Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 2 (2023)

Analysis and Prediction for IT Sector Growth Using Machine Learning and Stochastic Modeling Approaches

M. Vijayakanth[1], V. Veeramanikandan[2]

[1]Research Scholar, Dept. of Computer and Information Science, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India

[2]Assistant Professor, Dept. of Computer Science, Thiru Kolanjiappar Govt. Arts College, Vridhachalam-606 001, Tamil Nadu, India

Abstract

Data mining involves extracting valuable insights, patterns, correlations, and trends from large datasets stored in database repositories. Machine learning and data mining approaches are used to analyze and predict various research areas with the help of statistical, mathematical, and computational modeling or techniques. The main objective is to convert raw data into actionable knowledge. In this paper considers the information technology sector data for analysis and prediction using data mining, machine learning, stochastic models, and test statistics. It is used to find future predictions based on four different parameters: open, high, low, and close using familiar machine learning approaches and stochastic model are linear regression, multilayer perceptron, M5P, random forest, random tree, REP tree and proposed stochastic model**.** Numerical illustrations are provided to prove the proposed results with test statistics or accuracy parameters.

Keywords: *Machine Learning, Classification, Decision Trees, and Stochastic Modeling*

1. Introduction and Literature Review

Data mining encompasses diverse methods like clustering, classification, regression analysis, association rule mining, anomaly detection, text mining, and time series analysis. Its significance spans various fields such as business, marketing, finance, healthcare, and scientific research. Data mining provides valuable insights by analyzing structured and unstructured data (e.g., text, images, images, and videos). However, ethical considerations, privacy, and security should be prioritized, especially when dealing with sensitive or personal information.

I suspect that students learn more from our programming assignments than from our much sweated-over lectures, with their slide transitions, clip art, and joke attempts. A great assignment is deliberate about where the student hours go, concentrating the student's attention on material that is interesting and useful. The best assignments solve a problem that is topical and entertaining, providing motivation for the whole stack of work. Unfortunately, creating great programming assignments is both time consuming and error prone. The Nifty Assignments special session is all about promoting and sharing the ideas and ready-to-use materials of successful assignments [1].

The experimental results were carried out on six datasets obtained from different disciplines, and DRFLLS proves the dataset which has a small rate of missing values gave the best estimation to the number of nearest neighbors by DRFPC and in the second degree by DRFFSM1 when r = 4, while if the dataset has high rate of missing values, then it gave the best estimation to number of nearest neighbors by DRFFSM5 and in the second degree by DRFFSM3. After that, the missing value was estimated by LLS, and the results accuracy was measured by NRMSE and Pearson correlation. The smallest value of NRMSE for a given dataset is corresponding to DRF

correlation function which is a better function for a given dataset. The highest value of PC for a given dataset is corresponding to DRF correlation function which is a better function for a given dataset [2].

Predicting the Pharmaceutical stocks along with Nifty index is possibly one of the very toughest exercises in Indian Capital Markets. The present study focuses on Short- & Long-term dynamics of the Pharmaceutical industry in Indian capital market. The Pharmaceutical sector along with Nifty Index Regular closing monitory value are a sample to the analysis between June 2015 and June 2020. In the paper, ADF test is embarked to examine immovability of data and is evident that it is un-movable at initial difference level. The co-integration test of Johansen is applied to assess long-term balance of Nifty Index analysis with the pharmaceutical sector and to define the co-integration of the variables. Granger causality test is used to regulate causal & short-term relationship of the variables with the corresponding bidirectional casualties amidst them [3].

. In this paper the NSE – Nifty Midcap50 companies among them top 4 companies having max Midcap value has been selected for analysis. The historical data has a significant role in, helping the investing people to get an overview about the market behavior during the past decade. The stock data for the past five years has been collected and trained using ARIMA model with different parameters. The test criterions like Akaike Information Criterion Bayesian Information Criterion (AICBIC) are applied to predict the accuracy of the model. The performance of the trained model is analyzed and it also tested to find the trend and the market behavior for future forecast [4].

Quinoline and its derivatives have become significant compounds because of their variety of applications in medicine, synthetic chemistry, coordination chemistry, as well as in the field of industrial chemistry. This review will summarize the different conventional methods of synthesis of quinolines and also entrenching special technique approaches such as microwave-assisted synthesis, multicomponent reactions, solvent-free reaction conditions, ionic liquids, ultrasound promoted synthesis, phase-transfer catalyst, photocatalyst, heterogeneous, homogeneous, biocatalysis, etc [5].

Data mining is a valuable tool for the practice of examining large pre-existing databases to generate previously unknown helpful information; in this paper, the input for the weather data set denotes specific days as a row, attributes denote weather conditions on the given day, and the class indicates whether the conditions are conducive to playing golf. Attributes include Outlook, Temperature, Humidity, Windy, and Boolean Play Golf class variables. All the data are considered for training purpose, and it is used in the seven-classification algorithm likes J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduce Error Pruning (REP) and Random Forest (RF) are used to measure the accuracy. Out of seven classification algorithms, the Random tree algorithm outperforms other algorithms by yielding an accuracy of 85.714% [6].

From the last twenty years, the application of Internet based technologies had brought a significant impact on the Indian stock market. Use of the Internet has eliminated the barriers of brokers and geographical location because now investors can buy and sell their shares by accessing the stock market status from anywhere at any time. Before investing money, it is very important for investors to predict the stock market. In today's digital world Internet based technologies such as Cloud Computing, Big Data analytics, and Sentiment analysis have changed the way we do business. Sentiment analysis or opinion mining makes use of text mining, natural language processing (NLP), in order to identify and extract the subjective content by analyzing user's opinion, evaluation, sentiments, attitudes and emotions. In this research work importance of sentiment analysis for stock market indicators such as Sensex and Nifty has been done to predict the price of stock. Finally, we draw conclusions and provide suggestions for future work [7].

