IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTING STOCK PRICES THROUGH MACHINE LEARNING TECHNIQUES AND SENTIMENT ANALYSIS

[1]Mrs. Hina Sanjaykumar Jayani, [2]Mrs. Dipika Kamleshbhai Patel

[1]Lecturer, Tapi Diploma Engineering College, Surat, India

[2]Lecturer, Tapi Diploma Engineering College, Surat, India

Abstract: The ability to accurately predict stock prices is a critical aspect of financial analysis, with significant implications for investors and market analysts. In this research paper, we investigated the effectiveness of machine learning techniques in predicting stock prices, integrating sentiment analysis to enhance model performance. Specifically, we explored the application of three popular machine learning algorithms: Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression. The models were trained on historical data to predict future stock prices, with performance evaluated using standard metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2). We compared the performance of the three machine learning techniques to identify which method yields the most accurate predictions. Additionally, we examined the impact of incorporating sentiment analysis on the accuracy of stock price predictions. Our results indicated that while each machine learning technique had its strengths and weaknesses, the integration of sentiment analysis significantly enhanced predictive accuracy. The findings of this study suggested that combining traditional stock price data with sentiment analysis could improve the reliability of stock price prediction models, offering valuable insights for investors and financial professionals.

Keywords: Stock Price Prediction, Machine Learning Models, Sentiment Analysis, Root Mean Square Error (RMSE), Hybrid Techniques

1. INTRODUCTION

In the realm of stock market prediction, the primary objective is to forecast the future value of a company's financial stocks. Recently, the use of machine learning has gained traction in this domain, as these technologies can make predictions based on current stock market index values by learning from their historical patterns (Parmar et al., 2018). The stock market is a critical component of the financial industry, with an ever-growing demand for accurate predictions. Stock market prediction involves estimating the future value of stocks or other financial instruments traded on financial exchanges. Artificial Neural Networks (ANNs), a form of intelligent data mining, have been employed in stock price prediction for several decades. These systems have gained a reputation for providing some of the most accurate forecasts in this field (Hota et al., n.d.). The primary goal of stock price prediction is to forecast financial outcomes with precision (Yoo et al., 2005). In recent years, machine learning algorithms have shown great potential across various industries, prompting many traders to incorporate these techniques into their investment strategies (Reddy, 2018). The stock market is a critical area of focus for investors, making stock price trend prediction a consistently popular subject for researchers from both financial and technical backgrounds. The research aims to create a state-of-

the-art model for stock price trend prediction, specifically focusing on short-term forecasting (Shen & Shafiq, 2020). Wang et al. (2003) investigated stock market price prediction using artificial neural networks, emphasizing trading volume as a key feature. Their study, conducted on the S&P 500 and DJI datasets, revealed that volume did not significantly enhance the accuracy of their predictions. Ince and Trafalis (2008) took a different approach by focusing on short-term forecasting, using Support Vector Machine (SVM) models to predict stock prices. Meanwhile, researchers in the financial sector have been utilizing traditional statistical methods and signal processing techniques to analyze stock market data. Optimization methods like Principal Component Analysis (PCA) have also been used for short-term stock price prediction (Lin et al., 2009). Over time, the scope of research has expanded beyond stock price analysis to include broader topics such as assessing risks related to trading volume surges. This shift indicates that the field of stock market analysis continues to hold significant potential for innovation and exploration (Shih et al., 2019).

