Insight on Stock Market using Data Mining

Ashvi Rebecka.K
Department of CST
Karunya Institute of Technology And Sciences
Email: ashvirebecka@karunya.edu.in

Joy Jeffery Camillo
Department of CST
Karunya Institute of Technology And Sciences
Email: jojeff374@gmail.com

Volga.A Mohanan
Department of CST
Karunya Institute of Technology And Sciences
Email: volgamohanan@gmail.com

Abstract

A stock market is the aggregation of buyers and sellers of stocks, which represent ownership claims on businesses.Rise and Fall of stock prices are affected by many factors like Government policies, Institutional Buyers, Retail Buyers and many more.There are many algorithms that help investors to have a better insight of stock market.We use three best algorithms and get the average of it.We then show that we have achieved the maximum accuracy possible.

I. INTRODUCTION

A stock exchange is a place or an organization through which individuals and organizations can trade stocks. Many large companies have their stocks listed on a stock exchange. This makes the stock more liquid and thus more attractive to many investors. The exchange may also act as a guarantor of settlement.

The stock market is one of the most important ways for companies to raise money, along with debt markets which are generally more imposing but do not trade publicly.This allows businesses to be publicly traded, and raise additional financial capital for expansion by selling shares of ownership of the company in a public market. The liquidity that an exchange affords the investors enables their holders to quickly and easily sell securities. This is an attractive feature of investing in stocks, compared to other less liquid investments such as property and other immoveable assets. Some companies actively increase liquidity by trading in their own shares.

History has shown that the price of stocks and other assets is an important part of the dynamics of economic activity, and can influence or be an indicator of social mood. An economy where the stock market is on the rise is considered to be an up-andcoming economy. The stock market is often considered the primary indicator of a country's economic strength and development.

Rising share prices, for instance, tend to be associated with increased business investment and vice versa. Share prices also affect the wealth of households and their consumption. Therefore, central banks tend to keep an eye on the control and behavior of the stock market and, in general, on the smooth operation of financial system functions.

Exchanges also act as the clearinghouse for each transaction, meaning that they collect and deliver the shares, and guarantee payment to the seller of a security. This eliminates the risk to an individual buyer or seller that the counterparty could default on the transaction.

The smooth functioning of all these activities facilitates economic growth in that lower costs and enterprise risks promote the production of goods and services as well as possibly employment. In this way the financial system is assumed to contribute to increased prosperity, although some controversy exists as to whether the optimal financial system is bank-based or market-based.

Recent events such as the Global Financial Crisis have prompted a heightened degree of scrutiny of the impact of the structure of stock markets, in particular to the stability of the financial system and the transmission of systemic risk.

Clearly the the stock market investment is uncertain,regardless of the Buyers and Sellers.This uncertainity makes the investment interesting,but also too risky at the same time.The problem is that,the investors cant rely only on the brokerages,as they say it out of experience without proper techniques. Therefore we are in need of a good technique to help the investors.[2]There are many machine learning algorithms and concepts in data mining to help the investors get a better idea on how,where,when to invest.Though the accuracy of these techniques are considerably better,they cant be completely relied on.We propose a model that is a combination of three best algorithms.With the help of the proposed model we get a better accuracy that can be relied on.

II. EXISTING TECHONOLOGIES

In this section we give a general overview about Stock Market, adding the details that will be needed later in this paper.

he stock market works by buyers and sellers who bid on shares of stocks. These are a small piece of ownership of a public corporation. Stock prices usually reflect investors' opinions of what the company's earnings will be.

Traders who think the company will do well in the future bid the price up, while those who believe it will do poorly bid the price down. Sellers try to get as much as possible for each share, hopefully making much more than what they paid for it.

