Time Series Analysis: Microsoft Stock

—

Forecasting Microsoft Stock Price Using ARIMA

Dataset Introduction

This dataset contains daily stock trading data for Microsoft Corporation, starting from April 1st, 2015.

- Date: the date of the stock record
- Open: the stock's opening price
- High and Low: the highest and lowest prices during the day
- Close: the closing price, which is also the value we're trying to forecast
- Volume: the number of shares traded on that day

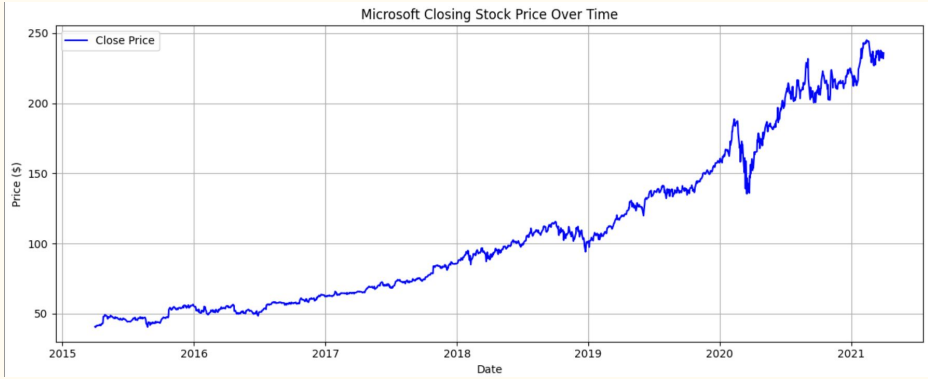
	Date	Open	High	Low	Close	Volume
0	4/1/2015 16:00:00	40.60	40.76	40.31	40.72	36865322
1	4/2/2015 16:00:00	40.66	40.74	40.12	40.29	37487476
2	4/6/2015 16:00:00	40.34	41.78	40.18	41.55	39223692
3	4/7/2015 16:00:00	41.61	41.91	41.31	41.53	28809375
4	4/8/2015 16:00:00	41.48	41.69	41.04	41.42	24753438

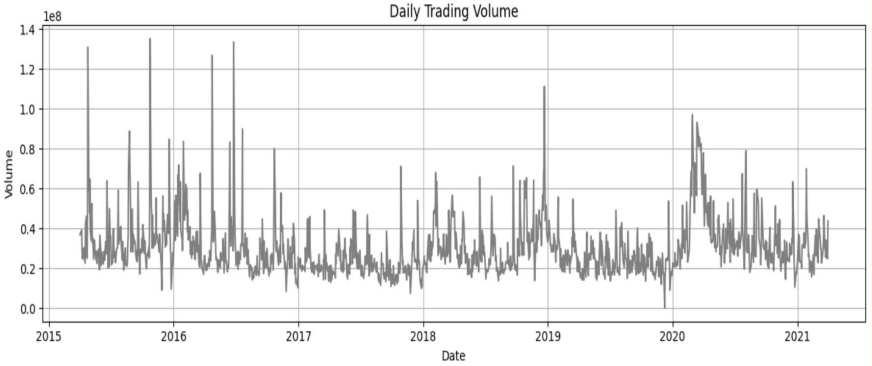
Problem Statement

- Predicting stock prices is crucial for informed financial decisions
- Microsoft (MSFT) is a globally important, high-volume stock
- Goal: Use historical data to forecast short-term closing prices
- Approach: Apply time series models (ARIMA, Prophet)
- Evaluate model accuracy for practical forecasting use

Exploratory Data Analysis(1)

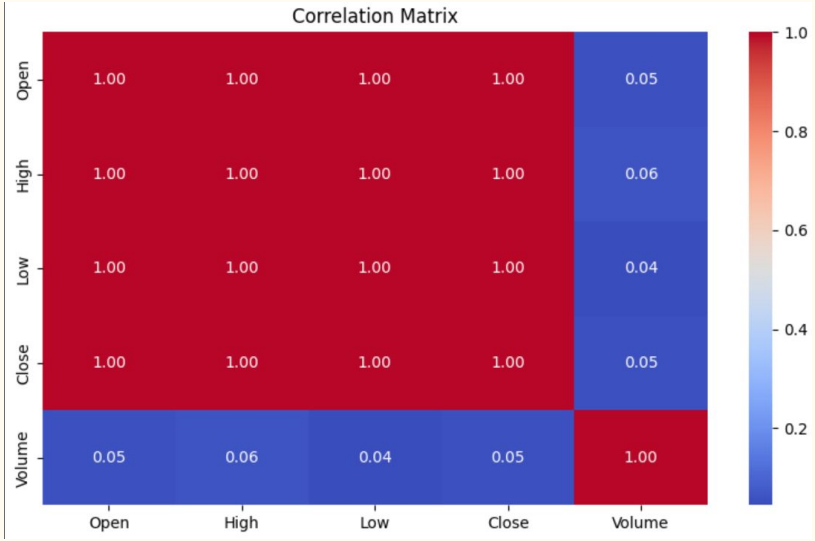
- Microsoft stock has shown a strong upward trend from 2015 to 2021, with noticeable acceleration after 2019
- Despite price growth, the trading volume remained relatively volatile, with occasional spikes during major events





