Proceedings of the International Conference on Applied Artificial Intelligence and Computing (ICAAIC 2022)
IEEE Xplore Part Number: CFP22BC3-ART; ISBN: 978-1-6654-9710-7

Scalability of Data Mining Algorithms for Non-Stationary Data

Satyajit S. Uparkar
Inter Institutional Computer Centre
RTM Nagpur University
Nagpur, India
uparkarss2204@gmail.com

Ujwal A. Lanjewar
Prerna College of Commerce
RTM Nagpur University
Nagpur, India
ualanjewar@gmail.com

2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) | 978-1-6654-9710-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAAIC53929.2022.9792711

***Abstract*—Streaming data is increasing day-by-day in real world applications such as stock market, financial data etc. The prime quality of this data is the online advent and collection at high speed and vulnerability to transforms in the data distributions due to the dynamic environment. The biggest challenge of analyzing this data is the lack of unrestricted availability of non-stationary dataset for evaluation and comparison. Extracting useful information from non-stationary dataset is also challenging because of its scaling nature. In this paper, different classification techniques of data mining are applied for analyzing non-stationary dataset using different existing and proposed cascade scaling method. The approach begins with the collection of a real time dataset of National Stock Exchange of India NSEI from online Yahoo finance site, which is then pre-processed. The novel approach of three rules has been applied to fix the binary target value based on the buy/sell of the shares on the particular day. Once the dataset has been pre-processed, it is decomposed into smaller datasets of equal size subsets. The particular data mining approach is then applied identically to each subset. The findings of the data mining approach on all subsets are pooled and aggregated for the final output. The performance of the suggested algorithm is assessed using a number of criteria, including accuracy, precision, recall, f-score, and execution time. The comparative analysis of the mentioned performance metrices are evaluated for the various data mining approaches. The next comparative analysis includes the non-scaling and proposed cascade scaling method for the various data mining approaches. Lastly, the performance comparison of existing scaling methods with proposed cascade scaling method for the accuracy parameter is evaluated where proposed method achieved substantial performance on the non-stationary dataset.**

***Keywords*—National stock exchange, non-stationary data, Cascade approach, Scalability, Pooling and Aggregation**

I. INTRODUCTION

In order to build high-performance, efficient, and scalable data mining algorithms, scalability is a fundamental issue that needs to be addressed. Data mining methods can easily extract information from enormous datasets after dealing with the scalability issue. Issues such as the vast size of the dataset, the entire data flow, and the difficulty of data mining techniques encourage the development of scalability-based parallel & distributed data mining algorithms. Scalability-based data mining algorithms that run the complete dataset in less time can be created using some scaling methodologies. Scalability [1] refers to an algorithm's ability to handle increasing amounts of input by adding additional assets to the system. In this case, the system can be scaled up or down depending on the size of the task. Scaling the approach has become a crucial feature as a result of the vast amount of data available today, which is why research has switched to discovering techniques to deal with scalable data. According to the scalability definition, we group the strategies into a variety of types utilizing a combination of scaling and data mining techniques. These tactics are simple to apply to a scalable dataset or a large volume of data [2].

Working with large datasets with hundreds of thousands or millions of instances and tens of thousands of attributes is becoming more common. When faced with these constraints, data mining algorithms become less successful unless they are scaled up appropriately. Both humans and machine learning systems have difficulty dealing with large datasets. Streaming data is becoming more prevalent in real-world applications like the stock market and financial data. The primary quality of this data is its online availability and high-speed gathering, as well as its vulnerability to data distribution transformations owing to the dynamic environment. The lack of an unconstrained non-stationary dataset for evaluation and comparison is the most difficult aspect of studying this data. Because of its scaling nature, extracting usable information from a non-stationary dataset is also difficult. In this study, various data mining classification algorithms are used to analyze a non-stationary dataset using various known data mining approaches and a novel cascade scaling method to build the proposed scaling mechanism. Thus, this research study is an attempt to deal with the scalability issue associated with data mining algorithms by integrating with the machine leaning of training and testing approach. In addition, the research study also aims to provide a novel approach of predicating the binary classification.