Fig1Graph of Crude Stock

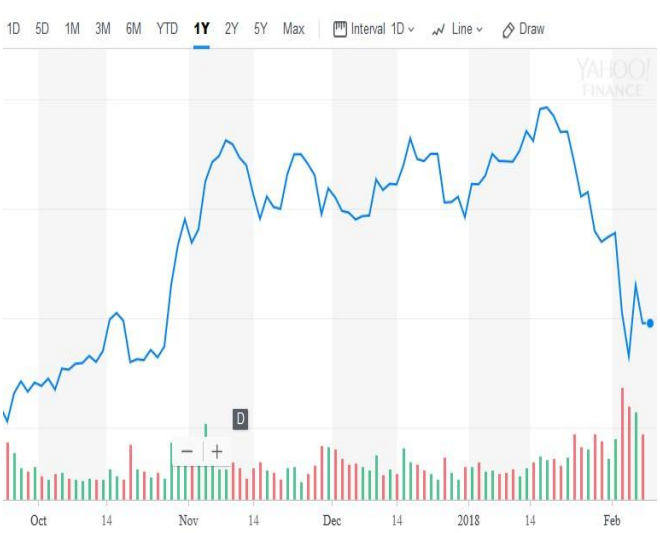
Fig2 Graph of Oil Stock

Fig.3 Graph of Stock Index

A. R-PROGRAMMING

R is a programming language and free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. [5]The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. R ranks 8th in the TIOBE index.

R is a GNU package. The source code for the R software environment is written primarily in C, Fortran, and R. R is freely available under the GNU General Public License, and precompiled binary versions are provided for various operating systems. While R has a command line interface, there are several graphical front-ends available.



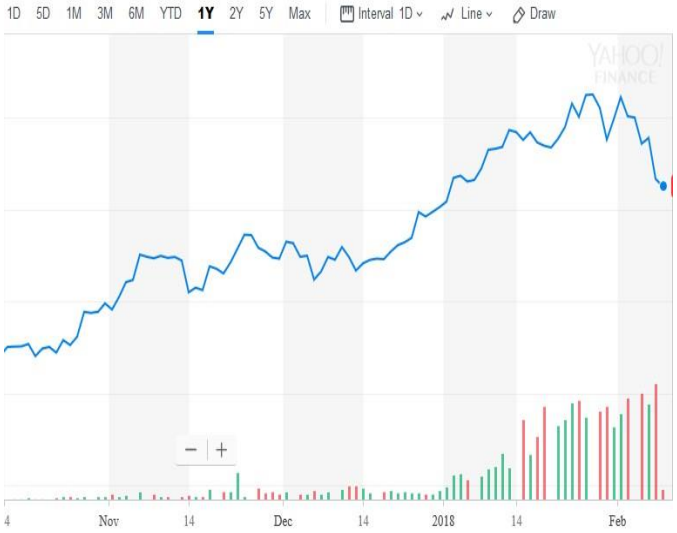


Fig.4 Graph of Gold stock

B . RAPID MINER

RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization. RapidMiner is developed on an open core model.

C. ORANGE

Orange is a component-based visual programming software package for data visualization, machine learning, data mining and data analysis.

Orange components are called widgets and they range from simple data visualization, subset selection and preprocessing, to empirical evaluation of learning algorithms and predictive modeling.

Visual programming is implemented through an interface in which workflows are created by linking predefined or userdesigned widgets, while advanced users can use Orange as a Python library for data manipulation and widget alteration.

D . WEKA TOOL

Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3),for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

- Free availability under the GNU General Public License Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection distribution is sequence modelling[4]. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes ,normally, numeric or nominal attributes, but some other attribute types are also supported.

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like this, and the bird sounds like this. Weka is open source software issued under the GNU General Public License.

WEKA	**ORANGE**	**RAPID MINER**	**R PROGRAMMING**
Regression	Visualization	Predictive Analytics	Graphics Visualization
Clustering	Open source	Data Transformation	Text Mining
Data pre- processing	Platform Independent	Independent	Spatial Data Analysis
Classification	Scripting Interface	Data Loading	Data Manipulation
Visualisation	Extendable	Data Visualization	Clustering
Association		Data Transformation	Graphics

