Front.Comput.Sci.
DOI

RESEARCH ARTICLE

A Regression-Based Share Market Prediction Model for Bangladesh

Rubaiyat Jahan Mumu (✉), Syeda Tasnim Fabiha (✉), Farzana Aktar(✉)
B M Mainul Hossain

Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh

Abstract Share market is one of the most important sectors of economic development of a country. Everyday almost all companies issue their shares and investors buy and sell shares of these companies. Generally investors want to buy shares of the companies whose market liquidity is comparatively greater. Market liquidity depends on the average price of a share. In this paper, a thorough linear regression analysis has been performed on the stock market data of Dhaka Stock Exchange. Later, the linear model has been compared with random forest based on different metrics showing better results for random forest model. However, the amount of individual significance of different factors on the variability of stock price has been identified and explained. This paper also shows that the time series data is not capable of generating a predictive linear model for analysis.

Keywords regression, share market, parametric method

1 Introduction

Development of a nation's capital market often reflects its economic development. As a component of this market, the stock market, which is also known as share market, plays an important role in industrial development through supply of capital. The importance of institutions like the stock exchanges cannot be overemphasized for the development of capital market. Share market is a place where shares of different companies are allowed to be brought and sold. Bonds, mutual funds and derivative contracts are also traded along with share of the companies. There are two types of share markets. One type is called the primary share market and the other type is secondary share market. In primary share market, a company get registered to issue shares to the public and thus raise money. Generally, companies are get listed on the stock exchange through primary market route. In secondary share market, investors buy shares from other investors at prevailing price.

The capital market is vital to the long-run growth and prosperity of the business sector, as it facilitates the transfer of funds from savers to investors. By requiring the disclosure of certain corporate financial data, an efficient market system allows investors to assess the risk-return trade-offs involved in a transaction and move funds towards comparatively more promising investments. The secondary markets allows investors and financial institutions to alter the liquidity, composition and risks of their portfolios in response to newer information on changes in market condition. A number of brokers conduct these transactions in secondary share market. An investor may sell all its shares and exit the financial market if it wants. Every investor wants to buy share at lowest price and sell at comparatively higher price. Therefore, investors want to predict the market change rate earlier. Those, who can predict the market, are the gainers.

In Bangladesh, investors become more interested in the secondary market for its liquidity, and the inherent return possibility due to the volatility of the market. The first stock market operator of Bangladesh is "Dhaka Stock Exchange", was set up in 1954. To integrate the entire economic role, it was reactivated in 1976 with the listing of nine companies. The capital market of Bangladesh has already experienced some major collapse 2010 [1]. The incident was popularly known as "2010-11 Bangladesh Share Market Scam". The major reasons of this incident was entrance of inexperi-

Received month dd, yyyy; accepted month dd, yyyy

E-mail:

enced investors and the sudden unsustainable growth of stock market in a very short period. This research has focused on the features and factors of a company which affects company movement. For this analysis, weekly time series data of certain companies has been collected from January, 2012 to May, 2015. A regression analysis has been conducted on these data and a report is generated finally.

A large number of studies had been conducted on the status and determinates of stock exchange market. A study was conducted focusing towards identifying the factors of stock price volatility and these effects on A-rated companies and B-rated companies [2]. It used multiple linear regression model using SPSS to determine the statistical significance of the independent variables. Another study was conducted to see the effectiveness of 17 variables for the movement of stock price in Bangladesh capital market by using SPSS. It was also conducted on Dhaka Stock Exchange of Bangladesh (DSEBd). Later this study found 5 core factors affecting price movement. [3].

As accurate stock market prediction is one of the major problem for investors, a study used neural networks to produce a successful model that could be used for stock market prediction. The prediction system was made up of several neural networks that learned the relationships between various technical and economical indexes and the timing for when to buy and sell stocks. The goal of that study was to predict the best time to buy and sell for one month in the future [4]. Researches on Karachi Stock Exchange (KSE) was done to predict the market performance on day closing using different machine learning techniques. It compared several machine learning techniques including Single Layer Perceptron (SLP), Radial Basis Function (RBF) and Support Vector Machine (SVM) [5].

