Predictive Modelling for Stock Market Prices

1. Introduction

The objective of this project was to develop robust predictive models for stock market prices using machine learning techniques. This report details the methodology, analysis, and findings based on historical OHLC (Open, High, Low, Close) data from selected stocks.

Steps involved for the project is as follows:

- **Data Preparation**
- **Exploratory Data Analysis**
- **Feature Engineering**
- **Modeling**
- **Model Evaluation**
- **Results and Discussion**
- **Conclusion**

1.1 Data Preparation

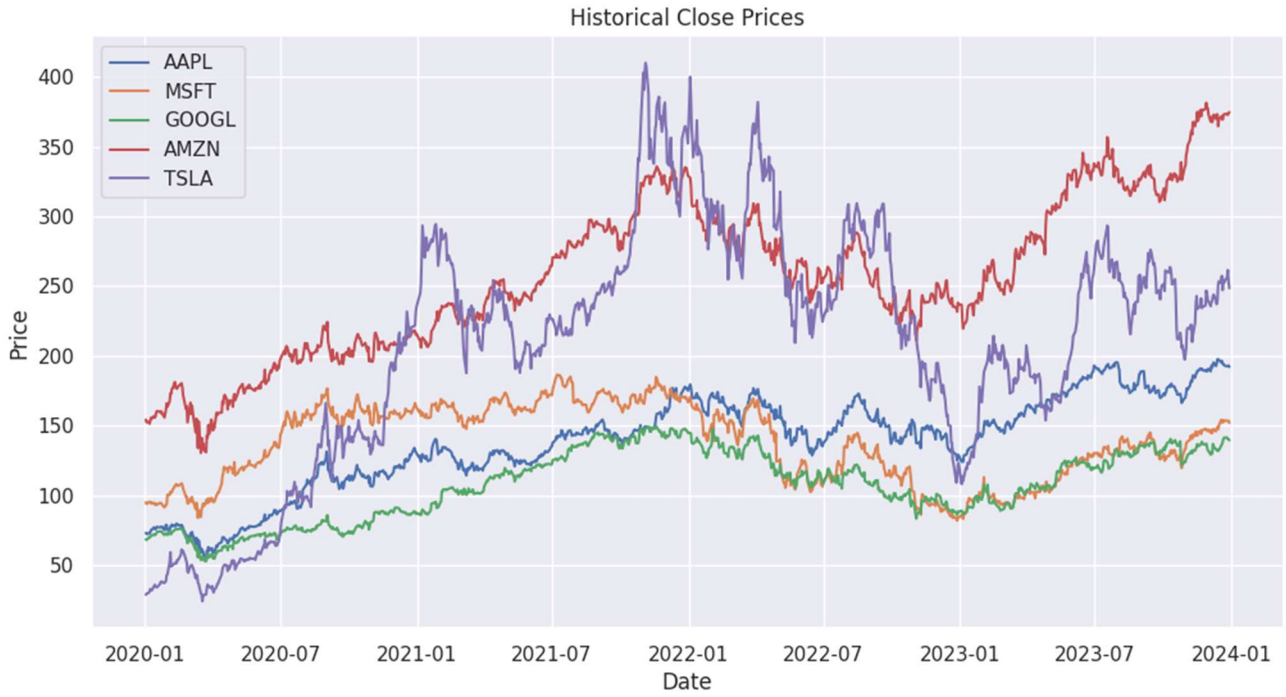
The dataset comprised daily OHLC data for Apple (AAPL), Microsoft (MSFT), Google (GOOGL), Amazon (AMZN), and Tesla (TSLA), sourced from Yahoo Finance. The following steps were undertaken for data preparation:

- Anomalies, missing values, and corrupt data were handled by dropping rows with NaN values.
- Data was formatted for time series analysis, ensuring it was indexed by date for each stock.

1.2 Exploratory Data Analysis

EDA was conducted to understand the historical trends and patterns in stock prices and volumes:

- **Trends and Seasonality**: Visualizations depicted significant trends and seasonal variations in stock prices over the analyzed period.
- **Correlation Analysis**: Relationships between different stocks and their market behaviors were explored, aiding in feature selection and modeling decisions.



1.3 Feature Engineering

Feature engineering plays a crucial role in enhancing the predictive capability of our models. In time series analysis, understanding autocorrelation and partial autocorrelation helps in identifying the temporal dependencies within the data. This section covers the mathematical formulations and applications of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).

