International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Noble Forecasting Technique for Time Series Data Analysis Using Machine Learning

V V Sesi Kumar

Student, Department of Information Technology, GMR Institute of Technology, Rajam, India

ABSTRACT:

Time series data is the data which changes over a period of time .Stock price prediction is one of the most browsed topics in that time series analysis. A stock market is a public market where you could buy as well as sell the shares of a publicly enrolled company. A stock(most commonly known as shares) is a primary stake that shows the authority of a particular individual or a group of people. If a person has more than half of the shares of a company then he is primary key holder of that company .Stock price prediction involves predicting the future stock values of a company based on its present trend as well as the earlier activities .A better stock prediction enhances the investors and other stake holders in enhancing their company . This prediction is achieved by using some machine learning algorithms . The machine learning algorithms that we use in this paper are Logistic Regression ,Support Vector Machine(SVM) ,Decision Tree ,Random Forest ,Long Short Term Memory(LSTM) .The above Mentioned algorithms are used because they provide more accurate results than the remaining machine learning algorithms .Here, this paper provides an overview of the applications of machine learning in stock market forecasting to determine what can be done in the future.

Keywords: stock price prediction, machine learning, Support Vector machine(SVM) ,Long Short Term Memory(LSTM) ,Logistic Regression ,stock market

INTRODUCTION:

Time series data analysis is the analysis of the datasets that change over a period of time .In this prediction of stocks is one of the subsets ,coming to the Stock price prediction it is one of the most widely studied topics and it is gaining more attraction of researchers as it is one of the fields where one can make huge profits in marketing area by predicting the stock values of a particular company . From several years the stock price prediction is under the limelight as it produces significant profits. The stock values of the company is a dynamic one i.e for every now and then the stock values changes .Many researchers have worked on this topic and used various machine learning algorithms to predict the stock values .Various models are introduced in both industry as well as in academia for the stock price prediction from machine learning algorithms to data mining and then to the latest one is the deep learning techniques .Our paper mainly focuses on the machine learning algorithms, in that we mainly use the supervised machine learning algorithms because of the results they produce in prediction. However ,it is difficult to analyse which algorithm is best to predict the values but according to the latest trends Logistic Regression ,Support vector machine(SVM) one of the supervised machine learning algorithms which produces more accurate results ,best computational time and also we can use Long Short Term memory(LSTM) and also the regression models are also used by various researchers which also provides the promising results .

This prediction is used in our day to day life for example now a days there are many apps like Upstox, Groww, stock edge where many investors have to predict the stock prices if their prediction goes in their way they can make the enormous profits . Machine learning makes the prediction by training and testing the models with the historical datasets, social media data, crawled financial news or trends.

Stock market plays an vital role in our life and the reasons are listed below:

1. Stock market helps companies raise capital
2. Stock market helps create personal wealth
3. It helps increase investment in the economy
4. Market serves as an indicator of the state of the economy
5. Stock Market also affects non-investors in the economy

LITERATURE SURVEY:

[1]. M. Nabipour, P. Nayyeri, H. Jabani, S. S. and A. Mosavi, "Predicting Stock Market Trends Using Machine Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis," in IEEE Access, vol. 8, pp. 150199-150212, 2020, doi: 10.1109/ACCESS.2020.3015966. This study compares nine machine learning models (Decision Tree, Random Forest, Adaptive Boosting (Adaboost), eXtreme Gradient Boosting (XGBoost), Support Vector Classifier (SVC), Naïve Bayes, K-Nearest Neighbors (KNN), Logistic Regression and Artificial Neural Network (ANN)) and two powerful learning methods (Recurrent Neural Network (RNN) and Long short-term memory (LSTM).Four stock market groups, namely diversified financials, petroleum, non-metallic minerals and basic metals from Tehran stock exchange, are chosen for experimental evaluations.We have approached to take our input values in two forms they are continuous data and the binary data .After comparing the nine machine learning algorithm models .Logistic Regression, Random Forest ,Support vendor machine(SVM) have shown promising results where as the LSTM model precisely gave the more accurate results accounting more than 90% accuracy compared to the remaining algorithms .

