Enhancing Intraday Trading through Machine Learning: A NSE Nifty 50 Analysis

***Abstract*—This paper explores the intersection of machine learning (ML), intraday trading, and the economic landscape of India, focusing on the National Stock Exchange (NSE) Nifty 50 index. The study evaluates various ML models, including K-Nearest Neighbors (KNN), Random Forest Regressor (RF), Decision Tree Regressor, Support Vector Regression (SVR), and Linear Regression (LR), revealing LR as the most proficient model with a remarkable R2 score exceeding 0.997 across $CLOSE$, $HIGH$, and LOW values.**

A key observation is that traditional ML techniques primarily rely on historical data for predictions. However, for intraday strategies, the incorporation of $OPEN$ prices on the day of prediction proves to be a valuable addition. Upon closer examination, the approach utilizing combined data, utilizing both $OPEN$ prices and historical data, demonstrates superior performance with the highest R2 scores (0.99940 and 0.99903) and lowest errors in the $HIGH$ and LOW cases. Conversely, the second approach, excluding $OPEN$ prices, exhibits the least error and the highest R2 score (0.99998) for $CLOSE$ price prediction. In conclusion, incorporating $OPEN$ data in intraday strategies emerges as a compelling and insightful method, significantly enhancing predictive accuracy.

The findings highlight the pivotal role of ML in predicting stock prices within India's NSE Nifty 50, not only providing accurate forecasts but also optimizing intraday trading decisions. As financial markets continue to evolve, the integration of ML and intraday strategies emerges as a promising avenue for achieving more informed and data-driven investment practices in the economic landscape.

I. Introduction

The economic vitality of any system relies on financial activities, propelling progress and well-being. However, in a global landscape where a significant portion of the population grapples with poverty, it becomes crucial to extend beyond conventional financial interactions to ensure inclusive growth opportunities for everyone [1]. India is transitioning to a circular economy, emphasizing corporate contributions to sustainable practices and resource efficiency in alignment with governmental goals [2].

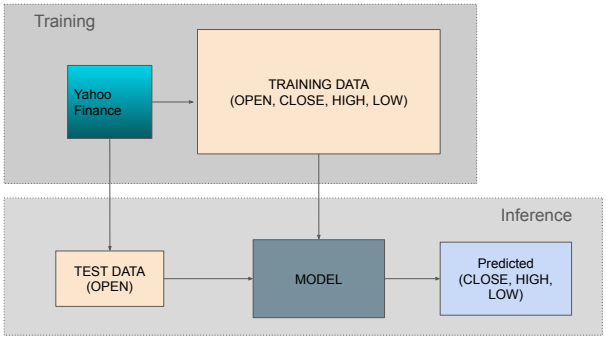
Trading, whether in traditional stock markets or the rapidly expanding realm of cryptocurrencies, is a dynamic force that molds global financial landscapes. It stands as a pivotal catalyst for economic activity, offering individuals and corporations avenues to engage in markets. As nations like India shift towards sustainable practices, the influence of trading on environmental, social, and governance (ESG) factors becomes increasingly important [3]. Inclusive trading practices have the potential to contribute not just to financial prosperity but also to ethical and sustainable investments, aligning with overarching economic objectives.

India's stock market journey, spanning from the establishment of the Bombay Stock Exchange (BSE) [4] in 1875 to contemporary electronic trading, mirrors the nation's economic growth. This transformation is in harmony with India's shift toward a circular economy and sustainable financial practices. Serving as a vital player, the stock market directs investments towards environmentally conscious and socially responsible endeavors. This historical context underscores the pivotal role of financial markets in India's economic advancement, fostering the vision of inclusive growth and sustainable financial practices.

