Stock Prediction using Machine Learning a Review Paper

Nirbhey Singh Pahwa
Student, Information Technology (I.T.)
Vidyalankar Institute of Technology
Mumbai, Maharashtra, India.

Neeha Khalfay
Student, Information Technology (I.T.)
Vidyalankar Institute of Technology
Mumbai, Maharashtra, India.

Vidhi Soni
Student, Information Technology (I.T.)
Vidyalankar Institute of Technology
Mumbai, Maharashtra, India

Deepali Vora
Assistant Professor,
Information Technology (I.T.)
Vidyalankar Institute of Technology
Mumbai, Maharashtra, India

ABSTRACT
Every day more than 5000 trade companies enlisted in Bombay stock Exchange (BSE) offer an average of 24,00,00,000+ stocks, making an approximate of 2000Cr+ Indian rupees in investments. Thus analyzing such a huge market will prove beneficial to all stakeholders of the system. An application which focuses on the patterns generated in this stock trade over the period of time, and extracting the knowledge from those patterns to predict future behavior of the BSE stock market is essential. An application representing the information in visual form for user interpretation to buy and to sell a specific company's stock is a key requirement.

Such an application based on machine learning algorithms is the right choice in current scenario. This paper surveys the machine learning algorithms suitable for such an application; as well it discusses what are the current tools and techniques appropriate for its implementation.

General Terms
Support Vector Machine (SVM), Support Vector Regression (SVR) and stock market.

Keywords
Machine learning, review paper, stock prediction, machine learning algorithms, supervised learning, unsupervised learning, supervised learning algorithms, regression, classification, regression algorithm, Support Vector Machine (SVM), Support Vector Regression (SVR), classification, linear regression, logistic regression, types of regression, types of classification, types of programming languages for machine learning, types of libraries for machine learning, types of libraries for graphing, types of libraries for analysis.

1. INTRODUCTION
Machine learning can be defined as the data which is obtained by knowledge extraction. Machines don't have to be programmed explicitly instead they are trained to make decisions that are driven by data. Instead of writing a code for every specific problem, data is provided to the generic algorithms and logic is developed on the basis of that data. When a machine improves its performance based on its past experiences it can be said that machine has truly learnt.

The technique for most accurate prediction is by learning from past instances, and to make a program to do this is best possible with machine learning techniques. Any machine learning technique (supervised or unsupervised) is efficient enough to generate rules for programs, in consideration with present ones to take a better decision. In this scenario, the decision is whether the stock will increase or decrease (Stock analysis).

2. MACHINE LEARNING ALGORITHMS

2.1 Unsupervised learning
When the dataset is not well defined or very hard for interpretation, it is called unsupervised learning. The labels for the data are not defined. There no right way to divide data set except performing iterations. Thus, in supervised learning the input is used to generate a structure by looking at the relation of the input itself.

For example, Classification of animals. [4]

According to this research, unsupervised learning is not advisable for prediction.

2.2 Supervised learning
Supervised learning can be said as function approximation, training examples lead to function generation. If the learning is done with right training set, a well behaved function can be expected. Supervised learning grows consistently with the data. It is a type of induction learning, and it causes biased supervised learning sometimes.

E.g.: The function generated with supervised learning will be

X^2, if X is the input value and the output is self-multiplied.

Since, there is well defined data available from BSE itself and which is in well-defined numeric form it would be beneficial to use supervised learning algorithms. Supervised learning algorithms are of two variants: [3]

1. Regression.
2. Classification

2.2.1 Regression algorithm
The method of Support Vector Classification (SVC) can be used to solve regression problems. When Support Vector Machine (SVM) is used to solve regression problems the method is referred as Support Vector Regression (SVR).