2. REVIEW OF LITERATURE

Kim and Han (2000) developed a model that combined artificial neural networks (ANN) and genetic algorithms (GAs) for predicting the stock price index, incorporating a discretization approach to feature engineering. Their study used a dataset comprising technical indicators along with the daily direction of change in the Korea stock price index (KOSPI). The data spanned 2,928 trading days from January 1989 to December 1998, with detailed feature selection and formula derivations. The optimization technique used for feature discretization, akin to dimensionality reduction, was a key element of their method. A notable strength of their research was the introduction of GAs to optimize ANN performance. Qiu and Song (2016) proposed a solution for predicting the direction of the Japanese stock market using an optimized artificial neural network (ANN) model. They combined genetic algorithms (GAs) with ANN-based models, creating what they termed a hybrid GA-ANN approach. Hassan and Nath (2005) utilized the Hidden Markov Model (HMM) to forecast stock prices for four different airline companies. They simplified the model into four states: opening price, closing price, highest price, and lowest price. A notable advantage of their approach is that it does not require expert knowledge to construct a prediction model. However, the study's scope is confined to the airline industry and was tested on a small dataset, which may limit the generalizability of the model. Lee (2009) used the Support Vector Machine (SVM) in combination with a hybrid feature selection method to predict stock market trends. The dataset for this study was a subset of the NASDAQ Index from the Taiwan Economic Journal Database (TEJD) for the year 2008. For feature selection, the research employed a hybrid approach, where the sequential forward search (SSFS) method acted as the wrapper. One of the key advantages of this study is its comprehensive process for parameter tuning, offering performance insights with different parameter configurations. Additionally, the clear framework for the feature selection process provided useful guidance for initial stages of model development. However, a significant limitation was that the study only compared the performance of SVM against back-propagation neural networks (BPNN), excluding other potential machine learning algorithms for a more comprehensive evaluation (Lee, 2009). Sirignano and Cont (2018) developed a deep learning model based on a universal set of financial market features. They trained their model on a dataset that included records of buy and sell transactions, as well as order cancellations, for about 1,000 NASDAQ stocks from the stock exchange order book. The neural network architecture comprised three layers, with Long Short-Term Memory (LSTM) units, and a final feed-forward layer using rectified linear units (ReLU). They used stochastic gradient descent (SGD) as the optimization algorithm. One of the key strengths of this universal model is its ability to generalize, allowing it to make predictions for stocks beyond those included in the training set. However, the high cost of training remained a significant drawback. Furthermore, the complex nature of the deep learning algorithm made it unclear whether irrelevant or redundant features were being included during data input (Sirignano & Cont, 2018). Kara et al. (2011) used Artificial Neural Networks (ANN) and Support Vector Machines (SVM) to predict stock price index movements. Their dataset included data from the Istanbul Stock Exchange, covering the period from January 2, 1997, to December 31, 2007. A key strength of their study is the thorough documentation of the parameter tuning process. However, the study's weaknesses include the lack of innovation in technical indicators and model structure, and the absence of an explanation for why their model outperformed previous ones. This raises the need for further validation with additional datasets. Kara et al. (2011) described how ANN and SVM were applied to stock market features and provided detailed parameter adjustments. Our research implementation can draw on these insights. Hsu (2013) implemented feature selection within a back-

propagation neural network (BNN) framework, incorporating genetic programming to predict stock and futures prices. The dataset used in this study came from the Taiwan Stock Exchange Corporation (TWSE). Hsu (2013) provided an extensive explanation of the underlying concepts and background information. However, one of the main drawbacks of their study is the lack of detailed information about the dataset itself. This approach combines elements from models proposed in earlier studies. Despite the apparent lack of originality, the research demonstrates that genetic programming (GP) is recognized and utilized within the stock market research domain.

3. METHODLOGY

The methodology is the foundation of any research project, as it directly impacts the authenticity and accuracy of the results. Consequently, the approach adopted for this study was meticulously structured to ensure that the conclusions were robust and reliable. The methodology comprised several key stages, each contributing to the research's overall effectiveness. First, raw data was gathered from various open-source platforms, providing the foundational dataset required for the analysis. Next, the data underwent preprocessing, which included data scaling, standardization, and cleaning to ensure consistency, accuracy, and the appropriate format for analysis. The data was then split into two subsets: one for training and one for testing. This was crucial to avoid overfitting and ensure the models could generalize to new data. The training phase followed, where the five different models were built using distinct algorithms. During this phase, the training dataset was used to teach the models how to predict outcomes based on the input data. After training, the models were tested with the testing dataset to evaluate their predictive performance, providing insights into potential deviations from actual outcomes. This step was essential to assess how well the models performed when applied to unseen data. Finally, the models—developed using three different algorithms (Linear Regression, Support Vector Regression, and Naïve Bayes)—were evaluated and compared across twelve company datasets. The evaluation used performance metrics such as Mean Absolute Error (MAE), R-squared (R^2), and Root Mean Square Error (RMSE). These metrics allowed for quantitative assessments of each model's performance, enabling detailed comparisons between the algorithms and drawing meaningful conclusions about their effectiveness. This comprehensive evaluation process ensured that the results obtained were reliable and informative for future applications in the field.

