JES
Journal of Engineering Sciences
Vol 13, Issue 03, March /2022
ISSN NO:0377-9254

SUPERVISED MACHINE LEARNING EXAMINATION OF THE STOCK MARKET

Miss. Rangu Shirisha, PG scholar, Department Of Computer Science, Siddhartha Institute of Technology & Sciences ,TS, India. Email: shirishadarvin@gmail.com

Mrs.M.Sowjanya, Associate Professor, Department Of Computer Science, Siddhartha Institute of Technology & Sciences ,TS, India.

Dr. A. Sathyanarayana, Associate Professor, Department Of Computer Science, Siddhartha Institute of Technology & Sciences ,TS, India.

ABSTRACT :

The stock exchange, sometimes known as the stock market, is one of the most challenging and complicated methods to do business. This body is used by small enterprises, brokerage firms, and the financial sector to create money and divide risks; it's a complicated notion. This research, on the other hand, suggests using a machine learning algorithm to forecast future stock values for exchange by merging open source libraries and pre-existing algorithms in order to help make this unpredictable business model more predictable. We'll see whether this simple implementation produces adequate results. The outcome is purely based on statistics, and it is based on a variety of assumptions that may or may not hold true in reality, such as prediction time.

Keywords : Stock Market ,Data Analysis.

I INTRODUCTION

Advances in fundamental components of information technology have revolutionised the way businesses work over the last few decades. Financial markets, being one of the most intriguing inventions, have a major influence on a country's economy. The global stock market capitalization reached 68.654 trillion US dollars in 2018, according to the World Bank. Because of technology improvements, stock trading has been a major focus of attention in recent years. Investors are looking for tools and strategies that might help them increase earnings while reducing risk. Stock Market Prediction

(SMP) is a difficult task due to its non-linear, dynamic, stochastic, and incorrect nature. SMP is a kind of time-series forecasting in which previous data is analysed and future values are predicted. Financial market forecasting has been a source of worry for analysts from a variety of professions, including economics, mathematics, material science, and computer science. Stock market forecasting relies heavily on the capacity to make money via stock trading. The price of a share, the performance of the corporation, government regulations, the country's Gross Domestic Product (GDP), inflation, natural catastrophes, and other variables all have an impact on the stock market. According to the Efficient Market Hypothesis, stock market values are significantly impacted by fresh information and follow a random walk pattern, making them hard to predict using past data alone. This was a widely accepted notion in the past. Researchers were able to demonstrate that stock market values might be forecasted to a degree thanks to technological advancements. Economic and commercial trends may be predicted using historical market data and information collected from social media platforms. Stock market prediction algorithms' performance is heavily influenced by the quality of the attributes they employ. Academics have looked at a number of approaches for increasing stock-explicit traits, but feature extraction and selection processes need further study. Figure 1 depicts an overview of the paper.

A stock trend is highly significant in a fast-money-spinning economy like India. Nation growth impacts market performance in emerging nations like India and others; if the stock market falls, country moneyspinning falls as well; if the stock market rises, country moneyspinning grows as well. To put it another way, stock growth is necessary for country development. Because of misconceptions that purchasing or selling stock is phoney and dishonest, only around 10% of people in any given place participate in the stock market. People will develop more faith in the stock market if their misgivings about it can be removed. A strategy may help to boost awareness and attract more people to a certain stock trend. The more likely it is, the larger the desire output of anticipating how to alter people's attitudes is. Machine learning algorithms may also be used to predict future trends. This forecast is made using four different predicting models: linear regression, support vector machine, K-nearest neighbour, and Random forest. We're using these

algorithms to forecast future stock market trends, which will help individuals invest their money for greater profit and a more exact stock value, and only by doing so will we be able to boost the country's development and economy. We examine all methods in this research to see which one is optimal for large and small data sets. We offer stock market data sets from Google, Amazon, MCD, and IBM, among other firms. This stock market data was acquired from a number of sources, including Kaggle, the National Stock Exchange of India, and others.

II. LITERATURE SURVEY

Mu yen chen et al. released a research in 2019 that quantified the impact of news items on market prices using a deep learning approach called LSTM (long short-term memory). They think that this research can predict stock market developments.

