



Development of stock market trend prediction system using multiple regression

Muhammad Zubair Asghar¹ · Fazal Rahman¹ · Fazal Masud Kundi¹ · Shakeel Ahmad²

Published online: 14 February 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

The Stock market trend prediction is an efficient medium for investors, public companies and government to invest money by taking into account the profit and risk. The existing studies on the development of stock-based prediction systems rely on data acquired from social media sources (sentiment-based) and secondary data sources (financial-sites). However, the data acquired from such sources is usually sparse in nature. Moreover, the selection of predictor variables is also poor, which ultimately degrades the performance of prediction model. The problems associated with existing approaches can be overcome by proposing an effective prediction model with improved quality of input data and enhanced selection/inclusion of predictor variables. This work presents the results of stock prediction by applying a multiple regression model using R software. The results obtained show that the proposed system achieved a prediction accuracy of 95% on KSE 100-index dataset, 89% on Lucky Cement, 97% on Abbot Company dataset. Furthermore, user-friendly interface is provided to assist individuals and companies to invest or not in a specific stock.

Keywords Stock market · Prediction · Data sparseness · Multiple regression · Stock predictors · R

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10588-019-09292-7>) contains supplementary material, which is available to authorized users.

✉ Muhammad Zubair Asghar
zubair@gu.edu.pk

¹ Institute of Computing and Information Technology, Gomal University, Dera Ismail Khan (KP), Pakistan

² Faculty of Computing and Information Technology in Rabigh (FCITR), King Abdul Aziz University (KAU), Jeddah (Rabigh), Saudi Arabia

1 Introduction

Prediction of different events like flood, weather, earthquake, election, and stock are considered as the challenging tasks and have a major impact on human life. Stock prices prediction is a complicated and dynamic process due to rapid prices movement and market behavior. The prediction techniques in the stock market can play a crucial role in bringing more people and existing investors at one place (Zhang et al. 2013). The Stock market forecasting and analysis have gained the attention of researchers in the field of data mining to develop applications which could assist individuals as well as business organizations in decision making to invest money or not on a specific stock by taking into account the profit and risk.

The major objective of stock market prediction is to design software applications that can (i) analyze stock-related data from different financial resources, (ii) process that data using some prediction model, and (iii) present information to users. Design of such applications receive the considerable attention from financial experts, economists, businessmen, and others. Many people, business and organization consider a stock business as a profitable resource and wish to take part in the particular business. However, before purchasing and selling the stock, they have to estimate the profit and demand of particular stock using online financial data and trends. Therefore, the stock prediction is a challenging task due to different factors, such as retrieval and analysis of relevant stock data, fitting data into appropriate model and interpretation of results (Ladan et al. 2014).

The existing studies (Ladan et al. 2014; Kamley et al. 2013; Javaid 2010; Reddy 2010; Yuan and Luo 2014; Park et al. 2010; Ariyo et al. 2014; Devi et al. 2013) on stock-based prediction are based on the sentiment-based and secondary data acquired from social sites and financial sources. The sentiment-based approaches for stock market prediction mainly rely on unstructured data extracted from social media sites. The major limitation of such approaches is the low accuracy of prediction model due to noisy nature of input data. The other category of stock-oriented prediction techniques is based on analytical models using secondary data. The analytical approaches provide a more accurate prediction as compared to sentiment-based techniques due to well-structured nature of input data. However, such methods for the development of stock-based prediction systems rely on data that is usually sparse in nature. Moreover, the selection of predictor variables is also poor, which ultimately degrades the performance of prediction model. The problems associated with existing approaches can be overcome by proposing an effective prediction model with improved quality of input data and revised selection/inclusion of predictor variables.

The proposed method is based on multiple regression analysis of stock trend prediction, supported by a revised set of stock indicators. The key addition to the state of art methods (Ladan et al. 2014; Kamley et al. 2013; Javaid 2010) is in the way it preprocesses the input data with a revised set of stock indicators. Our system can take dataset as input, apply preprocessing steps, perform prediction, and gives a recommendation about the stock investment to user on the basis of prediction. The proper selection of stock indicators and applying preprocessing

steps provides efficiency of the proposed system with respect to the state-of-the-art methods in terms more accurate prediction of the stock trend.

The main of the contributions of this work include the development of a multiple regression-based stock market trend prediction system using a revised set of predictors (Appendix A of Supplementary Material). Following is the synopsis of contributions presented in this work.

- Applying different pre-processing steps for dimensionality reduction, such as data cleaning using fill in missing values and data reduction using volume compaction.
- To develop a multiple regression-based prediction model for predicting stock trend using a revised set of predictors.
- To provide a user-friendly decision support interface for individuals and companies to invest or not in the stock market.
- Evaluating performance of the proposed system with respect to baseline methods.

To achieve aforementioned objectives, the intention is to contribute knowledge, beneficial for stock market investors by developing a stock trend prediction system based on multiple regression model to assist the individuals and companies interested to take a decision regarding making an investment in the stock market.

The rest of paper is structured as follows. Section 2 presents literature review. In Sect. 3, we describe the proposed method. Experiment design is presented in Sect. 4. The final section outlines the work with a discussion on how it can be expanded in future.

2 Related work

The literature review section deals with the discussion on relevant studies on stock market trend prediction. The prior works are classified into two major categories, namely (i) analytical approach, and (ii) sentiment-based approach. Figure 1 shows proposed classification scheme of the literature review.

The first level shows that there are two major approaches, namely (i) analytical, and (ii) sentiment-based for stock market prediction. The level two shows related data sources for each approach. The last level depicts the relevant models used for both of the aforementioned approaches of stock prediction. The rest of the section is based on the aforementioned literature review classification scheme.

2.1 Analytical approaches

These approaches deal with the acquisition of secondary and historical stock data from different financial sources, such as Yahoo Finance, Google Finance, and Pakistan Stock Exchange. Analytical approaches are further classified into Regression driven and Autoregressive Integrated Moving Average (ARIMA) techniques.

