Stock Market Forecasting in the Digital Age: Comparing Classical Statistical Models with Deep Learning Approaches for Financial Time Series Analysis

Anik Das (Roll No. 29954322012)
Adrika Mukherjee (Roll No. 29954322010)
Ananya Datta (Roll No. 29954322013)

BSc Data Science
Calcutta Institute of Engineering And Management (CIEM)

Introduction

- **Problem Background:** Financial markets are volatile and complex. Accurate stock price prediction is challenging due to economic indicators, geopolitical events, company performance, and investor sentiment.
- **Importance:** Crucial for informed investment decisions, risk management, and maximizing returns for individual and institutional investors.
- **Goals:** Examine the effectiveness of classical statistical methods and deep learning approaches (AR, ARIMA, SARIMAX, LSTM) in stock price forecasting using historical Yahoo stock data (2015-2020).

Data Description

- **Dataset Source:** Kaggle
- **Timeframe:** November 2015 – November 2020 (5-year span)
- **Variables:**
  - Date, Open, High, Low, Close, Volume, Adj Close
- **Notable Events:** Captured key economic phases, including COVID-19 pandemic and subsequent market volatility.

Objectives

- To explore and analyze historical stock price data to identify patterns and trends.
- To implement and evaluate various time series forecasting models and machine learning algorithms for stock price prediction.
- To compare the performance of different models using appropriate evaluation metrics.
- To develop a predictive model that offers insights into future stock price movements.

Preprocessing Steps

- **Data Loading and Inspection:**
  - Ensured completeness and quality of the dataset.
- **Missing Values:**
  - No missing values found in the dataset.
- **Train-Test Split Strategy:**
  - Dataset split into training (first 80%) and testing (remaining 20%) sets.
  - Dataset converting to time series framework.

Summary Statistics

- **Closing Price Statistics (Nov 2015 - Nov 2020):**
  - Minimum: $1,829
  - Maximum: $3,626.91
- **Statistics for other columns**

	Date	High	Low	Open
count	1825	1825.000000	1825.000000	1825.000000
mean	2018-05-23 00:00:00	2660.718673	2632.817580	2647.704751
min	2015-11-23 00:00:00	1847.000000	1810.099976	1833.400024
25%	2017-02-21 00:00:00	2348.350098	2322.250000	2341.979980
50%	2018-05-23 00:00:00	2696.250000	2667.840088	2685.489990
75%	2019-08-22 00:00:00	2930.790039	2900.709961	2913.860107
max	2020-11-20 00:00:00	3645.989990	3600.159912	3612.090088
std	NaN	409.680853	404.310068	407.169994

	Close	Volume	Adj Close
count	1825.000000	1.825000e+03	1825.000000
mean	2647.856284	3.869627e+09	2647.856284
min	1829.079956	1.296540e+09	1829.079956
25%	2328.949951	3.257950e+09	2328.949951
50%	2683.340088	3.609740e+09	2683.340088
75%	2917.520020	4.142850e+09	2917.520020
max	3626.909912	9.044690e+09	3626.909912
std	407.301177	1.087593e+09	407.301177

Figure: Summary Statistics for all the columns

Closing Price Time Series

- Overall upward trend from 2015 to 2020, price increasing from \$2,000 to over \$3,500.
- Significant drop in early 2020 during the COVID-19 pandemic, falling below \$2,250.
- Strong recovery following the pandemic-induced drop, reaching new highs by late 2020.

Figure: Stock Closing Price Over Time

Daily Returns Analysis

- Daily returns represent the percentage change in closing price, providing insights into day-to-day volatility.
- Relatively symmetric variation in daily returns,with return falls within -2% to +2%.
- Noticeable deviation during COVID-19 outbreak: sharp and abrupt decline in returns, heightened volatility, and unprecedented dip with daily returns exceeding ±5%.

