Journal of Advanced Zoology
ISSN: 0253-7214
Volume 45 Issue S-4 Year 2024 Page 118-127

Machine Learning Techniques For Stock Price Prediction - A Comparative Analysis Of Linear Regression, Random Forest, And Support Vector Regression

Shital Sameer Pashankar[1*], Jyoti Dashrath Shendage[2], Dr. Janardan Pawar[3]

[1]Computer Science, Indira College of Commerce and Science, Pune, India.*
Email: srbhigwankar@gmail.com
[2]Computer Science, Indira College of Commerce and Science, Pune, India.
Email: shendagejyoti@gmail.com
[3]Computer Application, Indira College of Commerce and Science, Pune, India.
Email: janardanp@iccs.ac.in

***Corresponding Author:** Shital Sameer Pashankar*
Email: srbhigwankar@gmail.com

Abstract

In the dynamic and rapidly evolving stock market, the ability to generate accurate and timely predictions holds paramount significance for companies and investors alike. Machine learning algorithms can identify complex patterns in the stock market. Machine learning algorithms play a critical role in this situation, leveraging their capacity to analyze extensive datasets and provide valuable insights, thereby forecasting future trends. Stock price prediction is still an arduous task because of the financial markets' well-known volatility. In recent years, there has been a significant increase in the use of machine learning techniques for stock price prediction. This is because these algorithms can handle large amounts of data and identify complex patterns that are difficult for humans to recognize. The proposed methodology focuses on using linear regression (LR), support vector regression (SVR), and random forest machine learning models to predict Tata Consultancy Services (TCS) stock prices. Machine learning presents a promising method in this field of stock price prediction, which is essential for traders and investors to make well-informed judgements. via data on TCS stock prices, the study evaluates the models via feature engineering and hyperparameter tuning. An understanding of how well these algorithms anticipate stock values is given by the analysis and findings. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) standard metrics are used to compare the time series data. After Applying hyperparameter tuning on Support Vector Regression and Random Forest the standard metrics RMSE shows decrease in error rate. In the proposed model Linear Regression has better performance than Support Vector Regression and Random Forest.

Keywords: Stock market, Linear Regression, Random Forest, Support Vector Regression (SVR), Prediction.

INTRODUCTION

The linear regression model demonstrates the lowest RMSE, indicating better predictive performance compared to Random Forest and SVR. Random Forest follows closely, while SVR appears to struggle in capturing the complexity of the TCS stock price data.

In the contemporary landscape characterized by technological advancements and enhanced computing capabilities, machine learning has become pervasive across various industries. The realm of stock price prediction has particularly witnessed a surge in the application of machine learning techniques. Basically, the stock market is one where there are buyers and sellers interested in buying stocks of a certain company, and the prices of these stocks vary widely over time [1]. Traditional methods of stock price prediction, like technical and fundamental analysis, rely on historical trends and human judgment. Although effective to a certain extent, these methods exhibit limitations. For instance, technical analysis predominantly concentrates on price patterns and trends in the stock market. In contrast, machine learning algorithms encompass a broader spectrum of factors, concurrently analyzing historical data, real-time market data, and economic indicators. This comprehensive approach enables machine learning algorithms to identify intricate patterns, setting them apart from traditional methods. An advantage of machine learning in stock price prediction is in its ability to continuously learn and adapt.

The stock market is characterized as dynamic, unpredictable and non-linear in nature [6]. As the market undergoes constant evolution, traditional methods may become outdated. Machine learning algorithms can assimilate new data continuously, adjusting their predictions in real-time. This adaptability contributes to their increasing accuracy and reliability over time. The surge in the adoption of machine learning techniques for stock price prediction is attributed to their capability to handle vast datasets and discern complex patterns that may elude human recognition. Machine Learning algorithms expert in adapting to changing market conditions, providing investors with a competitive edge in the volatile stock market environment. In recent years, there has been a significant increase in the use of machine learning techniques for stock price prediction. This is because these algorithms can handle large amounts of data and identify complex patterns that are difficult for humans to recognize. They can also adapt to changing market conditions and update their predictions in real-time, giving investors an edge in the volatile stock market.

The Linear regression algorithm predicts the relationship between an independent variable and dependent variables. Although straightforward to comprehend, linear regression assumes a linear relationship between variables, which may not always hold in the intricate dynamics of the stock market.

The random forest machine learning algorithm enrol an ensemble of decision trees and also approaches the challenge of overfitting by randomly selecting subsets of data to build multiple decision trees. The combination of these results enhances prediction accuracy for stock price prediction.

