STOCK MARKET PREDICTION USING REGRESSION

Rohan Taneja[1]**, Vaibhav**[2]

[1,2] *Dept. of Computer Science & Engineering, IMS Engineering College, Ghaziabad, India*

Abstract *- Stock market prediction is the model of determining future values of a company's stock prices. It helps people who have a great extent in investing their money in stocks and to achieve higher profits. It has been a great mystery for peoples to predict the stock prices as it depends on many factors of a company profile .Stock market keeps varying day by day .In this paper, a regression model 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) offers an average of 240000000+ stocks, making an approximate of 2000Cr+ Indian rupees in investment. 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 extraction the beneficial information from those patterns to predict the future behavior of the BSE stock market is necessary .An Application representing the graphical view of stocks values in visual form for user or investors to invest in a particular stock for higher profit is a key requirement .Many Researchers are analyzing the data to predict the stock prices but all of them has it's own short coming .In this Model ,We proposed the application of Machine Learning using Python to predict Stock prices and it could be used to guide an investors decisions. The algorithm can be used for training set of market data collected by web scrapping for the period of any days.*

Key Words: **Machine Learning, Bombay Stock Exchange, Regression.**

1. INTRODUCTION

Stock market is a collection of buyers and sellers which show their interest with the trading of stocks which are released by the companies for elevating the capital and are bought by the investors in order to get a portion of the company. Stock Market is always aggressive, you don't know what will happen next as it is very difficult to predict the future stock price of the companies since it keeps fluctuating every day. As we all know a company's stock depends on many factors and taking into account the nonlinearities and discontinuities of the factors which are considered to impact stock markets. Some of the reasons are like company related news, political events natural disasters etc. stock price prediction is one of the most important issues to be investigated in academic and financial researches [1]. Multiple Regression Analysis, Principal Component Analysis, Hurst Exponent Analysis, and Grey Relation Analysis can be used during this initial step [Ince and Trafalis, 2007; Hurst, 1951; Kung and Wen, 2007].This Model of Prediction will focus exclusively on predicting the daily trend (price movement) of individual stock . This Model will make no attempt on deciding how much money to allocate to each prediction. In this present model Regression, Web Scrapping, Differential Evolution Optimization-based Fuzzy type-2 Clustering are considered where regression which can be linear and multiple is a predictive method.

The remainder of the paper is organized as follows: Section 2 describes the web scrapping method to extract data from the Bombay Stock Exchange [BSE] website itself. Section 3 describes the application of Regression which consist of Multiple and linear regression analysis for the prediction. The structure of the Differential Evolution Optimization based Fuzzy type-2 Clustering system that is implemented in this paper is then presented in Section 4. The experiments and empirical results of proposed model are presented and described in Section 5. Finally, Section 6 provides a conclusion.

2. WEB SCRAPPING

Web Scrapping or Web Harvesting is a method of extracting data from the websites. It may access the World Wide Web directly using the Hypertext transfer protocol, or through a web browser. In this proposed model web scrapping is used to extract the stock prices of a certain company you want from the Bombay stock exchange website at any time, as you know prices fluctuates every day and will save it in a csv file which will be our training data for the regression analysis and it there will be uncertainty in the prices along with dates Then, here the Differential Evolution Optimization based Fuzzy type-2 Clustering system will start working.

3. REGRESSION

Regression is a predicting method whose outcome is based on the given input. The simplest regression Technique is linear regression whereas advanced regression technique is multiple regression.

3.1 Linear Regression

Linear regression is a statistical Technique use to predict the relationship between the dependent and independent variable and is represented as V=Y+WX where V is the dependent variable and X is the independent variable, Y is a constant and W is the slope of the regression line.

3.2 Multiple Regression analysis

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 , in initials to make our prediction strong and more exact we have the challenge to determine only those inputs variable which have the strongest forecasting ability, basically the main factors on which a company profile depends. IN our proposed model multiple regression analysis is performed on 24 economic variables to both reduce the dimensionality of the variable set and identify which variables have a strong relationship with the market price of the BSE index for the subsequent testing period [2]. Example variables include past prices, volume, technical indicators, T-bill rates, certificate of deposit rates, credit ratings, producer/consumer price indexes, industrial production levels, and money supply levels.

The input variables including date and prices are collected through web scrapping. The variables with inappropriate t-statistics and p-values are excluded from the list of inputs. According to the experiment results, the Multiple Regression Analysis method identified the 2-month T-bill (T-Bill3) rate, 2- month Certificate of Deposit (CDR3) rate, past BSE Index price level, past Money Supply (M1) level, recent Industrial Production (IP) reading, and the recent Producer Price Index (PPI) as significant and relevant variables in the regression model (with p-values less than 0.005). Analysis of the t-statistics and significance value of each variable suggested that these variables contain relevant information about the future stock prices. Given that the absolute t-statistics values of each variable was greater than 1 and the p-values (or significance values) are less than 0.005, the selected variables are believed to have strong forecasting ability.

