Application of Genetic Algorithm and Machine Learning Techniques for Stock Market Prediction

Shibendu Mukherjee
Dept. of Computer Science and Engineering
University Visvesvaraya College of Engineering
Bangalore, India

S M Dilip Kumar
Dept. of Computer Science and Engineering
University Visvesvaraya College of Engineering
Bangalore, India

Abstract**—** *In a financial research it is crucial to compute the expected momentum of the stocks. The objective is to reduce the risk involved in share investments and maximizing the returns of an investment. In this work, Genetic Algorithm (GA) is used to select high quality stocks with investment value from a vast pool of stocks. For the genetic algorithm to efficiently select the stocks a cogent fitness function is defined. Once defined, the elitist stock is determined. The resultant stock is clustered and a logistic regression model is built upon it. This gives a binary output for the user/customer whether to buy the stock or sell it. The experiments were conducted using RStudio and the results reveal that the proposed technique generates a higher accuracy for the prediction.*

Index Terms**—***Stock Market prediction, Machine Learning, Clustering, Genetic Algorithm, Logistic Regression.*

I. Introduction

Stock markets dwell in the area of uncertainty and it is therefore very difficult to predict the price. The price and volume are very volatile entities and they are dominated by many factors. Market internal, fundamental and policies are some of the key factors. Investors generally adapt several strategies for the investments. These strategies can be classified in two types of stock analysis [14]. The first one is fundamental analysis, the second being the technical analysis. Fundamental analysis deals with economics and financial science. It makes use of economic and financial data as an indicator such as macroeconomic indicators, foreign exchange rates, current ratio, etc. Whereas technical analysis is only based on transaction volume and stock price. The reason to go with technical analysis is to learn from the history of the stock prices as it is expected to repeat itself through the variation of volume and price. An investor faces a difficult time selecting perspective stocks from a pool of options available to them. Selecting them wisely with a right strategy is the key to generate profit.

There has been numerous approaches addressed for the issue. The use of artificial intelligence, hybrid systems and many machine learning approaches are attempted for it. Ibrahim et al. [3] applied artificial neural network to select valuable stocks. Qisen et al. [2] used fuzzy time series and GA in combination to forecast stocks. Phayung et al. [5] used support vector regression to predict the stock market price. Recently Shipra et al. [1] utilized neural network with machine learning techniques to train the model. However there are certain drawbacks to these approaches. In case of fuzzy systems they lack the ability to learn, neural network has overfitting problem and is often easily trapped in local minima, and in the support vector approach choosing the kernel is a daunting task. In order to overcome these issues, genetic algorithm can be used instead to select the stocks.

The primary goal of this work is to build a model which gives an accuracy high enough to instill the confidence among the investors. So, at first we need to ascertain the independent variables which increases the confidence of the model. For the same we need to check whether or not the variables are correlated to each other. Clustering based models are particularly well suited to analyzing stock market data which are noisy and often contain non-linear relationships and high-order interactions.

There are powerful and flexible tree based models that may be less appropriate in identifying simple first order relationships. In such situations, a linear factor model is a suitable alternative and would, in most cases, be preferable to using a decision tree. This suggests that modeler should undertake some preliminary analysis of the dataset in order to identify the complexity of the underlying relationships before deciding upon the most appropriate modeling methodology [15].

In situations where a linear model is required, logistic regression could be used as a suitable substitute. It belongs to the generalized linear model family and has been used extensively in economics and finance to forecast a specific event, such as bankruptcy or the probability of default [15]. The model is also able to produce smooth probabilities derived from continuous inputs which may be regarded as desirable when the output is a simple ranking of stocks, even if this is an approximation to the underlying model. Being a linear technique, however, logistic regression shares the usual weakness of the classical modeling approach. Specifically, it requires a valid mean function assumption and is affected by multicollinearity as well as being sensitive to outliers and missing data [15].

In this work, we highlight the benefits of a hybrid technique that combines Genetic Algorithm with clustering and logistic regression in order to deliver enhanced predictions of future stock returns. Analytic results demonstrate that the hybrid approach enhances predictability of the stocks without introducing additional volatility.

Rest of the paper is structured as follows:Section II, describes background and motivation is presented. In Section III, the proposed methodology is presented in detail. While Section IV describes the experiment and analysis. Finally, Section V concludes the paper.