Additionally, its ability to handle large datasets with numerous features, without requiring extensive pre-processing and accommodating missing values, renders the random forest algorithm robust and efficient. Support vector regression (SVR) is a important machine learning technique, widely accepted for recognizing pattern of time series dataset. Operating by identifying the best hyperplane that distinctly separates training data into different classes, SVR excels in handling problems characterized by a non-linear relationship between variables. This makes it a suitable choice for predicting stock prices that deviate from a linear pattern. Renowned for its high accuracy, SVR proves versatile in catering to both linear and non-linear stock price prediction scenarios.

LITERATURE SURVEY

In recent years, the financial markets have witnessed a surge in the adoption of advanced technologies, with machine learning (ML) emerging as a powerful tool for analyzing and predicting stock prices. As traditional financial models face challenges in capturing the complexities of dynamic market conditions, machine learning techniques offer a promising avenue to enhance the accuracy and efficiency of stock price predictions. This literature review gives an overview of the current state of research in stock price prediction using machine leaning, highlighting key methodologies, challenges and future directions.

Malti Bansal et al. [1] developed model using machine learning algorithms K-Nearest Neighbour, Linear regression, Support Vector Regression, Decision Tree Regression and Short-Term Memory for predicting stock prices. Dataset of 12 companies over the last 7 years was collected and used for prediction of stock prices. The models were tested on three essential performance metrics, Symmetric Mean Absolute Percentage Error (SMAPE), R-Squared value (R2), and Root Mean Square Error (RMSE). Out of five algorithms the long Short-

term Memory algorithm shows the best predictive performance as it has the least value of error i.e. SMAPE (1.59), R2 (-0.11), and RMSE (22.55).

Md. Mobin Akhtar et al. [2] proposed a model using SVM and Random Forest classifier for prediction of stock prices. Experimental results show SVM model 78.7% accuracy and random forest 80.8% accuracy for predicting stock prices.

Dr. Poorna Shankar et al. [3] designed a model using machine learning algorithms as Artificial intelligence multi-layer perceptron's, accumulative neural networks, Naïve Bayes, back-propagation networks, single-layer LSTMs, vector bearers, cyclic neural networks. Historical dataset of Tata Motors India Ltd extracted from NSE India. In the proposed work LSTM shows less error percentage than other algorithms.

Abdulhamit Subasi et al. [4] designed model using four datasets (NASDAQ, NYSE, NIKKEI, and FTSE) and 7 classifiers (Random Forest, Bagging, AdaBoost, Decision Trees, SVM, K-NN, and ANN) have been used for stock market price prediction. The experimental result shows Random Forest with leaked Bagging with leaked data shows highest prediction percent i.e. 93%.

Wasiat khan et al. [5] developed model using deep leaning for prediction of stock market predicted using machine learning algorithms on information contained in social media and financial news. experimental results show that highest prediction accuracies of 80.53% and 75.16% are achieved using social media and financial news, respectively.

Mehar Vij, Deeksha et al. [6] designed a model using ANN and Random Forest Machine Learning models for predicting stock Market. The experimental result shows comparative analysis based on RMSE, MAPE and MBE values clearly indicate that ANN gives better prediction of stock prices as compared to RF. Results show that the best values obtained by ANN model gives RMSE (0.42), MAPE (0.77) and MBE (0.013).

Srinath Ravikumar et al. [7] The proposed system works in two methods Regression and Classification. In regression, the system predicts the closing price of stock of a company, and in classification, the system predicts whether the closing price of stock will increase or decrease the next day. The machine learning algorithms used for experiment was Support Vector Machine (linear), Support Vector Machine (poly), Support Vector Machine (rbf), Support Vector Machine (sigmoid), K – Nearest Neighbours , Logistic Regression, Naïve Bayes , Decision Tree Classification, Random Forest Classification. Dataset used for experiment was from Yahoo Finance. The Logistic Regression Model gave maximum mean accuracy of 68.622%.

M Umer Ghani et al.[8] proposed a model using (Linear, Logistic), K-Nearest Neighbour (KNN), Decision Tree (DT), Artificial Neural Network (ANN) and Simple Moving Average (SMA) with the help of Time Series Forecasting (TSF), Three month Moving Average(3MMA), Exponential Smoothing (ES) for prediction of stock price. Dataset used for experiment was obtained from Yahoo Finance GOOGLE, FB,AMAZON,AAPLE. Exponential Smoothing predictions results are best rather than Linear Regression(LR) and Three Months Moving Average(3MMA).