[2]. Z. K. Lawal, H. Yassin and R. Y. Zakari, "Stock Market Prediction using Supervised Machine Learning Techniques: An Overview," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411609. This paper reviews studies on supervised machine learning models in stock market predictions. The study discussed how supervised machine learning techniques are applied to improve accuracy of stock market predictions. Here they used Support Vector Machine(SVM) algorithm and found that this algorithm produce more accurate results than the other algorithms like Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Naïve Bayes, Random Forest, Linear Regression and Support Vector Regression (SVR) .This study showed that the classification algorithm in the supervised machine learning is mostly used to predict the stock prices ahead of the regression algorithm .In the 38 case studies mentioned in this paper 24 studies have used the SVM algorithm to predict the stock prices.

[3]. X. Yuan, J. Yuan, T. Jiang and Q. U. Ain, "Integrated Long-Term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market," in IEEE Access, vol. 8, pp. 22672-22685, 2020, doi: 10.1109/ACCESS.2020.2969293.This Paper Used Six machine learning models including Logistic Regression (LR), Support Vector Machine (SVM), k-NearestNeighbor (KNN), Random Forest (RF), Gradient Boosting Decision Tree (GBDT) and Long Short-term Memory (LSTM) are used to predict the direction of the closing price.To evaluate the performance of the prediction models, two commonly used evaluation criteria are used in this study: Accuracy, and F1 score. Accuracy is used to evaluate the overall classification ability of the model. F1 is an indicator used in statistics to measure the accuracy of a binary model, which also considers the accuracy and recalls of the classification model. We can see that RF and GBDT have good predictive ability in most cases in short-term prediction. Although SVM will take too long time to process big data, the prediction level is still significant in some cases. One of the machine learning methods, the SVM method, is not suitable for predicting large-scale stock data. We intend to incorporate more suitable machine learning methods for prediction, such as reinforcement learning methods, into the ensemble model.

[4]. Diviesh Chaudhari, Pranay Bafna, Shubham Jadhav, Jay Chaudhari, Prof. V.T. Patil, Prof. V.O.Patil",” Stock Market Prediction Using Machine Learning Algorithm”, International Journal of Scientific Research & Engineering Trends Volume 8, Issue 3, May-Jun-2022.This Paper has taken dataset from about fifty stocks of Indian National Stock Exchange's NIFTY 50 index was taken, by collecting stock data from January 1, 2013, to December 31, 2018, and lastly by the calculation of some technical indicators.It is reported that the average accuracy for the prediction of the trend of fifty stocks obtained by support vector machine is 87.35%, perceptron is 75.88%, and logistic regression is 86.98%. Since the stock data are time series data, another dataset is prepared by reorganizing previous dataset into the supervised learning format which improves the accuracy of the prediction process which reported the results with support vector machine of 89.93%, perceptron of 76.68%, and logistic regression of 89.93%, respectively .By the above results we can say that for the time series data we use SVM(support vector machine) which provide more accurate results than any other supervised machine learning algorithms.

[5]. S.Chen and C. Zhou, "Stock Prediction Based on Genetic Algorithm Feature Selection and Long Short Term Memory Neural Network," in IEEE Access, vol. 9, pp. 9066-9072, 2021, doi: 10.1109/ACCESS.2020.3047109.In this paper, the multi-factor model is introduced into stock forecasting. The stock market has a large number of stock factors which describe the change of stock price. In this research, a large number of typical stock factors are selected. However, typicality does not mean that it can be applied to all. It is basically GA-LSTM model in which Chinese stock data is taken for forecasting.GA is proposed for feature selection to select the factors more suitable for the current scene. And combined with LSTM deep learning network model, the complex nonlinear relationship between factors and stocks is mined to predict stock prices. Although the stock price prediction model proposed in this paper can effectively improve the prediction accuracy and has strong robustness, there are still some shortcomings as follow: Firstly, we only use Chinese stock data for experiment, so further research can include data from different stock markets .Secondly, in the design of model parameters in this paper, trial and error is usually adopted instead of systematic method to find the optimal size of parameters, such as the selection of number of factors.

