ISE 537 Financial Analytics
Final Project Report

Stock Price Prediction: Comparison of ARIMA and LSTM

Ankur Ojha

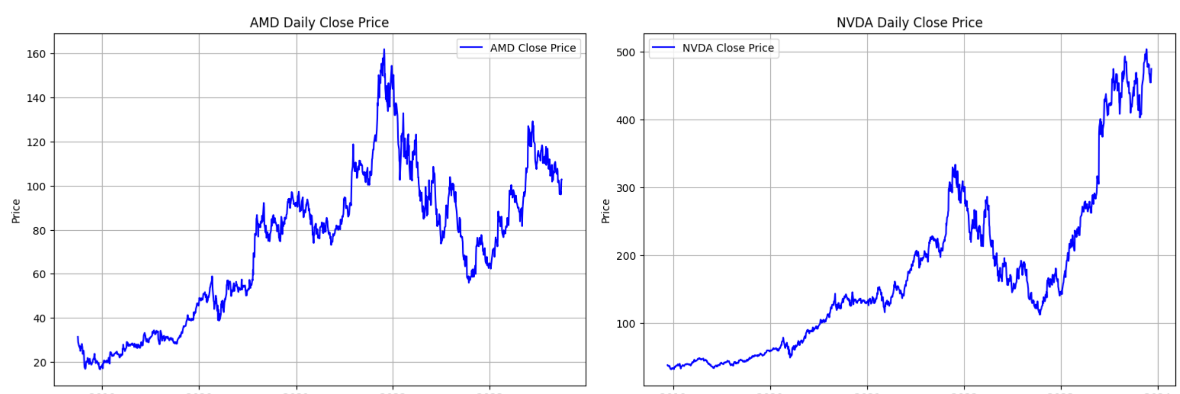
December 16, 2023

Abstract

In this report, we delve into the predictive analysis of stock prices for two major technology companies, NVIDIA Corporation (NVDA) and Advanced Micro Devices, Inc. (AMD), covering the period from 2018 Q4 to 2023 Q3. Our primary objective is to explore the auto-regressive structure and trend of the daily closing stock prices, assess the impact of market dynamics, including the 2020 financial upheaval, and forecast future price movements for 2022-23. Our analysis reveals that the stock prices exhibit non-stationary characteristics over time. By applying ARIMA and LSTM models, we found distinct behaviors in the price series, with notable differences in response to market changes, particularly during periods of high volatility. The ARIMA model showed a strong ability to capture linear trends, while the LSTM model excelled in identifying non-linear patterns, especially in the face of rapid market shifts. Our forecasts for 2022-23, using a novel Hybrid ARIMA-LSTM approach, demonstrate a significant improvement in prediction accuracy, indicating the model's robustness in adapting to complex market behaviors.

1 Data

For this project stock price prediction application, I am focusing specifically on the stocks of NVIDIA (NASDAQ: NVDA) and AMD (NASDAQ: AMD) for the time period of Q4 2018 to Q3 2023. These two companies are at the forefront of the semiconductor industry, making them particularly interesting for analysis.



(a) AMD Stock prices

(b) NVIDIA Stock prices

Figure 1: Stock prices for the AMD and NVIDIA between Q4 2018 to Q3 2023

There was lot of change in the economic outlook in the last five years. The booming AI industry has significantly increased the demand for high-performing chips, catapulting Nvidia and AMD into the spotlight. Additionally, the surge in cryptocurrency popularity has led to increased demand for powerful GPUs, a specialty of both companies, further influencing their stock prices. Other than that there was a pandemic later followed by a supply chain issue impacting these stocks and the overall market. With these factors, the data from NVIDIA and AMD stocks provide a rich and complex landscape for predictive analysis, making them ideal choices for my project.

2 Data Preprocessing

In this section, we delve into the preliminary data analysis conducted on two selected stocks, focusing on their daily closing prices. The analysis includes checking for missing values, understanding seasonality and trends, and conducting stationarity tests (ADF) to inform subsequent transformation and modeling choices.

2.1 Missing Values and Time Series Decomposition

Initial data checks confirmed that there were no missing values within the datasets for both NVDA and AMD stocks. A decomposition analysis was performed to dissect the time series into its constituent components—trend, seasonality, and residuals. Seasonality is found in both stocks.