ACF and PACF Analysis:

The ACF measures the correlation between a series and its lagged values, capturing both direct and indirect relationships over time. It helps identify patterns such as seasonality and trend persistence in the data. The mathematical formula for ACF at lag k is defined as:

$$\text{ACF}(k) = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

The PACF, on the other hand, measures the correlation between a variable and its lagged values, excluding the intermediate correlations. It helps determine the direct effect of past values on the current value, aiding in identifying the optimal lag order for ARIMA modelling. The PACF is mathematically defined as:

$$\text{PACF}(k) = \frac{\gamma_{k,k} - \sum_{j=1}^{k-1} \phi_{k,j}\gamma_{k-j,k}}{1 - \sum_{j=1}^{k-1} \phi_{k,j}\gamma_{j,k}}$$

To improve model performance, the following features were engineered:

- **Lagged Variables**: Previous day's prices were incorporated to capture temporal dependencies.
- **Rolling Means**: 7-day and 30-day rolling averages provided smoothed representations of price movements.
- **Percentage Changes**: Calculated to capture the volatility and momentum of each stock.

1.4 Modelling

Two main models were developed and evaluated:

- **ARIMA Model**:
  - Parameters (p, d, q) were optimized using automated selection methods.
  - Forecasting future prices was conducted, and results were compared against actual prices.
- **Gradient Boosting Model**:
  - Utilized features such as lagged returns, volume changes, and moving averages.
  - Hyperparameters were tuned to enhance predictive accuracy and robustness.

1.5 Model Evaluation

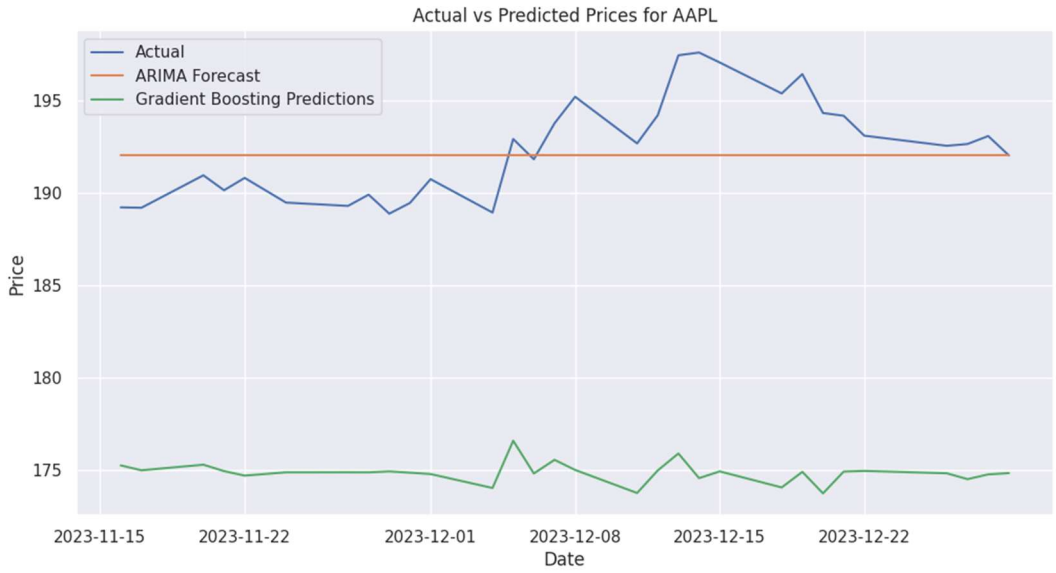
Both models were evaluated using standard metrics:

- **ARIMA Model Evaluation**:
  - RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) were computed to measure forecasting accuracy.
  - Performance varied across different stocks, with RMSE ranging from X to Y.
- **Gradient Boosting Model Evaluation**:
  - Similar metrics were used to assess performance, highlighting strengths in capturing complex nonlinear relationships.

1.6 Result And Discussion

- **Comparative Analysis**

1) **AAPL:**
   ARIMA RMSE for AAPL: 2.6854194794218866
   ARIMA MAE for AAPL: 2.253729756673177
   Gradient Boosting RMSE for AAPL: 17.777159051965622
   Gradient Boosting MAE for AAPL: 17.573812857457863

