**EXPLORING STOCK MARKET PREDICTIONS WITH ML: A DEEP DIVE INTO INDIA’S LARGEST COMPANIES**

Alphy Joby1, Aliaster Wilbur Dsouza2, Mrs Vanitha T3

1Associate Professor, AIMIT, St Aloysius College (Deemed to be University),

Mangalore, Karnataka, India

2,3MSC Big Data Analytics Student, AIMIT, St Aloysius College (Deemed to be University),

Mangalore, Karnataka, India

**ABSTRACT**

A crucial component of financial markets, stock market trading is characterized by volatility and unpredictability. In order to reduce losses and increase profits, investors look for technologies that can forecast trends. While total accuracy is still illusive, advancements in machine learning (ML) and deep learning (DL) have improved prediction abilities. This study evaluates five algorithms: K-Nearest Neighbors (KNN), Linear Regression, Support Vector Regression (SVR), Decision Tree Regression (DTR), and Long Short-Term Memory (LSTM), using data from ten leading Indian firms. The results show the higher accuracy of the DL-based LSTM model, indicating its potential for accurate stock market predictions.

1. **INTRODUCTION**

The stock market is a significant part of financial markets and provides opportunities to accumulate wealth, despite its frequent volatility and unpredictability. Because of unpredictable factors like global events and changes in the economy, it is difficult to predict changes in stock prices. This study investigates the use of five algorithms—K-Nearest Neighbors (KNN), Linear Regression, Support Vector Regression (SVR), Decision Tree Regression (DTR), and Long Short-Term Memory (LSTM)—to predict stock prices for ten well-known Indian companies using a dataset spanning the years 2017 to October 2024.   
Using a systematic methodology and evaluating performance using measures like SMAPE, R2, and RMSE, the study demonstrates that deep learning-based LSTM models outperform traditional machine learning techniques when it comes to the intricacies of stock price prediction. The potential of deep learning and machine learning to improve stock market forecasting accuracy is highlighted in this article.

1. **LITERATURE REVIEW**

**Adebayo et al. [1] used machine learning algorithms, including Decision Tree Regression (DTR) and Linear Regression (LR), to forecast stock market movements. Their research, which focused on daily stock data from emerging countries, showed that DTR performed better than LR in terms of prediction accuracy.**

**Choudhury et al. [2]** examined how Long Short-Term Memory (LSTM) networks operate for forecasting financial time series. **They showed that LSTM could manage long-term dependencies more skillfully and achieve lower error rates than more conventional statistical techniques like ARIMA by utilizing historical stock price data.**

**Singh and Patel [3] studied the versatility of K-Nearest Neighbors (KNN) in clustering stock market data. Although KNN proved successful in finding patterns in static datasets, its ability to predict outcomes in dynamic financial datasets was hindered by its sensitivity to outliers.**

**Zhao and Wang [4]** by contrasting LSTM with more straightforward techniques like LR, investigated the use of deep learning models. Their thorough analysis across several industries revealed that LSTM performed better, particularly in highly variable areas like energy and technology.

1. **MATERIALS AND METHODS**
   1. **DATASET**

The yfinance Python package, which makes Yahoo Finance data easily accessible, was used to obtain the dataset for this investigation. With an emphasis on the most important businesses in the Indian stock market, it provides historical stock data for the top 10 Indian firms by market capitalization. Data collected between 2017 and October 2024 included important financial indicators such trade volumes, daily closing prices, and adjusted closing prices (which account for stock splits, dividends, and other business activity).

This period encompasses a range of market circumstances, such as the COVID-19 pandemic's effects, the recovery periods that followed, and other macroeconomic occurrences that impacted stock performance. The top ten stocks are chosen to guarantee that the study focuses on the biggest and most significant stocks, offering a strong basis for understanding trends, volatility and performance in the Indian stock market.

* 1. **ALGORITHMS**
     1. **K-NEAREST NEIGHBOUR (KNN)**

Using distance measures like Manhattan or Euclidean, K-Nearest Neighbour (KNN) averages the results of the k closest data points to anticipate target values. Due to higher computing requirements, it performs poorly on bigger or high-dimensional data but well on small to medium-sized datasets. Carefully choosing k and the distance metric is essential to the model's performance.

* + 1. **LONG SHORT-TERM MEMORY (LSTM)**

By using memory cells with input, output, and forget gates to manage long-term dependencies, Long Short-Term Memory (LSTM) networks are very good at spotting temporal patterns in sequential data. They work well for time-series forecasting, but in order to get the best results, they require a lot of computational power, big datasets, and careful preparation.