We made an attempt to construct an optimal portfolio using Sharpe's single index model. For this study, we collected monthly closing prices of Nifty Midcap 150 scrips from July 2011 to June 2016. In this empirical study, we considered only 25 scrips out of 150 for construction of an optimal portfolio. Nifty Midcap contains different sectors scrips. Risk can be spread among the selected scrips. Risk and return was studied for individual securities. Sharpe's single index model was formulated using the excess returns to beta ratio, cut - off rate, which finally led to the construction an optimal portfolio and determined the percentage of fund investments in various scrips. In general, investors take investment decisions based on global information and market efficiency and make portfolio choices to generate better returns. This study will fundamentally help the investors to take the right investment decisions. The present study identified an optimal portfolio from the selected 25 companies, which served to maximize the returns for the investors [8].

To report the initial experience of noninvasive prenatal diagnosis of fetal Down syndrome (The NIFTY test) in a clinical setting. Methods: The NIFTY test was offered as a screening test for fetal Down syndrome to pregnant women with a singleton pregnancy at 12 weeks of gestation or beyond. A satisfaction questionnaire was sent to the first 400 patients. Results: During a 6-month period, 567 NIFTY tests were performed. Over 90% of those studied were ethnic Chinese, and the mean age of the women studied was 36 years. The test was performed at 12–13 weeks of gestation in 49.21%. The median reporting time was 9 days. The test was positive for trisomy 21 in eight cases, and for trisomy 18 in 1 case; all were confirmed by fetal karyotyping. There was no false-positive result. Of the questionnaires, 182 completed responses were received. Over 95% had complete or almost complete resolution of anxiety. Except for one, all were satisfied with the NIFTY test, and all indicated that they would recommend the test to their friends. Conclusion: The NIFTY test was a highly specific test. Unnecessary invasive tests and associated fetal losses could be avoided in almost all women who have a normal fetus [9].

2. Backgrounds and Methodologies

A data mining decision tree is a widely used machine learning technique for classification and regression tasks. It visually depicts a sequence of decisions and their possible outcomes in a tree-like structure. Each internal node represents a decision based on a specific feature, and each branch corresponds to the potential result of that decision. The tree's leaf nodes represent the final decision or the predicted outcome. The "CART" (Classification and Regression Trees) algorithm is the most used algorithm for building decision trees [10].

The authors discuss that the study proposes a new hybridization of seasonal trend decomposition methods based on Loess (STDL) and Optimal Kernel Extreme Learning Machines (OKELM) for the short- and medium-term prediction of the daily close price of the CRUDE OIL index. ELM parameter tuning is done by using the Gray Wolf Optimization Algorithm (GWO) to improve the predictive performance of the ELM further. The validation of the proposed work is done using two measures of performance, namely MASE and SMAPE [11] and [12].

2.1 Stochastic Model for Expected Share Growth Prediction (SMESGP)

A stochastic model is one in which the variables' epistemic uncertainties are considered. The delays are due to natural variations in the process being modeled. The variable is a quantity whose value changes in time series datasets. A discrete random variable is a variable whose values are obtained by counting. A continuous random variable is a variable that is used to determine whose values are obtained by measuring. A random variable is a crucial variable whose value is a numerical outcome of a number.

The proposed model using discrete random variables X, Y, and Z has a countable number of possible values denoting the primary fields at the i-th decision epoch, i = 1, 2, … n. Y is another discrete random variable using secondary domains. ‘W’ denotes the continuous random variable [13].

The Laplace transform L(.) is a simplification on a large class of functions. The inverse Laplace transform takes a part of a complex variable s and a function of a real variable time t. Laplace transform provides an alternative functional description that often simplifies the process of analysing the behaviours of the system based on a set of specifications [13]. Laplace transformation from the time domain to the frequency domain transforms differential equations into algebraic equations and convolution into multiplication. It has many applications in the sciences and technology [14].

In mathematics, convolution fk (.) is an operation on two functions of f and g to produce a third function that expresses how the shape of one is modified by the other. The term convolution refers to both the result purpose and the process. Convolution is like cross-correlation. For continuous functions, the cross-correlation operator is the adjoint of the convolution operator. Convolution has applications that include probability, statistics, computer vision, natural language processing, image and signal processing, engineering, and differential equations. A convolution is integral that expresses the overlap of one function ‘g’ as it is shifted over another function f. It, therefore, "blends" one function with another [15].

Now the probability that the threshold level is not reached till ‘t’.