The past research study claims that the data set related to stock exchange are most venerable for the prediction. The Data set used in this research study is readily available on Yahoo finance site. The data set of seven attributes can be downloaded from 18th Sept 2007 till date [3].

The NSEI data set with initial seven attributes are given in the following table-

TABLE I. NSEI DATA SET WITH INITIAL ATTRIBUTES

Sr. No.	Attributes	Description
1	Date	In DD-MM-YYYY format and provides the dates of transactions.
2	Open	Open stock price of the day.
3	High	Highest stock price of the day
4	Low	Lowest stock price of the day
5	Close	Close price of stock at the closing time of the day.
6	Adj Close	Adjusted close price adjusted for addition of dividend and/or capital gain distributions.
7	Volume	Transaction of number of shares

The preprocessing steps and the innovative rules for determining the final binary Target value of buy/sell, are discussed in the section of proposed methodology.

II. Literature Review

In this research study, we look at the many solutions that have been used to address the scalability problem. To produce a broad assessment of popular strategies for scaling up data mining technology, we focus on basic concepts rather than specific implementations. Following this broad examination, we may build a taxonomy of the algorithms used in the suggested research technique. Many method-based publications, as well as a brief synopsis of recent research effort, are appraised here. Several studies have been conducted on implementing machine learning algorithms for stock market forecasting.

Luca and Honchar (2017), worked on the Google Stock dataset. A study used RNN, LSTM, and Gated Recurrent Unit (GRU) and discovered that LSTM outperforms other methods. Two scaling method were used for data transformation viz. Normalization and min-max scaling. [4]. Patel J. and et. al (2015) the Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest, and Naive-Bayes approaches are compared [5]. Ballings M and et. al (2015) compared Random Forest, AdaBoost, Kernel Factory, Neural Networks, Logistic Regression, Support Vector Machine, KNN, and other machine learning approaches using a dataset of European organizations [6]. Usmani M. and et .al (2016), performed a study using machine learning techniques on the Karachi Stock Exchange (KSE). It compared SLP, MLP, RBF, and SVM. When compared to the others, MLP performs the best [7]. Milosevic N. (2016), test various machine learning techniques (SVM, Nave Bayes, Random Forest), and the random forest algorithm yielded the greatest F-score [8]. Bhardwaj A. and et. al. (2015), used Unsupervised learning as a major contributor to process that gives in one investigation. The internet-based technologies including the opinion mining, cloud computing and big data analysis were integrated for processing the stock exchange data set. [9]. Roondiwala M and et.al. (2017) performed the comparative analysis of RNN and LSTM approach on the stock exchange data set. The end results indicate LSTM may be used to anticipate Nifty values [10]. Yang B and et.al (2017) introduces a multilayer feed forward network technique based on the Chinese Stock dataset. A Bagging approach was used to calculate the predictive indices. The error between the actual price and the predictive price was studies for the accuracy of the model [11].

Ashwini Pathak and Sakshi Pathak (2020) performed the comparative study of performance metrics using various machine learning algorithms. On the basis of the various performance metrics, it was observed that the Random Forest was the best algorithm for the prediction of the stock market price [12]. Nti. I.K. and et. al (2020), provided a novel approach of ensemble regressors in combination of the classifiers. The combinations were tested for their accuracy, time and error metrics. Altogether, four stock market data sets over the period of 2012 to 2018 were used for this research study [13]. Dev Shah and et. al. (2019) studied on Taxonomy used in stock market analysis including statistical approaches, pattern recognition, machine learning, sentimental analysis and hybrid approaches. The summary of literature based on these parameters reflects the accuracy rates. The authors had quoted the problem and open challenges associated with the mentioned parameters [14]. Nusrat Rouf, Majid Bashir Malik and et. al, (2021) had talked about the available literature on stock market prediction during the last decade. The two perceptions of fundamental analysis and technical analysis were the base for any traditional model optimization [15]. R. Seethalakshmi (2018), worked on the NIFTY 50 data set to predict the closing stock price using simple linear regression analysis. Two models were studies and compared for their optimal solution. For the comparative analysis various data mining algorithms have been used in this study [16]. Budiharto, W. (2021), worked on the Open High Low close (OHLC) model for short terms gain using LSTM approach. The finding for accuracy of their model were explore during the covid period. LSTM provides an accuracy upto 94.57% for short term epoch of one year. The option provides the better decision for the short-term investor to study the market trend based on the historic data [17].

