ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

A Machine Learning Approach to Stock Forecasting

Dhruv Pathak[1], Hima Karan Kadali[2], Prayag Saraiya [3]

U.G. Student, Department of Computer Engineering, MPSTME Engineering College, Vile Parle , Maharashtra, India[1]

U.G. Student, Department of Computer Engineering, MPSTME Engineering College, Vile Parle, Maharashtra, India[2]

U.G. Student, Department of Computer Engineering, MPSTME Engineering College, Vile Parle, Maharashtra, India[3]

ABSTRACT: The stock market has always been a promising avenue for lucrative investing, but most of the profit making depends on the analysis of the current and past market scenario followed by subsequent predictive actions. The currently overblown market economy has given rise to numerous variables which need to be considered before making a beneficial transaction in the stock market. Manually analysing all these variables and affecting factors is too cumbersome and error prone. Therefore a Machine Learning approach is best suited for analysis of such a seemingly chaotic system. Machine learning can give a prediction of various aspects of a particular stock or an index, such as future values of the opening price, closing price, index value etc. This will help investors and traders make better and faster decisions.

KEYWORDS: Stock market; prediction; machine learning

I. INTRODUCTION

Stock markets are a place where stocks are bought and sold and their prices depend on their demand. Analysing the intrinsic factors behind their demand as well as the statistical data regarding a stock can give us an accurate idea about its future state. The analysis of the intrinsic factors to predict a stock price is called fundamental analysis while the statistical analysis to predict a stock price comprises technical analysis. Application of Machine learning algorithms is generally focused on technical analysis but incorporation of the concepts of fundamental analysis into machine learning can be beneficial. This paper talks about how various efforts have been taken in the application of Machine Learning to Stock forecasting and also suggests new potent ideas that can be worked upon. The structure of this paper is as follows, section II provides an introduction to various stock market concepts, section III provides an overview machine learning and primarily involved machine learning concepts, Sections IV and V comprise the literature review and comparative study of related work. Section VI consists of the proposed methodology which will be followed to improve the application of machine learning for stock forecasting.

II. STOCK MARKET CONCEPTS

A stock market, equity market or share market is the aggregation of buyers and sellers (a loose network of economic transactions, not a physical facility or discrete entity) of stocks (also called shares), which represent ownership claims on businesses [7].

When a company or firm wishes to raise revenue for business growth, it can either take a loan or issue stocks. When a company issues stocks, they are open for the public to purchase and the owners of these stocks also in a way own a proportional part of the company. Thus they are entitled to a part of the profit made by the company, which is usually received as dividend per share. These stocks or shares can be freely traded among individuals in a stock market. The smallest unit ofsuch a market is a transaction. The more a stock is in demand, the more times it gets transacted, and its price goes up.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

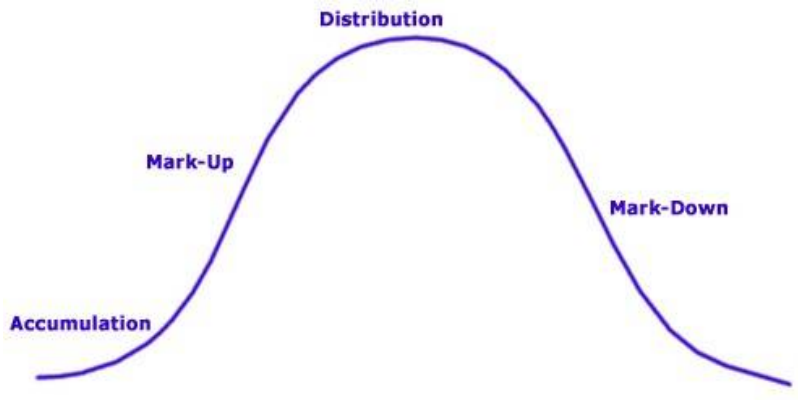


Figure 1. Market Phases
(source:http://i.investopedia.com/inv/articles/site/techanalysis/050504_2.gif)

The stock market usually follows a cyclic trend, going through four phases as shown by the figure below.