II. Background And Motivation

Recent studies, such as artificial neural networks, fuzzy systems, and evolutionary algorithms have attempted to determine the optimal time to trade in a stock market. Two main methods have emerged: one is to forecast the price of stock and then make a buy or sell decision, and the other is to use a combination of trading rules through technical analysis to create a buy or sell signal. Also there are hybrid methods of these two. The difference between these two methods is that the first may predict the point of price, while the second will find rising or falling fluctuations of the stock price.

1) Genetic Algorithm: GA imitates the natural selection process in biological evolution with selection, crossover and mutation, and the sequence of operations like chromosome evolution, computation of fitness, and checking the termination condition [14]. GA is based on the survival-of-the-fittest by gradually manipulating the problem solutions to obtain more superior solutions in population. Optimization is performed in the representation rather than in the problem space directly. Till date, GA has become a popular optimization method as they often succeed in finding the best optimum by global search in contrast to most common optimization algorithms. There has been interesting insight given by Chengxiong et al. [14] and Tejas et al. [6], which can help to develop a hybrid of these two approach. The aim of this work is to identify the quality of each stock using GA so that investors can choose some good ones for investment. Here we use stock ranking to determine the quality of stock, the stocks with a high rank are regarded as good quality stock. Fig. 1 illustrate the procedures in a GA.

After evolving the fitness of the population, the best chromosomes with the highest fitness value are selected by means of the selection method. The fittest chromosomes are more likely selected followed by crossover step, offspring chromosomes are created by some crossover techniques.The mutation prevents the GA from converging too quickly in a small area of the search space. This is specially helpful in case there is no satisfactory solution for a specific problem. By applying the mutation operator, it is possible to produce new chromosomes. The evaluation and reproduction steps are repeated until a certain number of generations, or a defined fitness or until a convergence criterion of the population are reached. In the ideal case, all chromosomes of the last generation have the genes representing the optimal solution. Through the process of GA optimization, the stocks are ranked according to the fundamental financial information and price return. Investors can select the top n stocks to construct a portfolio.

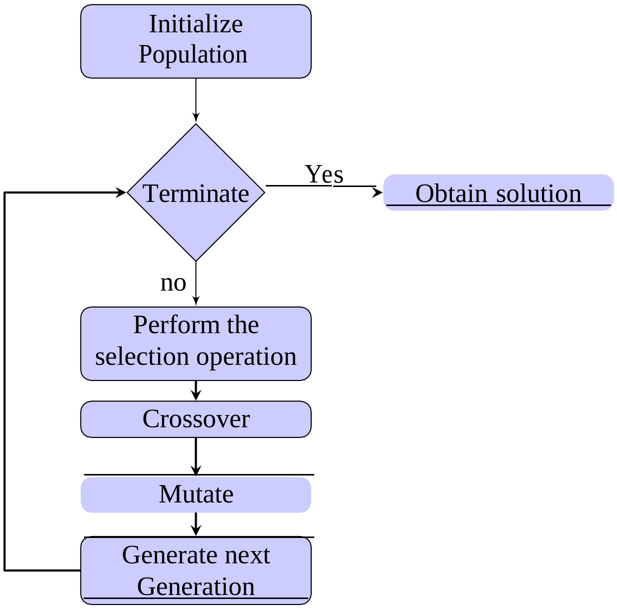

Fig. 1. Flowchart of Genetic Algorithm

2) Clustering: According to Vladimir Estivill-Castro, the notion of a “cluster” cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms [8]. However, different researchers employ different cluster models, and for each of these cluster models again different algorithms can be given. Some of the most used cluster models are as under:

i) Hierarchical clustering : It is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect “objects” to form “clusters” based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram [10], where the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don’t mix.

ii)k-means Clustering : The number of clusters is fixed to k, in k-means clustering. Clusters are represented by a central vector, which may not necessarily be a member of the data set. Here the k cluster centers are found and assigned the objects to the nearest cluster center, such that the squared distances from the cluster are minimized [8].

$$D = 1 - C \qquad \text{Eq. (1)}$$

where D = distance and C = correlation between spot clusters. The distance measure between two clusters is calculated as given above. For finding the spot correlation data, a visual representation dendrogram can be used. The bottom of the dendrogram referred to as leaf nodes contains the individual spots [10]. For highly correlated data, the correlation value is close to 1 and so from Eq. (1) we will have a value close to 0. Thus, the highly correlated clusters are near the bottom of the dendrogram. Correlation value of 0 arises when spot clusters are not correlated and the corresponding distance value comes to 1. Whereas in case of negatively correlated spots, it will have a correlation value of -1 and $D = 1 - -1 = 2$.