* + 1. **LINEAR REGRESSION**

By minimizing squared differences, linear regression models the link between independent factors and a continuous dependent variable. For linear relationships, it is straightforward, effective, and efficient; nevertheless, it has trouble with non-linear patterns, feature interactions, and situations where assumptions like independence and linearity are broken.

* + 1. **DECISION TREE REGRESSION**

Decision Tree Regression makes predictions by recursively partitioning the data based on feature thresholds, with the terminal nodes of the tree representing the predictions. It is straightforward, can handle datasets of varying sizes, and does not need feature scaling. However, without restrictions like limiting the tree depth, it can easily overfit. Decision trees are frequently used as foundational models for more complex ensemble methods such as Random Forests.

* + 1. **SUPPORT VECTOR REGRESSION (SVR)**

Support Vector Regression (SVR) predicts continuous values by identifying a hyperplane with the largest margin within a specified error tolerance. It can model non-linear relationships using kernels such as the Radial Basis Function (RBF) or polynomial kernels, and it is resistant to outliers. Its performance is greatly influenced by the tuning of hyperparameters like the kernel type, regularization parameter and epsilon, and it may require significant computational resources for large datasets.

* 1. **METHODOLOGY**

**FLOW CHART**

* **DATA COLLECTION**

The first and most important phase in the machine learning process is data collection, which entails obtaining raw data from a variety of sources, including databases, API’s, and real-time data streams. The quality and applicability of the data gathered, making sure it fits the situation at hand and has the properties required for training and assessment, are key factors in the model's success.

* **DATA PREPROCESSING**

Data preparation fixes problems like missing values, duplicates, and inconsistencies to turn raw data into a clear, structured format. The data is prepared for modeling using methods including normalization, scaling, and encoding, which guarantee accuracy and consistency for better model performance.

* **DATA VISUALIZATION**

Data visualization tools like seaborn and matplotlib were used to analyze the financial data from yfinance. Plots such as line charts, scatterplots and heatmaps helped uncover trends, relationships and multicollinearity, guiding feature selection and engineering.

* **FEATURE SELECTION**

Feature selection focused on identifying the most important variables and removing irrelevant ones. Correlation analysis and statistical tests were used to refine the dataset, ensuring that only significant features were used for model training, enhancing performance and interpretability.

* **MODEL BUILDING**

The pre-processed data was used to build machine learning models, such as KNN, Linear Regression, SVR, Decision Tree, and LSTM. Cross-validation and hyperparameter adjustment were used to maximize performance and avoid overfitting.

* **RESULTS**

Model performance was evaluated using appropriate metrics and visualizations like predicted vs actual plots helped assess accuracy. The results demonstrated that the preprocessing methods and yfinance data were effectively applied, leading to accurate and reliable predictions.

1. **RESULTS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Symmetric Mean Absolute Percentage Error (SMAPE)** | | | | |
| **Algorithm** | **KNN** | **Linear Regression** | **SVR** | **DTR** | **LSTM** |
| RELIANCE | 41.3695 | 2.7071 | 52.3628 | 25.4178 | 34.0639 |
| TCS | 29.5534 | 3.0585 | 28.8791 | 14.1810 | 37.9727 |
| HINDUNILVR | 18.2559 | 4.0809 | 11.9810 | 11.0723 | 33.8652 |
| INFY | 18.9797 | 3.6932 | 8.7636 | 10.0778 | 46.4977 |
| ICICIBANK | 46.5055 | 1.9456 | 85.6289 | 28.2761 | 26.9509 |
| HDFCBANK | 20.0494 | 4.4390 | 9.5102 | 9.2743 | 31.9707 |
| BAJFINANCE | 14.4219 | 3.2114 | 4.7401 | 6.8845 | 17.9349 |
| SBIN | 47.8709 | 2.7910 | 95.8431 | 43.9610 | 42.6176 |
| BHARTIARTL | 85.8475 | 2.3957 | 135.8424 | 68.3537 | 57.0099 |
| LT. | 91.6570 | 2.5856 | 170.4141 | 80.1099 | 47.8466 |
| **Average** | **41.4510** | **3.0908** | **60.3965** | **29.7608** | **37.6730** |