S(t) = P[T > t] = P [The Total antigenic diversity due to ' k' contacts does not cross the threshold level and total due to 'k' contacts does not cross the threshold].

$$S(t) = P\left[\sum_{i=1}^{k} x_i < z_1 \cap \sum_{i=1}^{k} y_i < z_2\right]$$
$$= P\left[\sum_{i=1}^{k} x_i < z_1\right] P\left[\sum_{i=1}^{k} y_i < z_2\right]$$

= Pr [That there are k contacts in (0,t) and the total antigenic diversity does not cross the threshold and the virulence does not cross the threshold]

$$S(t) = \sum_{k=0}^{\infty} [F_k(t) - F_{k+1}(t)] \left[\int_0^{\infty} g_k(x)\overline{H(x)}\,dx\right] \left[\int_0^{\infty} q_k(y)\overline{M(y)}\,dy\right] \quad \ldots$$

(1)

where $H(x) = 1 - \overline{H(x)}$ and $M(y) = 1 - \overline{M(y)}$

It is assumed that,

$Z_1 \sim \exp(\theta)$ and $H(x) = 1 - e^{-\theta x}$

$Z_2 \sim \exp(\lambda)$ and $M(y) = 1 - e^{-\lambda y}$

$\overline{H(x)} = e^{-\theta x}$ and $\overline{M(y)} = e^{-\lambda y}$

Hence $S(t) = \sum_{k=0}^{\infty} [F_k(t) - F_{k+1}(t)] \left[\int_0^{\infty} g_k(x) e^{-\theta x}\,dx\right] \left[\int_0^{\infty} q_k(y) e^{-\lambda y}\,dy\right]$ … (2)

$$S(t) = \sum_{k=0}^{\infty} [F_k(t) - F_{k+1}(t)][g_k^*(\theta) q_k^*(\lambda)]$$
$$= 1 - [1 - g^*(\theta) q^*(\lambda)] \sum_{k=1}^{\infty} F_k(t) [g^*(\theta) q^*(\lambda)]^{k-1}$$

$L(t) = 1 - S(t)$

$$= [1 - g^*(\theta) q^*(\lambda)] \sum_{k=1}^{\infty} F_k(t) [g^*(\theta) q^*(\lambda)]^{k-1} \quad \ldots (3)$$

Taking Laplace transform of both sides we have

$$l^*(s) = [1 - g^*(\theta) q^*(\lambda)] \sum_{k=1}^{\infty} f_k^*(s) [g^*(\theta) q^*(\lambda)]^{k-1} \quad \ldots (4)$$

we assume that $f(.) \sim \exp(\eta)$ and $f^*(s) = \frac{\eta}{\eta + s}$

$g(.) \sim \exp(\beta)$ and $g^*(\theta) = \frac{\beta}{\theta + \beta}$

$q(.) \sim \exp(c)$ and $q^*(\lambda) = \frac{\alpha}{\lambda + \alpha}$

we obtain the E(T) which means to finding the expected estimations based on different problems.

$$l^*(s)=\left[1-g^*(\theta)q^*(\lambda)\sum_{k=1}^{\infty}\frac{\eta}{\eta+s}\left[g^*(\theta)q^*(\lambda)\right]\right] \quad \ldots (5)$$

$$E(T)=\left.\frac{-dl^*(s)}{ds}\right|_{s=0}$$

$$1-g^*(\theta)q^*(\lambda)=1-\frac{\beta}{\beta+\theta}.\frac{\alpha}{\alpha+\lambda}$$

$$=\frac{\beta\lambda+\theta\alpha+\theta\lambda}{\beta\alpha+\beta\lambda+\theta\alpha+\theta\lambda} \quad \ldots (6)$$

$$\left[g^*(\theta)q^*(\lambda)\right]^{k-1}=\left[\frac{\beta\alpha}{(\beta+\theta)(\alpha+\lambda)}\right] \quad \ldots (7)$$

$$E(T)=\frac{\beta\lambda+\theta\alpha+\theta\lambda}{\beta\alpha+\beta\lambda+\theta\alpha+\theta\lambda}\left[\frac{\beta\alpha}{(\beta+\theta)(\alpha+\lambda)}\right] \quad \ldots (8)$$

Symbolic representation refers to using symbols, such as letters, mathematical symbols, numbers, or other abstract forms, to represent or convey information, ideas, or concepts. This method of communication is fundamental to human language, mathematics, and many other fields. The parameter 'open' values are assigned as 'λ' symbol, the 'high' values are assigned as 'β', 'low' parameter assigned as 'θ' and 'close' named as 'α'. In this assumption is very useful for applying numerical values easily to the proposed methodology SMESGP.

2.1.1 Data Normalization

Normalization, also called one of the familiar preprocessing steps, is accustomed to carrying all values into the range [0, 1]. This is additionally called unity-based standardization. This can be summed up to limit the scope of importance in the dataset between any self-assertive point 'a' and 'b' and assign (0.1 and 0.9), respectively. Many researchers take into consideration normalization techniques for converting the range. Authors discuss agricultural data with various factors that help the growth of agriculture sectors using a stochastic approach [16] [17].

$$X'=a+\frac{(X-X_{\min})(b-a)}{X_{\max}-X_{\min}} \quad \ldots (9)$$

2.1 Linear Regression

Linear regression is a statistical technique employed to comprehend and forecast the connection between two variables by discovering the optimal straight line that most effectively aligns with the data points. It aids in ascertaining how alterations in one variable correspond to changes in another, proving valuable for predictions and trend recognition. The core idea of linear regression is to find the best-fitting straight line (also called the "regression line") through a scatterplot of data points. This line represents a linear equation of the form:

$$\mathbf{y = m_x + b} \quad \ldots (10)$$

Where y is the dependent variable (the one you want to predict or explain), x is the independent variable (the one you're using to make predictions or explanations), m is the slope of the line, representing how much, y changes for a unit change in x, b is the y-intercept, indicating the value of y when x is 0.