3) Logistic Regression: In statistics, logistic regression [11] model is a probability model that was developed by statistician D. R. Cox. The logistic function describes the model which gives the probabilities of the possible outcomes for a single trial. Logistic Regression [4] is used specifically to deal with problems whose dependent variable is binary. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables, by estimating probabilities. Logistic regression is a special case of generalized linear model and thus analogous to linear regression. The key differences is the estimated probabilities and are restricted to $[0, 1]$ through the logistic distribution function because logistic regression predicts the probability of the instance being positive.

III. Proposed Method

In this section a detailed explanation is being given of the proposed method is presented.

A. Generate the initial population

The initial population is being created from the available stocks. An individual is characterized by a fixed-length binary bit string, which is a chromosome. All the individuals of the initially created population are evaluated by means of a fitness function .

B. Define the fitness function

The fitness function is used to create a genetic pool. Subsequent to the fitness evaluation of the individuals in the initial population, new population is evolved. The formulation of new generation is executed in stages of reproduction, crossover and mutation. All-encompassing objective of this step is to procure a new population of individuals which have superior fitness values.

The fitness function is built using the returns of the previous month which is then coupled with the weight of the stock holdings. Then the overall value of the function is computed and the iteration continues for fixed number of times (200 in this case). Only those stocks are selected which gives the highest value for this function. During reproduction phase the individuals are winnowed based on their fitness values i.e. individuals with diminished fitness values are eliminated, whilst the others with elevated fitness values are duplicated to the next generation.

Selection, crossover, mutation followed with fitness evaluation on the evolved species is continued in a loop. After many generations are created and killed, only the best of the best are left. The algorithm will then return those survived individuals as the values we want to find predominantly, such as the combination of optimal parameters, trading rules, or stock weights.

C. Clustering the data obtained

The elitist stocks are now selected for further processing where the data is divided into training set and testing set. The outcome variable or the dependent factor is being removed. There is a need to preprocess the data so that there is no dominant factor while classifying the data. We used mean and standard deviation of the variables for normalization, then they are divided into clusters. To decide the number of clusters we need to build a dendrogram to visually help us. And finally it is subjected to split into train set and test set according to the clusters in which they fall.

Next, we form clusters to segment the stocks into similar groups. Cluster analysis [7] or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It improves the success rate and yield, and have important practical value of guidance for investment decisions. The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

D. Build Logistic Regression model

Use the split data from above step to measure out of sample accuracy. To make sure that the sample is split same for all experiments, we need to set a seed value. This initializes the random number generator. The logistic response function is given as under :

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}} \qquad \text{Eq. (2)}$$

Here dependent variables is denoted by Y . Then $P(Y = 1)$ and $P(Y = 0)$ are the two possible outcomes i.e their probabilities. We can compute the other probability by $P(Y = 0) = 1 - P(Y = 1)$. The independent variables in the Eq. (2) are x_1 , x_2,...., x_k . The coefficients are selected to predict a high probability for good returns and to predict a low probability for poor returns. Another useful way of taking this logistic response function is by considering the odds.

$$Odds = \frac{P(Y = 1)}{P(Y = 0)} \qquad \text{Eq. (3)}$$

The $Odds$ >1 obtained from Eq. (3) if $y = 1$ is more likely and the $Odds < 1$ if $y = 0$ is more likely. And is equal to 1 if both are equally likely. Now if we substitute the logistic response function i.e Eq. (2) for the probabilities in the Odds equation then Eq. (4) is obtained, which is e raised to the power of linear equation.

$$Odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k} \qquad \text{Eq. (4)}$$

$$log(Odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \qquad \text{Eq. (5)}$$

Eq. (5) is called as the “*logit* ” and looks like linear regression equation which helps us understand how the coefficients or β effect the predictions of the probabilities. A positive β value will increase the logit which in turn increases the odds of 1 whereas a negative β value decrease the logit and which in turn decreases the odds of 1. Returning to the model, β values are checked and any positive value in this is indicative of higher returns and negative value for lower returns. Thereafter model comparison can be done based on the AIC (Akaike information criterion) value. This has a limitation of building the model on the same training set. Once the models are built, a model can be selected based on value of the AIC. From which choosing the least value of AIC is indicative of the best model. Build a prediction model on the training set. This should give the probabilities as outcome. Here, we are interested in a binary prediction. That is to say whether the stock have moved up or went down. For this we need to use a threshold Value ‘t’.

$$P(stock = 1) = \begin{cases} \geq t, & \text{predict stocks to soar;} \\ < t, & \text{predict stocks to plummet;} \end{cases} \qquad \text{Eq. (6)}$$

The *Threshold value* is selected based on which errors are better. If ‘t’ is large, then we rarely predict stock moving up. So the investor is not able to invest in the stock market. On the other hand if the value of ‘t’ is small then it results in too many false positives which is precarious for the investor. With no preference between errors, select $t = 0.5$. This predicts the more likely outcome.