Exploratory Data Analysis(2)

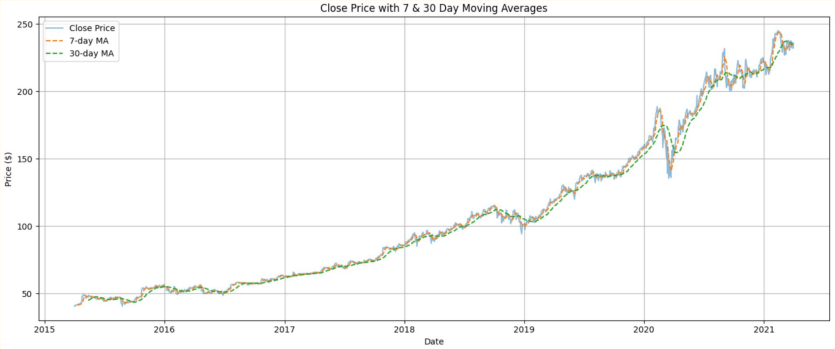
- Strong positive correlation observed among Open, High, Low, and Close prices (correlation ≈ 1.0)
- Volume has weak correlation with price-based features, suggesting it behaves independently

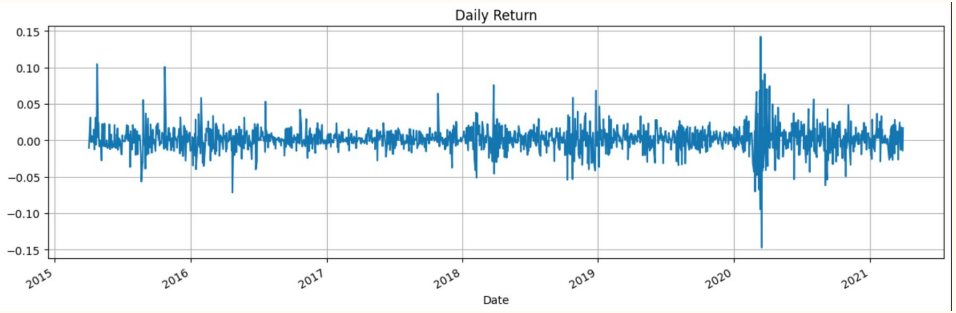


Feature Engineering

• The 7-day and 30-day moving averages captured short- and mid-term trends, clearly showing upward momentum with occasional dips.

• Daily returns revealed higher volatility during early 2020, likely reflecting market reaction to COVID-19.





Time Series Modeling(1)

- Performed Augmented Dickey-Fuller (ADF) test to check stationarity
- ADF Statistic: 1.68
- P-value: 0.998. Since P-value $>$ 0.05 Failed to reject the null hypothesis
- Conclusion: The time series is non-stationary and requires differencing

```
ADF Statistic: 1.6830285474635915
p-value: 0.9980864714246888
```

Time Series Modeling(2)

- Model Configuration
  - Model type: ARIMA(5,1,0) on Close price
  - Observations: 1,482
  - Covariance type: OPG
- Model Fit & Criteria
  - AIC: 6662.12
  - BIC: 6693.93
  - HQIC: 6673.98

SARIMAX Results

Dep. Variable:	Close	No. Observations:	1482
Model:	ARIMA(5, 1, 0)	Log Likelihood	-3325.062
Date:	Thu, 24 Jul 2025	AIC	6662.124
Time:	04:50:31	BIC	6693.927
Sample:	0	HQIC	6673.980
	- 1482		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.2705	0.012	-23.380	0.000	-0.293	-0.248
ar.L2	0.0213	0.012	1.705	0.088	-0.003	0.046
ar.L3	0.0503	0.014	3.481	0.001	0.022	0.079
ar.L4	-0.0138	0.014	-0.972	0.331	-0.042	0.014
ar.L5	-0.0127	0.013	-1.002	0.316	-0.038	0.012
sigma2	5.2192	0.081	64.475	0.000	5.061	5.378

