Stock Prediction Using a Machine Learning

Abhishek shukla
Computer Science and Engineering, Raj Kumar Goel Institute Technology, Ghaziabad, Uttar Pradesh, India

***Abstract* - Stock prediction is a very important for stock price surveillance. While as a canonical pattern recognition problem, it's very difficult due to various predictions of stock value of a company. To solve this problem In Stock Market Prediction, the aim is to predict the future value of the financial stocks of a company. The recent trend in stock market prediction technologies is the use of machine learning which makes predictions based on the values of current stock market indices by training on their previous values. Machine learning itself employs different models to make prediction easier and authentic. This focuses on the use of Regression and LSTM based Machine learning to predict stock values. Factors considered are open, close, low, high and volume. The data stored can be visualized through a web application that uses HTTP GET requests for requesting the data stored in MySQL to a HTML template for rendering dynamic High Charts.**

I.INTRODUCTION

A correct prediction of stocks can lead to huge profits for the seller and the broker. it is brought out that prediction is chaotic rather than random, which means it can be predicted by carefully analyzing the history of respective stock market. Machine learning is an efficient way to represent such processes. It predicts a market value close to the tangible value, thereby increasing the accuracy.

Machine learning can be defined as the data which is obtained by knowledge extraction. Machines do not have to be programmed specifically 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 generated 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).

II. MACHINE LEARNING ALGORITHMS

A. Unsupervised learning

When the data set is not well defined for interpretation, it is called unsupervised learning. The labels for the data are not defined. There is 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.

B. 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 $X2$, 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:

1. Regression.
2. Classification

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.

1.1 Regression problems
Input is mapped by labels.
Input mapped to large and many data set.

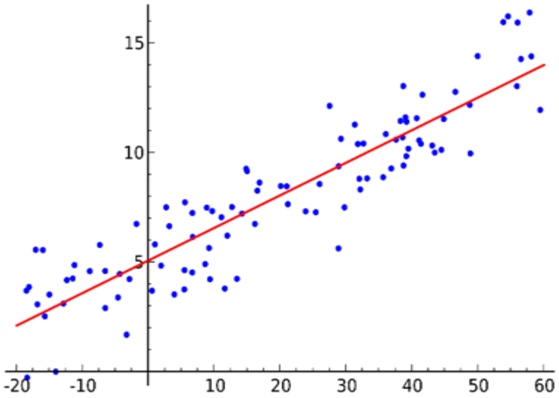


Fig 1

1.2 Types of regression
The seven types of regression are briefly explained in following:
A. 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.
Advantages:
1. It is implemented, when relationships of the independent variables and the dependent variable are linear (almost), and it shows optimal results.
2. If the datasets are well defined, there is no better regression than linear regression.

Disadvantages:
1. A linear relationship between the given independent and the taken dependent variables is essential.
2. It can suffer from multiple correlation, homoscedastic, etc.

B. Logistic 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.
Advantages:
1. Mostly used in classification problems. Linear relationship between the dependent and the independent variables is not necessary.

Disadvantages:
1. It needs huge sample sizes.
2. Since, maximum likelihood calculations are less accurate at low sample sizes in comparison to the ordinary least square.

C. Polynomial regression
If the power of independent variable is greater than one, the regression equation is called as polynomial regression.
Advantages:
1. The best fit line need not be a straight line. It is instead a curve which fits accurately into the data points.

Disadvantages:
1. Higher polynomials can end up producing weird results on extrapolation.

D. Stepwise 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:
1. It includes and excludes predictors as required for each step.
2. The aim of this technique is to maximize the prediction power with minimum numbers of predictor variables.
3. Higher extensity of data set can be handled by this technique.

Disadvantages:
1. It often has many potential predictor variables but very less data to estimate coefficients correctly.
2. Adding more data does not help much, if at all.

E. Ridge regression
It is a technique used when the data suffers from multiple correlation.

In multiple regression, 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.

Advantages:

1. It diminishes the value of coefficients but does not reach zero, which shows the no feature selection feature.

Disadvantages:

1. Normality cannot be assumed.

F. Lasso regression

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.

Advantages:

1. It shrinks coefficients exactly to zero, which shows the feature selection.
2. This is a regularization method and uses L1 regularization.