1. Accumulation phase: Investors begin to buy stocks of a company as the company starts stabilizing and shows promise. Overall market sentiment begins to switch from negative to neutral. The market is said to have a bearish (generally negative) sentiment.
2. Mark-up phase: Once the growth becomes evident, more buyers begin buying stocks. At the same time early buyers having reaped benefits of the growth start selling stocks.
3. Distribution phase: Overall sellers for a stock go up. Sellers at this stage either make large profit or break even. The market is said to have a bullish (generally positive) sentiment.
4. Mark-down phase: The stock value depreciates as supply of stock increases after the distribution phase. Toward the end of this phase investors again start coming in seeing the depreciated price as an opportunity for profit.

Stock exchanges also offer indices, whose values are dependent on the behaviour of underlying stocks and sectors. These indices are often used as traders as an aid for investing.

III. MACHINE LEARNING CONCEPTS

Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.[8]

Machine learning is classified into three categories: supervised, unsupervised and reinforcement learning as shown by the figure below.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

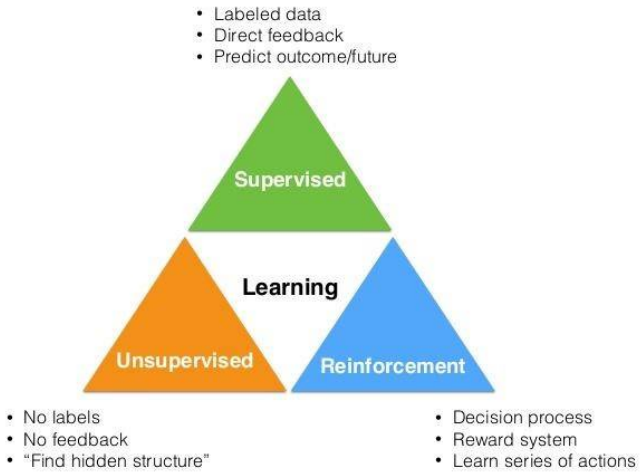


Figure 2. Machine learning types (source:http://www.techjini.com/wpcontent/uploads/2017/02/mc-learning.jpg)

In this paper the purpose of machine learning is prediction. Therefore supervised learning is employed.

Supervised learning is the machine learning task of inferring a function from labelled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a desired output value. A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new examples.[9] This task of mapping new examples is further classified into a regression problem or classification problem. Regression is used for prediction of continuous values while classification is used to predict discrete categories as outputs.

The application of supervised learning to a data set is shown by the following block diagram.

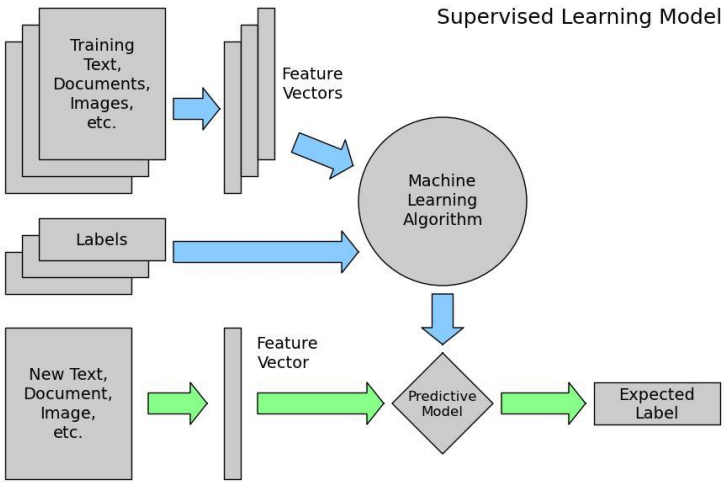


Figure 3.Supervisedlearningmodel (source:http://ogrisel.github.io/scikit-learn.org/sklearntutorial/_images/plot_ML_flow_chart_12.png)

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

The most widely used learning algorithms are Support Vector Machines, linear regression, logistic regression, naive Bayes, linear discriminant analysis, decision trees, k-nearest neighbour algorithm, and Neural Networks (Multilayer perceptron).[9]

IV. LITERATURE REVIEW

The Stock market has been a vertical for machine learning research since a long time. Several researchers have applied and tested different machine learning algorithms on historical stock data. Research has been focused on the choice of algorithm, optimization of a promising algorithm, feature selection, data preparation and sentiment analysis.