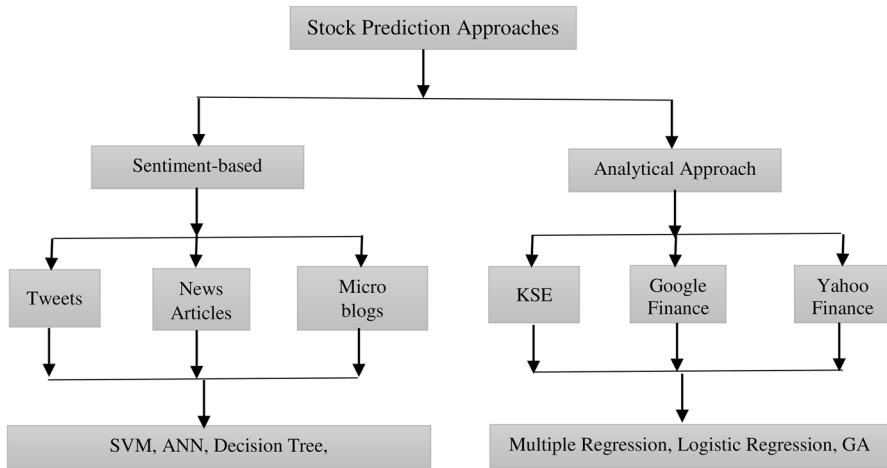


Fig. 1 Classification of stock prediction approaches

The regression driven approaches are based on different regression models such as linear regression, multiple regression, logistic regression and genetic algorithm. The collected data is passed through a specialized pre-processing module to prepare it for further processing. Finally, pre-processed data is made an input to the prediction module.

In data mining and statistics, multiple regression model is a technique for estimating the relationship between a dependent and one or more independent variables (Zhang et al. 2013). In multiple regression, an attempt is made to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation. The model associates each value of the independent variable x with a value of the dependent variable y . Following are some related works based on regression paradigm.

Ladan et al. (2014) introduced a multiple regression model to establish the relationship between different macroeconomic variables, namely: inflation rate, exchange rate and interest rate and stock prices. They compiled the dataset from CBN Statistical Bulletin 2010 available at www.cenbank.org. It is reported that exchange rate is significant, whereas interest rate is not significant and there exists a high correlation between the macroeconomic variables, such as inflation rate, exchange rate, and stock market returns. Moreover, it is observed that 54.2% of the total variation in all Index can be justified by the microeconomic variable. However, they used only limited number of macroeconomic variables, which if increased, can produce better results.

Kamley et al. (2013) applied the multiple regression approach to predict the stock market price from the stock market data (yahoo finance) using three variables, namely: open, high and close. Different pre-processing steps are applied for the cleansing, integration, and transformation of data. They achieved a prediction accuracy of 89%. However, the major limitation of their work includes lack sufficient training of the model due to poor selection of stock indicators.

Javaid (2010) used multiple regression model to predict market shares of companies on different parameters, namely: KIBOR, dividend, earning per share, gross domestic product, and inflation. They used Karachi-Stock-Exchange (KSE) 30-Index dataset and achieved up to 62% prediction accuracy. However, more enhanced variables can be used to predict the stock trend.

Reddy (2010) developed a statistical analysis tool (SAS) system to forecast stock market data by collecting secondary data over different periods from the National Stock Exchange (Nifty). They identified the seasonal differences and reported that the data series is likely to more accurate to build the model such as auto regression (AR), moving average (MA). Finally, forecasting is made for the market index by presenting change in market indices in a graphics mode. The limitation of AR model includes the huge difference between predicted and real-time values, due to some external factors, which often leads to inaccurate prediction.

Yuan and Luo (2014) analyzed the factors: price, trading volume, and position which affect the price movements. The required data is acquired from the “wind information terminal”, including the index of opening price, closing price, highest and the lowest price, volume and total holdings. The decision tree model is applied to predict price movement based on the selected variables by acquiring a prediction accuracy of about 70%. Moving average price has more significant with respect to volume and interest. They reported that the volume and position change has a limited role in price forecasting.

Park et al. (2010) describes the procedure to analyze financial data using multiple regression model. They used MRDDV model that emerges quantitative as well as qualitative variables to estimate the behavior of data. S&P closing price and XHB daily closing price are used as quantitative variables and consumer sentiments and housing construction data are used as qualitative variables. The model indicates the highs of y , the model shows lows in the stock data in both cases. However the lac of their study to include only one quantitative and one qualitative variable instead of using several quantitative variables for more effective analysis.

Unlike regression models, Autoregressive Integrated Moving Average (ARIMA) techniques, introduced by Box and Jenkins in 1970, are considered as effective for short-term prediction, such as financial time series forecasting. Following are some the selected studies.

Ariyo et al. (2014) introduced a short-term stock prediction system using ARIMA model using stock-related data from New York Stock Exchange (NYSE) and Nigeria Stock Exchange (NSE). The model achieved results with respect to comparing methods.

Devi et al. (2013) developed a stock trend forecasting system using ARIMA model by highlighting seasonal trend and flow. The model produces efficient results in terms of effective time series analysis for stock trend prediction with respect to baseline studies.

In their work on Indian stock prediction, Mondal et al. (2014) performed experiments on 56 Indian stocks using ARIMA model. The proposed model is evaluated using Akaike information criterion (AIC). Experimental result show effectiveness of the model in terms of improved prediction accuracy over previous different periods of data.

2.2 Sentiment-based approaches

In these approaches, the stock prediction is performed on the basis of user-generated reviews about a particular stock, posted on different online forums and review sites. User reviews about a particular stock are extracted, pre-processed and classified using supervised and unsupervised learning approaches. In the following paragraphs, some of the prior works performed on sentiment-based approaches are presented.

Rao and Srivastava (2012), analyzed the relationship between tweets analogy such as bullishness, volume, agreement with the financial market indicators like volatility, trading, and stock prices, and achieved a stable correlation. On daily basis, the sum of positive and negative tweets are calculated using Naïve Bayesian classifier. For forecasting, they incorporated an Expert Model Mining System (EMMS) with R square and mean absolute percentage error. The prediction accuracy increases with the increase in time windows and vice versa. However, they conducted the experiments by capturing sentiments of a particular index or company instead of multi-company indices.