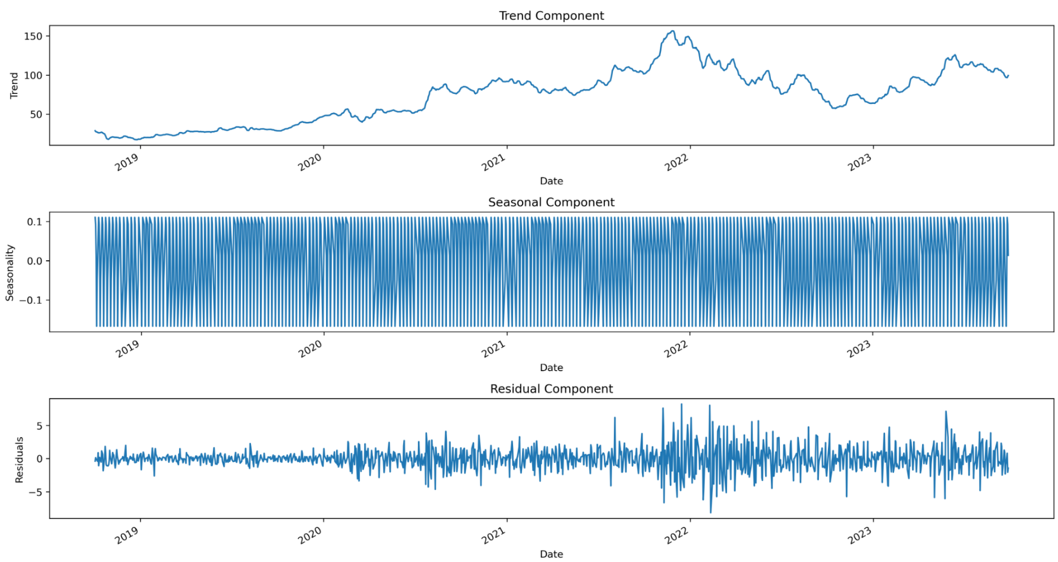


Figure 2: AMD Stock Data breakdown for Trend, Seasonality, and Residual

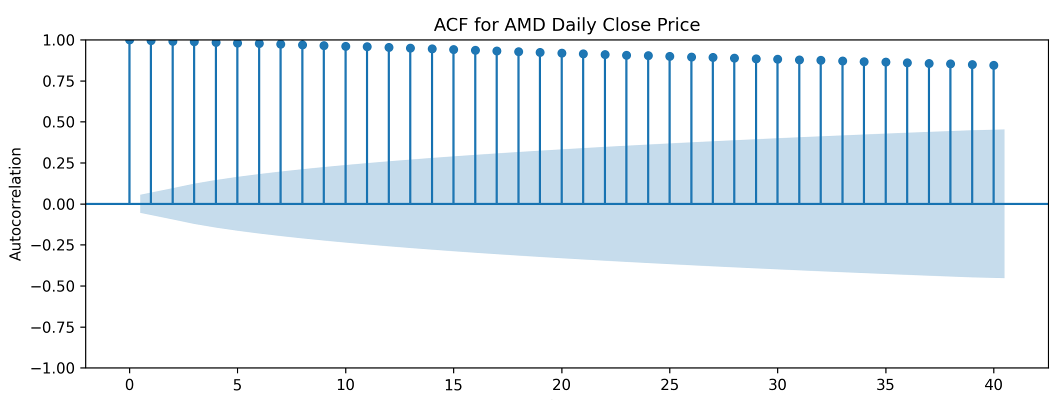


Figure 3: ACF Plot for AMD Stock Close Price with lags upto 40.

The ACF plot for AMD's daily close price indicates significant positive autocorrelation at the first lag, with a gradual decay and no apparent seasonality, suggesting an AR(1) model may be appropriate.

2.2 Stationarity Testing

The Augmented Dickey-Fuller (ADF) test was applied to the daily closing prices of AMD stock to test for stationarity. The ADF test yielded a statistic of -1.54 and a P-value of 0.512. Given the P-value exceeded common significance levels (0.05), there was insufficient evidence to reject the null hypothesis, leading to the preliminary conclusion that AMD's stock prices were non-stationary.

