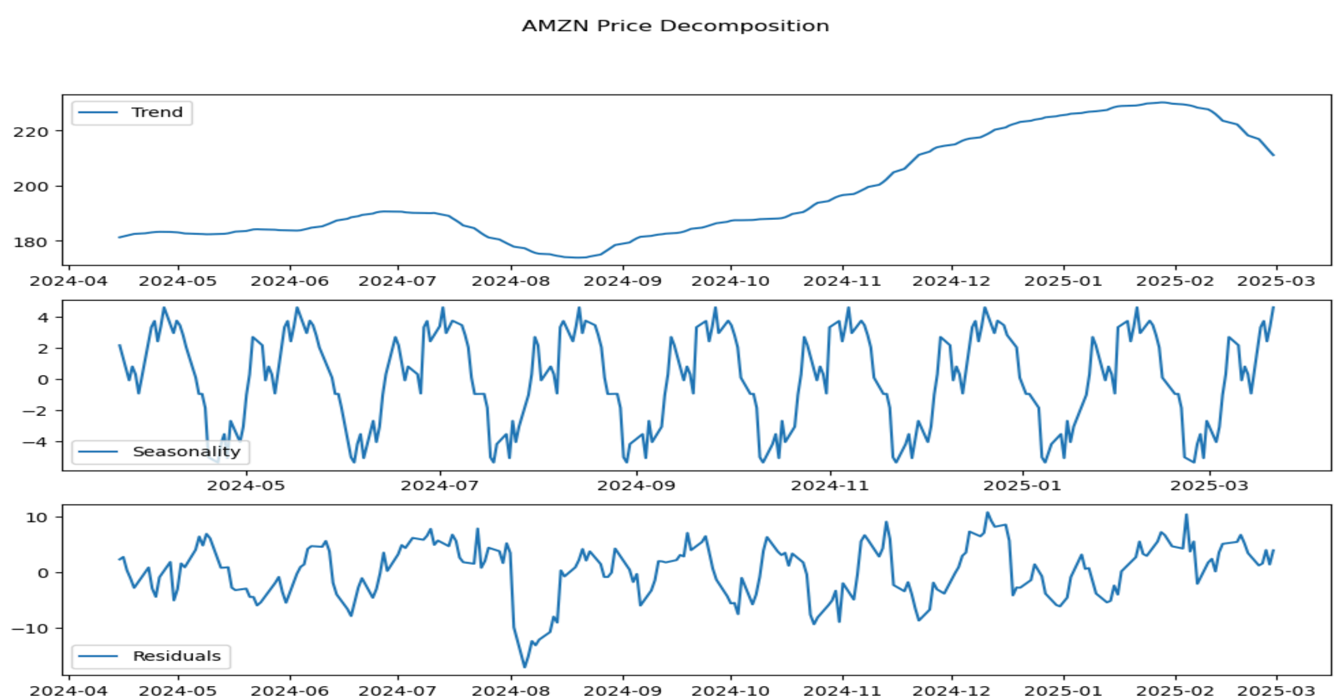


GOOGL



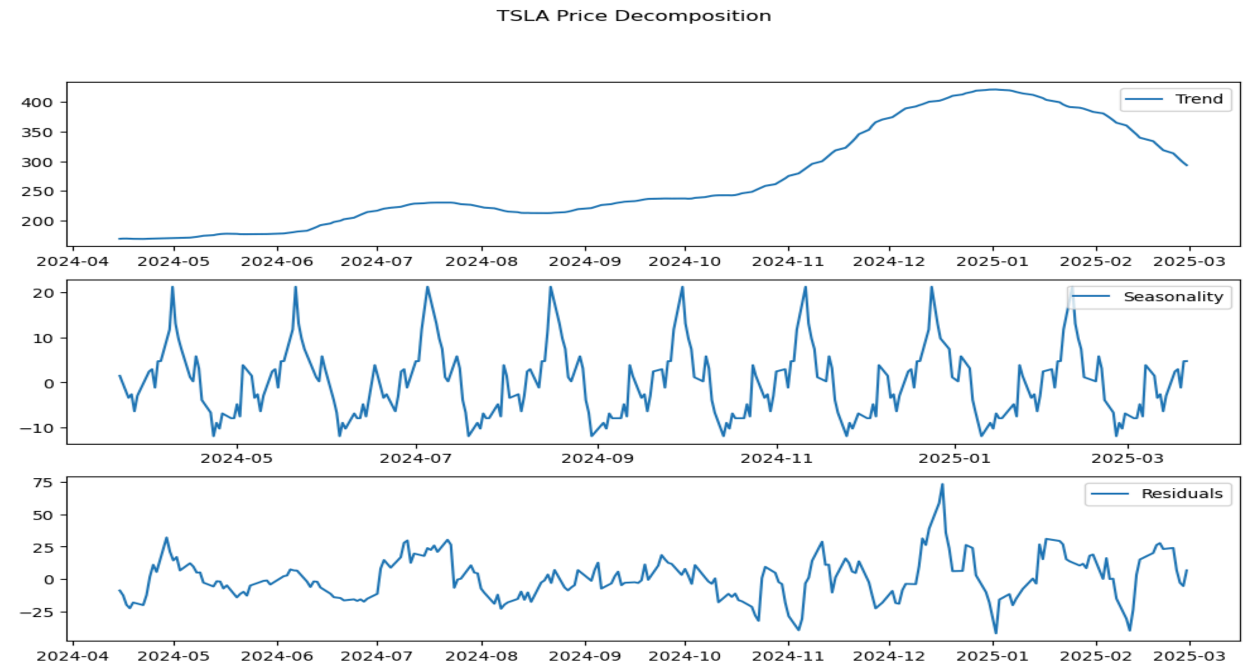
GOOGL Price Decomposition
Trend
Seasonality
Residuals
190
180
170
160
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
2025-01
2025-02
2025-03

AMZN



AMZN Price Decomposition
Trend
Seasonality
Residuals
220
200
180
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
2025-01
2025-02
2025-03

TSLA



5. Stationarity Tests (ADF & KPSS Tests)

To assess stationarity (a key assumption for time-series models like ARIMA), we applied two tests:

- **Augmented Dickey-Fuller (ADF) Test:** Checks for unit roots (if present, data is non-stationary).
- **KPSS Test:** Tests whether a time series is stationary around a deterministic trend.

Findings:

- Most stock prices were **non-stationary**, indicating trends over time.
- After **first-order differencing**, stationarity was achieved, making the data suitable for ARIMA modeling.

Before first-order differencing

```
📊 Stationarity Tests for AAPL

ADF Test Results:
ADF Test: p-value = 0.2957
❌ The series is likely non-stationary.

KPSS Test Results:
KPSS Test: p-value = 0.0100
❌ The series is likely non-stationary.
--------------------------------------------------

📊 Stationarity Tests for MSFT

ADF Test Results:
ADF Test: p-value = 0.1386
❌ The series is likely non-stationary.

KPSS Test Results:
KPSS Test: p-value = 0.1000
✅ The series is likely stationary.
--------------------------------------------------

📊 Stationarity Tests for GOOGL

ADF Test Results:
ADF Test: p-value = 0.1855
❌ The series is likely non-stationary.

KPSS Test Results:
KPSS Test: p-value = 0.0143
❌ The series is likely non-stationary.
--------------------------------------------------

📊 Stationarity Tests for AMZN

ADF Test Results:
ADF Test: p-value = 0.5035
❌ The series is likely non-stationary.

KPSS Test Results:
KPSS Test: p-value = 0.0100
❌ The series is likely non-stationary.
--------------------------------------------------

📊 Stationarity Tests for TSLA

ADF Test Results:
ADF Test: p-value = 0.6159
❌ The series is likely non-stationary.

KPSS Test Results:
KPSS Test: p-value = 0.0100
❌ The series is likely non-stationary.
--------------------------------------------------
```