The previous works on stock market prediction considered various multi-dimensional factors for generating a predictive model. However, there have not been many works on the thorough analysis of stock market data of Bangladesh. In this study we propose to run a predictive analysis for stock market price of Dhaka Stock Exchange. As a result of analysis, we will be able to tell the amount of effect of each factor on the variability of the stock price and predict future stock prices of the companies with a significant confidence level.

2 Methodology

For generating an efficient predictive model for the stock market dataset, two different types of methods are being followed in this work. One is parametric: linear regression and another is non-parametric: random forest. Initially all the predictors have been considered as important contributors for the model generation. Later, analysis is performed on the identification of the best predictors, which contribute most for describing the variability of the stock closing prices.

2.1 Linear Regression

Linear regression is an approach of linear modeling between a dependent variable Y and one or more independent variables. If the number of independent variable is one, then it is called simple linear regression. The example equation for simple linear regression is given below:

$$\hat{y} = \beta_0 + \beta_1 x \tag{1}$$

where $\hat{y}$ is the response variable, x is the predictor, c is the intercept and β is the slope of a linear equation.

When linear Regression is performed for multiple predictors, it is Multiple Linear Regression. It gives each predictor a separate slope coefficient. If we have n distinct predictors. Then multiple linear regression model is:

$$\hat{y} = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \tag{2}$$

Here, ϵ is the error term in regression models. In a linear Regression model, attempt is made to predict $\beta_0, \beta_1, \cdots, \beta_n$ by minimizing the Residual Sum of Squares (RSS) using the Least-Squares method.

$$RSS = \sum_{i=1}^{k} (y_i - \hat{y}_i)^2 \tag{3}$$

Like the simple linear regression, the multiple linear regression also aims to reduce RSS by substituting the value of $\hat{y}$ into Equation 3.

$$RSS = \sum_{i=1}^{k} (y_i - (\beta_1 x_1 + beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + c)) \tag{4}$$

After fitting the linear regression model to a particular data-set, some problems may arise based on different data properties such as non-linearity of response-predictor relationships, non-constant variance of error terms, out-liars, high leverage points, co-linearity etc. These problems can result in small R square and t values and bigger p values indicating lack of significant relationship between the predictors and response. To overcome these problems, different data manipulation techniques are performed on our data-set such as transformation ($\log x$, $\sqrt{x}$, x^2), use of weighted least

squares, removing out-liars, high leverage points, dropping or combining co-linear variables using Variance Inflation Factor (VIF) etc.

2.2 Random Forest

This tree-based approach involves producing multiple trees that are combined to generate a single prediction with probable high accuracy. In order to avoid the high variance problem of decision trees, bagging increases the prediction accuracy at the expense of loss of interpretability. Random Forest is an improvement for the bagging concept [6]. Instead of bootstrapping of training samples used by Bagging, a random sample of m predictors is chosen as split candidates from the full set of n predictors where typically $m \approx \sqrt{n}$.

2.3 Evaluation Metrics

Three metrics Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Pearson Correlation Coefficient are used to compare the two above-mentioned approaches for model generation.

2.3.1 MAE

MAE gives the average magnitude of errors in a set of predictions without considering their direction. ItâĂŹs the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y_j}) \tag{5}$$

2.3.2 RMSE

This metric also calculates average magnitude of error. The errors are squared before they are averaged giving the large errors relatively higher weights. This means RMSE should be more useful when large errors are particularly undesirable.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y_j})^2} \tag{6}$$

2.3.3 Pearson Correlation Coefficient

This metric gives an idea about the strength of relationship between the response and the predictors. The amount of variability in the responses is explained using correlation coefficient. When there are multiple predictors, then it is termed as R square.