Radu Iacomin[1] has performed an experimental study on feature selection in order to improve the results from the SVM algorithm on historical stock data. 16 forex stocks and their reverses were chosen with their data from the period of 4 November 2008 to 7 January 2014. The main goal was to detect the direction and price of 8 January 2014 and to validate the algorithms for a real prediction [1]. In order to improve the SVM algorithm, feature selection was introduced and this was done using two techniques namely Genetic Algorithm (GA) and Principal Component Analysis (PCA). The features selected by both algorithms were presented using the following tables.

Indicator	Description
Commodity Channel Index	Identifying new trends or announcing the instability variation
Accumulative Swing Index	Generating signals from the maximums and minimums of the price
Williams R	Identifying the current close price level depending on the last maximums
Swing Index	Generates the trend of the price in the short future
Polarized Fractal Efficiency	Explains the geometry of fractals in the form of an oscillator
Time series forecast	Illustrates statistic trends if the safety price of a period
Ravi	Uses 2 moving averages calculated in percentage to generate a decision
DMX	Uses a Jurik moving average on directional technical indicators of the price
Volatility Breakout	Generates the security and volatility of a stock
Schaff Trend Cycle	Generates trends by combining stochastically moving averages

GASVM indicators

Indicator	Description
Moving Average	Exponential moving average
Relative strength index	Momentum oscillator that measures price variations and the speed of variations
Williams R	Identifying the current close price level depending on the last maximums
Bollinger Bands	Measures the price volatility and uses it for dynamic levels of support and resistance
Rate of Change	Illustrates the variation of the price for a given period
Stochastic Oscillator	Momentum oscillator that measures the distance between the close and the range
Ichimoku	Determines the tendency of buying or selling of the stock
Linear Regr. Slope	Determines the slope of linear regression of a period
Aroon Oscillator	Determines the level of overbought and oversold in the last period

PCASVM indicators

Figure 4. Selected indicators and their description [1]

Therefore the GASVM and PCASVM solve the problem of prediction better and also eliminate issues false predictions and redundant features.

Pankaj Kumar and Dr. Anju Bala[2] have applied various various machine learning algorithms to the stock data set with the objective of stock market prediction. They treat the problem as a binary classification problem. Three models namely Decision tree model, Linear model and
Random forest model were used. These models were used for training and testing of stock data following which their accuracy was checked and the three models were compared. They used different graphs, namely ROC plot graph, H-

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

measure, Area under curve (AUC) graph and smoothed score distribution graph, to compare the models. They also use the True Positive Ratio (TPR) and False Positive Ratio (FPR) for comparison. The evaluation matrix generated is as follows.

Figure 5. Evaluation Matrix [2]

Evaluation matrix	Decision tree	Linear model	Random Forest
TPR	0.717	0.656	0.627
FPR	0.685	0.603	0.549
AUC	0.517	0.538	0.554
Accuracy	51.87	52.83	54.12

They conclude that Random Forest is more efficient to predict binary classification due to its accuracy [2] but also suggest that ensemble of different methods can further improve accuracy.

Yahya Eru Cakra and Bayu Distiawan Trisedya [3] incorporate sentiment analysis into the prediction of stock prices. They use the Naïve Bayes and Random Forest algorithm to classify tweets for Sentiment analysis. Then the result of Sentiment analysis is used for stock price prediction using a linear regression model. The authors use two datasets to accomplish the task. The first is tweet dataset collected using Twitter REST API and the second is the Stock Price dataset gathered via Yahoo Finance CSV API.

The tweets were then classified into three classes: positive, negative and neutral using different supervised learning algorithms. The positive tweet percentage was calculated post tweet classification.