After first-order differencing

```
📊 Stationarity Tests for AAPL (After Fir

ADF Test Results:
ADF Test: p-value = 0.0000
✅ The series is likely stationary.

KPSS Test Results:
KPSS Test: p-value = 0.1000
✅ The series is likely stationary.
-----------------------------------------

📊 Stationarity Tests for MSFT (After Fir

ADF Test Results:
ADF Test: p-value = 0.0000
✅ The series is likely stationary.

KPSS Test Results:
KPSS Test: p-value = 0.1000
✅ The series is likely stationary.
-----------------------------------------

📊 Stationarity Tests for GOOGL (After Fi

ADF Test Results:
ADF Test: p-value = 0.0000
✅ The series is likely stationary.

KPSS Test Results:
KPSS Test: p-value = 0.1000
✅ The series is likely stationary.
-----------------------------------------

📊 Stationarity Tests for AMZN (After Fir

ADF Test Results:
ADF Test: p-value = 0.0000
✅ The series is likely stationary.

KPSS Test Results:
KPSS Test: p-value = 0.1000
✅ The series is likely stationary.
-----------------------------------------

📊 Stationarity Tests for TSLA (After Fir

ADF Test Results:
ADF Test: p-value = 0.0000
✅ The series is likely stationary.

KPSS Test Results:
KPSS Test: p-value = 0.1000
✅ The series is likely stationary.
-----------------------------------------
```

6. Feature Engineering

To improve predictive accuracy, additional features were created:

- **Lagged Features:** Previous day's closing price (`Lag_1`).
- **Rolling Window Features:**
  - **5-day Moving Average** (`Rolling_Mean`) to capture short-term trends.
  - **5-day Rolling Standard Deviation** (`Rolling_Std`) to measure volatility.
- **Daily Percentage Change (`Pct_Change`)** to capture momentum.

These features were used in machine learning models such as **XGBoost** to enhance prediction accuracy.

Model Development

ARIMA Model

We implemented an **ARIMA (AutoRegressive Integrated Moving Average)** model to forecast stock prices. The ARIMA model was trained using historical closing prices, and the best order (p,d,q) was selected using **ACF/PACF plots** and **GridSearch optimization**.

Model Training

1. **Data Preparation:** The closing prices of Apple Inc. (AAPL) were used for ARIMA training.

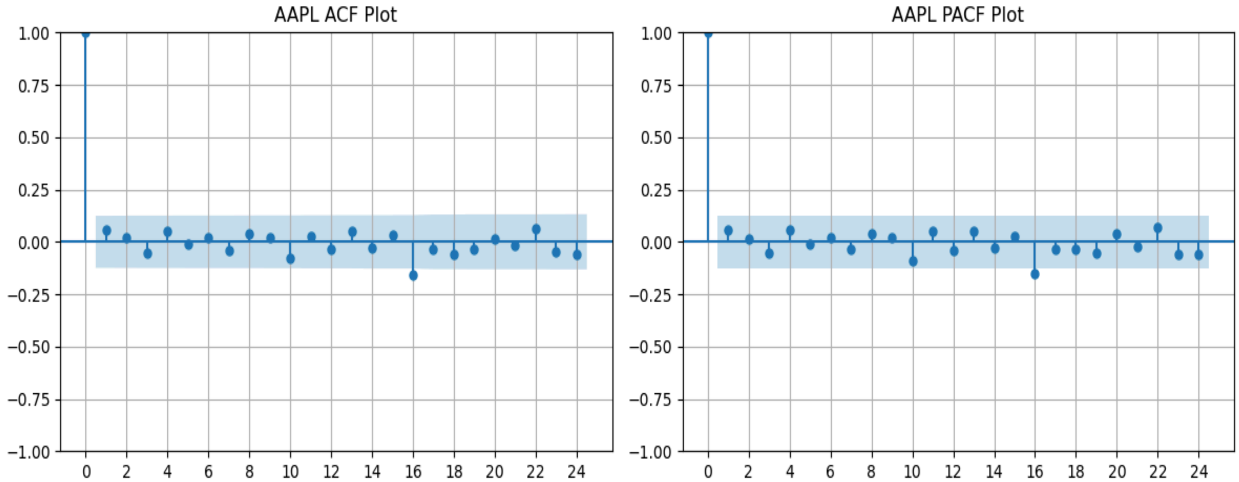
2. **Train-Test Split:** 80% of the data was used for training, while the remaining 20% was used for testing.