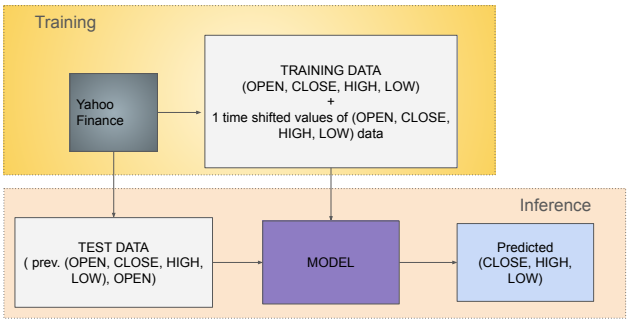
Originating in 1992, the National Stock Exchange (NSE) [5] of India has emerged as a prominent player in the country's financial landscape. Its establishment marked a shift towards electronic trading, introducing modern technologies for seamless and efficient transactions. The NSE Nifty 50 Index, a flagship index, comprises the 50 most liquid and large-cap stocks listed on the NSE, serving as a benchmark for market performance. Beyond the Nifty 50, the NSE hosts various sectoral and thematic indices, contributing to a holistic representation of India's diverse economic sectors. These indices play a pivotal role in guiding investment decisions, providing investors with a comprehensive view of the Indian stock market and facilitating diversified portfolio management.

Machine Learning (ML) and Artificial Intelligence (AI) have revolutionized stock price prediction, leveraging advanced algorithms to analyze vast financial data and identify patterns. These technologies provide valuable insights, enhance decision-making, and mitigate risks for traders and investors. AI/ML has also led to the development of algorithmic trading strategies, automating trades based on real-time analysis [6]. Forecasting index prices accurately is a challenging task due to the dynamic and often unpredictable nature of factors such as economic indicators, geopolitical events, and market sentiment. The inherent volatility of financial markets presents an additional layer of complexity to forecasting, as market conditions can rapidly change, resulting in fluctuations in index prices. Nonetheless, AI/ML continues to play a growing role, contributing to more informed and data-driven investment strategies in the financial landscape.

The rest of the paper is structured as follows: Section II provides an overview of related studies conducted thus far. Section III provides a comprehensive discussion of dataset for the analysis of various models, elucidated in Section IV. The subsequent Section V explain the performance metrics used. The findings of this study are presented in Section VI, and finally, Section VII concludes the study.



(a) System Model Employing $OPEN$ Data Approaches



(b) System Model Merging Traditional and $OPEN$ Data Approaches

Fig. 1: System Models for $OPEN$ and Integrated Approach: Incorporating NSE Index Values from Yahoo Finance

II. Related Works

[6] gives an elaborate discussion of algorithmic trading based on liquidity, volatility, investors' emotions, and the discovery of stock prices. [7] studies two algorithms Gaussian Process and SMOreg to predict stock prices in the Telecommunication Sector Companies in Indonesia Stock Exchange. [8] explains that direction of movement of stock or index is also essential rather than just checking how close the price is with the predicted one. Research has been conducted to improve upon the State of the art ML models (Support Vector Machine [9], K-Nearest Neighbors [10]) to predict stock prices.

Other works also try to analyse sentiments of the trader to predict the pricing of a stock or the index [11]–[13]. [11] have used Random Forest and Deep Learning algorithms. [12] use artificial neural network (ANN) for predicting the next day Bombay Stock Exchange index value. [13] uses Bayesian model algorithm first and analyse the correlation between between investor sentiment and the GEM stock price in China.

Conventional machine learning approaches advocate utilizing the preceding n values to predict the $n+1$th value, necessitating a consistent availability of historical data. We propose a novel approach of using ML for Stock Trading, diverging from conventional methods by utilizing the opening price to forecast $CLOSE$, $HIGH$, and LOW prices rather than relying on historical data. Our comparative analysis includes six ML models: K-Nearest Neighbors (KNN), Random Forest Regressor (RF), Decision Tree Regressor, Support Vector Regression (SVR) without initialization, SVR with initialization, and Linear Regression (LR). Notably, LR outperforms the other models.

It is imperative to underscore the potential application of Intraday Strategies, such as shorting at predicted $HIGH$ prices and longing at predicted LOW prices. A key aspect of our exploration involves leveraging the $OPEN$ price on the day of prediction to ascertain the anticipated $HIGH$ or LOW prices. This facet serves as a compelling point of investigation. In our analysis, we present a comparative study involving three models: one exclusively utilizing $OPEN$ prices, traditional models relying solely on historical data, and our innovative combined model incorporating both previous data and the $OPEN$ price on the day of prediction. This comparative framework elucidates the efficacy of our approach in enhancing predictive capabilities for devising strategic trading decisions.