The classified tweet data along with model corresponding stock features from 1 to 5 days before observation day were applied to different prediction models. The prediction models designed were Price Fluctuation Prediction, Margin Percentage Prediction, and Stock Price Prediction. Then the models were evaluated based on the value of coefficient of determination (R^2). The research showed that Random Forest algorithm was better at classifying tweet data with an accuracy of 60.39% accuracy and the price prediction model had the highest value of R^2 closest to 1. Theyalso found that percentage of positive tweets decreases R^2 value of a model.

Trevor M. Sands, Deep Tayal, Matthew E. Morris and Sildomar T. Monteiro [4] have conducted experiments to develop a prediction algorithm with improved accuracy. Their work aimed at boosting the Support Vector Machine model with Particle Swarm Optimization (PSO) for thepurpose of short term stock prediction. The optimized SVM performance is then compared to the Naïve Bayes classifier and artificial neuralnetwork models in terms of accuracy and computation time required. [4]

Daily stock data (opening price and adjusted closing price) of apple and google, from January
2005 to March 2014 was taken as the data set which further was split into three for cross validation. Then SVM parameters were optimized using PSO following which SVM was applied to the data sets.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

The results were then compared to other algorithms such as Naïve Bayes classifier and ANN, as shown by the table given below.

AVERAGE PERFORMANCE RESULTS FOR APPLE DATASET

Algorithm	Accuracy (%)	Time (s)
NBC	49.4 ± 1.65	0.158
ANN	85.1 ± 0.263	20.4
LS-SVM	50.3 ± 1.57	0.163
SVM + PSO	96.1 ± 0.503	23.5

AVERAGE PERFORMANCE RESULTS FOR GOOGLE DATASET

Algorithm	Accuracy (%)	Time (s)
NBC	47.7 ± 2.09	0.145
ANN	84.2 ± 0.343	24.2
LS-SVM	47.4 ± 2.57	0.327
SVM + PSO	97.5 ± 0.429	16.2

Figure 6. Performance comparison SVM+PSO [4]

Therefore the results proved PSO optimized SVM to be highly robust and accurate in providing reliable predictions for the short term which is useful for casual interday traders.

Alexander Porshnev, Ilya Redkin and Alexey Shevchenko [5] used a lexicon-based Sentiment analysis to evaluate public psychology followed by application of machine learning algorithms i.e Support vector machine and Neural Network to predict DJIA and S&P500 indicators.

Sentimental analysis of twitter posts was performed natural language processing concepts. A dictionary approach was used i.e measuring the tweets with words like “hope”, “worry” and ‘’fear”. Then more complex dictionaries were created to sort these tweets into eight basic emotions using a sentimental analyser which uses a ‘gold standard’- happy, loving, calm, energetic, fearful, angry, tired, sad. A gold standard for emotional sorting was developed through a manual process, using linguistic experts. Therefore three data sets were now available - Basic data set , Basic&WHF (worry, hope and fear) , Basic&8EMO (eight emotions) [5].

35,00,000 tweets per day and opening and closing stock prices from yahoo for historical data for DJIA(Dow Jones Industrial Average) were used to develop a sentiment analysed training dataset. Then Neural networks Algorithm and Support Vector Machine algorithms were applied on the datasets i.e(Basic data set , Basic&WHF ,Basic&8EMO), to perform a comparison.

It was found that the SVM algorithm on the Basic&8EMO dataset gave the best accuracy (64.10%) in DJIA prediction. Although not highly accurate in terms of prediction, the research finds the hypothesis that, ‘inclusion of people psychology via sentiment analysis of twitter feed into prediction’ to be promising.

Mehak Usmani, Syed Hasan Adil, Kamran Raza and Syed Saad Azhar Ali [6] performed research using machine learning techniques with the main objective to predict market performance of Karachi Stock Exchange (KSE). In their research they attempted to predict the stock market by three variants of Artificial Neural Network (Single Layer Perceptron, Multilayer Perceptron and Radial Basis Function) and by the Support Vector Machine algorithm.