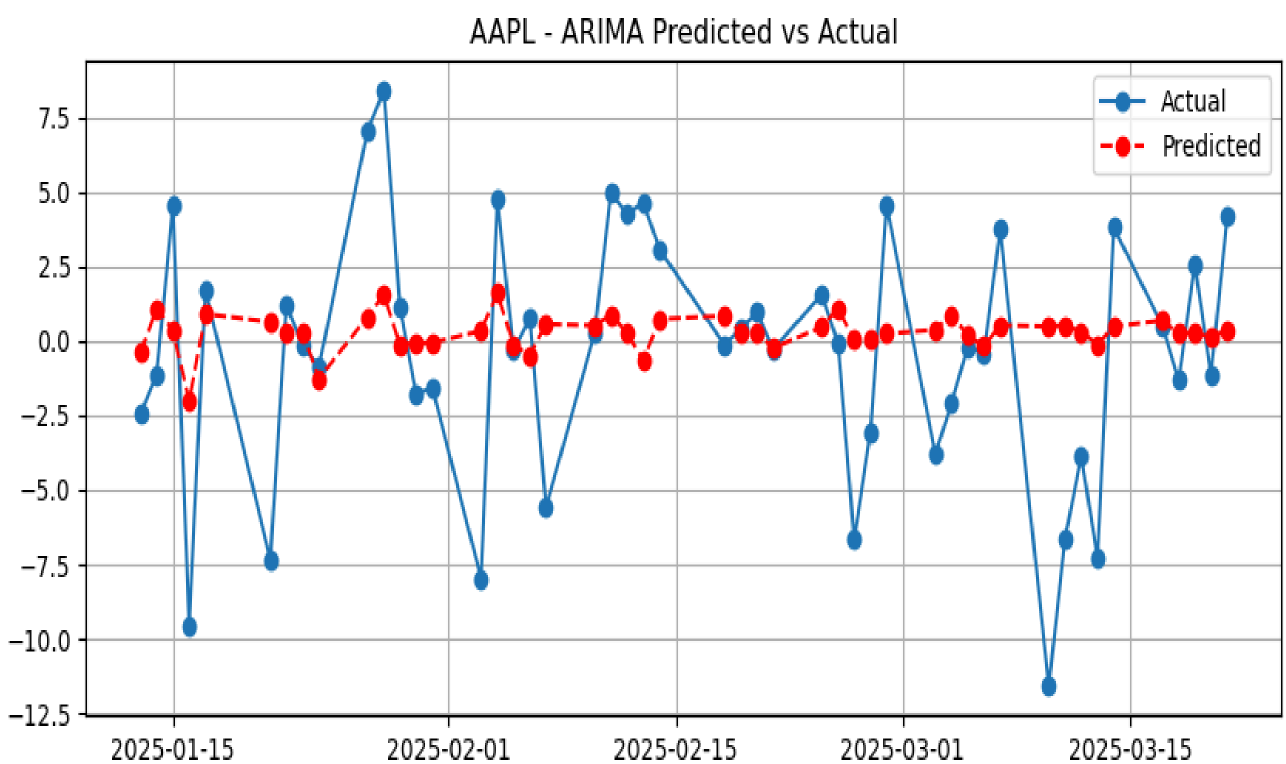
3. **Differencing:** Since stock prices are often non-stationary, **first-order differencing** was applied to remove trends.

4. **Model Order Selection:** The ARIMA (p, d, q) parameters were set as (7,1,12) based on ACF and PACF analysis.

5. **Training:** The ARIMA model was fitted on the training dataset, and future prices were forecasted on the test set.
6. **Prediction:** The model forecasted the stock prices for the test set.
7. **Evaluation:** Performance was measured using **MAE, RMSE**