BSE t+1 = -33.690 + (11.999*T-Bill3) + (1.365*IP) + (.059*M1) – (.781*PPI) – (9.124*CDR3) + (.955*BSE) (1)

Table 1: Model Calculation Summary

R	R Squared	Adjusted R Square	Std. Error of Estimate
0.996	0.994	0.994	35.64

The relationship generated from the regression (Equation 1) indicates that the BSEt+1 is a function of the intercept (-33.690) plus various coefficients multiplied times the relevant variables. For this formula, positive changes in T-Bill3t, SP500t, IPtí1, and M1tí1 have positive effects on the prediction of the stock market level for the next month (BSEt+1), while the positive changes of CD3t and PPtí1 have negative effects. The R-squared value of the model is 0.994, implying that the equation explains 99.4 percent of the variation of the future stock market price.

4. DIFFERENTIAL EVOLUTION OPTIMIZATION-BASED FUZZY TYPE-2 CLUSTERING

Various researchers are applying Different types of clustering analysis in the various fields of science. There are two main types of clustering techniques first one is agglomerative hierarchical clustering and the other one is nonhierarchical clustering technique. The agglomerative hierarchical clustering methods are used as an explanatory statistical technique to determine the number of clusters of data sets, include Single, Complete, and Average linkage methods. These methods are appropriate for both qualitative and quantitative variables [Johnson and Wichern, 2002]. Fuzzy clustering is a well-established paradigm used to generate the initial type-2 fuzzy “If-Then” model [Hwang and Chung-Hoon Rhee, 2007]. For the proposed model, Fuzzy type-2 Differential Evolution-based Clustering is used since it has been proven to produce results that better suit the application of type-2 If-Then rules [Aliev et al., 2011]. This removes the uncertainty in choosing the “m parameter” existing in Fuzzy c-means by suggesting a solution for a range of its values covering {1.4, 2.6}, a meaningful range for “m.” Adequate choice of m is very important as it plays a visible role in forming the shape of resulting fuzzy clusters [Aliev et al., 2011]. As the experiments have shown, using a type-2 Fuzzy Clustering method provides better location of the cluster centers, and subsequently results in a better fuzzy rule model. This in turn allows capturing more uncertainty, while delivering higher robustness against the imprecision of the data. The objective function is as follows (with n data vectors, P = {p1, p2, . . . , pn} inputs; prototype vj of the jth cluster generated by the fuzzy clustering; membership degree uij of the I th data belonging to the j th cluster represented by the prototype vj):

$$J_{m1} = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m_1} \left\| \mathbf{p}_i - \mathbf{v}_j^{(1)} \right\| \rightarrow \min \quad , \quad J_{m2} = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m_2} \left\| \mathbf{p}_i - \mathbf{v}_j^{(2)} \right\| \rightarrow \min \qquad (2)$$

subject to constraints:

$$0 < \sum_{i=1}^{n} u_{ij} < n \quad (j = 1, 2, \ldots, c) \quad \text{and} \quad \sum_{j=1}^{c} u_{ij} = 1 \quad (i = 1, 2, \ldots, n)$$

The vector $\tilde{\mathbf{v}}_i$ is formed as:

$$\tilde{\mathbf{v}}_i = [\min(\mathbf{v}_i^{(1)}, \mathbf{v}_{Ind_i}^{(2)}), \max(\mathbf{v}_i^{(1)}, \mathbf{v}_{Ind_i}^{(2)})] \text{ where } Ind_i = \arg\min_j \left\| \mathbf{v}_i^{(1)}, \mathbf{v}_j^{(2)} \right\| \qquad (3)$$

Figure 1:Equations of Fuzzy Clustering.

In the proposed approach, minimization of the objective function (Equation 1)is done by Differential Evolution [Price et al., 2005] since it acts as a global search algorithm and is expected to be more advantageous than standard Fuzzy c-means for the case of a large number of highly-dimensional data vectors. Fragments of the Fuzzy type-2 IF-THEN model (rules) discovered by fuzzy clustering are shown below:

1) IF BSE is about 1222 AND T-Bill3 is about 1.69 AND CDR3 is 2.96 AND PPI is about 136.93 AND M1 is about 498.56 AND IP is about 12.37 THEN BSE is about 1044 ….. …… …………

To……..

