Regression based Stock Market Prediction

Shivraj Chaudhary[1], Varun Arora[2], Vir Bhadra Pratap Singh[3]

[1] *Graduation student, IT department, IMSEC, U.P., INDIA*
[2] *Graduation student, IT department, IMSEC, U.P., INDIA*
[3] *Assistant Professor, IT department, IMSEC, U.P., INDIA*

ABSTRACT

Stock Market Prediction is the method of determining future values of a company's stock. Stock Market Prediction has always attracted people interested in investing in share market and stock of a company for large profits but it is very difficult to predict the stock values of a company as it depends on many factors. Stock market keep on varying day by day. In this paper, a regression system is developed to predict the stock values of a company using regression.

Every day more than 6000 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 investors 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. In This Model, We proposed the application of Machine Learning Using Python to predict Stock Market prices and it could be used to guide an investor's decisions. The algorithm can be used for training set of market data collected for the period of one thousand or two hundred or three days.

General Terms-*Stock Market, Regression, linear regression and web scrapping .*

Keywords- *Machine learning, Stock market prediction, supervised learning, unsupervised learning, web scrapping and regression.*

INTRODUCTION

A collection of buyers and sellers of stock is the stock market, were stocks are released by the companies for elevating the capitals and are bought by the investors in order to get a portion of the company. Stock markets are always aggressive. It is very difficult to predict the future stock price of the companies since it keeps fluctuating every day. In data mining and Machine learning a number of algorithms were designed to overcome this incertitude. This project will focus exclusively on predicting the daily trend (price movement) of individual stocks. The project will make no attempt to deciding how much money to allocate to each prediction. In this present model Regression, Web Scrapping are considered, were regression is a predictive method. The predictive method makes prediction about values of data, are set of supervised Learning used for classification, regression and outliers detection. The most reliable way to forecast the future is to try to understand the present but the amount of data available nowadays is huge and generally beyond human comprehension. Data analysis comes handy to solve this problem. Data analysis can be used to better understand the present scenario of the Stock market so as to understand and try to create a better future scope for investment. With Data analysis, we can add a degree of certainty to the unpredictable and volatile nature of stock prices In this model, regression analysis is a supervised process for estimating the trend among variables. More Specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Although, will use data of Bombay Stock Exchange and by the help of web scrapping the data is obtained in a csv file as stock market prices varies day to day.

1. Proposed Model

This paper uses regressive model to predict the future price of a stock. If the output variable depends linearly on its previous values then it is called an linear regression. Linear regressive model define the current value of output variable as a linear combination of its own past values and present values of the input variables. The correlation technique finds the related stocks of the selected stock. The Moore and Penrose technique is used to estimate the coefficients of the

regression equation. The linear regression model is a regression equation. The regression equation is solved to find the coefficients, by using those coefficients we predict the future price of a stock. Regression analysis is a statistical tool for investigating the relationship between a dependent or response variable and one or more independent variables. Initially we choose a stock exchange from a group of stock exchanges and then we select a stock from that stock exchange and its related stocks from the same stock exchange to retrieve their past values. Now we prepare the input data by using that historical data. In this model the input data is grouped into two sets as training data set and testing data set. The training data set is used to train a model and to estimate the unknown coefficients of the auto regression equation. These coefficients are estimated by using regression technique. The estimated coefficients are used to predict the future price of a stock. By using web scrapping, the training data is obtained and regression analysis is done on training data to predict the stock values. If the data is not linear then Novelty detection technique is used to make the training data linear.

2. Method

The main Terminologies used in the Stock market prediction model are regression, web scrapping, support vector machine, novelty detection, data pipelines.

2.1 Regression

Regression is used for predicting an outcome based on a given input. The simplest regression technique is linear regression and advanced regression technique is multiple regression. If a single descriptive variable is used then it is known as simple linear regression and if more than one descriptive variable is used then the technique is multiple regression.

3.1.2 Linear Regression

Linear Regression is statical technique used to predict the relationship between the dependent and an independent variable. Generalities represented as V=Y+WX, where Vis the dependent variable, X is the independent variable, Y is a constant and W is the slope of regression line..

3.1.3 Multiple Regression

Multiple regression is a technique for modeling the association among the scalar dependent variable "V" and one or more descriptive variables indicated by "U". It predicts the future value of variable with respect to other variables.

$$V = w0+w1\ y1+ \ldots +wnyn+ \epsilon$$

where , V implies the dependent variable,we implies the co-efficients, y1 to yn implies the independent variables, and ϵ implies the random error. In this work Multiple regression technique is used for predicting the future stock price In our method we are using web scrapping (also termed screen scraping,webdata extraction, web harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a locale file in your computer or to database in table(spreadsheet)format, with this proposed technique we are using data pipelines to transform data from one representation to another through a series of steps. We will be importing different packages in this model such as numpy, csv, matplotlib etc.Main Algorithm used here is Scikit

learn, Figure 1 specifies the working of algorithm. We will apply the regression technique to find the predicted prices ,matplotlib to construct a graph for the regression line .

