Proceedings of the International Conference on Edge Computing and Applications (ICECAA 2022)
IEEE Xplore Part Number: CFP22BV8-ART; ISBN: 978-1-6654-8232-5

# Stock Market Prediction using Machine Learning Models

A.Yasmin, Research Scholar
Department of Computer Science
School of Computing Sciences
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Pallavaram, Chennai, India
yasmin.a@jbascollege.edu.in

S. Kamalakkannan, Associate Professor
Department of Information Technology
School of Computing Sciences
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Pallavaram, Chennai, India
kannan.scs@velsuniv.ac.in

P.Kavitha, Research Scholar
Department of Computer Science
School of Computing Sciences
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Pallavaram, Chennai, India
pkavikamal@gmail.com

2022 International Conference on Edge Computing and Applications (ICECAA) | 978-1-6654-8232-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECAA55415.2022.9936188

**Abstract— Stock market prediction is the needed emerging economic statistics from business to normal middle-class peoples, to make their investment as a profitable one. This article has utilized the dynamic dataset of the company. The dataset includes the closing price of the stock of the last 290 working days. The dataset is downloaded using the yahoo finance (https://finance.yahoo.com), so the data is pretty accurate. Further, some technical analysis and machine learning techniques are used to predict the future prices and exchange of company's stock. The machine learning models includes Linear Regression, Decision Tree, Random Forest, SVR, LSTM, Lasso Regression, KNN, Bayesian Ridge, Gradient Boosting, and Ada Boost are used in this article and suitable technique for the dataset is chosen for performing effective prediction of stock market.**

***Keywords: Stock Market, Share Market, Stock Prediction, machine learning***

## I. INTRODUCTION

In recent days, data analysis is the emerging concept over the university which helps to predict the future trends in the specific field. In that sequence, stock market analysis for predicting the future price of the certain stock was considers as the major requirement in current days. Day to day process on stock exchange increases the data in huge manner on daily basis and fluctuations within a day hikes the existing data with additional volume. And to analyze these changes and predict the future price or market trends of certain company stock was a challenging task in this scenario. To provide the solution for these and to provide the accurate prediction to the end user or stake holders is addresses in this paper. This paper attempts to use machine learning approaches to predict the future trend of stock market with the existing data [1].

Stock cost investigation has been a basic area of examination and is one of the top uses of AI. Stock Price Prediction with the assistance of AI provides you with the worth of the stock from here on out. The fundamental point behind foreseeing stocks is to acquire the most extreme benefits. Anticipating how the securities exchange will perform is a hard assignment to do. There are different factors likewise associated with the expectation, for instance, physical (area of organization, item sending off area), mental (notoriety of an organization among the clients - premium in the event of Apple, an incentive for cash if there should be an occurrence of Xiaomi). All the mix of these variables turns the offer costs dynamic and unpredictable. This way of behaving of market makes it truly challenging to anticipate stock costs with high exactness. Precise expectation of securities exchange returns is an exceptionally provoking undertaking because of unstable and non-straight nature of the monetary securities exchanges. With the presentation of man-made consciousness and other operational perspective with computational techniques may predict the stock cost in an effective and productive manner. In this paper, LSTM and some machine learning models calculations have been identified to forecasting the next day including stock price of five companies based on their shutting cost of previous fifteen days market prices. We are Focusing on the Close Price of the Stock as the Close value timings are fixed as because of AMO (After Market Orders) and PMO (Post Market Orders) the It are not fixed to Open Price timings. [2]

## II. RELATED WORK

In [3], Artificial Neural Network and Random Forest strategies have been used for anticipating the following day shutting cost for five organizations having a place with various areas of activity. The monetary information: Open, High, Low and Close costs of stock are utilized for making new factors which are utilized as contributions to the model. The models are assessed utilizing standard vital markers: RMSE and MAPE.

In [4], the NIFTY updates are considered from stock market exchange form Indian stock market for the period of six months and try to predict the stock prices for the next some days by applying LSTM approach which was the model utilize the existing information and predict or forecast. The experiments are conducted for some limited company stocks to provide an efficient result.