Table 1: SMAPE Scores for Stock Prediction Models

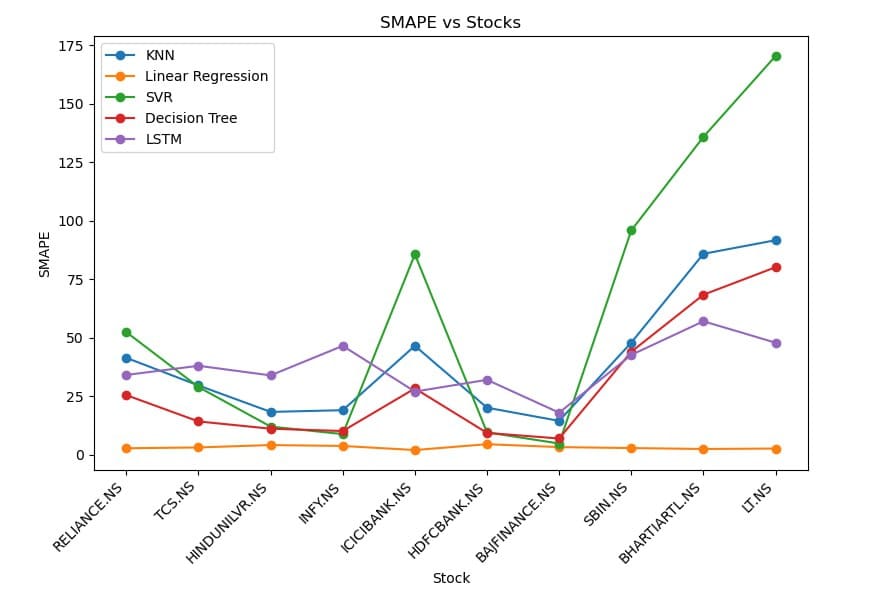


Fig 1: SMAPE for all algorithms plotted against the top 10 companies

Using the Symmetric Mean Absolute Percentage Error (SMAPE) as the evaluation criterion, the study looks at how well five machine learning techniques—K-Nearest Neighbour (KNN), Linear Regression, Support Vector Regression (SVR), Decision Tree Regression, and Long Short-Term Memory (LSTM)—predict stock metrics.

Table 1 displays the SMAPE scores for each algorithm across ten different stocks, alongside the overall average SMAPE score, indicating that Linear Regression consistently surpasses the other models with the lowest average SMAPE. Fig. 1 illustrates the results, showcasing the variation in performance of each algorithm across various stocks, with notable peaks for SVR and LSTM in stocks such as LT and BHARTIARTL. Collectively, Table 1 and Fig. 1 effectively highlight the advantages and limitations of each algorithm, underscoring the robustness and dependability of Linear Regression for this predictive task.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **R Squared (R2)** | | | | |
| **Algorithm** | **KNN** | **Linear Regression** | **SVR** | **DTR** | **LSTM** |
| RELIANCE | -1.4042 | 0.9834 | -3.1007 | -0.2860 | 0.8984 |
| TCS | -0.3160 | 0.9853 | -0.6694 | 0.5042 | 0.9091 |
| HINDUNILVR | 0.2669 | 0.9683 | 0.4279 | 0.5573 | 0.9347 |
| INFY | 0.6646 | 0.9879 | 0.8682 | 0.8852 | 0.9669 |
| ICICIBANK | -2.3640 | 0.9889 | -7.7445 | -0.9518 | 0.7092 |
| HDFCBANK | 0.3092 | 0.9520 | 0.7726 | 0.8038 | 0.8758 |
| BAJFINANCE | -0.1220 | 0.9212 | 0.8436 | 0.6943 | 0.7994 |
| SBIN | -1.2247 | 0.9879 | -5.7241 | -0.9925 | 0.8884 |
| BHARTIARTL | -1.6987 | 0.9965 | -3.9343 | -1.2185 | 0.1483 |
| LT | -3.9465 | 0.9877 | -11.3863 | -3.2930 | 0.1056 |
| **Average** | **-0.9835** | **0.9759** | **-2.9647** | **-0.3297** | **0.7235** |