Qasem et al. (2015) used a machine learning techniques to classify twitter sentiments. Stock sentiments are collected mainly from Twitter, Google, Facebook, and Tesla. After pre-processing, the neural network and logistic regression model are trained for further analysis. Results show that uni-gram-based TF-IDF outperforms bi-gram-based TF-IDF with an accuracy of 58%. The limitation of their work includes not using the clustering techniques.

Nasseri et al. (2015) proposed an intelligent trading support system using text mining techniques and reported that stock-related microblogs affect the stock prices. The stock-related sentiments are collected from “StockTwits” microblogging site. After performing feature selection for extracting relevant terms in a tweet, decision tree model is applied to identify the decision trend of important terms. Results show that user’s sentiments act as a valuable resource for stock trading decision. The limitation of their work includes the limited length of N-grams for extracting tweets.

Enke et al. (2011) proposed a three-layer stock market prediction system based on regression analysis. The first layer chooses only those variables which have a positive and significant relationship with the target. After selecting the appropriate parameters, next stage is to apply type-2 fuzzy clustering technique on these parameters for constructing a cluster of related data and to extract fuzzy rules from such clusters. Finally, the extracted fuzzy rules are passed through a neural network for effective prediction of the stock market behavior.

In the classification of Indian stock market data, Soni and Shrivastava (2010) applied three supervised machine learning algorithms: classification and regression tree (CART), LDA and QDA, which provides a comprehensive mode of analysis of stock market data in form of a binary tree, linear surface, and quadratic surface. The performance of the three machine learning algorithms is evaluated in terms of misclassification and correct classification rate. The misclassification rate of CART algorithm is 56.11%, which is smaller than the other two machine learning algorithms, indicating a better classification performance. Table 1 gives an overview of

Table 1 Overview of selected studies

Study	Methods	Datasets	Results	Limitations
Ladan et al. (2014)	Multiple regression R square	Datasets: CBN Statistical Bulletin (2010) Parameters: (Inflation rate, exchange rate, interest rate)	54% Accuracy	They used only limited number of macroeconomic variables, which if increased and revised, can produce better results
Kamley et al. (2013)	Multiple regression R square	Dataset: Yahoo finance, Infosys Parameters: (Open, High and Close)	54% Accuracy	To obtain more promising results stock market data needs extended set of parameters with enhanced level of data normalization
Reddy (2010)	Regression-based EMMS model	Datasets: DJIA, NASDAQ, Twitter Parameters: (Returns, Volatility, Trading Volume, Index close price)	91% Accuracy	Capturing sentiments of a particular index or company instead of multi-company indices
Javaid (2010)	Multiple regression analysis	Datasets: Karachi stock exchange 30-Index Parameters: (KIBOR, dividend, earning per share, gross domestic product, and inflation)	62% Accuracy	Prediction accuracy can be improved by introducing revised set of predictors
Devi et al. (2013)	Neural network, Logistic regression	Twitter, Facebook	58% Accuracy	Incorporation of clustering techniques can improve the results
Mondal et al. (2014)	Decision tree	StockTwits	Better results	Limited length of N-grams for extracting tweets
Rao and Srivastava (2012)	Multiple Regression, Type 2 Fuzzy clustering, Neural network	Economic variables	Better prediction result	Further refinements can be made by supplying the Differential Evolution optimization to the Fuzzy Neural Network inference system
Yuan and Luo (2014)	Decision tree	Wind information terminal	70% Accuracy	Volume and position change have a limited role during price forecasting
Qasem et al. (2015)	SVM, correlation	NASDAQ, S&P and DJIA	62.23 Accuracy	Lack of short-sale mechanism that can maximize the profit even when the market is bullish

the selected studies conducted on stock prediction. The proposed system is shown in Fig. 2.

3 Methods

The proposed system implements a multiple regression technique using a revised set of stock predictors supported by computation and preprocessing steps for the prediction of a stock market trend in the form user-friendly recommendation interface to end users.

3.1 Data collection

We extracted three different historical local stock quotes from a financial website (ksestocks.com/QuotationsData), spanning over the duration of 2 years (01/01/2014 to 01/12/2015) and having the different attributes like stock closing, stock opening, high, low, close and volume. Each stock is represented by a stock symbol like KSE Stand for Karachi stock exchange, LUCK represents Lucky Cement Stock (Table 2). In this work, we used Karachi stock exchange as a data collection resource, as it holds a bulk of historical data relevant to local stock. Each data set consist of 24 months historical stock related instances. Table 3 shows the detail of KSE 100 index stock quotes, where the “Symbol” column represents a specific company or stock equity, such as KSE stand for Karachi stock exchange, “Date” column shows the time and day of stock transaction. The “Open”, “High” and “Low” columns indicate the starting value, peak value and lowest value of an equity for a specified date, and “close” represents the final worth of an equity at the time of closing stock market. Similarly “Volume” column shows the number of stock traded on a specific day.

To acquire historical data from KSE stock/quotation, following steps are performed: (i) open browser and write the desired address in the address bar, (ii) select a specific script from a date to date and MS excel format option, (iii) click the prepare summary file button, and (iv) after selecting all of the appropriate options mentioned above, click on the start download image to successfully download a specific dataset in excel format.

3.2 Computation and pre-processing of stock market indicators

Proper selection, computation and preprocessing (Appendix E, F, G of Supplementary Material) of stock market indicators assist in developing the quality input data and improving the efficiency the stock prediction process. Computation and pre-processing of Stock Market Indicators aim at the preparation of acquired dataset for subsequent processing to achieve the results efficiently (Zhang et al. 2013). The basic objective is to derive relevant information from existing attributes to be used as an input parameter to train the multiple regression model. In this work, we propose to use three predictor variables, namely: *Return*, *Volatility*, *ROI (Return on Investment)* and *Volume*; and one predicted variable *Return*,

