Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

ISSN NO. 2320-5407

Article DOI: 10.21474/IJAR01/22239
DOI URL: http://dx.doi.org/10.21474/IJAR01/22239

RESEARCH ARTICLE

A MACHINE LEARNING FRAMEWORK FOR FORECASTING TOP GAINERS IN THE "NIFTY50" INDEX

Vijayalakshmi P[1], Dr. Praveena Murala[2], Bogani Ramadevi[2], Mupparaju Lavanya[2] and Popuri Anusha[2]

1. PG Student.
2. Assistant Professor, Department of CSE, QIS College of Engineering and Technology (A), Ongole, Andhra Pradesh, India.

Manuscript Info

Manuscript History
Received: 04 September 2025
Final Accepted: 06 October 2025
Published: November 2025

Key words:-
NIFTY50, stock forecasting, machine learning, ensemble models, Ridge regression, XGBoost,LSTM, consensus ranking

Abstract

This research introduces a machine learning framework for forecasting top gainers in the NIFTY50 index. The framework refreshes five years of historical stock data daily from Yahoo Finance and applies interpolat ion,normalization, and feature engineering (SMA, EMA, RSI, MACD). Six different models were considered namely Ridge Regression, Decision Tree, Random Forest, XGBoost, LightGBM, & LSTM are evaluated using MAE, RMSE, and R^2. Ridge Regression provided strong baselines (NTPC with $R^2 \approx 0.965$), ensembles delivered stability under volatility, and LSTM captured sequential trends with sufficient data. A consensus approach merged per-model rankings into a final top-5 picklist (IOC.NS, ONGC.NS, NTPC.NS, CIPLA.NS, BPCL.NS). Results demonstrate that combining diverse models offers more reliable stock selection than relying on a single regression baseline.

Introduction:-

Financial markets are inherently volatile, making short-term stock prediction a challenging problem.The NIFTY50, which consists of 50 of the most actively traded companies on the National Stock Exchange of India, is commonly used as a benchmark for making investment decisions. Capturing its price dynamics requires models capable of handling non-linear dependencies, market noise, and temporal patterns. Machine learning (ML) has shown strong potential for financial forecasting, offering tools to learn complex relationships beyond traditional statistical methods. However, relying on a single model such as linear regression limits robustness. This paper proposes a unified ML pipeline that refreshes five years of NIFTY50 data from Yahoo Finance, engineers indicators such as SMA, EMA, MACD, and RSI, and evaluates six model families: Ridge Regression, Decision Tree, Random Forest, XGBoost, LightGBM, and LSTM. To improve reliability, we introduce a consensus mechanism that merges model rankings into a final top-5 picklist, providing investors with practical, reproducible insights.

Literature Review:-

Stock market prediction has been widely explored through statistical methods, machine learning approaches, and deep learning techniques and deep learning approaches. Linear regression provided simple baselines but was limited in handling nonlinear financial data. Decision trees introduced nonlinearity but often overfit, while Random Forest improved generalization by aggregating multiple trees [1]. Boosting frameworks such as XGBoost [2] and LightGBM [3] Reached top performance with efficient learning and feature importance ranking. More recently, recurrent neural networks, particularly LSTMs [4], have been applied These models help identify patterns over time in financial data, but they work best when large datasets are available and careful tuning. Overall, prior work demonstrates that while advanced architectures improve accuracy, simpler models remain competitive in financial forecasting. This motivates the present study to evaluate classical, ensemble, and deep models for forecasting top gainers in the NIFTY50 index, India's leading benchmark.

Table I — Summary of Related Work

Author & Year	Technique(s)	Key Findings
Fama [1]	Efficient Market Hypothesis	Stock markets are largely efficient, making prediction difficult.
Breiman [2]	Random Forest	Ensemble decision trees improve stability and reduce overfitting
Chen & Guestrin [3]	XGBoost	Boosting yields higher predictive accuracy than single models.
Ke et al. [4]	LightGBM	Gradient boosting improves scalability and efficiency.
Hochreiter & Schmidhuber [5]	LSTM	Captures temporal dependencies in sequential financial data.