III. Proposed Methodology

For analysis or to predict the non-stationary data set, it is necessary to convert it to a stationary instant of time. As an initial step of the data preprocessing, the scope of this research study has included the NSEI data set till 31st December 2021. Fig. 1 and 2 of Open and Close price over the time period are the proofs of non-stationarity of the NSEI data set.

Fig. 1. Non stationarity in Open price of NSEI data set

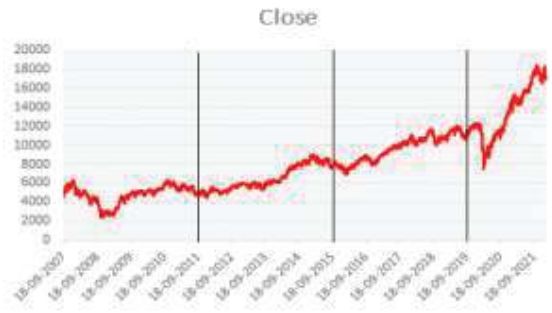

Fig. 2. Non stationarity in Close price of NSEI data set

The visual test indicates that the Open and Close price varies in the four quadrants. The same is also true for the remaining attributes of High and Low. The variation thus proves the non-stationarity characteristics of the NSEI

data set. The summary statistics is another way to prove the non-stationarity. Table II and III are based on the average and standard deviation for the four quadrants of time period are-

TABLE II. VARIATION IN AVERAGE VALUES OVER THE TIME PERIOD

Years	Open	High	Low	Close
2007- 2011	4845.647	4899.065	4787.297	4844.658
2011 - 2014	6051.83	6087.042	6010.335	6049.316
2015- 2018	8941.941	8978.515	8890.101	8934.093
2018 - 2021	12615.15	12680.34	12519.87	12601.85

TABLE III. VARIATION IN STANDARD DEVIATION VALUES OVER THE TIME PERIOD

Years	Open	High	Low	Close
2007- 2011	948.9262	943.672	953.9221	948.2095
2011 - 2014	996.8293	997.1834	997.3834	997.3586
2015- 2018	1026.594	1024.492	1029.367	1027.176
2018 - 2021	2433.782	2431.609	2429.132	2431.656

Thus, the variations in the of Average and Standard Deviation values over the time period, also prove the non-stationarity nature of the NSEI data set.

For dimensionality reduction, the approach of the feature selection method for selecting the most relevant input variables, from the original dataset has been used. After the proof for non-stationarity, Date column is eliminated. The column of Adj. Closed includes the duplicate values of Close price and hence removed. The column of Volume contains maximum zero values in initial 30% of the records and hence Volume column is also dropped. For row reduction, the records of "null" values for Open, High, Low and Close are also eliminated.

The Target attribute is a binary classifier based on buy/sell of the stock price on the day. The target variable value is based on the rules applied from the theories of the stock market applied for buy or sell of the stock. The attribute of final Target is the outcome of the most frequent value of buy/sell derived from the following three innovative rules-

Rule 1: If Open > Close, buy the shares, otherwise sell the shares on that particular day.

Rule 2: If Open=Low, buy or if Open=High, sell otherwise, if Open > average (High, Low), then sell, otherwise buy the shares on that particular day.

