MANIPAL INSTITUTE OF TECHNOLOGY
ICT DEPT – VI SEM DWDM LAB PROJECT REPORT

'InvestiQue'

Applying Random Forest Regression for Personalized Investment Strategies Based on Acceptable Risk Factor

Date: 14th April 2023
Batch: IT-B1

Members:
ADITYA CHOUBEY 200911076
SHREYANSH RAI 200911094
RAZA ABBAS 200911104

Abstract

Millions of people all over the world invest in stocks daily, taking the risk of losing their investment to the erratic nature of the stock market. This project aims to develop a helpful computer aid for new investors looking to enter the stock market. The stock recommendation takes into consideration the desired investment period and the amount of risk accepted by the investor to recommend stocks registered under the National Stock Exchange of India. For the primary regression model, the following models were tested - ARIMA, Linear Regression, Decision Tree, and Random Forest. Our findings show that for an interactive and responsive recommendation system, Random Forest gives the best output given a vast training dataset.

I. Introduction

In recent years, the rise of the stock market has piqued the curiosity of investors looking to profit from its prospective returns. This has resulted in a rise in the development of stock price prediction and recommendation systems that use a variety of computer techniques to deliver significant insights to investors. Based on historical data, market trends, and other pertinent criteria, stock price prediction and recommendation systems seek to estimate future stock prices and discover investment possibilities. For investors, accurate stock forecasts and recommendations are essential since they have a big influence on their financial decisions and investment choices.

The stock market is volatile and unpredictable. It is influenced not only by historical patterns but also by contemporary events. However, machine-learning-based methods provide a quick and effective technique for estimating the state of the stock market in a few days. It does not take into account what is going on in the world right now. Nonetheless, it gives us useful patterns mined from previous data, which proves to be a significant aspect of long-term investment.

In this research, we offer a stock recommendation system for long-term investments that is responsive to the user's risk tolerance and investment period preferences. We analyze and compare the effectiveness of four key algorithms in predicting stock values across different markets and timeframes, showing their strengths and limits. Furthermore, we analyze the strategies utilized in this research, such as data collecting, preprocessing, and mining approaches, to provide insights into their effectiveness and implementation issues.

II. Literature Survey

To develop our Stock value prediction and Stock Recommendation System, we surveyed multiple IEEE conference and journal research papers. We have limited our literature survey to focus on the 4 primary algorithms we used for comparison-Linear Regression, Decision Tree, Random Forest, and ARIMA.

In [1], the authors use historical data of the stocks to train and test their proposed model. They also evaluate the performance of various time-series algorithms such as ARIMA, Exponential Smoothing, and Prophet to predict future stock prices. Their findings show that ARIMA and LSTM give relatively low errors in predicting stock prices, and ARIMA is better than LSTM in predicting the trend of stock prices.

In [2], the use of machine learning algorithms for stock price forecasting in Indonesia's telecommunications industry is investigated by the authors. With an average MAPE of 4.03%, XGBoost was found to perform better than other algorithms in terms of accuracy.

In [3], the performance of the authors' LQ45 stock price prediction algorithms on the Indonesia Stock Exchange is compared to SMO regression, random forest, and linear regression. With an average MAPE of 2.34%, Random Forest outperformed the competition.

In [4], the application of a Long Short-Term Memory (LSTM) deep learning model for stock price forecasting in the Indian stock market is investigated by the authors. The LSTM model proved successful in accurately predicting stock values, according to the authors, who trained and tested it using historical stock prices.

In [5], the application of decision tree regression for stock price forecasting on the Indonesia Stock Exchange, notably during the COVID-19 outbreak, is examined by the authors. The decision tree regression model was proven to be highly accurate at forecasting stock prices by the authors using COVID-19 data and historical stock prices for training and testing.

In [6], the authors investigate the application of different machine learning algorithms for stock price forecasting in the Indian stock market. The accuracy of Random Forest, with an average MAPE of 0.065%, outperformed that of Linear Regression, Decision Tree, and other algorithms, according to the authors' comparison of their performance.

In [7], the authors look into how the Extreme Gradient Boosting (XGBoost) algorithm might be used to predict stock prices on the Chinese stock exchange. To evaluate the stock data, the authors also employed data visualization tools. With an average MAPE of 1.85%, the results demonstrated that XGBoost performed better than other methods.