Fig. 2 Proposed system

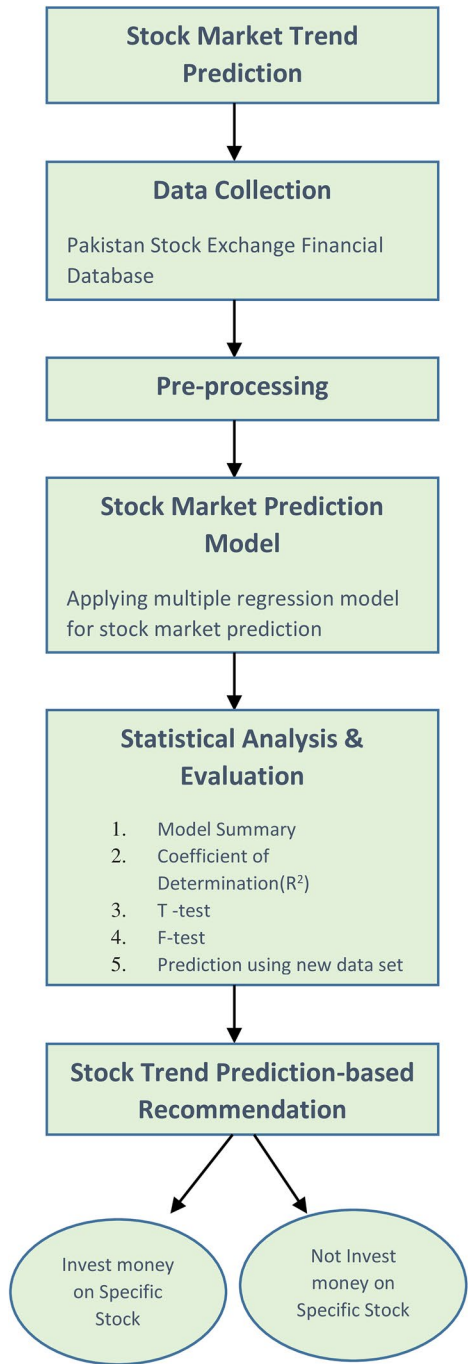


Table 2 Datasets used

Dataset	Title	Description	URL
Dataset#1	KSE 100 index	This dataset contains Karachi stock exchange 100 company stock indices (Appendix B of Supplementary Material)	www.ksestocks.com/QuotationData
Dataset#2	Lucky Cement	This dataset contains data related to Lucky Cement company stock (Appendix C of Supplementary Material)	www.ksestocks.com/QuotationData
Dataset#3	Engro Fertilizer Limited	Contains stock data related to Engro Fertilizer Limited. (Appendix D of Supplementary Material)	www.ksestocks.com/QuotationData

Table 3 Sample KSE 100-INDEX stock quotes

Symbol	Date	Open	High	Low	Close	Volume
KSE-100	1-Jan-14	25,261.66	25,628.41	25,261.66	25,608.85	198,346,580
KSE-100	2-Jan-14	25,615.12	25,979.06	25,615.12	25,962.18	257,114,700
KSE-100	3-Jan-14	25,968.76	26,136.28	25,968.76	26,046.71	170,698,120
KSE-100	6-Jan-14	26,052.69	26,299.69	26,052.69	26,169.83	149,219,040
KSE-100	7-Jan-14	26,172.73	26,289.33	26,051.19	26,259.57	200,616,520
KSE-100	8-Jan-14	26,265.01	26,432.02	26,265.01	26,341.08	188,892,540
KSE-100	9-Jan-14	26,343.68	26,566.73	26,305.51	26,373.24	250,510,720

to effectively predict future stock market trend regarding stock securities. The aforementioned variables have been selected due to their efficacy reported in the recent studies (Ladan et al. 2014; Kamley et al. 2013; Javaid 2010). Later, the effectiveness of the selected variables is verified by applying t test (see Sect. 4). For this purpose, following computations and preprocessing steps are applied (Algorithm 1) on input data to reduce data sparseness from the collected data to obtain stock indicators (predictor and predicted variables) with a minimum level of noise for subsequent processing (Simon and Raoot 2012). Let us take the brief description of the aforementioned indicators.

3.2.1 Computation of stock change

The Stock change computation is performed to identify stock return or gain. Stock change can be computed (Eq. 1) by subtracting the previous stock closing price from the current days close and dividing it by previous day close (Rao and Srivastava 2012).

$$\text{Stock_change} = (\text{current day stock closing price} - \text{previous say stock closing price}) / (\text{previous say stock closing price}) \quad (1)$$

Stock change computation of 2 January 2014 is illustrated with the following example, using Eq. 1 $\text{Stock_change} = (321.67 - 310.6)/310.6 = 0.035641$. A sample list of entries for stock change computations is presented in Table 4.

Table 4 Stock change computation

Symbol	Date	Open	High	Low	Close	Volume	Change
LUCK	1-Jan-14	301.9	312	301.88	310.6	275,100	
LUCK	2-Jan-14	311.99	322.95	311	321.67	643,800	0.035641
LUCK	3-Jan-14	323.5	325.99	316.35	317.6	218,700	-0.01265
LUCK	6-Jan-14	317.99	321.49	315	316.43	122,100	-0.00368
LUCK	7-Jan-14	316.5	317.45	309.53	314.79	249,300	-0.00518
LUCK	8-Jan-14	314	317.8	311	312.82	478,800	-0.00626

Table 5 Stock gain computation

Symbol	Date	Open	High	Low	Close	Volume	Change	Return
LUCK	1-Jan-14	301.9	312	301.88	310.6	275,100		
LUCK	2-Jan-14	311.99	322.95	311	321.67	643,800	0.035641	0.0356406
LUCK	3-Jan-14	323.5	325.99	316.35	317.6	218,700	-0.01265	0
LUCK	6-Jan-14	317.99	321.49	315	316.43	122,100	-0.00368	0
LUCK	7-Jan-14	316.5	317.45	309.53	314.79	249,300	-0.00518	0
LUCK	8-Jan-14	314	317.8	311	312.82	478,800	-0.00626	0

Table 6 Stock volatility computation

Symbol	Date	Open	High	Low	Close	Volume	Change	Return	Volatility
LUCK	1-Jan-14	301.9	312	301.88	310.6	275,100			
LUCK	2-Jan-14	311.99	322.95	311	321.67	643,800	0.035641	0.03564	0.316325
LUCK	3-Jan-14	323.5	325.99	316.35	317.6	218,700	-0.01265	0	0.276532
LUCK	6-Jan-14	317.99	321.49	315	316.43	122,100	-0.00368	0	0.27513
LUCK	7-Jan-14	316.5	317.45	309.53	314.79	249,300	-0.00518	0	0.27881
LUCK	8-Jan-14	314	317.8	311	312.82	478,800	-0.00626	0	0.278975

3.2.2 Computation of stock gain/return

The stock gain or return is the profit obtained by the investors, after purchasing some stock security to do business in the stock market. It is the dependent variable, whose value is to be predicted on the basis of other predictors used in the multiple regression model. If the value of “change” (Table 5) is greater than zero, then its value is assigned to “Return” column using computations performed in MS-Excel software.