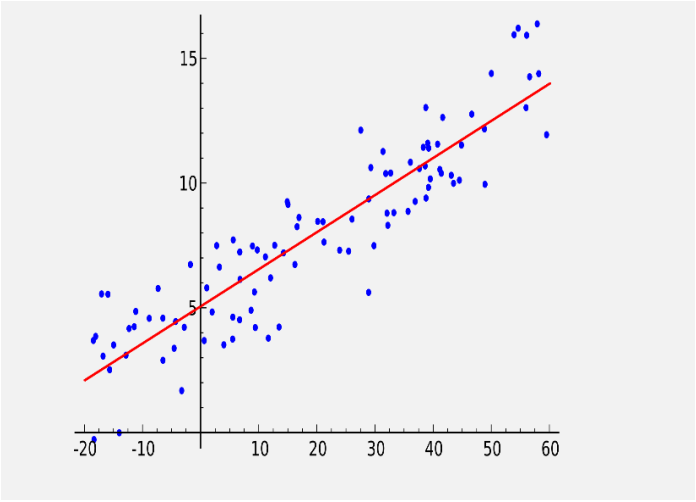
The model produced by SVC depends only on the training data, because the factor of cost of model building does not care about training points that lie outside the margin.

Similarly, the model produced by SVR only depends on the training (Subset) data, as the cost factor for building the model does not consider any training data close to the model prediction. [5]

2.2.1.1 Regression problems

Input is mapped by labels.

Input mapped to large and many data set.



Use regression when the output can be mapped to many outlets.

Figure 1

2.2.1.2 Types of regression

The seven types of regression are briefly explained and compared in the following tables 1 (a) and 1 (b).

2.2.2 Classification algorithm

Classification is a type of supervised learning (machine learning) in which some decision is taken or prediction is made on the basis of information which is currently available and the procedure of carrying out classification is a formal method which is used for constantly making such judgements in different and new situations. The formation of a classification method from a data set for which the true classes are known is also known as pattern recognition, supervised learning or discrimination (in order to differentiate it from unsupervised learning in which the classes are always inferred from the data). Classification is used in many situations like the most difficult situations arising in science, industry and commerce can be determined by classification or decision problems which use complex and often very extensive data.

2.2.2.1 Classification problems

Input is mapped to label.

Input to small and discreet data set.

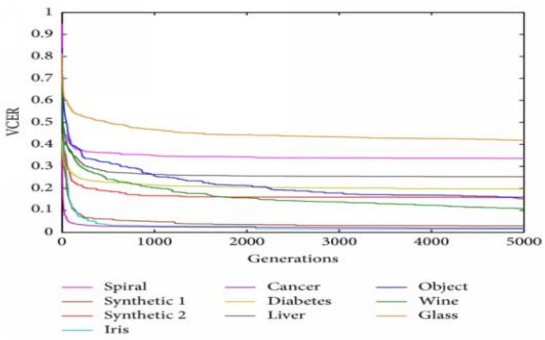


Figure 2

2.2.2.2 Types of classification

The different types of classifications are briefly explained and compared in the following table 2. [7]

3. ALGORITHMS AND TOOLS FOR THIS SYSTEM

3.1 Linear regression

The most commonly known modelling technique is linear regression. In this technique, the first (dependent variable) is continuous, the second variables (independent variable) can be continuous or discrete and this leads to a linear line which is the nature of this regression.

It establishes a relationship between the first variable (dependent variable (Y)) and one second variables (independent variables (X)) and making a straight line which is best fit after computation (which is the regression line).

It is given by an equation:

$Y = (a + b) * X + e$, where 'a' is the intercept, 'b' is the slope of the line and 'e' is the error term. Given equation is also used to predict the value of target variable, on given predictor variable(s).

The major difference between the simple linear regression and multiple regression is that, multiple regression supports more than one independent variables, but simple linear regression has only one independent variable which it can handle.

To obtain best fit line, following procedures are to be done. This can be accomplished by the least square method. It is the most easy and common way for making a regression line. It computes the best-fitting line for the taken data by reducing to the minimum the addition of the squares of the vertical deviations, from each point to the produced line. Since, the deviations are first squared, when summed; positive and negative values do not cancel out.

The following equation is used for calculating the line plotting:

$$\min(w) \left|\left|Xw - y\right|\right|^2$$

Points to consider before considering linear regression:

- A linear relationship between the given independent and the taken dependent variables is essential.
- It can suffer from multicollinearity, heteroscedasticity, etc.
- Outliers can impact linear regression in a huge way, which can even lead to wrong predictions.
- Step wise approach also uses selection of most significant independent variables.