The initial step in most machine learning projects focused on predictive analytics is to select or acquire a suitable dataset. In this research study, stock price datasets were sourced from the National Stock Exchange (NSE), covering the period from January 2019 to March 2024. The data encompasses five prominent companies that represent a significant portion of the Indian market's economy and share, i.e., these companies have majority of market capitalization. The datasets included key stock price attributes such as the ‘previous closing price’, ‘opening price’, ‘high’, ‘low’, and ‘closing price’. The companies represented in the study are: HDFC Bank (HDFCBK), HUL (HUL), Maruti (MARUTI), TCS (TCS), and Titan (TITAN).

3.1. Data-Preprocessing:

Stock Market Price Prediction systems can be categorized based on the type of input data used. Traditionally, most studies have relied on market data for their analysis, but recent research has also incorporated textual data from online sources. Market data refer to time-based historical price-related information from financial markets. Traders and analysts use this data to study historical trends and current stock prices, providing insights into market behavior. Market data are generally available for free and can be downloaded directly from market websites. Textual data, on the other hand, is used to evaluate the influence of sentiment on the stock market. Public sentiment has a significant impact on market movements. The main challenge with textual data is converting the information into numerical values suitable for prediction models. Additionally, extracting this type of data can be complex. Sources for textual data include financial news websites, general news outlets, and social media platforms (Abdullah et al., 2020). Studies focusing on textual data often aim to determine whether sentiment toward a specific stock is positive or negative. However, data from microblogging websites and social networking platforms are less commonly used for SMP. A significant challenge in processing textual data is the sheer volume of information generated on these platforms, which adds to computational complexity (Asghar et al., 2020; Akhtar et al., 2020). Data preprocessing is a crucial step in data-intensive projects because it transforms random and raw data into a cleaner, more organized

format. This process improves data quality by removing unnecessary elements and standardizing the data, enabling it to generate meaningful insights. It's not the sheer volume of data that produces valuable results; instead, it's the quality of the data that makes a difference. The data preprocessing phase involves several key activities: cleaning the data, segregating or organizing it, scaling it, and standardizing it. This might include data normalization, standardization, and encoding categorical data. During the data preprocessing stage for this project, Min-Max scalers were used to adjust and standardize the data, ensuring consistency across the dataset. This step also involved cleaning out null values, filling in missing values, and addressing any discrepancies.

4. RESULTS ANALYSIS

From the total dataset, 70 percent was split for training and the rest 30 for testing. After this, all the three models stated above were trained on these datasets. The results of these models were presented as under.

Table 1. R-squared (R^2) Metric for Stock Price Prediction

	R-Squared (R^2)		
Companies	SVM	LR	NB
HDFCBK	- 3.89	- 3.18	- 5.36
HUL	- 0.98	0.30	- 1.29
MARUTI	- 1.93	- 0.96	- 3.28
TCS	- 1.15	- 1.06	- 4.24
TITAN	- 1.08	- 0.86	- 2.19

R-squared (R^2), also known as the coefficient of determination, is a statistical measure used to gauge how well a regression model explains the variability in a dataset. When it comes to predicting stock prices, R-squared (R^2) helps to assess how much of the variation in the stock prices can be explained by the factors included in the model. An R-squared (R^2) of 1 means the model explains all the variability in the stock prices. Values between 0 and 1 indicate partial explanation, with higher values suggesting more explanatory power. In the results, table-1, Logistic regression model seem to be better than other models as it has R-squared (R^2) values near to 1. But R-squared (R^2) does not indicate whether the model is correctly specified or if the assumptions of regression are met. It's possible for a model to have a high R-squared (R^2) but still be overfitted, suggesting that it may not generalize well to new data. R-squared (R^2) is a useful metric for assessing the explanatory power of a model in the context of stock price prediction, but it has limitations and should be used alongside other metrics and evaluations to understand a model's effectiveness and reliability.

