IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 7, July 2025

DOI: 10.17148/IJARCCE.2025.14731

Review Paper on Predicting Stock Prices with Machine Learning Using Random Forest Algorithm

Mayur D. Nikam[1], Rohit N. Nikam[2], Sunita N. Deore[3]

PG student, K.T.H.M. College, Nashik, India[1]

PG student, K.T.H.M. College, Nashik, India[2]

Assistant Professor, K.T.H.M. College, Nashik, India[3]

Abstract: The precise forecasting of stock market prices presents a tough challenge, owing to the crucial volatility and details of the market. In this research paper, we tackle this challenge by introducing a stock market prediction model grounded in the Random Forest algorithm. Our study centers on historical trading data encompassing a diverse array of stocks and ETF funds, harnessing the capabilities of AI technology and machine learning methodologies to forecast and scrutinize stock prices through regression analysis. The outcomes underscore the Random Forest model's capacity to achieve commendable accuracy in stock prediction, thereby offering invaluable insights for both institutional and individual stock investments. These models rely on technical indicators as inputs, with the closing value of stock prices serving as the predicted variable. The results not only underscore the effectiveness of our proposed approach in constructing predictive models for stock price projection but also highlight the potential of Machine Learning algorithms to reveal valuable insights into the dynamics of stock market activity. Moreover, our paper investigates the exploration of diverse Machine Learning models, encompassing Linear Regression, Support Vector Regression, Decision Tree, Random Forest Regressor, and Extra Tree Regressor. Their implementation has proven instrumental in achieving precision in stock price prediction and has furnished fresh perspectives into the intricate interplay between buyers and sellers in the stock market. The evaluation of these models is grounded in their accuracy in predicting stock prices, using both closing values and stock prices as crucial metrics.

Keywords: stock, random forest, prediction, tree, regressor, Machine Learning

I. INTRODUCTION

Academia and industry have been interested in and conducted a great deal of study on stock price prediction. Numerous methods for predicting changes in the equities market have surfaced since the advent of artificial intelligence (AI) and machine learning algorithms. The use of the Random Forest algorithm is one such well-known method. A potent ensemble learning technique that has drawn a lot of interest in stock price prediction is the Random Forest algorithm. This algorithm creates accurate forecasts by integrating several decision trees and influencing the collective intelligence of a wide range of predictors. Because of its capacity to manage big datasets, identify intricate patterns, and reduce overfitting, it has been extensively used in the banking industry.

There are numerous crucial phases involved in using the Random Forest algorithm to anticipate stock prices. First, historical stock market data is gathered and pre-processed, including price, volume, and other pertinent factors. After that, the algorithm builds a collection of decision trees, each of which has been trained on a distinct portion of the data. The method generates a reliable and accurate prediction of future stock values by combining the forecasts of several separate trees. When it comes to stock price prediction, the Random Forest algorithm has a number of advantages over alternative machine learning methods. Its capacity to manage high-dimensional feature spaces makes it possible to incorporate a variety of elements that affect stock prices, including macroeconomic indicators, market sentiment, and business financials.

We use to utilize historical trading data of multiple stocks and ETF funds as the research objects. These datasets enable us to capture the diverse dynamics of the stock market and evaluate the effectiveness of the Random Forest algorithm in different financial contexts. Furthermore, we address the presence of missing values in the datasets and employ suitable preprocessing techniques to ensure the reliability of our predictions. By influencing AI technology and machine learning methods, specifically regression prediction, we always analyze and predict stock prices. We investigate the impact of various model parameters, such as the number of estimators, maximum depth, minimum samples split, and minimum samples leaf, on the prediction accuracy. Through severe evaluation using appropriate metrics like R2 score, we assess the performance of the Random Forest model.

The accurate prediction of stock prices has significant implications for guiding both institutional and individual stock investments. By achieving desirable accuracy in stock prediction, the Random Forest model offers valuable insights into market behavior and facilitates informed decision- making. We explore the interpretability of the model's predictions and examine its potential in understanding the activity between buyers and sellers in the stock market. Implications for Stock Investments: Both institutional and individual stock investments are significantly impacted by the ability to predict stock prices with accuracy. The Random Forest model provides useful insights into market behaviour and enables well-informed decision-making by predicting stocks with an acceptable level of accuracy. We investigate the model's capability for comprehending the activity between buyers and sellers in the stock market as well as the interpretability of its predictions.