III. Dataset

We utilize NIFTY 50 index for our analysis. The dataset for our analysis is sourced from the Yahoo Finance Website [14] and spans from October 01, 2007, to October 01, 2023, with a daily frequency. The data is structured in CSV format, comprising columns for $DATE$, $OPEN$, $HIGH$, LOW , and $CLOSE$ prices. This comprehensive dataset allows us to delve into historical stock market trends and patterns. In preparation for model evaluation, a careful split of the dataset was performed. For testing purposes, 20% of the data covering the period from July 29, 2020, to September 29, 2023, was reserved. Conversely, the initial 80% of the dataset, spanning from October 03, 2007, to July 29, 2020, was allocated for training our models. This approach ensures a robust evaluation process and facilitates the development of accurate predictive models based on a substantial historical training dataset. To address missing data entries, we employ the mean-refilling technique.

IV. Models

This paper introduces an innovative prediction approach, departing from the conventional use of historical data for future price prediction. Instead, our method predicts $Param$ prices solely based on the $OPEN$ prices of the same day. The $Param$ values considered include $HIGH$, LOW, and $CLOSE$ prices for that day. The model is trained using $Param$ and $OPEN$ prices from the training dataset, with predictions made on $Param$ prices using only $OPEN$ values from the test dataset. The integrated system model is depicted in Figure 1a. We employ the $OPEN$ value for testing to predict $CLOSE$, $HIGH$, and LOW values.

Furthermore, we introduce a methodology that integrates both historical data and a predictive approach based on the

$OPEN$ prices of the same day. The schematic representation of this integrated system model is depicted in Figure 1b. In this scenario, we utilize the $OPEN$ value of the day of prediction, along with the ($OPEN$, $CLOSE$, $HIGH$, LOW) values from the preceding day, pertinent to the intraday trade execution.

A. K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm commences by computing the Euclidean distance between matrices, as demonstrated in Equation (1). Following this, it identifies K instances with the smallest Euclidean distances. Prediction for the target variable involves averaging the values of its K-nearest neighbors. The selection of K holds pivotal importance in KNN. Smaller K values introduce heightened variability in predictions, whereas larger K values may result in underfitting. Therefore, determining the optimal K value is crucial. In our context, we observe the KNN graph leveling off after a certain point, even with significant alterations in the K value. The observed flattening of the KNN graph after a certain point is attributed to its sensitivity to outliers, noise, and irrelevant features. Outliers exert a disproportionate influence on the neighbors, affecting the predictions as depicted in the graph. The specified number of neighbors for the illustrated graph is 5.

$$d(X,Y) = \sqrt{\sum_{i,j}(X_{ij} - Y_{ij})^2} \tag{1}$$

B. Random Forest Regressor (RF)

Random Forest (RF) functions by creating an array of decision trees during training, each based on diverse parameters. The final prediction is determined through the majority or average votes across all trees, considering their individual predictions. In Equation (2), where Y' signifies the ultimate prediction and X represents the input data, absolute feature importance is computed by averaging predictions from all t_i trees to identify the most influential predictors. Assuming $t(X)$ and $t_i(X)$ represent decisions from the final and i-th tree, respectively, the calculation is succinctly expressed. In our specific context, despite substantial adjustments in the $n_{\text{estimators}}$ value, the RF model demonstrates a leveling-off phenomenon after a certain point. This observed flattening is attributed to RF's inherent sensitivity to outliers, noise, and irrelevant features.

$$Y' = t(X) = \frac{1}{n_{\text{estimators}}} \sum_{i=1}^{n_{\text{estimators}}} t_i(X) \tag{2}$$

C. Decision Tree Regressor

The Decision Tree Model operates independently, recursively partitioning the dataset based on selected features. Each internal node represents a decision, and every leaf node signifies the final prediction, ensuring transparency in decision-making. Despite the model's inherent simplicity and interpretability, similar to RF and KNN, it exhibits suboptimal performance.