Methodology:-

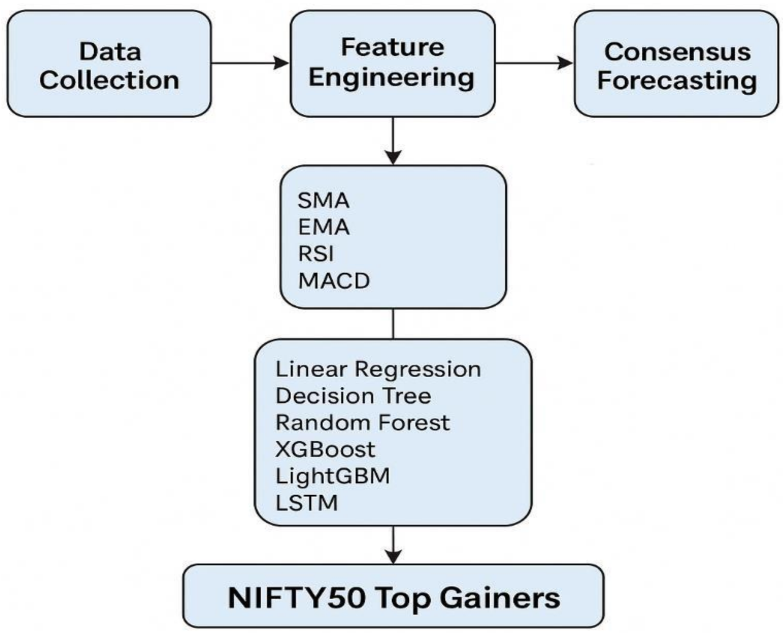
The proposed framework is designed as a reproducible pipeline for forecasting top gainers in the NIFTY50 index. It integrates data acquisition, feature engineering, predictive modeling, and consensus-based ranking.

Data Collection and Preprocessing:-

Daily OHLCV data for all NIFTY50 constituents are automatically refreshed from Yahoo Finance, ensuring the dataset remains current up to the latest trading day. Missing values are handled through interpolation and forward–backward filling. Features are normalized before modeling.

Feature Engineering:-

Technical indicators are computed to enrich the dataset: simple and exponential moving averages (SMA, EMA), moving average convergence– divergence (MACD), relative strength index (RSI), lagged closing prices and returns, rolling volatility, and day-of-week dummy variables. The prediction target is the next-day closing price.



Models:-

Six model families are implemented:

- Ridge Regression (RidgeCV): a regularized linear model used as a baseline.
- Decision Tree: rule-based predictor with high interpretability but prone to overfitting.
- Random Forest: ensemble of trees to reduce variance and improve robustness.
- XGBoost: boosting framework offering high predictive accuracy through sequential correction.
- LightGBM: scalable gradient boosting optimized for speed.
- LSTM: recurrent neural network model that learns patterns and relationships across long sequential datasets.

Evaluation and Consensus Ranking:-

Models are trained with an 80/20 temporal split and Analyzed using mean absolute error (MAE), root mean square error (RMSE), and R^2. Each model generates a top-5 ranked list of stocks by expected return. These lists are combined into a consensus table, where stocks are prioritized by the number of model votes, weighted expected return, and median return. The final output is a robust top-5 shortlist of investment candidates.

Results and Analysis:-

The models were assessed using MAE, RMSE, and R^2 on an 80/20 temporal split.

Model-wise Results:-

Ridge Regression served as a strong baseline, achieving $R^2 \approx 0.965$ for NTPC with low error, as shown in Table II. The predicted vs actual closing price of NTPC is illustrated in Fig. 2, confirming that Ridge Regression tracks short-term price movements effectively. Decision Trees highlighted high-return opportunities such as IOC and BPCL but tended to overfit. Random Forests improved stability across stocks, with IOC predictions shown in Fig. 3. Gradient boosting methods (XGBoost, LightGBM) delivered high accuracy and efficiency. LSTM captured sequential dependencies, demonstrated by ONGC results in Fig. 4, though training time was higher.

