IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

A review on stock prediction using machine learning

Aiswarya S. Kumar
aiswaryaskumar97@gmail.com
College of Engineering, Chengannur, Kerala

Greeshma Merin Varghese
greeshmamerinvarghese@gmail.com
College of Engineering, Chengannur, Kerala

Radhu Krishna R.
mail2radhukrishna@gmail.com
College of Engineering, Chengannur, Kerala

Reshma K. Pillai
reshmakpillai3@gmail.com
College of Engineering, Chengannur, Kerala

ABSTRACT

The goal of this review is to describe the various methods used to predict the stock market, gold price and fuel price. The following paper describes the work that was done on investigating the application of regression, SVM, ELM, ANFIS techniques on the stock market price prediction. The report describes the various technologies with their accuracy level and efficiency in the test phase. It was found that support vector regression was the most effective out of the models used, although there are opportunities to expand this research further using additional techniques to incorporate the current affairs into the prediction features.

Keywords— *Stock prediction, Linear Regression, Fuzzification, SVM*

1. INTRODUCTION

The stock exchange is thought to be a fancy adaptive system that's tough to predict because of the big range of things that verify the day to day worth changes. We tend to try this in machine learning that tries to work out the link between variable quantity and one or a lot of freelance variables. Here, the independent variables square measure the options and also the variable quantity that we'd prefer to predict is that the worth. it's apparent that the options that we tend to square measure exploitation don't seem to be really freelance, we all know that the amount and outstanding shares don't seem to be freelance further because the price and also the come on investment not being freelance.

This study aims to use completely different models to predict the worth changes and to judge the various model's success by withholding knowledge throughout coaching and evaluating the accuracy of those predictions exploitation legendary knowledge. These analysis considerations closing costs of the stocks. The model for the stock exchange was solely involved with the price for stocks at the top of a business day, high-frequency commerce is a locality of active analysis, however this study most popular a simplified model of the stock exchange.

2. MOTIVATION

Stock market price prediction is an issue that has the capacity to be worth billions of dollars and is actively studied by the largest financial corporations in the world. It is a relevant problem because it has no clear solution. Several attempts can be made at approximation using many machine learning techniques. The project allows methods for real-world machine learning applications including acquiring and analyzing a large data set and using a variety of methods to train the system and predict potential outcomes.

3. METHODOLOGY

3.1 Linear Regression

The regression method is finished through the sci-kit-learn machine learning library. This is often the core for the worth prediction practicality. There square measure some extra steps that have got to be done in order that the information will be fed into the regression algorithms and come plausible results. Especially, each coaching dataset should be normalized to a Gaussian usually distributed or normal-looking distribution between -1 and one before the input matrix is suited to the chosen regression model [1]. There square measure one or two necessary details to notice concerning the method the information should be pre-processed so as to match into regression models.

Firstly, dates square measure usually portrayed as strings of the format ”YYYY-MM-DD” once it involves info storage. This format should be born-again to one whole number so as to be used as a column within the feature matrix. This is often done by victimisation the date’s ordinal worth. In Python, this is often quite easy. The columns within the information Frame square measure hold on as numpy datetime64 objects, that should be born-again to vanilla Python date-time objects that square measure successively born-again to associate whole number victimisation the to ordinal() constitutional perform for date time objects. Every column within the feature matrix is then scaled victimisation scikitlearn’s scale() perform from the pre-processing module. Mean absolute error methodology is employed to gauge the performance of the regression model.

3.2 Logistic regression

Using the set of options chosen supply regression algorithm[2] was tried, mistreatment the linear model from Python's scikit-learn library, in an endeavour to classify as accurately as doable whether or not the subsequent day's London PM gold worth fix would be higher or not up to the present day's. The total set of gold worth fixes technical indicator options were used. This technique conjointly used l-1 norm for the penalty instead of the default l-2, because it achieved higher results. The typical results of running supply regression with tenfold cross-validation were as follows:

Table 1: Logistic Regression (Gold Price Features Only)

Precision	55.03%
Recall	76.27%
Accuracy	60%

So the recall of supply regression is over seventy fifths on the average that is extremely high. This suggests that nearly all positive examples were properly classified as positive. However, the preciseness is simply over fifty fifth, which is comparatively low. This is often as a result of the rule made several false positives; even for negative examples, it's additional doubtless to predict one than zero.