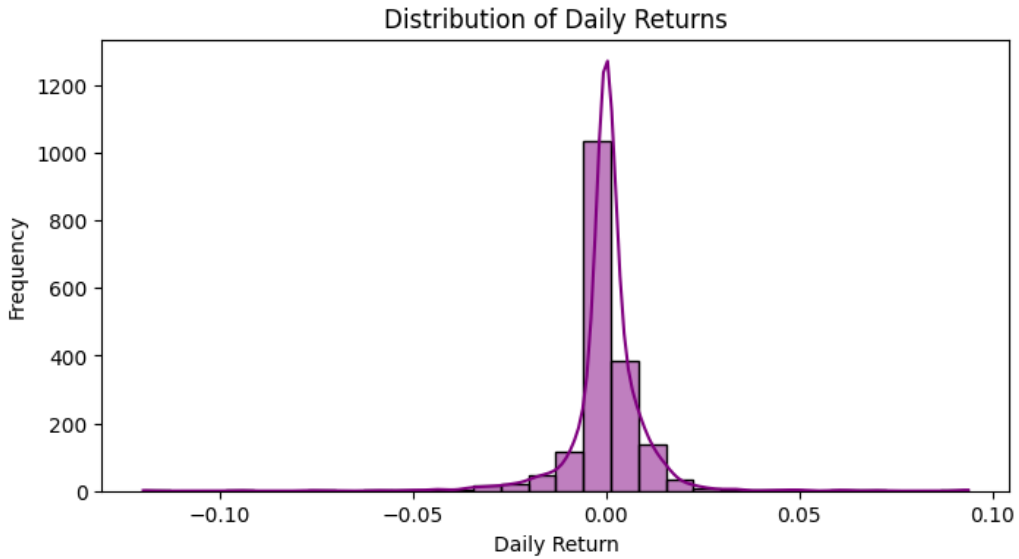

Figure: Distribution of Daily Returns

Volume Traded Over Time

- Trading volume is an important indicator of market activity.
- Volume varies significantly over time, with several notable spikes.
- Highest trading volumes occurred during the COVID-19 market crash in March 2020.
- Periods of high volatility in prices generally coincide with increased trading volumes.
- Average daily trading volume is approximately 3.87 billion shares.

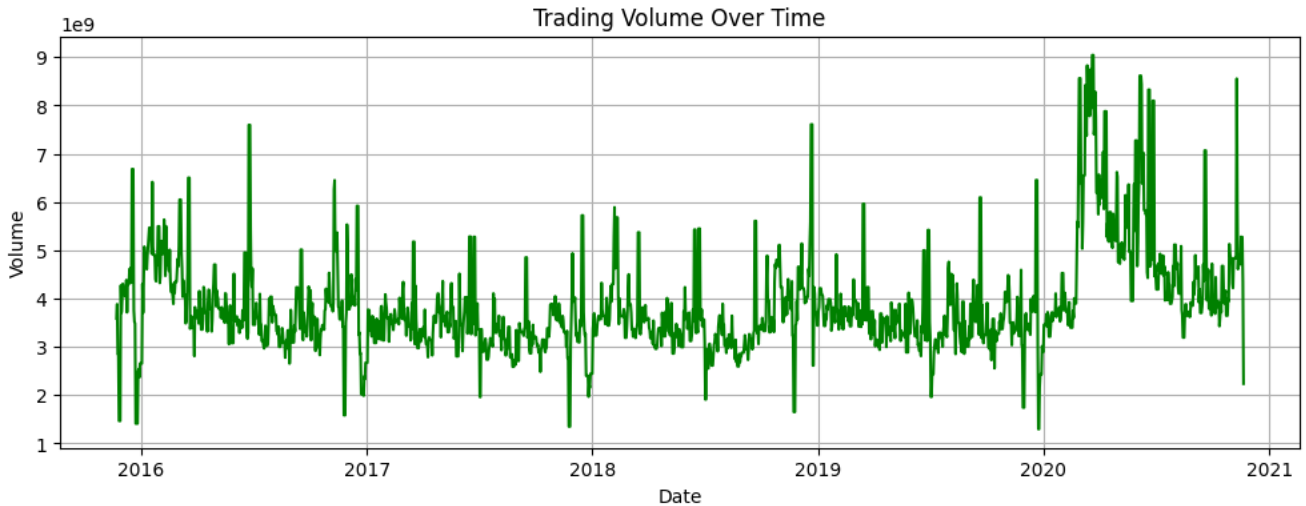

Figure: Trading Volume Over Time

Moving Averages (MA50 & MA200)

- Moving averages smooth price data to identify trends and filter noise.
- MA50 (orange line) responds more quickly to price changes than MA200 (green line).
- During strong uptrends, price consistently stayed above both moving averages.
- During COVID-19 crash, price fell significantly below both moving averages before recovering.
- MA200 acted as a support level during several pullbacks.