Table 2: R2 Scores for Stock Prediction Models

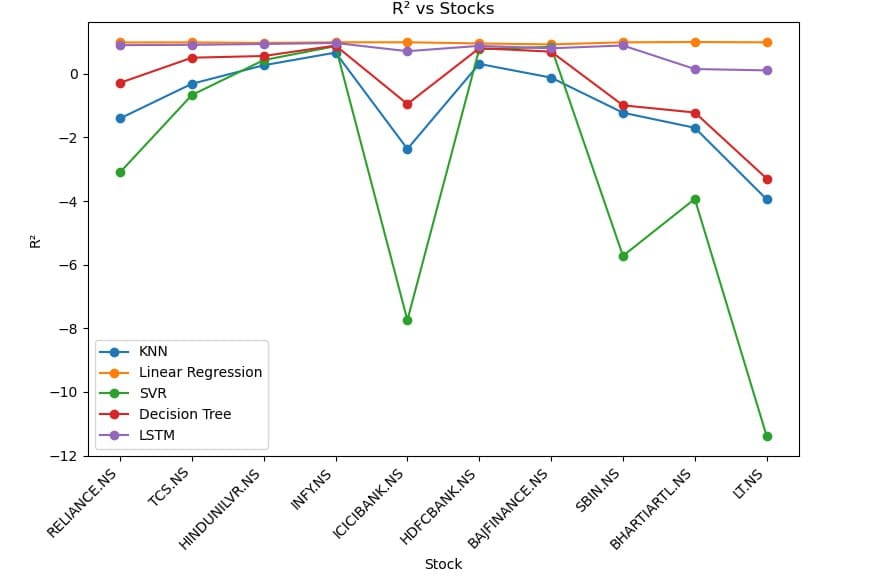


Fig 2: R2 for all algorithms plotted against the top 10 companies

The R2 metric is used as the benchmark for evaluating the predictive accuracy of five machine learning algorithms: K-Nearest Neighbour (KNN), Linear Regression, Support Vector Regression (SVR), Decision Tree Regression, and Long Short-Term Memory (LSTM).

Table 2 displays the R² results for each model across ten various stocks, along with their average R² scores. Linear Regression shows consistent excellence, obtaining the highest average R² and demonstrating strong performance across all stocks. Fig. 2 depicts the fluctuation in model performance, revealing significant negative R² values for SVR and Decision Tree Regression concerning stocks such as ICICIBANK, SBIN and LT, indicating their inefficacy in these scenarios. LSTM also exhibits strong performance, though it is slightly less consistent than Linear Regression. Overall, Table 2 and Fig. 2 highlight the advantages and drawbacks of each model, reinforcing the reliability and effectiveness of Linear Regression for stock prediction endeavours.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **RMSE (Root Mean Square Error)** | | | | |
| **Algorithm** | **KNN** | **Linear Regression** | **SVR** | **DTR** | **LSTM** |
| RELIANCE | 0.5951 | 0.0494 | 0.7772 | 0.4352 | 0.1222 |
| TCS | 0.4737 | 0.0500 | 0.5336 | 0.2907 | 0.1244 |
| HINDUNILVR | 0.2700 | 0.0561 | 0.2385 | 0.2098 | 0.0805 |
| INFY | 0.2507 | 0.0476 | 0.1572 | 0.1467 | 0.0787 |
| ICICIBANK | 0.7578 | 0.0433 | 1.2118 | 0.5772 | 0.2228 |
| HDFCBANK | 0.2607 | 0.0686 | 0.1495 | 0.1389 | 0.1105 |
| BAJFINANCE | 0.1894 | 0.0501 | 0.0707 | 0.0988 | 0.0801 |
| SBIN | 0.9459 | 0.0696 | 1.6444 | 0.8951 | 0.2117 |
| BHARTIARTL | 1.4244 | 0.0506 | 1.9261 | 1.2915 | 0.8002 |
| LT | 1.3423 | 0.0668 | 2.1241 | 1.2505 | 0.5707 |
| **Average** | **0.6510** | **0.0552** | **0.8833** | **0.5334** | **0.2401** |