IV. Experiment and Analysis

The experiment were conducted using RStudio and results are presented. R programming language is used which gives software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. The R community is noted for its active contributions in terms of packages [13].

The experimental data contains monthly stock returns from NASDAQ stock exchange [12]. The NASDAQ is the second-largest stock exchange in the world, and it lists many technology companies. Data was obtained from infochimps, which is a website giving access to datasets. Each observation in the dataset is the returns per month of a particular company in a year. The data was collected in a span of years between $2000 - 2009$. The tickers of the companies that are listed on the exchange for the entire period $2000 - 2009$, and those stock price which never fell below $1 are selected. Thus, one observation is for say “X” stock in 2000, and another observation isfor same “X” in the year 2001. The goal is to predict the stock return in December will be positive or not, using the returns of the stock for the first 11 months. The data contains about 11580 stocks index movement value.

Confusion Matrix : In the field of machine learning, a confusion matrix, also known as a contingency table or an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class [9]. It is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. This allows more detailed analysis than mere proportion of correct guesses (accuracy). A typical table containing elements of the confusion matrix is given below.

	Predicted = 0	Predicted =1
Actual = 0	True Negative	False Positive
Actual = 1	False Negative	True Positive

TABLE I
Confusion Matrix

To compute outcome measures, the formulas are listed below:

$$OverallAccuracy = \frac{TN + TP}{N} \qquad \text{Eq. (7)}$$

$$Sensitivity = \frac{TP}{TP + FN} \qquad \text{Eq. (8)}$$

$$Specificity = \frac{TN}{TN + FP} \qquad \text{Eq. (9)}$$

where TN stands for True Negative, FP stands for False Positive, FN for False Negative, TP for True Positive and N stands for total number of elements. Sensitivity using Eq. (8) also known as the true positive rate tells us the positive percentage of actual to predicted outcome that we were able to classify. Using Eq. (9) specificity can be calculated, which tells us the negative percentage of actual to predicted outcome that we were able to classify. The accuracy of the matrix can be computed by the Eq. (7) using which we can compare the results.

The plot shown in Fig. 2 depicts the curve of evaluation function versus the variable value. This is useful to look at correlations between the variable and the evaluation values. It shows the minimal and mean evaluation value, indicating how far the GA has progressed. And the plot in Fig. 3 is the binary chromosome of the gene selection frequency, i.e. the times one gene in the chromosome was selected in the current population. It will make histograms for each variable to indicate the selected values in the population.

To pick the number of clusters needed, we just need to draw a horizontal line across the dendrogram. Number of vertical lines that line crosses is the number of clusters needed to be. The farthest the horizontal line can move up and down