The research focused on incorporating different factors affecting the market, as input attributes for the model. The selected attributes were continuous numeric values, were of different range and were normalized between [+1,-1]. The factors considered were as follows [6]:

1. Market History: historical closing index of KSE-100 after applying statistical techniques including ARIMA and SMA over the data.
2. News: business, financial, political and international event based news.
3. General Public Mood: Twitter as a public sentiment source.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

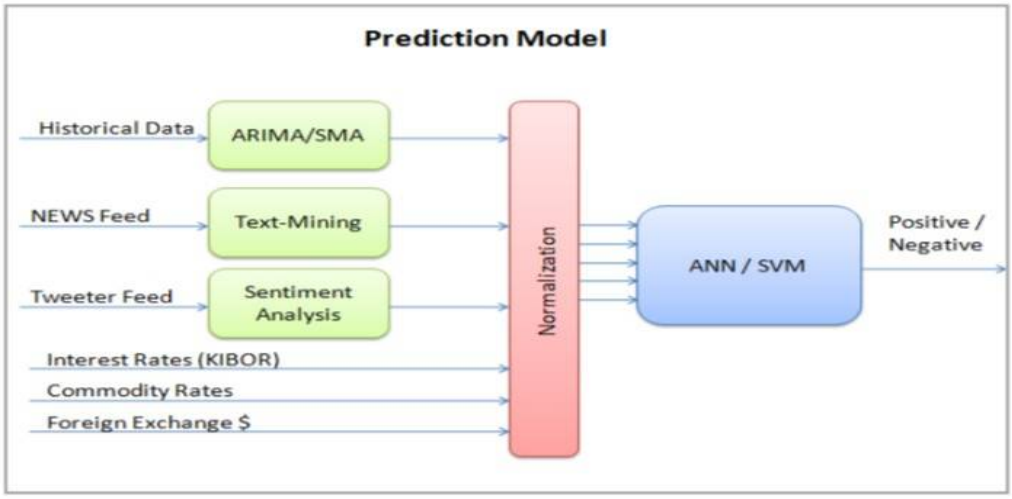
(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

4. Commodity Price: Commodities including Gold, Silver and Petrol were used as they have an impact on market behaviour.
5. Interest Rate: 1-week rates of the daily issued Karachi Inter Bank Offer Rate (KIBOR).
6. Foreign Exchange: Historical exchange rate between Pakistan Rupee (PKR) and US Dollar (USD).

The data was then prepared and used by the aforementioned algorithms.



7. Developed model [6]

TABLE I
COVARIANCE AND CORRELATION OF ATTRIBUTES & MARKET PERFORMANCE

S. No.	Attributes	Covariance	Correlation
1	Oil rates	422.5897	0.276812
2	Gold Rates	-586.7435	-0.07018
3	Silver Rates	4.9080036	0.0238194
4	FEX	0.045220645	0.00020486
5	SMA	-1500.61	-0.03858
6	ARIMA	-19741.62879	-0.131519799
7	KIBOR	-0.1984227	-0.0104532
8	NEWS	2.88594651	0.06580184
9	Twitter	-9.2570117	-0.1456551

TABLE II
COMPARISON OF MACHINE LEARNING TECHNIQUES

Data set used for verification	Machine Learning Algorithms			
	SLP	MLP	RBF	SVM
Training Set	83%	67%	61%	100%
Training Set	60%	77%	63%	60%
Average	71.5%	72%	62%	80%

Figure 8. Performance analysis of results

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

Therefore the method employed resulted in giving an idea about the factors affecting market behaviour, and to what extent. Oil rates has the highest covariance showing that it has the most influence on market behaviour.

Out of the algorithms employed MLP proved to have the highest accuracy of 77% in predicting the market behaviour using input parameters.

V. PRIMARY RESULTS

Following is a table comprising the main outcomes of the research work done in the aforementioned literature survey.