Introduction to Machine Learning:

Machine learning is one of the subfields of artificial intelligence(AI) . Machine learning is defined as a science of programming machines to think and act like humans or it can be defined as branch of the Artificial Intelligence (AI) which mimics the human intelligence .

Types of Machine Learning :

Based on the types of the inputs given and type of techniques used to implement a program in the Machine Learning ,It is categorized into 4 types

(1) ***Supervised Learning***

(2) ***Unsupervised Learning***

(3) **Reinforcement Learning**

(4) **Semi-Supervised Learning**

Supervised Machine Learning:

Here we give Labelled input data to the machines and based on that input data the machines will predict the current request of the user. For example to differentiate between a dog and a cat .we give Hundreds of dogs &cats examples like their pictures and their characterstics to the machines as input and now the machine will understand the difference between a dog and a cat and it classifies the dogs and cats based on the input provided by the user.

Classification and Regression are 2 types of techniques that are used in the supervised Machine Learning .Some of the famous algorithms used in this type of learning to predict the model are Linear Regression, Decision-Tree, Random Forest ,Support Vector Machine(SVM) ,Naïve-Bayes etc.,

Unsupervised Machine Learning:

Here we don't give any Labelled input to the machines .Consider the above Supervised Example as stated in the previous example instead of giving specific details to the machines we give all the pictures of dogs and cats to the machine and the machine will now classify the dogs and cats by some parameters like colour, size, shape etc., from the huge amount of input data given to the machines .

Clustering & Association are the types of the techniques used in the Unsupervised Learning. K-means ,Hierarchical ,Genetic are some of the algorithms that are used in the Unsupervised Learning

Reinforcement Learning:

Here the users teach the machines where it made a mistake and make better decisions in future by Reinforcement learning .

Semi-Supervised Learning:

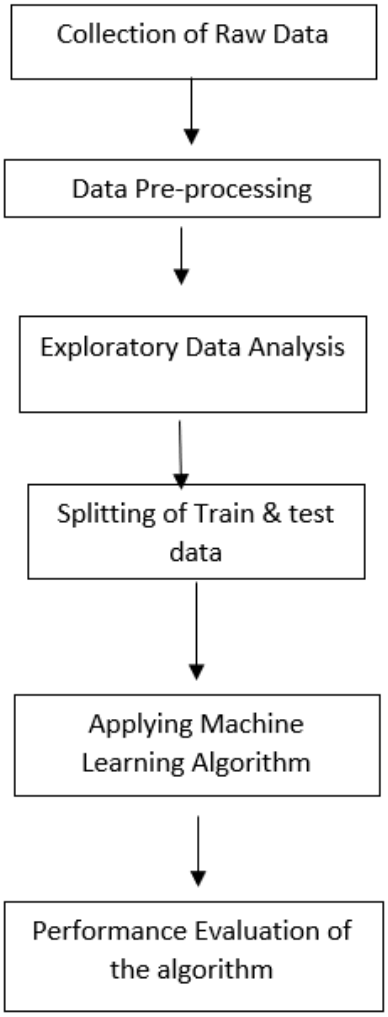
Semi-Supervised learning is a type of Machine Learning algorithm that represents the intermediate ground between Supervised and Unsupervised learning algorithms. It uses the combination of labeled and unlabeled datasets during the training period.

Applications of Machine Learning :

1. Virtual Personal Assistants like Google Assistant, Alexa ,Siri
2. Traffic predictions using Google maps
3. Email spam filtering
4. Online fraud detection

5. Stock Market Trading(Widely used)
6. Health diagnosis (Disease Identification, Personalised Treatment)
7. Automatic Translation (Bill Boards)

METHODOLOGY:



Data Collection:

Data collection is the process of gathering data from various offline and online sources by scraping, collecting, and loading it. The most challenging aspect of a machine learning project, especially when done at scale, can be large volumes of data collection or data production.