Andrea Picasso et al. presented a report in 2019 in which the authors used a range of application and automation methodologies to predict market trends using a combination of economic and elemental analysis. A neural network, a machine learning technique, is used to tackle the problem of trend stock and charts using predictive data. As an input, the emotion of a news storey is employed. According to their research, the most challenging accomplishment in the use of astral one-off news information is the difficulty. To solve this difficulty in the future, the correct feature fusion method will be necessary.

Gangadhar Shobha et al. published a paper in 2018 that included a complete review of machine learning approaches to help readers comprehend equations and ideas. The author spoke about three different kinds of machine learning algorithms, as well as measures including accuracy, confusion matrix, recall, RMSE, precision, and frequency of mistakes. Because many people are unclear whether to use most machine learning techniques for prediction or for other reasons, the author feels that this overview will assist those who are new to machine learning.

Suryoday Basak et al. created an experimental methodology for forecasting stock values in 2018, whether they rise or fall. In this experiment, the author employs two algorithms: a random forest classifier and Gradient enhanced decision' n trees, and they achieve more accuracy than previous research publications, with results ranging from 50% to 67 percent realism, according to the author of this article. In the future,

they may use the built improved tree model for short-term data windows.

Arash Negahdari kia et al. conducted a number of tests and models for stock prediction using historical data in 2018, including the HyS3 graph-based semi-supervised model and the ConKruG network views Kruskal based graph algorithm. They hope that in the future, social media data, such as Twitter data, will be used to help these algorithms better predict stock values.

In this paper, Bruno Miranda et al. work on the prediction of financial market values using machine learning models such as support vector machine (SVM) and neural networks with data from the North American market. New models may have future opportunities for North American market data for prediction purposes.

K. Hiba Sadia et algoal .'s in this research is to first preprocess the raw data, then compare random forest with SVM techniques. The author's major purpose is to identify the best algorithm for stock trend prediction, and in the end, the random forest algorithm is the best-fit method for future market forecasting. They hope that adding additional parameters to the model will increase the accuracy of the findings in the future.

A. Akash et al. published two new algorithms in 2019: "LS SVM" (least square support vector machine) and "PSO" (parametric support vector machine) (particle swarm optimization). To minimise overfitting and certain technical signs, the work "PSO" generally picks the bestunbounded parameter with the "LS SVM," which will essentially increase the result accuracy. On the other side, the proposed method is being compared to an artificial neural network model.

III SYSTEM ANALYSIS

EXISTING SYSTEM

As we all know, the stock market is an important trading platform that affects everyone on a personal and national basis. The basic notion is easy to understand. Companies will list their stock on the stock market as a minor commodity known as Stocks. They do so in order to raise money for the firm. Individuals, on the other hand, sometimes lose money and sometimes gain more money in the stock market since we cannot predict the best fit for future sales.

PROPOSED SYSTEM

Our approach, which is based on historical year-by-year and month-by-month projections, can forecast future sales. Now,

if we try to graph the stock exchange price over time (say, 6 months), it is really difficult to predict the future result on the graph, but we can easily draw a graph and see the analysis using our offered technique.

IV IMPLEMENTATION

Architecture:

