University of Applied Sciences
HOCHSCHULE EMDEN·LEER

ProjectPlan

Comparision of ARIMA,SARIMA and LSTM Time series models on stock market data

Author:	Sandesh Nonavinakere Sunil
Matriculation No.:	7026004
Course of Studies:	Business Intelligence and Data Analytics

Author:	Nishan Poojari
Matriculation No.:	7026796
Course of Studies:	Business Intelligence and Data Analytics

First examiner:	Prof. Dr. Elmar Wings
Submission date:	April 13, 2025

University of Applied Sciences Emden/Leer · Faculty of Technology · Mechanical Engineering Department
Constantiaplatz 4 · 26723 Emden · http://www.hs-emden-leer.de

Contents

List of Figures

List of Tables

Acronyms

1 Overview

Stock market data analysis using machine learning involves leveraging historical and real-time data to predict stock prices, detect patterns, and optimize trading strategies. The process begins with data collection and preprocessing, where stock prices, trading volume, and financial indicators are gathered, cleaned, and normalized. Advanced deep learning models like Long Short-Term Memory (LSTM) networks, ARIMA and SARIMA are used for time-series forecasting.

2 Problem Statement

Stock market price prediction is a complex task due to its highly volatile and non-linear nature. Traditional statistical methods often fail to capture long-term dependencies and intricate patterns in financial time-series data. This project aims to develop a robust stock market forecasting model using Long Short-Term Memory (LSTM), AutoRegressive Integrated Moving Average (ARIMA), SARIMA

3 Data

NIFTY 50 is a Indian stock market index that represents the performance of the top 50 companies listed on the National Stock Exchange (NSE) of India. It is one of India's leading stock indices and serves as a benchmark for the overall stock market performance.

3.0.1 Key Features of NIFTY 50

- Composition – Includes 50 large, liquid, and financially stable companies from various sectors.
- Market Representation – Covers about 65
- Sector Diversity – Comprises companies from banking, IT, pharmaceuticals, energy, and other key sectors.

3.0.2 Data Availability

The data for this project has been downloaded from Kaggle

4 Data Exploration

4.0.1 Data Loading and Initial Inspection

- File Source: The data is loaded from a CSV file. This file is expected to contain daily market data spanning from 2000 to 2024.
- Date Conversion: The Date column is converted to a datetime object and set as the DataFrame index. This is essential for time-series analysis because it allows us to work with date ranges and ensures proper temporal ordering.
- Initial Preview: By printing the first 20 rows (using df.head(20)), we can visually inspect the data and verify that the values look as expected. For example:
  - The first few rows show daily values for Open, High, Low, and Close.
  - Fundamental ratios such as P/E, P/B, and Div Yield
  - Data Summary: The df.info() function is used to review:
    - The total number of entries (rows) in the dataset.
    - The data type of each column.
    - Non-null counts for each column, which help identify missing values.

4.0.2 Detailed Column Summary

- Column: Date
  - Data Type: object
  - Number of Entries (non-null): 5970 / 5970 (100.00 Perc)
  - Missing Values: 0 (0.00 Perc)
- Column: Open
  - Data Type: float64
  - Number of Entries (non-null): 5970 / 5970 (100.00 Perc)
  - Missing Values: 0 (0.00 Perc)
- Column: High
  - Data Type: float64
  - Number of Entries (non-null): 5970 / 5970 (100.00 Perc)
  - Missing Values: 0 (0.00 Perc)
- Column: Low
  - Data Type: float64
  - Number of Entries (non-null): 5970 / 5970 (100.00 Perc)
  - Missing Values: 0 (0.00 Perc)
- Column: Close
  - Data Type: float64
  - Number of Entries (non-null): 5970 / 5970 (100.00 Perc)
  - Missing Values: 0 (0.00 Perc)
- Column: P/E
  - Data Type: float64
  - Number of Entries (non-null): 5970 / 5970 (100.00 Perc)
  - Missing Values: 0 (0.00 Perc)
- Column: P/B
  - Data Type: float64
  - Number of Entries (non-null): 5970 / 5970 (100.00 Perc)
  - Missing Values: 0 (0.00 Perc)
- Column: Div Yield
  - Data Type: object
  - Number of Entries (non-null): 5970 / 5970 (100.00 Perc)
  - Missing Values: 0 (0.00 Perc)

4.0.3 Data Transformation

- Handling Placeholder Values: In many datasets, missing or invalid values might be represented by placeholders such as "-". We replace these placeholders with NaN
- Filtering Weekends: Since stock markets operate only on weekdays, the data is filtered to remove weekends. The DayOfWeek column is derived from the index (where Monday is 0 and Friday is 4), and only rows with values less than 5 are retained
- Creating a COVID Dummy Variable: A new binary column, COVIDdummy, is created to capture the market conditions during the COVID period (set to 1 between 2020-03-01 and 2020-12-31).

Final Dataset Size: After all cleaning and filtering, the final dataset size is printed. For example: Final dataset size after cleaning: 5944 rows.

5 Data Storage

We plan on Using Git for storing stock market data in CSV

- Version Control – You can track changes to the data over time, see previous versions, and revert if needed.
- Collaboration – If multiple people need to update or review the file, Git provides a structured way to manage changes.
- Backup and Security – GitHub/GitLab/Bitbucket provide cloud-based storage with some level of protection.

6 Success Criteria

The performance of these models will be evaluated using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared values to determine the most effective approach for stock price forecasting. The insights gained from this study will assist us in making data-driven decisions, improving risk assessment, and optimizing investment strategies.