2.4 Support Vector Machine

3 Experimentation

In this section the data description is provided after being explored. Moreover, the Paired T-Test performed to select the best predictor is also presented.

3.1 Data

Most of the data was collected from the Stock market related websites of Bangladesh like Lanka Bangla Financial Portal, Dhaka Stock Exchange Bangladesh Ltd. and Trading Economics [7] [8] [9]. From these sites, stock market related data from 2013 to 2017 for five pharmaceutical companies are collected and aggregated to form the training data. Data was collected on quarter basis from 4th quarter of 2013 to 1st quarter of 2017. The five selected companies are: ACI Pharmaceuticals, Beximco Pharmaceuticals Ltd, Renata Limited, Glaxo SmithKline Bangladesh Ltd and Square Pharmaceuticals Limited. The descriptions of the predictors and the response are given in Table 1. Stock prices change because of supply and demand. If more people want to buy a stock than sell it, then the price scales up. Usually, best performing companies tend to have increased stock price. If a company is earning is better than expected, the stock price jumps up. The possible predictors are shown in table 1. This is a time series dataset i.e. readings of the same predictors are taken at unit intervals of time. The current ratio (CR) considers the current total assets of a company (both liquid and non-liquid) relative to that company's current total liabilities. The current ratio is also known as the working capital ratio. The Asset Turnover ratio is used as an indicator of the efficiency of a company in deploying its assets for revenue generation.

Table 2 shows F-statistic, R square and residual sum of squares, RSE values. From the data, TATO, DTA, ROE have comparatively larger F-statistic value. This indicates the existence of relationship between the response and the predictors. RSE estimates the standard deviation of the response from the population regression line. These values indicate the amount of error with respect to the mean values.R square statistic records the percentage of variability in the response that is explained by the predictors.

Figure 2 shows correlation among the predictors of the dataset. It is evident that three predictors DTA, TOTA, ROE are significant relationship with stock price each having correlation value 0.59, .71 and .43 respectively.

Table 1: Predictor Name and Data Description

Type	Predictor Name	Acronym	Data Description
Predictor	Term	Term	Term means the quarter number
	Stock Price	Stock Price	Stock price means the price of each share of a company
	Return on Asset	ROA	Return on assets (ROA) is an indicates the profitability of a company related to its total assets.
	Return on Equity	ROE	Return on equity indicates the net income returned as a percentage of shareholders equity.
	Current Ratio	CR	The current ratio means the liquidity ratio to measure a company's ability to pay short-term and long-term obligations
	Total Asset Turn Over	TATO	Asset turnover is the the value of a company's sales or revenues generated relative to the value of its assets
	Debt to Total Asset	DTA	Total debt to total assets is a leverage ratio that defines the total amount of debt relative to assets
	Panel	Panel	Panel is the unique code for each company