AAPL Stock's ACF and PACF Plot



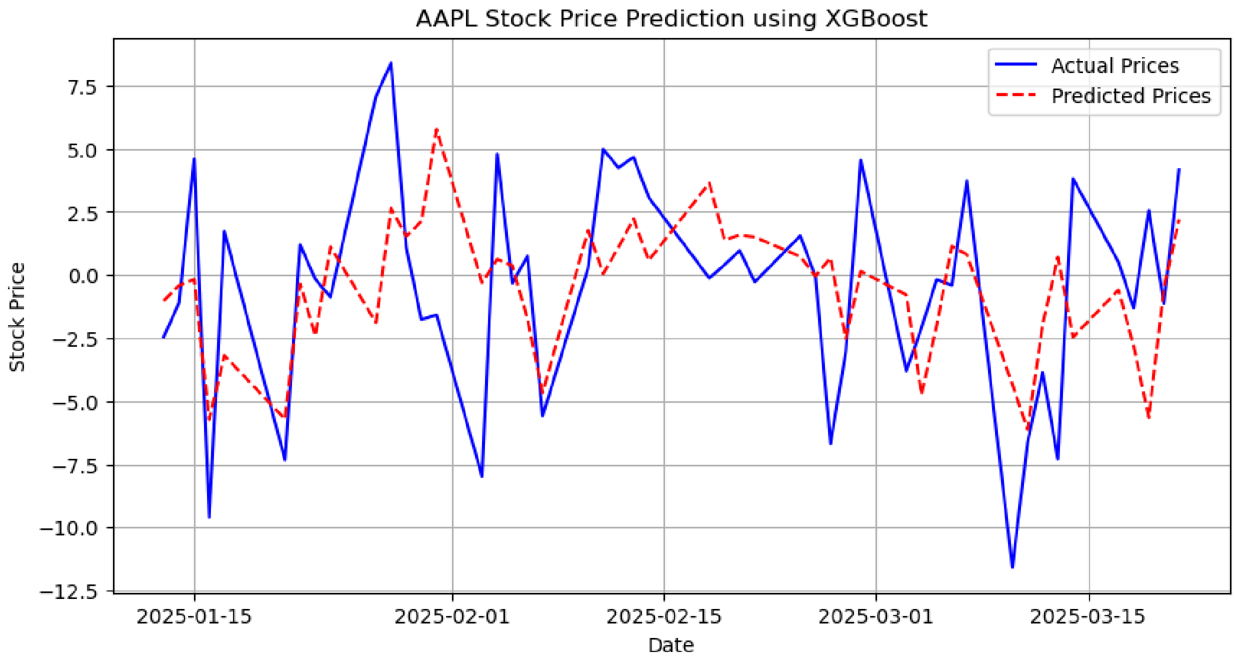


XGBoost Model

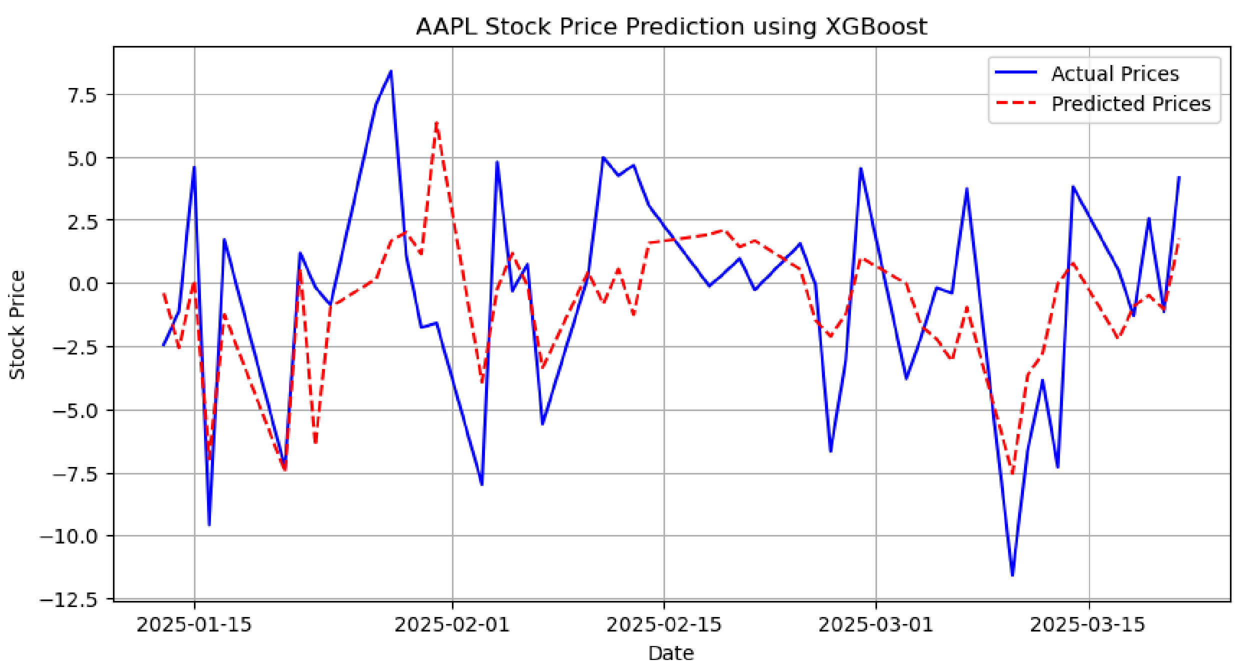
To enhance predictive accuracy, we implemented an **XGBoost (Extreme Gradient Boosting)** model, a powerful tree-based algorithm. The features included **lagged returns, moving averages, and volume changes**.