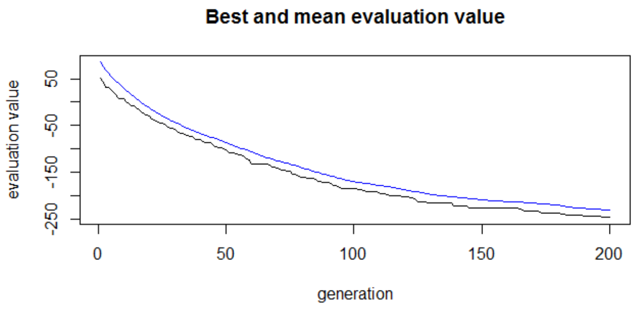

Fig. 2. Evaluation Function vs. Variable value

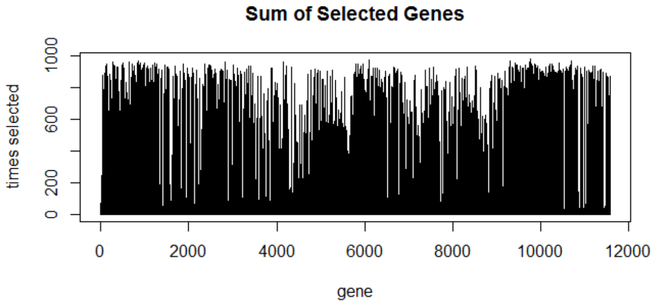

Fig. 3. Gene Selection Frequency

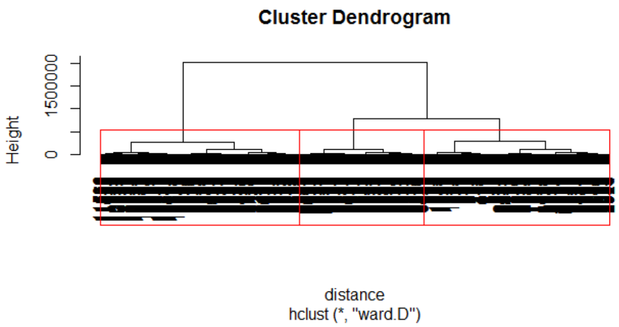

Fig. 4. Dendrogram

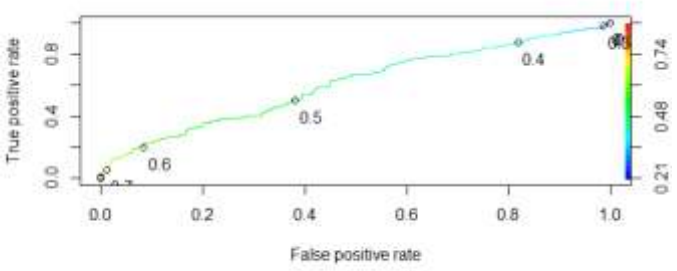

Fig. 5. ROC of 1^{st} cluster

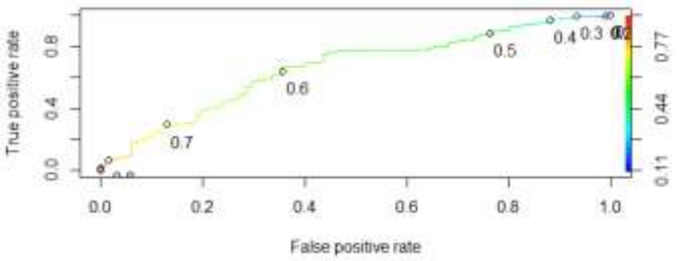

Fig. 6. ROC of 2^{nd} cluster

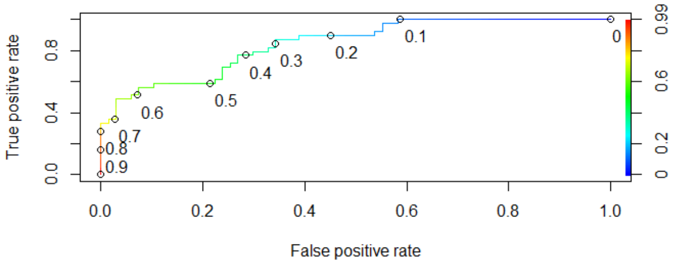

Fig. 7. ROC of 3^{rd} cluster

without hitting one of the horizontal lines of the dendrogram the better that choice is. The applications of the model needs to be considered for number of clusters to be made. To get a finer model the number of clusters need to be increased and for a coarser it can be decreased. From the Fig. 4 and the applicability of the model, we chose three clusters.

To choose the threshold value we need the Receiver-Operator Characteristics (ROC) Curve. We obtain three graphs as shown in Figs. 5, 6, and 7 since there are three models built. These curves have true positive rate on Y -axis and false positive rate on the X-axis. ROC value always starts from a value $(0,\ 0)$ which corresponds to a threshold value $t = 1$. Thus we will not catch any positive value for December stock which we are trying to predict. And it end up having a sensitivity of zero. But it will result in labeling all investment as bad and as a result the investor won't lose money *i.e* false positive $rate = 0$.

The ROC curve always ends at the $(1,\ 1)$ which corresponds on the threshold value of 0. If $t = 0$ then all the prediction will suggest the outcome of a positive returns. The sensitivity will be equal to $'1'$. Thus we can formulate the relation between the $'t'$, specificity, and sensitivity. For a high threshold the specificity is high and sensitivity is low and in case of low threshold specificity is low and sensitivity is high. After considering these graphs we can set the threshold value to 0.5.