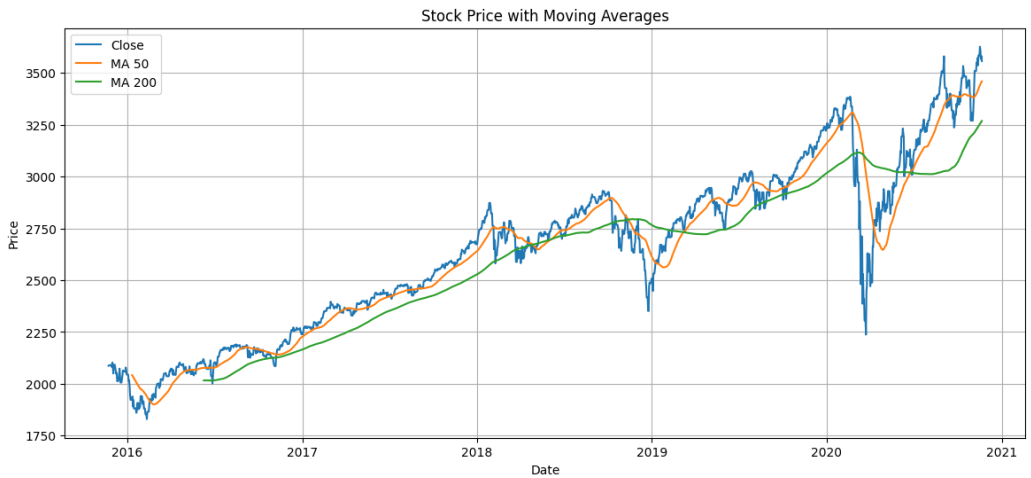

Figure: Stock Price with Moving Averages

Bollinger Bands

- Bollinger Bands consist of a middle band (20-day MA) and two outer bands (two standard deviations away).
- Help identify periods of high/low volatility and potential overbought/oversold conditions.
- Periods of high volatility (wide bands) occurred during market stress (early 2016, late 2018, COVID-19 in 2020).
- Periods of low volatility (narrow bands) observed in mid-2017 and early 2019.

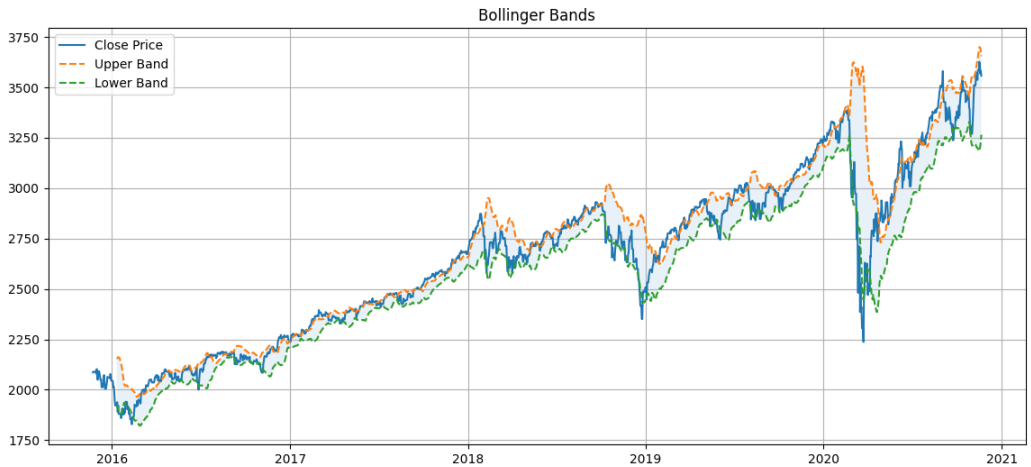

Figure: Bollinger Bands

COVID-19 Crash Visualization

- Time series plot of Yahoo's daily closing stock prices (Nov 2015 - Nov 2020), overlaid with a smoothed trend line.
- Highlights the 2020 market crash due to the COVID-19 pandemic.
- Blue line: actual closing prices (significant short-term fluctuations and overall volatility).
- Red trend line: clear long-term upward trajectory.

Figure: Yahoo Stock Close Price and Trend with 2020 Market Crash

Time Series Decomposition

- Breaks down a time series into its constituent components: trend, seasonality, and residual (or random) components.
- **Trend Component:**
  - Clear long-term upward movement in stock price.
  - Significant dip during COVID-19 pandemic, followed by strong recovery.
- **Seasonal Component:**
  - Minimal or negligible seasonality detected, typical for daily financial time series.
- **Residual Component:**
  - Captures irregular fluctuations due to external shocks or market events.