3.2.3 Determination of dispersion/volatility

The stock volatility is an important indicator that acts as a predictor variable in the proposed regression model to predict stock return, showing how much variability exists in the stock return. Normally, Volatility measures the dispersion of returns and can be calculated by using the Standard Deviation or variance of the stock change of the same security multiplied by square root of 360 (360 days in a year). If the volatility is high ($\geq 50\%$), it indicates that the stock is riskier (Rao and Srivastava 2012). A sample list of entries for stock volatility computations is presented in Table 6.

3.2.4 Stock volume

The stock volume is considered as a strong indicator to effectively forecast the stock market return, enabling the stock customers to make an investment or not (Shen et al. 2012). Therefore, “stock volume” as a predictor variable, plays a pivotal

role in making the about decision stock investment. The stock-volume refers to the total number of shares of a specific stock traded on a stock exchange. In the stock exchange, the term “high volume” means bullish and “low volume” represent bearish. A sample list of entries for stock volume is presented in Table 7.

3.2.5 Volume reduction by computation of average

As proposed by the Simon and Raoot (2012), volume reduction can improve the efficiency of the prediction model, therefore, we applied volume reduction by taking the monthly average of each independent and dependent variable (Eqs. 2–6).

In this study, we use a daily stock quotation to predict the future stock fluctuation based on different parameters. These daily quotations become a bulk of data with sparse nature of high and low values. To minimize such sparseness of data, we take the monthly average of each parameter. Due to such inconsistencies in data, prediction accuracy is decreased, as low-quality data will generate low quality mining results (Zhang et al. 2013). To increase the efficiency of the stock prediction model, data sparseness issue in terms of variant nature of data in volume, ROI, return and volatility, needs to be addressed (Nasseri et al. 2015). For this purpose, we propose to address the aforementioned issue by taking average of each parameter (volume, ROI, return, and volatility) as shown in Eqs. 2–6. It is observed experimentally (see Sect. 4) that the average values are less prone to sparseness and increase the accuracy of the result. The formula to calculate monthly average is given as follows:

$$\text{monthly}_{average} = \text{sum_of_values} / \text{total_value} \quad (2)$$

$$\text{monthly_volatility}_{average} = \text{sum_of_values} / \text{total_value} \quad (3)$$

$$\text{monthly_volume}_{average} = \text{sum_of_values} / \text{total_value} \quad (4)$$

$$\text{monthly_ROI}_{average} = \text{sum_of_values} / \text{total_value} \quad (5)$$

$$\text{monthly_return}_{average} = \text{sum_of_values} / \text{total_value} \quad (6)$$

where the *sum_of_values* represents the summation of daily stock volume, stock return, ROI and stock volatility; and *total_values* shows the number of days in a month on which stock market transactions occur. For example, using Eq. 3 and its

Table 7 Sample LUCK stock volume listing

Symbol	Date	Open	High	Low	Close	Volume	Change	Return	Volatility
LUCK	1-Jan-14	301.9	312	301.88	310.6	275,100			
LUCK	2-Jan-14	311.99	322.95	311	321.67	643,800	0.0356	0.03564	0.3163
LUCK	3-Jan-14	323.5	325.99	316.35	317.6	218,700	-0.012	0	0.2765
LUCK	6-Jan-14	317.99	321.49	315	316.43	122100	-0.003	0	0.275
LUCK	7-Jan-14	316.5	317.45	309.53	314.79	249,300	-0.005	0	0.278
LUCK	8-Jan-14	314	317.8	311	312.82	478,800	-0.006	0	0.2789

corresponding data values of the excel columns, we compute monthly-volatility average as follows:

$$\text{monthly_volatility}_{\text{average}} = \text{sum}(H249 : H270)/22 = 0.270753017$$

A sample list of entries of average stock volatility, volume, return and ROI computations, taken from the KSE dataset are presented as follows (Table 8).

3.2.6 Return-on-investment (ROI)

The last parameter “*Return-on-Investment (ROI)*”, shows the profit gained by the investor with respect to traded stock (Kamley et al. 2013). It plays an important role in predicting the stock market trend, and can be computed by applying the following formula:

$$\text{ROI} = \text{average_monthly_return}/\text{average_monthly_volume} \quad (7)$$

where *average_monthly_return* shows the profit and *average_monthly_volume* represents the quantity of stock traded on a stock market. For example, if the *average_monthly_return*=0.007393368, and the *average_monthly_volume*=426,763.6364,

