Stock Price Prediction

using ARIMA and XGBoost Model

2024-08-03

B C Samrudh

I. Introduction

This report presents the findings of stock price prediction project, which aimed to develop models for forecasting stock prices based on historical data. Five stocks have been used: Apple (AAPL), Google (GOOGL), Morgan Stanley (MS), JP Morgan(JPM), and Goldman Sachs (GS).

II. Methodology

The approach involves the following steps:

1. Data Collection and Preparation
2. Exploratory Data Analysis (EDA)
3. Feature Engineering
4. Model Development (ARIMA and Gradient Boosting)
5. Model Evaluation

III. Data Collection and Preparation

Data of 5 Stocks - Apple, Google, Morgan Stanley, JP Morgan and Goldman Sachs was collected from Yahoo Finance using `yfinance library`

Augmented Dickey–Fuller test Results:

Metric	AAPL	GOOGL	MS	JPM	GS
ADF Statistic	0.2600	0.1529	−0.8182	−0.8664	−1.1432
p-value	0.9754	0.9695	0.8137	0.7988	0.6976
1% Critical Value	−3.4330	−3.4330	−3.4330	−3.4330	−3.4330
5% Critical Value	−2.8627	−2.8627	−2.8627	−2.8627	−2.8627
10% Critical Value	−2.5674	−2.5674	−2.5674	−2.5674	−2.5674

IV. Model Development

ARIMA Model:

I used `auto_arima function` of the `pmdarima` to find the optimal p,q,d values instead of ACF and PACF Plots.

```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=5298.559, Time=1.21 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=5300.881, Time=0.23 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=5299.351, Time=0.17 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=5299.632, Time=0.35 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=5301.885, Time=0.09 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=5297.974, Time=1.07 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : AIC=5297.632, Time=0.44 sec
 ARIMA(0,1,3)(0,0,0)[0] intercept   : AIC=5294.344, Time=0.62 sec
 ARIMA(1,1,3)(0,0,0)[0] intercept   : AIC=5294.000, Time=2.06 sec
 ARIMA(2,1,3)(0,0,0)[0] intercept   : AIC=5272.445, Time=9.55 sec
 ARIMA(3,1,3)(0,0,0)[0] intercept   : AIC=5272.005, Time=9.56 sec
 ARIMA(3,1,2)(0,0,0)[0] intercept   : AIC=5271.683, Time=5.50 sec
 ARIMA(3,1,1)(0,0,0)[0] intercept   : AIC=5294.429, Time=1.75 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=5301.005, Time=2.53 sec
 ARIMA(3,1,2)(0,0,0)[0]             : AIC=5273.014, Time=11.12 sec

Best model:  ARIMA(3,1,2)(0,0,0)[0] intercept
Total fit time: 46.272 seconds
                               SARIMAX Results
==========================================================================================
Dep. Variable:                           y   No. Observations:                 1996
Model:                 SARIMAX(3, 1, 2)   Log Likelihood               -2628.842
Date:                  Sat, 03 Aug 2024   AIC                           5271.683
Time:                          07:32:46   BIC                           5310.872
Sample:                      01-31-2014   HQIC                          5286.074
                           - 01-03-2022
Covariance Type:                     opg
==========================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0477      0.027      1.734      0.083      -0.006       0.102
ar.L1          0.5766      0.014     42.405      0.000       0.550       0.603
ar.L2         -0.9181      0.012    -74.331      0.000      -0.942      -0.894
ar.L3         -0.0772      0.012     -6.398      0.000      -0.101      -0.054
ma.L1         -0.6246      0.009    -71.570      0.000      -0.642      -0.607
ma.L2          0.9731      0.009    103.374      0.000       0.955       0.992
sigma2         0.8169      0.014     59.225      0.000       0.790       0.844
==========================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):              2377.44
Prob(Q):                              0.89   Prob(JB):                         0.00
Heteroskedasticity (H):               5.82   Skew:                            -0.04
Prob(H) (two-sided):                  0.00   Kurtosis:                         8.35
==========================================================================================
```

XGBoost Model:

I used `Bayesian optimization` instead of GridSearch and RandomSearch for Hyperparameter Tuning for the following reasons.

- Efficiency: Bayesian optimization explores the hyperparameter space more efficiently than GridSearchCV.
- Complex Search Spaces: Bayesian optimization handles complex relationships between hyperparameters better.
- Early Stopping: Bayesian optimization can terminate the search early if performance plateaus.