Sr. no	Authors	Paper Title	Primary results of research
1	R. Iacomin	Stock market prediction	Benefits of optimizing SVM by addition of a new algorithm for feature selection before feeding the classification predictor. The optimized algorithms, GASVM and PCASVM and show more accuracy and efficiency than SVM.
2	P. Kumar and A. Bala	Intelligent stock data prediction using predictive data mining techniques	Among different predictive models, namely random forest model, linear model and decision tree model, the Random Forest model is best suited to predict a binary classification.
3	Y. E. Cakra and B. D. Trisedya	Stock price prediction using linear regression based on sentiment analysis	Random forest algorithm is better than other classification algorithms at classifying tweet data for the purpose of Sentiment analysis. A price prediction model shows a higher coefficient of determination than other prediction models.
4	T. M. Sands, D. Tayal, M. E. Morris, and S. T. Monteiro	Robust stock value prediction using support vector machines with particle swarm optimization	PSO optimized SVM is highly robust and very accurate (95%) in providing reliable predictions for the short term and is beneficial for interday trading.
5	Porshnev, I. Redkin, and A. Shevchenko	Machine Learning in Prediction of Stock Market Indicators Based on Historical Data and Data from Twitter Sentiment Analysis	Sentimental analysis of twitter posts for stock prediction is a promising approach. SVM Algorithm applied on a dataset classified under the Basic&8EMO (eight emotions) scheme had highest accuracy in terms of prediction.
6	M. Usmani, S. H. Adil, K. Raza and S. S. A. Ali	Machine Learning in Prediction of Stock Market Indicators Based on Historical Data and Data from Twitter Sentiment Analysis	Application of the Multi-Layer Perceptron algorithm on an assorted dataset of market history, news, general public mood, Commodity price, interest rate and foreign exchange gave a prediction of the Market behaviour with an accuracy of 77%.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

VI. FUTURE SCOPE

The performed literature survey has shown that although chaotic the stock market can be predicted to a lucrative extent. Most of the existing work in this field focuses on technical analysis and short term stock market prediction, be it prices, indices or trends.

We are inclining our research towards a long term prediction model, which would be developed using fundamental as well as technical factors. Literature review shows that fundamental analysis can be included through the subject of Sentiment analysis. We would like to further explore the same and also incorporate other intrinsic macro-economic factors (GDP, interest rates, news etc.).
We plan to tap into the cyclic nature of the economy in order to develop models based on the past data and apply the appropriate one to the present scenario if the parameters show a similar trend as they showed in the past.

VII. CONCLUSION

The Support Vector Machine algorithm seemed to be the most promising base algorithm for stock prediction. Artificial Neural Networks were also beneficial but had high latency. Most researches showed that an optimized version of the SVM was best suited for the job. Therefore it can be used to develop predictive models. Also fundamental analysis for stocks can be incorporated into machine learning through techniques such as feature selection and Sentiment analysis. Thus despite being a pool of seemingly chaotic data, the stock market can be predicted through carefully developed and tested predictive models.

REFERENCES

1. R. Iacomin, "*Stock market prediction*", 2015 19th International Conference on System Theory, Control and Computing (ICSTCC), 2015.
2. P. Kumar and A. Bala, "*Intelligent stock data prediction using predictive data mining techniques*", 2016 International Conference on Inventive Computation Technologies (ICICT), 2016.
3. Y. E. Cakra and B. D. Trisedya, "*Stock price prediction using linear regression based on sentiment analysis*", 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2015.
4. T. M. Sands, D. Tayal, M. E. Morris, and S. T. Monteiro, "*Robust stock value prediction using support vector machines with particle swarm optimization*", 2015 IEEE Congress on Evolutionary Computation (CEC), 2015.
5. Porshnev, I. Redkin, and A. Shevchenko, "*Machine Learning in Prediction of Stock Market Indicators Based on Historical Data and Data from Twitter Sentiment Analysis*", 2013 IEEE 13th International Conference on Data Mining Workshops, 2013.
6. M. Usmani, S. H. Adil, K. Raza and S. S. A. Ali, "*Stock Market Prediction Using Machine Learning Techniques*", 2016 3rd International Conference On Computer And Information Sciences (ICCOINS), 2016.
7. Stock Market, https://en.wikipedia.org/wiki/Stock_market
8. Machine Learning. http://www.techjini.com/blog/machine-learning/
9. Supervised Learning, https://en.wikipedia.org/wiki/Supervised_learning