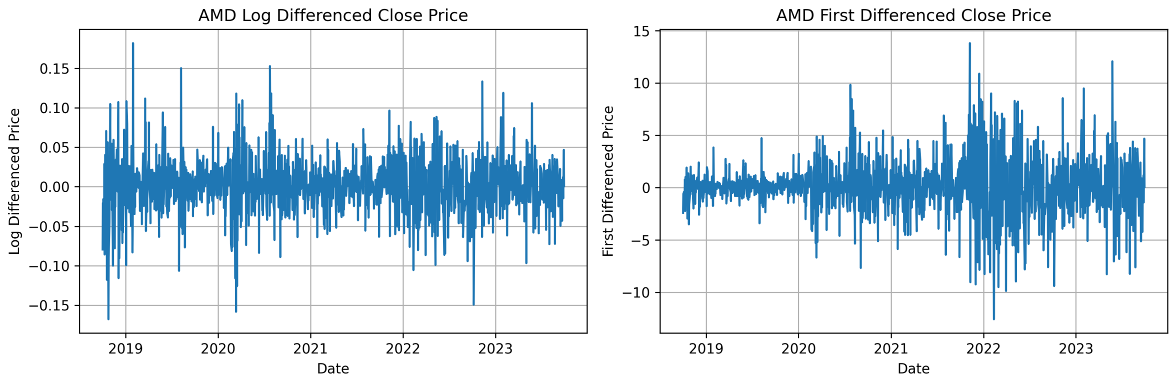


Figure 4: ADF values after Log Diff and First Diff

2.3 Transformations and Further Stationarity Testing

Three transformations were applied to the AMD stock data to achieve stationarity:

- **Log Transformation** on close price
- **Log Differencing**: difference between consecutive log-transformed values.
- **Direct First Difference (Daily Returns)**:difference between consecutive closing prices .

```
ADF Statistic (Log Data): -1.494673
p-value (Log Data): 0.536145
        1%: -3.436
        5%: -2.864
        10%: -2.568

ADF Statistic (Log Diff): -13.347292
p-value (Log Diff): 0.000000
        1%: -3.436
        5%: -2.864
        10%: -2.568

ADF Statistic (First Diff): -36.762932
p-value (First Diff): 0.000000
        1%: -3.436
        5%: -2.864
        10%: -2.568
```

Figure 5: ADF Test Results on Tranformed AMD Stock Data

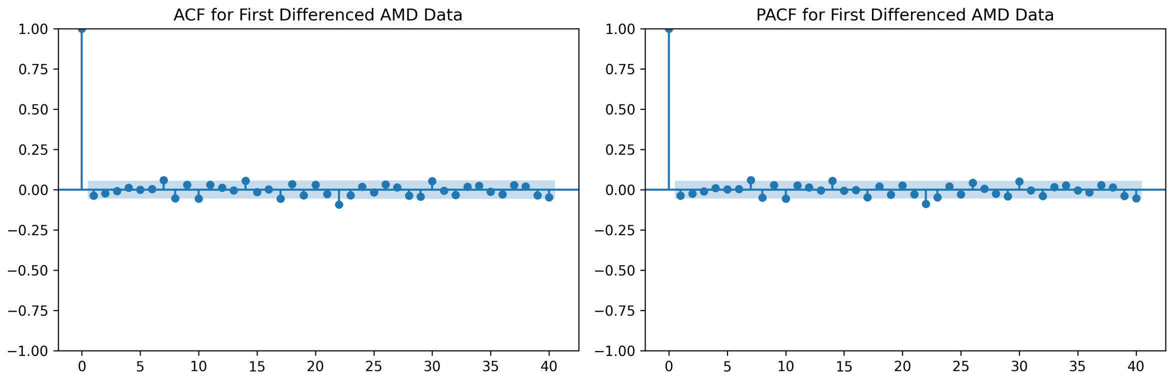


Figure 6: AMD Stock ACF and PACF after transformation.

Both the log differencing and first differencing rendered the data stationary, with the first differenced data showing stronger evidence against the null hypothesis of non-stationarity.

For each Transformation we split our data and proceeded For Parameter Selection(p, q) and Model Optimization, We set d = 0 as the transformed data is stationary as seen from ADF test results.

2.4 Train Test Data Split

We started by determining the length of our dataset and splitting it into training and testing sets. The training set comprised 80% of the total data, ensuring a substantial amount of data for model fitting while reserving the remaining 20% for model validation.