Disadvantages:

1. If group of predictors are very interrelated, it picks only one among all of them and reduces others to zero.

G. Elastic Net regression

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:

1. It allows Elastic-Net to extend some of Ridge‟s stability under rotation.
2. There is no limit to the number of selected variables.

Disadvantages:

1. It suffers with double shrinkage sometimes.
2. It does encourage group effect when there is highly correlated variables.

1.2.1 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 decisions 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 (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.

1.2.3. Classification problems

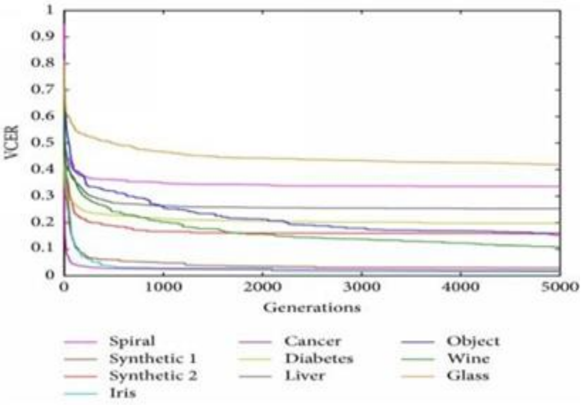


Fig 2

1.2.2.1 Types of classification

The different types of classifications are briefly explained in the following:

A. Support Vector Machine (SVM)

A Support Vector Machine (SVM) implements classification by finding the hyperplane that maximizes the margin between the two classes. The vectors or cases that represent the hyperplane are the support vectors.

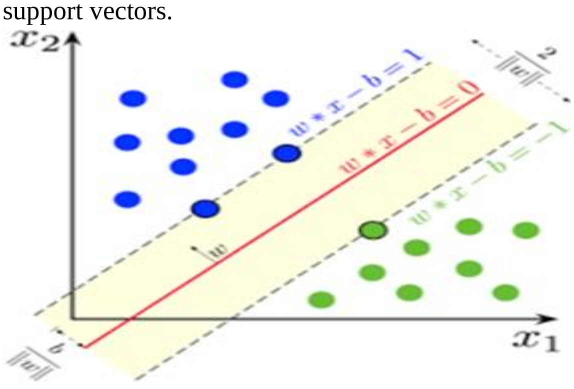
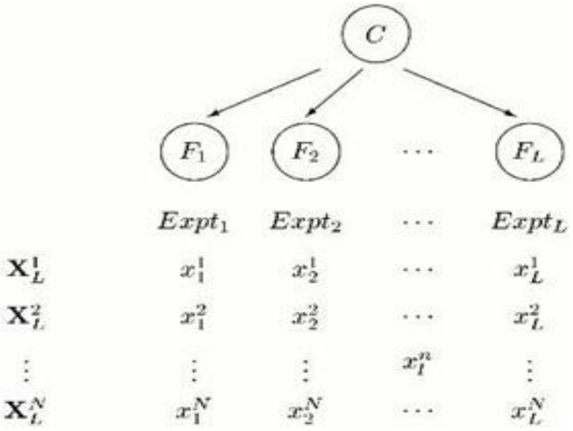


Fig.3

Advantages:
1. Widen the margin between two classes in the feature space characterized by a kernel function.
2. They are robust with respect to high input dimension.
3. High accuracy.
4. Good for large feature sets.

Disadvantages:
1. Difficult to combine background knowledge.
2. Sensitive to outliers. Hard to interpret.
3. Memory-intensive.

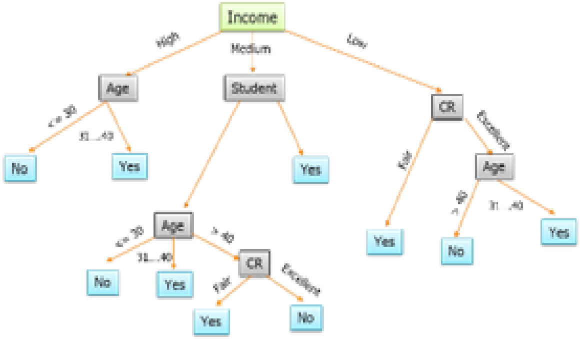
B. Bayesian's Classifier
The Naïve Bayesian classifier is classification method based on Bayes‟ theorem with independence assumptions between predictors. This model is easy to build, with no perplexing iterative parameter estimation which makes it particularly useful for very large datasets. Although it is simple, the Naïve Bayesian classifier often does surprisingly well and is widely used because it often outruns more practiced classification methods.