7 Keywords used for references

The below are the keywords used to identify references related to our project

- ARIMA
- LSTM
- Time series analysis
- Stock market prediction
- COVID - 19 stock market prediction

8 References

1. Kuang, S., (April 2023) *A Comparison of Linear Regression, LSTM model and ARIMA model in Predicting Stock Price: A Case Study: HSBC's Stock Price.* School of Economics and Management, South China Normal University, Guangzhou, China. Available at: https://www.researchgate.net/publication/370573980_A_Comparison_of_Linear_Regression_LSTM_model_and_ARIMA_model_in_Predicting_Stock_Price_A_Case_Study_HSBC's_Stock_Price [Accessed 6 Apr. 2025]. doi: http://dx.doi.org/10.54691/bcpbm.v44i.4858

2. Bao, W., Yue, J. and Rao, Y., (2017) *A deep learning framework for financial time series using stacked autoencoders and long-short term memory.* PLoS ONE, 12(7), p.e0180944. Available at: https://doi.org/10.1371/journal.pone.0180944 [Accessed 6 Apr. 2025].

3. Åkesson, A. and Holm, A., (31 October 2024) *A comparison of LSTM and ARIMA forecasting of the stock market.* KTH Royal Institute of Technology. Available at: https://kth.diva-portal.org/smash/get/diva2:1942061/FULLTEXT01.pdf [Accessed 6 Apr. 2025].

4. Zakamulin, V. (2016) *Market Timing with Moving Averages: Anatomy and Performance of Trading Rules.* [Working paper, revision of May 29, 2016]. Available at: https://ssrn.com/abstract=2459384 doi: 10.2139/ssrn.2459384 [Accessed 6 Apr. 2025].

5. Chong, E., Han, C. and Park, F.C. (15 October 2017) *Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies.* Available at: https://www.researchgate.net/publication/316373382_Deep_Learning_Networks_for_Stock_Market_Analysis_and_Prediction_Methodology_Data_Representations_and_Case_Studies doi: http://dx.doi.org/10.1016/j.eswa.2017.04.030 [Accessed 6 Apr. 2025].

6. Pilla, P. and Mekonen, R. (2025) *Forecasting S&P 500 Using LSTM Models.* Available at: https://zenodo.org/records/14759118 doi: https://doi.org/10.5281/zenodo.14759117 [Accessed 6 Apr. 2025].

7. Ho, M.K., Darman, H. and Musa, S. (2021) *Stock Price Prediction Using ARIMA, Neural Network and LSTM Models, Journal of Physics: Conference Series, 1988(1), 012041.* Available at: https://doi.org/10.1088/1742-6596/1988/1/012041 doi: 10.1088/1742-6596/1988/1/012041 [Accessed 6 Apr. 2025].

8. Kashifa, K. and Ślepaczuk, R. (26 June 2024) *LSTM-ARIMA as a Hybrid Approach in Algorithmic Investment Strategies.* Available at: https://arxiv.org/abs/2406.18206v1 doi: https://doi.org/10.48550/arXiv.2406.18206 [Accessed 6 Apr. 2025].

9. Xiao, R., Feng, Y., Yan, L. and Ma, Y. (31 August 2022) *Predict Stock Prices with ARIMA and LSTM.* Available at:https://arxiv.org/abs/2209.02407 doi: https://doi.org/10.48550/arXiv.2209.02407 [Accessed 6 Apr. 2025].

10. Wang, C., 2024. *Stock price forecasting by ARIMA, linear regression, LSTM and decomposition linear models. Proceedings of the 4th International Conference on Signal Processing and Machine Learning.* Available at:https://doi.org/10.54254/2755-2721/55/20241572 doi: https://doi.org/10.54254/2755-2721/55/20241572 [Accessed 6 Apr. 2025].

11. Wang, Z., 2024. *Stock price prediction using LSTM neural networks: Techniques and applications. Proceedings of the 6th International Conference on Computing and Data Science.* Available at: https://doi.org/10.54254/2755-2721/86/20241605 doi: https://doi.org/10.54254/2755-2721/86/20241605 [Accessed 6 Apr. 2025].

12. Majumder, A., Rahman, M.M., Biswas, A., Zulfiker, M.S. and Basak, S., 2022. *Stock Market Prediction: A Time Series Analysis.* In: S. C. Satapathy, V. Bhateja and S. Das, eds. Smart Systems: Innovations in Computing. Singapore: Springer, pp.389–401. Available at: https://www.researchgate.net/publication/354362969_Stock_Market_Prediction_A_Time_Series_Analysis doi: https://doi.org/10.1007/978-981-16-2877-1_35 [Accessed 6 Apr. 2025].

13. Ruan, J., Wu, W. and Luo, J., n.d. Stock Price Prediction Under Anomalous Circumstances. Department of Computer Science and Goergen Institute for Data Science, University of Rochester. available at : https://www.researchgate.net/publication/354985130_Stock_Price_Prediction_Under_Anomalous_Circumstances doi: http://dx.doi.org/10.48550/arXiv.2109.15059 [Accessed 6 Apr. 2025].

14. Qiu, J., Wang, B. and Zhou, C., 2020. *Forecasting stock prices with long-short term memory neural network based on attention mechanism.* Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0227222 doi: https://doi.org/10.1371/journal.pone.0227222 [Accessed 6 Apr. 2025].

15. Fu, W., Li, Z., Zhang, Y. and Zhou, X., 2023. *Time Series Analysis in American Stock Market Recovering in Post COVID-19 Pandemic Period. Account and Financial Management Journal, 8(2), pp.3081–3086.* Available at: https://www.researchgate.net/publication/368690273_Time_Series_Analysis_in_American_Stock_Market_Recovering_in_Post_COVID-19_Pandemic_Period doi: https://doi.org/10.47191/afmj/v8i2.03 [Accessed 6 Apr. 2025].