3.2 Logistic regression

It is used to find the probability of how much chance there is for cases such that the event is success and the same event is a failure. Logistic regression can be implemented when the dependent variable is binary (one of two values) in nature, that is, it can have at most two values. In this example the value of 'Y' can be 0 to 1, it is represented by the equation: [9]

Odds = probability of event occurring divided by the probability of event not occurring

$$\text{Odds} = \frac{\text{p}}{1-\text{p}}$$

$$Ln(Odds) = Ln\left(\frac{P}{1-P}\right)$$

$$Log(P) = Ln\left(\frac{P}{1-P}\right) = B0 + (B1 * 1) + \cdots + (Bk * k)$$

Where, P is the probability of interested characteristic presence. As there is a binomial distribution of dependent variable implemented, there has to be a link function which will be best fitted for the distribution, which is the logit function. The above equation has, the parameters selected to max the chance of getting the sample values instead of minimizing the addition of squared errors (as seen in the ordinary regression). [8]

Points to consider before considering logistic regression:

- It is widely used for classification problems and does not really need linear relationship between the dependent and the independent variables. It can take many different types of relationships since it enforces a non-linear log transformation for predicting the odds ratio.
- To remove over fitting as well as under fitting, all significant variables should be included. A better approach to make sure this practice is by using a step wise method to compute the logistic regression.
- It needs huge sample sizes. Since, maximum likelihood that calculations are less accurate at low sample sizes in comparison to the ordinary least square.
- No multi collinearity i.e. the independent variables need not be inter-related with each other. But there are still options to consider interaction impacts of categorical variables in computation and modelling.
- It will be called as Ordinal logistic regression when the values of dependent variable are ordinal. [6]

3.3 Tools for implementation

The different types of development software which can support the system are briefly explained and compared in the following table 3. [1]

The different types of libraries are briefly explained and compared in the following table 4. [2]

The different types of tools are briefly explained and compared in the following table 5.

4. SYSTEM DIAGRAM

With above knowledge in consideration and undertaking the tables as reference, a proposed system and its diagram is shown below. The system will work on a comma separated variable (CSV) file, which will have a record of all the dates and its crude data of open, high, low, etc. Out of this crude data, knowledge will be extracted by performing data pre-processing and refining to predict a close information for requested date of future. The CSV files are provided by the BSE itself.

Once the knowledge is available, it will be feed to the SVM algorithm to perform stock prediction and give a data visualization using python, this investment prediction will be sub-divided into different time frames (months, days, hours) and a suitable advice from the prediction can be obtained by the consumer. The system diagram is as show below:

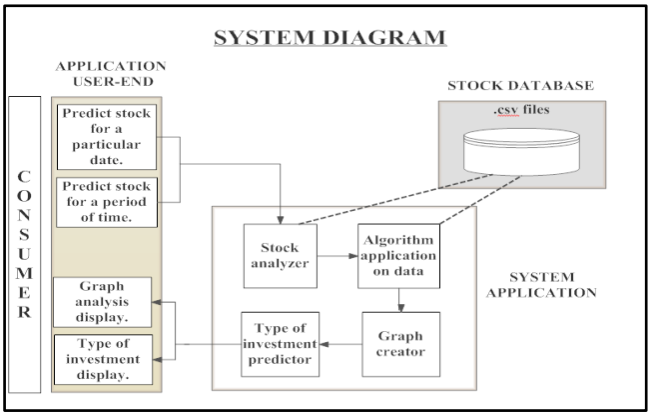


Figure 3

5. CONCLUSION

This paper summarizes important techniques in machine learning which are relevant to stock prediction. The paper recommends use of linear regression and logistic regression for stock prediction and stock analysis and this study recommends SVM to obtain accurate results. A constraint to this conclusion is the necessity of the dataset used in prediction to be classification friendly. The paper summarizes the tools which can be used for implementation of machine learning algorithms. All the tools support regression and classification algorithms, users can choose any tool based on their familiarity and convenience. The paper proposes a system to extract knowledge from data and performing a prediction to advise the consumer for investments.