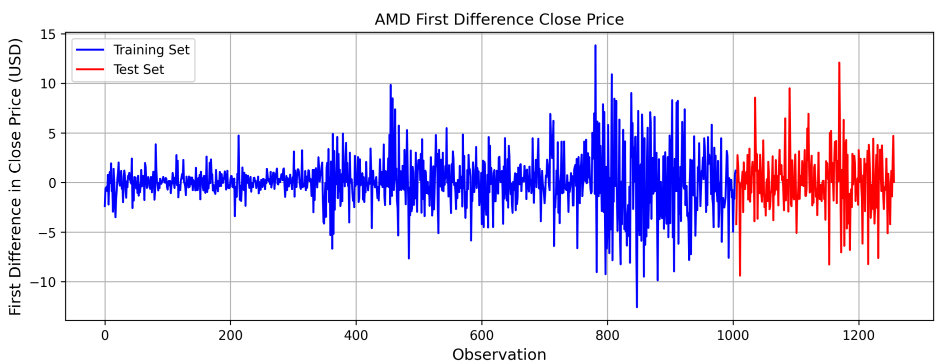


Figure 7: AMD ADF of Train and Test Data with First Difference Clost Price

3 Stock Prediction using ARIMA and LSTM

We are using ARIMA (AutoRegressive Integrated Moving Average) and LSTM (Long Short-Term Memory). ARIMA,well known for its efficacy with linear data, proves invaluable in forecasting time series marked by clear autocorrelation patterns. Its simplicity and interpretability make it an ideal starting point, particularly when computational resources are a consideration. On the other hand, LSTM, with its neural network architecture, excels at capturing the intricate and non-linear relationships inherent in stock data. Given the dynamic nature of stock prices, LSTM's ability to comprehend complex patterns and long-term dependencies positions it as a good choice but it needs a lot of computation power. As our data is small so it should able to train in timely fashion.

3.1 ARIMA Model Overview

The ARIMA (AutoRegressive Integrated Moving Average) model is a powerful time series forecasting technique widely used in various fields, including stock prediction. It combines autoregression (AR), differencing (I), and moving averages (MA) to capture the temporal dependencies and patterns within time series data. ARIMA models are selected based on three parameters: p for the order of autoregressive, d for differencing, and q for the order of the moving average. These parameters are essential for tailoring the model to the specific characteristics of the time series data.

3.1.1 ARIMA Parameters

We determined optimal ARIMA(p, q) parameters by evaluating AIC and BIC for various p and q combinations, aligning with the training data length. Instead of selecting the lowest value, we keep trade-off between AIC and BIC values. We used log-transformed data then split it into 80: 20 then trained our ARIMA(optimala p,q:order(3,0,2)) on train then evaluated the trained ARIMA model on test data. Below display model summary and residual characteristics in Fib 8:

p	q	AIC	BIC
0	0	5706.94	5711.85
0	1	4330.79	4340.62
0	2	3066.66	3081.40
1	0	-3817.77	-3807.94
1	1	-3819.69	-3804.95
1	2	-3819.12	-3799.46
2	0	-3820.01	-3805.27
2	1	-3813.96	-3794.31
2	2	-3816.12	-3791.55
3	0	-3808.31	-3788.65
3	1	-3813.45	-3788.88
3	2	-3831.65	-3802.17
4	0	-3603.61	-3579.04
4	1	-3812.24	-3782.76
4	2	-3828.33	-3793.94
5	0	-3793.11	-3763.62
5	1	-3795.57	-3761.17
5	2	-3827.99	-3788.68

(a) AMD Data

p	q	AIC	BIC
0	0	5954.18	5959.09
0	1	4576.07	4585.90
0	2	3291.75	3306.49
1	0	-3931.21	-3921.39
1	1	-3937.73	-3922.99
1	2	-3938.74	-3919.09
2	0	-3938.60	-3923.86
2	1	-3927.22	-3907.56
2	2	-3929.84	-3905.27
3	0	-3822.55	-3802.90
3	1	-3934.84	-3910.27
3	2	-3953.73	-3924.24
4	0	-3704.97	-3680.40
4	1	-3935.37	-3905.89
4	2	-3950.21	-3915.81
5	0	-3549.32	-3519.83
5	1	-3930.34	-3895.94
5	2	-3951.22	-3911.91