Pawee Werawithayaset et al. [9] Proposed a model to predict the closing price of the stock Exchange of Thailand (SET).Machine learning algorithms used are Multi-Layer Perceptron, Sequential minimal optimization algorithm (SMO) and Partial Least Square(PLS) Classifier to predict the closing price of the stock. Dataset used for experiment was based on 100 stock data from SET 100. To compare algorithms the error values of Mean Absolute Error(MAE) and Root Mean Square Error (RMSE) are used. The experimental result shows that Partial Least Square is the Best algorithm of the three algorithm to predict the stock closing price.

Nagaraj Naik et al. [10] designed a model by considering 33 different combinations of technical indicators to predict the stock market price. Or this they used Boruta feature selection technique to predict stock. For this Artificial Neural Network and regression prediction model is used model performance is evaluated using metrics is Mean absolute error (MAE) and Root mean square error. The experimental results are better than the existing method by decreasing the error rate in the reduction to 12%. Dataset used National Stock Exchange, India (NSE) for experiment.

Ashish Sharma et al. [11] developed a regression analysis model for prediction of stock market. In this Polynomial regression, RBF regression, Sigmoid regression, Linear regression are used to help the stock brokers and investors for investing money in the stock market as stock prediction is dynamic in nature.

METHODOLOGY

The methodology section covered data pre-processing, feature engineering, model selection, training, and evaluation. The time series analysis involved extracting relevant temporal features and creating lag variables to capture the sequential nature of stock prices. Three machine learning models—Linear Regression, Random Forest, and SVM—were chosen for their distinct characteristics. The dataset was split into training and testing sets, and standard scaling was applied to the features

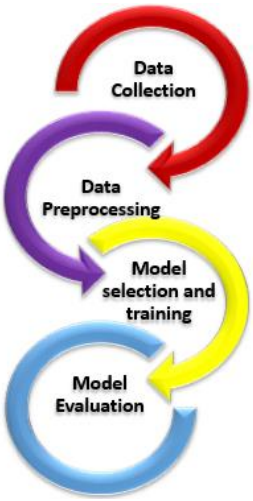

Figure 1 showing methodology used for research.

DATA COLLECTION

The collected raw data sometimes has impurities or noise. So, data pre-processing method is used to convert raw data into clean data [10]. The historical data has been extracted from yahoo finance. The data set includes 18 years of data from The dataset includes 18 years of data from Tata Consultancy Services (TCS). The data contains information about stocks such as Date, Open, High, Low, Close, Adj. Close and Volume [3].

Table 1 TCS Data representation from Jan 2005 to Jan 2024

	Date	Open	High	Low	Close	Adj Close	Volume
0	1/3/2005	166.932495	169.587494	166.932495	169.157501	119.614983	4734168.0
1	1/4/2005	168.750000	169.250000	166.645004	167.657501	118.554291	4979104.0
2	1/5/2005	167.524994	167.524994	160.037506	165.625000	117.117065	9604320.0
3	1/6/2005	165.000000	168.250000	161.562500	165.057495	116.715744	8006888.0
4	1/7/2005	166.250000	168.125000	164.082504	167.494995	118.439430	7623136.0
...	...	...	...	...	...	...	...
4712	1/24/2024	3880.000000	3883.649902	3805.600098	3841.800049	3841.800049	2657709.0
4713	1/25/2024	3839.899902	3861.000000	3778.699951	3810.300049	3810.300049	2205154.0
4714	1/29/2024	3800.800049	3820.000000	3780.850098	3801.000000	3801.000000	1646747.0
4715	1/30/2024	3807.449951	3847.800049	3787.000000	3800.550049	3800.550049	1411621.0
4716	1/31/2024	3811.199951	3834.000000	3796.449951	3815.949951	3815.949951	2459358.0

4717 rows × 7 columns

In this research we considered data from January 2005 to 31 January 2024 then we evaluated our approach with different prediction methodologies.

PRE-PROCESSING

'When we have collected the data, we must check that the data collected have missing data or not have missing data [1].