Model Parameters:

1. ARIMA (Auto-Regressive Integrated Moving Average)
   `ARIMA(3,1,2)(0,0,0)[0] intercept`
2. XGBoost
   `'n_estimators': 325, 'learning_rate': 0.0455081099701603, 'max_depth': 5, 'min_child_weight': 14, 'subsample': 0.9939195480488889, 'colsample_bytree': 0.9742302863939567`

V. Key Findings

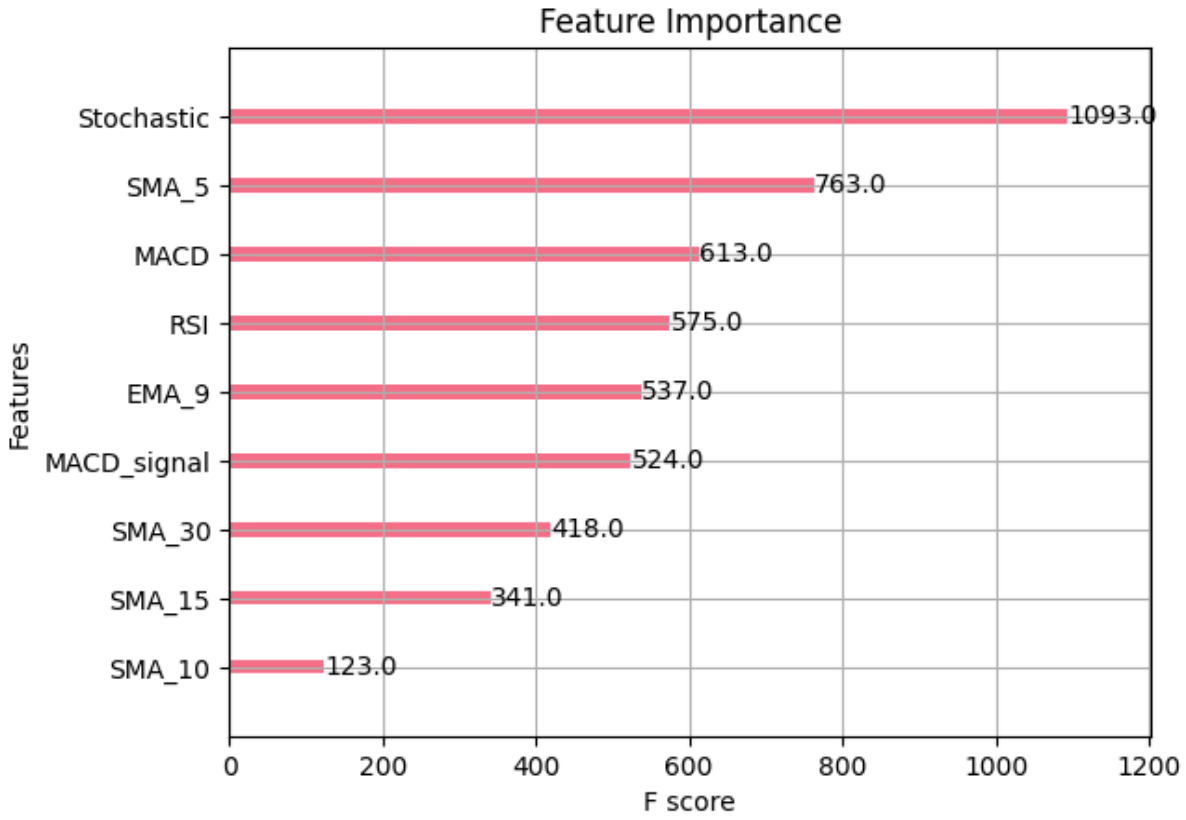
Data Analysis

- The stock prices of all five companies showed an overall upward trend over the analyzed period.
- We observed varying levels of volatility across the stocks, with some periods of high volatility coinciding with major market events.
- The distribution of daily returns for most stocks appeared to be approximately normal but with fat tails, indicating the presence of extreme price movements.

Feature Engineering

- **Exponential Moving Average (EMA)** - 9 days: Calculates the EMA over a 9-day period to capture short-term trends.
- **Simple Moving Averages (SMA)**: Computes SMAs for various window sizes (5, 10, 15, and 30 days) to analyze price trends over different periods.
- **Relative Strength Index (RSI)**: Measures the speed and change of price movements over a 14-day period, indicating overbought or oversold conditions.
- **Moving Average Convergence Divergence (MACD)**: Computes the difference between the 12-day and 26-day EMAs to assess the strength and direction of a trend.
- **MACD Signal Line**: Calculates the 9-day EMA of the MACD to identify potential buy or sell signals.
- **Stochastic Oscillator**: Measures momentum by comparing the closing price to its price range over a specific period. The `Technical Analysis Library (ta)` was used to calculate Stochastic Oscillator.