2.2 Some notable implementation of regression model

- **sklearn.linear_model.Ridge**

L2regularized least squares linear model

- **sklearn.linear_model.ElasticNet**

L1+L2regularized least squares linear model trained using Coordinate Descent

- **sklearn.linear_model.LassoLARS**

L1regularized least squares linear model trained with Least Angle Regression

- **sklearn.linear_model.SGDRegressor**

L1+L2regularized least squares linear model trained using Stochastic Gradient Descent

2.3 Feautres

For Machine Learning is about building programs with tunable parameters (typically an array of floating point values) that are adjusted automatically so as to improve their behavior by adapting to previously seen data. Regression is labeled as a supervised learning which is continuous we might wish to determine the age of an object based on such observations: this would be a regression problem: the label(age) is a continuous quantity. A supervised learning

algorithm makes the distinction between the raw observed data X with shape (n_samples, n_features) and some label given to the model during training. In scikit-learn this array is often noted y and has generally the shape (n_samples,). After training, the fitted model will try to predict the most likely labels y_new for new a set of samples X_new. If y has floating point values (e.g. to represent a price, a temperature, a size...), the task to predict y is called regression. Recent studies in stock market prediction suggest that there are many factors which are considered to be correlated with future stock market prices. Nonetheless, using too many financial and economical factors can overload the prediction system [Thawornwong and Enke, 2003; Hadavandi et al., 2010; Chang and Liu, 2008;Esfahanipour and Aghamiri, 2010]. As a result, one of the initial and most challenging steps of stock market prediction is determining the manageable amount of the input variables which have the strongest forecasting ability and can be used as inputs to a prediction system i.e Multiple Regression Analysis. Companies from the best clustering technique under goes the regression technique for predicting the future stock.The technique used is multiple regression. The fig.2. shows the predicted values of TCS for the month of January using the multiple regression technique, this data set contains 21 days stock price and its close price is taken for fitting the price which is one of the attribute which means that the close price is taken as the dependent variable and rest of the attribute as the independent variables. The fig.3 gives the result for the prediction i.e. it results the future price of February which is approximately equal to the stockprice of February. This will help the buyers and sellers to choose the correct company for their investment.

2.4 Future Scope

Warren Buffet has earned in Billions and he is one of the top 3 richest person in earth only due to investing and he didn't earn much till he turned 49, but he has started his investments long back. Now leaving the past, US has become one of the wealthiest country and its economy is worth 18 trillion dollars. After US, the country which has grown rapidly is our neighbour China, they had opened up or liberalised their economy slowly and in the last 30 years it has grown enormously as manufacturing hub. Its economy is worth 11 trillion dollars. India is one of the emerging economies and its economy is worth 1–2 trillion dollars, so Government is steering the economy forward by changing so many policies. So for us to become 10 trillion economy in the future, growth of companies has to be equally there. Any midcap company can become bluechips company if it had good management and as a investor if we had stayed invested with those companies, you can see the potential of equities.

Another point is sensex was started at 179 points i think in 1979, from that day till now it is at 27–28000, so if someone has started investing consistent amount in this for the past decades he would have earned more than real estate or gold.

3. Application

Using new statistical analysis tools of complexity theory, researchers at the New England Complex Systems Institute (NECSI) performed research on predicting stock market crashes.It has long been thought that market crashes are triggered by panics that may or may not be justified by external news. This research indicates that it is the internal structure of the market, not external crises, which is primarily responsible for crashes. The number of different stocks that

s

move up or down together were shown to be an indicator of the mimicry within the market, how much investors look to one another for cues. When the mimicry is high, many stocks follow each other's movements - a prime reason for panic to take hold. It was shown that a dramatic increase in market mimicry occurred during the entire year before each market crash of the past 25 years, including the financial crisis of 2007–08.