In [5], they present an examination of the expectation by contributing various classifiers. The consequences of the examination are done on a precision premise. In [6] they use both unpredictable forest areas and LSTM associations (even more unequivocally CuDNNLSTM) as planning procedures to look at their ampleness in deciding out-of-test directional improvements for the period of 25 years from 1993 to 2018. They present a multi-incorporate setting containing not simply of the benefits with respect to the end costs, yet also concerning the underlying expenses and intraday returns. In [7], the Black-Scholes Option assessing model (BSOPM) has for a long while been being utilized for valuation of significant worth decisions to find the expense of stocks. In this work, using BSOPM, they have composed a close to logical strategy and numerical methodology to notice the expense of call decision and put decision and considered these two expenses as buying cost and selling cost of stocks in the wild grandstands with the objective that we can expect the stock expense (close expense). In [12], the customer opinion mining was discussed using machine learning model which supports the prediction to get the real opinion of the customer. In [13], author compares the price prediction models of stock market using neural network techniques. In [14], price prediction for bitcoin was discussed using LSTM model. In [15], general classification approaches was given to label the classifier output which makes the customers to understand the overall prediction view of the tuple in the dataset.

## III. DATASET

- The Dataset is dynamic dataset of last 290 working days using the yahoo finance in python.
- The Dataset has the Open, High, Low, Close, Adj Close, Volume.
- Steps for getting the dataset.
  1. We are taking the name of the company.
  2. Converting the name of the company to the yahoo ticker (it is the name by which the yahoo has stored the details) using the google search.
  3. Downloading the dataset.
  - The dataset of last 270d (9 months) is downloaded using yahoo finance.
  - We choose the 9 months data because it is recommended practice in the field of the stock market.

- Dataset is as follows:

- Open:
  - It is showing the opening price of the stock at 9:15 AM IST for Indian Stock market
  - There is the difference in last day close and open price due to AMO's. (After Market Orders)
- Close:
  - It is the closing price of the stock at 3:30 PM IST for Indian Stock market.
- High:
  - It represents highest value of the stock for that day.
- Low:
  - It represents lowest value of the stock for that day.
- Volume:
  - It is the number of the stocks changing hands.
  - There is no doubling in counting
- Adjacent Price:
  - The changed shutting cost alters a stock's end cost to mirror that stock's worth in the wake of representing any corporate activities.
  - It is in many cases utilized while looking at verifiable returns or doing a point-by-point examination of past execution
- From seeing the dataset, we got some findings like:
  - The Adjacent Close and Close price are completely same in the dataset.
  - There are some days where are 0 volumes. Meaning no transaction of the stock on that day.
  - Only Close Price can be predicted as the Open price depends on AMO, high price or low price depends on the relative strength of the bulls and bears.

## IV. ANALYSING THE DATASET

### DATA PREPROCESSING:

1. Converting to Indian Rupees.

   - The dataset of the foreign companies has the data in their currency. So to convert the currency to Indian Rupee I had used the exchange – rate apis.

### A. Exchange Rate Apis:

- They gather conversion standard information from various national banks and business sources and afterward utilize our own calculation to mix these different datasets.
- This interaction lessens the effect of a wrong distant transformation rate provided by one source. They possibly support a cash code in ExchangeRate-API assuming they have no less than 3 information hotspots for that money.
- Their trade rates are classed as demonstrative midpoint rates. These are precise enough for errands like cost assessments in an internet business store or details on a dashboard.

2. Removal of outliers.
- The outliers like zero volume data and the null value data are removed from the dataset

3. MinMax Scalarization

There will be many techniques are there in machine learning to evaluate the given umbers using existing or new mathematical operations and that will be scaled to high and low level to showcase the standardization of the technique. The given minmax scalarization utilize the weight information about the required features and that captures the required distance measures to chosen the nearest or closest k neighbor.

Based on the scaling factors there are two mathematical models are in practices where one is standardization and other one is normalization. In this standardization scales the given information from the variable 0 to 1. In these negative values cannot come into the scaling because there will be some values are present to standardize the existing values.