2.2 Multilayer Perception

A Multilayer Perceptron (MLP) is an artificial neural network consisting of multiple layers of interconnected nodes or neurons. It's a fundamental architecture in deep learning and is used for various tasks, including classification, regression, and more complex tasks like image recognition and natural language processing. The architecture of an MLP typically includes three types of layers:

i. **Input Layer:** This layer consists of neurons receiving input data. Each neuron corresponds to a feature in the input data, and the values of these neurons pass through the network.

ii. **Hidden Layers:** These layers come after the input layer and precede the output layer. They are called "hidden" because their activations are not directly observed in the final output.
iii. **Output Layer:** This layer produces the network's final output. The number of neurons in the output layer depends on the problem type.

2.3 M5P

M5P is a machine learning algorithm used for regression tasks. It is an extension of the decision tree-based model called M5, which Ross Quinlan developed. The M5 algorithm combines decision trees and linear regression to create more accurate and flexible regression models. M5P, specifically, stands for M5 Prime.

Steps involved in the M5P

Step 1. Building the initial decision tree (M5 model): Recursive Binary Splitting and Pruning (optional)
Step 2. Linear Regression Model: Leaf Regression Models and Model Parameters
Step 3. Piecewise Linear Regression: Piecewise Prediction
Step 4. Model Evaluation: Training and Testing.

2.4 Random Forest

Random Forest is a popular machine learning ensemble method for classification and regression tasks. It is an extension of decision trees and is known for its high accuracy, robustness, and ability to handle complex datasets. Random Forest is widely used in various domains, including data science, machine learning, and pattern recognition. The main idea behind Random Forest is to create an ensemble (a collection) of decision trees and combine their predictions to make more accurate and stable predictions. Random Forest is an ensemble learning method combining multiple decision trees to make more accurate and robust predictions for classification and regression tasks. The steps involved in building a Random Forest are as follows:

Step 1. Data Bootstrapping
Step 2. Random Feature Subset Selection
Step 3. Decision Tree Construction
Step 4. Ensemble of Decision Trees
Step 5. Out-of-Bag (OOB) Evaluation
Step 6. Hyperparameter Tuning (optional)

2.5 Random Tree

In machine learning, a Random Tree is a specific type of decision tree variant that introduces randomness during construction. Random Trees are similar to traditional decision trees but differ in how they select the splitting features and thresholds at each node. The primary goal of introducing randomness is to create a more diverse set of decision trees, which can help reduce overfitting and improve the model's generalization performance. Random Trees are commonly used as building blocks in ensemble methods like Random Forests. The critical characteristics of Random Trees are Random Feature Subset, Random Threshold Selection, No Pruning and Ensemble Methods

Steps involved in Random Tree

Step 1. Data Bootstrapping:
Step 2. Random Subset Selection for Features:
Step 3. Decision Tree Construction:
Step 4. Voting (Classification) or Averaging (Regression):

2.6 REP Tree

REP (Repeated Incremental Pruning to Produce Error Reduction) Tree is a machine learning algorithm for classification and regression tasks. A decision tree-based algorithm constructs a decision tree using incremental pruning and error-reduction techniques. It is an extension of decision trees that incorporates pruning to reduce overfitting and improve the model's generalization performance. Below are the steps involved in building a REP Tree.

Step 1. Recursive Binary Splitting
Step 2. Pruning
Step 3. Repeated Pruning and Error Reduction
Step 4. Model Evaluation

2.7 Accuracy Metrics

The predictive model's error rate can be evaluated by applying several accuracy metrics in machine learning and statistics. The basic concept of accuracy evaluation in regression analysis is comparing the original target with the predicted one and using metrics like R-squared, MAE, MSE, and RMSE to explain the errors and predictive ability of the model [18]. The R-squared, MSE, MAE, and RMSE are metrics used to evaluate the prediction error rates and model performance in analysis and predictions [19] and [20].

R-squared (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The values from 0 to 1 are interpreted as percentages. The higher the value is, the better the model is.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad \ldots (11)$$

MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaging the absolute difference over the data set.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad \ldots (12)$$

RMSE (Root Mean Squared Error) is the error rate by the square root of MSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad \ldots (13)$$

Relative Absolute Error (RAE) is a metric used in statistics and data analysis to measure the accuracy of a forecasting or predictive model's predictions. It is particularly useful when dealing with numerical data, such as in regression analysis or time series forecasting.

$$RAE = \frac{\sum|y_i - \hat{y}_i|}{\sum|y_i - \bar{y}|} \quad \ldots (14)$$

Root Relative Squared Error (RRSE) is another metric used in statistics and data analysis to evaluate the accuracy of predictive models, especially in the context of regression analysis or time series forecasting.

$$RRSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}} \quad \ldots (15)$$

Equation 11 to 15 used to find the model accuracy which is used to find the model performance and error. Where Y_i represents the individual observed (actual) values, $\hat{Y}_i$ represents the corresponding individual predicted values, $\bar{Y}$ represents the mean (average) of the observed values and Σ represents the summation symbol, indicating that you should sum the absolute differences for all data points.

Numerical Illustrations

The corresponding dataset was collected from the open-source Kaggle data repository. The IT dataset includes 5 parameters and the daily instance between January 1996 and August 2022 with different data categories: date, open, high, low, and close [21]. The secondary dataset namely Nifty 50 data which is includes date of observation, open price of the index on a particular day, high price of the index on a particular day, high price of the index on a particular day and close price of the index on a particular day. A detailed description of the parameters is mentioned in the following Table 1.