Table 2 TCS Data after data preprocessing

Date	Open	High	Low	Close	Adj Close	Volume	Year	Month	Day	Weekday	Dayofyear
2005-01-03	166.932495	169.587494	166.932495	169.157501	119.614983	4734168.0	2005	1	3	0	3
2005-01-04	168.750000	169.250000	166.645004	167.657501	118.554291	4979104.0	2005	1	4	1	4
2005-01-05	167.524994	167.524994	160.037506	165.625000	117.117065	9604320.0	2005	1	5	2	5
2005-01-06	165.000000	168.250000	161.562500	165.057495	116.715744	8006888.0	2005	1	6	3	6
2005-01-07	166.250000	168.125000	164.082504	167.494995	118.439430	7623136.0	2005	1	7	4	7

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a vital procedure involving the initial examination of data through the exploration of patterns, the verification of assumptions, and the utilization of summary statistics and graphical representations. EDA proves valuable in identifying outliers, discerning patterns, and uncovering trends within the provided dataset. It plays a pivotal role in the discovery of meaningful patterns inherent in the data.

Figure 2 Showing Open and Close stock price

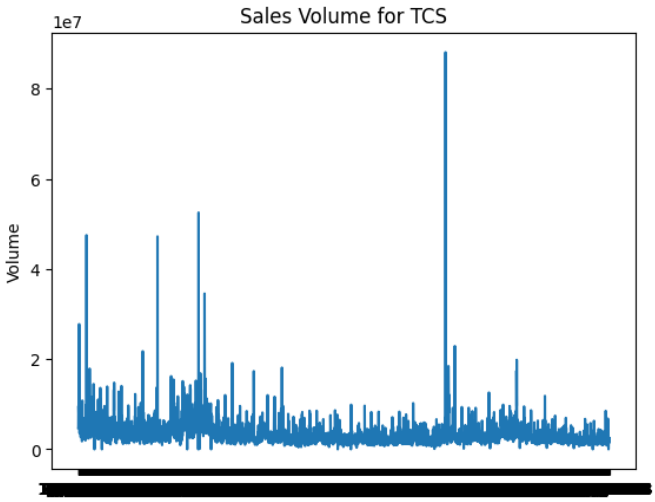

Figure 3 Sales Volume for TCS

MODEL SELECTION AND TRAINING

The following are Machine Learning algorithms used for prediction of stock market price prediction in the proposed methodology:

1) Linear Regression

Linear regression is a fundamental and widely used technique in machine learning and statistics for predicting a continuous outcome variable based on one or more predictor variables. The goal of linear regression is to find the best-fitting linear relationship (line) that minimizes the difference between the predicted and actual values of the outcome variable. The general form of a simple linear regression equation with one predictor variable is:
y=mx+b
Where
y is the dependent variable (outcome),
x is the independent variable (predictor),
m is the slope of the line,
b is the y-intercept.

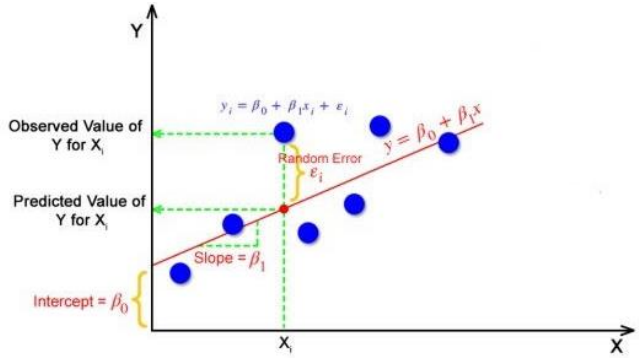

Figure 4 Linear Regression

2) Random Forest

To address the weaknesses of decision trees random forest can be used which combines the power of multiple decision trees into one. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and improve the model's performance.

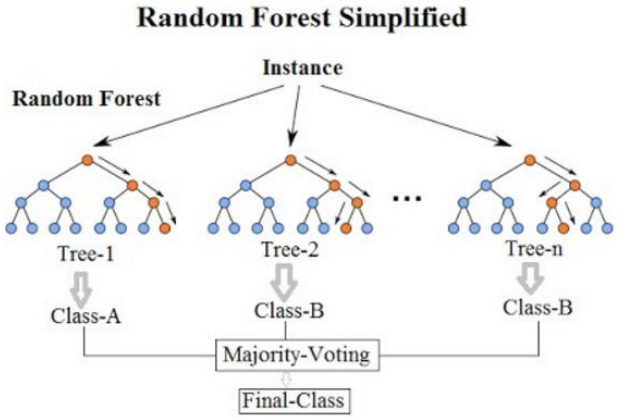

Figure 5 Random Forest

Ensemble Learning

An ensemble method is a technique that combines predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. A model comprised of many models is called an ensemble model.

