PREDICTING THE STOCK PRICE USING LINEAR REGRESSION

Sasidhar Reddy Bommareddy  
Department of Computer Science,  
REVA Institute of Technology and Management, Bangalore.

K Sai Smaran Reddy  
Department of Computer Science,  
REVA Institute of Technology and Management, Bangalore.

Kaushik P  
Department of Computer Science,  
REVA Institute of Technology and Management, Bangalore.

K V Vinay Kumar  
Department of Computer Science,  
REVA Institute of Technology and Management, Bangalore.

Dr. Vishwanath R Hulipalled  
School of Computing and Information Technology  
REVA University, Bangalore

Abstract: Stock price prediction is always a predominant goal for every investor which helps them to knowing the future prices considering the previous records. We made our effort in predicting the same using statistical modeling approach, the Linear Regression Technique for finding Open, Close, High and Low values of TCS from National Stock Exchange Of India (NSEI).

Keywords: Linear Regression, National Stock Exchange of India, Prediction, Stock Market.

I. INTRODUCTION

Stock Market is a public market for trading of company stock and derivatives at an agreed price. It is generally a dynamic market where the prices vary and it becomes difficult for an investor for predicting the prices considering external factors like factors like political situations, public image on the company according to efficient market hypothesis [16].

Stock Market became one of the integral parts of global economy to the extent that any fluctuation in the market influences personal and corporate financial lives and the economic health [2]. For the past decades, predicting the stock prices has been a trendy topic in financial applications.

If the accuracy of prediction is more, decisions can be taken easily for the future [1]. The risk of falling of stock prices is very rare due to market fluctuations, but there is a risk again [2]. Successful prediction of stock prices can yield significant profits. There are two basic methodologies investors rely on when the objective of analysis is to determine what stock to buy and at what price: Fundamental Analysis, Technical Analysis. Fundamental Analysis is the examination of the underlying forces that affect the well being of the economic, industry groups and companies. To forecast the future stock fundamental analysis combines economic, industry and company analysis along with political climate to derive a stock current fair value and forecast future value. On the other hand technical analysis is a methodology that is used to predict the direction of price movements based on current and past movements represented in a chart. This analysis basically uses charts to understand the pattern and trend of market.

In this paper we use linear regression, which is statistical learning technique where we predict the value depending on the criterion variable.

The main motivation behind this work is that it is very crucial for stock market investors to estimate the behavior or trend of stock market prices in order to invest in a company/category which is and is going to trend in profits in coming future. The unpredictable volatile market index makes it a highly challenging task to accurately forecast its path of movement. In this context we are comparing two predictive techniques to know the technique with best efficiency in terms of predicted values vs. the actual market values.

In the following sections, in section 2 we look at the literature review of some reference papers where we understand their work, later in section 3 we will go through the Linear regression. In further sections, we will look at the results we observed with conclusion and future scope of the work.

II. LITERATURE SURVEY

Muhammad Waquar et al [1] study, underlined the utilization of principal component analysis (PCA) to improve the performance of machine learning model in classification of high dimensional data. But as it was investigated they concluded that PCA does not always guarantee the improve of accuracy. Dinesh Bhuriya et al [2] used linear regression, polynomial and RBF regression to predict the stock prices using 5 variables and compared the above models and concluded that linear regression is best among all other used.

P ASamarak et al [3] in their research work stated that abnormality detection is key step for achieving the self-healing concept. With their study, they explored applications of supervised machine learning techniques to detect abnormal behavior in systems. When used different techniques author
found Random forest algorithm provides high level accuracy with oversampling techniques. Eswara Reddy et al [4] in their research paper for highly volatile financial TSD, they proposed ARIMA-GARCH model which is suitable for multi-step ahead forecasting, involves MA filter based decomposition as a preprocessing step on given TSD.