Fig. 2 — Predicted vs Actual Closing Price of NTPC (Ridge Regression)

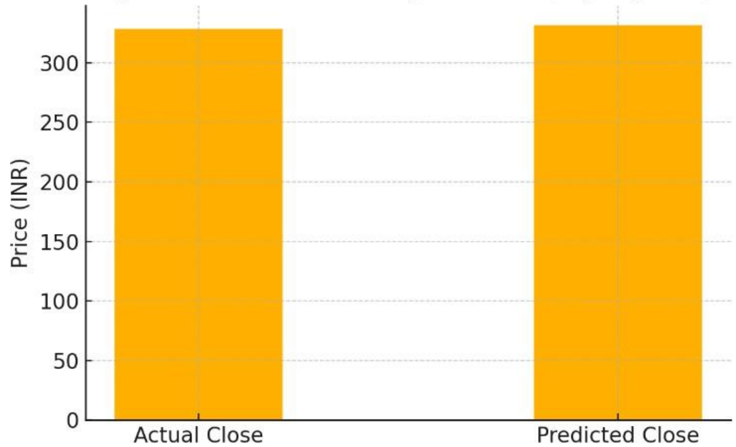


Fig. 3 — Predicted vs Actual Closing Price of IOC (Random Forest)

Fig. 4 — Predicted vs Actual Closing Price of ONGC (LSTM)

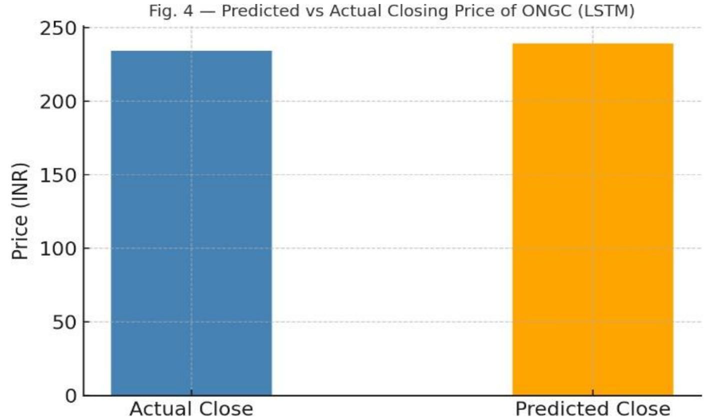


Table II — Top-5 stocks predicted by Ridge Regression

Ticker	Last Date	Actual Close	Predicted Close	Expected Return (%)	R^2
NTPC.NS	2025-09-05	328.70	331.67	0.90	0.965

POW ERG RID. N S	2025 -09- 05	285.35	287.43	0.73	0.93 6
ADANI ENT.NS	2025 -09- 05	2281.4 0	2297.95	0.73	0.93 6
M&M.N S	2025 -09- 05	3561.3 0	3582.08	0.58	0.91 8
ONGC. NS	2025 -09- 05	234.13	235.37	0.53	0.90 0

Comparative Performance:-

Table III summarizes the comparative performance of all six models. Ridge performed best on stable stocks, ensembles reduced variance, and boosting methods improved predictive accuracy. LSTM's performance varied but added sequential insights.

Table III — Comparative Performance of Models on NIFTY50 Data

Model	**Best R^2 (example stock)**	**Test MAE (avg range)**	**Test RMS E (avg range)**	**Remarks**
Ridge Regression	0.965 (NTP C)	Low (3–45)	Low (5–67)	Strong baseline; robust for stable stocks.
Decision Tree	0.692 (NTP C)	Hig h (9–25)	High (13–29)	Captures non-linearities but prone to overfitting.