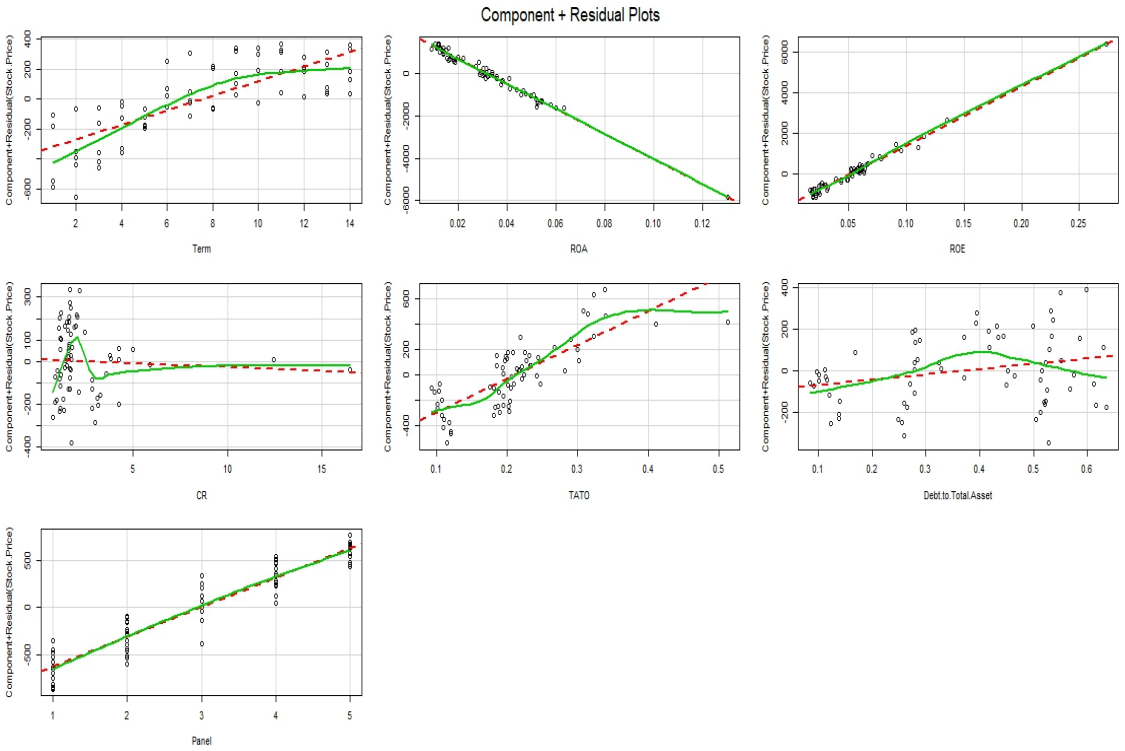
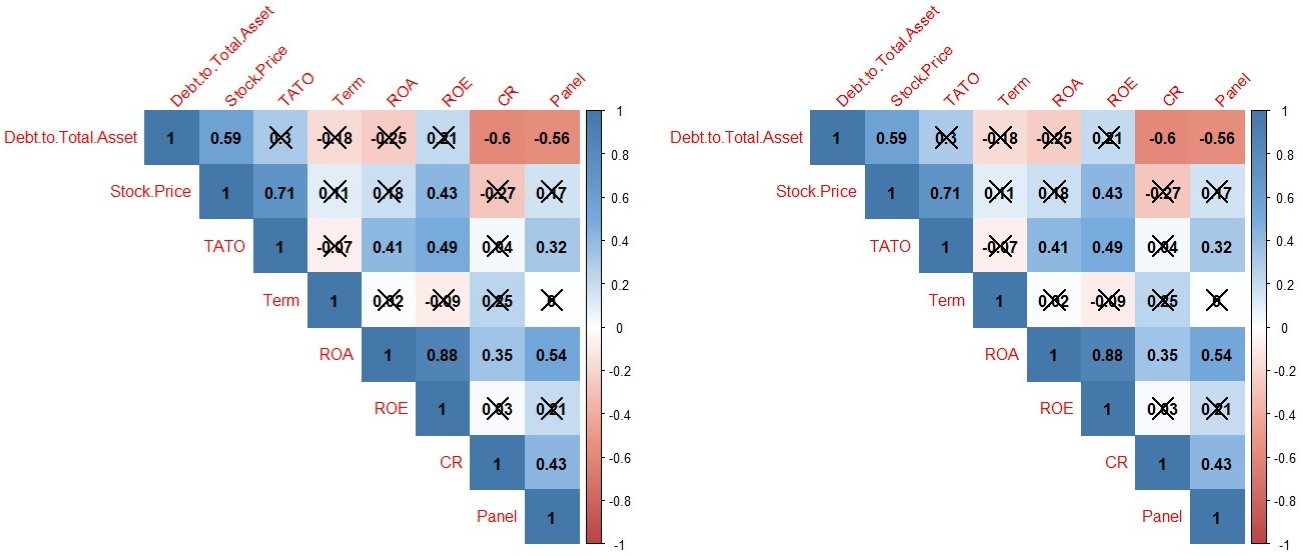


Fig. 1: Component Residual Plots for the Predictors

4 Fitting to Linear Model

Figure 1 shows the component residual plots for each of the predictors. It is evident that in most cases, the difference between residual line and component line is small which indicates linearity.However, for Term, CR, TATO, DTA the difference is larger. Therefore, these predictors are transformed for a linear relationship according to Tukey power transformation [10]. As a solution to this issue, we performed transformation and found that neither using square root nor taking log value improved results. Therefore, no transformation was applied on the predictors.