In [8], the use of machine learning algorithms for stock price prediction in the Indian stock market is investigated by the researchers. With an average MAPE of 0.53%, Random Forest outperformed Linear Regression, Decision Tree, and other algorithms in terms of accuracy, according to the authors' comparison.

In [9], the authors evaluate the effectiveness of Long Short-Term Memory (LSTM) and Linear Regression models for stock price forecasting in the Indian stock market. The authors discovered that the LSTM model beat the linear regression model in terms of accuracy and provide a unique method to reduce the mean square error in the linear regression model.

In [10], the authors investigate the use of linear regression and decision tree regression for stock price prediction in the Bangladesh stock market. The authors used historical stock prices to train and test the models and found that decision tree regression outperformed linear regression in terms of accuracy.

In [11], the application of linear regression and decision tree regression for stock price prediction in the Bangladesh stock market is examined by the authors. The models were trained and tested using historical stock prices,

and the authors discovered that, in terms of accuracy, decision tree regression surpassed linear regression.

In [12], the authors investigate the use of data mining methods to examine Netflix's stock price changes. The authors identified the elements influencing the stock price using historical stock data and discovered that corporate news, market trends, and financial performance are key drivers.

In [13], the authors suggest an improved Random Forest (RF) model forecast stock prices in the Chinese stock market. The model was trained and tested using historical stock data and technical indications by the authors, who discovered that the optimized RF model performed more accurately than the conventional RF model.

In [14], the Random Forest (RF) algorithm's effectiveness for predicting stock market trends in the Indian stock market is assessed by the authors. The RF model outperformed the other models in terms of accuracy, according to the authors' comparison of it with other machine learning models.

In [15], the authors suggest a stock trading strategy for the Chinese stock market based on the Autoregressive Integrated Moving Average (ARIMA) model and a greedy algorithm. In terms of return on investment, the author discovered that the suggested approach performed better than the buy-and-hold strategy.