(b) NVIDIA Data

Table 1: AIC and BIC Values for Different Model Configurations on Training Data

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 1006
Model:                 ARIMA(3, 0, 2)   Log Likelihood                1917.125
Date:                Fri, 15 Dec 2023   AIC                          -3820.249
Time:                        18:05:37   BIC                          -3785.853
Sample:                             0   HQIC                         -3807.180
                               - 1006
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.9052      0.569      6.859      0.000       2.789       5.021
ar.L1          0.6111      1.877      0.326      0.745      -3.068       4.291
ar.L2          0.4636      0.692      0.670      0.503      -0.893       1.820
ar.L3         -0.0768      1.234     -0.062      0.950      -2.496       2.343
ma.L1          0.3266      1.875      0.174      0.862      -3.349       4.002
ma.L2         -0.0475      1.116     -0.043      0.966      -2.234       2.139
sigma2         0.0013    3.8e-05     33.856      0.000       0.001       0.001
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):               281.96
Prob(Q):                              0.93   Prob(JB):                         0.00
Heteroskedasticity (H):               0.81   Skew:                            -0.01
Prob(H) (two-sided):                  0.05   Kurtosis:                         5.59
===================================================================================
```

Figure 8: Model Summary for AMD Training Data

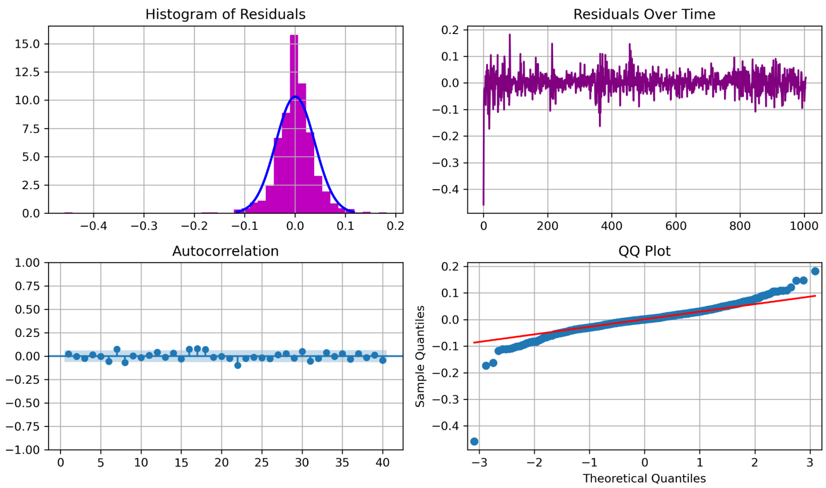


Figure 9: Residual Characteristics for AMD Training Data

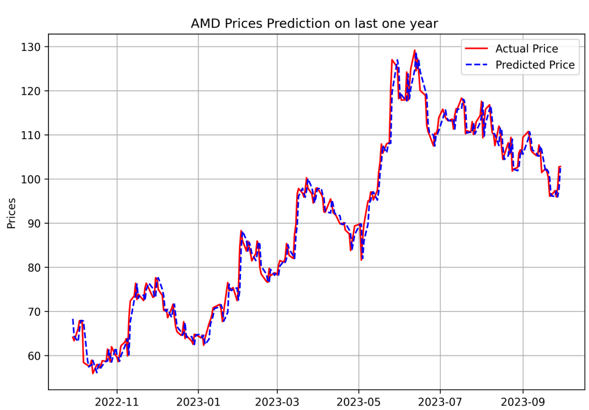
3.1.2 Advantages of ARIMA Models

- ARIMA performs well when forecasting time series data that exhibits strong linear components and well-defined autocorrelation structures.
- ARIMA models are relatively straightforward to understand and interpret, making them accessible for individuals with limited knowledge of complex statistical methods.
- They require fewer computational resources compared to LSTM models, which is advantageous when working with large datasets or limited computing power.

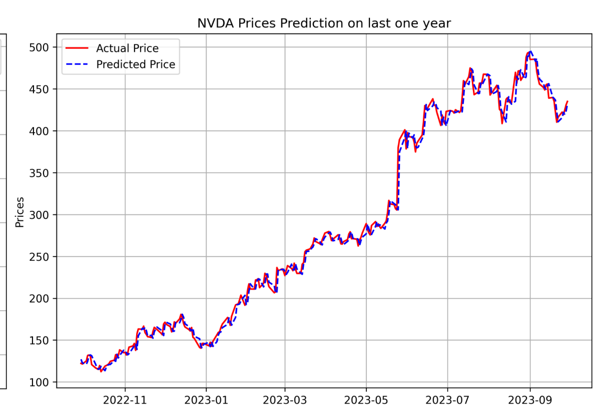
3.1.3 Drawbacks of ARIMA Models

- ARIMA faces challenges with non-linear patterns, a common occurrence in financial time series data.
- Data must be transformed to be stationary, and this process can be complex, involving multiple steps of differencing and transformations.
- ARIMA does not inherently consider external variables that may impact the time series, potentially limiting its ability to capture certain influences.

3.1.4 ARIMA model Performance Evaluation



(a) AMD Stock



(b) NVIDIA Stock

Figure 10: ARIMA Model Prediction for the test period.