Observing the residual versus fitted plot shown in figure 3, we find no significant pattern. This shows evidence of linearity in the model. Also after removal of outliers, the plot generated more effective results.From the plot, we do not observe any funnel shape that represents absence of non-constant variance of errors in the data called Heteroscedasticity. Therefore, the model satisfies the homoscedasticity property of fitted Linear Model.After log-transforming the data,



(a) 0.05 Significance Level

(b) 0.01 Significance Level

Fig. 2: Pearson correlation of the Features (Non-significant features are marked with a cross mark (X))

Table 2: Coefficients of Linear Regression Analysis (****p<0.001, ***p<0.01, **p<0.05, *p<0.1)

Predictors	**F-stat**	R^2	**RSE**
Panel	1.861	0.02783	565.20
DTA	33.85	0.3425	464.80
ROE	14.59	0.1834	518.00
ROA	2.122	0.03027	594.00
TATO	67.9	0.5109	400
CR	5.272	0.7506	551.30
Term	0.7566	0.01151	569.00

we found the shape disappeared leaving some outliers. Three outliers 3, 38, 48 were found and removed from the dataset.

In figure 4a the Variance Inflation Factor (VIF) values for the predictors including ROA generated very high values for ROA and ROE. This means that the ROA and ROE values must be collinear. We excluded ROA from the predictor set. Then the VIF values of all the predictors became less than 4 indicating absence of multicollinearity in the model. From figure 4b, normal Q-Q plot shows that standardized residuals closely follow a linear trend. This confirms normality of residuals.

Figure 5 shows the autocorrelation function plot for residual time series. The first line indicates the correlation of the residual with itself. This is why it is larger. The next lags of the residuals are small and down the dotted blue line, which indicates that current values of residuals are not dependent on previous values.

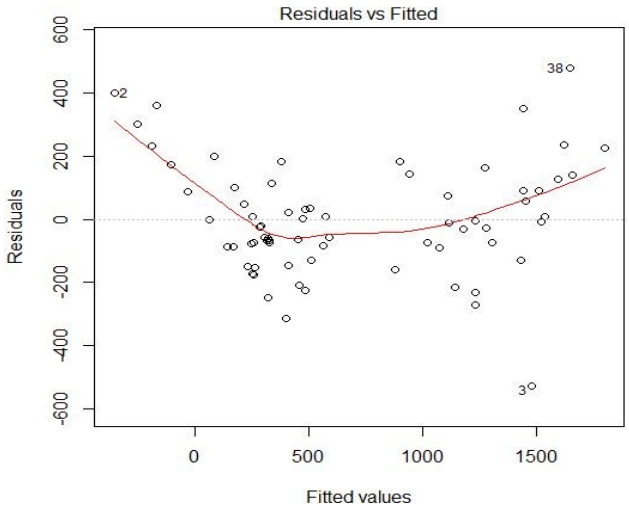
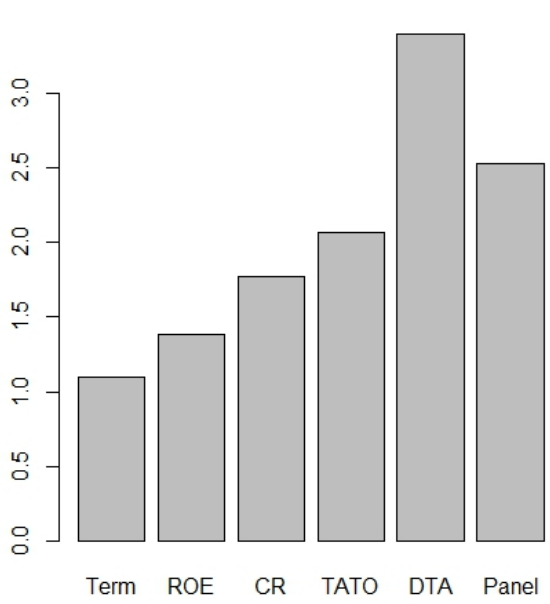