III. Methodology

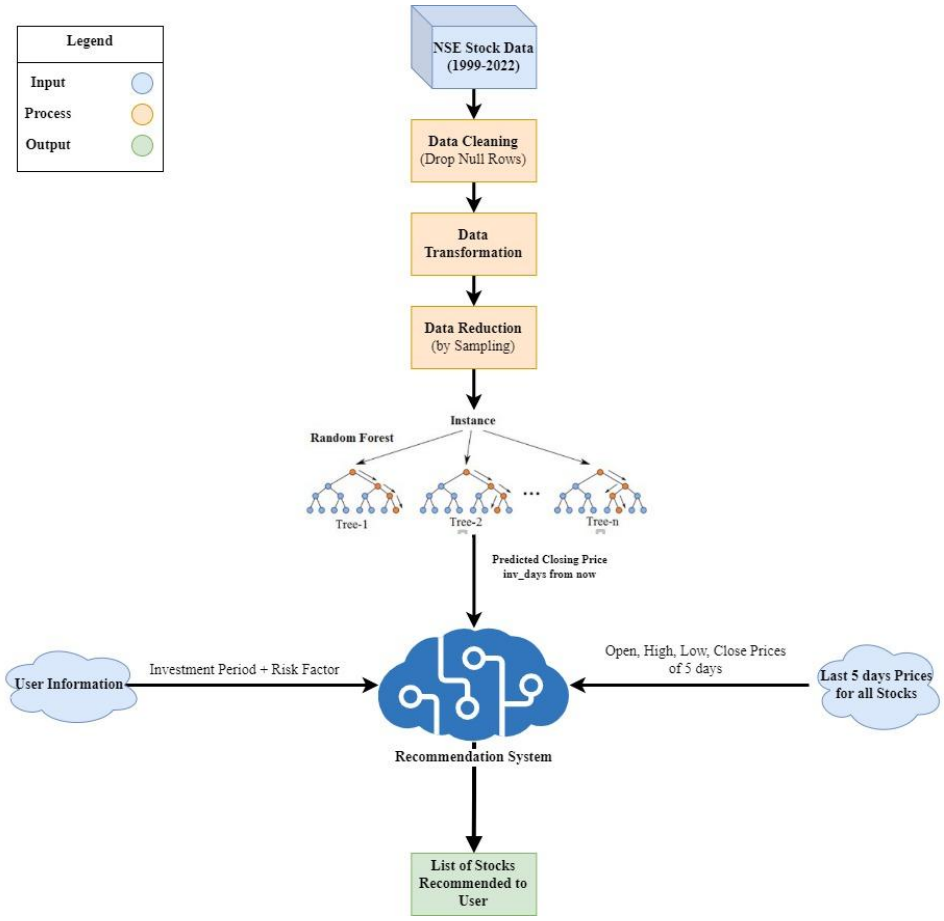

FIGURE 1(a). Block Diagram for Methodology

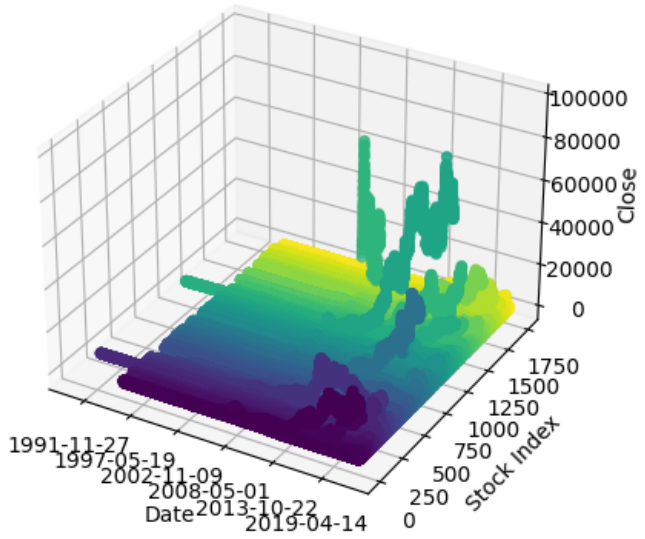

FIGURE 1(b). Raw NSE Stock Dataset from 1991-2022

A. Data Collection

The raw data for the proposed stock price prediction system was the consolidated historical stock prices from India's National Stock Exchange(NSE), from 1991 to 2022. The raw data contains the daily Open, High, Low, Close, Volume, and % Change values in INR for stocks of all companies registered under NSE. Fig. 1(b) shows the Close prices plotted against the Date for each Stock.

Note that in our dataset, each stock is denoted with an index **i**. This was done so that we can use the Stock Index value for our training process. The Stock Name to Stock Index mapping is stored in an array and is constant for the entire system.

B. Data Preprocessing

Once the raw data is collected, it is passed through a 3-stage process, as shown in Fig.1(a), before any data-mining algorithms are performed on it. The three stages are namely, data cleaning, data transformation, and finally data reduction.

The raw data, being a vast consolidated dataset of all companies registered since 1999, must first be cleaned before we use it for any further steps. Any rows in the dataset containing empty(or NaN) values are dropped from the dataset. This ensures the system only processes and trains based on valid data, and not on incomplete/invalid data.

Once the dataset is cleaned to ensure the purity of data, it is passed through a data transformation procedure. The aim of this procedure is very simple - to generate a complete training input(**X**) and output(**y**) for the regression step(explained in Section III[C]). The training input **X** is generated considering a window of 5 days. For each company with a mapping index **i**, the training input subset **Xi** is the list of Open, Close, High, and Low of 5 consecutive days for all days the company's stock was traded. This is done so that our model is considerate of recent trends in the company's stock prices. The training output **y** is the Close price for each stock's day **(inv_days+5)** from that day.

On transforming the dataset to generate **X** and **y**, it is randomly sampled such that all stocks are equally sampled(the ratio of the number of data points for all stocks remains the same). This is done to reduce the training

time for the Random Forest Regressor. For testing purposes, 10% of the dataset is sampled and used for further analysis steps.

C. Data Mining using Random Forest Regression

As shown in Fig. 1(a), the proposed stock prediction system's central algorithm is the Random Forest Regression algorithm. The RF Regression is fitted on **X** and **y** obtained after the steps described in Section III[B]. The final results are demonstrated in Fig. 3. For the actual prediction step, stock prices for the last **5 days** are fetched for all the stocks. This data is arranged accordingly(in a similar way as described in Section III[B]) and fed into the trained RF model. The output is consolidated and passed onto the final step of the stock recommendation system.

D. Generating Recommended Stocks Based on Risk Factor

Based on the acceptable risk factor entered by the user, the following recommendation process provides the list of stocks satisfying the constraints. First, the predicted prices from the previous step are used to develop the metric used in the procedure - the percentage change in Closing price inv_days from now. This metric is then mapped to the risk factor, which is then used to simply filter out any stocks which have a higher predicted risk factor.