Table.1 Features of the various algorithms

III.COMPARISON TABLE

Procedure	R Programming	RapidMiner	Weka	Orange
Partitioning of dataset into training and testing sets.	Pass (but limited partitioning methods)	Pass (but limited partitioning methods)	Pass (but limited partitioning methods)	Pass (but limited partitioning methods)
Descriptor scaling	Pass	Pass	Fail (cannot save parameters for scaling to apply to future datasets)	Fail (no scaling methods)
Descriptor selection	Fail (no wrapper methods)	Fail	Pass (but is not part of Knowledge Flow)	Fail (no wrapper methods)
Parameter optimization of machine learning/statistical methods	Fail (not automatic)	Pass	Fail (not automatic)	Fail (not automatic)
Model validation using cross-validation and/or independent validation set	Pass (but limited error measurement methods)	Pass	Pass (but cannot save model so have to rebuild model for every future dataset)	Pass (but cannot save model so have to rebuild model for every future dataset)

Table 2. Comparison table on the various algorithms

IV. PROPOSED METHODOLOGIES

The stock market consists of various stocks,but for our proposed technology we are taking into account AAPL,SPY,SPY,CRUDE.There are various data generated per day namely,Open,High,Low,Close,Adjacent Close and Volume.The

main data is the one that is at the end of the day,the Close data.All the Close data of these stocks are taken for the past 10 years,and the forecast is evaluated for 5 days.The Mean Absolute Error and Root Mean Square Error is noted for each stock with regard to each algorithm are got.

Linear Regression,SMOreg,Multilayer Perceptron are one of the best Time Series machine learning algorithms.They produce good accuracy individually,but the investors don't rely on these techniques completely,as it is too risky.Therefore,we propose this methodology,were we combine the three machine learning algorithms to get better accuracy and that eventually gives the investors more insight in forecasting the Stock Market.

A . Linear Regression

When using opinion as primary data, it is necessary to make a suitable analysis of it. A famous example using opinion as data is sentiment analysis which is a linear regression. Stock price dataset was gathered using Yahoo Finance CSV API. Information being collected were the open stock price and close stock price of the companies for each day. Retrieved data was CSV formatted. This data were then being used as the predicted value, also combined with the result of sentiment analysis to create prediction model.

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.This term should be distinguished from multivariate, where multiple correlated dependent variables are predicted, rather than a single scalar variable. In linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, linear regression refers to a model in which the conditional mean of given the value of X is an affine function of X. Less commonly, linear regression could refer to a model in which the median, or some other quantile of the conditional distribution of y given X is expressed as a linear function of X. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis. Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.

These models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.Linear regression models [3]are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares loss function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

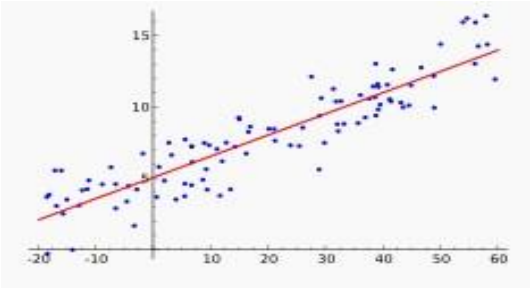


Fig 5 Linear Regression

B. Multilayer Perceptron

They are the neural network,Neural network is a classification algorithm which inspired by human neural network. This method is quite similar to classification algorithms which use linear approach. The algorithm creates a function which calculates weight for each feature Below is the linear equation for Neural Network. There are many methods in determining value of w. One of which is gradient descent. It is an iterative method which updates value of w by minimizing the value of square error.

Below is the function of gradient descent. If the learning rate value is high, the appropriate weight will be reached quickly but the function is more unstable. In the other hand, if the learning rate value is low, the appropriate weight will be reached slowly but the function is more stable. Since the output of the created function is continuous, another function is needed to transform the result into discrete. This function is called activation function.In spite of all the features mentioned for neural networks, building a neural network for prediction is somehow complicated. In order to have a satisfactory performance one must consider some crucial factors in designing of such a prediction model. One of the main factors is the network structure including number of layers, neurons, and the connections. Other factors to be considered are the activation functions in each neuron, the training algorithm, data normalization, selecting training and test set and also evaluation measurements. In the suggested model two neural networks, a multilayer Perceptorn feedforward and an Elman recurrent are used and the back propagation algorithm is used to train these networks. The inputs to the neural networks are the lowest, the highest and the average value in the d previous days. Other information available about the stock market is not used because our goal is to predict the value of the stock share only based on the stock value history. In other words, the proposed model can be viewed as a time series prediction model.