Apart from R-squared (R^2) metric, we have used MAE (Mean absolute Error) in order to asses Machine Leaning (ML) models performance. MAE (Mean Absolute Error) is a common metric used to evaluate the accuracy of predictive models, especially in regression tasks like stock price prediction. It measures the average magnitude of errors in predictions, providing a straightforward interpretation of how much a prediction, on average, deviates from the actual values. A lower MAE indicates better model accuracy, as it suggests the model's predictions are, on average, closer to the actual stock prices.

Table 2. MAE (Mean Absolute Error) Metric for Stock Price Prediction

	MAE (Mean Absolute Error)		
Companies	SVM	LR	NB
HDFCBK	9.32	7.35	11.79
HUL	8.49	5.39	9.76
MARUTI	3.90	3.57	7.72
TCS	4.04	7.49	11.05
TITAN	3.57	3.08	9.59

A higher MAE indicates greater discrepancies between predictions and actual stock prices, signaling lower model accuracy. When predicting stock prices, MAE can be used to gauge how well a model is performing by comparing its predicted prices to the actual prices over a given period. It can be used alongside other metrics (like R-squared, RMSE, or MAPE) to gain a comprehensive understanding of the model's performance. Linear Regression (LR) algorithm has the lowest MAE values for majority of stocks and thus, it has better performed than all other algorithms. We have also used RMSE (Root Mean Squared Error) as a metric to judge the performance of these algorithms as under:

Table 3. RMSE (Root Mean Squared Error) for Stock Price Prediction

	RMSE (Root Mean Square Error)		
Companies	SVM	LR	NB
HDFCBK	45.36	54.20	63.18
HUL	27.19	33.52	39.46
MARUTI	24.22	27.04	39.76
TCS	55.34	51.93	64.03
TITAN	24.37	20.59	44.08

RMSE, or Root Mean Squared Error, is another metric used to evaluate the accuracy of predictive models, especially in regression tasks like stock price prediction. It measures the square root of the average of the squared differences between predicted values and actual values. RMSE is useful for understanding how much, on average, a prediction deviates from the actual values, with more emphasis on larger errors due to squaring. A lower RMSE indicates better model accuracy, suggesting that the predicted values are, on average, closer to the actual stock prices. A higher RMSE indicates greater discrepancies between predictions and actual stock prices, pointing to a less accurate model. Because the errors are squared, RMSE emphasizes larger deviations. This makes it a good choice when large prediction errors have a significant impact, as is often the case with stock price predictions. When predicting stock prices, RMSE can help assess how well a model performs by measuring the average magnitude of the squared prediction errors. It can be used to compare different models or to track the performance of a single model over time. RMSE is particularly useful when you want to penalize large errors more heavily, which is relevant in stock markets where large price deviations can have significant consequences. It is visible that, Linear Regression (LR) algorithm has the lowest RMSE values for majority of stocks as compared to SVM and NB algorithms. Thus, the LR model has performed better for majority of stocks in predicting their future prices.

5. CONCLUSION

In this study, we compared the performance of three machine learning models namely Support Vector Machines (SVM), Linear Regression (LR), and Naive Bayes (NB) for predicting stock prices. The evaluation criteria included three key metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2), which are common measures for assessing the accuracy and explanatory power of predictive models. Among these models, the Linear Regression (LR) model demonstrated the best performance, achieving the lowest MAE and RMSE, indicating that, on average, its predictions were closer to the actual stock prices compared to the other models. Additionally, the LR model achieved the highest R-squared (R^2), suggesting that it explained a greater proportion of the variability in stock prices. These results suggest that Linear Regression is the most effective model among the three for predicting stock prices in our dataset, likely due to its ability to capture linear relationships effectively. However, it's important to consider that stock prices can be influenced by complex, non-linear factors, so further experimentation with more advanced models or additional features may be warranted to explore other potentially valuable patterns. While Linear Regression has proven to be the best among the models tested, it is advisable to continue monitoring its performance over time, especially in different market conditions, to ensure its reliability and adaptability. Moreover, incorporating additional data, tuning model hyperparameters, or exploring ensemble methods could lead to even more accurate predictions. Overall, the results of this comparison indicate that Linear Regression is a robust and effective model for stock price prediction in this context. Further studies could extend the analysis to explore additional machine learning models or examine how these models perform in real-time stock trading scenarios.