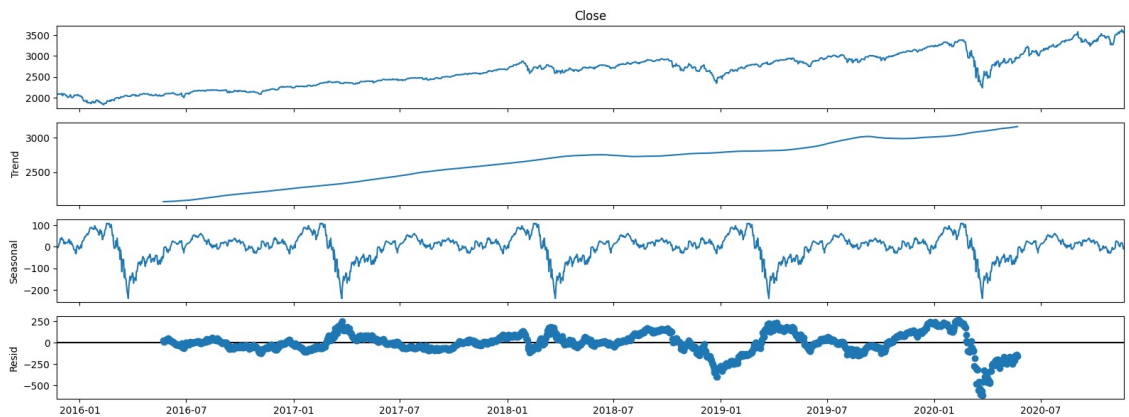

Figure: Time Series Decomposition

ACF and PACF Plots

- **Insights:**
  - ACF shows a slow decay,in the original price series.
  - PACF cuts off after lag 2, suggesting an AR(2) model for the time series.

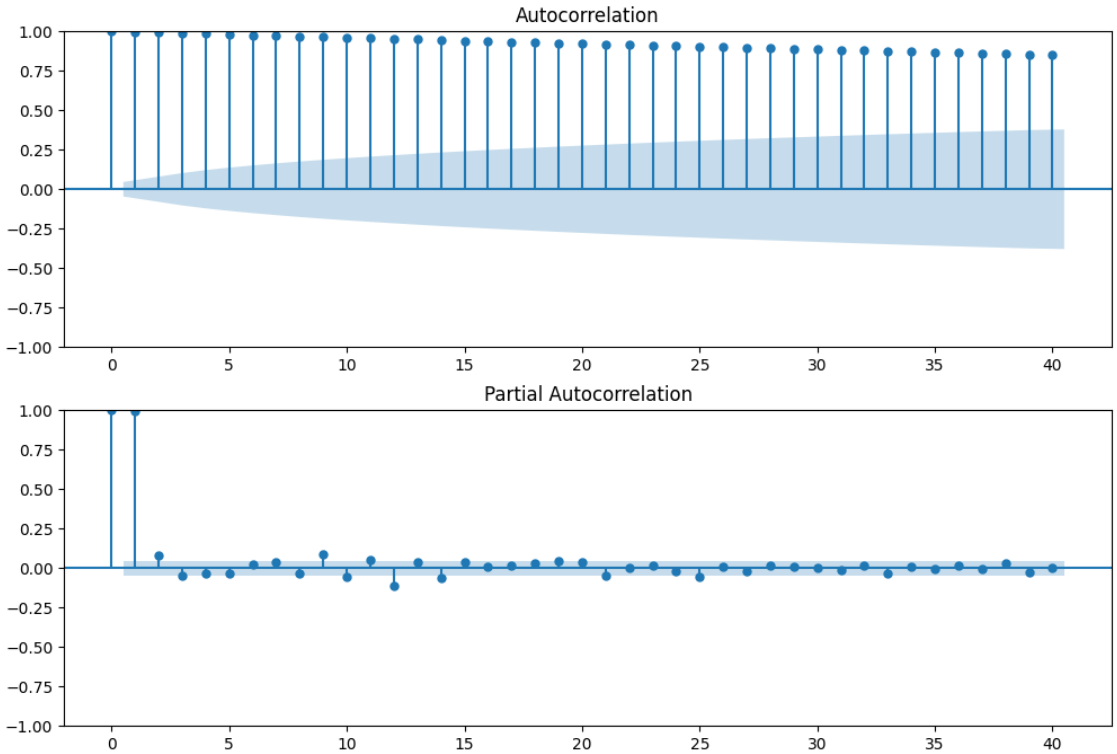

Figure: ACF and PACF Plots

AR(2) Model

- **Autoregressive (AR) Model:** Predicts future values based on past values.
- **AR(2) Model:** Current value is a linear combination of the two previous values plus a random error term.
- **Equation:**

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$$

Results and Forecast of AR(2) Models

- **Efficiency of the model**:
  - RMSE for AR(2): 26.9785
  - AIC: 17195.1613
  - BIC: 17217.1942
- The AR(2) model showed moderate performance on the training data, capturing some of the patterns in the daily returns but missing many of the extreme movements.
- The AR(2) model results show:
  - The model captures some of the patterns in the daily returns but misses many of the extreme movements.
  - The predictions are much less volatile than the actual returns
  - The RMSE,AIC, and BIC values show good performance of the model.