3.3 ELM

A single layer feed forward network with x1,x2,….xm ,input nodes,h1,h2,….,hn ,hidden nodes and ti be the target node .Let (ai, bi) be the weights connecting from the input layer to hidden layer and β1,β2,...,βn be the weights of the nodes connecting from hidden layer to the output layer. Let 'g' be the piecewise continuous activation function[5]. The hidden layer outputs are given as:

$$\sum_{i=1}^{N}[\beta_{\mathrm{i}}g(a_{\mathrm{i},}b_{\mathrm{i}},x_{\mathrm{j}})] = t_{\mathrm{j},}j = 1, \dots, N$$

Equation (1) can be rewritten as $\beta H = T$. Here H is called the hidden layer output matrix,which can be expressed as follows,

$$H(\, a_{1,}\dots\dots\dots, a_N\colon b_{1,}\dots\dots\dots b_N\colon x_1\, \dots\dots\dots x_M) =$$

$$\begin{pmatrix} G(a1, b1, x1)\,. & G(aN, bN, x1) \\ G(a1, b1, xM)\,. & G(aN, bN, xM) \end{pmatrix}$$

$$\beta = [\beta_1\beta_{2\dots\dots\dots}\beta_{\mathrm{n}}]^{\mathrm{T}} \text{ and } T = [t_1 t_{2\dots\dots\dots} t_{\mathrm{n}}]^{\mathrm{T}}$$

In the hidden output matrix, each value represents the hidden output values of their corresponding node. The three-step ELM learning algorithm is as follows:

In a Single layer feed forward neural network, (xi , ti) be the training pair and g(a, b, x) be the hidden node output function.
Step 1: Initially hidden nodes are chosen randomly, (ai, bi) where i = 1, 2, 3, ...,N.
Step 2: Calculate the hidden layer output matrix H
Step 3: Calculate the output weights Î¸ where Î¸= T. Here represents the Moore-Penrose inverse of hidden layer output matrix H.

3.4 ANFIS

The rule base contains the subsequent 2 Sugeno-type fuzzy if-then rules:
Rule 1: if x is A_1 and y is B_1 then $f_1=p_1x +q_1y+r_1$
Rule 2: if x is A_2 and y is B_2 then $f_1=p_2x+q_2y+r_2$

Wherever, x and y area unit the inputs, Ai and metallic element area unit the fuzzy sets, fi is that the output area unit the ensuing parameters that area unit determined throughout the coaching method. The five-layer ANFIS structure consists of fuzzification, inference, social control, consequent, and output. ANFIS [4], as conferred in figure 1, incorporates a five-layer network to implement a Takagi-Sugeno-type fuzzy system layer1 layer2 layer3 layer4 layer 5.

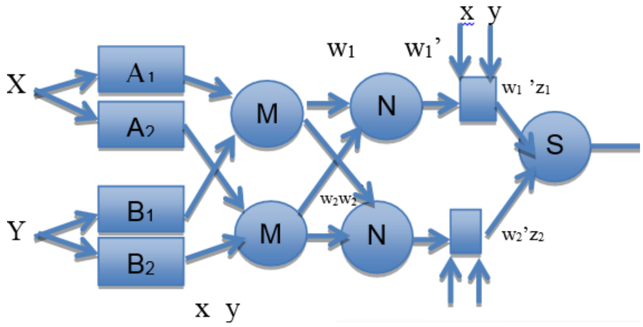

Fig. 1: The five-layer ANFIS structure

The ANFIS structure is described below:
Layer 1: Parameters in the first layer are called premise parameters; each node in this layer generates fuzzy membership grades. where ( ) Ai x and ( ) Bi x can adopt any fuzzy membership function (MF). In this paper, the Gaussian function is used to develop the prediction model.
Layer 2: Each node in the second layer calculates the firing strength of each rule via multiplication
Layer 3: Node i in this layer calculates the ratio of the ith rules firing strength to the sum of all rules' firing strengths.
Layer 4: In this layer, parameters are called consequent parameters and every node computes the contribution of ith rule towards the overall output:
Layer 5: Finally, the single node calculates the overall output as the summation of all incoming signals:

ANFIS uses the hybrid-learning algorithm, consists of the combination of gradient descent, and least-squares methods. The former is employed to determine the nonlinear input parameters and the latter is used to identify the linear output parameters. The main objective of the learning algorithm for ANFIS architecture is to tune all the modifiable parameters in order to match the ANFIS output with the training data. ANFIS applies two phases, including forwarding pass and backward pass to recognize the pattern of the given data set.

3.5 SVM

Two key parts within the implementation of SVM are the techniques of mathematical programming and kernel functions. The parameters are found by resolution a quadratic programming downside with linear equality and difference constraints; instead of by resolution a non-convex, at liberty optimisation downside. SVM algorithm [3] developed by Vapnik is predicated on applied math learning theory. SVM is used for each classification and regression task. In classification case, we have a tendency to associate realize an optimum hyper plane that separates 2 categories. So as to search out Associate in Nursing optimum hyper plane, we'd like to attenuate the norm of the vector w, that defines the separating hyper plane. This is often cherish increasing the margin between 2 categories.

$$f(x) = sign(\textstyle\sum_{i=1}^{N} \alpha i\; yi\; K(xi, x) +$$
$$\frac{1}{Ns}\textstyle\sum_{0<\alpha i<C}\; [yi \sum_{i=1}^{N} \alpha i\; yi\; K(x_{\mathrm{I}}, x_{\mathrm{j}})])$$

Mathematically, we are going to get quadratic programming down side wherever the amount of variables is up to the number of observations. Think about the matter of separating the set of coaching vector happiness to 2 separate categories.

Table 2: Comparison of table

Method	Generalization	Risk	Overfitting problem	Long term prediction	Accuracy	Efficiency in training
SVM	High	Less	Less vulnerable	Easy	Higher than ANFIS	100%
ANFI*S	Higher than ELM	Comparatively Less	Vulnerable than SVM	Easy	Higher than ELM	78%
ELM	Medium	Medium	Less vulnerable than Regression	Difficult than ANFIS	Lower than ANFIS	69%
Logistic Regression	Low	High	More vulnerable	Difficult	Low	60%
Linear Regression	Very Low	High	More vulnerable	Difficult	Low	57%

4. CONCLUSION

From the analysis of the result, it is clear that the prediction of the price of the fuel, the gold and the stock can be done efficiently with the SVM method. The SVM method is combined with sentiment analysis technique to obtain the correct predictions with respect to the current affairs collected from twitter. The combination of these two techniques makes this project unique from others as it concentrates more on the current affairs as well as on the higher accuracy rate. The SVM method is simple but when it is used with Twitter sentiment analysis it can outreach other methods because of its efficiency and simplicity.

5. REFERENCES

[1] Lucas Nunno, Stock Market Price Prediction Using Linear and Polynomial Regression Models University of New Mexico Computer Science Department Albuquerque, New Mexico, United States, lnunno@cs.unm.edu

[2] Megan Potoski, Predicting Gold Prices, CS229, Autumn 2013

[3] Abdolreza Yazdani-Chamzini1, Siamak Haji Yakhchali, Forecasting Gold Price Changes by Using Adaptive Network Fuzzy Inference System, Young Researchers Club, South Tehran Branch, Islamic Azad University, accepted 27 March 2012.

[4] Mr Sachin Sampat Patil, Prof. Kailash Patidar, Assistant Prof. Megha Jain, SSSIST, Sehore, Madhya Pradesh, India, Stock Market Prediction Using Support Vector Machine, International Journal of Current Trends in Engineering & Technology Volume: 02, Issue: 01 (JFAB,2016)18.

[5] S. Kumar Chandar, M. Sumathi, S.N. Sivanadam, Forecasting Gold Prices Based on Extreme Learning Machine, International Journal of Computers Communications and Control, ISSN1841-9836, 11(3):372-380, June 2016.