With the help of our paper, we hope to improve knowledge of stock market dynamics and further this area. In this study, we investigate the effectiveness and possible benefits of using the Random Forest algorithm to predict stock prices. We examine the pertinent literature, including earlier studies that have sparked this algorithm in other financial settings. We also go over the algorithm's drawbacks and restrictions in an effort to shed light on its usefulness in real-world situations.

II. LITERATURE REVIEW

In financial engineering, stock price prediction has been a hot topic that has drawn constant research efforts to create efficient methods and strategies. The abundance of financial data, such as technical indicators, sentiment research from social media platforms, and quarterly financial ratios, has created new opportunities to investigate the connection between these variables and stock market activity. Machine learning techniques, especially the Random Forest algorithm, have drawn interest recently due to their potential for precise stock price prediction. The purpose of this literature review is to examine the main ideas and conclusions from a number of studies that employ machine learning and Random Forest approaches to predict stock prices.

Utilizing Financial Ratios and Technical Analysis:
Several papers explored the use of financial ratios and technical indicators as input features for stock price prediction models. Loke [1] and Zi et al. [2] used quarterly financial ratio data to predict stock price movements. They found that while the Random Forest method exhibited weak accuracy over multiple quarters, it achieved high accuracy in specific periods, emphasizing the non-stationary nature of stock price signals. Furthermore, Du et al. [3] used historical trading data and applied Random Forest to analyse stock prices, demonstrating the potential of machine learning techniques in guiding institutional and individual stock investments.

Optimization and Ensemble Approaches:
Efforts to optimize the Random Forest algorithm were explored in multiple papers. Zi et al. [2] proposed a prediction model based on weighted random forest and the ant colony algorithm, achieving lower prediction error compared to traditional Random Forest and regression algorithms. Sharma and Juneja [4] introduced LSboost, combining predictions from an ensemble of trees in a Random Forest. Their approach outperformed Support Vector Regression, offering an effective model for stock market index prediction. Shrivastav and Kumar [5] developed an ensemble model comprising Deep Learning, Gradient Boosting Machine (GBM), and Random Forest techniques. Their findings showcased the superior performance of the ensemble model in terms of accuracy and error reduction.

This paper explores the use of recurrent neural networks (RNNs) [6] in finance for predicting stock closing prices and analyzing real-time sentiment data. The authors have implemented a web app using Django and React, which displays live prices and news, bridging the frontend with a machine learning model built using Keras and TensorFlow. The authors investigate various machine learning techniques, including random forest and support vector machine, for stock prediction. They emphasize the significance of data preprocessing and [7] explore the applicability of the prediction system in real-world scenarios. Their work involves data preprocessing of stock market prices from the previous year and examines the use of prediction systems in practical scenarios.

The study applies Artificial Neural Networks (ANNs) [8] and Random Forest techniques to predict the closing prices of stocks across various sectors. The models employ financial data, including open, high, low, and close prices, and achieve efficient stock closing price prediction, as evidenced by low RMSE and MAPE values. Shen and Shafq [9] focus on deep learning for predicting stock market price trends. They conduct extensive feature engineering and introduce a customized deep learning-based system for price trend prediction.

DOI: 10.17148/IJARCCE.2025.14731

Their work demonstrates high accuracy in stock market trend prediction, highlighting the importance of feature engineering and data preprocessing. The authors aim to forecast stock index prices during the COVID-19 period, employing machine learning models such as the autoregressive deep neural network (AR-DNN) and autoregressive random forest (AR-RF) [10]. Their models outperform traditional time-series forecasting models, offering valuable insights for investors and policymakers during turbulent times. The study also emphasizes the implications for investment decisions and financial policies.