Next is normalization, where everyone studied in databases and know the different normalization process to normalize the data based on the requirement. The mean and standard deviation are evaluated for the given dataset and the unnormalized or missing values are normalized by applying normalization.

Normalization is a rescaling of the information from the first reach so that all values are inside the new scope of 0 and 1.

Normalization expects that you know or the value can be appraising the base without reduce the qualities of the bae value. The researcher can evaluate the qualities by accessing the information of the normalized values in the dataset.

Minmax Scaler scales each and every information within the range of [0,1] but not likely in range of [-1,1]. But it may also consider if there are any negative attributes existed in dataset.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

## V. PREDICTION METHODS MACHINE LERARNING MODELS

### *A.* Linear Regression

Linear regression is the method that relapse the factors and found the connectivity between the variables called as independent and dependent [8]. The linear regression line represented by the variables x and y, where x is the independent and y is the dependent as per the equation given below:

y = a0 + a1 * x
Here the target is to derive the best attributes or values which is suitable for a0 and a1.

### *B.* Decision Tree

The most remarkable and broadly involved device for order and expectation is the choice tree. A choice tree is a flowchart-like tree structure in which each inner hub addresses a property test, each branch mirrors the test's decision, and each leaf hub (terminal hub) stores a class name.

By isolating the source set into subgroups in view of a characteristic worth test, a tree can be "prepared." Recursive apportioning is the most common way of rehashing this strategy on each inferred subset. High-layered information can be dealt with through choice trees. The exactness of the choice tree regressor is by and large great.

### *C.* Random Forest

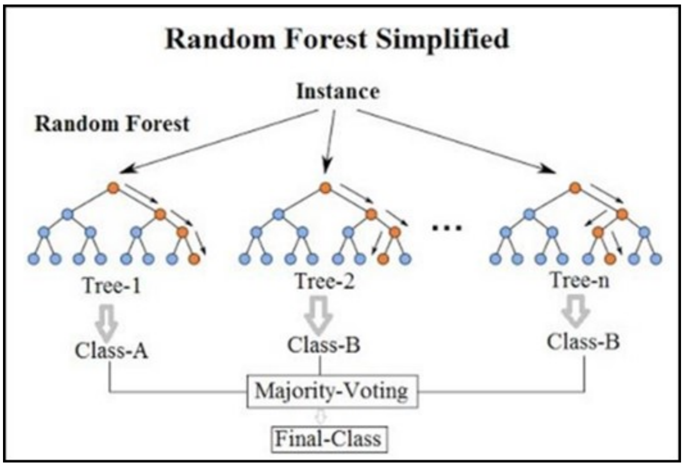

**Figure 1. Random forest classifier**

The Decision Tree is a really seen and untraveled computation and subsequently a lone tree not provide the required result with the features existent in the dataset. And to overcome this, Random Forest is furthermore a "Tree"- based estimation that uses the qualities components of different Decision Trees for choosing.

Subsequently, it will in general be suggested as a 'Woodlands' of trees and hence the name "Sporadic Forest". The term 'Erratic' is a direct result of how this estimation is boondocks of 'Indiscriminately pursued Choice Trees'.

Decision Tree computation has a critical weight in that it causes over-fitting. This issue can be confined by executing Random Forest in place of decision tree which was more accurate and time-consuming classification process in decision making which is shown in figure 1. And also results of decision tree was combined with random forest to achieve final result which was most suitable decision making or prediction process in machine learning.

### *D.* SVR

Support Vector Machines (SVMs) are notable in order issues. The utilization of SVMs in relapse isn't also reported, be that as it may. These sorts of models are known as Support Vector Regression (SVR) shown in figure 2.

The issue of SVR is to observe a capacity thaplanning from an info space to genuine numbers based on a preparation test. So how about we currently plunge profound and comprehend how SVR functions as a matter of fact.