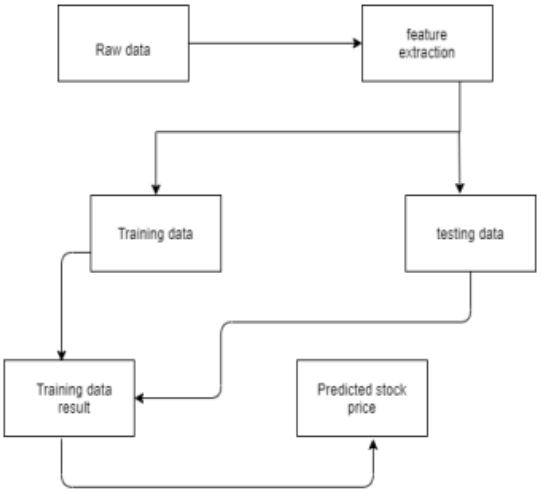


Fig-1.. architectures of the system model

Kaggle is an online community for data analysis and predictive modelling. It also contains information from data miners who have provided datasets from a variety of industries. Various data scientists compete to create the best accurate data prediction and visualisation models. It allows users to construct models using their own data and engage with a variety of data scientists to address real-world data science challenges. The dataset for the proposed study was obtained via Kaggle. On the other hand, this data set is accessible in raw format. A compilation of stock market data for a few companies makes up the data set. The first step is to convert raw data into processed data. Because the raw data acquired has a variety of properties, but only a few of those features are useful for prediction, feature extraction is used. In the first step, feature extraction, the important attributes are selected from the whole list of characteristics available in the raw dataset. Feature extraction starts with a set of measured data and works its way up to derived values or features. These characteristics are intended to be both informative and non-redundant, facilitating learning and generalisation. Feature extraction is a dimensionality reduction technique that reduces a large number of raw variables to a smaller number of manageable features while still properly and totally depicting the original data set. Following the feature extraction technique, the data created after feature extraction is separated into two portions by a classification procedure.

- **MODULES:**

  - trainer

- cosmonaut
- admin
- artificial intelligence

Trainer

Cosmonaut training is one of the most important parts of the human flight programme. As part of the shift to advanced digital smart technologies, computer-assisted training, and artificial intelligence, crew training for orbital space stations is being developed. The crew training process includes planning, activity organisation, and performance management. In order to accomplish the goals of crew training, it is essential to make the most of available resources. For next training sessions, software to download emergency scenario scenarios. Emergency preparedness training. Detection of crew flaws at all stages of the training process.

Cosmonaut

The development of computer-assisted control in the area of cosmonaut training has been prompted by the rising demand for quick decision-making. The complexity of cosmonaut training increases as the number of accomplished flight missions increases. The complexity of material and technical infrastructure, as well as communication equipment, is significant. Understand the goals, objectives, and functions of the whole crew training system by using integrated and special purpose simulators.

Admin:

Authorizing trainers and cosmonauts is the administrator's role. To administer, all of the information must be gathered. The need to collect large amounts of data (the number of flying operations aboard surpasses a thousand) and account for all factors affecting the planning and administration of training sessions led to automated crew training control. Based on this data, the system will be able to generate additional evaluations and exercises, allowing for a more effective understanding of the training content.

Artificial Intelligence

In many domains of science and technology, including manned space programmes, advanced digital and smart technologies, robotic systems, novel materials and design techniques, large data processing systems, computer-aided learning, and artificial intelligence (AI) are all important. Several AI-based technology concepts and pilot systems have been

developed over several decades by the industry (3-D computer vision, automated systems for planning and assessing cosmonaut actions, inquiry and communications system).

V RESULT AND DISCUSSION

Store the data set and see the number of samples with features.

```
# Set start and end date for stock prices
start_date = datetime.date(2009, 3,8)
end_date = datetime.date.today()
# Load data from Quandl
#data = quandl.get('FSE/SAP_X', start_date=start_date, end_date=end_date)
# Save data to CSV file
data = pd.read_csv("data/sap_stock.csv", low_memory = False, skiprows = 1, encoding = "ISO-8859-1")
#data.to_csv('data/sap_stock.csv')
print("The GTD dataset has {} samples with {} features.".format(*data.shape))
```

```
The GTD dataset has 2550 samples with 11 features.
```

Display the dataset 1st 5 rows.

	Date	Open	High	Low	Close	Change	Traded Volume	Turnover	Last Price of the Day	Daily Traded Units	Daily Turnover
0	2009-03-09	25.16	25.82	24.48	25.59	NaN	5749357.0	145200289.0	NaN	NaN	NaN
1	2009-03-10	25.68	26.95	25.68	26.87	NaN	7507770.0	198480965.0	NaN	NaN	NaN
2	2009-03-11	26.50	26.95	26.26	26.64	NaN	5855095.0	155815439.0	NaN	NaN	NaN
3	2009-03-12	26.15	26.47	25.82	26.18	NaN	6294955.0	164489409.0	NaN	NaN	NaN
4	2009-03-13	26.01	26.24	25.65	25.73	NaN	6814568.0	176228331.0	NaN	NaN	NaN