Thus, this literature review highlights the growing interest in using Random Forest and machine learning techniques for stock price prediction. Researchers have explored the use of financial ratios, technical analysis, and sentiment analysis from social media platforms to enhance prediction accuracy. Efforts to optimize the Random Forest algorithm and the application of ensemble models have demonstrated promising results in improving prediction performance. However, challenges such as the non-stationary nature and volatility of the stock market persist. Further research is needed to refine these models and address these challenges, potentially leading to more reliable stock price predictions.

III. PROPOSED ALGORITHM

Random Forest Algorithm: -

The Random Forest algorithm is learning method that combines the predictions of multiple decision trees to make accurate predictions or classifications. It is widely used in machine learning for tasks such as regression and classification. Random Forest is known for its ability to handle complex problems, large datasets, and mitigate overfitting. By creating an ensemble of decision trees and introducing randomness in feature selection, it improves the model's performance and generalization ability.

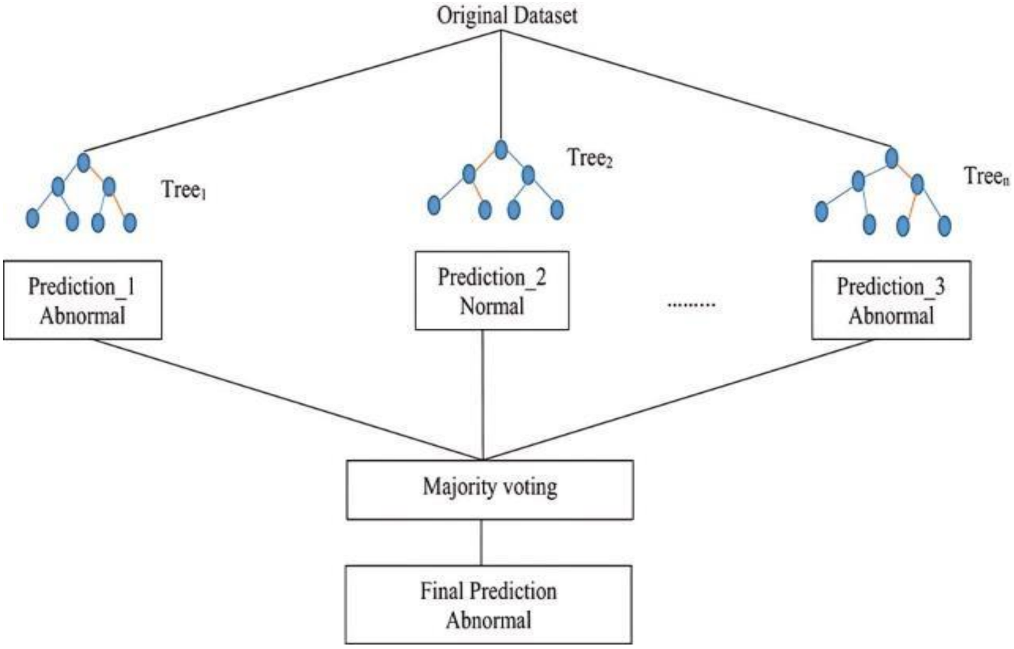

Fig 3.1 Random Forest Algorithm Architecture (source: https://www.geeksforgeeks.org)

1. Initialize the number of decision trees (n_estimators) and other hyper parameters such as the maximum depth of trees and minimum samples per leaf.
2. For each tree in the forest:
a. Randomly select a subset of the training data with replacement (bootstrap sampling).
b. Randomly select a subset of features for the tree.
c. Grow a decision tree using the selected data and features:
If the maximum depth is reached or the number of samples are below the minimum samples per leaf, stop growing the tree.

- Otherwise, find the best feature and threshold to split the current node based on information gain or impurity decreases.
- Split the node into child nodes based on the chosen feature and threshold.

DOI: 10.17148/IJARCCE.2025.14731

- Recursively apply the splitting process to each child node until stopping criteria are met.

3. Prediction:

- For a new test instance, pass it through each decision tree in the forest.
- Aggregate predictions from all trees using averaging (for regression) or voting (for classification).
- Obtain the final prediction for the test instance.