Additionally, all datasets have shortcomings. This is why the machine learning process relies so heavily on data preparation. Data preparation, in a nutshell, is a set of procedures for enhancing the machine learning capabilities of your dataset. In a larger sense, selecting the most effective data collection method is part of data preparation. And the majority of machine learning time is spent using these methods. The creation of the initial algorithm can take months!

In order to employ data analysis to discover recurrent patterns, data collecting enables you to record a record of prior events. Utilizing machine learning techniques, you may create prediction models from these patterns that track trends and foretell future changes.Effective data collection techniques are essential to creating high-performing predictive models since predictive models are only as good as the data on which they are based. For the work at hand, the data must be accurate and contain pertinent information. For instance, a debt default model might profit from rising petrol prices over time but not from tiger population increases.

Data Pre-processing:

Preparing raw data to be acceptable for a machine learning model is known as data preparation. In order to build a machine learning model, it is the first and most important stage. It is not always the case that we come across the clean and prepared data when developing a machine learning project. Additionally, any time you work with data, you must clean it up and format it. Therefore, we use a data pre-treatment activity for this.Data pre-processing involves elimination of various tasks like removing duplicate values ,missing values or either filling the missing values by using the functions like fillna() etc.,

Exploratory Data Analysis:

In statistics, exploratory data analysis (EDA) is a method of examining data sets to highlight their key features, frequently utilising statistical graphics and other techniques for data visualisation. EDA differs from traditional hypothesis testing in that it is primarily used to explore what the data can tell us beyond the formal modelling. A statistical model can be utilised or not. John Tukey has championed exploratory data analysis since 1970 in order to motivate statisticians to investigate the data and maybe develop ideas that could inspire future data gathering and experiments. EDA differs from initial data analysis (IDA), which is more specifically focused on addressing missing values, transforming variables as necessary, and validating assumptions for model fitting and hypothesis testing. IDA is a part of EDA. We mainly use barplots ,histograms, piecharts ,scatter plot, area plot ,box plot etc.,

Train & Test data:

One of the rapidly developing technologies in the world allows computers or other machines to convert a vast amount of data into forecasts. These predictions, however, heavily rely on the quality of the data, and if we are not using the proper data for our model, it won't provide the desired outcome. In projects involving machine learning, the original dataset is typically split into training data and test data. Our model is trained using a portion of the original dataset—the training dataset—and its generalizability to a different dataset, often known as the test set, is then assessed. Consequently, the two fundamental ideas in machine learning are the train and test datasets, where the training dataset is utilised to fit the model and the test dataset is used to evaluate the model. Usually, the test dataset is approximately 20-25% of the total original data for an ML project and the training data is 70-80% in the ML project.

Algorithms used:

Decision tree;

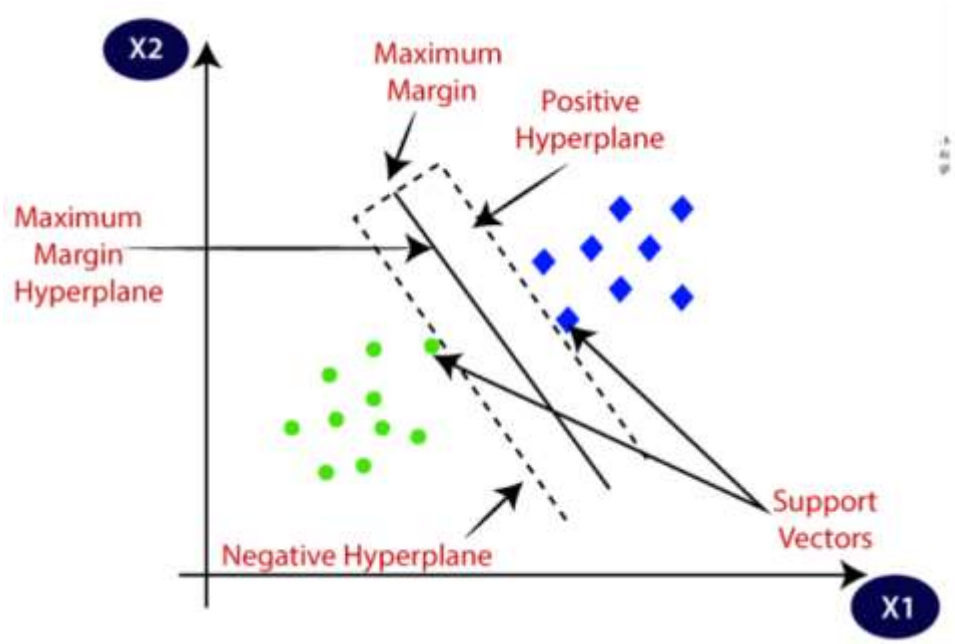
Decision Tree is a common supervised learning approach employed for both regression and classification problems. The goal of technique is forecasting a target by using easy decision rules shaped from the dataset and related features. Being easy to interpret or able to solve problems with different outputs are two advantages of using this model; on the contrary, constructing over-complex trees that cause overfitting is a typical disadvantage.