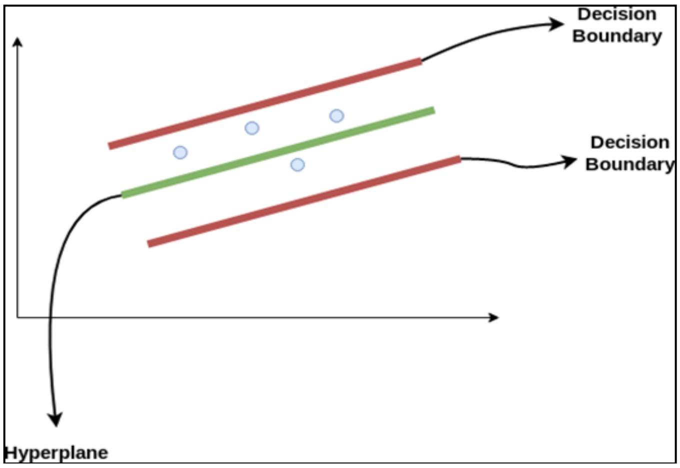

**Figure 2. Support vector machine classifier**

Consequently, we will take those focuses that are inside the choice limit and have the least blunder rate, or are inside the Margin of Tolerance. This gives us a superior fitting model [9].

SVR is a strong calculation that permits us to pick how open minded we are of mistakes, both through an OK blunder margin(ϵ) and through tuning our resistance of falling external that satisfactory blunder rate.

Ideally, this instructional exercise has shown you the intricate details of SVR and has passed on you adequately certain to add it to your demonstrating weapons store.

### *E.* LSTM

LSTM addresses Long Short-Term Memory, a model at first proposed in 1997. LSTM is a Gated Recurrent Neural Network. The key part is that those associations can store information that can be used for future cell taking care of. We can consider LSTM a RNN with some memory pool that has two key vectors:

- Short-term state: keeps the result at ongoing time step.
- Long-term state: stores, peruses, and dismisses things implied for the long haul while going through the organization.

The choice of perusing, putting away, and composing dependson some enactment work. The result from those actuate capacities is a worth between (0, 1). The disregard and result entrances pick whether to keep the oncoming new information or dispose of them. The memory of the LSTM block and the condition at the outcome entryway conveys the model decision. The outcome then, is passed to the association again as a data making a tedious gathering.

### *F.* Lasso

The word “LASSO” stands for Least Absolute Shrinkage and Selection Operator. It is a measurable equation for the regularization of information models and element choice.

Regularization is a significant idea that is utilized to keep away from infg of the information, particularly when the prepared and test information are a lot shifting.

Regularization is carried out by adding a "punishment" term to the best fit got from the prepared information, to accomplish a lesser change with the tried information and furthermore limits the impact of indicator factors over the result variable by packing their coefficients.

In regularization, what we do is ordinarily we keep similar number of highlights however diminish the size of the coefficients. We can decrease the size of the coefficients by utilizing various sorts of relapse procedures which utilizes regularization to beat this issue.

There are two principal regularization strategies, to be specific Ridge Regression and Lasso Regression. The two of them contrast in the manner they allocate a punishment to the coefficients. In this blog, we will attempt to see more about Lasso Regularization strategy. It is utilized over relapse strategy for a more exact expectation. This model purposes shrinkage.

### *G.* KNN

K Nearest Neighbor Algorithm likewise has a place with a sort of learning calculation called managed learning calculation. It tends to be utilized for characterization (most ordinarily and here and there for Regression). It is a truly adaptable calculation which can likewise be utilized for figuring missing qualities and furthermore helps in resampling datasets.

The calculation's learning is:

– An Instance based learning: Like different calculations here we don't gain from preparing information collection. All things considered, we utilize the total informational index to prepare the model and after this we utilize this model to anticipate inconspicuous information.

– We don't prepare the model in earlier as we do in different calculations. We train the model when there is prerequisite for expectations. Thus it is called as Lazy Learning.

– No predefined type of planning, Thus it is non- parametric

### *H.* Bayesian Ridge

Bayesian relapse permits a characteristic instrument to endure lacking information or ineffectively appropriated information by forming direct relapse utilizing likelihood merchants instead of point gauges. Result or reaction 'y' is accepted to drawn from a likelihood circulation instead of assessed as a solitary worth.