Input:

- Training dataset: A collection of labelled data instances used to train the model.
- Testing dataset: Unlabelled data instances used for prediction and evaluation.
- Split the node into child nodes based on the chosen
- Unlabelled data instances used for prediction and evaluation.

Output:

- Predicted labels or values for the test instances in the testing dataset

IV. SYSTEM ARCHITECTURE

Data Collection and Preprocessing: Reputable sources, including financial databases or APIs, provided the stock market data used in this investigation. Relevant characteristics including stock prices, volume, and other indicators are included in the dataset. The gathered data goes through preprocessing procedures to guarantee data quality. This include dealing with missing values, cleansing the data, and making sure the data is consistent.

Feature Selection: When it comes to stock market forecasting, feature selection is essential. To find the most pertinent features that significantly affect the stock price forecast, feature selection is done in this study. To ascertain the feature relevance, methods like Select from Model with Random Forest Regressor are used. This procedure enhances the model's effectiveness and interpretability while lowering its dimensionality.

- Random Forest is a popular choice for stock prediction due to its ability to handle high-dimensional data and perform feature selection internally. It works by training multiple decision trees on random subsets of data and averaging predictions, reducing overfitting and variance.

- **Pros:** Random Forests are robust to overfitting, especially with high-dimensional data, and provide feature importance scores, which are useful for feature selection.

- **Cons:** Random Forests may struggle with sequential dependencies in time-series data. While they can capture nonlinear relationships, they may not handle temporal dependencies as well as algorithms like LSTMs.

Optimising hyperparameters with Grid Search CV:

Hyperparameters are those that the user must set and that the model does not learn. To maximise the model's performance, these hyperparameters must be adjusted. GridSearchCV was utilised for hyperparameter tuning in this project. GridSearchCV uses cross-validation to assess the model's performance after conducting an exhaustive search across a given parameter grid. The evaluation metric was used to determine the optimal set of hyperparameters.

Data Splitting, Model Training, and Evaluation:

The dataset was divided into training and testing sets to assess the performance of the Random Forest model for stock market prediction. An 80-20 split was utilized, where 80% of the data was allocated for training the model, and the remaining 20% was reserved for evaluating the model's predictive capabilities.

Evaluation Metric-R2 Score:

The coefficient of determination, or R2 score, was selected as the evaluation metric to gauge how well the stock market prediction model performed. The R2 score calculates the percentage of the target variable's (stock prices') volatility that can be accounted for by the input features. Better prediction accuracy and a better model fit to the data are indicated by a higher R2 score.

Where:

✓ SSR (Sum of Squared Residuals) is the sum of the squared differences between the predicted values and the actual values.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 7, July 2025

DOI: 10.17148/IJARCCE.2025.14731

✓ SST (Total Sum of Squares) is the sum of the squared differences between the actual values and the mean of the actual values.