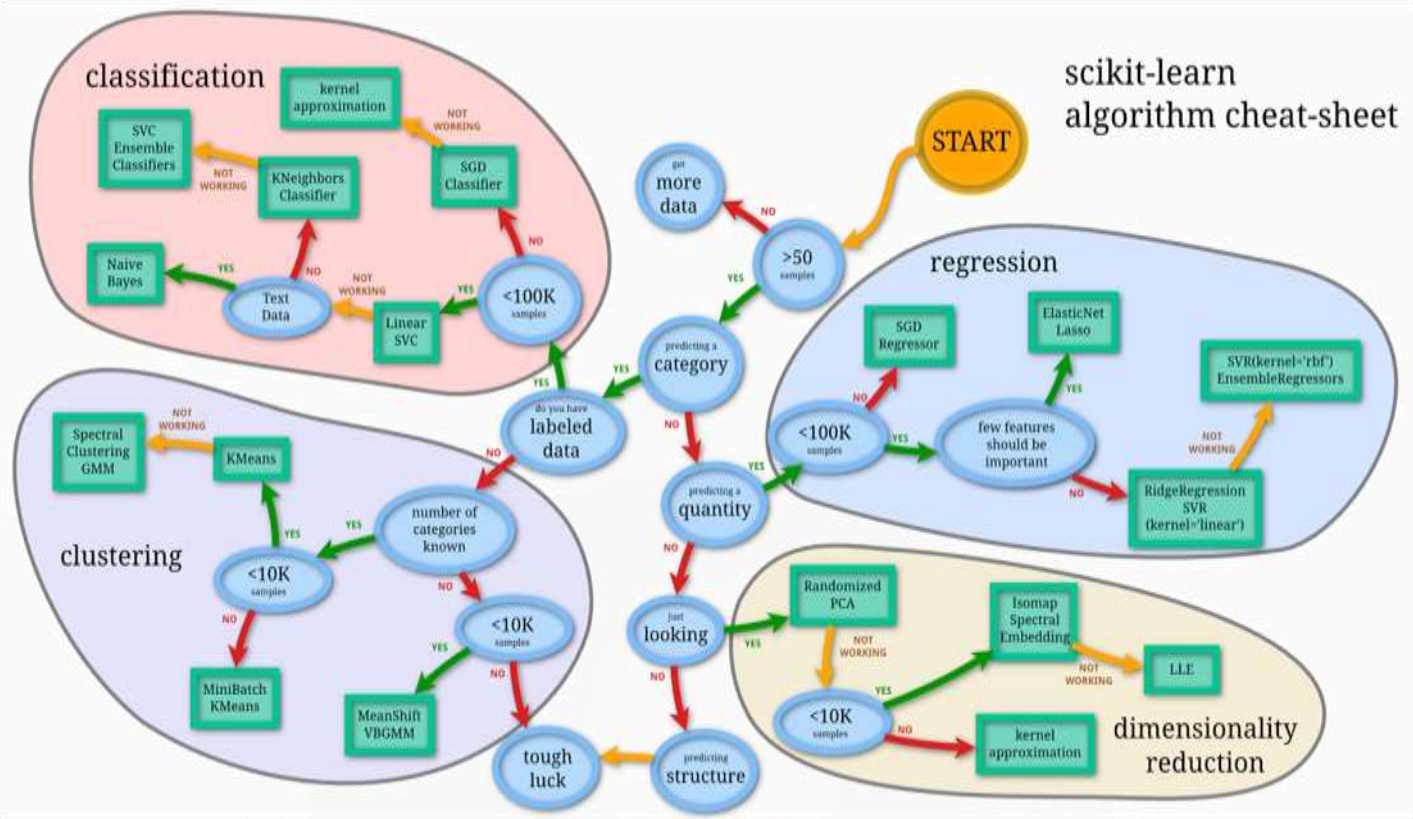

Fig 1: Scikit Learn Algorithm.

4. 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.

4.1 Platform and Technologies used

In, proposed model many technologies is used that are describes below.

4.1.1 Python

Python is a widely used high-level programming language for general-purpose programming, created by Guido

	fit	lwr	upr
1	2538.601	2484.866	2592.336
2	2567.754	2514.280	2621.227
3	2533.174	2479.633	2586.716
4	2440.331	2388.195	2492.467
5	2411.601	2361.118	2462.083
6	2421.809	2371.525	2472.094
7	2480.512	2429.403	2531.620
8	2484.126	2432.161	2536.092
9	2483.787	2431.781	2535.793
10	2505.974	2453.653	2558.295
11	2545.378	2492.513	2598.242
12	2484.026	2431.525	2536.527
13	2508.713	2456.197	2561.230
14	2486.602	2434.546	2538.657
15	2486.327	2434.021	2538.634
16	2495.515	2443.135	2547.895
17	2502.837	2450.657	2555.018
18	2488.573	2436.312	2540.834
19	2521.211	2469.042	2573.381
20	2526.357	2473.695	2579.020
21	2477.643	2424.451	2530.835

Fig2. Training Data of a company's stock (TCS)

van Rossum and first released in 1991. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such as C++ or Java.The language provides constructs intended to enable writing clear programs on both a small and large scale.

4.1.1.1 Web Scrapping

Web scraping (web harvesting or web data extraction) is data scraping used for extracting data from websites Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

4.1.1.2 Novelty Detection

Novelty detection is one of the fundamental requirements of a good classification system. A machine learning system can never be trained with all the possible object classes and hence the performance of the network will be poor for those classes that are under-represented in the training set. A good classification system must have the ability to differentiate between known and unknown objects during testing. For this purpose, different models for novelty detection have been proposed..

5. ACKNOWLEDGMENTS

Our thanks to the faculty of IMS Engineering College Ghaziabad who have contributed towards development of the Proposed model.

6. REFERENCES

[1] http://www.ijcaonline.org/archives/volume163/number5/pahwa-2017-ijca-913453.pdf.

[2] https://en.wikipedia.org/wiki/Web_scraping

[3] Dr. P. K. Sahoo, Mr. Krishna Charlapally, "Stock Price Prediction Using Regression Analysis".

[4] Han Lock Siew, Md Jan Nordin, "Regression Techniques For The Prediction Of Stock Price Trend"

[5] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-

[6] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[7] https://www.ijser.org/researchpaper/Stock-Price-Prediction-Using-Regression-Analysis.pdf

[8] https://www.quantinsti.com/blog/machine-learning-trading-predict-stock-prices-regression/