Stock Prediction Using ML

Anant Singhal, Hitanshu Yadav, Akshat Singh, Mali Mandar Mukund and Aviral Jain

School of Computer Science and Engineering, VIT Bhopal University

Abstract

Stock Market has always been a hot spot for investors and companies to grasp the change regularity of the stock market and predict its trends. Predictions on stock market prices are a great challenge due to the fact that it is an immensely complex. There are many studies aiming to take on the of predicting the trends of stock market. This article studies the usage of Linear Regression, SVM (Supported Vector Machine), LSTM (Long Short Term Memory), Random Forest and GBM (Gradient Boosting Machine) to predict the future trends of stock market based on the price history. To achieve this goal, five different prediction model was built with the help of above mentioned algorithms. The models were tested with data of five different companies and were compared to each for better accuracy. The result that were obtained are promising, getting up to an average of 49.6898% accuracy when predicting if the price of a particular stock is going to go up or not in the near future.

1. INTRODUCTION

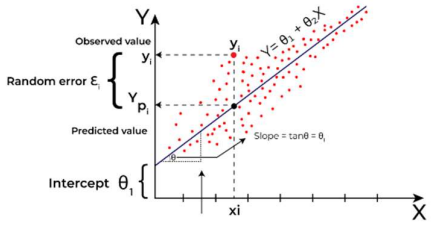
Predictions of stock trends has been proven a very difficult task, due to its inherit complex and chaotic nature. The number of variable to be considered are immense and the signal to noise ratio very insignificant. The stock prices are influenced by many factors such as politics, economy, society and market. This makes the task of predicting stock market prices behaviour in the future a very hard one.

The forecast of the stock market is directly related to the profits of the investor. The more accurate the forecast, the more effectively it can avoid risks and invest the company which can give better profits. Stock Prediction also plays an important role in the research of a country's economic development. Therefore, the research on the intrinsic value and prediction of the stock market has great theoretical significance and wide application prospects.

There has been a huge amount of studies towards this topic, presenting us with a variety of approaches to reach the goal. There are numerous variables effecting the stock prices, including geopolitical events, macroeconomic indicators, investor sentiment, and technological advancements. These factors create an environment where forecasting future price movements becomes impossible to predict accurately, like navigating a maze without a map.

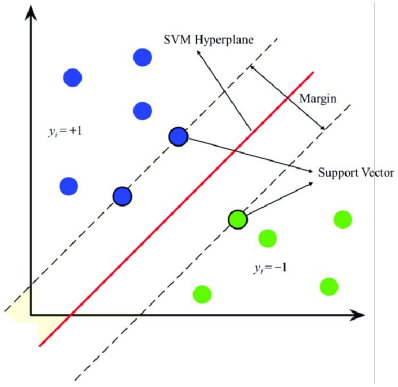
2. BACKGROUND AND RELATED WORK

First of all, we have taken one of the most basic Machine learning algorithm, Linear Regression. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The goal is to find the best-fitting line that minimizes the sum of squared differences between the observed and predicted values. The equation of the line is represented as $y = mx + b$, where y is the dependent variable, x is the independent variable, m is the slope, and b is the y-intercept.



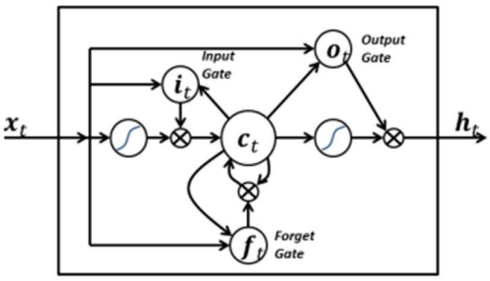
(Fig. 1. Linear regression)

Second model we have taken is SVM (Supported Vector Machine). SVM is a supervised machine learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates data points into different classes or predicts continuous outcomes. SVM aims to maximize the margin between classes while minimizing classification errors. It achieves this by identifying support vectors, which are data points closest to the decision boundary. SVM can handle linear and non-linear relationships. It's effective for high-dimensional data and offers robust performance even with limited training samples.