Rule 3: Calculate Typical Price (TP) as an average of (Open, High and Low) values of the day. Then apply the rule, if Current TP > Previous TP, then, sell, otherwise buy the shares. The first binary value for this column was adjusted as zero. [3]

The synthesized data set attributes used for this study are given in the following table-

TABLE IV. NSEI DATA SET AFTER PRE-PROCESSING

Sr. No.	Attributes	Description
1	Open	Open stock price of the day.
2	High	Highest stock price of the day
3	Low	Lowest stock price of the day
4	Close	Close price of stock at the closing time of the day.
5	Target	Indicates Buy or Sell of the NSEI stock on the day.

Thus, the NSEI data set having the binary Target value of buy/sell, as the dependent variable whereas Open, High, Low, Close, as the set of independent variables, is then examined for the predictive analysis. The initial accuracy of the complete model was tested using Logistic Regression and imposing the boosting concept [3].

Scalability is an issue that must be addressed in data mining algorithms in order to construct high-performance, efficient, and scalable data mining algorithms. After coping with the scalability issue, data mining technologies can readily extract information from massive databases. In the proposed work, a new scalability method is suggested and applied to different data mining algorithms. The proposed scaling methodology is developed using a cascade approach. The proposed method's technique is depicted in Fig. 3. As an initial step, divide the dataset into equal-sized subsets. For improved performance, the value of the number of subgroups, designated as "n," is identified. The maximum number of subgroups should never be more than n. To find the greatest value of n, apply (1). Overfitting may occur if the number of subsets is bigger than n; consequently, reduce subset division to avoid overfitting.

$$n = math\_sqrt(len(dataset)) \quad (1)$$

After checking for overfitting, divide the dataset into *n* subsets and split each subset into train and test segments. Apply the same data mining strategy to each subgroup and compute the results separately. Combine and aggregate the results from each subgroup to get a final output for all performance measures, including accuracy, precision, recall, and F-score [18]. In terms of performance, the proposed scaling methodology is compared to non-scaling data mining methods. Determine which data mining method performed the best on the given scaling method.

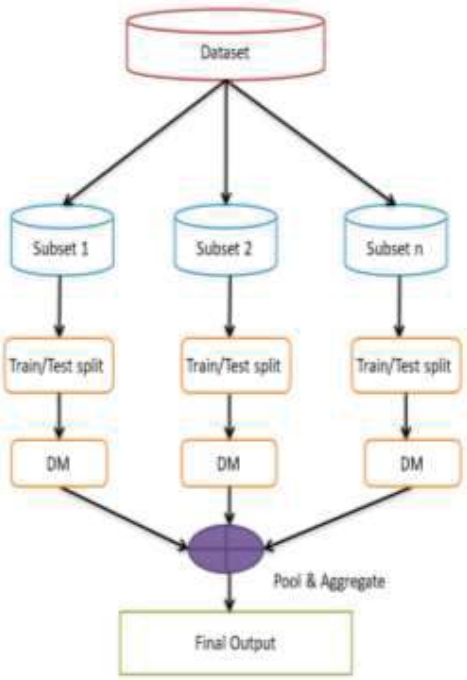

Fig. 3. Procedure for proposed scalable cascade approach

IV. Results And Discussion

The proposed cascade scalability strategy is tested and deployed on the NSEI dataset using multiple data mining methods. The suggested method's performance is assessed using accuracy, precision, recall, F-score, and execution time. Table V shows the results of the suggested cascade methodology on numerous data mining methods for the NSEI Dataset. When compared to other data mining methods, logistic regression performed the best. Several data mining methods were outperformed by the proposed strategy. Because it takes more resources and processing capacity to assemble and combine a large number of decision trees and their outputs, the random forest and Adaboost approaches took substantially longer to execute than other methods for all datasets evaluated in this study.