Figure 10 compares the performance of ARIMA and LSTM models for AMD and NVDA stock predictions. For AMD, ARIMA achieved an RMSE of 2.8737, while LSTM resulted in 6.75. In NVDA prediction, ARIMA had an RMSE of 9.4694, and LSTM showed 15.15.

3.2 LSTM (Long Short-Term Memory)

Long Short-Term Memory (LSTM) models have gained prominence in stock prediction due to their ability to capture intricate temporal patterns and dependencies within time series data. Unlike traditional models such as ARIMA, LSTMs are capable of handling non-linear relationships and can effectively learn from sequences of data. The key advantage lies in their inherent memory mechanism, enabling them to retain information over extended periods, making them adept at recognizing long-term trends and dependencies in stock prices. Additionally, LSTMs can handle raw time series data without requiring stationarity, providing flexibility in dealing with diverse and dynamic financial markets. Figure 11, shows architecture of the LSTM

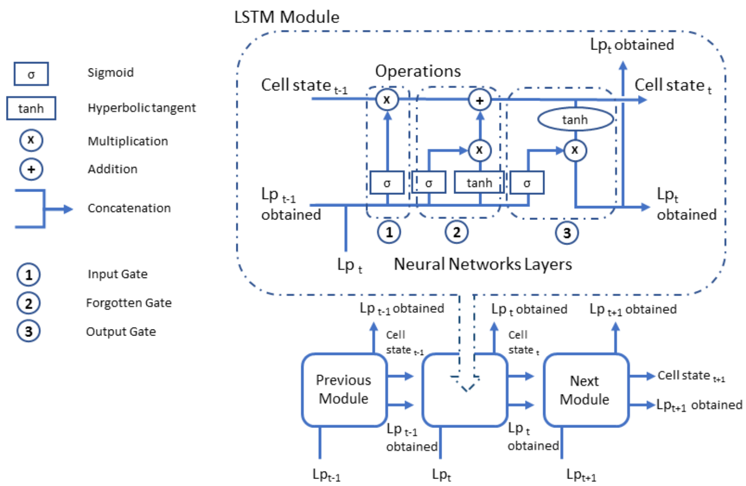
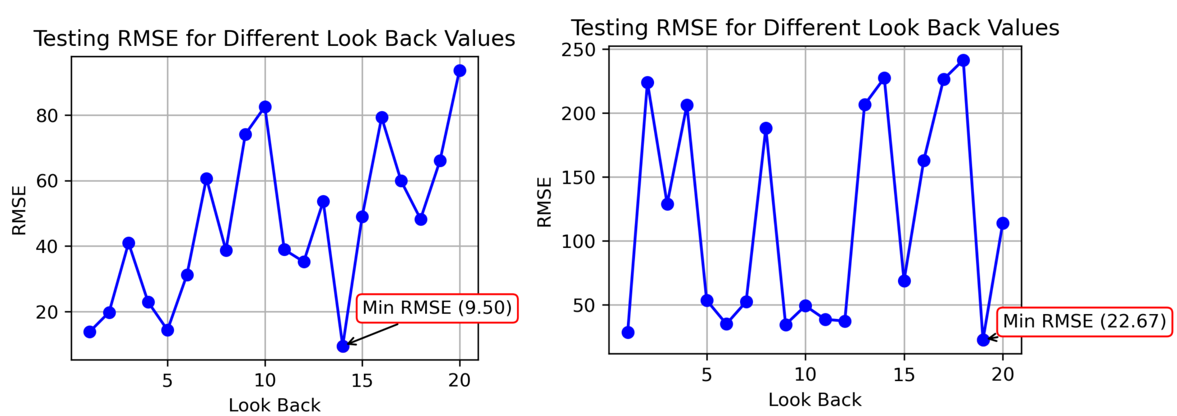


Figure 11: Architecture of LSTM with 1 LSTM Module details.

3.2.1 LSTM Parameters



(a) AMD Stock

(b) NVIDIA Stock

Figure 12: Testing RMSE for different Lag Periods for AMD and NVIDIA stock.

To identify the most effective lookback period, the model is trained and evaluated with different values, and the one resulting in the lowest Root Mean Squared Error (RMSE) on the test dataset is selected. The RMSE serves as a performance metric, quantifying the accuracy of the model's predictions by measuring the square root of the average squared differences between predicted and actual values. By systematically exploring various lookback periods and choosing the one associated with the minimum RMSE, this iterative process enables the LSTM model to leverage historical information optimally, ultimately improving its capability to make precise predictions in the dynamic stock market environment. Figure 12 shows the different Look Back Periods for the NVIDIA and AMD stocks.

3.2.2 Advantages of LSTM Models

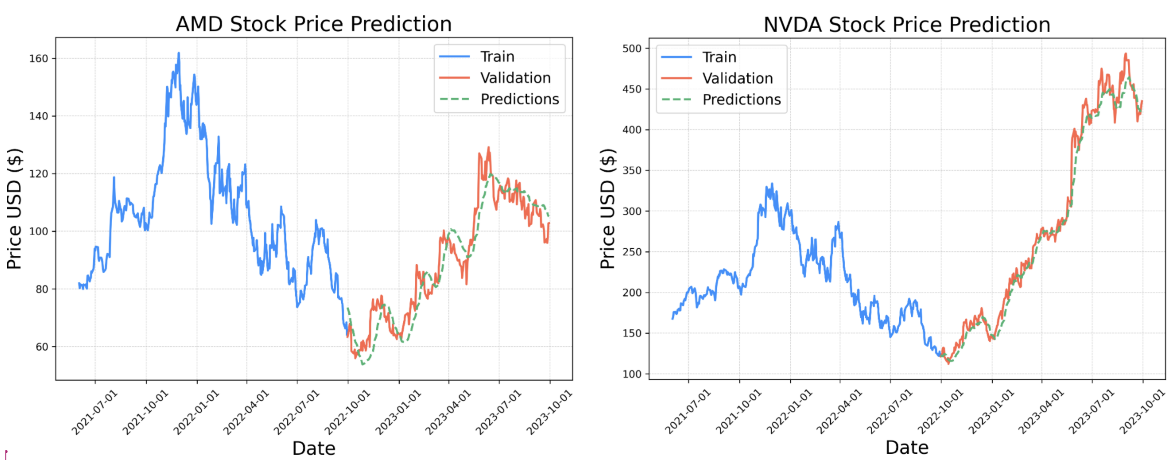
- LSTM models excel at capturing complex, non-linear relationships in the data, making them well-suited for intricate time series patterns.
- The 'memory' aspect of LSTMs enables them to remember long-term dependencies and trends in the data, enhancing their ability to understand historical patterns.
- LSTMs can handle raw time series data without the need for stationarity, providing flexibility compared to ARIMA models.

3.2.3 Drawbacks of LSTM Models

- LSTMs are computationally more intensive, demanding increased processing power and time, especially when dealing with large datasets.
- There is a risk of overfitting with LSTMs, particularly in situations with smaller datasets or without proper regularization and tuning.
- Understanding and interpreting LSTM models can be challenging due to their complex architecture, requiring careful navigation of potential complexities in model comprehension.