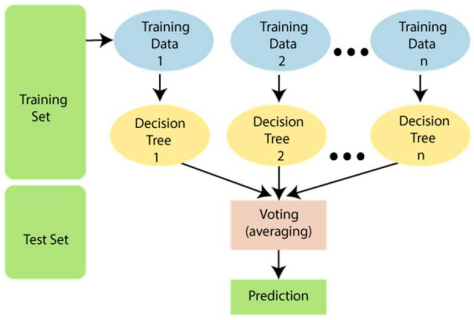
(Fig. 2 SVM)

Third algorithm we are using is LSTM (Long Short Term Memory). LSTM is a type of recurrent neural network architecture designed to model sequential data and overcome the vanishing gradient problem. It utilizes a memory cell with gated units, including input, output, and forget gates, to retain information over long sequences. LSTM can capture long-range dependencies and handle time lags in data, making them well-suited for tasks such as time series prediction or stock prediction. LSTM can effectively learn and remember patterns in sequential data, making them a powerful tool for sequential modelling and prediction tasks.



(Fig. 3 LSTM)

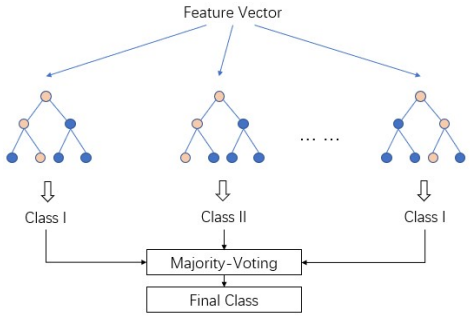
Forth model is implemented using Random Forest. Random Forest is a powerful machine learning algorithm that constructs multiple decision trees during training and combines their predictions through averaging or voting. It operates by randomly selecting subsets of features and data points for each tree, reducing overfitting and increasing generalization. This ensemble method is robust against noise and outliers, making it suitable for various tasks, including classification and regression. Its simplicity, scalability, and ability to handle high-dimensional data make Random Forest a popular choice for both beginners and experts in the field of machine learning.



(Fig. 4 Random Forrest Algorithm)

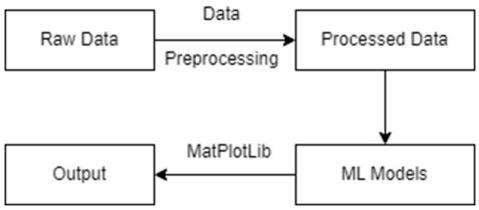
The Final Algorithm used is GBM (Gradient Boosting Machine). It is a powerful ensemble learning technique that builds a strong predictive model by sequentially adding weak learners, usually decision trees, to correct the

errors of the previous models. It minimizes a loss function by optimizing the model's predictions against the true values. GBM trains each subsequent model to focus on the mistakes made by the previous ones, gradually improving accuracy. GBM is highly effective in handling complex relationships in data and is less prone to overfitting. It's widely used in regression and classification tasks due to its robustness and performance.



3. METHODLOGY AND DEVELOPMENT

A data flow diagram was designed to represent the flow of raw data to the final output, going through the machine learning algorithms to achieve the goal.



1. Data Processing

The data of five companies are gathered in form of a csv (Comma Separated value) files. The data of TCS, HDFC Bank, Asian Paints, Maruti Suzuki and ONGC which are part of Nifty 50 Index. The data used is of approximately twenty years i.e. 2001 to 2021.

In order to reduce random variation and noise, exponential smoothing was performed to make analysing data easier. The closing price of each stock at the end of each day was used for predictions.

The Pre-processing of the data was completed with the help of NumPy and Pandas.

The design methodology is centred around leveraging machine learning to provide accurate predictions of stock market movements. The methodology follows a structured process:

- **User Research**: Understanding the needs and preferences of investors through data analysis and feedback collection.
- **Ideation**: Generating ideas for project features and functionalities based on user research insights.
- **Prototyping**: Creating low-fidelity and high-fidelity prototypes of the user interface to visualize user interactions.
- **User Testing**: Gathering feedback from users to iterate on design and functionality improvements.

The working formula of the used models are:

- Linear Regression
  Y = a + bX
  Where, X is the independent variable plotted along x-axis, Y is the dependent variable plotted along the y-axis, b is the slope line and a is the intercept.