3) Support Vector Regression

Support vector Regression (SVR) is a type of regression algorithm that is an extension of Support Vector Machines (SVM) for regression tasks. SVR is particularly useful when dealing with non-linear relationships between the input features and the target variable. It aims to find a hyperplane in a high-dimensional space that best represents the relationship between the input features and the output variable. The regression problem is a generalization of the classification problem, in which the model returns a continuous valued output.

Terminology

i) Hyperplane: In SVM this is basically the separation line between the data classes. In SVR, it is defined as the line that will help us predict the continuous value or target value.

ii) Boundary line: These are two parallel lines drawn to the either side of support vector with the error threshold value, (epsilon) are known as the boundary line. These lines create a margin between the data points.

iii) Support Vectors: These are the data points which are closest to the boundary. The data points or vectors that are the closest to the hyperplane and affect its position.

iv) Kernel: The function used to map a lower dimensional data into a higher dimensional data. This is important function because SVR performs linear regression in a higher dimension. There are many types of kernel functions like polynomial kernel, Gaussian kernel, sigmoid kernel etc.

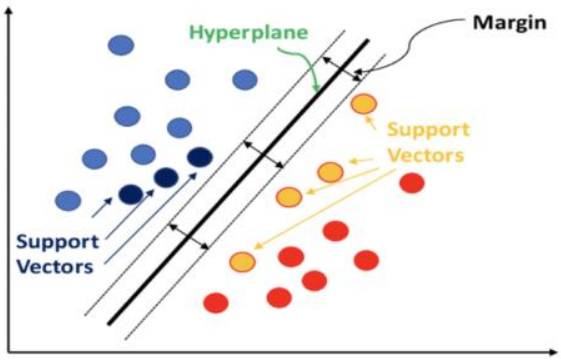

Figure 6 Support Vector Regression

Support Vector regression is a non-linear algorithm to find the optimal hyperplane that divides the given input dataset in N-dimensional space (N is the number of features). The optimal hyperplane is maximum margin of separation between all training data points and hyperplane. While searching for optimal hyperplane, the algorithm tries to find support vector or boundary points. These support vectors are picked such that the hyperplane will be at a possible maximum distance from both support vectors. Above fig. 1 shows the optimal hyperplane with maximum margin of separation and two decision boundaries on either side of optimal hyperplane along with potential support vectors.

Hyperparameters

The hyperparameters that are optimized for Support Vector Regression and Random Forest Regressor are shown in table 3 and table 4. For the sake of simplicity, hyperparameter tuning is optimized on the metric RMSE. To ensure the reliability of the results, a sensitivity analysis of the hyperparameters is conducted. A range of possible good hyperparameter values from the validation set is found. The hyperparameter value is then iterated through the range, to see whether the results on the test set hold. This is done to ensure that the optimal hyperparameters are not chosen at random and that the conclusions from this study hold. Grid search is employed in this study to systematically explore different combinations of hyperparameters for Random Forest and SVR. This process aims to improve the models' performance by fine-tuning their configurations. The hyperparameter grids are defined as follows:

Table 3 Hyperparameters Random Forest Regressor

Parameter	Values		
n_estimators	50	100	150
max_depth	None	10	20
min_samples_split	2	5	10
min_samples_leaf	1	2	4

Table 4 Hyperparameters Support Vector Regression

Parameter	Values				
C	0.1	1	10	100	1000
γ	0.0001	0.001	0.01	0.1	1

ANALYSIS REPORT

The standard metrics used to compare the time series data will be Mean Squared Error, Root Squared Error and Mean Absolute Percentage Error. The lower the value of MSE and RMSE the better the fit of the model.

Mean Squared Error: It measures the average of squares of errors i.e. the average squared difference between the actual values and the predicted values.

Equation 1 Mean Squared Error Equation

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

Root Mean Squared Error: It measures the difference between values predicted by a model and the values measured. It represents the square root of the difference between predicted values and the observed values of these differences. These deviations are called Residuals.

Equation 2 Root Mean Squared Error Equation

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

Where N= Total predictions
The models were trained and evaluated using a dataset split into training and testing sets. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were chosen as evaluation metrics. The results of the evaluation are presented in the table 5 below:

Table 5 Comparative analysis of MSE and RMSE values obtained using LR, RF and SVR models.

Model	**MSE**	**RMSE**	**After Hyperparameter tuning RMSE**
Linear Regression (LR)	74.92	8.66	8.66
Random Forest (RF)	118.54	11.02	10.79
Support Vector Regression (SVR)	489807.75	27.91	10.53

The linear regression model demonstrates the lowest RMSE, indicating better predictive performance compared to Random Forest and SVR. Random Forest follows closely, while SVR appears to struggle in capturing the complexity of the TCS stock price data. Fig. *7* represents graphs showing original price of stock vs predicted price of stocks using *LR, Tuned* RF, *Tuned SVR*.

Figure 7 LR, RF, SVR Stock Price Prediction graph.

CONCLUSION

In conclusion, this research explores the application of machine learning models for TCS stock price prediction. As part of this research, three algorithms i.e., Linear Regression, Random Forest and Support Vector Regression algorithm were chosen for the prediction of stock prices of TCS company. The dataset was starting from Jan 2005 upto Jan 2024. Feature engineering, hyperparameter tuning, and thorough analysis were performed to understand the models' performance. The models were tested on two essential performance metrics, namely, R-squared value (R2), and Root Mean Square Error (RMSE). The results suggest that linear regression and random forest models show promise in predicting TCS stock prices, with linear regression exhibiting superior performance due to least value of errors MSE(74.92) and R2(8.66). Future work could involve further refinement of features, exploring additional models, incorporating sentiment analysis and considering external factors that may influence stock prices. Overall, this study contributes to the ongoing efforts to leverage machine learning for accurate and reliable stock price predictions.

REFERENCES

1. Malti Bansal, Apoorva Goyal, Apoorva Choudhary " Stock Market Prediction with High Accuracy Using Machine Learning Techniques" Malti Bansal et al. / Procedia Computer Science 215 (2022) 247–265.
2. Md. Mobin Akhtar, Abu Srwar Zamani, "Stock market prediction based on statistical data using machine learning algorithms" 1018-3647 The Author(s). Published by Elsevier B.V. on behalf of King Saud University 2022.
3. Dr. Poorna Shankar, "Stock Price Prediction Using LSTM, ARIMA and UCM", Journal of Development Economics and Management Research Studies (JDMS), A Peer Reviewed Open Access International Journal, ISSN 2582 5119 (Online), 09 (11), 55-66, January-March, 2022.
4. Abdulhamit Su basi, Faria Amir, Kholoud Bagedo, Asmaa Shams, Akila Sarirete "Stock Market Prediction Using Machine Learning" Abdulhamit Subasi et al. / Procedia Computer Science 194 (2021) 173–179 ,18th International Learning & Technology Conference 2021.
5. Wasiat Khan, Mustansar Ali Ghazanfar, Muhammad Awais, Amin Karami Alyoubi, Ahmed S. Alfakeeh " Stock market prediction using machine learning classifiers and social media news", Journal of Ambient Intelligence and Humanized Computing March 2020.
6. Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar "Stock Closing Price Prediction using Machine Learning Techniques" 1877-0509 © 2020 The Authors. Published by Elsevier B.V, International Conference on Computational Intelligence and Data Science (ICCIDS 2019).
7. Srinath Ravikumar , Prasad Saraf "Prediction of Stock Prices using Machine Learning (Regression, Classification) Algorithms" 2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, 2020. 978-1-7281-6221-8/20/$31.00 ©2020 IEEE.
8. M Umer Ghani "Stock Market Prediction Using Machine Learning(ML)Algorithms" ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal Regular Issue, Vol. 8 N. 4 (2019), 97-116 eISSN: 2255-2863 DOI: http://dx.doi.org/10.14201/ADCAIJ20198497116
9. Pawee Werawithayaset , Suratose Tritilanunt "Stock Closing Price Prediction Using Machine Learning" Seventeenth International Conference on ICT and Knowledge Engineering 2019.
10. Pashankar, S. S., & Dhoke, A. A Case Study: On Indira College Of Commerce & Science Student's Campus Placement Determination Using Logistic Regression Analysis For Prediction.
11. Nagaraj Naik, Biju R. Mohan , K. Somani et al. "Optimal Feature Selection of Technical Indicator and Stock Prediction Using Machine Learning Technique", Springer Nature Singapore Pte Ltd. 2019. (Eds.): ICETCE 2019, CCIS 985, pp. 261–268, 2019.
12. Ashish Sharma, Dinesh bhuriya "Survey of Stock Market prediction Using Machine Learning Approach", International Conference on Electronics, Communication and Aerospace Technology ICECA 2017 IEEE.