TABLE V. Performance Evaluation of Proposed Cascade Scaling Method on Different Data Mining Methods for NSEI Dataset

Data Mining Methods	Accuracy	Precision	Recall	F-Score	Execution Time
Decision Tree	0.95	0.95	0.73	0.82	0.0015
Random Forest	0.96	0.95	0.72	0.82	0.139
AdaBoost	0.96	0.95	0.74	0.83	0.127
SVM	0.96	0.95	0.73	0.82	0.021
Logistic Regression	0.98	0.96	0.73	0.83	0.023
K-NN	0.96	0.96	0.74	0.84	0.0027
Naive Bayes	0.93	0.95	0.74	0.83	0.0026

The high values of Accuracy and Precision parameters provide the consistency of the algorithms for the analysis of the non-stationary data set. The Recall and F-score values are substantially less in case of all the data mining approaches. In compared to other methods, logistic regression has the highest accuracy (98%). The longest execution durations are for random forest and Adaboost, whereas the shortest execution times are for other approaches. The tunning of the hyper-parameters and probability threshold of the various data mining methods can lead to achieve higher recall.

Figures 4, 5, 6, 7 and 8 demonstrate a comparison of the proposed method's accuracy, precision, recall, F-score, and execution time in sec., on the NSEI dataset.

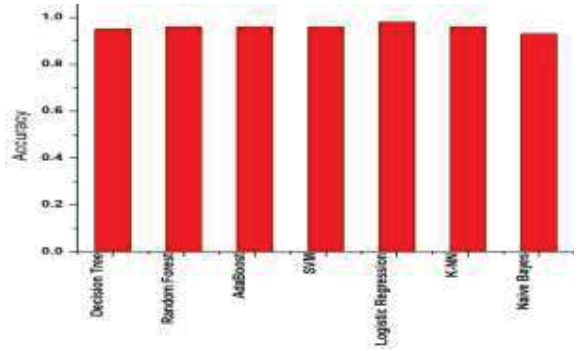

Fig. 4. Comparison of accuracy of proposed scaling method for data mining methods on NSEI dataset

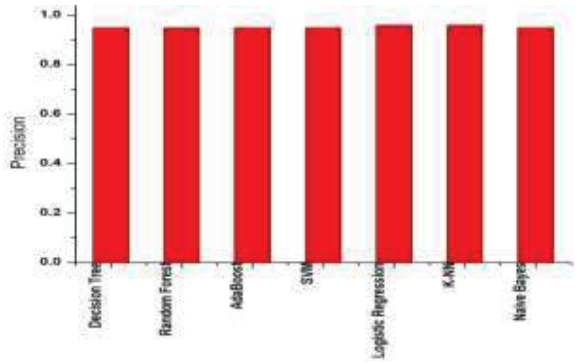

Fig. 5. Comparison of precision of proposed scaling method for data mining methods on NSEI dataset

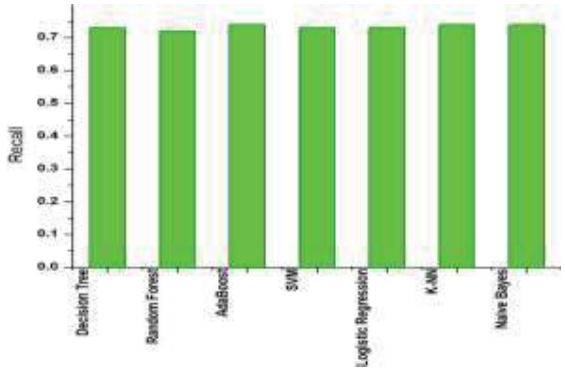

Fig. 6. Comparison of recall of proposed scaling method for data mining methods on NSEI dataset

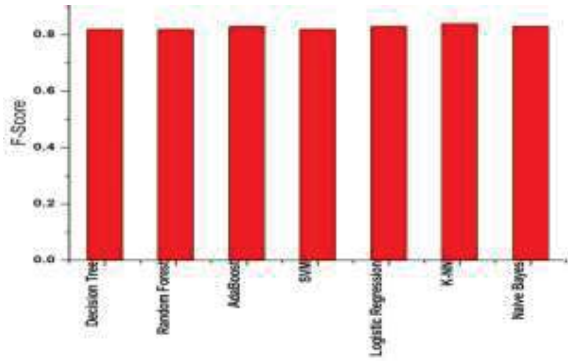

Fig. 7. Comparison of F-score of proposed scaling method for data mining methods on NSEI dataset