Results generated by different models are given below:

From the Tables II, III, IV, V and VI we can compute the accuracies of the respective models as given below:

0	1
5256	6324

TABLE II
BASELINE

	Predicted = 0	Predicted =1
Actual = 0	417	1160
Actual = 1	787	3640

TABLE III
LOGISTIC REGRESSION

	Predicted = 0	Predicted =1
Actual = 0	467	1110
Actual = 1	353	1544

TABLE IV
CLUSTER THEN LOGISTIC REGRESSION

	Predicted = 0	Predicted =1
Actual = 0	91	92
Actual = 1	68	122

TABLE V
GA THEN LOGISTIC REGRESSION

	Predicted = 0	Predicted =1
Actual = 0	98	85
Actual = 1	63	127

TABLE VI
GA + CLUSTER + LOGISTIC REGRESSION

- Baseline model where prediction is always True or False = 54.60564%
- Logistic regression model without Genetic Algorithm application = 56.70697%
- Logistic Regression model on clustered data without Genetic Algorithm Application = 57.88716 %
- Logistic Regression model with Genetic Algorithm Application = 57.10456 %
- Logistic Regression model on clustered data with Genetic Algorithm Application = 60.32172 %

It is observed a modest improvement of the proposed model over other models. Since predicting stock returns comes under hard problem, this is a good increase in accuracy. By investing in stocks for which we are more confident that they will have positive returns (by selecting the ones with higher predicted probabilities), this genetic selection cluster then predict model can give us an edge over other models.

V. CONCLUSION AND FUTURE WORKS

In this paper, we attempted to make a prediction for the NASDAQ stocks. In the proposed system, we used Genetic Algorithm to select the best stocks, then cluster them and finally build a logistic regression model. The experiment reveals that there is a significant high results accuracy than other existential models. The fitness function can have stocks with weights as per the fundamentals of the stock. The accuracy can be further enhanced by ensembling with other models.

REFERENCES

[1]Shipra Banik, Mohammed Anwer, et al. Dhaka stock market timing decisions by hybrid machine learning technique. In *Computer and Information Technology (ICCIT), 2012 15th International Conference on*, pages 384–389. IEEE, 2012.

[2]QiSen Cai, Defu Zhang, Bo Wu, and Stehpen CH Leung. A novel stock forecasting model based on fuzzy time series and genetic algorithm. *Procedia Computer Science*, 18:1155–1162, 2013.

[3]Ibrahim M Hamed, Ashraf S Hussein, and Mohamed F Tolba. An intelligent model for stock market prediction. *International Journal of Computational Intelligence Systems*, 5(4):639–652, 2012.

[4]David W Hosmer Jr and Stanley Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004.

[5]Phayung Meesad and Risul Islam Rasel. Predicting stock market price using support vector regression. In *Informatics, Electronics & Vision (ICIEV), 2013 International Conference on*, pages 1–6. IEEE, 2013.

[6]Tejas P Patalia and GR Kulkarni. Design of genetic algorithm for knapsack problem to perform stock portfolio selection using financial indicators. In *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on*, pages 289–292. IEEE, 2011.

[7]Ruizhong Wang. Stock selection based on data clustering method. In *Computational Intelligence and Security (CIS), 2011 Seventh International Conference on*, pages 1542–1545. IEEE, 2011.

[8]Nimrat Kaur Sidhu, Rajneet Kaur. ”Clustering In Data Mining”International Journal of Computer Trends and Technology (IJCTT),V4(4):710-714 April Issue 2013 .ISSN 2231-2803. www.ijcttjournal.org . Published by Seventh Sense Research Group.

[9]wikipedia. Confusion matrix, 2015.

[10]D.Radha Rani, A.Vini Bharati, P.Lakshmi Durga Madhuri, M.Phaneendra Babu, A.Sravani. ”Analysis of Dendrogram Tree for Identifying and Visualizing Trends in Multi - attribute Transactional Data”. International Journal of Engineering Trends and Technology(IJETT). V3(1):14-18 Jan-Feb 2012. ISSN:2231-5381. www.ijettjournal.org. published by seventh sense research group

[11]wikipedia. Logistic regression, 2015.

[12]wikipedia. Nasdaq stock market, 2015.

[13]wikipedia. R language, 2015.

[14]Chengxiong Zhou, Lean Yu, Tao Huang, Shouyang Wang, and Kin Keung Lai. Selecting valuable stock using genetic algorithm. In *Simulated Evolution and Learning*, pages 688–694. Springer, 2006.

[15]Min Zhu, David Philpotts, Ross Sparks, and Maxwell J. Stevenson. A hybrid approach to combining cart and logistic regression for stock ranking. *The Journal of Portfolio Management*, 38(1):100–109, 2011.