Table 1 (a): Types of regression

Properties	Linear regression	Logistic regression	Polynomial regression	Stepwise regression
Features	In this technique, the first (dependent variable) is continuous, the second variables (independent variable) can be continuous or discrete, and this leads to a linear line which is the nature of this regression.	It is used to find the probability of how much chance is there for cases such that the event is a success and the same event is a failure. Logistic regression is used when the dependent variable is binary in nature, that is, it can have at the most two values.	If the power of independent variable is greater than one, the regression equation is called as polynomial regression.	Stepwise regression is used when there are many independent variables. An automatic process (steps) is used for the selection of independent variables, and no involvement or human intervention is needed.
Advantages	It is implemented, when relationships of the independent variables and the dependent variable are linear (almost), and it shows optimal results. If the datasets are well defined, there is no better regression than linear regression.	Mostly used in classification problems. Linear relationship between the dependent and the independent variables is not necessary.	The best fit line need not be a straight line. It is instead a curve which fits accurately into the data points.	It includes and excludes predictors as required for each step. The aim of this technique is to maximize the prediction power with minimum numbers of predictor variables. Higher dimensionality of data set can be handled by this technique.
Disadvantages	A linear relationship between the given independent and the taken dependent variables is essential. It can suffer from multicollinearity, heteroscedasticity, etc.	It needs huge sample sizes. Since, maximum likelihood calculations are less accurate at low sample sizes in comparison to the ordinary least square.	Higher polynomials can end up producing weird results on extrapolation.	It often has many potential predictor variables but very less data to estimate coefficients correctly. Adding more data does not help much, if at all.

Table 1 (b): Types of regression

Properties	Ridge regression	Lasso regression	Elastic Net regression
Features	It is a technique used when the data suffers from multicollinearity. In multicollinearity, even if the least squares estimates are not biased, their variances are huge which separate the observed value far from the true value, by involving a bias to the regression estimates, standard errors can be reduced.	Lasso regression is Least Absolute Shrinkage and Selection Operator. It penalizes the exact size of the coefficients. It also is capable of reducing the variability and improving the accuracy for the linear regression models.	It is a hybrid of Ridge Regression and Lasso techniques. It is trained with L1 and L2 prior as regularize. Elastic-net is most useful, if there are multiple features which are interrelated.
Advantages	It diminishes the value of coefficients but does not reach zero, which shows the no feature selection feature.	It shrinks coefficients exactly to zero, which shows the feature selection. This is a regularization method and uses L1 regularization.	It allows Elastic-Net to extend some of Ridge's stability under rotation. There is no limit to the number of selected variables.
Disadvantages	Normality cannot be assumed.	If group of predictors are very interrelated, it picks only one among all of them and reduces others to zero.	It suffers with double shrinkage sometimes. It does encourage group effect when there is highly correlated variables.

Table 2: Types of classification

Properties	Support Vector Machine (SVM)	Bayesian's Classifier	Decision Tree
Definition	A Support Vector Machine (SVM) implements classification by finding the hyperplane that maximizes the margin	The Naïve Bayesian classifier is classification method based on Bayes' theorem with independence	Decision tree is used to build classification models in the form of a tree structure. It breaks down a

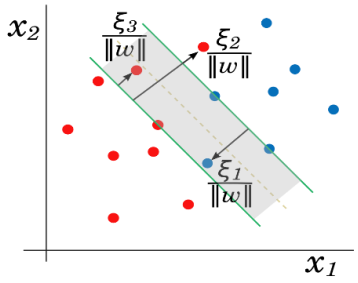
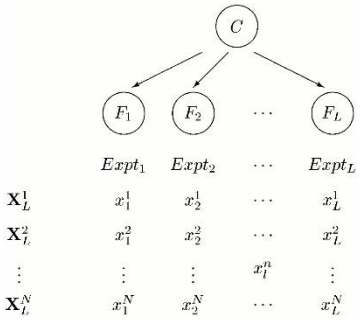
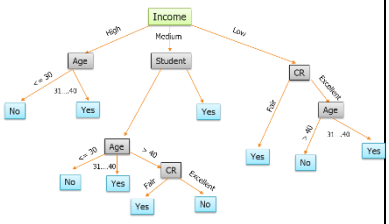
	between the two classes. The vectors or cases that represent the hyperplane are the support vectors.	assumptions between predictors. This model is easy to build, with no perplexing iterative parameter estimation which makes it particularly useful for very large datasets. Although it is simple, the Naïve Bayesian classifier often does surprisingly well and is widely used because it often outruns more practiced classification methods.	dataset into smaller subsets simultaneously and an associated decision tree is incrementally developed. The topmost decision node in a tree which resembles or correlates to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.
Diagram			
Advantages	Widen the margin between two classes in the feature space characterized by a kernel function. They are robust with respect to high input dimension. High accuracy. Good for large feature sets.	Easy to implement. Satisfactory results obtained in most of the cases.	Easy to understand. Easy to generate rules. Reduces problem complexity.
Disadvantages	Difficult to combine background knowledge. Sensitive to outliers. Hard to interpret. Memory-intensive.	Assumptions: Class conditional independence, which causes loss of accuracy. Practically, dependencies exist among variables which cannot be modelled by Naïve Bayesian Classifier.	Training time is comparatively expensive. A document is only connected with one branch. While constructing the tree, once a mistake is made at a higher level node of the tree, any sub-tree below it is wrong. Does not handle continuous variable properly. May suffer from overfitting.