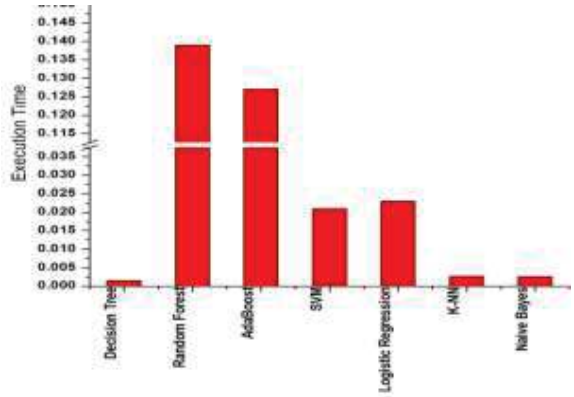

Fig. 8. Comparison of execution time of proposed scaling method for data mining methods on NSEI dataset

For various data mining approaches, the suggested scaling method is compared to non-scaling methods. On the NSEI dataset, tests were conducted using Decision Tree, Random Forest, AdaBoost, SVM, and Logistic Regression. Figures 9, 10, 11, 12 and 13 demonstrate a comparison of the suggested scaling method's accuracy, precision, recall, F-score, and execution time in sec., with non-scaling. For all data mining approaches except logistic regression, the scaling strategy significantly improves recall and F-score when compared to non-scaling. Scaling approaches take significantly less time to process data than non-scaling methods, implying that scaling methods outperformed non-scaling data mining methods on all measures.

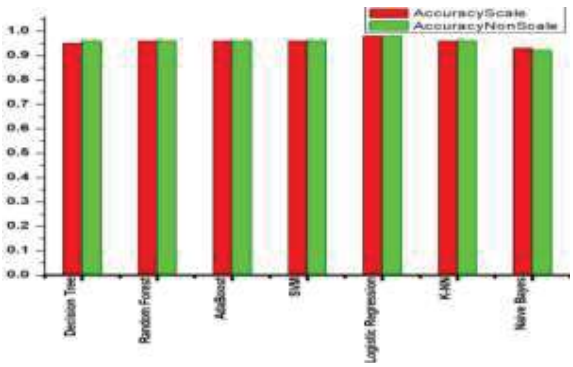

Fig. 9. Comparison of accuracy of proposed scaling method with non-scalingon NSEI dataset

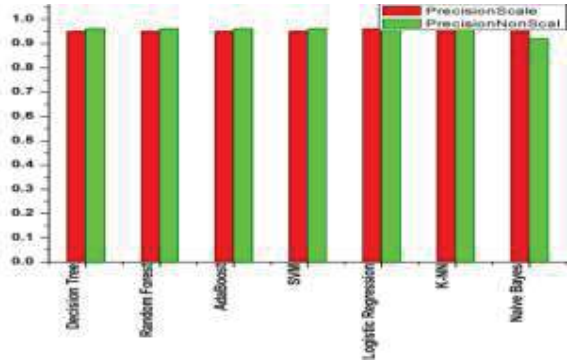

Fig. 10. Comparison of precision of proposed scaling method with non-scalingon NSEI dataset

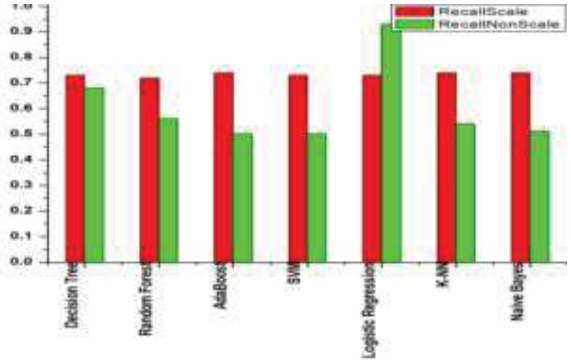

Fig. 11. Comparison of recall of proposed scaling method with non-scalingon NSEI dataset