D. Support Vector Regression (SVR)

The objective of Support Vector Regression (SVR) is to find a function $f(x)$ that approximates the mapping from input variables x to output y. SVR maps input features to a higher-dimensional space without explicitly calculating transformed feature vectors, providing flexibility in handling noise and outliers. The optimization process involves a trade-off between achieving a small error and a wide margin, controlled by hyperparameters such as the regularization parameter C and tolerance parameter ε. The kernel matrix H represents the similarity between each pair of input data points. Two different analyses within this model were conducted - one without the initialization of any values and another with the initialization $y = x$. In the initial scenario, a distinct graph closure is not evident; however, in the subsequent case, a more proximate graph closure is observed. This occurrence is attributed to the $CLOSE$ proximity of the $HIGH$, $CLOSE$, and LOW values of the index to the $OPEN$ Value of the Index in the latter case. Initializing $y = x$ proves more effective compared to other models (with parameters: $C = 1.0$, $kernel\_type =$ "$linear$", $\varepsilon = 0.1$). To manage computation time, the maximum iteration count is set to 500,000.

E. Linear Regression (LR)

Linear Regression (LR) is a fundamental regression model that captures the linear relationship between a dependent variable output and num independent variables $\text{input}_1, \text{input}_2, \ldots, \text{input}_{num}$. LR is mathematically expressed by Equation (3). Here, output represents the output for a sample, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_{num}$ are regression coefficients, $\text{input}_1, \text{input}_2, \ldots, \text{input}_{num}$ are the respective variables, and ϵ denotes the residual unexplained error.

For our specific case, only one feature, $OPEN$, is considered, rendering $\text{input}_2, \text{input}_3, \ldots, \text{input}_{num}$ and $\beta_2, \beta_3, \ldots, \beta_{num}$ irrelevant. Thus, our equation simplifies to (4). The model demonstrates superior performance, outperforming other experimented models, and its predictions are utilized to harness the benefits of ML for Price Prediction. Additionally, we apply this model to both the traditional approach and our integrated approach, as demonstrated by the following equations 5 and 6.

$$\text{output} = \beta_0 + \beta_1 \text{input}_1 + \beta_2 \text{input}_2 + \ldots + \beta_{num} \text{input}_{num} + \epsilon \tag{3}$$

$$\text{output} = \beta_0 + \beta_1 OPEN + \epsilon \tag{4}$$

$$\begin{aligned} \text{output} = &\beta_0 + \beta_1 OPEN(\text{prev}) + \beta_2 HIGH(\text{prev}) \\ &+ \beta_3 LOW(\text{prev}) + \beta_4 CLOSE(\text{prev}) + \epsilon \end{aligned} \tag{5}$$

$$\begin{aligned} \text{output} = &\beta_0 + \beta_1 OPEN(\text{prev}) \\ &+ \beta_2 HIGH(\text{prev}) + \beta_3 LOW(\text{prev}) \\ &+ \beta_4 CLOSE(\text{prev}) + \beta_5 OPEN(\text{same}) + \epsilon \end{aligned} \tag{6}$$

V. Performance Metrics

- **Mean Squared Error (MSE):** MSE evaluates the proximity of the regression line to a set of points, aiming for minimal values to achieve precise predictions.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

- **Mean Absolute Error (MAE):** MAE quantifies the absolute deviation between attributes, irrespective of their direction.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

- **R-squared (R^2) Score:** The R-squared score indicates the ratio of explained variance to the total variance inherent in the data. It is formulated as:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{9}$$

In Equations (7) - (9), the notations represent i as the i^{th} input sample, y_i as the actual label value, $\hat{y}_i$ as the predicted output value, $\bar{y}$ denoting the mean of the observed true values, and n as the total number of samples. Favorable model performance is characterized by lower MSE and MAE, and higher R2 scores.