This model uses a three layer neural network in which the input layer has 3d neurons which get the lowest, the highest and the average stock value in the last d days. In the hidden layer there are h neurons which are fully connected to the input and output layers. There is one neuron in output layer which predicts the expected stock value of the next day of the stock market. n this paper the lowest, the highest and the average value of the stock market in the last d days are used to predict the next day's market value[7]. The stock market data have been extracted from Tehran Stock Market website. In this method in contrast with other methods the disorders in the market caused by social or political reasons are not omitted from the data set because we want to predict the value based on the value history.

The simulation data was extracted in 2000 to 2005. In this period of time 1094 companies' shares were traded in Tehran Stock Market. The data used as input to the system are the lowest, the highest, and the average value in the last d days The prediction system predicts the next day's value using the above data. In neural networks applications the input data is usually normalized into the range of [0, 1] or [1,1] according to the activation function of the neurons. So in this paper the value of the stock market is normalized into the range of using the and then the neural networks are trained and tested using the back propagation algorithm.

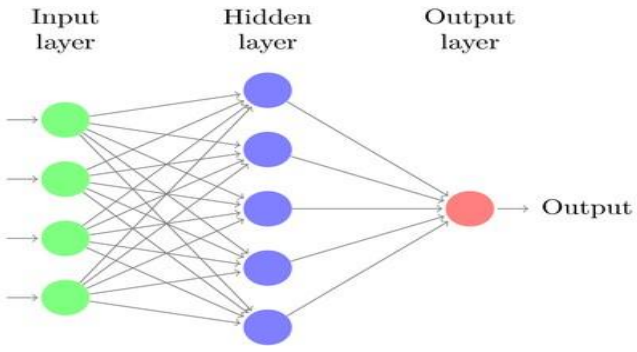


Fig 6 Multilayer Perceptron

C. SMOReg

Support Vector Machines (SVM) are supervised learning models used for classification and regression analysis. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. Support Vector Regression (SVR) is a SVM algorithm to handle nonlinear prediction [5]. SMOreg is an iterative optimization algorithm proposed by Smola and Scholkopf for using SVR regression.And the regression of SMOreg uses constraints structural risk minimization as the model and has the good ability to model regression, prediction with non-linear data.SVM

is an instance based algorithm which created a linear function which maximizes the distance between classes. The algorithm uses instance data on the edge of the class to create the class function. This instance data is called support vector.The linear line is the classifier function. Black points represent instances that are used to create the function and are called support vectors. Distance between the dashed line is called margin. The aim of SVM algorithm is to construct a function that maximize margin.Support Vector Machines (SVM) are supervised learning models used for classification and regression analysis. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. Support Vector Regression (SVR) is a SVM algorithm to handle nonlinear prediction.

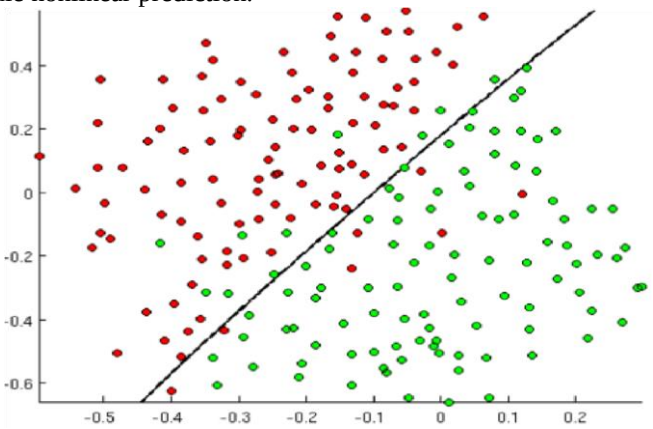


Fig 7 SMOReg

V. RESULTS AND ANALYSIS

The various Time series algorithms were used and the results of the close value are calculated.The close value is forecasted for five days.The close value of the stocks of CRUDE,SPY,GLD,AAPL are forecasted individually using the algorithms .Inorder to forecast ten years of data are fed to the weka tool.The average of all three are taken and found to be more accurate and reliable than each algorithm individually.The proposed system takes into account the best of the algorithms and gives better accuracy,thus helping the investors.