Table 1. IT sector sample dataset

Date	Open	High	Low	Close
12-Aug-22	30184.95	30195	29819.15	29885.9
11-Aug-22	30164.2	30388.9	30073.95	30233.7
10-Aug-22	30013.45	30013.45	29456.8	29701.15
8-Aug-22	29999.15	30036.75	29694.15	29967.7
5-Aug-22	29891.85	30080.2	29799.2	29973.55

4-Aug-22	29694.05	29972.35	29278.15	29782.9
3-Aug-22	29002.3	29453.35	28929.9	29416.95
2-Aug-22	29131.1	29153.05	28871.2	29024.65
1-Aug-22	29376.65	29376.65	28928.95	29220
29-Jul-22	28983.25	29355	28909.15	29152.3
28-Jul-22	28225.1	28703.4	28170.45	28663.05
27-Jul-22	27420.65	27908.25	27339.95	27878.8
26-Jul-22	28157.5	28157.5	27383.95	27418.85
25-Jul-22	28006.65	28320.85	27890.65	28216
22-Jul-22	28429.55	28549.25	27862.05	28168.15
21-Jul-22	28116.55	28369.95	27848.75	28342.9
20-Jul-22	27701.05	28175.5	27631.4	28146.35
19-Jul-22	27071.8	27399.5	27065.8	27346.1
18-Jul-22	26801.75	27376.2	26785.8	27322.95

Table 2: Machine Learning Models with R-squared

ML Approaches	Open	High	Low	Close
Linear Regression	0.9999	1.0000	1.0000	0.6318
Multilayer Perceptron	0.9999	0.9999	0.9999	0.6235
M5P	0.9999	1.0000	1.0000	0.6526
Random Forest	0.9996	0.9996	0.9996	0.6318
Random Tree	0.9998	0.9998	0.9998	0.6319
REP Tree	0.9996	0.9996	0.9996	0.6322
SMESGP	0.9999	1.0000	1.0000	0.7048

Table 3: Machine Learning Models with Mean Absolute Error (MAE)

ML Approaches	Open	High	Low	Close
Linear Regression	44.9885	38.9000	39.5621	3940.2086
Multilayer Perceptron	56.4962	54.4408	51.0757	4175.1262
M5P	56.4085	40.3445	41.4212	3287.9042
Random Tree	172.3620	172.4273	172.1771	4637.6290
Random Forest	133.1774	124.9607	124.6882	4635.9466
REP Tree	172.8076	179.8567	171.0901	4632.4136
SMESGP	38.8457	33.4751	34.1654	3154.2451

Table 4: Machine Learning Models with Root Mean Squared Error (RMSE)

ML Approaches	Open	High	Low	Close
Linear Regression	86.1156	72.4948	74.4105	8259.7400
Multilayer Perceptron	94.2393	85.8482	84.0979	8339.5901
M5P	104.0219	69.4361	75.1667	8073.8867
Random Tree	230.6526	221.1492	221.594	8481.3391
Random Forest	177.0957	161.5143	160.3190	8479.6440
REP Tree	225.0978	228.7301	218.3946	8481.0855
SMESGP	81.2132	67.9854	68.2122	7521.5451

Table 5: Machine Learning Models with Relative Absolute Error (RAE)

ML Approaches	Open	High	Low	Close
Linear Regression	0.7581	0.6499	0.6733	52.2130
Multilayer Perceptron	0.9520	0.9096	0.8693	55.3260
M5P	0.9505	0.6741	0.7050	43.5691
Random Tree	2.9045	2.8809	2.9305	61.4547
Random Forest	2.2442	2.0878	2.1222	61.4324
REP Tree	2.9120	3.0050	2.9120	61.3856
SMESGP	0.6912	0.5874	0.6185	38.4235

Table 6: Machine Learning Models with Root Relative Squared Error (RRSE)

ML Approaches	Open	High	Low	Close
Linear Regression	1.0791	0.9008	0.9421	77.5091
Multilayer Perceptron	1.1809	1.0667	1.0648	78.2584
M5P	1.3035	0.8628	0.9517	75.7650
Random Tree	2.8903	2.7479	2.8057	79.5886
Random Forest	2.2192	2.0069	2.0299	79.5727
REP Tree	2.8207	2.8421	2.7652	79.5862
SMESGP	0.7854	0.4214	0.5487	34.2356

Table 7: Machine Learning Models with Time Taken to Build Model (Seconds)

ML Approaches	Open	High	Low	Close
Linear Regression	1.9100	0.0900	0.0400	0.1700
Multilayer Perceptron	5.9500	6.0600	5.5100	8.0200
M5P	1.2300	0.2500	0.1000	0.2400
Random Tree	0.0100	0.0800	0.0300	3.7700
Random Forest	3.4600	1.8000	2.9800	10.4600
REP Tree	0.1400	0.0500	0.0200	1.3200
SMESGP	0.0100	0.0700	0.0300	0.0800