Random Forest:

A large number of decision trees form a random forest model. This model basically averages the prediction results for trees called forests. The algorithm also contains three random ideas that randomly select the training data when the formed trees, randomly select a subset of variables when splitting nodes, and consider only a subset of all variables to split each node of each basic decision tree. Each base tree is learned from a random sample of the dataset during the training process of the random forest.

SVM:

One of the most well-liked supervised learning techniques for both classification and regression issues is called a support vector machine, or SVM. However, machine learning classification issues are where it is most frequently applied. The SVM algorithm's goal is to provide the best lines or decision boundaries that can categorise the n-dimensional space, making it simple to assign additional data points to the appropriate category in the future. A hyperplane is the name of this best-case decision boundary. SVM picks extrema/vectors to assist in the creation of hyperplanes. The approach is known as a support vector machine, and these extreme situations are known as support vectors. Think about the following graph, which is divided into two distinct categories using decision boundaries or hyperplanes:



LSTM:

A specific kind of RNN called LSTM has a wide range of applications, including document categorization, time series analysis, and voice and speech recognition. In contrast to feedforward networks, RNN predictions are reliant on prior estimates. RNNs are not frequently used in experimental studies because they have a few flaws that lead to estimates that are unworkable. Without delving too far, LSTM uses designated gates for forgetting old information and learning new information to solve problems. Four neural network layers that interact in a particular way make up the LSTM layer. Three components make up a typical LSTM unit: a cell, an output gate, and a forget gate. Cell's primary function is to identify values over arbitrary time intervals and the task of controlling the information flow into the cell and out of it belongs to the gates

RESULTS AND DISCUSSIONS :

We use five machine learning models to forecast the stock direction of up and down. To evaluate the performance of the prediction models, two commonly used evaluation criteria are used in this study: Accuracy, and F1 score. Accuracy is used to evaluate the overall classification ability of the model.

The formula is as follows :

$$\text{Accuracy} = TP + TN / TP + TN + FP + FN$$

TP (True Positives) representing model prediction is true, and the real sample is also true; TN (True Negatives) representing model prediction is false, and the real sample is also false; FP (False Positives) representing model prediction is true while the real sample is false; FN (False Negatives) representing model prediction is false while the real sample is true.

Precision is used to estimate the accuracy of positive samples in the prediction data and recall is used to evaluate the coverage of positive samples in the prediction data of the model. The formula is shown as