IV. Future Scope

Our stock recommendation system takes into account only the historical prices in NSE to suggest stocks suitable for the investor. However, this means that the stock price prediction only reacts to the actual stock's performance. It does not take into consideration the news on the company and industry. To prevent this, our proposed system could be augmented using NLP techniques to take into consideration the current affairs to predict the risk factor and closing price of a company's stock. With our current model, it would be incredibly trivial to develop an NLP model that parallelly predicts another metric based on the most recent news of a company. The metric, denoting whether the company's image in the world is good or bad, could then be used to predict the stock price and risk factors in the future.

V. Results

In the below figures, the stock value predictions are plotted out, taking TCS as our dataset. The 4 figures denote the 4 different prediction algorithms tested- Linear Regression, Decision Tree, Random Forest, and ARIMA. It must be noted that, even though ARIMA gives the best results, it fails to be scalable, as it has an O(n2T) time complexity and takes a considerable amount of time to produce the prediction outputs. Linear Regression fails to follow the stock rising and falling trends, and as a result, fails in basic prediction tasks. Decision Trees produce an acceptable, yet flawed prediction output. Thus, we base our entire Stock Recommendation System based on Random Forest Regression. This enables us to produce satisfactory results optimally, such that our recommendation system feels as responsive as possible.

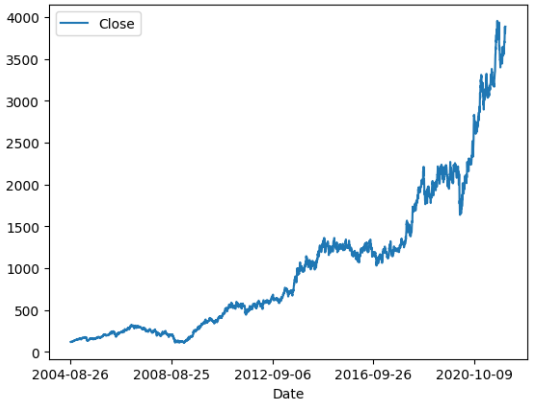

FIGURE 2. TCS Stock Value Close Price Dataset

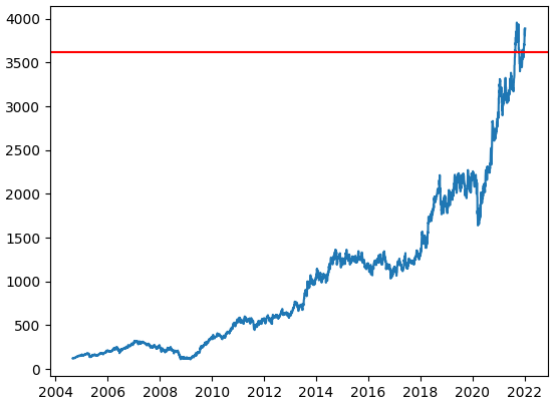

FIGURE 3. TCS Stock Value Prediction of 100 days into the future using a Random Forest Regressor. Model Score = 0.9971745428788035

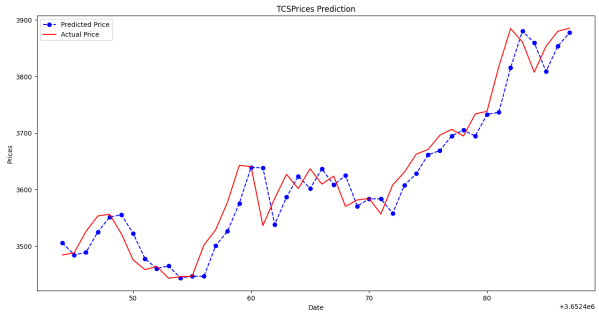

FIGURE 4. TCS Stock Value Prediction of 100 days into the future using Auto-Regressive Integrated Moving Average(ARIMA) with parameters (4,1,0)

FIGURE 5. TCS Stock Value Prediction of 100 days into the future using Linear Regression. Model Score = 0.96703893

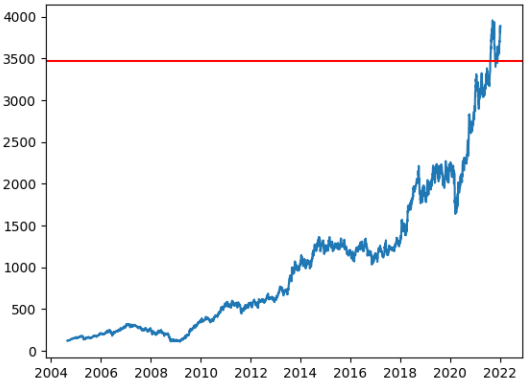

FIGURE 6. TCS Stock Value Prediction of 100 days into the future using Decision Tree Regressor. Model Score = 1.0