VI. Results and Analysis

The outcomes of our study are meticulously presented in Tables I, II, and III, offering a comprehensive overview of the metrics associated with $CLOSE$, $HIGH$, and LOW values, respectively. Accompanying these tabular representations, we provide illustrative plots in Figure 2a, 2b, and 2c. Upon careful examination, it is evident that Linear Regression (LR) emerges as the most proficient model in our analysis. LR demonstrates a remarkable R2 score of 0.997, 0.999, and 0.998 across the respective categories, surpassing other models significantly, as they register scores below 0.9. Even when compared to the decent performance of Support Vector Regression (SVR) with initialization, LR outshines it.

Model	MAE	MSE	R2 Score
Linear Regression	93.95	14875.98	0.997
Support Vector Machine	9292.10	92177844.25	-17.78
SVR (y=x)	373.81	160036.77	0.967
K-Nearest Neighbors	4205.26	21829563.18	-3.45
Decision Tree	4299.83	22704721.43	-3.63
Random Forest	4259.20	22330743.84	-3.55

TABLE I: Evaluation Metrics for Predictions on $CLOSE$

Model	MAE	MSE	R2 Score
Linear Regression	52.02	6101.04	0.999
Support Vector Machine	9360.82	93462468.97	-18.04
SVR (y=x)	3002.33	9520590.40	-0.94
K-Nearest Neighbors	4226.02	22050699.51	-3.49
Decision Tree	4187.38	21684140.01	-3.42
Random Forest	4203.75	21841254.36	-3.45

TABLE II: Evaluation Metrics for Predictions on $HIGH$

Model	MAE	MSE	R2 Score
Linear Regression	64.52	9000.15	0.998
Support Vector Machine	9252.46	91443610.45	-17.62
SVR (y=x)	537.29	319745.80	0.935
K-Nearest Neighbors	4150.20	21366762.23	-3.35
Decision Tree	4215.91	21963591.79	-3.47
Random Forest	4190.59	21734165.16	-3.43

TABLE III: Evaluation Metrics for Predictions on LOW

TABLE IV: Performance Metrics for Different Approaches

	Adj Close	High	Low
$OPEN$ Data Approach			
MAE	93.95	52.02	64.52
MSE	14875.98	6101.04	9000.15
R2 Score	0.99697	0.99876	0.99817
Traditional Approach			
MAE	7.39	50.52	58.37
MSE	109.97	4719.30	5724.48
R2 Score	0.99998	0.99904	0.99883
Integrated Approach			
MAE	88.42	39.45	51.14
MSE	13083.49	2966.18	4783.71
R2 Score	0.99733	0.99940	0.99903

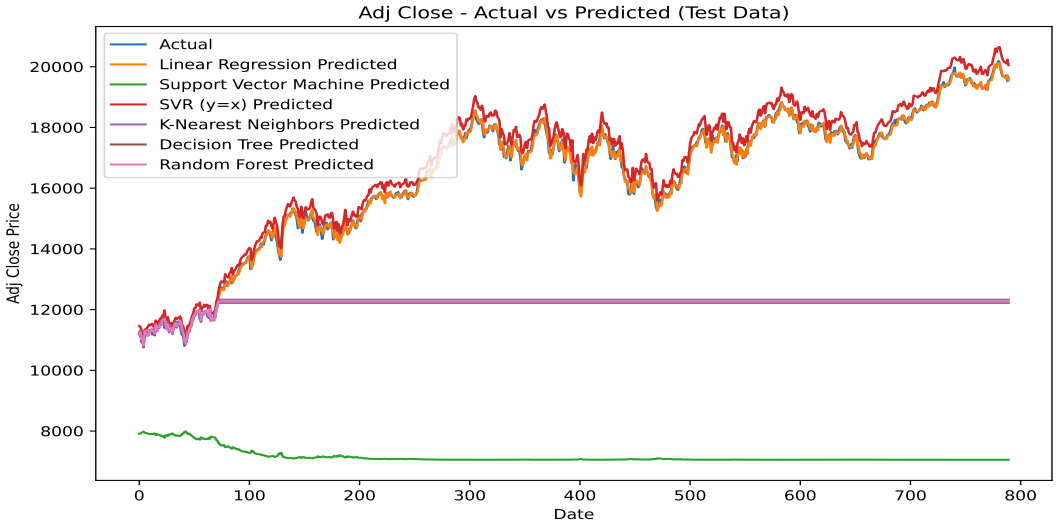
The inferior performance of the alternative models can be attributed to the presence of outliers in the dataset. Notably, LR exhibits outstanding Mean Absolute Error (MAE) values, with scores of 93.95, 53.02, and 64.52, while all other models exhibit MAE values in higher orders. SVR with initialization, though presenting decent results, records MAE values of 373.81, 3002.33, and 537.29. The observed trends in Mean Squared Error (MSE) align with the aforementioned findings, further underscoring the superior predictive capabilities of LR. Our comprehensive evaluation establishes LR as the optimal choice in this predictive modeling scenario.