✓

SSR = Σ(y_pred - y_actual)^2
SST = Σ(y_actual - y_mean)^2

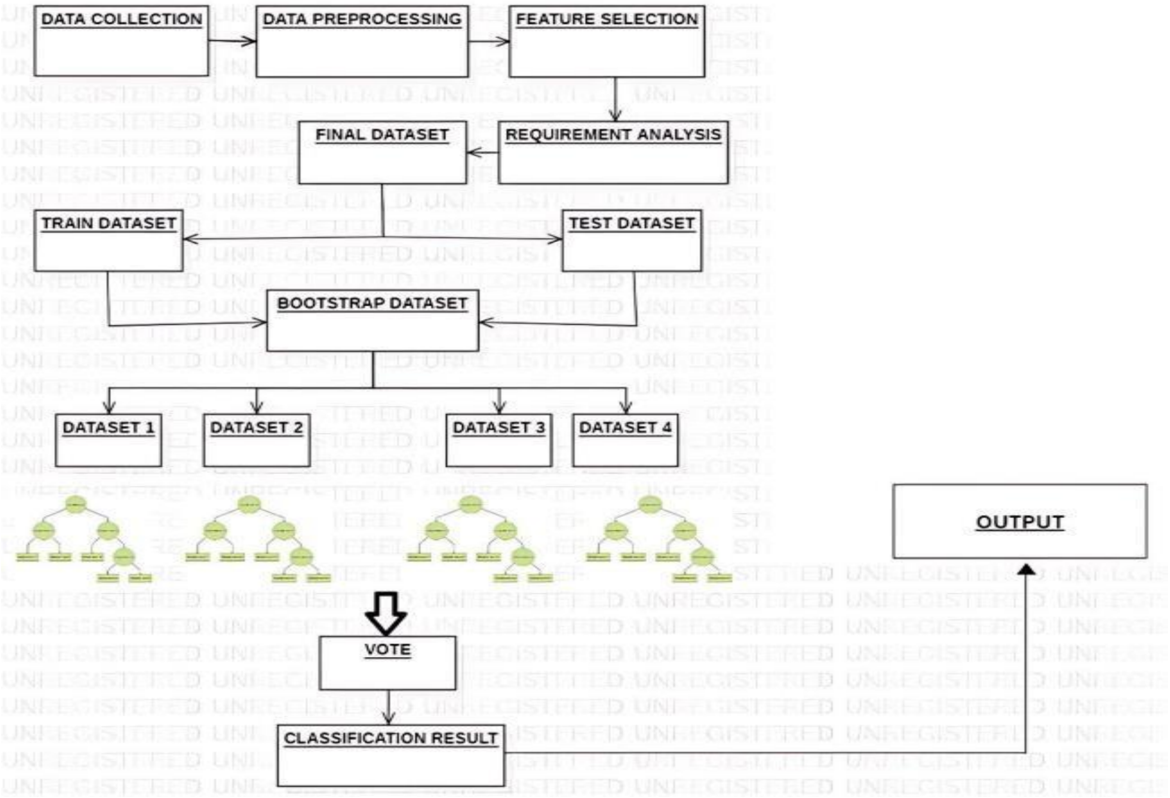

Fig 3.2 System Architecture

V. CONCLUSION

In conclusion, this research paper has explored the application of the random forest algorithm, combined with the grid search technique, for stock market prediction. The incorporation of grid search has enhanced the performance of the random forest algorithm by systematically exploring a range of hyperparameter combinations and selecting the optimal ones based on cross-validation. This process has resulted in improved prediction accuracy and generalization capabilities of the model. Additionally, by adjusting the model's parameters to the unique features of the stock data, the grid search technique has improved the model's resilience to overfitting.

When combined with grid search, the random forest method has demonstrated remarkable prediction accuracy and resilience, making it a useful tool for stock market forecasting. The feature importance analysis carried out in this study has shed light on the major variables affecting stock market behaviour and offered useful data for making investing decisions. But it's important to recognise that stock market forecasting is difficult by nature because it depends on a number of uncontrollable variables. The effectiveness of the grid search method and the random forest algorithm in stock market prediction has been shown in this study.

Predictive modelling in the financial industry is advanced by the conclusions and insights gained from this study. Investors and financial analysts can improve their decision-making, risk management, and comprehension of stock market dynamics by utilising the random forest algorithm's capability and optimising its performance through grid search. Future studies should keep examining cutting-edge methods and adding more variables to enhance the precision and resilience of predictive models because stock market forecasting is still a dynamic and complicated field.

DOI: 10.17148/IJARCCE.2025.14731

REFERENCES

[1]. Du, S., Hao, D., and Li, X.,. "Research on stock forecasting based on random forest," in 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA), Oct. 2022.

[2]. Rajkar, A., Kumaria, A., Raut, A., Kulkarni,N. "Stock Market Price Prediction and Analysis." 2021.

[3]. Zi, R., Jun, Y., Yicheng, Y., Fuxiang, M., and Rongbin, L., "Stock price prediction based on optimized random forest model," in 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML), May 2022.

[4]. VivekaPriya, N., and Geetha, S.. "Stock Prediction Using Machine Learning Techniques.",2022.

[5]. Park, S., Yang, J.-S., "Machine learning modeling to forecast uncertainty between capital sudden stop and boom.",2023.