Fig. 3: Regression Diagnostic Plots

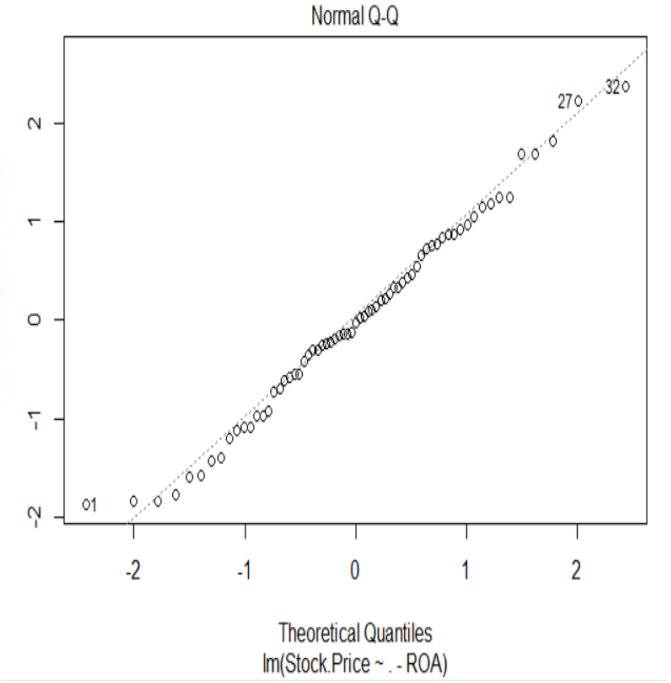
After we are done with data pre-processing to fit them into a linear model, it is found that our model satisfies the conditions of linear modelling such as measurement error free observation of predictors and responses, linearity, normality, homoscedasticity, no multicollinearity and no autocorrelation.

5 Application of Evaluation Metrics

Though the MAE, RMSE and Correlation coefficient values for linear regression and random forest both showed signifi-



(a) Variance Inflation Factors of the Predictors



(b) Regression Diagnostic Plots

Fig. 4: VIF and Normal Q-Q

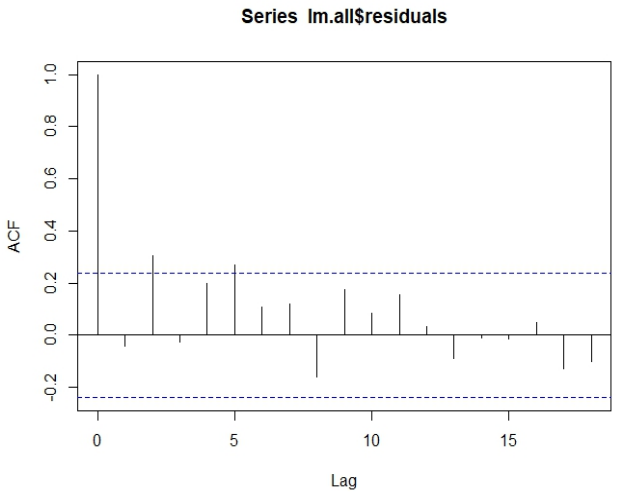


Fig. 5: Autocovariance and Autocorrelation Functions Plot

cant results. On this dataset, random forest outperforms linear regression as showed in table 3.