Our investigation extends to a comparative analysis between our proposed methodology and the conventional practice of relying on historical data. This comparative assessment is succinctly presented in Table IV and Figure 3. Notably, we discern a notable impact attributed to the inclusion of the $OPEN$ value alongside the utilization of preceding data.

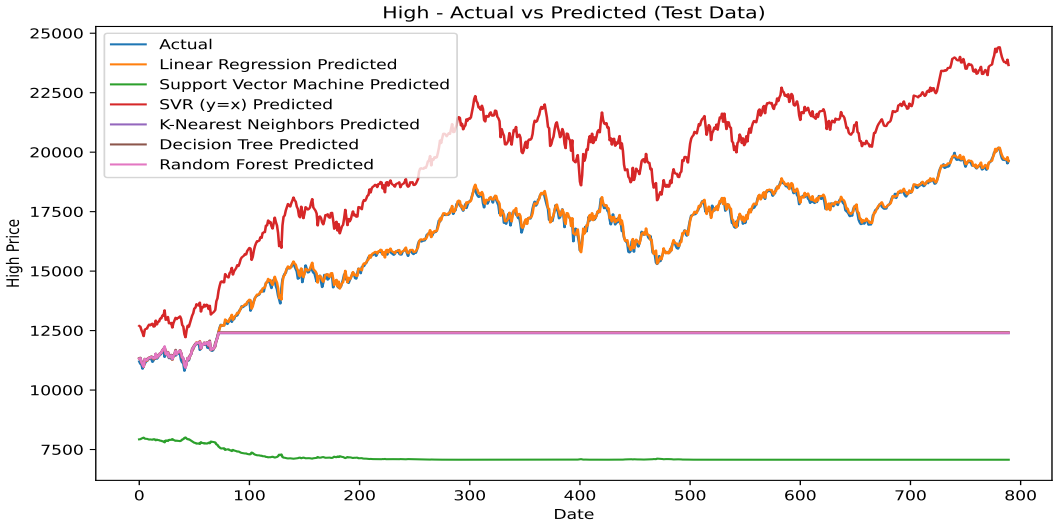
Upon closer inspection, it becomes apparent that the Approach with Combined data yields the highest R2 scores and the lowest errors in the $HIGH$ and LOW cases. Conversely, the second approach, which excludes the consideration of $OPEN$ prices on the day of prediction, exhibits the least error and the highest R2 score for $CLOSE$ data. This trend is visually depicted by the accompanying Figure 3. In light of these findings, we draw the conclusion that incorporating $OPEN$ data in the formulation of intraday strategies proves to be a compelling and insightful avenue for enhancing predictive accuracy.