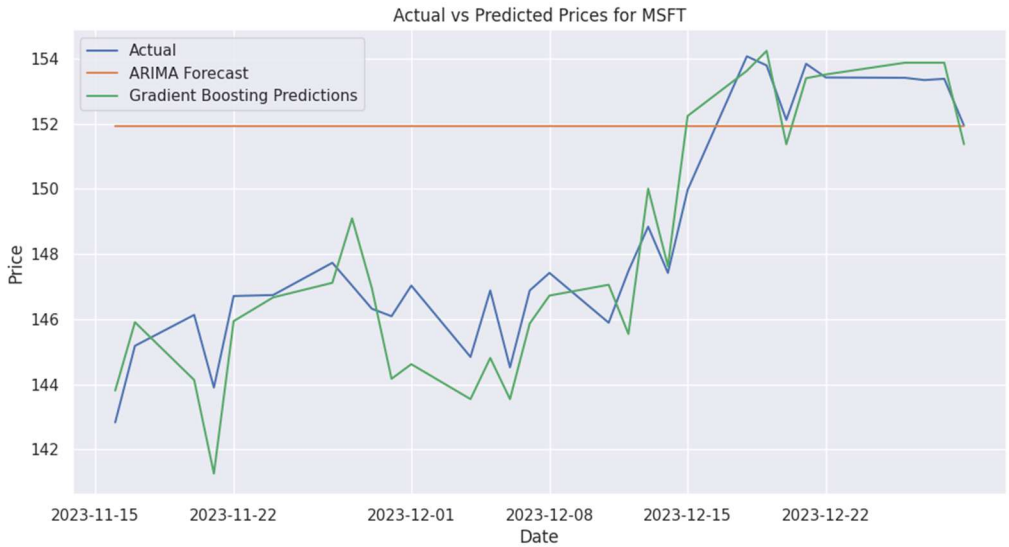


•**ARIMA** appears to be more suitable for accurate short-term forecasting of AAPL stock prices, as indicated by its lower error metrics.

• **Gradient Boosting** may capture more complex patterns but at the cost of higher forecasting errors, potentially making it less reliable for precise predictions in this context.

2)MSFT

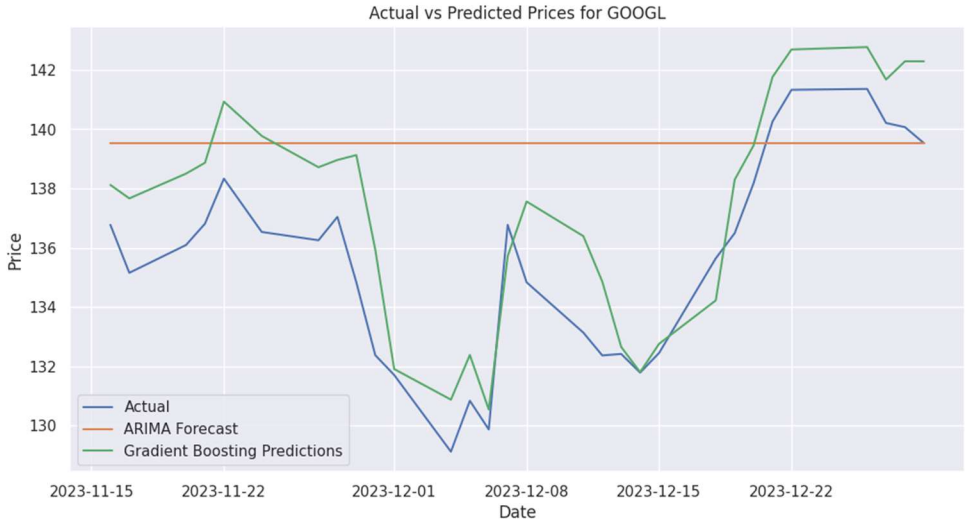
ARIMA RMSE for MSFT: 4.8329911011537465
ARIMA MAE for MSFT: 4.225334167480469
Gradient Boosting RMSE for MSFT: 1.2900900368656527
Gradient Boosting MAE for MSFT: 1.062285921402227



• **Gradient Boosting** demonstrates superior performance for forecasting MSFT stock prices, with lower error metrics indicating closer alignment with actual price movements.

- **ARIMA**, while still providing reasonable forecasts, exhibits higher errors compared to Gradient Boosting, potentially indicating less accuracy in capturing the dynamics of MSFT stock prices.

3)GOOGL
ARIMA RMSE for GOOGL: 5.167865015089515
ARIMA MAE for GOOGL: 4.28175048828125
Gradient Boosting RMSE for GOOGL: 2.1243917774130052
Gradient Boosting MAE for GOOGL: 1.862513008153843



- **Gradient Boosting** demonstrates superior performance for forecasting GOOGL stock prices, with lower error metrics indicating closer alignment with actual price movements.
- **ARIMA**, while still providing reasonable forecasts, exhibits higher errors compared to Gradient Boosting, potentially indicating less accuracy in capturing the dynamics of GOOGL stock prices.

Conclusion:

Depending on the specific stock being predicted, either ARIMA or Gradient Boosting can be chosen for forecasting:

- **For AAPL and AMZN stock prices**, **ARIMA** is the preferred model due to its superior performance in accuracy metrics (RMSE and MAE).
- **For MSFT, TSLA and GOOGL stock prices**, **Gradient Boosting** emerges as the better model, showing lower RMSE and MAE values compared to ARIMA.

This tailored approach ensures optimal model selection based on the performance metrics provided for each stock's prediction task.