Table 3: Tools (Types of programming language)

Properties	**Python**	**Java**	**R**	**.Net**
Features	Free and Open Source. It is a high-level language. It provides portability. It is interpreted. It provides extensive libraries.	Platform independent. Portable. Multi-threaded. Distributed. Robust.	Free and Open Source. Flexible and powerful. It is cross-platform. Interactive language. Ease due to the package system.	Windows application. Console application. Languages supported IntelliSense Assemblies.

	Indentation.			
Advantages	Efforts required to write a program in Python is less as compared to other languages. Portable. Its integration with other languages like C/C++. Python being a general purpose language helps us attain maximum flexibility. Its emphasis on productivity and easiness in readability.	Elimination of pointers and the replacement of the complex multiple inheritance by interface provides simplicity. The "Write Once Run Anywhere" feature provided by Java Networking capability i.e. the distributiveness Reliability i.e. the robustness offered by Java.	Free and Open Source Graphical capabilities. Packages available for data mining, spatial analysis. It can easily import data from CSV les, SAS, SPSS or from Microsoft Excel, MySQL or SQLite directly. Graphics output in the form of PNG, PDF, JPG and SVG formats and table output for LATEX and HTML is provided by R.	Compatible with multiple languages. Ease in interfacing with Microsoft or Windows Seamless integration of language. The drag and drop capability as well as the blueprints offered. Easiness provided by the separation of the HTML code from the source code.
Disadvantages	Pace decreases as it is an interpreted language. Design of the language is a drawback since the testing required is more and the errors are shown during the runtime. Absence felt in mobile computing and browsers because of security.	Since it is a multi-platform language, it is slow and also occupies more memory space. Problems caused for low level programming. Limitations for applet caused due to security.	Poses difficulty for the unexperienced users. Provides several simple GUIs which include point and click interactions but lacks to deliver the commercial offerings. R commands occupy all the available memory which is a drawback while doing data mining.	Does not support multi-platform. Excessive use of system resources. Application pauses from execution due to garbage collection. Installation process has to be manually done since it is not predefined.

Table 4: Tools (Types of library for analysis)

Properties	**Scikit-learn**	**Pandas**	**Theano**	**NLTK**
Features	Tools available for data analysis and mining. Primary focus on modelling of data. Open source and can be used commercially. Built using SciPy, matplotlib and NumPy. Accessibility to one and all. Reusability.	It provides label for data. The table format for data which includes label for columns and indexed rows.	Integrated with NumPy. Optimizes speed and stability. Ability to perform derivative for functions with more than one input. Self-verification and unit testing. High-speed performance.	Freely and openly available It provides Python program which incorporates human language data. Interface having an ease of use provides many resources, text processing libraries for parsing, tagging, classification, tokenization, stemming and semantic reasoning. It has predefined graphical demonstrations and sample data.
Advantages	Cleanliness offered by the API design. Robustness. High speed. Ease of Use. Well documented. Active development and well supported.	Easy to perform an operation since Pandas provide table format for data. Supports different data types. Built-in functionality for many data processing applications. Due to its support in storage and memory functions, large scale of data can be handled.	It provides differentiation automatically even when it is not needed. Numpy's syntax is supported and borrowed by its mature and an intuitive tensor interface. It saves a lot of development time which helps one to focus on one's system rather	It is fully self-contained. It provides raw versions of real-world data in the form of trained models as well as functions that can be used as building blocks for common NLP tasks.