Harun Ercan [5] in his experimental results found that the forecast values are close to actual values on using NARX model on Baltic Stock Market. Quibin Liang et al [6] in their research on stock market trend prediction proposed a model to forecast the stock prices considering the importance of future representation for machine learning based model. They investigated Restricted Boltzmann Machine for feature extraction. On experimental results, they found that features extracted by Bernoulli Restricted Boltzmann Machine gave a higher direction accuracy. In addition, it is found that this method is only effective for trend prediction.

Kai Chen et al [7] work yielded good accuracy of results when normalization was considered. Also, SSE index accuracy was also improved. But this result was only in case of Shanghai Securities ETF180. Poonam Somani et al [8] in their research on stock market prediction used Hidden Markov Model. They surveyed on various techniques like neural networks, support vector machine and concluded that Markov Model is more efficient in extracting the information from the dataset.

Eslam Nader Desokey et al [9] in their research paper on Enhancing stock prediction clustering using K-Means with genetic algorithm (GA) has observed accuracy of 89.31% by selection of new model using new centroid selection optimization for K-Means with GA. Samaranawickrama et al [10] found that when considering the forecast error or test error MLP models produce the highest and the lowest errors. The forecasting accuracy of the best feed forward networks is approximately 99%. SRNN and LSTM networks generally produce lower errors compared with feed forward networks but in some occasions, the error is higher than feed forward networks. Compared to other two networks, GRU networks are producing comparatively higher forecasting errors.

Rohit Verma et al [11] study used a neural network for predicting the stock market but the accuracy was found to be accurate until when large data with sudden variations is not considered. Bihui Luo et al [12] used an algorithm in based on the derivation of the theory of calculus, which has strong versatility, and it is widely used in the field of application. The main disadvantage here is, the structure of Hidden layer is difficult to determine.

R.M. Kapila Taranga Rattnayanka et al [13] in their study focused mainly on identifying the suitable hybrid forecasting approach based on ANN with traditional ARIMA approach under the high volatility. R.M.C.D.K. Rajasinghe et al [14] work which was based on Random Walk Hypothesis, suggests the unpredictable nature of prices in financial market. Closing prices were predicted and it was found that actual and predicted are closely moving.

Felix Ming Fai Wong et al [15] proposed a unified latent factor model to model the joint correlation between stock price and newspaper content, which allows us to make predictions on individual stocks, even those that do not appear in the news, they used ADMM algorithm to formulate the sparse matrix factorization problem. Bing Yang et al [16] in their research on Stock market index prediction using deep neural network ensemble (DNN) after experiment on Shanghai stock market using DNN algorithm they concluded that the accuracy of prediction for close values is not satisfactory.

III. DATA AND PREPROCESSING

Data Collection: We have collected the history data of TCS from the National Stock Exchange of India (NSEI) from Quandl consisting of Open, Close, High, Low and Volume as attributes. Each attribute has its own value and time at which they are recorded. Total data of 100 months is collected in which 99 months is used for training and the rest 1 month for testing the performance of the model.

Pre-Processing: Data pre-processing is a collective name for all methods that aim to ensure the quality of the data. In this stage we basically perform pre-processing on the data by selecting the best features which are extracted from the data collected. The data collected is then modified for just selecting only the required type of the data and removing the unwanted data. The various pre-processing stages include Feature generation which is used for selecting only the required type of the data. The data is then trained and a model is built based on this for later prediction process. The data model built is then scaled and only the required data is selected. This is the pre-processing technique.

Prediction Process: In this process, Linear Regression algorithm is applied on trained data where we predict the prices of Open, Close, High, Low values. This is the stage where the actual algorithm/technique is applied and the prediction process takes place.

Linear Regression Mathematical Calculation:

Linear regression model tries to produce the best possible straight line for the dataset. For determining the best fit we attempt to minimize the distance between all points and their distance to our line. In this regression technique we predict the value of one variable of Y from the other values of X, where Y is the Criterion Variable and X is called the Predictor Variable that we are basing our Prediction.