VII. Conclusion

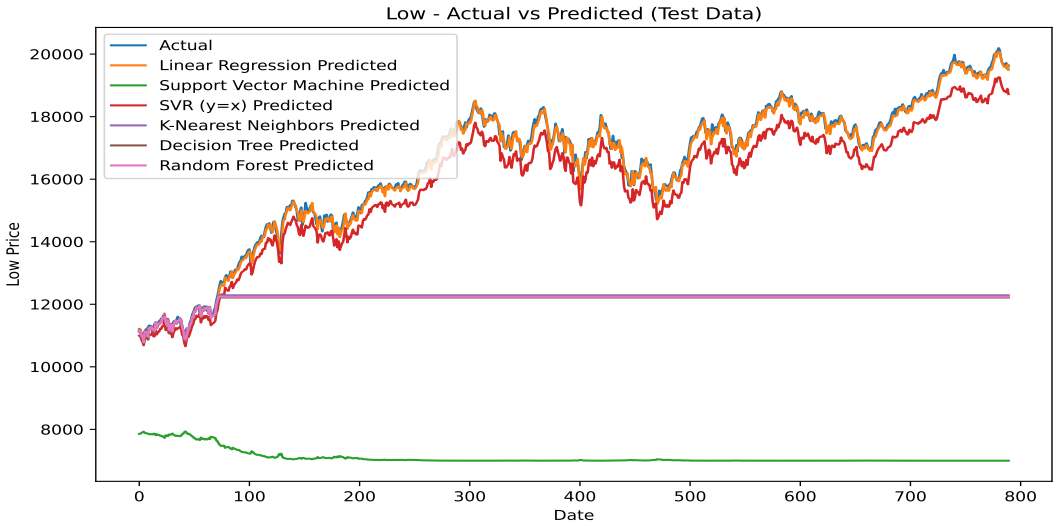
In conclusion, our study delves into the dynamic realm of stock market predictions, specifically focusing on India's National Stock Exchange (NSE) Nifty 50 index and emphasizing the significance of intraday trading strategies. The integration



(a) Comparative Performance of Multiple ML Models in Predicting $CLOSE$ Prices



(b) Comparative Performance of Multiple ML Models in Predicting $HIGH$ Prices



(c) Comparative Performance of Multiple ML Models in Predicting LOW Prices

Fig. 2: Combined Figure for ML Model Comparison

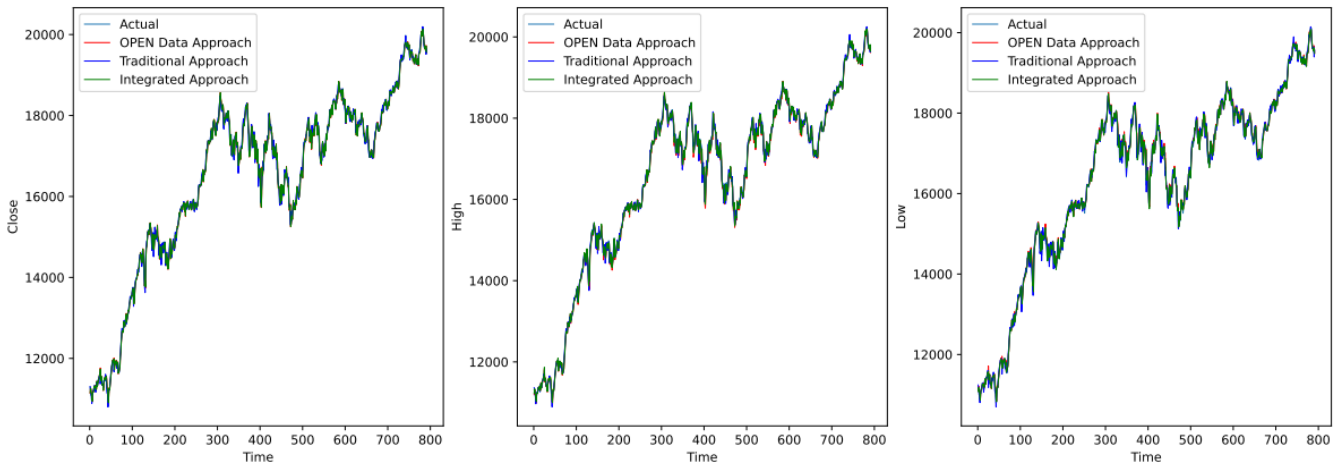


Fig. 3: Plot illustrating the comparison between actual data and predictions across various models.

of machine learning (ML) techniques has proven instrumental in forecasting stock prices, offering valuable insights for investors and traders.

Our comprehensive analysis explored various ML models, including K-Nearest Neighbors (KNN), Random Forest Regressor (RF), Decision Tree Regressor, Support Vector Regression (SVR), and Linear Regression (LR). Among these, Linear Regression emerged as the most proficient model, showcasing exceptional predictive capabilities with a remarkable R2 score of 0.997, 0.999, and 0.998 across $CLOSE$, $HIGH$, and LOW values.