The possible reason behind this is that the stock market data is a time series data as shown in figure 6. This kind of time-series data have inherent collinearity among predictors [11] [12] [13]. Because of this, the model is more biased towards having better performance using tree based method

Table 3: MAE, RMSE and Correlation Coefficient values (* indicates statistically significant at 0.05 level compared to Linear Regression)

Metrics	Linear Regression	Support Vector Machine	Random Forest
MAE	220.8		119.9
RMSE	283.7		195.9
Correlation Coefficient	0.88		0.9

random forest.

6 Results and Discussion

As discussed previously, the data is fit to a linear regression model. To evaluate the model, 10-fold cross validation is performed. The R2 is 0.88 that indicates data fits well to the linear model.

The F-Statistic is 56.21 on 6 and 60 degree of freedom. The p-value of the F-Test is 2.2e-16 which is much less than the 0.05 confidence interval. For this, the null hypothesis that the fit of the intercept-only model and the fitted model is equal, can be rejected. Table 4 shows the linear regres-

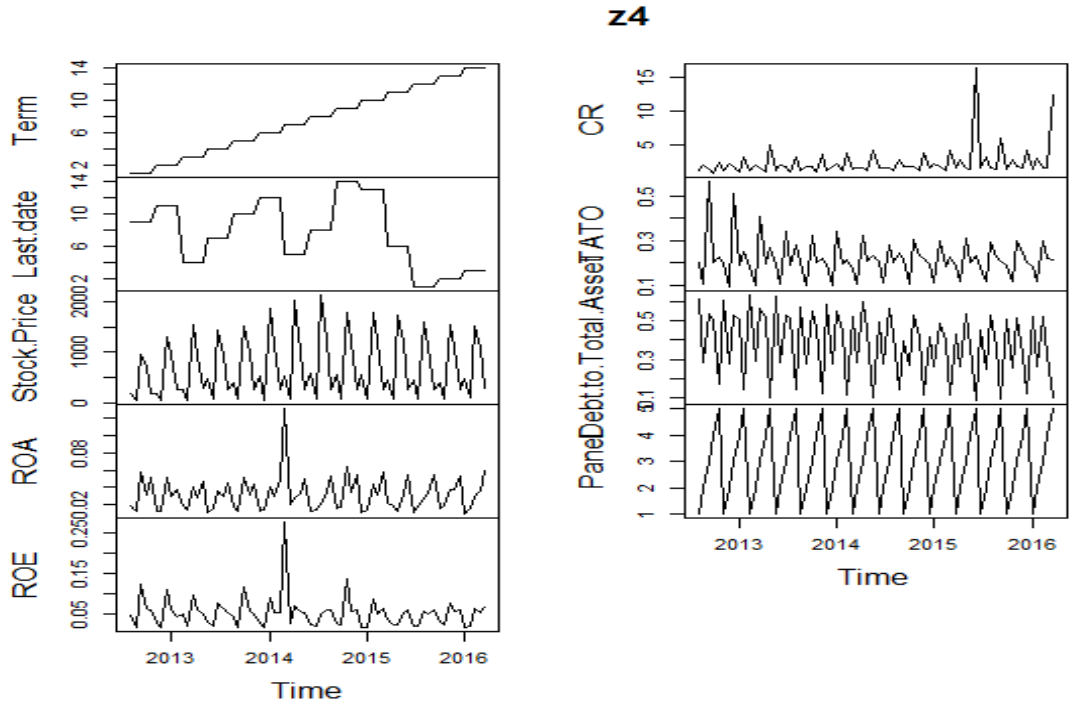


Fig. 6: Time Series Plot

Table 4: Coefficients of Linear Regression Analysis with Principal Components (***p<0.0001, *p<0.05)