Ljung-Box (L1) (Q):	0.06	Jarque-Bera (JB):	7885.27
Prob(Q):	0.81	Prob(JB):	0.00
Heteroskedasticity (H):	22.71	Skew:	-0.47
Prob(H) (two-sided):	0.00	Kurtosis:	14.26

Time Series Modeling(3)

- Significant Coefficients
  - AR(1): -0.2705, $p < 0.001$
  - AR(3): 0.0503, $p < 0.01$
  - Other lags (AR2, AR4, AR5) are not statistically significant
- Residual Diagnostics
  - Ljung-Box Q ($p = 0.81$): No autocorrelation
  - Heteroskedasticity Test ($p < 0.01$): Heteroskedasticity exists
  - Jarque-Bera $p = 0.00$: Residuals not normally distributed

SARIMAX Results

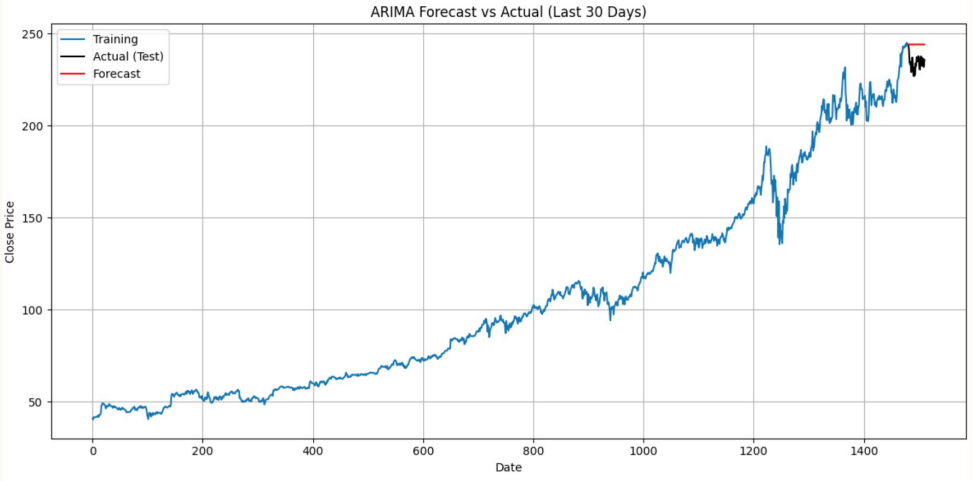
Dep. Variable:	Close	No. Observations:	1482
Model:	ARIMA(5, 1, 0)	Log Likelihood	-3325.062
Date:	Thu, 24 Jul 2025	AIC	6662.124
Time:	04:50:31	BIC	6693.927
Sample:	0	HQIC	6673.980
	- 1482		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.2705	0.012	-23.380	0.000	-0.293	-0.248
ar.L2	0.0213	0.012	1.705	0.088	-0.003	0.046
ar.L3	0.0503	0.014	3.481	0.001	0.022	0.079
ar.L4	-0.0138	0.014	-0.972	0.331	-0.042	0.014
ar.L5	-0.0127	0.013	-1.002	0.316	-0.038	0.012
sigma2	5.2192	0.081	64.475	0.000	5.061	5.378

Ljung-Box (L1) (Q):	0.06	Jarque-Bera (JB):	7885.27
Prob(Q):	0.81	Prob(JB):	0.00
Heteroskedasticity (H):	22.71	Skew:	-0.47
Prob(H) (two-sided):	0.00	Kurtosis:	14.26

Time Series Modeling(4)

- The ARIMA model overestimated the stock price during the last 30 business days.
- Predicted values are consistently higher than actual observations.
- This indicates the model may not fully capture recent short-term volatility.



Time Series Modeling(5)

- The ARIMA(5,1,0) model forecasts a slightly upward trend in the near future.
- Forecast line (in red) shows a flattening pattern, suggesting limited short-term growth.
- The forecast continues from a historical strong upward momentum seen since 2015.
- No strong volatility is projected in the forecast range.

Time Series Modeling(6)

- Root Mean Squared Error (RMSE): 10.61
- Mean Absolute Error (MAE): 9.91
- Indicates the model's average prediction error is around $9–10 per day over the test period.
- Relatively low error given the price range (~$200+), but still room for refinement.
- RMSE > MAE suggests presence of larger occasional errors (outliers).

```
RMSE: 10.61
MAE: 9.91
```

Result/Conclusion

- Successfully implemented ARIMA(5,1,0) to forecast stock closing prices.
- Forecast captures short-term trend reasonably well for the next 30 business days.
- Evaluation metrics indicate low average error:
  - RMSE = 10.61, MAE = 9.91
- Forecasted values are slightly overestimated compared to actual test values.
- ARIMA model performs well on stable time series, but may underperform during volatile periods.