- SVM (Supported Vector Machine)

$$\vec{X}.\vec{w} - c \geq 0$$

$$\text{putting } -c \text{ as } b, \text{we get}$$

$$\vec{X}.\vec{w} + b \geq 0$$

$$\text{hence}$$

$$y = \begin{cases} +1 & \text{if } \vec{X}.\vec{w} + b \geq 0 \\ -1 & \text{if } \vec{X}.\vec{w} + b < 0 \end{cases}$$

- LSTM (Long Short Term Memory)

$$f_t = \sigma_g\,(W_f \times x_t + U_f \times h_{t-1} + b_f)$$

$$i_t = \sigma_g\,(W_i \times x_t + U_i \times h_{t-1} + b_i)$$

$$o_t = \sigma_g\,(W_o \times x_t + U_o \times h_{t-1} + b_o)$$

$$c'_t = \sigma_c\,(W_c \times x_t + U_c \times h_{t-1} + b_c)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot c'_t$$

$$h_t = o_t \cdot \sigma_c(c_t)$$

f_t *is the forget gate*
i_t *is the input gate*
o_t *is the output gate*
c_t *is the cell state*
h_t *is the hidden state*

- Random Forest

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2$$

Where *N* is the number of data points, *fi* is the value returned by the model and *yi* is the actual value for data point *i*.

- GBM (Gradient Boosting Machine)

$$\begin{aligned} L &= -\left[\sum_{i=1}^{n} y_i \log(p) + (1-y_i) log(1-p)\right] \\ &= -y*\log(p) - (1-y)*\log(1-p) \\ &= -y*log(p) - log(1-p) + y*log(1-p) \\ &= y*\left[log(p) - log(1-p)\right] - log(1-p) \\ &= -y*\left[\frac{\log(p)}{\log(1-p)}\right] - log(1-p) \\ &= -y*log\left(\frac{p}{1-p}\right) - log(1-p) \end{aligned}$$

4. EVALUATION

This study adopts a rolling window approach, constructing a new model daily with fresh training and validation data. Training utilizes the preceding 10 months of trading data, while validation employs the preceding week's data. Subsequently, predictions are made using the most recent model each day.

Given the temporal nature of the data, various supervised learning algorithms were considered. Among them are Linear Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting Machine (GBM), and Long Short-Term Memory (LSTM) neural networks. Each model was evaluated for its ability to classify input data considering prior instances.

Google's TensorFlow, Scikit-Learn (SVM & ensemble) and XGBoost was utilized for model development. For instance, the LSTM model comprises an input layer incorporating technical indicators and pricing data, feeding into an output layer with sigmoid activation. Other models were similarly structured with appropriate input features and activation functions.

Numerous experiments were conducted using these models, with results averaged across selected stocks from the Nifty 50 stock data for the years 2000 - 2021 (TCS, HDFC, Asian Paints, Maruti, and ONGC).

Evaluation of model performance involved assessing algorithmic performance and financial outcomes, compared against established baselines. Metrics such as accuracy, precision, recall, and F-measure were employed, calculated based on true positives (-*tp*), true negatives (-*tn*), false positives (-*fp*), and false negatives (-*fn*).

A.Trading operations

When the predicted class is ”1”, in other words, in the case that the network predicts that the stock price will go up, then the strategy is to open a “buy” position on the current moment and close it on j. In that case, profit was defined as close(j)−close(i).

The financial results were calculated based on hypothetical trades of sets of a hundred stocks per operation disregarding costs and taxes.

B. Baselines

For comparison, the baselines chosen are based on other classical machine learning algorithms in addition to other simplistic investment methods.

The machine learning methods consist of approaches that are traditional and widely used but less complex than the one this project is based on. Using the exact same input, trying

to do the same predictions, the models chosen were Linear Regression, SVM, LSTM, Random Forest, and Gradient Boosting Machine Model.

The other baselines are the following investment strategies:

- Buy and hold: Buy at the first time step and sell at the latest.
- Optimistic: If prices went up on the previous time step, then perform a buying operation and sell it on the following step.
- Pseudo-random (following class distribution): Decides whether or not to perform a trading operation based on probabilities according to the class distribution.

C. Empirical Results

Experiments were carried out to predict price movement using the stock data from the past 20 years.