ALGORITHMS	Linear Regression	SMO Reg	Multi Perceptron	My Model
Forecast value 1	269.3122	268.8507	268.8507	269.0045
Forecast value 2	269.0772	268.7009	268.7009	268.8263
Forecast value 3	269.4499	269.1928	269.1928	269.2785
Forecast value 4	269.496	269.8445	269.8445	269.7283
Forecast value 5	269.2231	269.9261	269.9261	269.6918

Table 3. Result analysis of spy

ALGORITHMS	Linear Regression	SMO Reg	Multi Perceptron	My Model
Forecast value 1	163.8273	163.2753	163.278	163.973
Forecast value 2	163.683	163.825	163.936	163.6432
Forecast value 3	164.7883	164.572	163.538	163.347
Forecast value 4	163.293	163.286	163.926	164.763
Forecast value 5	164.8233	163.538	164.5426	164.234

Table 4. Result analysis of aapl

ALGORITHMS	Linear Regression	SMO Reg	Multi Perceptron	My Model
Forecast value 1	12.7366	12.6673	12.5278	12.6439
Forecast value 2	12.6986	12.6308	12.2735	12.5343
Forecast value 3	12.6898	12.5919	11.9606	12.4141
Forecast value 4	12.6868	12.5372	11.6433	12.2891
Forecast value 5	12.6768	12.5086	11.3408	12.1754

Table 5. Result analysis of gld

ALGORITHMS	Linear Regression	SMO Reg	Multi Perceptron	My Model
Forecast value 1	124.8784	125.4938	123.9642	124.7788
Forecast value 2	124.5249	125.6216	122.8808	124.3424
Forecast value 3	124.1537	125.7457	121.7828	123.8941
Forecast value 4	123.8412	125.8799	120.7407	123.4873
Forecast value 5	123.5693	126.0125	119.6836	123.0885

Table 6. Result analysis of crude oil

VI. CONCLUSION

This research model shows that the outcomes got by the Time series techniques individually can be improved when the close data of the stocks namely
SPY,AAPL,GLD,CRUDE is taken and an average of the techniques are calculated.The technique's individual outcome are not accurate enough and reliable.Therefore,this proposed model gives better accuracy and reliability.This is achieved by the usage of the combination of the best machine learning algorithms.In this research which utilized the WEKA Tool, the combined average has

outperformed the other techniques individually.The close data of the various stocks are noted and evaluation of the Mean Absolute Error and Root Mean Square Error are calculated,and found to the best when combining the three algorithms.

Since there are many other evolving machine learning algorithms, further research can be conducted to compare the effects of various Time series algorithm with regard to the insight of stock market.

VII. REFERENCES

[1] Beechey M, Gruen D, Vickrey J. (2000). The Efficient Markets Hypothesis: A Survey. Reserve Bank of Australia

[2] Lo, A.W. and Mackinlay, A.C. A Non-Random Walk Down Wall Street 5th Ed. Princeton University Press, 2002

[3] Graham, Benjamin; Dodd, David (December 10, 2004). Security Analysis. McGraw-Hill. ISBN 9780071448208.

[4] Walsh, Ciaran (2003) Key Management Ratios, Third Edition, Prentice Hall.Shefrin, Hersh (2002) Beyond Greed and Fear: Understanding behavioral finance and the psychology of investing. Oxford University Press.O'Shaughnessy, James (2009).

[5] Predicting the Markets of Tomorrow: A Contrarian Investment Strategy for the Next Twenty Years, Penguin Group. ISBN 1591841089.

[6] Kirkpatrick and Dahlquist. Technical Analysis: The Complete Resource for Financial Market Technicians. Financial Times Press, 2006, page 3. ISBN 0-13153113-1.MacKay, D.J.C. (2003).

[7] Information Theory, Inference, and Learning Algorithms, Cambridge University Press. ISBN521-64298-1.Alpaydın, Ethem (2004) Introduction to Machine Learning (Adaptive Computation and Machine Learning), MIT Press.