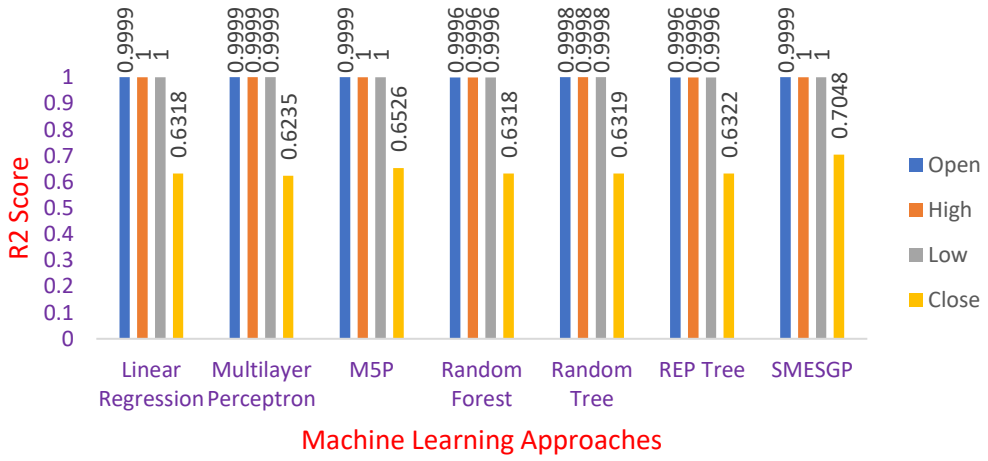

Fig. 1. R2 Score for Machine Learning Approaches

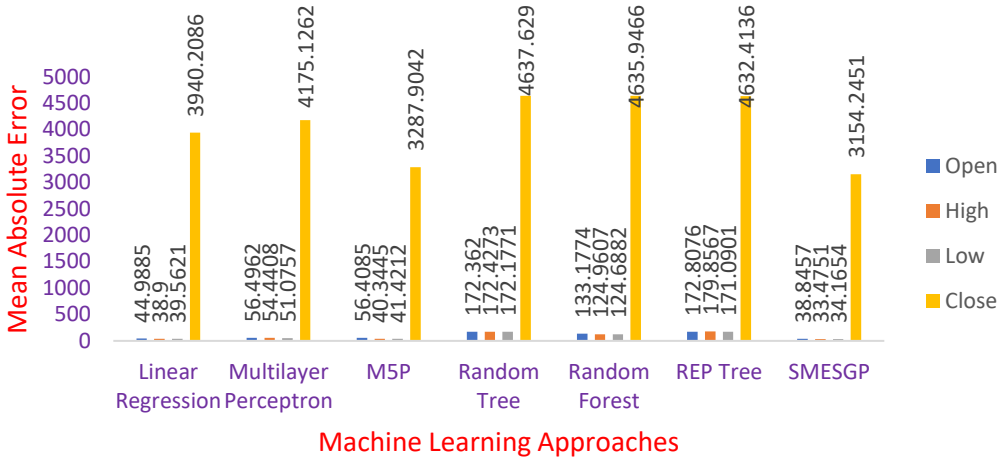

Fig. 2. Machine Learning Models with MAE

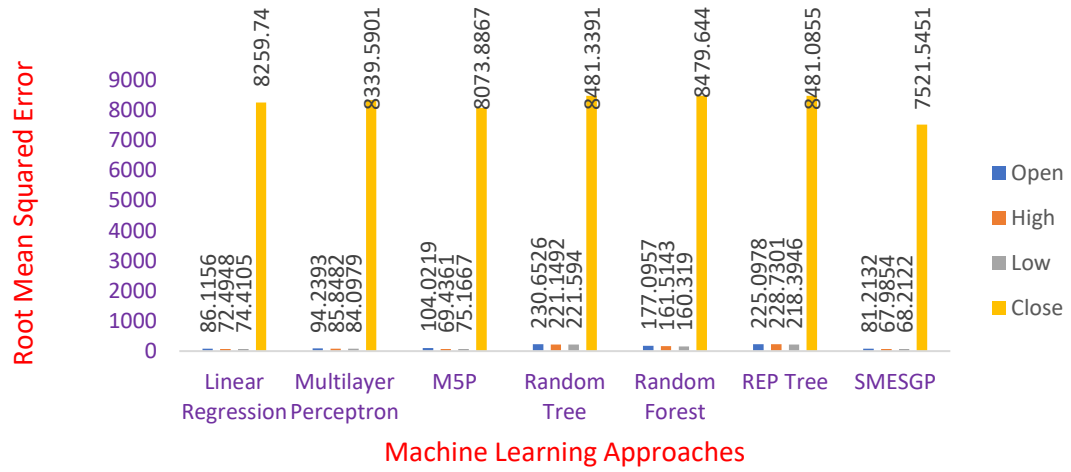

Fig. 3. Machine Learning Models with RMSE

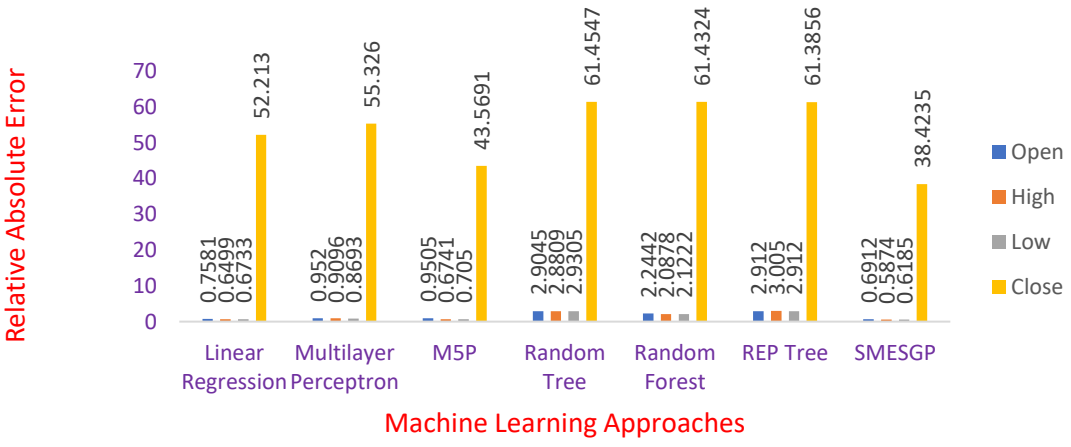

Fig. 4. Machine Learning Models with RAE

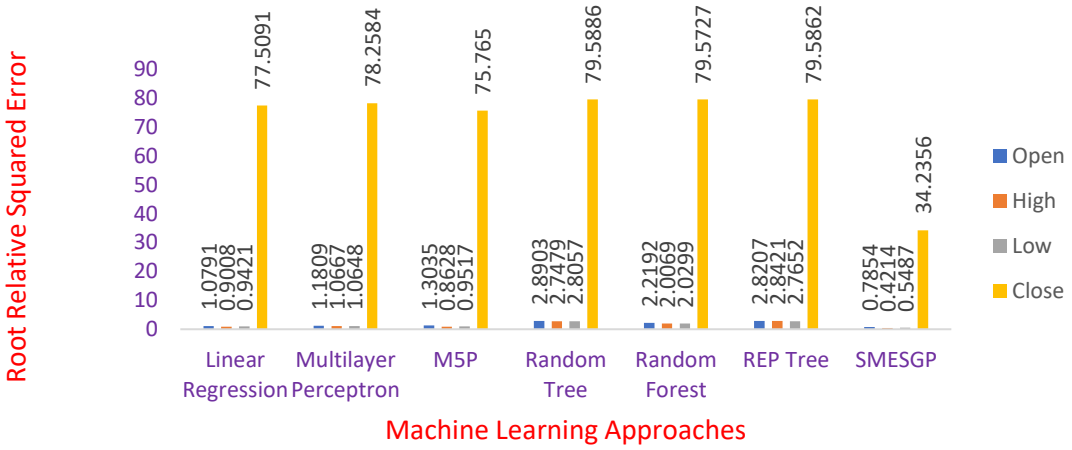

Fig. 5. Machine Learning Models with RRSE

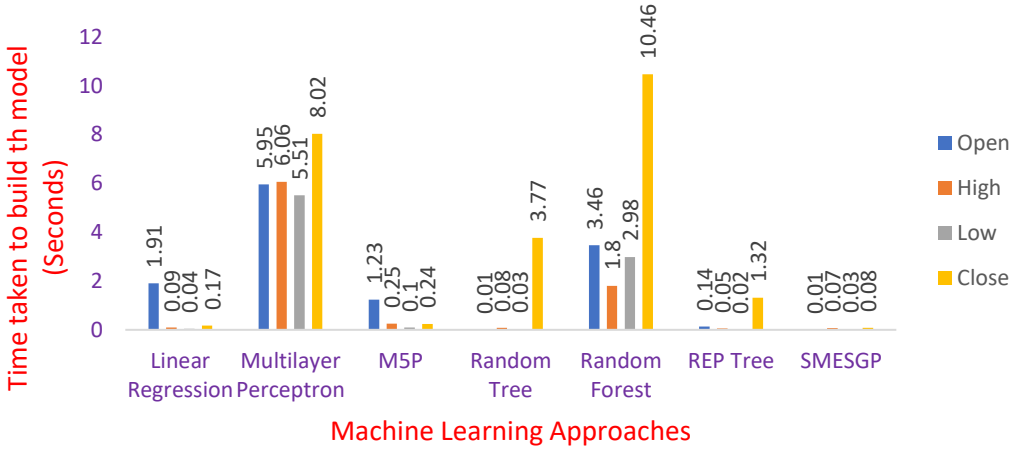

Fig. 6. Machine Learning Models and its Time Taken to Build the Model (Seconds)

3. Result and Discussion

Table 1 shows the sample dataset, which includes five parameters among five, only four having continuous data, namely open, low, high, and close, which is suitable for analysis and prediction purposes. The dataset indicates the average values of various IT sector shares called NIFTY 50. Based on the dataset, it is evident that different machine learning classifications, decision trees, and proposed stochastic modeling approaches are used to find the hidden patterns and which is the best or influencing parameter to decide future predictions. Related results and numerical illustrations with accuracy performance are shown in Table 1 to Table 7 and Figure 1 to Figure 6.

They are based on Equation 11, Table 2, and Figure 1, which is used to find the R2 score or correlation coefficient by comparing 4 parameters. Numerical illustrations suggest that there may be a significant difference from one parameter to another. In this case, using six different