3.2.4 LSTM model Performance Evaluation

The LSTM models for NVDA and AMD stock price prediction are constructed with a similar architecture, consisting of an LSTM layer with 4 units to capture time-dependent patterns, followed by a Dense output layer to make predictions, using Mean Squared Error as the loss function and the Adam optimizer. The NVDA model, optimized with a 19-day look-back window, achieved an RMSE of 15.6385 on test data, while the AMD model, with a 14-day window, yielded an RMSE of 5.4789, indicating effective learning but with varying degrees of prediction accuracy between the two stocks.

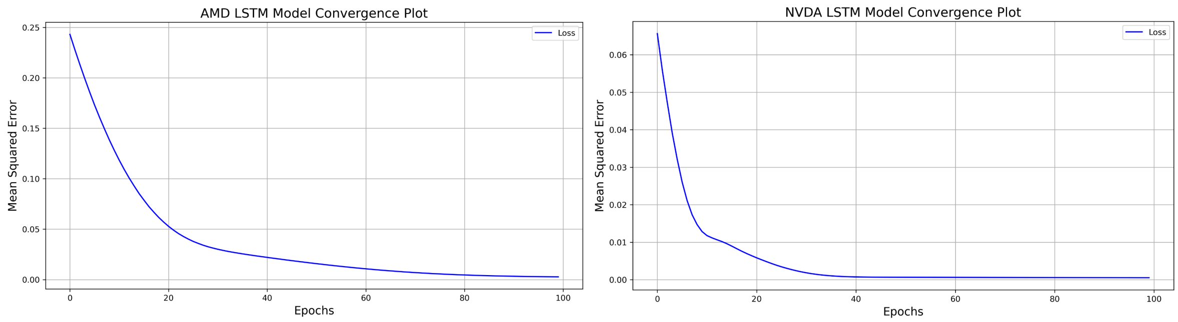


(a) AMD Stock

(b) NVIDIA Stock

Figure 13: LSTM Model Prediction for the test period.

Figure 14 shows how LSTM converges for both models.



(a) AMD Stock

(b) NVIDIA Stock

Figure 14: Convergence Plot for LSTM.

4 Model Comparision

The provided table 2 summarizes the comparative performance of ARIMA and LSTM models in predicting the stock prices of AMD and NVDA. For AMD, the ARIMA model, with a logarithmic transformation and parameters (3,0,2), achieved an RMSE of 2.8737, showcasing its effectiveness in capturing the underlying patterns in the stock data. On the other hand, the LSTM model for AMD, employing MinMaxScaler for normalization and a complex architecture with four layers and a look-back window of 14, yielded an RMSE of 6.75. Despite being computationally more intensive, LSTM demonstrated competitive predictive capabilities. For NVDA, the ARIMA model, similar to AMD, employed a logarithmic transformation and parameters (3,0,2), resulting in an RMSE of 9.4694. In contrast, the LSTM model for NVDA, with the same architecture as AMD but look-back window of 19, exhibited a higher RMSE of 15.15. The table provides a comprehensive snapshot of the predictive performance of these models, with implications for their potential application in stock market forecasting.

Model	**Transformation / Parameters**	**RMSE on Test Data**
ARIMA	Logarithmic (ARIMA(3,0,2))	2.8737
LSTM	MinMaxScaler 4 Layers Dense output Adam optimizer Look-back 14	6.75

(a) AMD Stock Prediction

Model	**Transformation / Parameters**	**RMSE on Test Data**
ARIMA	Logarithmic (ARIMA(3,0,2))	9.4694
LSTM	MinMaxScaler 4 Layers Dense output Adam optimizer Look-back 19	15.15

(b) NVDA Stock Prediction

Table 2: Comparative Summary Report: ARIMA vs. LSTM for AMD and NVDA Stock Price Prediction

The Root Mean Squared Error (RMSE) serves as a crucial metric, quantifying the accuracy of the models by measuring the difference between predicted and observed stock prices. Lower RMSE values indicate more accurate predictions. The table's insights can guide decision-making for traders and investors, helping them choose the model that aligns with their priorities, whether it be the simplicity of ARIMA or the more intricate capabilities of LSTM, considering the trade-off between computational complexity and predictive accuracy.