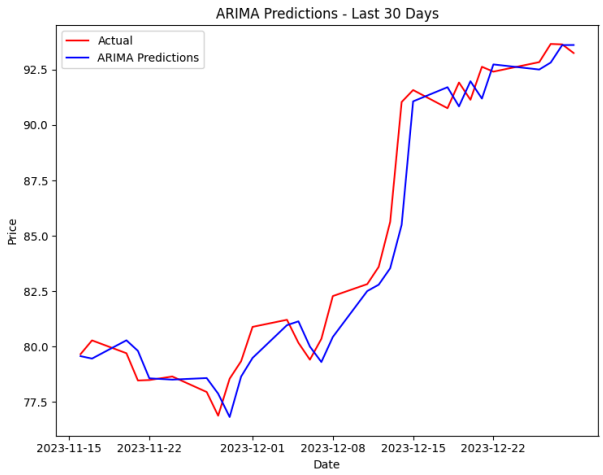
Feature Importance Graph

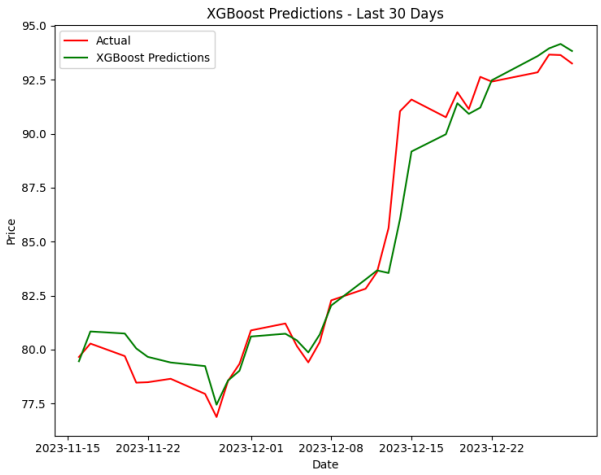


Model Performance

Metric	ARIMA Model	XGBoost Model
Mean Squared Error (**MSE**)	2.589	1.662
Root Mean Squared Error (**RMSE**)	1.609	1.289
Mean Absolute Error (**MAE**)	1.211	0.981
Mean Absolute Percentage Error (**MAPE**)	1.401	1.140

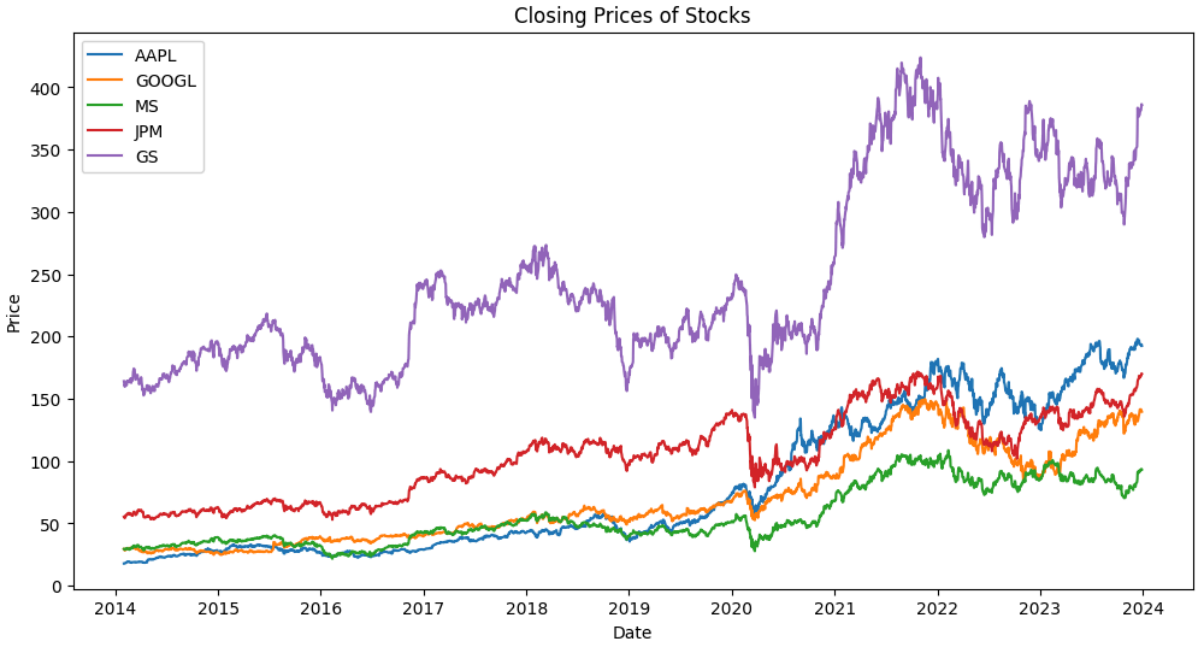
The Evaluation Metrics and the Graph show that the **XGBoost Model is better** in Accuracy and Reliability than the ARIMA Model.