Predictors	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-1691.838	167.744	-10.086	$1.55e^{-14}$***
Term	42.113	7.348	5.731	$3.44e^{-07}$ ***
ROE	342.091	902.390	0.379	0.706
CR	-32.597	15.769	-2.067	0.403
TATO	2367.872	824.498	8.515	$3.02e^{-05}$
DTA	2536.810	308.046	8.235	$1.95e^{-11}$***
Panel	213.175	31.151	6.843	$4.63e^{-11}$***

sion results. The null hypothesis for the t-test is that there is no relationship between the predictor and the response. For p values 0.001, 0.01, 0.05 and 0.1, we observe if any value is larger than t. It is evident that we can reject the null hypothesis in case of most of the predictors. At 0.001 confidence level, Term, TATO, DTA are Panel are related to the response. This means that, time sequence of stock data, total asset turnover of a company, Debt on total asset and company type impact the stock price of a day. The signs of their slope estimates indicate that stock price changes in both direction with the changes of these variables.

In this regression analysis, ROE estimate is positive showing the increased stock price of shares results in generation of more profit from the money shareholders have invested. However, stock price does not depend much on the amount of profit a company is making because the relationship is reverse. Increased stock price may lead to greater profitability. The estimate of CR is negative also explains the fact clearly that having large values for CR decreases stock price of shares to make greater amount of sales. Generally, the higher the asset turnover ratio, the better the company is performing, the greater is its demand of shares depicting significant increase in its stock price. In this analysis, the maximum estimate is achieved for debt to total asset (DTA) value. The reason behind this can be explained this way, suppose, a company having large debt to total asset ratio tends to be on the list of companies holding greater potential of investment opportunities. This increases their amount of debt though depicting potential lack of financial flexibility. Thus, shareholders tend to buy shares from these companies creating a high demand for their shares and thus increasing the stick price.

From these findings, we get some general decision criteria for predicting stock prices of share market. Due to fast changing rate of stock price, it is very difficult to predict exact stock price. However, the total asset turnover and debt to asset value indicating the current value of a company are potential predictors for predicting the stock price. Again, the main theory of price movement of a stock is what investors feel a company is worth. Therefore, though we can make some idea about public emotion at a given time, it is not possible to quantify this kind of variable.

7 Conclusion

This paper presents a price predictive model for Bangladesh for predicting stock price and analyzing influence of multiple factors on stock price. It has been shown that considering the most significant factors, prediction can be done with an acceptable accuracy measuring 119.9 in MAE, 195.9 in RMSE and 0.9 in Correlation Coefficient using a random forest model. Regression analysis results show that factors related to return on equity, total asset turnover, debt to total asset, company sector are the most significant ones. So, the specific factors belonging to these classes can be influenced to predict stock price of companies' shares in Bangladesh.

References

1. Syed Tashfin Chowdhury. (12 october 2011) bangladesh starts market rescue fund.
2. Samiul Parvez Ahmed, GM Wali Ullah, and Rashida Akter Tanzia. Factors affecting the stock price variability of dhaka stock exchange (dse). *Independent Business Review*, 7(2):61, 2014.
3. Md Shariful Islam, Mohammad Abdus Salam, and Md Mahmud Hasan. Factors affecting the stock price movement: A case study on dhaka stock exchange. *International Journal of Business and Management*, 10(10):253, 2015.
4. Takashi Kimoto, Kazuo Asakawa, Morio Yoda, and Masakazu Takeoka. Stock market prediction system with modular neural networks. In *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*, pages 1–6. IEEE, 1990.
5. Mehak Usmani, Syed Hasan Adil, Kamran Raza, and Syed Saad Azhar Ali. Stock market prediction using machine learning techniques. In *Computer and Information Sciences (ICCOINS), 2016 3rd International Conference on*, pages 322–327. IEEE, 2016.
6. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
7. Lankabangla financial portal. Online; Accessed: 1 November, 2017.
8. Dhaka stock exchange ltd. Online; Accessed: 4 November, 2017.
9. Trading economics. Online; Accessed: 4 November, 2017.
10. John W Tukey. Exploratory data analysis. 1977.
11. James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.
12. Problems that may occur in time series multiple regression. Online; Accessed: 6 November, 2017.
13. Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.