7) IF BSE is about 88.79 AND T-Bill3 is about 5.19 AND CDR3 is 5.90 AND PPI is about 97.40 AND M1 is about 679.06 AND IP is about 50.49 THEN BSE is about 899.20.

5. TOOLS/PACKAGES AND PLATFORM USED

In the proposed system python is used as the programming language in which several packages are imported such as

1. import Beautiful soup as Bs4
2. import numpy
3. import scipy
4. import panda
5. import scikit
6. import tkinter
7. import matplotlib

Here, Beautiful Soup is used for web scrapping whereas numpy and pandas are used for the mathematical equations and functions and tkinter is used to develop interface so that the end user can interact and matplotlib is used to represent the predicted data on the graph.

6. COMPUTER GENERATED ENVIRONMENT

The goal of this proposed model is to determine the next stock price value by using the previous days data which is a numerical data. The dataset used for testing contains various records. As determined during regression analysis, the input included the 3- month T-bill (T-Bill3) rate, 3-month Certificate of Deposit (CDR3) rate, past BSE Index level, past Money Supply (M1) level, recent Industrial Production (IP) reading, and the recent Producer Price Index (PPI). The output was the next BSE Index value.

For the simulation, the Differential Evolution-based Fuzzy type-2 clustering model included seven clusters, max iteration =1000, exponent =2, and population size=200. Parameters of the type-2 neural network (that was initiated during the clustering procedure) are further adjusted by the Differential Evolution algorithm on the training series (80% of all data).

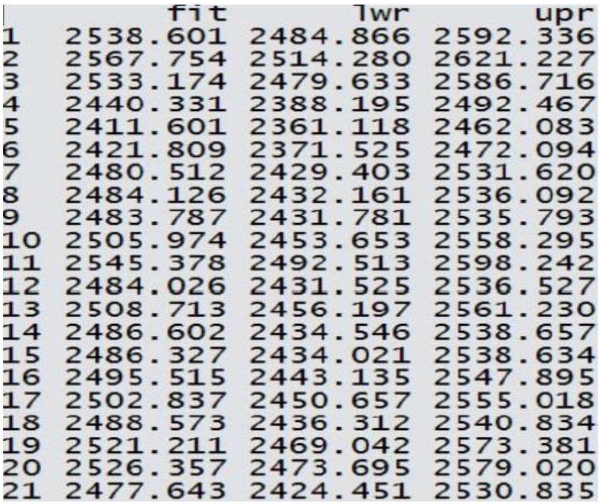
Table 2: Comparison of Performance of different Approach

	RMSE
TYPE -1 Approach	**0.948**
Type -2 (the suggested approach)	**0.909**

CONCLUSIONS AND RESULT

The proposed model uses regression analysis as a ML technique with web scrapping which is followed by Fuzzy type-2 Clustering. A prediction system has been built that uses machine learning technique to produce periodically forecasts about stock market prices. The use of Prediction and regression helps us to find errors and improve accuracy of the system. It can be very useful for the investors to use this to gain maximum profit.

There are many research directions which might be considered in the future work. Improving the accuracy of the predictive models is one of them. Accuracy can be improved by considering an entirely different aspect i.e., human sentiments. As future scope of stock market is limitless, the demand for its data analysis will be ever increasing. By changing only the training data, the proposed system can be used for any stock markets of other countries. With few altercations, the system can be used for various purposes such as predicting prices of commodities like gold, predicting the fuel consumption of a vehicle as well as monitoring health of a patient.



```
        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
```

Figure 2: Shows data of stock prices.

REFERENCES

[1] Aliev R. A., W. Pedrycz, B. Guirimov, R. R. Aliyev, U. Ilhan, M. Babagil, and S. Mammadli, "Type-2 Fuzzy Neural Networks with Fuzzy Clustering and Differential Evolution Optimization," Information Sciences (2011): pp. 1591-1608.

[2] Atsalakis G. S., and K. P. Valavanis, "Surveying stock market forecasting techniques – Part II: Soft computing methods," Expert Systems with Applications, Vol. 36, Issue 3, Part 2 (2009): pp. 5932-5941.

[3] Hwang C., and F. Chung-Hoon Rhee, "Uncertain fuzzy clustering: Interval Type-2 Fuzzy Approach to C-Means," IEEE Transactions on Fuzzy Systems, Vol. 15, No. 1 (2007): pp. 107-120.

[4] http://francescopochetti.com/stock-market-prediction-part-introduction/

[5] https://ac.els-cdn.com/S1877050911005035/1-s2.0-S1877050911005035-main.pdf?_tid=36d95515-dd48-4509-9ac8-8d7aab5028ae&acdnat=1523551353_d6564f13510bc5df1961e926a9039c75.