Table I to IV present a characterization of the Accuracy, Root mean square Error, mean square error and the Absolute mean error, for each stock data using the five ML models mentioned earlier.

Fig 1,2,3,4 Show the plots for Result metrics of each model when applied on the five stocks.

Table I: Result Metrics (Accuracy)

Stock	L.R	SVM	LSTM (Tolerance)	R.F	GBM
TCS	0.7170	-0.04705	50%	0.9997	0.9982
HDFC	-0.0011	-0.0015	38.89	0.9999	0.9885
Asian.p	0.4543	-0.0332	0.00%	0.9997	0.9845
Maruti	0.3774	-0.1019	27.78%	0.9991	0.9968
ONGC	0.0115	-0.0032	27.78%	0.9998	0.9869

Table II: Result Metrics (RMSE)

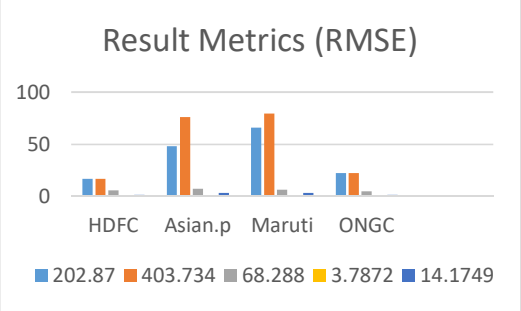
Stock	L.R	SVM	LSTM	R.F	GBM
TCS	250.562	482.025	99.713	7.2127	19.692
HDFC	23.544	23.5490	9.02	0.0994	2.054
Asian.p	59.508	81.8890	7.784	1.2778	10.106
Maruti	78.647	104.6340	8.068	2.7887	5.6970
ONGC	29.49	29.7090	6.547	0.2925	2.9430

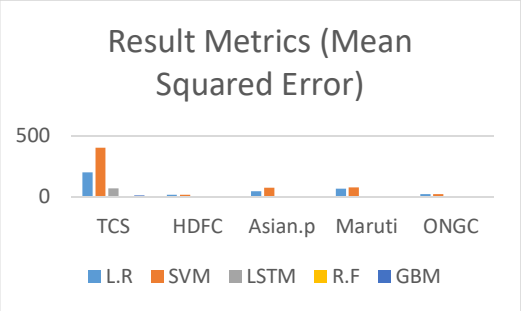
Table III: Result Metrics (Mean Squared Error)

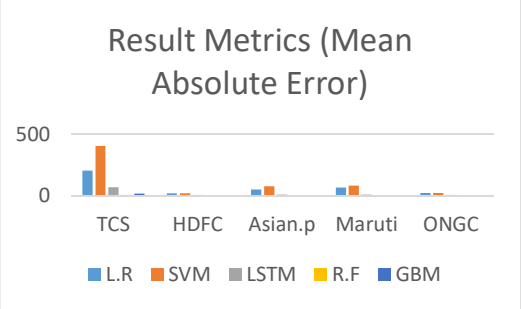
Stock	L.R	SVM	LSTM	R.F	GBM
TCS	62781.26	232347.87	9942.64	52.022	387.76
HDFC	554.3425	554.5718	81.3592	0.0098	4.2176
Asian.p	3541.201	6705.829	60.5936	1.6327	102.13
Maruti	6185.360	10948.29	65.0955	7.7769	32.45
ONGC	869.6482	882.6426	42.8644	0.0855	8.659

Table IV: Result Metrics (Mean Absolute Error)

Stock	L.R	SVM	LSTM	R.F	GBM
TCS	202.870	403.734	68.288	3.7872	14.1749
HDFC	16.9975	16.9471	5.7518	0.0696	1.5568
Asian.p	48.2625	76.2294	7.4659	0.3727	3.3138
Maruti	65.9406	79.7430	6.7241	1.1135	3.4061
ONGC	22.5264	22.3264	4.8589	0.1458	1.7786







5. CONCLUSIONS

This article represents a significant endeavour aimed at leveraging machine learning to provide insights into stock market trends and predicting the trends of stock market. We can observe that the implemented models can give

a very promising outcome with a few exceptions here and there.