Advantages:
1. Easy to implement.
2. Satisfactory results obtained in most of the cases.

Disadvantages:
1. Assumptions: Class conditional independence, which causes loss of accuracy.
2. Practically, dependencies exist among variables which cannot be modelled by Naïve Bayesian Classifier.

C. Decision Tree
Decision tree is used to build classification models in the form of a tree structure. It breaks down a data set 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.



Advantages:
1. Easy to understand. Easy to generate rules.
2. Reduces problem complexity.

Disadvantages:
1. Training time is comparatively expensive.
2. A document is only connected with one branch.
3. While constructing the tree, once a mistake is made at a higher-level node of the tree, any sub-tree below it is wrong.

III. ALGORITHMS AND TOOLS FOR THIS SYSTEM

A. Linear regression
The most commonly known modelling technique is linear regression. In this technique, the first dependent variable is continuous, the second variable 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 + bX 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 get 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 added; positive and negative values do not cancel out.

Points to consider before considering linear regression:

- Point A linear relationship between the given independent and the taken dependent variables is essential.
- It can suffer from multiple correlation, heteroscedasticity, etc.
- A linear relationship between the given independent and the taken dependent variables is essential. It can suffer from multiple correlation, 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.

B. Logistic regression

It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables, 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:

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

Odds = $p/1 - p$

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 good shaped for the distribution, which is the log it functions. The 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).

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 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.

C. Tools for implementation

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

The different types of libraries are briefly explained in following:

A. Scikit-learn

Features:

1. Tools available for data analysis and mining.
2. Primary focus on modelling of data.
3. Open source and can be used commercially.
4. Built using SciPy, matplotlib and NumPy.
5. Accessibility to one and all.

Advantages:
1. Cleanliness offered by the API design.
2. Robustness. High speed. Ease of Use.
3. Well documented.
4. Active development and well supported.

Disadvantages:
1. Restrictions on choice of language.
2. Not scalable enough.

B. Pandas

Features:
1. It provides label for data.
2. The table format for data which includes label for columns and indexed rows.

Advantages:
1.Easy to perform an operation since Pandas provide table format for data.
2. Supports different data types.
3.Built-in functionality for many data processing applications.
4. Due to its support in storage and memory functions, large scale of data can be handled.

Disadvantages:
1. Addition of syntactic noise.
2. The list throws away an extra holding space which is temporary for the values which are common to stay with the Pandas data frame.

C. Theano

Features:
1. Integrated with NumPy.
2. Optimizes speed and stability.
3. Ability to perform derivative for functions with more than one input.
4.Self-verification and unit testing.

Advantages:
1. It provides differentiation automatically even when it is not needed.
2. Numpy's syntax is supported and borrowed by its mature and an intuitive tensor interface.

Disadvantages:
1. Graph optimization results in the increase in the compilation time.
2. Theano scan are really slow in speed.

D.NLTK

Features:
1. Freely and openly available.
2. It provides Python program which
Incorporates human language data.
3.It has predefined graphical demonstrations and sample data.

Advantages:
1. It is fully self-contained.
2. 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.

Disadvantages:
1. It has to work with different variable types.
2. Mandatory usage of regular expressions, tagging, stemming, chunking and context-free and feature based grammars.
3. Raw text needs to be processed.
4. It has to discover parts of speech tags.
The different types of tools are briefly explained and compared in the following table 5.

A. Matplot lib

Features:
1. Works with labelled data similar to DataFrames in Panda.s
2. Not only can it cycle colors but also line styles and hatches.
3. Provides selection for backend Integration with LaTeX.
4. It can have multiple plots on the same axes.
5. Multiple subplots can be obtained in a single figure.
6. 3D plotting.