ARIMA Models

- **ARIMA(1,1,1) Model:**
  - Combines autoregressive (AR) and moving average (MA) components with an integration (I) component for non-stationarity.
  - Demonstrated better in-sample fit than the AR model, capturing overall trend and some fluctuations.
- **ARIMA(2,0,1) Model:**
  - Improved forecasting accuracy compared to ARIMA(1,1,1).
  - Captures overall direction of price movements and predicts changes better.

Forecast for ARIMA(2,0,1)

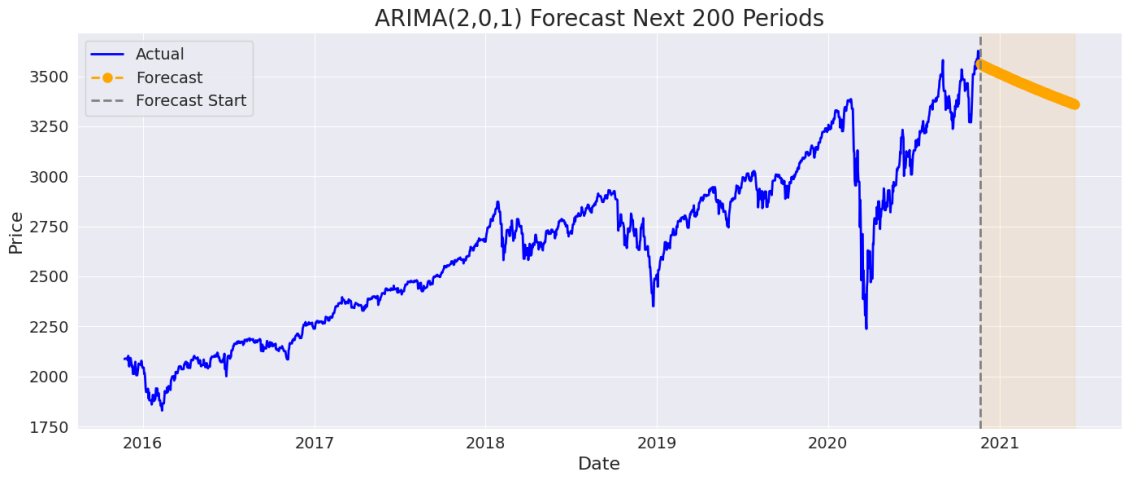

Figure: Forecast for ARIMA(2,0,1)

The ARIMA model results show:

- The model captures the overall direction of price movements and predicts changes better than the earlier ARIMA model.
- The model's performance metrics indicate poor predictive power: - RMSE: - 29.9667 .

Model Comparison of AR and ARIMA

- Multiple models implemented and evaluated for stock price prediction: AR(2), ARIMA(1,1,1), and ARIMA(2,0,1).
- The bar diagram below helps in comparing the performance metrics of the different models.