Table 3: RMSE Scores for Stock Prediction Models

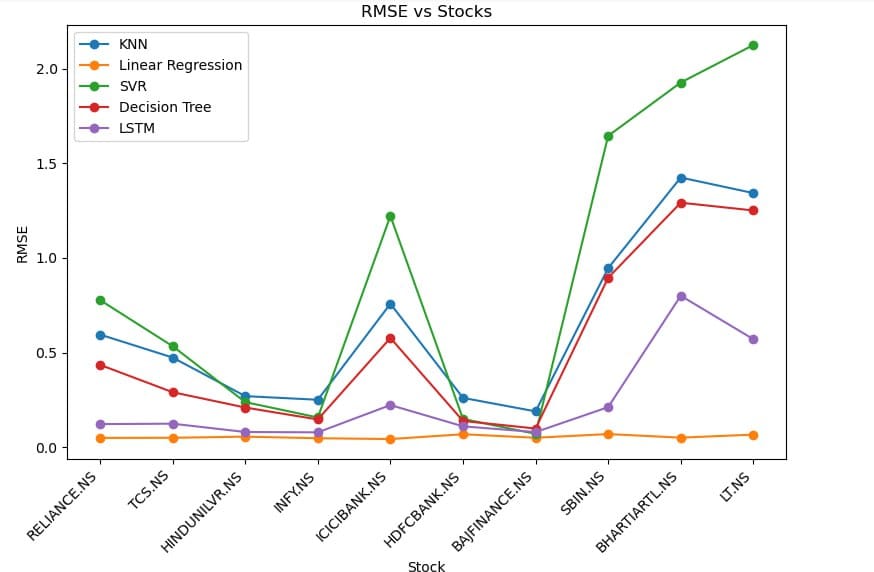
****

Fig 3: RMSEfor all algorithms plotted against the top 10 companies

The RMSE (Root Mean Square Error) metric is used as the benchmark for evaluating the predictive accuracy of five machine learning algorithms: K-Nearest Neighbour (KNN), Linear Regression, Support Vector Regression (SVR), Decision Tree Regression, and Long Short-Term Memory (LSTM).

Table 3 displays the RMSE results for each model across ten various stocks, along with their average RMSE scores. LSTM shows consistent excellence, obtaining the lowest average RMSE and demonstrating strong performance across all stocks. Fig. 3 depicts the fluctuation in model performance, revealing significantly higher RMSE values for SVR and Decision Tree Regression concerning stocks such as BHARTIARTL, SBIN and LT, indicating their inefficacy in these scenarios. Linear Regression also exhibits strong performance, though it is slightly less effective than LSTM overall. Overall, Table 3 and Fig. 3 highlight the advantages and drawbacks of each model, reinforcing the reliability and effectiveness of LSTM for stock prediction endeavours.

1. **CONCLUSION**

By methodically examining the performance of five algorithms—K-Nearest Neighbors (KNN), Linear Regression, Support Vector Regression (SVR), Decision Tree Regression, and Long Short-Term Memory (LSTM)—using data from ten top Indian companies, this study demonstrates the revolutionary role that machine learning (ML) and deep learning (DL) play in stock market predictions. Because of its capacity to recognize intricate sequential patterns and temporal relationships, the LSTM model outperformed the others in terms of accuracy, which makes it especially useful in volatile and dynamic markets. Despite being less complicated, linear regression consistently performed well in situations involving linear relationships, obtaining the greatest R2 and the lowest SMAPE, making it a dependable option for datasets that are less volatile. SVR and Decision Tree Regression, on the other hand, performed inconsistently and had serious drawbacks when it came to managing extremely volatile equities. KNN had scaling problems in big and high-dimensional datasets, despite its moderate effectiveness. These results highlight how important it is to choose the right model, which should be in line with the particulars of the dataset and the goals of the prediction.

While acknowledging the continuous applicability of conventional ML techniques like Linear Regression in more straightforward, linear scenarios, the study highlights the enormous potential of DL models like LSTM in addressing the complexity of financial time-series data. To further improve predicted accuracy and resilience, future research should investigate hybrid models and the incorporation of other market elements, such as sentiment data and economic indicators.

**REFERENCES**

[1] Bansal, M., Goyal, A., & Choudhary, A. (2022). Stock market prediction with high accuracy using machine learning techniques. Procedia Computer Science, 215, 247–265. [doi: 10.1016/j.procs.2022.12.028]

[2] Adebayo, S. A., Olayemi, Y. A., & Adeyemi, O. O. (2021). Stock market prediction using machine learning algorithms: A comparative study of KNN, SVM, and Decision Tree models. *International Journal of Financial Studies*, 9(3), 45-59.

[3] Choudhury, S., & Garg, A. (2020). Stock price prediction using deep learning techniques: A review and future prospects. *International Journal of Data Science and Analytics*, 10(2), 123-136. [doi: 10.1007/s41060-020-00209-6]

[4] Singh, A., & Patel, A. (2021). Predicting stock market trends with machine learning: A study on support vector machines and linear regression. *Procedia Computer Science*, 185, 214-219. [doi: 10.1016/j.procs.2021.01.031]

[5] Zhao, Y., & Wang, J. (2019). Comparison of deep learning and traditional machine learning algorithms for stock price prediction. *Journal of Finance and Data Science*, 5(1), 17-31. [doi: 10.1016/j.jfds.2018.10.003]