### *I.* Gradient Boosting (Best Model)

While concentrating on Machine Learning you probably go over this term called Boosting. It is the most confounded term in the field of Data Science. The guideline behind supporting calculations is first we fabricated a model on the preparation dataset, then a subsequent model is worked to correct the blunders present in the principal model.

When objective section is constant, we use Gradient Boosting Regressor while when it is an arrangement issue, we use Gradient Boosting Classifier. The main contrast between the two is the "Misfortune work". The goal here is to limit this misfortune work by adding powerless students utilizing slope plummet. Since it depends on misfortune work subsequently for relapse issues, we'll have different misfortune capacities like Mean squared blunder (MSE) and for characterization, we will have different for e.g log-probability. It is one of the most famous machine learning calculations for even datasets. It is sufficiently strong to track down any nonlinear connection between your model objective and elements and has extraordinary convenience that can manage missing qualities, exceptions, and high cardinality absolute qualities on your highlights with practically no unique treatment.

While you can assemble barebone slope supporting trees utilizing a few famous libraries, for example, XGBoost or LightGBM without knowing any subtleties of the calculation, you actually need to know how it functions when you begin tuning hyper-boundaries, modifying the misfortune capacities, and so on, to get better quality on your model

## VI. EXPERIMENTAL RESULTS

### *A.* Charts of different Algorithms

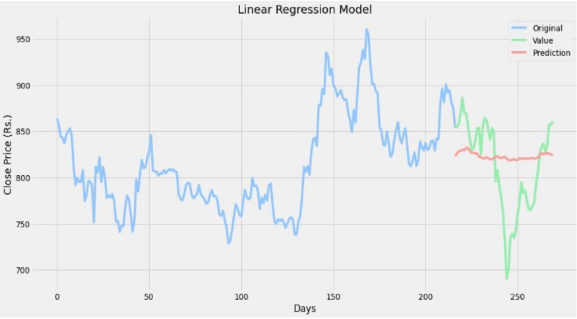

Figure 3. Predicted value of price for Linear Regression

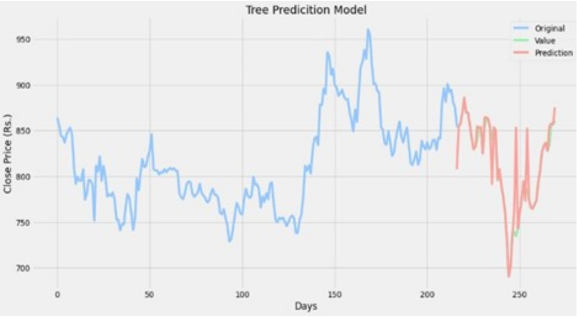

Figure 4. Predicted value of price for Decision tree

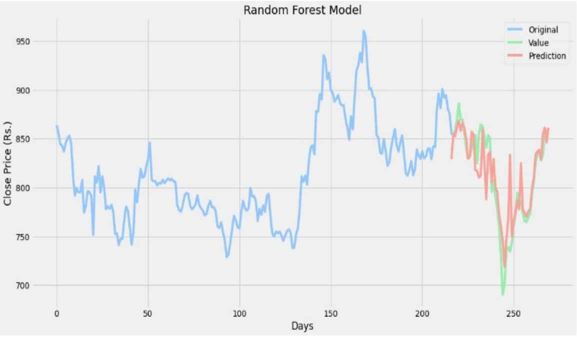

Figure 5. Predicted value of price for Random forest

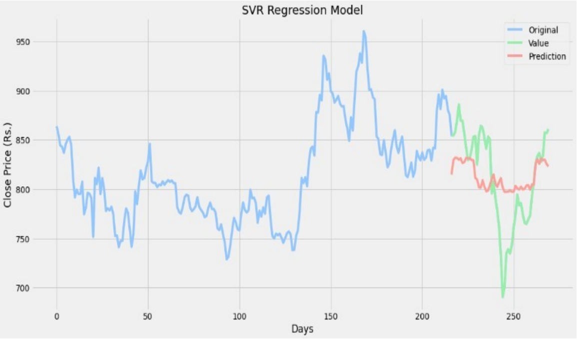

Figure 6. Predicted value of price for SVR

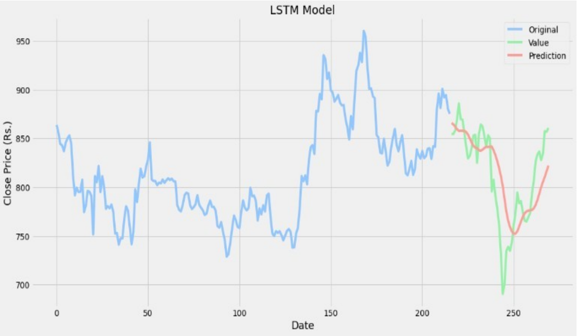

Figure 7. Predicted value of price for LSTM

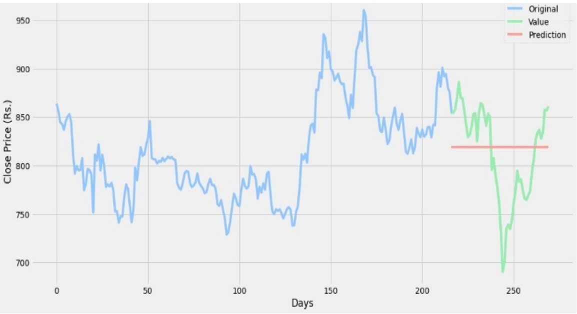

Figure 8. Predicted value of price for Lasso

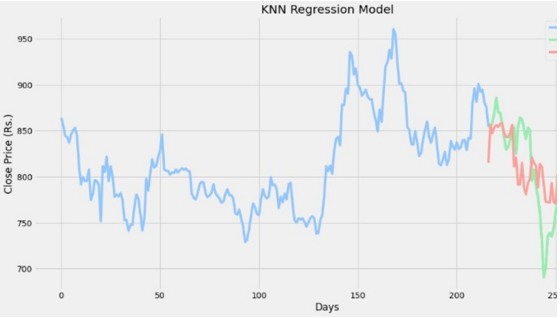

Figure 9. Predicted value of price for KNN

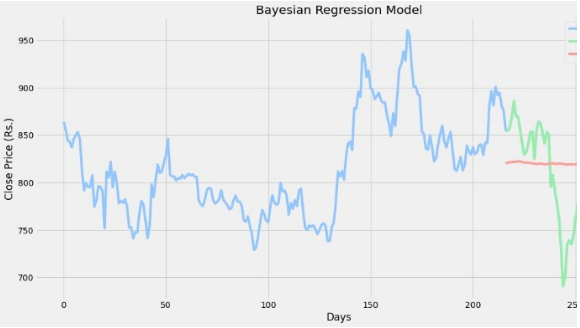

Figure 10. Predicted value of price for Bayesian Bridge

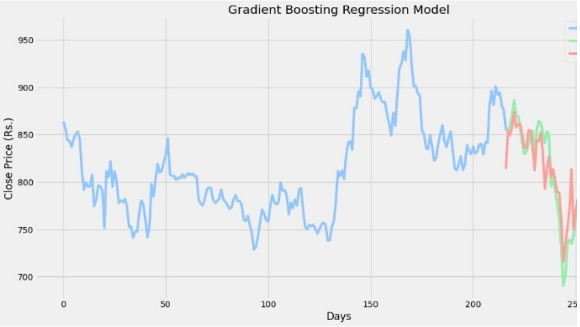

Figure11. Predicted value of price for Gradient Boosting

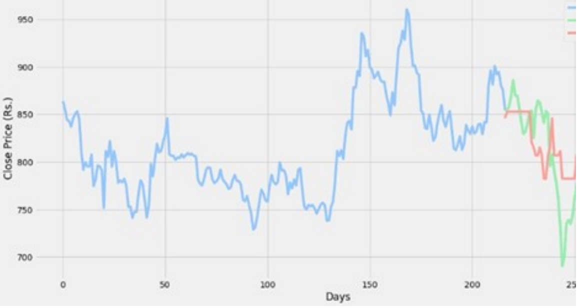

Figure 12. Predicted value of price for Ada Boost

Figure 3 to 12 shows the comparison price prediction values of different classifiers. In this, original value for previous 200 days are considered for experiment and price value for stock was predicted for next 25 days. Green line shows the originally obtained value for that 25 days and red line was the predicted valuje for the 25 days.