Table 8 Stock average volatility, volume, return and ROI computations

	Monthly return	Monthly volatility	Monthly volume	Monthly ROI
Jan-14	0.007393368	0.270753017	426,763.6364	1.73243E-08
Feb-14	0.005782105	0.245911877	357,247.3684	1.61852E-08
Mar-14	0.003800203	0.375555151	787,052.381	4.8284E-09
Apr-14	0.011094247	0.27815434	781,654.5455	1.41933E-08
May-14	0.005391991	0.314302113	477,109.5238	1.13014E-08
Jun-14	0.009571642	0.322219738	743,285.7143	1.28775E-08
Jul-14	0.004722581	0.372167292	721,826.3158	6.54255E-09
Aug-14	0.00727016	0.532336338	945,652.6316	7.68798E-09
Sep-14	0.010893525	0.32055854	715,072.7273	1.52341E-08
Oct-14	0.007446226	0.391723309	660,695	1.12703E-08
Nov-14	0.012286202	0.358565252	1,016,388.889	1.20881E-08
Dec-14	0.008647905	0.306632829	856,268.1818	1.00995E-08
Jan-15	0.006693448	0.284479878	774,845.4545	8.63843E-09
Feb-15	0.003353103	0.286502541	993,036.8421	3.37662E-09
Mar-15	0.006826056	0.450511578	827,995.2381	8.24408E-09
Apr-15	0.008961945	0.2956309	921,440.9091	9.72601E-09
May-15	0.002660617	0.304989628	505,640	5.26188E-09
Jun-15	0.009426019	0.341,118,816	899,845.4545	1.04752E-08
Jul-15	0.009810463	0.290,316,137	698,530	1.40444E-08
Aug-15	0.004691111	0.315966883	773,760	6.06275E-09
Sep-15	0.00477349	0.202594749	324,510	1.47098E-08
Oct-15	0.004008802	0.137766358	278,247.619	1.44073E-08
Nov-15	0.001253908	0.220166684	206,447.619	6.07374E-09
Dec-15	0.004798248	0.152533716	277,528.5714	1.72892E-08

then applying Eq. 7, we compute monthly ROI as follows: $ROI = 0.007393368/426763.6364 = 1.73243E - 08$. A sample list of entries of ROI computations is presented in Table 9.

3.2.7 Missing values

To handle data inconsistency for improving the efficiency of the prediction model, it is suggested that the missing values can be filled for data cleaning (inconsistency removal) (Feinberg and Genethliou 2005). There are three ways to fill in missing values: (i) by taking the average and put it to the missing place, (ii) to randomly select the entry and fill the missing value, and (iii) fill in the missing value with one step previous value. We have chosen the third option and therefore, the data cleaning is applied by manually filling the missing values with the relevant upper value. The missing values are mainly generated due to dividing by zero entry during the computation of dependent and independent variables at monthly basis.

Table 9 Sample computations for LUCK stock return on investment (ROI)

	Monthly return	Monthly volatility	Monthly volume	Monthly ROI
Jan-14	0.007393368	0.270753017	426,763.6364	1.73243E-08
Feb-14	0.005782105	0.245911877	357,247.3684	1.61852E-08
Mar-14	0.003800203	0.375555151	787,052.381	4.8284E-09
Apr-14	0.011094247	0.27815434	781,654.5455	1.41933E-08
May-14	0.005391991	0.314302113	477,109.5238	1.13014E-08
Jun-14	0.009571642	0.322219738	743,285.7143	1.28775E-08
Jul-14	0.004722581	0.372167292	721,826.3158	6.54255E-09
Aug-14	0.00727016	0.532336338	945,652.6316	7.68798E-09
Sep-14	0.010893525	0.32055854	715,072.7273	1.52341E-08
Oct-14	0.007446226	0.391723309	660,695	1.12703E-08
Nov-14	0.012286202	0.358565252	1,016,388.889	1.20881E-08
Dec-14	0.008647905	0.306632829	856,268.1818	1.00995E-08
Jan-15	0.006693448	0.284479878	774,845.4545	8.63843E-09
Feb-15	0.003353103	0.286502541	993,036.8421	3.37662E-09
Mar-15	0.006826056	0.450511578	827,995.2381	8.24408E-09
Apr-15	0.008961945	0.2956309	921,440.9091	9.72601E-09
May-15	0.002660617	0.304989628	505,640	5.26188E-09
Jun-15	0.009426019	0.341118816	899,845.4545	1.04752E-08
Jul-15	0.009810463	0.290316137	698,530	1.40444E-08
Aug-15	0.004691111	0.315966883	773,760	6.06275E-09
Sep-15	0.00477349	0.202594749	324,510	1.47098E-08
Oct-15	0.004008802	0.137766358	278,247.619	1.44073E-08
Nov-15	0.001253908	0.220166684	0.619	6.07374E-09
Dec-15	0.004798248	0.152533716	277,528.5714	1.72892E-08

Algorithm 1 Computations and Pre-Processing of Stock Indicators

Step1. Compute the change by subtracting the previous day stock closing from current day close and divide the same by previous day close using Eq. 1

Step2. Compute gain or return as described in section “*Determination of Dispersion/ volatility*”

Step3. Determine the dispersion/Volatility by using the standard deviation of calculated change as described in section “*Determination of Dispersion/ volatility*”

Step4. Apply data cleaning using fill-in missing value technique.

Step5. Apply Volume Reduction: by computing the average Return, Volatility and Volume for each month using Eq. 2 to Eq. 6

Step6. Calculate “Return on Investment” by dividing average monthly return to average monthly volume using Eq. 7

Step6. Repeat all above mentioned steps for each and every dataset preparation.

Step7. Finish

3.3 Stock market analysis/predictions system

The stock market analysis/prediction system aims at analyzing the stock-related financial data by applying some prediction model and presenting information to users/investors to invest or not. An economy with high stock prices is considered as rising economy. The prediction system assists in decision making process regarding the purchase and selling of shares. Prediction aims at forecasting a future event (response variable), based on different parameters (independent variables).

In the stock market environment, stock return can be predicted by using the historical stock quotations: *stock return*, *stock volatility*, *stock ROI*, and *stock volume values*. The basic idea is to find a linear combination of *stock volatility*, *ROI* and *stock volume* that best predicts the *stock return*. It assist the investors in making financial decisions whether to invest or not in gaining maximum profit. Multiple Regression analysis can be used to infer causal relationships between the independent and dependent variables (Ladan et al. 2014). The formal notation of multiple regression model with n variables is presented as follows:

$$Y_i = \beta_1 + \beta_2 X_2 + \beta_3 X_3 \dots \beta_n X_n + \mu \quad (8)$$

where Y is the dependent variable, X_2 and X_3 up to X_n are the independent variables, μ is the residual term. In Eq. 8, β_1 : It is the intercept term which gives a mean or average effect on Y_i of all the variables: X_2 and X_3 up to X_n . β_2, β_3 up to β_n : these are called the regression coefficients or slope, β_2 measures the change in the mean value of Y with respect to per unit change in X_2 , such that the values of X_3 and X_4 remain constant. Similarly β_3 measures the change in the mean value of Y with respect to per unit change in X_3 , such that the values of X_2 and X_4 remain constant; and β_4 measures the change in the mean value of Y with respect to per unit change in X_4 , such that values of X_2 and X_3 remain constant (Zhang et al. 2013).