Steps:

1. **Feature Engineering:**
   - Lag features (previous day's closing price)
   - Moving averages (5-day rolling mean)
   - Volatility measures (standard deviation of closing prices)
   - Percentage change in closing prices
2. **Train-Test Split:** 80% of the data was used for training, and 20% for testing.
3. **Hyperparameter Optimization:** Optuna was used for tuning parameters, and the best values found were:
   - `n_estimators`: 857
   - `max_depth`: 8
   - `learning_rate`: 0.4419
   - `subsample`: 0.73199
   - `colsample_bytree`: 0.60415
   - `gamma`: 4.4484
   - `reg_alpha`: 1.4281
   - `reg_lambda`: 1.6677
4. **Model Training:** XGBoost was trained with these optimized hyperparameters.
5. **Prediction & Evaluation:** The model predicted stock prices on the test set and was evaluated using **MAE, RMSE, MAPE, and R^2 score**.



Before Hyperparameter Tuning



After Hyper Parameter Tuning

Model Evaluation & Comparison

Evaluation Metrics

- **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual values.
- **Root Mean Squared Error (RMSE):** Measures the standard deviation of prediction errors, giving more weight to large errors.
- **R-squared (R2 Score):** Indicates how well the model explains variance in the actual values. A higher R2 score represents a better fit.

ARIMA Model Performance

For the ARIMA model, the evaluation metrics were as follows:

- **MAE:** 3.1521
- **RMSE:** 4.1892
- **R2 Score:** 0.0552

The ARIMA model performed reasonably well in predicting stock prices, but the low R2 score suggests that it does not explain much of the variance in stock prices.

XGBoost Model Performance

For the XGBoost model, the evaluation metrics were:

- **MAE:** 2.8341
- **RMSE:** 3.5096
- **R2 Score:** 0.3369

Compared to ARIMA, the XGBoost model achieved lower MAE and RMSE values, indicating better predictive accuracy. Additionally, the higher R2 score suggests that XGBoost explains a greater proportion of the variance in stock prices.

Model Comparison

Model	MAE	RMSE	R2 Score
ARIMA	3.1521	4.1892	0.0552
XGBoost	2.8341	3.5096	0.3369

From the above comparison, XGBoost outperforms ARIMA in all evaluation metrics, making it the preferred choice for stock price prediction in this study.

Implications for Trading Strategies

1. **Short-Term Trading & Algorithmic Trading:**
   - Given XGBoost’s superior accuracy, traders can integrate it into algorithmic trading systems to predict short-term price movements with greater confidence.
   - The model can be used to identify entry and exit points based on expected price trends, reducing the risk of false signals.
2. **Risk Management & Stop-Loss Adjustments:**
   - With lower error margins in predictions, traders can set more effective stop-loss and take-profit levels based on the forecasted price range.
   - XGBoost’s ability to capture trends helps in minimizing unexpected losses due to price fluctuations.
3. **Portfolio Allocation & Position Sizing:**
   - Investors seeking to allocate capital efficiently can use XGBoost-based forecasts to adjust portfolio weightings dynamically.
   - By analyzing price trends and volatility indicators, traders can make informed decisions on the size of positions taken in AAPL.

Conclusion

The comparison between ARIMA and XGBoost models highlights the importance of machine learning techniques in financial markets. The superior performance of XGBoost suggests that data-driven trading strategies, particularly in short-term trading, can yield better results compared to traditional statistical models. Traders and investors can leverage these insights to optimize their decision-making processes and enhance overall profitability.