decision tree approaches and the proposed stochastic model (SMESGP) used among these results, Linear Regression, Multilayer Perceptron, M5P, Random Forest, Random Tree, and REP Tree return a robust, strong positive correlation of nearly 0.9 when using open, high, low and close parameters. Comparatively, the proposed SMESGP approach reaches a maximum R2 score of 0.9999.

Further data analysis revealed a gradual improvement in test scores over time. The MAE is used to find model errors using Equations 12. Machine learning algorithms and the proposed model SMESGP will be used in

this case. The proposed model SMESGP returns a minimum error. Random tree and REP tree approaches return maximum error. The related numerical illustration is shown in Table 3 and Figure 2.

The RMSE (root mean square error) measures the difference between predicted and actual values using Equation 13. In this case, the proposed SMESGP returns a minimum error, and Random trees produce the maximum error for using RMSE test statistics. The related numerical illustration is shown in Table 4 and Figure 3.

Relative Absolute Error (RAE) measures accuracy using Equation 14 to compare the difference between predicted and actual values in percentage. In this research, ML and proposed stochastic model approaches. The proposed SMESGP approach returns a minimum error. The related numerical illustrations are shown in Table 5 and Figure 4. Similar error approaches are reflected in RRSE. Similar numerical examples are shown in Table 6 and Figure 5.