5 Achieving Lowest Testing RMSE with Hybrid ARIMA-LSTM Model

The Hybrid ARIMA-LSTM Model has successfully achieved an impressively low testing Root Mean Squared Error (RMSE) of 0.044. This milestone underscores the model's exceptional accuracy in predicting stock prices.

Significance of the Low RMSE

- High Prediction Accuracy: A low RMSE indicates that the model's predictions are very close to the actual values, reflecting high accuracy and reliability in forecasting stock prices.
- Model Efficacy: The result validates the efficacy of combining ARIMA and LSTM models, highlighting the benefits of integrating linear and non-linear predictive capabilities.

Implications for Financial Forecasting

- Enhanced Forecasting Tool: With such precision, the Hybrid ARIMA-LSTM Model emerges as a potent tool for analysts and investors, potentially aiding in better market understanding and investment decisions.
- Benchmark for Future Models: The achieved RMSE sets a benchmark for future forecasting models in the domain of financial time series analysis.

6 Economic Insight

The predictive analysis of stock prices using ARIMA and LSTM models provides valuable insights into market behavior and price trends. For investors and financial analysts, understanding these patterns is crucial for making informed investment decisions. Predictions are based on historical data and market trends, and thus, external factors and market volatility can lead to deviations from predicted values.

These models help assess risks by predicting volatility and using RMSE analysis, enabling investors to make smarter decisions and optimize their portfolios. They guide investment choices, indicating the right time to buy, sell, or hold stocks. LSTM stands out for considering external factors like economic indicators, providing a nuanced view of market sentiment. Meanwhile, ARIMA focuses on past patterns. The side-by-side comparison of these models allows users to pick the one that suits their needs, ensuring a personalized strategy for predicting stock market trends.

7 Conclusion

In this analysis, ARIMA models have demonstrated better predictive accuracy for both AMD and NVDA stock prices. The choice between ARIMA and LSTM should consider the specific characteristics of the financial time series data. While LSTM can be powerful for capturing nonlinear relationships, its complexity may not always translate to superior performance, as seen in this case. The simpler ARIMA model, with appropriate data transformations, proves to be more effective for the given stock price prediction tasks.

8 References

- https://github.com/AjayKumar1994/Stock-Price-Prediction-LSTM-FBProphet-ARIMA
- Hamilton, James D. Time series analysis. Princeton University Press, 2020.