Rando m Forest	0.81 6 (NT P C)	Med ium (9– 26)	Mediu m (10– 34)	Stable across tickers; less variance than DT.
XGBo ost	0.94 4 (HE R OM O TO C O)	Med ium (6– 11)	Mediu m (7– 13)	Boosting improves accuracy and generalizatio n.
Light GBM	0.94 0 (HE R OM O TO C O)	Med ium (8– 43)	Mediu m (9– 55)	Faster boosting with competitive accuracy.
LSTM	0.88 7 (UP L, WIP R O)	Med ium – Hig h (5– 41)	Mediu m– High (7– 53)	Captures sequential patterns; requires longer data.

Consensus Ranking:-

Since individual models produced varying top-5 lists, a consensus method was applied. The final shortlist, shown in Table IV, was derived by combining model votes, weighted expected return, and median return. The top-5 consensus picks (IOC, ONGC, NTPC, CIPLA, BPCL) are further analyzed in Figs. 5–6.

Fig. 6 — Consensus Vote Distribution of Final Top-5 Stocks

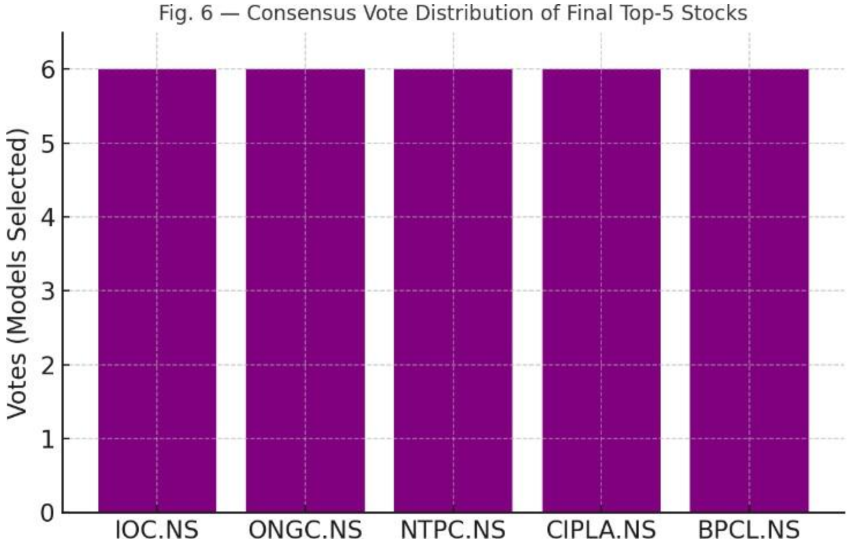


Table IV — Final Consensus Top-5 Picklist Across Six Models

Rank	Ticker	Models Selected (Votes)	Avg. Expected Return (%)	Weighted Return (%)	Remarks
1	IOC.NS	6	6.11	3.15	Consistently ranked high by all models.
2	ONGC.NS	6	3.55	2.80	Strong signals across ensembles and LSTM.
3	NTPC.NS	6	2.44	1.90	Stable baseline, supported by Ridge & RF.
4	CIPLA.NS	6	1.76	1.59	Reliable healthcare stock, moderate return.
5	BPCL. NS	6	2.20	1.46	Energy sector stock, robust across models.

Visual Analysis:-

- Fig. 1 — Proposed NIFTY50 forecasting pipeline.
- Fig. 2 — Predicted vs actual closing price of NTPC using Ridge Regression.
- Fig. 3 — Predicted vs actual closing price of IOC using Random Forest.
- Fig. 4 — Predicted vs actual closing price of ONGC using LSTM.
- Fig. 5 — Average expected returns of consensus top-5 stocks.
- Fig. 6 — Consensus vote distribution of final top-5 stocks.

Figures 2–4 confirm that the models accurately track price movements for representative stocks. Figure 5 shows the expected returns of consensus picks, while Figure 6 highlights the equal vote distribution across the final top-5.