Although the input dimension is very large, the algorithms used have demonstrated acceptable capabilities to achieve our goal of providing insights of stock market.

We have used RMSE (Root Mean Squared Error), MSE (Mean Squared Error) and MAE (Mean Absolute Error) as the performance metrics.

Some models have displayed a considerable gain in terms of accuracy than others, but other can have lower variance and would contribute to a more reliable model.

For example, Random Forest and GBM has shown accuracies of upwards of 99%, whereas Linear Regression fall behind way much.

Forthcoming Research

We intend to keep investigating ways to improve our models and its predictions by studying changes on the neural network architecture and different approaches for pre-processing the raw data as well as using new techniques to improve the accuracy for better predictions.

Given what we could observe so far, the key to get to better prediction results lies in improving the input like it is the exact same case for many other Machine Learning problems.

We also intend to evaluate the models using different and more realistic trading strategies, instead of simply buying and selling after a fixed amount of time. And also take into account for essential components of stock markets, like timing, execution booking and associated transactions costs.

ACKNOWLEDGMENT

The research for this article was supported by VIT Bhopal University.

We would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

We wish to express our heartfelt gratitude to Dr. S. Poonkuntran, Head of the Department, School of Computing Science and Engineering for much of his valuable support and encouragement in carrying out this work.

We would like to thank our internal guide, Dr. RAJNEESH KUMAR PATEL, for continually guiding and actively participating in our project, and giving valuable suggestions to complete the project work.

REFERENCES

1. “Machine Learning Stock Market Prediction Studies: Review and Research Directions” by Troy J. Strader, John J. Rozycki, Thomas H. Root and Yu-Hsiang John Huang.
2. “Stock market prediction using machine learning techniques” by Mehak Usmani, Syed Hasan Adil, Kamran Raza and Syed Saad Azhar Ali, published in 2016.
3. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by A. Géron.
4. “Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions” by Nusrat Rouf ,Majid Bashir Malik ,Tasleem Arif ,Sparsh Sharma ,Saurabh Singh ,Satyabrata Aich and Hee-Cheol Kim, available at https://www.mdpi.com/2079-9292/10/21/2717.
5. "Predicting Stock Market Trends Using Machine Learning Techniques" by S. Kumar.
6. “Stock Market Forecasting Using Machine Learning Algorithms” by Shunrong Shen, Haomiao Jiang and Tongda Zhang, available at https://masters.donntu.ru/2015/fknt/pozhydaev/library/article3.pdf.

7. “Stock market prediction using machine learning classifiers and social media, news” by Wasiat Khan, Mustansar Ali Ghazanfar, Muhammad Awais Azam, Amin Karami, Khaled H. Alyoubi & Ahmed S. Alfakeeh, in March 2020.
8. "Machine Learning for Stock Market Prediction: A Comprehensive Review" by A. Gupta.
9. “Stock Market’s Price Movement Prediction with LSTM Neural Networks” by David M. Q. Nelson, Adriano C. M. Pereira, Renato A. de Oliveira.
10. “Stock Market Prediction with High Accuracy using Machine Learning Techniques” by Malti Bansal, Apoorva Goyal and Apoorva Choudhary, available at https://www.sciencedirect.com/science/article/pii/S1877050922020993.
11. “Study on the prediction of stock price based on the associated network model of LSTM” by Guangyu Ding and Liangxi Qin , available at Study on the prediction of stock price based on the associated network model of LSTM | International Journal of Machine Learning and Cybernetics (springer.com)
12. “A systematic review of stock market prediction using machine learning and statistical techniques” by Deepak Kumar, Pradeepta Kumar and Sarangi Rajit, in 2022.
13. “A Machine Learning Model for Stock Market Prediction” by Osman Hegazy, Omar S. Soliman and Mustafa Abdul Salam, in 2013, available at https://arxiv.org/pdf/1402.7351.
14. “Stock market prediction: A big data approach” by Girija V Attigeri, Manohara Pai M M, Radhika M Pai and Aparna Nayak, in 2015, available https://ieeexplore.ieee.org/abstract/document/7373006.
15. “Stock Market Analysis using Supervised Machine Learning” by Kunal Pahwa and Neha Agarwal, in 2019, available at https://ieeexplore.ieee.org/abstract/document/8862225.