Table 1. R2 Score, MSE, MAPE and MAE for various classifiers

| | Model Name | R2 Score | MSE | MAPE | MAE |
|---|---|---|---|---|---|
| 1 | Linear Regression | 0.0398668 | 2231.04 | 0.0493334 | 38.6686 |
| 2 | Decision Tree | 0.803873 | 445.736 | 0.00876218 | 6.91389 |
| 3 | Random Forest | 0.761005 | 555.348 | 0.0190038 | 15.1189 |
| 4 | SVR | 0.281364 | 1669.88 | 0.0417576 | 33.1698 |
| 5 | LSTM | 0.505659 | 1148.69 | 0.0329098 | 25.958 |
| 6 | Lasso | -0.0202788 | 2370.8 | 0.0524584 | 41.392 |
| 7 | KNN | 0.445309 | 1288.92 | 0.0338929 | 26.9196 |
| 8 | Bayesian Ridge | -0.00337781 | 2331.52 | 0.0516967 | 40.7282 |
| 9 | Gradient Boosting | 0.810805 | 439.628 | 0.0193799 | 15.4267 |
| 10 | Ada Boost | 0.465004 | 1243.16 | 0.0349588 | 27.6497 |

Table 1 represents the consolidated values of the experimental results obtained for the ten different machine learning approaches. The individual graph for all approaches shown as individual graph. Based on existent studies R square score was more than 0.5 then that considered as better model. With that note linear regression, lasso, Bayesian ridge, SVR and Ada Boost may be considered as the lowest value. The other three parameters may will project that the model is not good if the value of that parameter is high. Based on these parameters, the same models lasso, linear regression and Bayesian ridge obtained highest value than others which shows that these models are not suitable for the predicting the given dataset. The best models are decision tree, random forest, LSTM and Gradient boosting based on the experimental values received.

**R2 Score:**
The R2 score shifts somewhere in the range of 0 and 100 percent. It is firmly connected with the MSE (see beneath), yet not the equivalent. Wikipedia characterizes R2 as

" … the extent of the difference in the reliant variable that is unsurprising from the free variable(s)."

One more definition is "(absolute fluctuation made sense of by model)/all out difference." So assuming it is 100 percent, the two factors are impeccably associated, i.e., with no change by any means. A low worth would show a low degree of connection, meaning a relapse model that isn't legitimate, however not in all cases.

Formula $R^2 = 1 - \frac{RSS}{TSS}$

$R^2$ = *coefficient of determination*
*RSS* = *Sum of squares of residuals*
*TSS* = *Total sum of squares*

Negative R² can be,

1. Model is not learning the trend that is present in train data.
2. Too little data has been used to evaluate the model compared to train data.
3. Too many outliers are present in the dataset.

**MSE:**
Mean square error (MSE) is the normal of the square of the mistakes. The bigger the number the bigger the error. Error for this situation implies the distinction between the noticed qualities y1, y2, y3, … and the anticipated ones pred(y1), pred(y2), pred(y3), … We square every distinction (pred(yn) - yn)) ** 2 so that negative and positive qualities don't counter balance one another.
Formula

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

*MSE* = *mean square error*
*n* = *number of data points*
$Y_i$ = *Observed values*
$\hat{Y}_i$ = *Predicted values*

**MAPE:**
The mean absolute percentage error (MAPE) is a proportion of how exact a figure framework is. It estimates this precision as

a rate, and can be determined as the normal outright percent mistake for each time-frame short real qualities separated by genuine qualities.