As the proposed stock prediction model is based on four variables, therefore, the general regression model (Eq. 8) can be formulated as follows:

$$Y_i = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \mu \quad (9)$$

where β_1 is the intercept, β_2 , β_3 , and β_4 are regression coefficients, and can be estimated using the least squares procedure (Zhang et al. 2013; Simon and Raoot 2012) to minimize the sum of squares of errors. Therefore, using Eq. 9, the value of β_1 can be computed as follows:

$$\beta_1 = \bar{Y} - \beta_2 \bar{X}_2 - \beta_3 \bar{X}_3 - \beta_4 \bar{X}_4 \quad (10)$$

Similarly, using Eq. 9, values of regression coefficients or slope: β_2 , β_3 and β_4 , can be derived as follows (Zhang et al. 2013):

$$\beta_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad (11)$$

$$\beta_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad (12)$$

$$\beta_4 = \frac{(\sum y_i x_{4i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad (13)$$

The proposed model uses: (i) three independent/predictors (volatility, volume and ROI), (ii) a dependent/predicted variable (return), (iii) β_1 (intercept), and (iv) three regression coefficients/slopes: β_2 , β_3 and β_4 . Using aforementioned setup, the Eq. 9 of our proposed regression model can be calibrated as follows:

$$\text{return} = \beta_1 + \beta_2 \times \text{volatility} + \beta_3 \times \text{volume} + \beta_4 \times \text{ROI} \quad (14)$$

Computing the regression coefficients: To compute the Regression coefficients, intercept line and other statistics, `lm()` function of R software (Nasseri et al. 2015) is applied on the KSE stock dataset, a partial listing is presented in Table 10. The implementation of `lm()` function in R is shown as follows:

$$\text{fit} <- \text{lm}(\text{Return} \sim \text{Volatility} + \text{Volume} + \text{ROI}, \text{data} = \text{tdata}) \quad (15)$$

where `lm()` is function; *Return* is predicted variable; *Volume*, *volatility* and *ROI* are the predictor variables; *tdata* is the training dataset; and `fit` the object which holds the result/summary (for detailed summary, see Sect. 4).

Intercept calculation: As mentioned earlier, the `fit` object contains a detailed summary of results obtained from the proposed prediction model, in which intercept value (β_1) is computed (Eq. 10), as follows:

$$\beta_1 = -3.378e^{-03} \quad (16)$$

Table 10 Monthly return, monthly volatility, monthly volume and monthly ROI quotation data

	Monthly return (Y_i) predicted variable	Monthly volatility (X2) predictor variable	Monthly volume (X3) predictor variable	Monthly ROI (X4) predictor variable
Jan-14	0.003322653	0.13218729	187,123,297.3	1.77565E-11
Feb-14	0.002399337	0.155792739	138,508,623.3	1.73227E-11
Mar-14	0.00452128	0.15949703	153,432,341.9	2.94676E-11
Apr-14	0.005423462	0.151713116	205,974,448.6	2.63307E-11
May-14	0.003065163	0.126045862	112,489,844.3	2.72484E-11
Jun-14	0.002463041	0.118697467	137,133,061	1.7961E-11
Jul-14	0.002228261	0.225426069	81,485,430.53	2.73455E-11
Aug-14	0.0038456	0.226194175	96,293,972.63	3.9936E-11
Sep-14	0.003590472	0.114401662	89,490,967.27	4.01211E-11
Oct-14	0.002670452	0.119953217	94,347,540.5	2.83044E-11
Nov-14	0.004288527	0.157518603	157,219,688.9	2.72773E-11
Dec-14	0.003882371	0.129336533	169,890,952.3	2.28521E-11

similarly, other regression coefficients: $\beta_2, \beta_3, \beta_4$ are also computed by the $lm()$ function of R (Eqs. 11–13) as follows:

$$\beta_2 = 8.439e^{-04} \tag{17}$$

$$\beta_3 = 2.638e^{-11} \tag{18}$$

$$\beta_4 = 1.177e^{+08} \tag{19}$$

Finally, putting the values of $\beta_1, \beta_2, \beta_3,$ and $\beta_4.$ in Eq. 14, we get the trained regression model as follows:

$$return = -3.378e^{-03} + 8.439e^{-04} \times volatility + 2.638e^{-11} \times volume + 1.177e^{+08} \times ROI \tag{20}$$

The Eq. 20 is a trained regression model to predict the stock return by using volatility, volume and ROI. To calculate return, we put the values of the slope ($-3.378e^{-03}$) and regression coefficients ($8.439e^{-04}, 2.638e^{-11},$ and $1.177e^{+08}$) along with the values of test variables: volume, ROI, and volatility. In this way, the stock return of the next month can be predicted using Eq. 20 and applying predict() function of R software, as follows:

We can predict the next month stock return, using Eq. 20, by applying the R predict function as follows:

$$predict(fit, newdata) \tag{21}$$

where fit is an object holding the values of trained model computed using Eq. 15 and newdata is a test data parameter computed as: newdata=data.frame (volatility=0.226194175, volume=96,293,972.63, ROI=3.99E-11), where volatility, volume and ROI are the test/predictor variables. After finding the values of fit and newdata, we get: predict(fit, newdata)=0.0040481.

In Table 10, the monthly return (0.0038456) for Aug-14 (highlighted in the red color font) shows that the prediction made by our proposed model ($\text{predict}(\text{fit}, \text{newdata}) = 0.0040481$) is quite accurate.