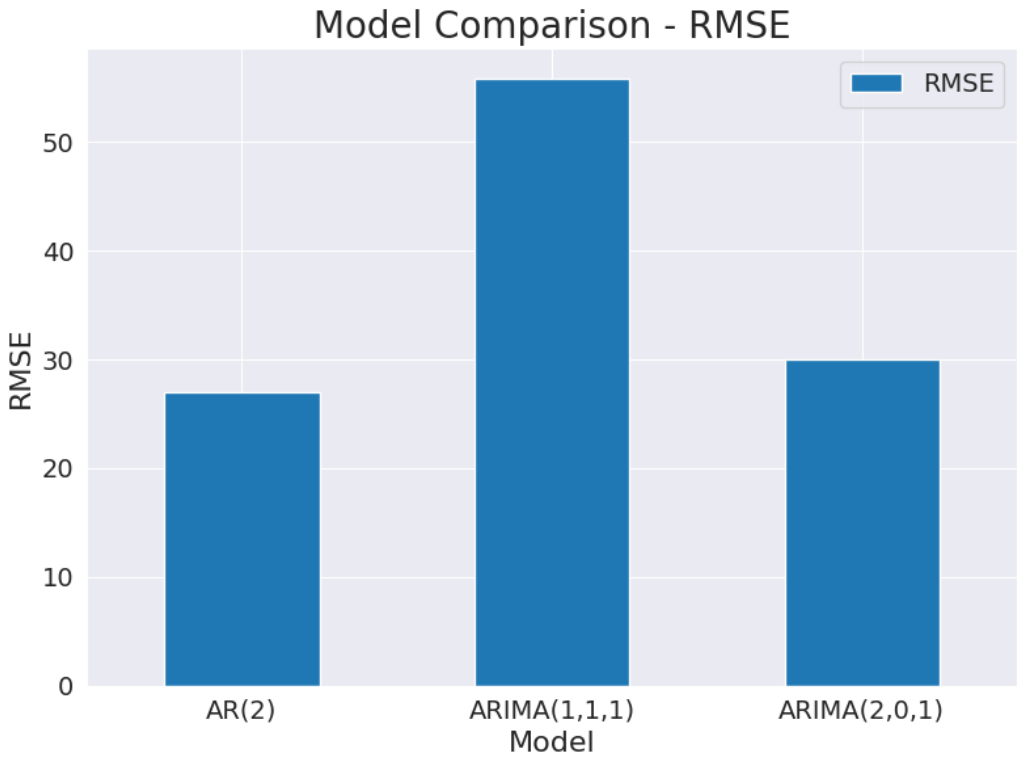

Figure: Model Comparison - RMSE

SARIMAX Models (3 types)

- **SARIMAX(1,1,1)(1,1,1,12):**
  - Combines non-seasonal and seasonal components (AR, differencing, MA).
  - Non-seasonal (1,1,1) for short-term patterns; seasonal (1,1,1,12) for yearly patterns.
  - Balances complexity with flexibility for trends and seasonal behavior.
- **SARIMAX(2,0,1)(1,1,1,12):**
  - Skips non-seasonal differencing (assumes stationary data).
  - Uses AR(2) for complex short-term patterns and MA(1) for noise.
  - Ideal for data with strong autocorrelation but no clear trend.
- **SARIMAX(2,1,2)(1,0,0,12):**
  - Sophisticated non-seasonal structure with AR(2) and MA(2) terms, plus first-order differencing.
  - Simpler seasonal component with AR(1) on a 12-period cycle.
  - Excels at capturing yearly patterns in data with mild seasonality but complex short-term fluctuations.

SARIMAX Model Results Comparison

To check the efficiency of the SARIMAX models, bar diagrams are drawn to show the RMSE, AIC and BIC of the models.