$$\text{Precision} = TP / TP + FP$$

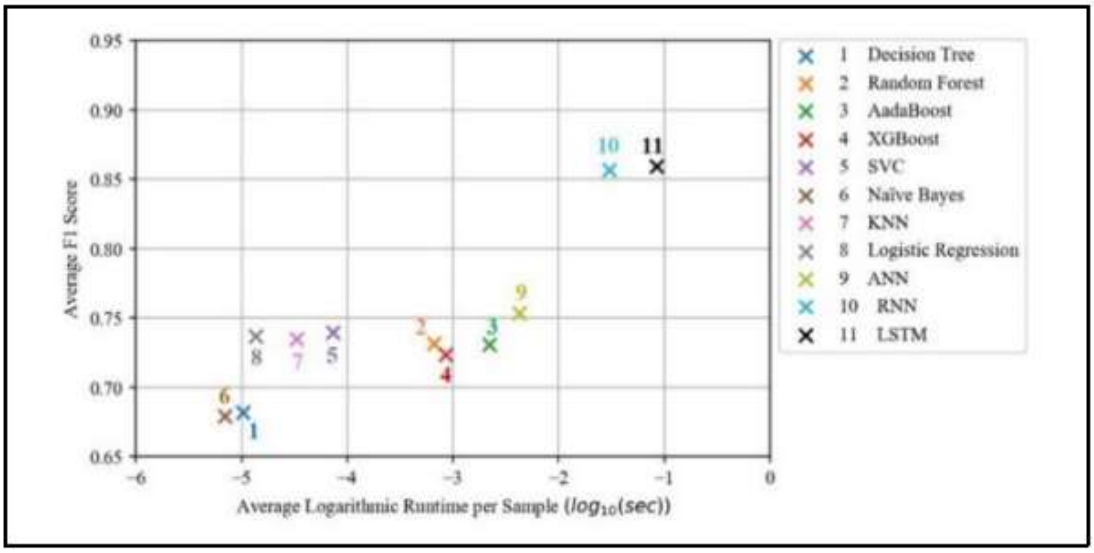
$$\text{Recall} = TP / TP + FN.$$

F1 is an indicator used in statistics to measure the accuracy of a binary model, which also considers the accuracy and recalls of the classification model.

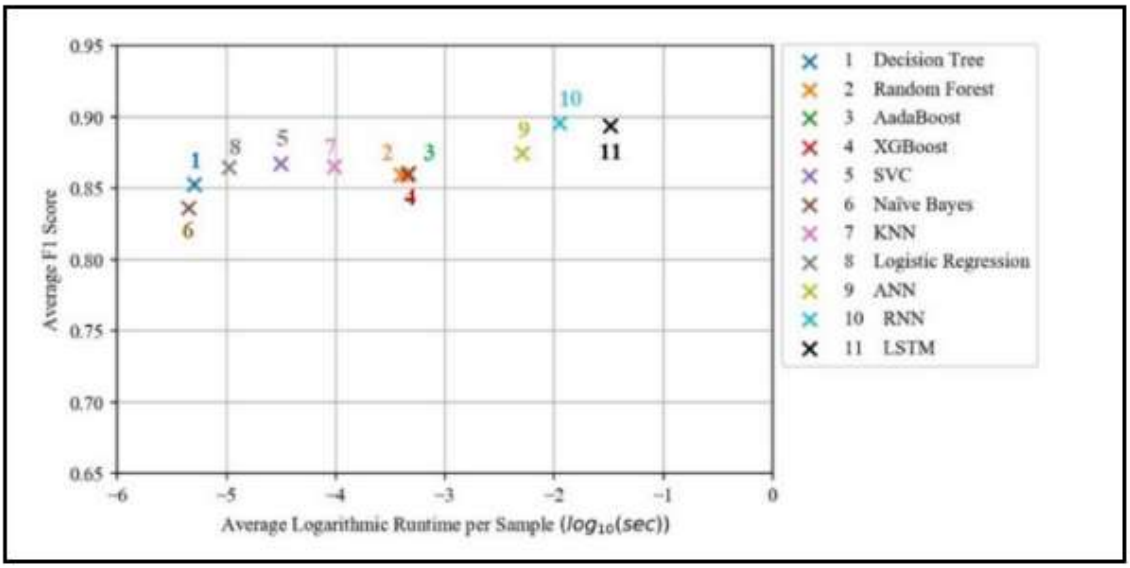
The formula is as follows:

$$F1 = 2 * \text{Precision} * \text{Recall} / \text{Precision} + \text{Recall} .$$

Here is the display of the results of this study, ten years of historical data of four stock market groups (petroleum, diversified financials, basic metals and non-metallic minerals) from November 2009 to November 2019 is employed, and all data is gained from www.tsetmc.com website. Figures 1-4 show the number of increase or decrease cases for each group during ten years. In the case of predicting stock market movement, there are several technical indicators and each of them has a specific ability to predict future trends of market; however, we choose ten technical indicators in this paper based on previous studies.