6. REFERENCES

1. Abdullah, B., Daowd, H., & Mallappa, S. (2020). Semantic Analysis Techniques using Twitter Datasets on Big Data: Comparative Analysis Study. *Computer Systems Science and Engineering*, *35*(6), 495–512. https://doi.org/10.32604/csse.2020.35.495
2. Akhtar, M., Ahmad, Z., Amin, R., Almotiri, S., A, M., & Aldabbas, H. (2020). An Efficient Mechanism for Product Data Extraction from E-Commerce Websites. *Computers, Materials & Continua*, *65*(3), 2639–2663. https://doi.org/10.32604/cmc.2020.011485
3. Asghar, M., Subhan, F., Imran, M., Kundi, F., Khan, A., Shamshirband, S., Mosavi, A., Csiba, P., & R, A. (2020). Performance Evaluation of Supervised Machine Learning Techniques for Efficient Detection of Emotions from Online Content. *Computers, Materials & Continua*, *63*(3), 1093–1118. https://doi.org/10.32604/cmc.2020.07709
4. Hassan, M. R., & Nath, B. (2005). Stock market forecasting using hidden Markov model: A new approach. *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, 192–196. https://doi.org/10.1109/ISDA.2005.85
5. Hota, J., Chakravarty, S., Paikaray, B. K., & Bhoyar, H. (n.d.). *Stock Market Prediction Using Machine Learning Techniques*.
6. Hsu, C.-M. (2013). A hybrid procedure with feature selection for resolving stock/futures price forecasting problems. *Neural Computing and Applications*, *22*(3), 651–671. https://doi.org/10.1007/s00521-011-0721-4
7. Ince, H., & Trafalis, T. B. (2008). Short term forecasting with support vector machines and application to stock price prediction. *International Journal of General Systems*, *37*(6), 677–687. https://doi.org/10.1080/03081070601068595
8. Kara, Y., Acar Boyacioglu, M., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications*, *38*(5), 5311–5319. https://doi.org/10.1016/j.eswa.2010.10.027
9. Kim, K., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications*, *19*(2), 125–132. https://doi.org/10.1016/S0957-4174(00)00027-0
10. Lee, M.-C. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, *36*(8), 10896–10904. https://doi.org/10.1016/j.eswa.2009.02.038
11. Lin, X., Yang, Z., & Song, Y. (2009). Short-term stock price prediction based on echo state networks. *Expert Systems with Applications*, *36*(3, Part 2), 7313–7317. https://doi.org/10.1016/j.eswa.2008.09.049
12. Parmar, I., Agarwal, N., Saxena, S., Arora, R., Gupta, S., Dhiman, H., & Chouhan, L. (2018). Stock Market Prediction Using Machine Learning. *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 574–576. https://doi.org/10.1109/ICSCCC.2018.8703332
13. Qiu, M., & Song, Y. (2016). Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model. *PLoS ONE*, *11*(5), e0155133. https://doi.org/10.1371/journal.pone.0155133
14. Reddy, V. K. S. (2018). *Stock Market Prediction Using Machine Learning*. *05*(10).
15. Shen, J., & Shafiq, M. O. (2020). Short-term stock market price trend prediction using a comprehensive deep learning system. *Journal of Big Data*, *7*(1), 66. https://doi.org/10.1186/s40537-020-00333-6
16. Shih, D.-H., Hsu, H.-L., & Shih, P.-Y. (2019). A Study of Early Warning System in Volume Burst Risk Assessment of Stock with Big Data Platform. *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 244–248. https://doi.org/10.1109/ICCCBDA.2019.8725738
17. Sirignano, J., & Cont, R. (2018). *Universal Features of Price Formation in Financial Markets: Perspectives From Deep Learning* (SSRN Scholarly Paper 3141294). https://doi.org/10.2139/ssrn.3141294

18. Wang, X., Phua, P. K. H., & Lin, W. (2003). Stock market prediction using neural networks: Does trading volume help in short-term prediction? *Proceedings of the International Joint Conference on Neural Networks, 2003.*, *4*, 2438–2442 vol.4. https://doi.org/10.1109/IJCNN.2003.1223946
19. Yoo, P. D., Kim, M. H., & Jan, T. (2005). Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation. *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, *2*, 835–841. https://doi.org/10.1109/CIMCA.2005.1631572