			than concentrating on optimizing graphs.	
Disadvantages	Restrictions on choice of language. Not scalable enough.	Addition of syntactic noise. The list throws away an extra holding space which is temporary for the values which are common to stay with the Pandas data frame.	Graph optimization results in the increase in the compilation time. Theano scan are really slow in speed.	It has to work with different variable types. Mandatory usage of regular expressions, tagging, stemming, chunking and context-free and feature based grammars. Raw text needs to be processed. It has to discover parts of speech tags.

Table 5: Tools (Types of library for graphing)

Properties	**Matplot lib**	**plot.ly**	**Bokeh**
Features:	Works with labelled data similar to DataFrames in Panda.s Not only can it cycle colors but also line styles and hatches. Provides selection for backend Integration with LaTeX. It can have multiple plots on the same axes. Multiple subplots can be obtained in a single figure. 3D plotting.	It is an online analytics and data visualization tool. Online graphing, analytics, and statistics tools are provided not only for individuals/collaboration, as well as scientific graphing libraries for Python Its graphical user interface which provides stats tools for analyzing and importing data into a grid. To create graphs more efficiently, they can be either embedded or downloaded. Responsible for providing API libraries for Python, MATLAB, R, Julia, Node.js and Arduino It helps with styling interactive graphs using IPython. Apps developed using Plotly for Google Chrome. Graphs and dashboards are created using the open source JavaScript library.	It provides elegant construction of novel graphics in the D3.js style and provides extension of this function to large or streaming datasets along with ugh performance. It provides aid to create interactive plots, data applications and dashboards efficiently. Its syntax is similar to R/ggplot users. It is fully open-sourced.
Advantages:	Default plot styles are available with built-in code. Deep integration with Python. The programming interface is Matlab-style.	Allowance for changing the colors and style for the graph. Also allows to change the plot type. Title for the graph and label for the axes can be provided. Hovering the mouse cursor on the line will provide the values of the points.	It provides visualization library that targets web browsers for presentation.
Disadvantages:	It is unpredictable for dynamic, interactive plots. Very much reliant on packages like Numpy. It works only for Python.	This technology is relatively new. It requires internet for viewing data into graphical representation.	Community support not prompt. Relatively new library with no history for credibility.

6. REFERENCES

[1] Author: W. HuangResearch paper: Forecasting stock market movement direction with support vector machine.Journal: Computers & Operations Research

[2] Author: J. MoodyResearch paper: Learning to trade via direct reinforcement. Journal: IEEE Transactions on Neural Networks

[3]https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

[4]https://azure.microsoft.com/en-in/documentation/articles/machine-learning-algorithm-choice/

[5] Author: Yusuf Perwej, Asif PerwejResearch paper: Prediction of the Bombay Stock Exchange (BSE) Market Returns Using Artificial Neural Network and Genetic Algorithm.Journal: Scientific Research

[6] Author: K. Senthamarai Kannan, P. Sailapathi Sekar,M.Mohamed Sathik and P. Arumugam Research paper: Financial Stock Market Forecast using Data MiningTechniquesJournal: International Multi-Conference of Engineers and Computer Scientists 2010 Vol I,IMECS 2010, March 17-19,2010, Hong Kong.ISSN:2078-0966

[7] Author: Zahid Iqbal, R. Ilyas, W. Shahzad, Z. Mahmood and J AnjumResearch paper: Efficient Machine Learning Techniques forStock Market PredictionJournal: Int. Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 3, Issue 6, Nov-Dec 2013, pp.855-867.

[8] Author: Marc-André MittermayeResearch paper: Forecasting Intraday Stock Price Trends with Text Mining TechniquesJournal: Hawaii International Conference on System Sciences –2004.

[9] Author: Prakash Ramani, Dr. P. D. MurarkaResearch paper: Stock Market Prediction Using Artificial Neural NetworkJournal: International Journal of Advanced Research in Computer Science and Software Engineering. ISSN: 2277-128x, Volume 3, Issue 4, April 2013