Average F-1 Score on continuous data



Average F-1 Score on continuous data

CONCLUSION

From the Research papers that we have gathered on the topic “STOCK MARKET PREDICTONS” we have used several supervised machine learning algorithms to forecast the stock values . Out of the many supervised machine learning algorithms available some of the algorithms that we used in our paper are SVM, Decision Tree ,Random Forest ,Linear Regression ,Logistic Regression and also we have used Long Short Term Memory(LSTM) which is one of the promising algorithm that has showed more accurate results than the remaining algorithm due to its long term memory capability.We have seen that if we take binary data instead of continuous data then the supervised algorithms that are mentioned above will show more better results with an accuracy of 85% .The RNN and LSTM techniques will show an accuracy of 90% whether it is binary dataset or the continuous data set .With this we can say that if binary data is taken then we can have more accurate results.When we apply the supervised machine learning algorithms on the short period of time for forecasting then SVM and Logistic Regression have shown better results than the LSTM algorithm . But when the stock data sets are taken for a long period of time and then Forecasting those stock prices for a long period of time the LSTM Technique has overpowered all the remaining supervised machine learning methods above with an accuracy of more than 90% .This prediction helps us to gain enormous profits in the stock market so that we can lead a happy life by careful investment of shares .

REFERENCES:

[1]. M. Nabipour, P. Nayyeri, H. Jabani, S. S. and A. Mosavi, "Predicting Stock Market Trends Using Machine Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis," in IEEE Access, vol. 8, pp. 150199-150212, 2020, doi: 10.1109/ACCESS.2020.3015966.

[2]. Z. K. Lawal, H. Yassin and R. Y. Zakari, "Stock Market Prediction using Supervised Machine Learning Techniques: An Overview," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411609.

[3]. X. Yuan, J. Yuan, T. Jiang and Q. U. Ain, "Stock Trend Prediction Using Candlestick Charting and Ensemble Machine Learning Techniques With a Novelty Feature Engineering Scheme," in IEEE Access, vol. 8, pp. 22672-22685, 2020, doi: 10.1109/ACCESS.2020.2969293.

[4]. Diviesh Chaudhari, Pranay Bafna, Shubham Jadhav, Jay Chaudhari, Prof. V.T. Patil, Prof. V.O.Patil",” Stock Market Prediction Using Machine Learning Algorithm”, International Journal of Scientific Research & Engineering Trends Volume 8, Issue 3, May-Jun-2022.

[5]. S.Chen and C. Zhou, "Stock Prediction Based on Genetic Algorithm Feature Selection and Long Short Term Memory Neural Network," in IEEE Access, vol. 9, pp. 9066-9072, 2021, doi: 10.1109/ACCESS.2020.3047109.

[6]. M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, ‘‘Evaluating multiple classifiers for stock price direction prediction,’’ Expert Syst. Appl., vol. 42, no. 20, pp. 7046–7056, Nov. 2015.

[7]. S. Basak, S. Kar, S. Saha, L. Khaidem, and S. R. Dey, ‘‘Predicting the direction of stock market prices using tree-based classifiers,’’ North Amer. J. Econ. Finance, vol. 47, pp. 552–567, Jan. 2019.

[8]. Y. Baek and H. Y. Kim, ‘‘ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module,’’ Expert Syst. Appl., vol. 113, pp. 457–480, Dec. 2018.

[9]. H. Chung and K.-S. Shin, ‘‘Genetic algorithm-optimized long short-term memory network for stock market prediction,’’ Sustainability, vol. 10, no. 10, p. 3765, 2018.

[10]. M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, and E. Salwana, ‘‘Deep learning for Stock Market Prediction,’’ Entropy, vol. 22, no. 8, p. 840, Aug. 2020.