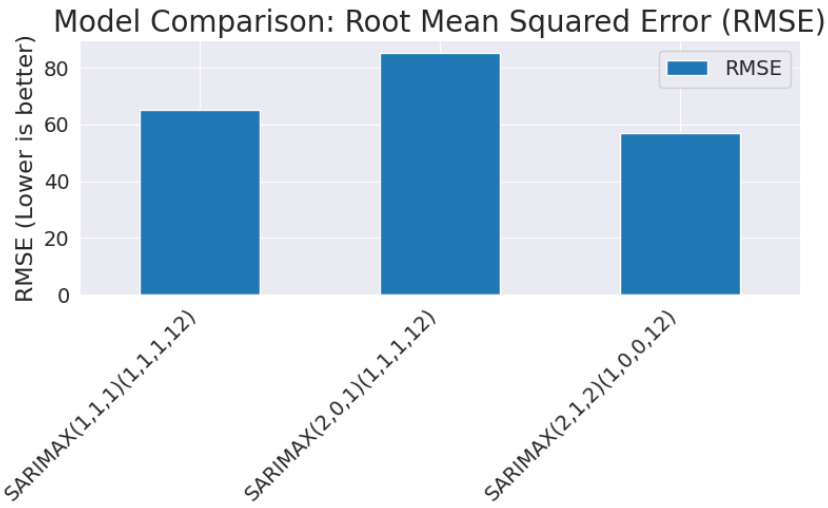

Figure: RMSE of the three SARIMAX models

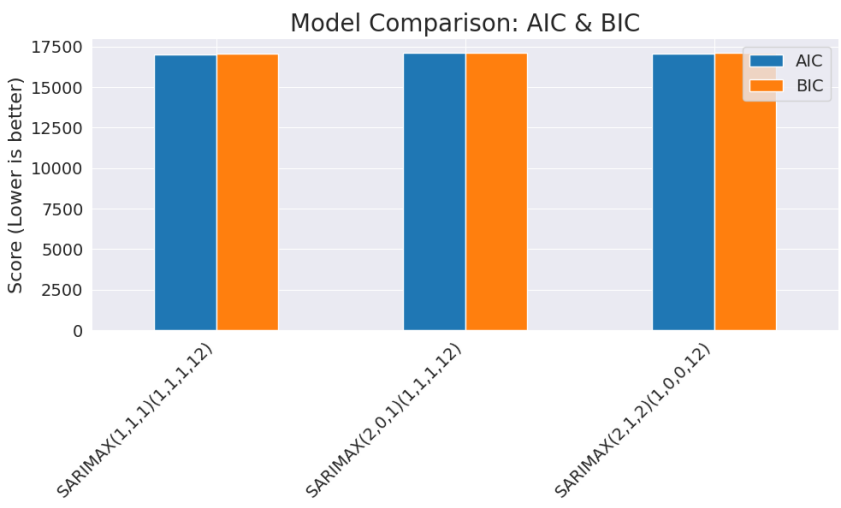

Figure: AIC and BIC of the three SARIMAX models

LSTM Model Overview

- **Long Short-Term Memory (LSTM):** A type of recurrent neural network (RNN) capable of learning long-term dependencies.
- **Key Features:**
  - **Memory Cells:** Maintain information over extended periods.
  - **Gates (Input, Forget, Output):** Regulate the flow of information into and out of the cell.
- **Suitability for Time Series:** Excellent for sequential data like stock prices due to their ability to remember past information.

LSTM Training Performance

- The LSTM model was trained on the historical stock price data.
- **Training Loss:** Measures how well the model fits the training data.
- **Observations:**
  - The model converged, showing a good fit without significant overfitting.
- **Performance Metrics:**
  - Root Mean Squared Error:56.66954104951507
  - Mean Absolute Error: 36.0699983134575
  - R-squared Score: 0.980255328086482

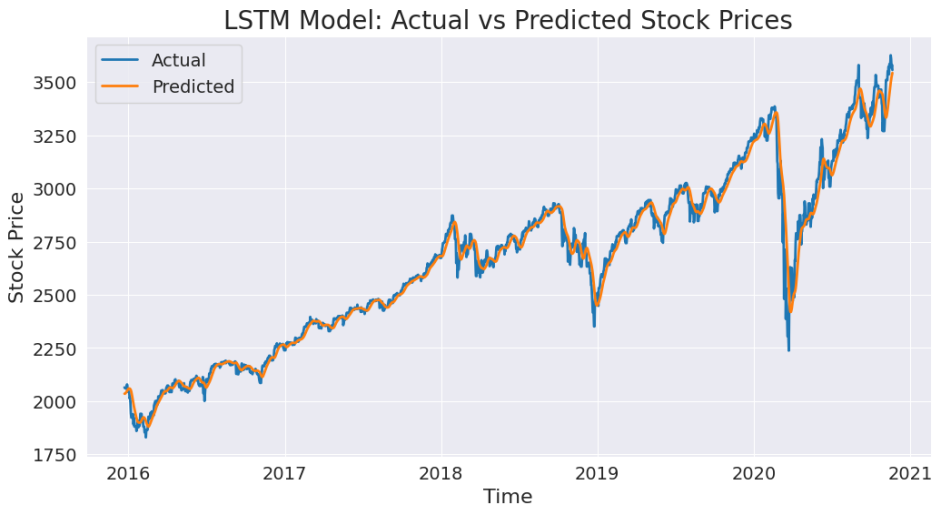

Figure: LSTM Model: Actual vs Predicted Stock Prices(Training Set)

LSTM Test Performance

- The trained LSTM model was evaluated on the test set.
- **Actual vs. Predicted Prices:**
  - The modelś predictions closely follow the actual stock prices on the test set.
  - Captures both the overall trend and short-term fluctuations effectively.
- **Performance Metrics:**
  - Root Mean Squared Error: 105.81455873855155
  - Mean Absolute Error: 78.52686838984157
  - R-squared Score: 0.8516092453913142