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

It is the most widely recognized measure used to estimate blunder, and works best assuming there are no limits to the information (and no zeros).

**MAE:**

In insights, mean absolute error (MAE) is a proportion of blunders between matched perceptions communicating a similar peculiarity. MAE is determined as

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}$$

It is thus an arithmetic average of the absolute errors

$|e_i| = |y_i - x_i|$ [11]

where yi is the prediction and xi the true value.

Note that elective definitions might incorporate relative frequencies as weight factors. The mean outright mistake involves similar scale as the information being estimated. The mean outright mistake is a typical proportion of figure blunder in time series analysis, here and there utilized in disarray with the more standard meaning of mean outright deviation. A similar disarray exists more generally.

## VII. Conclusion

Money related business areas give an extraordinary stage to monetary supporters and vendors, who can trade from any gadget that partners with the web. Expecting monetary trade is an inciting task in light of dependably changing stock characteristics which are dependent upon various limits which structure complex models. The overall assessment taking into account R2 Score, MAPE, MSE, MAE regards evidently show that Gradient Boosting gives better figure of stock costs when stood out from various models. The data by means of electronic amusement stages can either be delivered by individuals or bots. The assessments of bots can at times achieve wrong figures. The precision of the protections trade assumption structure can be furthermore improved by utilizing a Sentiment assessment anyway Machine Learning.

## *References*

[1] Nayak, Aparna, MM Manohara Pai and Radhika.M.Pai."Prediction models for Indian stock market." Procedia Computer Science 89 (2016):441-449.

[2] Reshma.R., Usha Naidu, V.Sathiyavathi, and L.SaiRamesh. "Stock Market Prediction Using Machine Learning Techniques." Advances in Parallel Computing Technologies and Applications, vol.40, pp.331-340, (2021).

[3] Vijh, Mehar, Deeksha Chandola, Vinay Anand Tikkiwal, and Arun Kumar. "Stock closing price prediction using machine learning techniques." Procedia computer science 167 (2020): 599-606.

[4] Mehtab, Sidra, Jaydip Sen, and Abhishek Dutta. "Stock price prediction using machine learning and LSTM-based deep learning models." In Symposium on Machine Learning and Metaheuristics Algorithms and Applications, pp. 88-106. Springer, Singapore, 2020.

[5] Abdulhamit Subasi, Faria Amir, Kholoud Bagedo, Asmaa Shams, Akila Sarirete, "Stock Market Prediction Using Machine Learning" Procedia Computer Science, Volume 194Issue C2021 pp 173–179https://doi.org/10.1016/j.procs.2021.10.071

[6] Ghosh, Pushpendu, Ariel Neufeld, and Jajati Keshari Sahoo. "Forecasting directional movements of stock prices for intraday trading using LSTM and random forests." Finance Research Letters 46 (2022): 102280.

[7] https://ideas.repec.org/a/eee/phsmap/v555y2020ics037843720301837.html
[8] https://towardsdatascience.com/ introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a

[9] https://www.analyticsvidhya.com/blog/2020/03/support - vector-regression -tutorial-for-machine-learning/

[10] https://towardsdatascience.com/understanding-adaboost- 2f94f22d5bfe

[11] https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70
[12] SA, Sadhana, S, Sabena, L, SaiRamesh, and A. Kannan. "Customer's opinion mining from online reviews using intelligent rules with machine learning techniques." *Concurrent Engineering* (2022): 1063293X221120084.
[13] Anand, C. "Comparison of Stock Price Prediction Models using Pre-trained Neural Networks." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 3, no. 02 (2021): 122-134.
[14] Andi, Hari Krishnan. "An Accurate Bitcoin Price Prediction using logistic regression with LSTM Machine Learning model." Journal of Soft Computing Paradigm 3, no. 3
(2021): 205-217.

[15] Sabena, S., S. Kalaiselvi, B. Anusha, and L. Sai Ramesh. "An Multi-Label Classification with Label Correlation." *Asian Journal of Research in Social Sciences and Humanities* 6, no. 9 (2016): 373-386.