Furthermore, our innovative approach introduced the utilization of intraday data, specifically focusing on $OPEN$ prices on the day of prediction. This departure from conventional reliance solely on historical data provided enhanced predictive accuracy. Upon closer inspection, it becomes apparent that the Approach with Combined data yields the highest R2 scores (0.99940 and 0.99903) and the lowest errors in the $HIGH$ and LOW cases. Conversely, the second approach, which excludes the consideration of $OPEN$ prices on the day of prediction, exhibits the least error and the highest R2 score (0.99998) for $CLOSE$ data. In light of these findings, we draw the conclusion that incorporating OPEN data in the formulation of intraday strategies proves to be a compelling and insightful avenue for enhancing predictive accuracy.

Our findings underscore the potential for leveraging ML techniques in the context of India's NSE Nifty 50, not only for accurate price predictions but also for optimizing intraday trading decisions. As financial markets continue to evolve, the intersection of machine learning and intraday strategies presents an exciting frontier for achieving more informed and data-driven investment practices in the economic landscape.

REFERENCES

[1] J. Antony and J. Joseph, "Financial literacy: Role and impact on financial inclusion," 09 2021.

[2] K. Dhiwar, M. Bedarkar, R. Goyal, R. Suri, R. Birajdar, and R. Modi, "Transitioning towards a circular economy: Role and challenges of corporates in india," in *2023 International Conference on Sustainable Islamic Business and Finance (SIBF)*, pp. 317–324, 2023.

[3] A. de Souza Barbosa, M. C. B. C. da Silva, L. B. da Silva, S. N. Morioka, and V. F. de Souza, "Integration of environmental, social, and governance (esg) criteria: their impacts on corporate sustainability performance," *Humanities and Social Sciences Communications*, vol. 10, no. 1, p. 410, 2023.

[4] BSE Limited, "Bse - bombay stock exchange." https://www.bseindia.com/, 2023. Accessed: 20 December 2023.

[5] NSE Limited, "Nse - national stock exchange." https://www.nseindia.com/, 2023. Accessed: 20 December 2023.

[6] V. Aggarwal, S. Batra, M. Yadav, and P. Kumar, "Literature review on algorithmic trading in financial markets," in *2023 International Conference on Sustainable Islamic Business and Finance (SIBF)*, pp. 76–80, 2023.

[7] J. H. Moedjahedy, R. Rotikan, W. F. Roshandi, and J. Y. Mambu, "Stock price forecasting on telecommunication sector companies in indonesia stock exchange using machine learning algorithms," in *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, pp. 1–4, 2020.

[8] Y. Wei and V. Chaudhary, "The directionality function defect of performance evaluation method in regression neural network for stock price prediction," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 769–770, 2020.

[9] Z. Liu, Z. Dang, and J. Yu, "Stock price prediction model based on rbf-svm algorithm," in *2020 International Conference on Computer Engineering and Intelligent Control (ICCEIC)*, pp. 124–127, 2020.

[10] Q. Yunneng, "A new stock price prediction model based on improved knn," in *2020 7th International Conference on Information Science and Control Engineering (ICISCE)*, pp. 77–80, 2020.

[11] E. P. Torres, E. A. Torres, M. Hernández-Álvarez, and S. G. Yoo, "Emotion recognition related to stock trading using machine learning algorithms with feature selection," *IEEE Access*, vol. 8, pp. 199719–199732, 2020.

[12] S. K. Khatri, H. Singhal, and P. Johri, "Sentiment analysis to predict bombay stock exchange using artificial neural network," in *Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization*, pp. 1–5, 2014.

[13] Y.-X. Zhang, X.-H. Liu, W.-J. Wang, and Y.-J. Liu, "A study of relationship between investor sentiment and stock price : Realization of investor sentiment classification based on bayesian model," in *2020 International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)*, pp. 34–37, 2020.

[14] "Yahoo finance." https://finance.yahoo.com/, 2023.