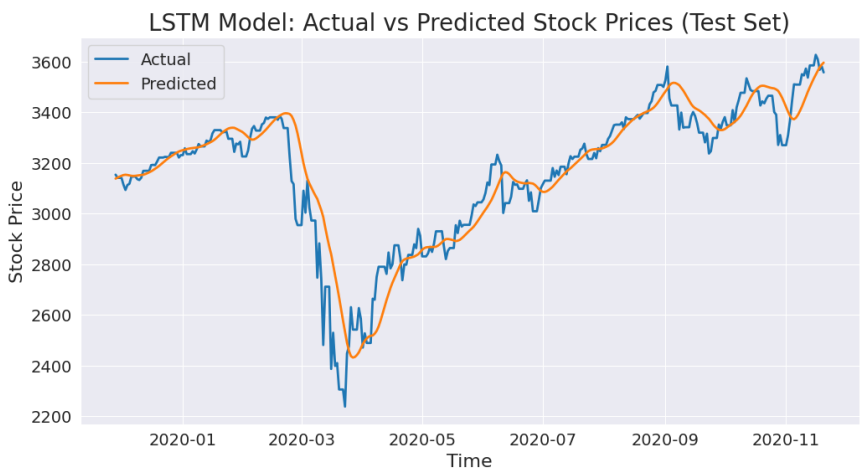

Figure: LSTM Model: Actual vs Predicted Stock Prices (Test Set)

Training History

- The plot illustrates the training and validation loss over epochs,
- **Observations:**
  - Both training and validation loss curves show a rapid decrease in the initial epochs, indicating that the model is learning effectively.
  - The curves then flatten out, suggesting that the model has converged and further training would yield diminishing returns.
  - The validation loss remains close to the training loss, which is a good indicator that the model is not overfitting to the training data and is generalizing well to unseen data.

Figure: Model Training History

Model Comparison Summary

- **Classical Models (AR, ARIMA, SARIMAX):**
  - Effective for capturing linear relationships and basic trends.
  - Struggle with high volatility and non-linear patterns.
  - SARIMAX models showed improved performance with seasonal components.
- **Deep Learning Model (LSTM):**
  - Demonstrated superior performance in capturing complex, non-linear patterns and long-term dependencies.
  - More robust in volatile market conditions.
- **Overall:**
  - LSTM generally outperforms classical models in terms of accuracy for stock price forecasting.
  - Classical models provide a good baseline and are simpler to interpret.

Continuation

Table: Model Performance Comparison (RMSE)

Model	**RMSE**
AR(2)	26.978460
ARIMA(1,1,1)	55.778002
ARIMA(2,0,1)	29.966678
SARIMAX(1,1,1)(1,1,1,12)	65.1399
SARIMAX(2,0,1)(1,1,1,12)	85.3284
SARIMAX(2,1,2)(1,0,0,12)	56.8642
LSTM(Train)	56.669541
LSTM(Test)	105.8145587

Limitations

- **Classical Models:**
  - **Linearity Assumption:** Struggle with non-linear stock market dynamics.
  - **Sensitivity to Volatility:** Performance degrades during high market volatility.
  - **Feature Engineering:** Limited ability to incorporate external factors without manual feature engineering.
- **Deep Learning Models (LSTM):**
  - **Data Intensive:** Require large datasets for optimal performance.
  - **Computational Cost:** Training can be computationally expensive and time-consuming.
  - **Interpretability:** Often considered a "black box" due to their complex internal workings, making it difficult to understand how predictions are made.

Future Scope

- **Hybrid Models:** Combine classical and deep learning approaches to leverage strengths of both.
- **External Factors:** Incorporate macroeconomic indicators, news sentiment, and social media data.
- **Advanced Deep Learning Architectures:** Explore attention mechanisms, Transformers, or Generative Adversarial Networks (GANs).
- **Real-time Forecasting:** Develop models for continuous, real-time stock price prediction.
- **Risk Management Integration:** Integrate forecasting models with portfolio optimization and risk management strategies.

References

The following resources are recommended:

- Box, G. E. P., Jenkins, G. M. (1970). Time Series Analysis: Forecasting and Control. Holden-Day.
- Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press.
- Murphy, J. J. (1999). Technical Analysis of the Financial Markets. New York Institute of Finance.
- Sezer, O. B., Gudelek, M. U., Ozbayoglu, A. M. (2020). Financial Time Series Forecasting with Deep Learning: A Systematic Literature Review. Applied Soft Computing, 90, 106181.
- Hochreiter, S., Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.
- Time Series Analysis, By James D.Hamilton
- Time Series, By Peter J Brockwell and Richard A Davies
- ChatGPT
- Wikipedia

Thank You