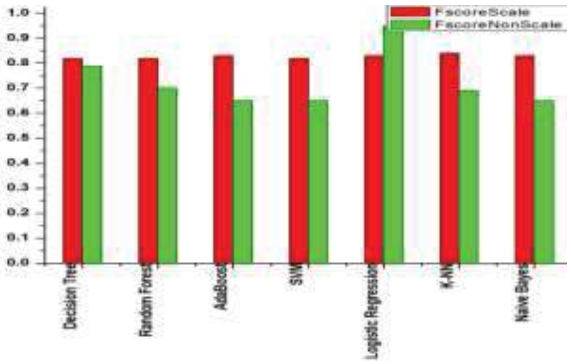

Fig. 12. Comparison of F-score of proposed scaling method with non-scalingon NSEI dataset

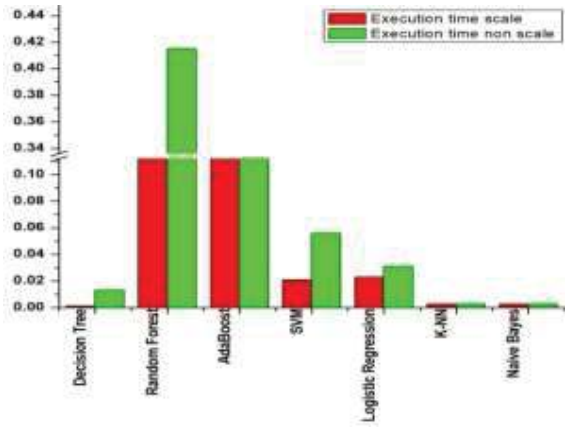

Fig. 13. Comparison of execution time of proposed scaling method with non-scalingon NSEI dataset

For various data mining methodologies, a comparison of existing approaches with the proposed method is also explored. Table VI shows a comparison of the proposed method with existing methods for the NSEI dataset. The table shows the performance various scaling algorithms, including the suggested method, on the NSEI dataset for nave bayes, decision tree, random forest, Adaboost, KNN, SVM, and logistic regression.

TABLE VI. COMPARISION OF EXISTING METHODS WITH PROPOSED IN TERMS OF ACCURACY MEASURE ON NSEI DATASET

Data Mining Methods	Decision Tree	Random Forest	Ada Boost	SVM	Logistic Regression	K-NN
MinMax	0.97	0.96	0.96	0.96	0.96	0.96
Standard Scaling	0.97	0.96	0.96	0.96	0.96	0.96
MaxAbs Scaling	0.96	0.96	0.96	0.96	0.96	0.96
Robust Scaling	0.96	0.96	0.96	0.96	0.96	0.96
Normalization	0.95	0.96	0.96	0.96	0.96	0.97
Proposed Cascade Method	0.95	0.96	0.96	0.96	0.98	0.96

Logistic regression approach provides a higher accuracy in case of the proposed cascade method. The change in various parameters of the data mining algorithm can provide better results. On all data mining approaches for recall, the proposed method performed better than existing scaling methods. In comparison to other scaling approaches, the Min Max, Standard scaling and normalization scaling method also produced more accurate results in case of Decision Tree, Random Forest and K-NN techniques respectively.

V. CONCLUSION

This research study is an attempt to provide initially data analysis steps for proving non-stationarity, data preprocessing and data cleaning. The novel approach of three rules for finalizing the binary Target of buy/sell, provides the scope of optimization for the predictivity analysis of the NSEI stocks.

Because of the vast amounts of data in any domain, data mining and knowledge discovery methods confront considerable obstacles. This problem is overcome by using the scalability idea, which reduces the execution time on huge datasets and remaining data mining methods successful. Thus, scalability factor for a growing data set is made robust by using the concept of training -testing model of machine learning. The research study reflects the readings of optimized outcomes for 80-20 ratio of training -testing of data set, to avoid the overfitting or else underfitting found for 60-40 and 70-30 ratios respectively.