Time taken is one of the significant tasks in machine-learning approaches. Based on Table 7 and Figure 6, a proposed stochastic model, namely SMESGP, takes minimum time to build the model. Subsequently, the random and REP trees take the next level of minimum time to build the models. Finally, multilayer perceptron takes the maximum time to build the model.

4. Conclusion and Further Research

It is essential to consider the limitations of this study. The findings presented in this study contribute to our understanding that all the parameters return robust positive correlations with minimum error except close. In this research, the proposed model, compared to the maximum number of machine learning approaches, returns a minimum error with less processing time. Finally, based on ML and Stochastic approaches, open, high, and low are considered for better future prediction. The close parameter returns the minimum R2 score and the maximum error. Finally, the close parameter is not suitable for future forecasts. In this research, clearly explain the proposed model SMESGP is the best model for analysis and prediction. Further, using different machine learning and model development, finding the suitable variable for future prediction with higher accuracy performance in the future. The proposed model is not only applicable in the field of share market analysis, but it will be adjected with symbolic representation and then applicable to other research areas like medical diagnosis, fraud detection, agriculture development, and other applicable areas.

5. Reference

[1] Parlante, N., Zelenski, J., Feinberg, D., Mishra, K., Hug, J., Wayne, K., Guerzhoy, M., Cheung, J.C.K. and Pitt, F., 2017, March. Nifty assignments. In Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education (pp. 695-696).

[2] Al-Janabi, S. and Alkaim, A.F., 2020. A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation. Soft Computing, 24(1), pp.555-569.

[3] Basha, M., Singh, A.P., Rafi, M., Rani, M.I. and Sharma, N.M., 2020. Cointegration and Causal relationship between pharmaceutical sector and Nifty–An empirical Study. PalArch's Journal of Archaeology of Egypt/Egyptology, 17(6), pp.8835-8842.

[4] Devi, B.U., Sundar, D. and Alli, P., 2013. An effective time series analysis for stock trend prediction using ARIMA model for nifty midcap-50. International Journal of Data Mining & Knowledge Mansagement Process, 3(1), p.65.

[5] Matada, B.S. and Yernale, N.G., 2021. Modern encroachment in synthetic approaches to access nifty quinoline heterocycles. Journal of the Indian Chemical Society, 98(11), p.100174.

[6] Rajesh, P. and Karthikeyan, M., 2017. A comparative study of data mining algorithms for decision tree approaches using the Weka tool. Advances in Natural and Applied Sciences, 11(9), pp.230-243.

[7] Bhardwaj, A., Narayan, Y. and Dutta, M., 2015. Sentiment analysis for Indian stock market prediction using Sensex and Nifty. Procedia computer science, 70, pp.85-91.

[8] Basha, S.M. and Ramaratnam, M.S., 2017. Construction of an Optimal Portfolio Using Sharpe's Single Index Model: A Study on Nifty Midcap 150 Scrips. Indian Journal of Research in Capital Markets, 4(4), pp.25-41.

[9] Lau, T.K., Chan, M.K., Salome Lo, P.S., Connie Chan, H.Y., Kim Chan, W.S., Koo, T.Y., Ng, H.Y.J. and Pooh, R.K., 2012. Clinical utility of noninvasive fetal trisomy (NIFTY) test–early experience. The Journal of Maternal-Fetal & Neonatal Medicine, 25(10), pp.1856-1859.

[10] Kohavi, R., & Sahami, M. (1996). Error-based pruning of decision trees. In International Conference on Machine Learning (pp. 278-286).

[11] Veeramanikandan, V., and Jeyakarthic. M., (2020). Hybridization Of Stdl with Optimal Kernel Extreme Learning Machine (Okelm) Based Short Term Crude Oil Price Forecasting In Commodity Futures Market. International Journal of Scientific & Technology Research, 9(2), 4029- 4036.

[12] Jeyakarthic. M., and Veeramanikandan, V., (2020). Forecasting of commodity future index a hybrid regression model based on support vector machine and grey wolf optimization algorithm. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(2), 2856-2862.s

[13] Korn, G. A. and T. M. Korn, Mathematical Handbook for Scientists and Engineers (2nd ed.), McGraw-Hill Companies, 1967.

[14] https://mathworld.wolfram.com/LaplaceTransform.html

[15] https://mathworld.wolfram.com/Convolution.html

[16] Rajesh, P. and M. Karthikeyan, Data Assimilation of Gross Domestic Product (GDP) in India using Stochastic Data Mining Approach. Journal of Computational and Theoretical Nanoscience, 2019, 16(4), pp. 1478–1484.

[17] Rajesh, P. and M. Karthikeyan, Data Mining Approaches to Predict the Factors that Affect Agriculture Growth using Stochastic Model. International Journal of Computer Sciences and Engineering, 2019, 7(4), pp.18-23.

[18] Akusok, A. (2020). What is Mean Absolute Error (MAE)? Retrieved from https://machinelearningmastery.com/mean-absolute-error-mae-for-machine-learning/

[19] S. M. Hosseini, S. M. Hosseini, and M. R. Mehrabian, “Root mean square error (RMSE): A comprehensive review,” International Journal of Applied Mathematics and Statistics, vol. 59, no. 1, pp. 42–49, 2019.

[20] Chi, W. (2020). Relative Absolute Error (RAE) – Definition and Examples. Medium. https://medium.com/@wchi/relative-absolute-error-rae-definition-and-examples-e37a24c1b566

[21] https://www.kaggle.com/datasets/debashis74017/stock-market-index-data-india-1990-2022