Information features in a dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2550 entries, 0 to 2549
Data columns (total 11 columns):
Date                    2550 non-null object
Open                    2242 non-null float64
High                    2543 non-null float64
Low                     2543 non-null float64
Close                   2550 non-null float64
Change                  11 non-null float64
Traded Volume           2504 non-null float64
Turnover                2497 non-null float64
Last Price of the Day   0 non-null float64
Daily Traded Units      0 non-null float64
Daily Turnover          7 non-null float64
dtypes: float64(10), object(1)
memory usage: 219.2+ KB
```

Descriptive statistics summary of data set:

	Open	High	Low	Close	Change	Traded Volume	Turnover	Last Price of the Day	Daily Traded Units	Daily Turnover
count	2242.000000	2543.000000	2543.000000	2550.000000	11.000000	2.504000e+03	2.497000e+03	0.0	0.0	7.0
mean	56.686886	61.562225	60.535073	60.995855	-0.070000	3.296818e+06	1.828440e+08	NaN	NaN	0.0
std	18.320821	21.184135	20.934460	21.097480	0.709761	2.004323e+06	9.350710e+07	NaN	NaN	0.0
min	25.160000	25.820000	24.480000	25.590000	-0.740000	0.000000e+00	1.767350e+05	NaN	NaN	0.0
25%	41.500000	43.430000	42.590000	42.950000	-0.500000	2.131686e+06	1.300462e+08	NaN	NaN	0.0
50%	56.560000	59.480000	57.580000	58.015000	-0.290000	2.852772e+06	1.626544e+08	NaN	NaN	0.0
75%	67.732500	78.365000	77.085000	77.762500	0.085000	3.876528e+06	2.104511e+08	NaN	NaN	0.0
max	100.100000	108.520000	107.020000	107.800000	1.250000	3.645671e+07	1.369431e+09	NaN	NaN	0.0

Display features in data set:

```
Index(['Date', 'Open', 'High', 'Low', 'Close', 'Change', 'Traded Volume',
       'Turnover', 'Last Price of the Day', 'Daily Traded Units',
       'Daily Turnover'],
      dtype='object')
```

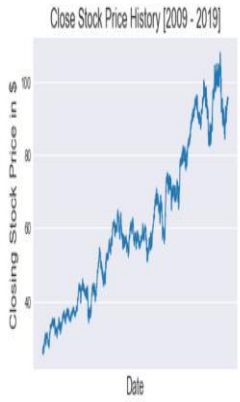
Reset index column so that we have integers to represent time for later analysis:

	index	Date	Close
0	0	2009-03-09	25.59
1	1	2009-03-10	26.87
2	2	2009-03-11	26.64
3	3	2009-03-12	26.18
4	4	2009-03-13	25.73

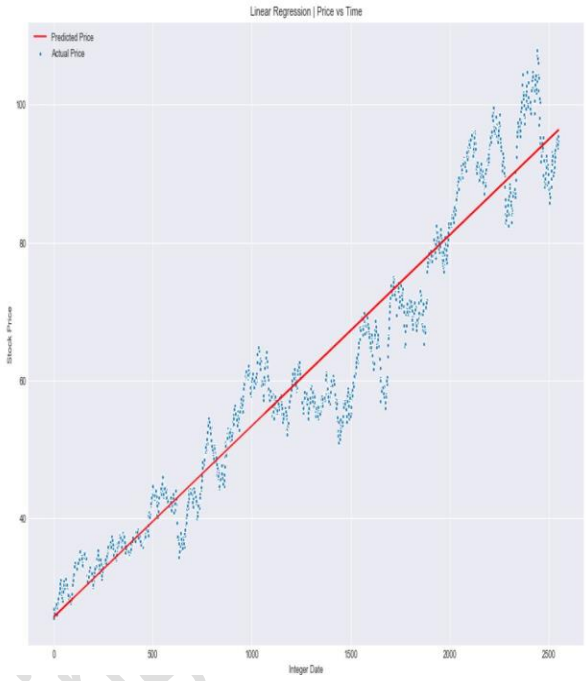
Check data types in columns:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2550 entries, 0 to 2549
Data columns (total 3 columns):
index    2550 non-null int64
Date     2550 non-null object
Close    2550 non-null float64
dtypes: float64(1), int64(1), object(1)
memory usage: 59.8+ KB
```

Create subplots to plot graph and control axes:



Linear Regression: Using linear regression plot the data:



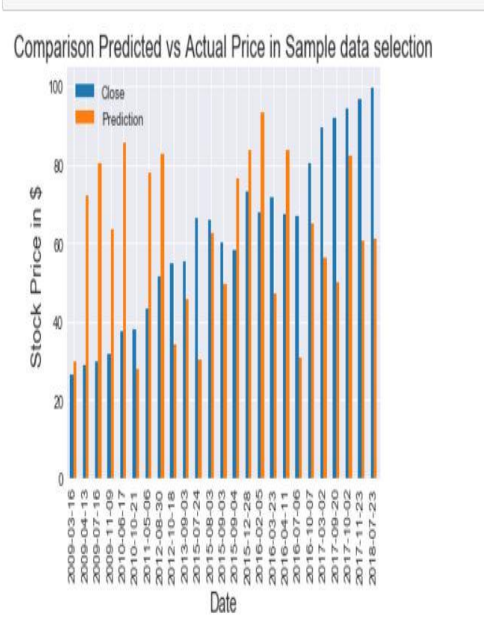
Generate array with predicted values:

```
[29.92726154 72.20558594 80.49490988 63.61130361 85.70692626 28.04206413
 77.94434867 82.57417173 33.86399732 45.53558721 30.065879   62.58553442
 49.25053506 76.33638616 83.73855837 93.30316291 47.11582622 83.71083488
 30.84213676 64.80341374 56.2091314  49.88817536 82.32466031 60.31220813
 60.89440145]
```

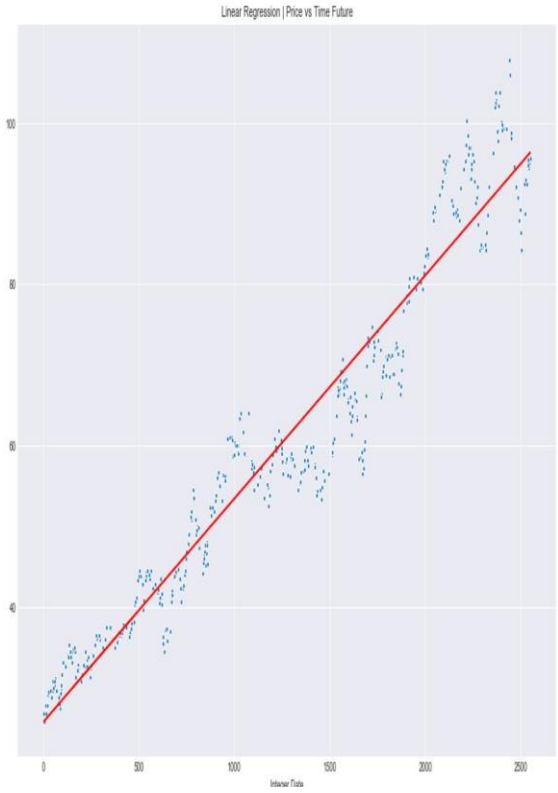
Display the details:

	index	Date	Close	Prediction
5	5	2009-03-16	26.480	29.927262
25	25	2009-04-13	29.005	72.205586
93	93	2009-07-16	29.890	80.494910
175	175	2009-11-09	31.500	63.611304
331	331	2010-06-17	37.440	85.706926

Comparison Predicted vs Actual Price in Sample data selection:

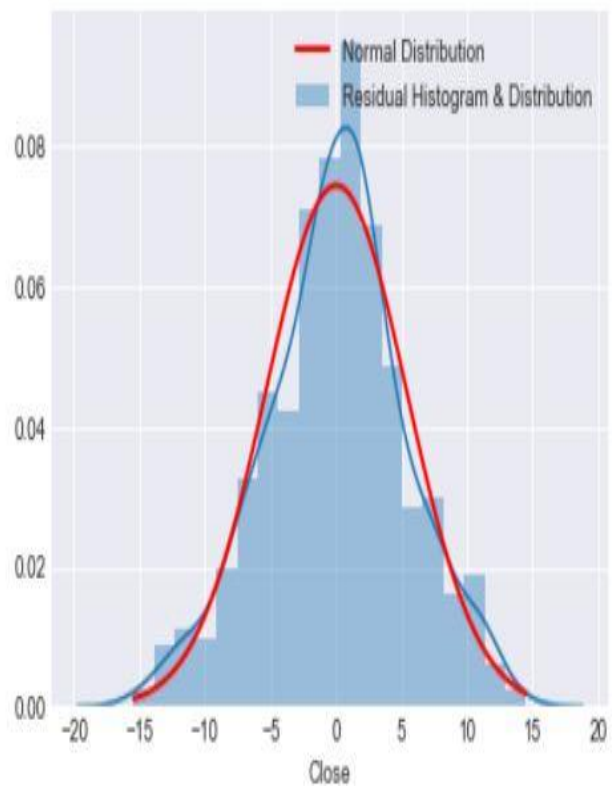


Price vs. Time Future:



Predicted Actual Price in future:

Residual Histogram & Distribution:



Add new column for predictions to df:

```
count    2550.000000
mean       60.995955
std        21.097480
min        25.590000
25%        42.950000
50%        58.015000
75%        77.762500
max       107.800000
Name: Close, dtype: float64
```

The value of R2 shows that are model accounts for nearly 94% of the differences between the actual stock prices and the predicted prices:

```
from sklearn.metrics import explained_variance_score
explained_variance_score(y_test, y_pred)
```

```
0.93664675240512
```

VI CONCLUSION

Machine learning is a highly powerful technology with a broad variety of applications, as we've seen so far. Machine learning, as we've seen, is strongly dependent on data. As a consequence, it's vital to understand that data is really valuable, and that data analysis, however easy it may seem, is a complex task. Machine learning has evolved into deep learning and neural networks, but the core premise remains the same for all of them. This article shows how to put machine learning into practise in a straightforward and concise manner. For dealing with and addressing various challenges in various circumstances, there are a range of techniques, strategies, and tactics. This essay focuses only on supervised machine learning and seeks to explain the fundamentals of this complex process in a straightforward manner.

Further Enhancement

More dimensions and variables will be incorporated, such as financial ratios, multiple occurrences, and so on. The bigger the number of parameters evaluated, the more accurate the result. The algorithms may also be used to analyse the content of public comments in order to spot patterns and connections between consumers and workers. To predict a company's overall performance structure, traditional algorithms and data mining technologies may be applied.

VII REFERENCES

[1] The film's cast includes Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. "A Machine Learning Approach to Building Domain-Specific Search Engine," Citeseer, IJCAI, 1999. "A Machine Learning Approach to Building Domain-Specific Search Engine," Citeseer, IJCAI, 1999. IJCAI

[2] Yadav, Sameer. (2017). 629-632 in Global Journal for Research Analysis on VOLATILITY IN THE STOCK MARKET - AN INDIA STOCK MARKET STUDY.

Montgomery, D.C., Peck, E.A., and Vining, G.G., 2012. An overview of linear regression analysis (Vol. 821). The publishing firm John Wiley & Sons is situated in the United Kingdom.

[4] H. Smith and N.R. Draper (1998). Regression Analysis in the Real World (3rd ed.). The ISBN number is 0-471-17082-8. John Wiley & Sons is the publisher.

[5] Robert S. Pindyck and Daniel L. Rubinfeld (1998, 4h ed.). Econometric Models and Economic Forecasts

[6] "Linear Regression," Yale University, 1997-1998. http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

Agarwal (n.d.) (n.d.) (n.d.) (n.d.) (n.d.) (n.d.) (n.d) (July 14, 2017). Economist with Intelligence, "An Overview of the Stock Market." On the 18th of December, 2017, I was able to get a hold of some information.

[8] Jason Brownlee, Machine Learning Mastery, "Linear Regression for Machine Learning," March 2016, https://machinelearningmastery.com/linear-regression-for-machinelearning/, accessed December 2018, https://machinelearningmastery.com/linear-regression-for-machinelearning/

[9] Google Developers, "Decending into ML: Linear Regression," https://developers.google.com/machinelearning/crash-course/descending-into-ml/linear-regression, https://developers.google.com/machinelearning/crash-course/descending-into-ml/linear-regression, Oct. 2018, Google LLC.

[10] Fiess, N.M., and MacDonald, R., 2002. One of the pillars of technical analysis is analysing the information content of High, Low, and Close prices. Economic Modelling, vol. 19, no. 3, pp. 353-374.

[11] T. Marwala and E. Hurwitz, 2012. Widespread errors when applying computational intelligence and machine learning in stock market modelling are common. An arXiv preprint is arXiv:1208.4429.