References

[1] Y. -T. Choy, M. H. Hoo, and K. -C. Khor, "Price Prediction Using Time-Series Algorithms for Stocks Listed on Bursa Malaysia," *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, IPOH, Malaysia, 2021, pp. 1-5.
doi: 10.1109/AiDAS53897.2021.9574445

[2] 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," *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, Manado, Indonesia, 2020, pp. 1-4.
doi: 10.1109/ICORIS50180.2020.9320758

[3] M. Lutfi, S. P. Agustin, and I. Nurma Yulita, "LQ45 Stock Price Prediction Using Linear Regression Algorithm, Smo Regression, And Random Forest," *2021 International Conference on Artificial Intelligence and Big Data Analytics*, Bandung, Indonesia, 2021, pp. 1-5.
doi: 10.1109/ICAIBDA53487.2021.9689749

[4] K. J, H. E, M. S, Jacob, and D. R, "Stock Price Prediction Based on LSTM Deep Learning Model," *2021 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Puducherry, India, 2021, pp. 1-4.
doi: 10.1109/ICSCAN53069.2021.9526491

[5] K. M. Hindrayani, T. M. Fahrudin, R. Prismahardi Aji and E. M. Safitri, "Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression," *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, 2020, pp. 344-347.
doi: 10.1109/ISRITI51436.2020.9315484

[6] P. S. Lakshmi, N. Deepika, V. Lavanya, L. J. Mary, D. R., Thilak, and A. A. Sylvia, "Prediction of Stock Price Using Machine Learning," *2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, Chennai, India, 2022, pp. 1-4.
doi: 10.1109/ICDSAAI55433.2022.10028862

[7] X. Er and Y. Sun, "Visualization Analysis of Stock Data and Intelligent Time Series Stock Price Prediction Based on Extreme Gradient Boosting," *2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, Chongqing, China, 2021, pp. 272-279.
doi: 10.1109/MLISE54096.2021.00057

[8] A. Vij, K., Saxena and A. Rana, "Prediction in Stock Price Using of Python and Machine Learning," *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2021, pp. 1-4.
doi: 10.1109/ICRITO51393.2021.9596513

[9] C. Ebenesh and K. Anitha, "A Novel Approach to Minimize the Mean Square Error in Predicting Stock Price Index using Linear Regression in Comparison with LSTM Model," *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Erode, India, 2022, pp. 1365-1370.
doi: 10.1109/ICSCDS53736.2022.9760764

[10] R. Karim, M. K. Alam and M. R. Hossain, "Stock Market Analysis Using Linear Regression and Decision Tree Regression," *2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, Sana'a, Yemen, 2021, pp. 1-6.
doi: 10.1109/eSmarTA52612.2021.9515762

[11] D. S. Kumar, T. B. C, S. N. R. C, V. A, A. S. Devi, and D. Kavitha, "Analysis and Prediction of Stock Price Using Hybridization of SARIMA and XGBoost," *2022 International Conference on Communication, Computing, and Internet of Things (IC3IoT)*, Chennai, India, 2022, pp. 1-4.
doi: 10.1109/IC3IOT53935.2022.9767868

[12] V. Gowri, B. Harish, F. Ahmed, and M. Srinath, "Netflix Stock Price Movements Insights from Data Mining," *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Mysuru, India, 2022, pp. 1-4.
doi: 10.1109/MysuruCon55714.2022.9972547

[13] R. Zi, Y. Jun, Y. Yicheng, M. Fuxiang, and L. Rongbin, "Stock price prediction based on optimized random forest model," *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, Hangzhou, China, 2022, pp. 777-783.
doi: 10.1109/CACML55074.2022.00134

[14] K. Lavingia, P. Khanpara, R. Mehta, K. Patel, and N. Kothari, "Predicting Stock Market Trends using Random Forest: A Comparative Analysis," *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2022, pp. 1544-1550.
doi: 10.1109/ICCES54183.2022.9835876

[15] Z. Wang, "Trading Method Based on ARIMA Model and Greedy Algorithm," *2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, Dalian, China, 2022, pp. 433-440.
doi: 10.1109/TOCS56154.2022.10016016