The model calculates the prediction with the following formulae

\[ y = a + bx \]  (1)

Here ‘a’ represents the index or the intercept and ‘b’ represents the slope or the coefficient of the variable ‘x’.

\[ a = \bar{y} - b\bar{x} \]  (2)

The value of ‘b’ is calculated by the formula:

\[ b = \frac{n\sum xy - (\sum x\sum y)}{n\sum x^2 - (\sum x)^2} \]  (3)
The values of the above terms are calculated as follows:

\[ \bar{y} = \frac{\sum y}{n} \]  
(4)

Where, \( \sum y = y_1 + y_2 + y_3 + \ldots + y_n \)

\[ \bar{x} = \frac{\sum x}{n} \]  
(5)

Where, \( \sum x = x_1 + x_2 + x_3 + \ldots + x_n \)

After calculating the values of (4) and (5) substitute those in equation (2). Now calculate equation (3) and finally substitute (2) and (3) in equation (1).

The following is the Algorithm which we have used for prediction using Linear Regression:

```
Algorithm: Linear Regression

Input: x_train: Open values of a company
       y_train: Close values of a company
       x_test: Open value for which close value is predicted
       n:- Size of training dataset

Output: predict_value:- Predicted value.

1: def linear_regression (x_train, y_train, x_test, n)
2: begin
3:   for i = 0 to n
4:     begin
5:       x_sum = 0, y_sum=0
6:       x_sum += x[i]
7:       y_sum += y[i]
8:     end
9:     x_mean = x_sum / n
10:    y_mean = y_sum / n
11:   for i = 0 to n
12:     begin
13:       xy_sum += x[i] * y[i]
14:       x2_sum += x[i] * x[i]
15:     end
16:     x_sum_sq = x_sum * x_sum
17:     l= n * xy_sum - (x_sum * y_sum)
18:     m = n * x2_sum - x_sum_sq
19:     b = l / m
20:     a = y_mean - ( b * x_mean)
21:     predict_value = a + b * x_test
22:     return predict_value
23: end
```

Model Evaluation: This is the final stage in the model. In this stage we are evaluating the performance of the algorithm in terms of Mean Squared Error (MSE), Mean Absolute Error (MAE) and R2-Score.

\[ MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y_i})^2 \]
\[ MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{Y_i}|}{n} \]

IV. RESULT ANALYSIS

![Fig 1](#): Comparison of actual open values with predicted open values for company TCS between 25-03-2018 to 24-04-2018.

![Fig 2](#): Comparison of actual high values with predicted high values for company TCS between 25-03-2018 to 24-04-2018.

![Fig 3](#): Comparison of actual low values with predicted low values for company TCS between 25-03-2018 to 24-04-2018.
Fig 4: Comparison of actual close values with predicted close values for company TCS between 25-03-2018 to 24-04-2018.

Table I
Error Analyses of Linear Regression Technique.

<table>
<thead>
<tr>
<th>Models</th>
<th>Open</th>
<th>High</th>
<th>Low</th>
<th>Close</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>326.843</td>
<td>2498.009</td>
<td>295.123</td>
<td>2277.900</td>
</tr>
<tr>
<td>MAE</td>
<td>13.925</td>
<td>36.536</td>
<td>13.525</td>
<td>34.945</td>
</tr>
<tr>
<td>R² Score</td>
<td>0.990</td>
<td>0.941</td>
<td>0.990</td>
<td>0.934</td>
</tr>
</tbody>
</table>

CONCLUSION

We performed study on TCS stock prices in National Stock Exchange of India (NSEI) using Linear Regression Technique by predicting the values of Open, Close, High, and Low. Our main goal for this study is to assist stock market investors understand the future prices of TCS as predicting the stock prices is always a challenging task as the market is dynamic in nature.

Future scope of this study involves considering more multiple companies from any stock exchange of different countries and also performing comparison of any different techniques of prediction so that one can understand which technique has less MSE, MAE and R² Score.

REFERENCES