In this work, a cascade technique is used to build the proposed scaling mechanism. A number of metrics are used to evaluate the suggested algorithm's performance, including accuracy, precision, recall, F-score, and execution time on NSEI dataset. The high precision and low recall value is an indication where most of the predicted labels are correct when compared to the training labels. When compared to non-scaling F-score and recall of the scaling method are significantly enhanced. Random Forest takes the longest to run since it has so many decision trees. As a result, trees are multiplied by the number of subgroups when utilizing the scaling strategy. The Random Forest's performance in terms of execution time can be improved in future research. Cascade method can also be helpful in high performance computing systems.

REFERENCES

[1] A. B. Bondi, "Characteristics of scalability and their impact on performance," in Proceedings of the 2nd international workshop on Software and performance, 2000, pp. 195-203.
[2] J. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A scalable parallel classifier for data mining," in Vldb, 1996, pp. 544-555.
[3] Uparkar S.S. , Lanjewar U.A. , Application of Predicitve Analysis for a Non Stationay Data set, ASPIRE 2022, Edition 3, 4-5 Feb., 2022, Page 27, ISBN no. 978 81 946772-3-9.
[4] Luca D P, Honchar O. Recurrent Neural Networks Approach to the Financial Forecast of Google Assets. International Journal of Mathematics and Computers in simulation, 2017, vol. 11, pp. 7–13.
[5] Patel J, Shah S, Thakkar P, Kotecha K. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications, 2015, 42(1), pp. 259-268.
[6] Ballings M, Poel D V D, Hespeels N, Gryp R. Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 2015, 42(20), pp. 7046–56.
[7] Usmani M, Adil S H, Raza K, Ali S S A. Stock market prediction using machine learning techniques. 3rd International Conference on Computer and Information Sciences (ICCOINS), 2016, pp. 322-327.
[8] Milosevic N. Equity Forecast: Predicting Long Term Stock Price Movement Using Machine Learning. arXiv, 2016.
[9] Bhardwaj A, Narayan Y, Vanraj, Pawan, Maitreyee D. Sentiment analysis for Indian stock market prediction using Sensex and nifty. Procedia Computer Science, 2015, 70, pp. 85–91.
[10] Roondiwala M, Patel H, Varma S. Predicting Stock Prices Using Lstm. International Journal of Science and Research (IJSR), 2017, vol. 6, pp. 1754–1756.
[11] Yang B, Gong Z J, Yang W. Stock Market Index Prediction Using Deep Neural Network Ensemble. 36th Chinese Control Conference (CCC), 2017, pp. 26–28.
[12] Ashwini Pathak and Sakshi Pathak, Study of Machine learning Algorithms for Stock Market Prediction, International Journal of Engineering Research & Technology, Vol. 9 Issue 06, June-2020, 295-300.
[13] Nti, I.K., Adekoya, A.F. & Weyori, B.A. A comprehensive evaluation of ensemble learning for stock-market prediction. J Big Data 7, 20 (2020). https://doi.org/10.1186/s40537-020-00299-5
[14] Dev Shah, Haruna Isah and Farhana Zulkernine, "Stock Market Analysis: A Review and Taxonomy of Prediction Techniques", Int. J. Financial Stud. 2019, 7, 26; doi:10.3390/ijfs7020026.
[15] Rouf, N.; Malik, M.B.; Arif, T.; Sharma, S.; Singh, S.; Aich, S.; Kim, H.-C. Stock Market Prediction Using Machine Learning

Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. Electronics 2021, 10, 2717. https://doi.org/10.3390/ electronics10212717.

[16] R. Seethalakshmi, "Analysis of stock market predictor variables using Linear Regression", International Journal of Pure and Applied Mathematics, Volume 119 No. 15 2018, 369-378.

[17] Budiharto, W. Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM). J Big Data 8, 47(2021). https://doi.org/10.1186/s40537-021-00430-0.

[18] AlZoman, R.M.; Alenazi, M.J.F. A Comparative Study of Traffic Classification Techniques for Smart City Networks Sensors 2021,21,4677. https://doi.org/10.3390/s21144677