Discussion:-

The results demonstrate that while Ridge Regression alone achieves high accuracy, ensemble methods provide greater robustness under volatility, and LSTM captures temporal patterns. The consensus framework integrates these strengths, producing a reliable and reproducible shortlist of investment candidates.

Conclusion:-

This paper presented a machine learning framework for forecasting top gainers in the NIFTY50 index. Historical price data was enriched with technical indicators and processed using six predictive models: Ridge Regression, Decision Tree, Random Forest, XGBoost, LightGBM, and LSTM. While Ridge Regression provided a strong baseline with high accuracy, ensemble methods improved stability under volatility, and LSTM captured temporal dependencies. To address variability across models, a consensus approach was introduced, producing a robust shortlist of five stocks— IOC, ONGC, NTPC, CIPLA, and BPCL—consistently selected by all models. The results demonstrate that consensus learning enhances reliability in stock prediction. Future work may incorporate sentiment analysis and macroeconomic variables to further improve forecasting performance.

References:-

1. E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," Journal of Finance, vol. 25, no. 2, pp. 383-417, 1970.
2. B. G. Malkiel, "The Efficient Market Hypothesis and Its Critics," Journal of Economic Perspectives, vol. 17, no. 1, pp. 59-82, 2003.
3. Y. Tang, Z. Song, and Y. Zhu, "A Survey on Machine Learning Models for Financial Time Series Forecasting," Neurocomputing, vol. 512, pp. 363-380, 2022.
4. National Stock Exchange of India, "NIFTY 50 Index Methodology," NSE Indices Limited, Mumbai, India, 2023.
5. Z. Fathali, Z. Kodia, and L. B. Said, "Stock Market Prediction of NIFTY 50 Index Applying Machine Learning Techniques," Applied Artificial Intelligence, vol. 36, no. 1, pp. 2111134, 2022.
6. G. Singh, "Machine Learning Models in Stock Market Prediction," arXiv preprint arXiv:2202.09359, 2022.
7. Singh, P. Gupta, and N. Thakur, "An Empirical Research and Comprehensive Analysis of Stock Market Prediction Using Machine Learning and Deep Learning Techniques," IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, 2021.
8. V. R. Jain, M. Gupta, and R. M. Singh, "Analysis and Prediction of Individual Stock Prices of Financial Sector Companies in NIFTY50," International Journal of Information Engineering and Electronic Business, vol. 13, no. 2, pp. 33-42, 2018.
9. A. Sachdeva et al., "An Effective Time Series Analysis for Equity Market Prediction Using Deep Learning Model," 2019 International Conference on Data Science and Communication (IconDSC),IEEE, 2019.
10. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
11. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.
12. G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Advances in Neural Information Processing Systems, vol. 30, 2017.
13. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
14. J. Bai and S. Ng, "Forecasting Economic Time Series Using Targeted Predictors," Journal of Econometrics, vol. 146, no. 2, pp. 304-317, 2008.
15. R. Cont, "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues," Quantitative Finance, vol. 1, pp. 223-236, 2001.
16. W. Lo and A. C. MacKinlay, "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test," Review of Financial Studies, vol. 1, no. 1, pp. 41-66, 1988.
17. D. M. Diebold and R. S. Mariano, "Comparing Predictive Accuracy," Journal of Business & Economic Statistics, vol. 13, no. 3, pp. 253-263, 1995.
18. J. H. Stock and M. W. Watson, "Forecasting Using Principal Components from a Large Number of Predictors," Journal of the American Statistical Association, vol. 97, no. 460, pp. 1167-1179, 2002.
19. R. F. Engle and C. W. J. Granger, "Co-integration and Error Correction: Representation, Estimation, and Testing," Econometrica, vol. 55, no. 2, pp. 251-276, 1987.
20. P. Samuelson, "Proof That Properly Anticipated Prices Fluctuate Randomly," Industrial Management Review, vol. 6, no. 2, pp. 41-49, 1965