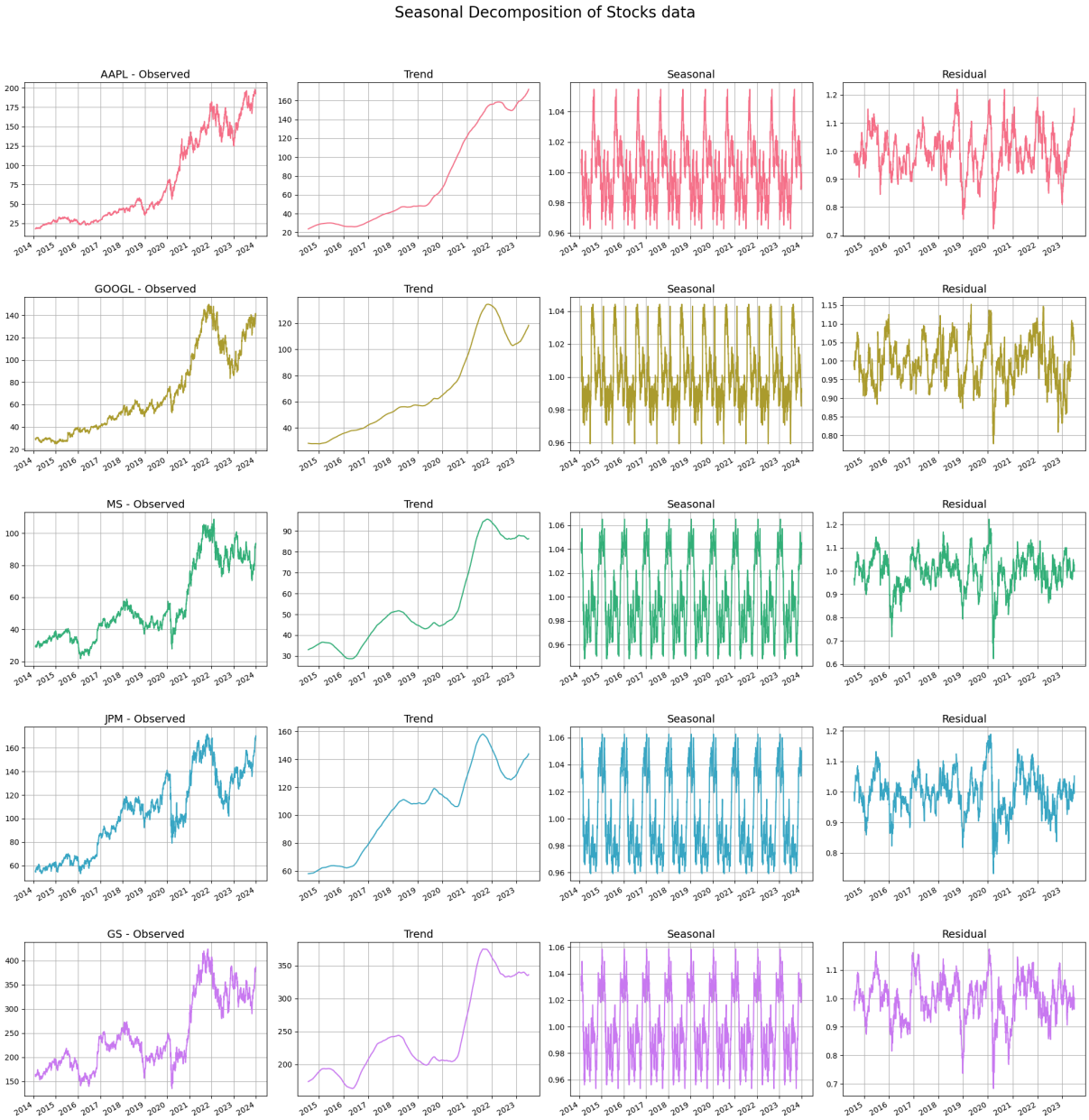


VI. Visualizations

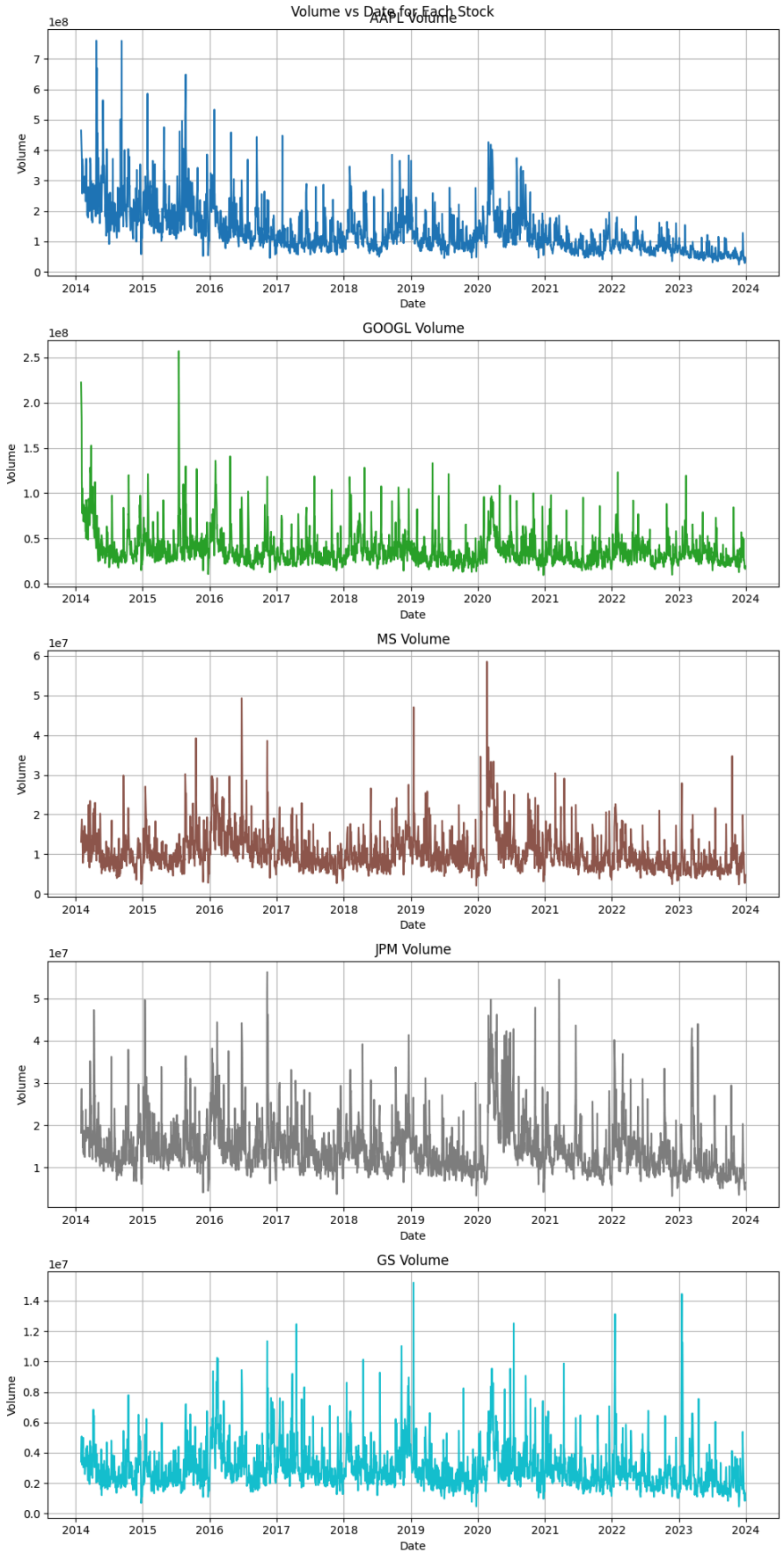
Closing Price vs Date Graph:



Seasonal Decomposition of the Stocks:



Volume vs Date Graph:



Volatility vs Date Graph: