https://doi.org/ 10.51153/kjcis.v6i2.150

Stock Prediction for ARGAAM Companies Dataset

Noman Islam[1], SalisKhizar Khan[2], Abdul Rehman[3], Usman Aftab[4], Darakhshan Syed[5]

Abstract

Economic forecasting provides excellent profit opportunities and is a major motivator for most researchers in this field. In the fast-growing business world, the behavior of stock prediction is challenging for most stockholders and commercial investors. It provides benefits to investors to invest more confidently. Machine learning is an emerging technology that provides the capability to learn on its own through real-world intercommunications. Regression is the fundamental technique in machine learning which is useful for real-time applications. This paper experiments with stock price prediction effectively by using three machine learning techniques i.e. linear regression, decision tree, and support vector machine. The techniques were applied to the ARAMCO and Saudi Dairy dataset and the performance is evaluated using various parameters such as R2 value, MAPE, and RMSE. The results substantiated the hypothesis.

Keywords: machine learning, linear regression, decision trees, support vector machine, stock price prediction, forecasting

1. Introduction

Financial leverage or stock price forecasting is an important financial topic that has gained a lot of interest from academics for several years. It is based on the notion that publicly available data from the past has some forecast patterns for future market returns. The stock market is one of the most highly followed markets in the world. In this modern technology era, a lot of innovation has been made in different sectors of this market through technology. By analyzing the previous stock data, we can predict stock prices and indexes. A modeling technique called stock market prediction uses fundamental characteristics of the stock price to predict its potential values. The accessibility of a vast quantity of previous data is essential for such a forecasting system to succeed in its pursuit of profitable money markets. The efficacy of these algorithms is severely constrained by the fact that the data used in this sort of investigation are financial time series information. Furthermore, as risks are inherent as a result of irregular market patterns, volatility, noise, etc., they cannot be disregarded when using such approaches[1]. The effective market hypothesis, which holds that the uncertainty-adjusted return, cannot be persistently achieved

[1]*Karachi Institute of Economics and Technology, Pakistan (noman.islam@gmail.com)*
[2]*NED University, Pakistan (khan.pg3200463@cloud.neduet.edu.pk)*
[3]*NED University, Pakistan (khan.pg3200463@cloud.neduet.edu.pk)*
[4]*NED University, Pakistan (aftab.pg3401118@cloud.neduet.edu.pk)*
[5]*Bahria University, Pakistan (darakshansyed.bukc@bahria.edu.pk)*

much above the viability of the entire [1] [2], is thus something that the regression techniques are bound to follow. This hypothesis makes the supposition that the stock value at the moment may be calculated as a consequence of stock price previous records and reasonable forecasts. Any departure from this premise could make the stock price unexpectedly.

To construct the stock market predictions with technological advances, a variety of machine learning (ML) methods was, nevertheless, conveniently simulated with the expansion of computing capacities. The famous artificial neural network (ANN) and its derivatives, genetic approach, support vector machine (SVM), support vector regression (SVR), and others have all been employed by researchers to forecast share prices. The difficulties faced by these predictive analytics are dealing with uncertainty while making more accurate stock price predictions, which reduces threats for traders and investors and produces a lucrative method. These models' accuracy draws more researchers, who are then inspired to suggest new, more accurate forecasting strategies. Currently, there has been a lot of fascinating work in the area of stock price prediction using ML techniques. The objective is to analyse market trends and predict stock prices including the index changes. Machine learning is a field of artificial intelligence that allows systems to learn from information and complete tasks without ever being given explicit step-by-step commands, depending solely on the data they've been fed [3-5]. ML is utilized in a variety of areas where creating a traditional technique to solve an issue is difficult. The stock market is a gathering of stock buyers and sellers. A stock is a fraction of a company's ownership held by a single entity or group of people[6-9]

One of the techniques for stock price prediction using machine learning is linear regression. It is the method in which historical data is fit to a straight line using the equation y = mx + c. Mathematically we can define linear regression as the method in which the values of the data set are fit to the straight line through the points which means the sum of the square of the distance between each point and the line is the shortest. The formula for the linear regression is known as the least square method and its hypothesis function is:

$$y=h(x)=\emptyset\_0+\emptyset\_1 x \quad \ldots \quad (1)$$

In addition, various other machine learning techniques can also be used. Decision tree is a technique that is based on a tree-like structure where at each node a feature is tested. Based on that, a value can be predicted at the leaf of the tree. Support vector machine is based on kernel technique where calculations of higher dimension are performed in the lower dimension using kernels. Other techniques include Naïve Bayes, random forest, logistic regression, and deep learning techniques. This research study examines publications and explained how these techniques provide precise and reliable future forecasts to predict stock prices in the financial market. We employed several methods in this research, and we found that not all of them were able to forecast the data we required. There seems to be a fundamental need for computerized and automated techniques to handle with effective and powerful utilization of a significant

amount of financial statistics to assist businesses and individuals in deciding crucial budgeting and planning.

In this paper, various machine-learning techniques have been applied to predict stock prices based on past data. The results are analyzed through root mean square error, R2 value and mean absolute precision error. The rest of the sections of the paper discusses the methodology. The next section presents the literature review. This is followed by the methodology and results section. The paper concludes with discussions on future work.

The main contributions of the paper are as follows:

The paper presented machine learning as an emerging technology that provides the capability to learn on its own through real-world datasets.
Use linear regression, decision tree, and support vector machine to predict the stock prices for Saudi companies by using historical data. The above-mentioned techniques were applied to the ARAMCO and Saudi Dairy datasets.
The performance is evaluated using various parameters such as R2 value, MAPE, and RMSE. The results are presented and analyzed to substantiate the hypothesis.

2. Literature Review

There has been a significant volume of research on employing machine learning for stock price prediction. Some of these researches are [10-18]. Selvamuthu et al.[19] presented a comparative study on the performance of several schemes on the tick dataset and 15-min dataset of an Indian company. This comparison concludes that the tick dataset has preferable accuracy to the 15-min dataset.

Since the advent of the virtual maturity level, analogical reasoning has transitioned to the area of technology that is most important and according to Zhang et al. [20], the most suitable strategy is to use synthetic neural networks and momentary neural networks, which may be considered to be a form of device awareness. Mobin et al. [21] proposed a workable tool to predict the movement of stocks with less finesse. The dataset of stock prices from the prior period was the first problem they examined. For the original study, the dataset has undergone pre-processing and refinement. Due to this, the structure may even concentrate on preparing the dataset's original information. After preparing the data, one can see how the vector machine is helped by the data set and the output it produces when using arbitrary woods. Similar to this, the suggested composition looks at the application of the soothsaying tool in real-world settings and issues linked to the sensitivity of the advice given. The accuracy level of the Stock Prediction model is 80.3%. Khattak et al. [22] proposed a model for predicting stock trends with a focus on lowering the sparsity of the data gathered through machine learning. Researchers use a k-nearest neighbor technique to categorize stock metrics after doing an outlier identification of the information provided for dimensionality reduction.
In [23], the analysts have used (ANN) and applied the mathematics and statistics technique ARIMA

on nearly three years of statistical data to forecast the KSE0-100 index. The research provides [24] a depth evaluation of the bottom-line relationship between large-scale-economic factors and the KSE market. In the same way, some researchers showed that a person's feeling plays a major role in analyzing and ultimately making a decision [25]. If you have data about the user's mood and their behavior by using social media, it will be very helpful to predict the user's decisions for investment in business as well as market performance. In [26], the researchers present the process of collecting data from different international financial markets. After that with the help of machine learning algorithms, they predicted the stock prices. In [27], the researchers show how the impact of financial and commercial news on stock prices. Some political events occurred, and they affected stock prices. Correspondingly, various data science techniques are also used in [28] the researchers have used variable moving average techniques by using the data of the Vietnamese stock market. Now a day's many companies and institutions are working on predictions of the stock market. These companies aim to propose the best method to predict the daily behaviors of stock market indexes and modify their systems accordingly. Majhi et al.[29] predicted stock market and SP500 indexes by using bacterial foraging optimization techniques to achieve short and long terms goals. After that, they compared their model to a Multilayer perceptron, and they used the liner combiner model to calculate their weights. The results show that the Majhi method is the very best fit in that situation because of the less complexity and more precision of the MLP method. Elon Musk tweets about Bitcoin, increasing the prices of Bitcoin in the cryptocurrency market. Stock prices are also depending on a few people's statements. In technological terms, these concepts are described as text analytics and text mining. This is another type of forecasting system [30] that are using a text-mining approach to predict real-time stock prices and market behaviors. Few institutions and researchers used text analytics and text mining approaches. Their findings investigate the effects of economic and financial news in predicting the stock market. Like a work about the effect of emotions like hope, fear, and worry on increasing or decreasing the amount of Dow Jones on the next day [31] or the effects of Facebook on the stock market. Other influential people's behaviors and mood affect stock markets. Therefore, there is a connection [32] between the capacity of investors and other activities of the stock market using the updated mood of people. Past research shows that 100 million American Facebook user mood impacted on the stock market between the periods of 10/09/2007 to 10/09/2010.

Another related work is based on Bollen et all's strategy [33]. The Bollen follows the text mining and text analytics approach. This approach received worldwide media coverage from various arbitrary resources. The main strategy they followed measuring the behavior and mood of people that are affecting the stock market. The authors gather all the data of Twitter users in 2008, then with the help of Google algorithms that are based on mood states namely Google Profile Mood Status to detect the public response for the presidential elections. The algorithms are categorized into six main parts that are calm, happy, vital, sure, alert, and Kind. The results obtained are very shocking for that research work.

Based on the research, we can see that there is still a gap in using machine learning techniques

to predict stock market prices. This research undertakes this problem and applies three popular machine-learning algorithms to predict stock prices.

3. Methodology

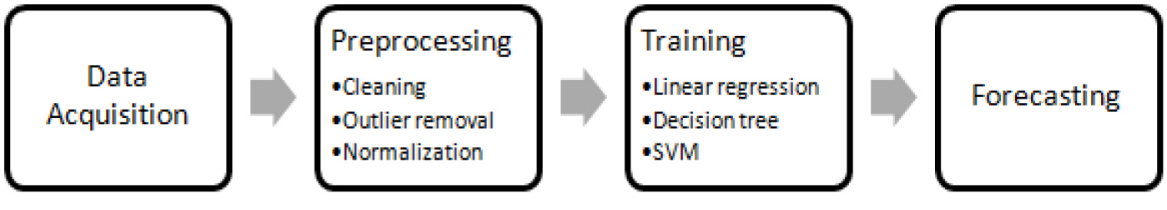
The overall implementation pipeline is shown in Figure 1a along with the dataset (Figure 1b). In this research, we have first collected the stock prices company data i.e. Saudi Dairy and Aramco Company, and then performed pre-processing on it. The complete procedure of pre-processing in this project can be divided into three text operations:
Data collection i.e. removing unwanted columns that are not used in stock predictions and cleaning data by removing null records.
Exploring (exploratory data analysis)
Predicting stock prices

The dataset comprises the following columns:
the date,
opening price,
closing price,
high and low price of the stock on that particular day,
adjusted closed price and the volume of the stock.

The records containing null values were removed using the panda's library. The data was standardized by subtracting the mean value and then normalized by the standard deviation as follows:

$$\frac{(x - \mu)}{\sigma} \quad \ldots \quad (2)$$

For outlier removal, those values are considered as outliers that are ±3 standard deviation from the mean.



a. Implementation pipeline

date	Open	High	Low	Close	adj_close	Volume
01/02/1993	11.72	11.88	11.72	11.85	11.85	204291
01/03/1993	11.88	11.94	11.62	11.94	11.94	203977
01/04/1993	11.88	12.07	11.88	12.07	12.07	328867
01/05/1993	12.13	12.13	12.03	12.13	12.13	178701
01/06/1993	12.03	12.19	10.01	10.01	10.01	272077
01/09/1993	12.03	12.07	9.92	9.92	9.92	333840
01/10/1993	12.03	12.03	11.98	12	12	136544
01/11/1993	12	12	11.88	11.91	11.91	76701
01/12/1993	11.91	11.91	11.78	11.78	11.78	190860
1/13/1993	11.85	12.7	11.78	11.82	11.82	531981
1/16/1993	11.78	11.78	11.65	11.65	11.65	277290
1/17/1993	11.65	11.72	11.56	11.59	11.59	309150
1/18/1993	11.62	11.62	11.62	11.62	11.62	10401
1/19/1993	11.56	11.65	11.4	11.62	11.62	652500
1/20/1993	11.59	11.68	11.59	11.65	11.65	105591
1/23/1993	11.68	11.68	11.62	11.68	11.68	215647
1/24/1993	11.65	11.68	11.56	11.56	11.56	125054
1/25/1993	11.56	11.62	11.53	11.53	11.53	175680
1/27/1993	11.53	11.56	11.49	11.56	11.56	257947
1/30/1993	11.56	11.59	11.47	11.56	11.56	130071
1/31/1993	11.53	11.59	11.53	11.56	11.56	267577
02/01/1993	11.59	11.59	11.56	11.56	11.56	75891

b. Dataset details

Figure 1: The Implementation Details

Table 1, 2, and 3 shows the dataset description and results after pre-processing. For evaluating techniques of machine learning, simple linear regression, and regression using the last values method, reasonable accuracy of the model is achieved, as discussed in the next sections. Considered performance evaluation parameters includes the: root mean square (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), the mean square error (MSE), and the correlation coefficient (R2). The basic linear regression model is used from sklearn with the following criteria: 60% training, 20% validation, and 20% testing division of the dataset.

Table 1: Columns removed from the dataset

S. No	Columns Removed
1	TCCompanyID
2	CompanyID
3	EntryNumber
4	Amount
5	ContractCount
6	TCMarketID
7	MarketID

Table 2: Main measures of the dataset

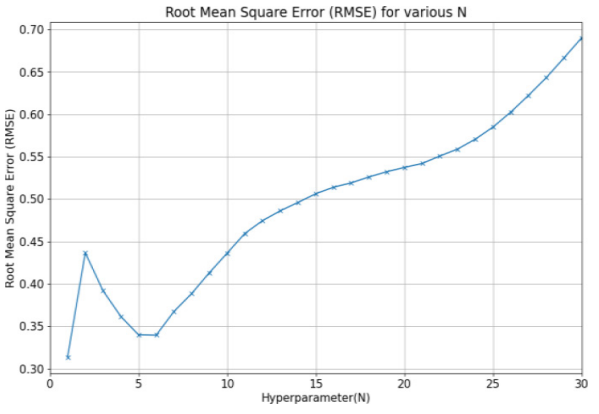
S. No	Main Measures of the Dataset
1	7345 rows x 7 columns
2	The length of the dataframe: 7345
3	num_cv = 1469
4	num_test = 1469
5	num_train = 4407
6	train. shape = (4407, 8)
7	cv.shape = (1469, 8)
8	train_cv.shape = (5876, 8)
9	test.shape = (1469, 8)

Table 3: Data Cleaning

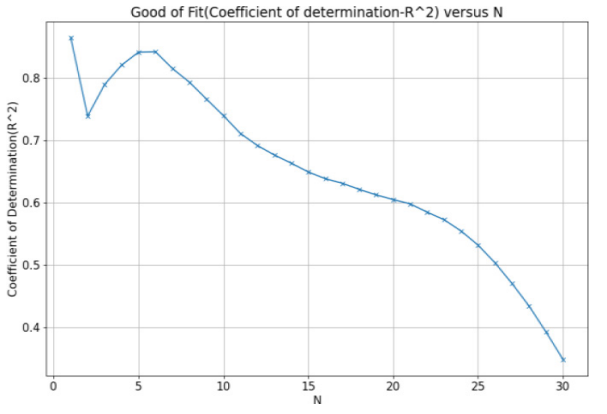
S. No	Columns Added for prediction using last values
1	Close Value – T1
2	Close Value – T2…
3	….
4	Close Value- T30

4. Results

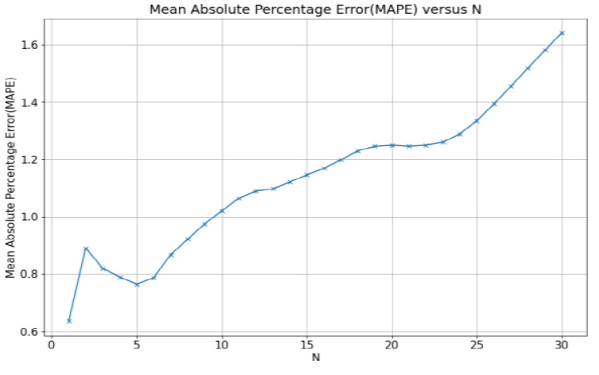
In this section the results obtained from the experiments are shown. By providing the proper simulation screenshots. Figure 2 shows the result of linear regression. The root mean square shown in Fig 2a initially decreases with the rise in the value of N. The model perfectly fits at n=5. Afterward, the model starts overfitting as evident from R2 square value shown in Fig 2b. The comparison of the predicted and actual prices is also shown in Figure 2d.



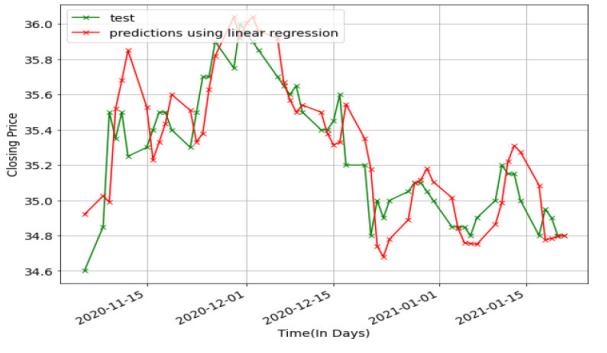
a. N vs. RMSE



b. N vs. R^2



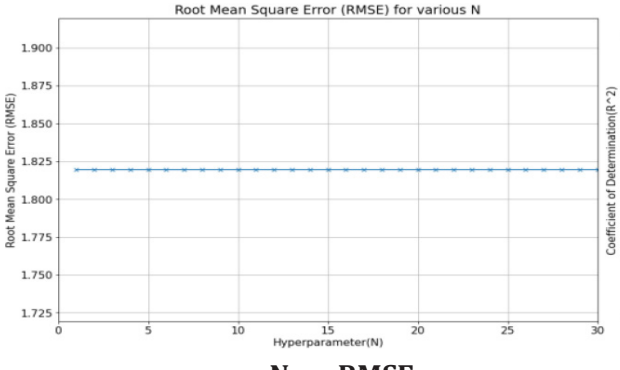
c. N vs. MAPE



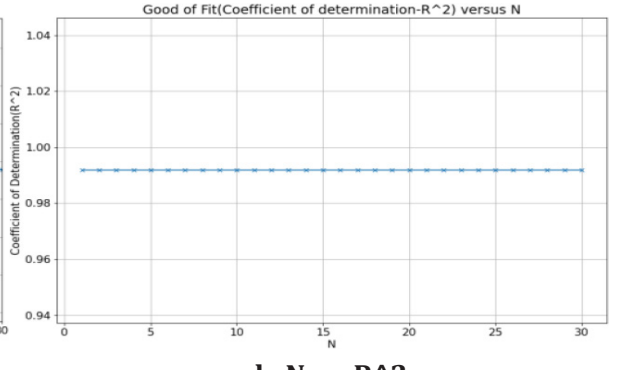
d. Time versus stock close value (test and predicted data)

Figure 2: Results for linear regression

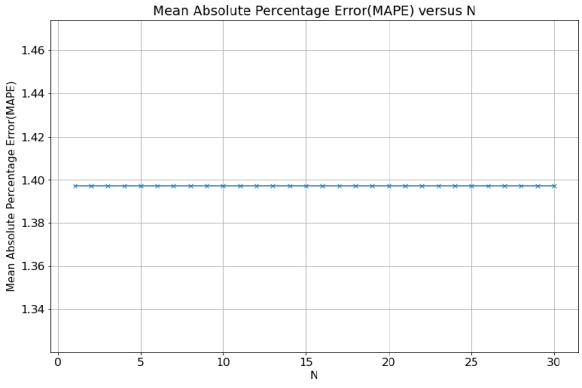
The result for the decision tree is shown in Figure 3. As shown in Figure the R2 value remains constant and doesn't vary much no matter how much past data is used for predicting future data. The root means the square error is around 1.8.



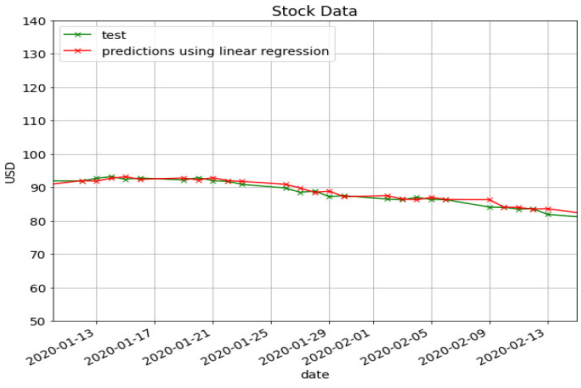
a. N vs. RMSE



b. N vs. R^2



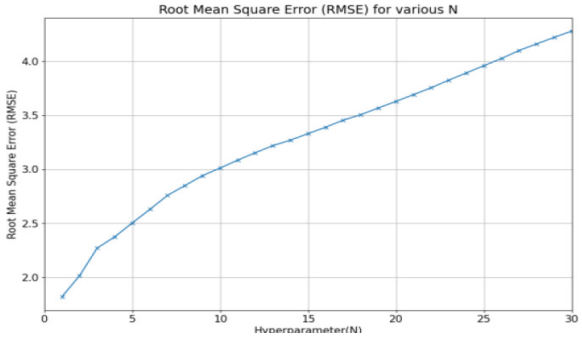
c. N vs. MAPE



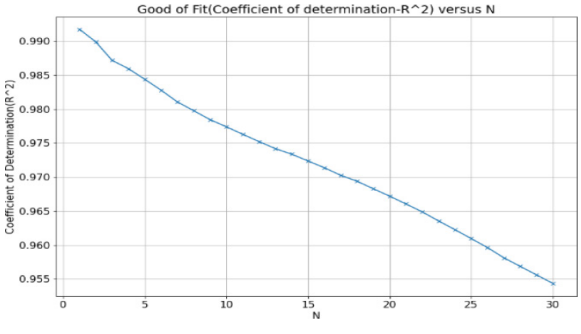
d. Time versus stock close value (Test and Predicted data)

Figure 3: Results for Decision Tree

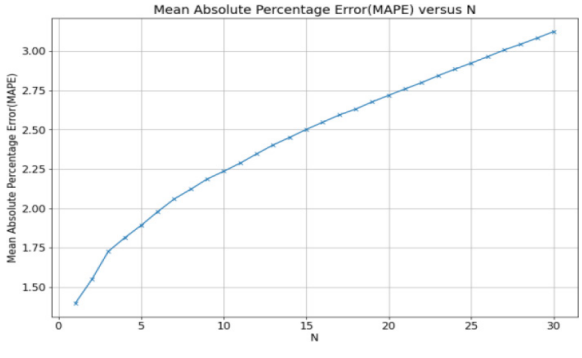
Figure 4 shows the result for the support vector machine. As can be seen, the model starts overfitting from the beginning.



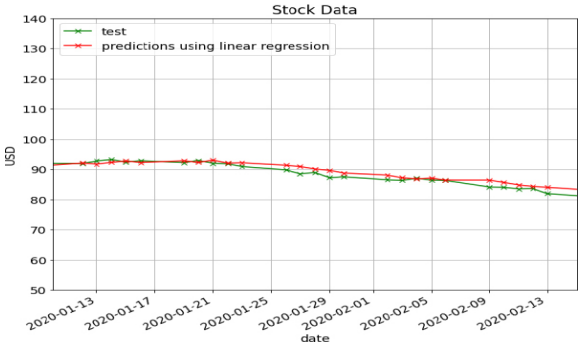
N vs. RMSE



N vs. R^2



N vs. MAPE



¬Time versus stock close value (Test and Predicted data)

Figure 4: Result for support vector machine

The results for basic linear regression are shown in Table 3. The table shows the RMSE, R2, and MAPE values for values of N i.e. the number of past data to be used for future prediction.

Table 3: Performance of linear regression for various hyper-parameter values

Data Set: ARAMCO

N	RMSE	R2	MAPE
1	0.313	0.865	0.636
5	0.339	0.841	0.789

Data Set: Saudi Dairy

N	RMSE	R2	MAPE
1	2.482	0.968	1.426
3	3.237	0.946	1.860

The criterion for the evaluation of results is given below:

1. R2> 95% is up to mark
2. MAPE<10 % is excellent <20% is good
3. RMSE <0.5 is up to mark or RMSE for testing data sets is considerably lower than training data sets for every instance

As can be seen, all of the techniques provide a good fit for regression as can be evident from Table 4. However, the decision tree provides the best result. A comparison has been made with other approaches to the literature review and similar results have been reported as shown in Table 5.

Table 4: Results for various machine learning regression techniques

Metrics	Linear regression	Decision Tree	Support vector machine
R2	0.80	0.99	0.95
RMSE	0.31	1.82	1.81
MAPE	0.60	1.39	1.39

Table 5: Comparison with other approaches

Technique	R2 value
Decision Tree	0.990
Multi-layer perceptron (MLP) [34]	0.969
Convolutional neural network (CNN) [34]	0.973
Recurrent neural network (RNN) [34]	0.975
Long short-term memory (LSTM) model [34]	0.977

5. Conclusion and future work

In this paper, linear regression, decision tree, and support vector machine were used to predict the stock prices for Saudi Company by using its real historical data. These three methods were used to design the model and acquire optimum results. It is found that simple linear regression models are not up to mark as confidence (R2) is just 86%-96% with MAPE 63%-142% and RMSE 0.313-2.48 which is well beyond the criteria of good regression. Whereas results obtained using decision tree values are highly accepted as confidence (R2) acquired for 99% with MAPE always less than 5% and RMSE is around 1.2 to 1.5%. Recommended other features selected to enhance the results of the model are news related to the industry, government policies impact, new Product launches, and dollar Prices.

References

[1] Malkiel, B.G., *The efficient market hypothesis, and its critics.* Journal of economic perspectives, 2003. 17(1): p. 59-82.

[2] Henrique, B.M., V.A. Sobreiro, and H. Kimura, *Stock price prediction using support vector regression on daily and up-to-the-minute prices.* The Journal of Finance and data science, 2018. 4(3): p. 183-201.

[3] Ahangar, R.G., M. Yahyazadehfar, and H. Pournaghshband, *The comparison of methods artificial neural network with linear regression using specific variables for prediction stock price in Tehran stock exchange.* arXiv preprint arXiv:1003.1457, 2010.

[4] Bikker, J.A., et al., F*orecasting market impact costs and identifying expensive trades.* Journal of Forecasting, 2008. 27(1): p. 21-39.

[5] Dzikevičius, A. and S. Šaranda, *EMA Versus SMA usage to forecast stock markets: the case of S&P 500 and OMX Baltic Benchmark.* Business: Theory and Practice, 2010. 11(3): p. 248-255.

[6] Edwards, R.D., J. Magee, and W.C. Bassetti, *Technical analysis of stock trends.* 2018: CRC Press.

[7] Janssen, C., C. Langager, and C. Murphy, *Technical Analysis: Indicators And Oscillators.* Website paper: Available at: http://www.investopedia. com/university/technical/techanalysis10. asp, 2006.

[8] Naeini, M.P., H. Taremian, and H.B. Hashemi. *Stock market value prediction using neural networks. in 2010 international conference on computer information systems and industrial management applications* (CISIM). 2010. IEEE.

[9] https://www.learn-stock-options-trading.com/stock-charts.html. *Introduction to Stock Charts.* 2009 [cited 2022 17th October].

[10] Dash, R.K., et al., *Fine-tuned support vector regression model for stock predictions.* Neural Computing and Applications, 2021: p. 1-15.

[11] Nguyen, G.L., et al., *A collaborative approach to early detection of IoT Botnet.* Computers & Electrical Engineering, 2022. 97: p. 107525.

[12] Nguyen, C.H., et al., *The linguistic summarization and the interpretability, scalability of fuzzy representations of multilevel semantic structures of word-domains.* Microprocessors and Microsystems, 2021. 81: p. 103641.

[13] Pham, D.V., et al., *Multi-topic misinformation blocking with budget constraint on online social networks.* IEEE Access, 2020. 8: p. 78879-78889.

[14] Le, N.T., et al., *Fingerprint enhancement based on tensor of wavelet subbands for classification.* IEEE Access, 2020. 8: p. 6602-6615.

[15] Le, N.T., et al., *Novel framework based on HOSVD for Ski goggles defect detection and classification.* Sensors, 2019. 19(24): p. 5538.

[16] Le, N.T., et al., *Automatic defect inspection for coated eyeglass based on symmetrized energy analysis of color channels.* Symmetry, 2019. 11(12): p. 1518.

[17] Vu, D.L., et al., HIT4Mal: *Hybrid image transformation for malware classification. Transactions on Emerging Telecommunications Technologies, 2020.* 31(11): p. e3789.

[18] Vu, D.-L., et al. *A convolutional transformation network for malware classification.* in 2019 6th NAFOSTED Conference on information and computer science (NICS). 2019. IEEE.

[19] Selvamuthu, D., V. Kumar, and A. Mishra, *Indian stock market prediction using artificial neural networks on tick data.* Financial Innovation, 2019. 5(1): p. 1-12.

[20] Zhang, X., et al., *Stock market prediction via multi-source multiple instance learning.* IEEE Access, 2018. 6: p. 50720-50728.

[21] Akhtar, M.M., et al., *Stock market prediction based on statistical data using machine learning algorithms. Journal of King Saud University-Science, 2022.* 34(4): p. 101940.

[22] Khattak, A., et al., *An efficient supervised machine learning technique for forecasting stock market trends, in Information and Knowledge in Internet of Things.* 2022, Springer. p. 143-162.

[23] Haritaoglu, I., et al., *Backpack: Detection of people carrying objects using silhouettes.* Computer Vision and Image Understanding, 2001. 81(3): p. 385-397.

[24] Saleem, F. and M.N. Alifiah, *Causal relationship between macroeconomic variables and stock prices in Pakistan.* Jurnal Kemanusiaan, 2017. 15(1).

[25] Huang, Y., et al. *Boosting financial trend prediction with twitter mood based on selective hidden Markov models.* in International Conference on Database Systems for Advanced Applications. 2015. Springer.

[26] Rosenzweig, E., *Successful user experience: Strategies and roadmaps.* 2015: Morgan Kaufmann.

[27] Asur, S. and B. Huberman, *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.* 2010.

[28] Ming, K.L.Y., M. Jais, and B.A. Karim, *Technical Analysis and Stock Price Prediction? Evidence from Malaysian Stock Market.* 2016.

[29] Abdelkarim, N., et al., *A New Hybrid BFOA-PSO optimization technique for decoupling and robust control of two-coupled distillation column process. Computational Intelligence and Neuroscience, 2016.* 2016.

[30] Gharehchopogh, F.S., T.H. Bonab, and S.R. Khaze, *A linear regression approach to prediction of stock market trading volume: a case study.* International Journal of Managing Value and Supply Chains, 2013. 4(3): p. 25.

[31] Zhang, X., H. Fuehres, and P.A. Gloor, *Predicting stock market indicators through twitter "I hope it is not as bad as I fear".* Procedia-Social and Behavioral Sciences, 2011. 26: p. 55-62.

[32] Wang, Z., S.-B. Ho, and Z. Lin. *Stock market prediction analysis by incorporating social and news opinion and sentiment.* in 2018 IEEE International Conference on Data Mining Workshops (ICDMW). 2018. IEEE.

[33] Bollen, J., H. Mao, and X. Zeng, *Twitter mood predicts the stock market. Journal of computational science, 2011.* 2(1): p. 1-8.

[34] Lu, W., Li, J., Wang, J. and Qin, L., 2021. *A CNN-BiLSTM-AM method for stock price prediction.* Neural Computing and Applications, 33(10), pp.4741-4753.