Advantages:
1. Default plot styles are available with built-in code.
2. Deep integration with Python.
3. The programming interface is Matlab- style.

Disadvantages:
1. It is unpredictable for dynamic, interactive plots.
2. Very much reliant on packages like Numpy.
3. It works only for Python.

B. plot.ly

Features:
1. It is an online analytics and data visualization tool.

2. Online graphing, analytics, and statistics tools are provided not only for individuals/collaboration, as well as scientific graphing libraries for Python
3. Its graphical user interface which provides stats tools for analyzing and importing data into a grid.
4. To create graphs more efficiently, they can be either embedded or downloaded.
5. Responsible for providing API libraries for Python, MATLAB, R, Julia, Node.js and Arduino
6. It helps with styling interactive graphs using IPython.
7. Apps developed using Plotly for Google Chrome.

Advantages:
1. Allowance for changing the colors and style for the graph.
2. Also allows to change the plot type.
3. Title for the graph and label for the axes can be provided.
4. Hovering the mouse cursor on the line will provide the values of the points.

Disadvantages:
1.This technology is relatively new.
2.It requires internet for viewing data into graphical representation.

C. Bokeh
Features:
1. 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.
2. It provides aid to create interactive plots, data applications and dashboards efficiently.
3. It is fully open-sourced.

Advantages:
1.It provides visualization library that targets web browsers for presentation.

Disadvantages:
1. Community support not prompt.
2. Relatively new library with no history for credibility.

IV. SYSTEM DIAGRAM

With the above knowledge in consideration and undertaking the table as reference, a given 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 take out 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/data 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:

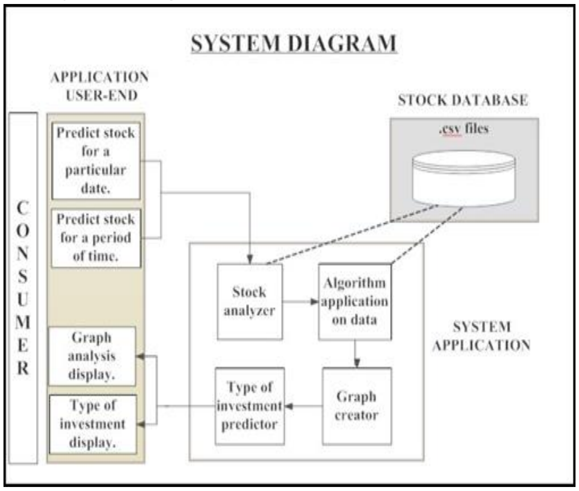

SYSTEM DIAGRAM
APPLICATION USER-END
CONSUMER
Predict stock for a particular date.
Predict stock for a period of time.
Graph analysis display.
Type of investment display.
STOCK DATABASE
.csv files
Stock analyzer
Algorithm application on data
Graph creator
Type of investment predictor
SYSTEM APPLICATION

Figure 3

V. CONCLUSION

This paper summarizes important techniques in machine learning which are pertinent to stock prediction. The paper recommends use of linear regression and logistic regression and Long Short-Term Memory (LSTM) for stock prediction and stock analysis and this study recommends SVM to acquire accurate results. A constraint to this conclusion is the necessity of the data set 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 informality and convenience. The paper proposes a system to extract knowledge from data and performing a prediction to advise the consumer for investments.

VI. HELPFUL HINTS

A. Figures and Tables

Properties	Python	Java	Json
Features	Free and Open Source. It is a high-level language. It provides portability. It is interpreted. It provides extensive libraries. Indentation.	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.
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 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 HTML is provided by Json.
Disadvantages	Pace decreases as it is an interpreted language. Absence felt in mobile computing and browsers because of security.	Since it is a multi-platform	Poses difficulty for the unexperienced users. Json commands occupy all the available memory which is a drawback while doing data mining.

Table 1: Tools (Types of programming language)

REFERENCES

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

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

[3] https://www.analyticsvidhya.com/blog/2015/08/comprehen sive-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 DataMining Techniques Journal: 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 Anjum Research paper: Efficient Machine Learning Techniques for Stock Market Prediction Journal: Int. Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 3, Issue 6, Nov-Dec 2013, pp.855-867.

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