3.4 Proposed algorithm and implementation

The complete algorithm of the proposed model is shown as follows:

Algorithm 2 Stock Market Trend Prediction using R

Input:	Read csv file of Stock Historical data
Output:	Model Regression equation to predict stock
Begin	
1. Get stock historical data from a financial site of www.ksestocks.com/QuotationData .	
2. Call algorithm 1 to preprocess the extracted data	
3. Save the preprocessed dataset with csv extension	
4. Read the .csv file using R function, read.csv (file. Choose(), header=T)	
5. Display the file contents using R interface	
6. Apply the $\text{lm}()$, R function to compute the Regression coefficients, intercept line and other statistics,	
7. $\text{fit} <- \text{lm}(\text{Return} \sim \text{Volatility} + \text{Volume} + \text{ROI}, \text{data}=\text{data1})$	
8. To display the summary of the proposed model by calling R, $\text{summary}()$, function that contains all the model statistics.	
9. Using Eq. 14, as a model equation to predict future stock trend	
10. Apply Predict function Of R software to predict stock return, using Eq. 21	
11. Call recommendation()	
End	
Function Recommendation(<i>fit, Volatility, Volume and ROI, Return[]</i>)	
Begin	
1. <i>Predicted_Val</i> <-- Apply predict function by using values of <i>fit, Volatility, Volume and ROI</i> .	
2. <i>counter</i> =0	
3. <i>Total_stock_instances</i> =0	
4. # <i>n</i> is number of instances	
5. for <i>i</i> =1 to <i>n</i>	
6. if (<i>Predicted_Val</i> > <i>Return[i]</i>)	
7. <i>counter</i> = <i>counter</i> +1.	
8. end if	
9. end for	
10. <i>threshold-value</i> =70% of <i>n</i>	
11. if (<i>counter</i> > <i>threshold-value</i>)	
12. print "Invest Money on Stock"	
13. else	
14. print "Not Invest Money on Stock"	
15. end if	
16. Return()	
End function	

The proposed stock prediction model is implemented using R software (Kabacoff 2015), a widely used platform for data analysis and visualization purposes. R contains a bundle of packages and libraries for statistical analysis, data extraction, and graphics. It is an open source software, and freely available from the Comprehensive R Archive Network (CRAN) at (<http://cran.r-project.org>). To perform experiments, we randomly split the dataset into training (80%) and testing (20%), this split ratio of training and testing data assists in achieving promising results.

The implementation procedure of experiments conducted on three datasets is presented as follows (step#1 to step#9): (1) *Data collection*: The required stock data is extracted using the first step of the algorithm 2, (2) *Pre-processing*: the data acquired at step#(1) is pre-processed by applying the second step of the algorithm 2, which calls the pre-processing algorithm (algorithm 1), (3) *File saving*: the preprocessed data obtained at step#(2) is saved as.csv(comma separated value) file, which is subsequently made input to the R software, (4) *File reading*: to read.csv file, we used “read.csv (file. Choose(), header=T)”, option of the R software, as shown in the Fig. 3, after applying the read.csv() function, contents of the input dataset are automatically stored in the data1 object to be used in subsequent analysis, (5) *Content display*: To display contents of the data1 object (read file), we write data1 and press enter; Resultantly, one can visualize contents of all of the variables (predictors and predicted), as shown in Fig. 4, (6) *Model training*: In this step, multiple regression model is trained by applying lmo() function on the input data (tdata object) along with required parameters, the model’s details are stored in the fit object, as shown in Fig. 5, (7) *Model statistics*: After training the multiple regression model (step# 6), the next task is to show the detailed summary of the model using summary() function. Figure 6 shows different parameters of the implemented model, (8) *Regression model generation*: Using the summary details generated in step# 7, we can build the model regression (using Eq. 14), by incorporating different parameters like slope/intercept, regression coefficients as follows:

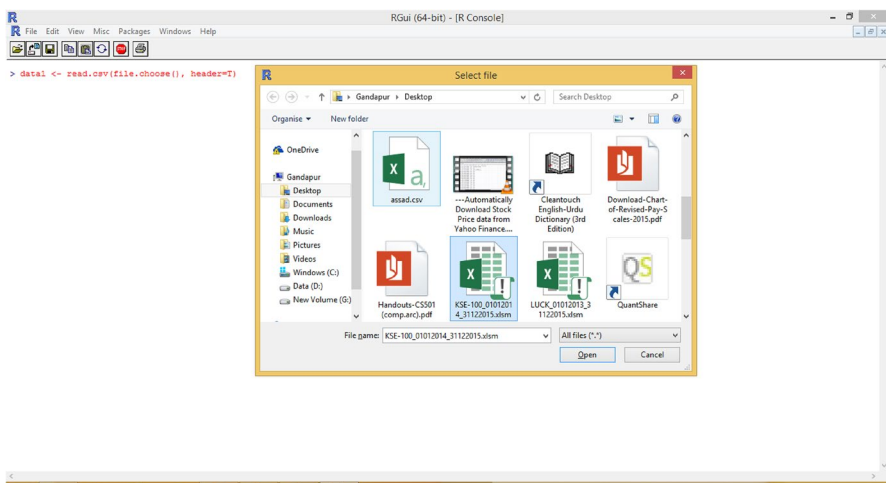
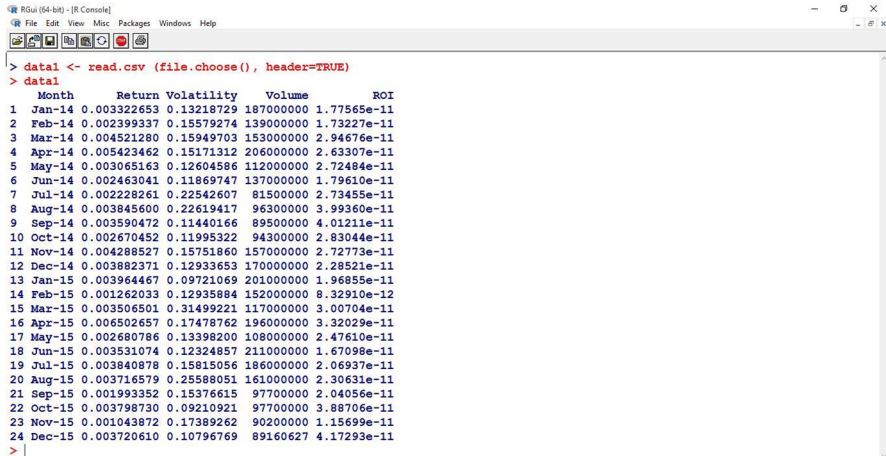


Fig. 3 Read dataset using R software

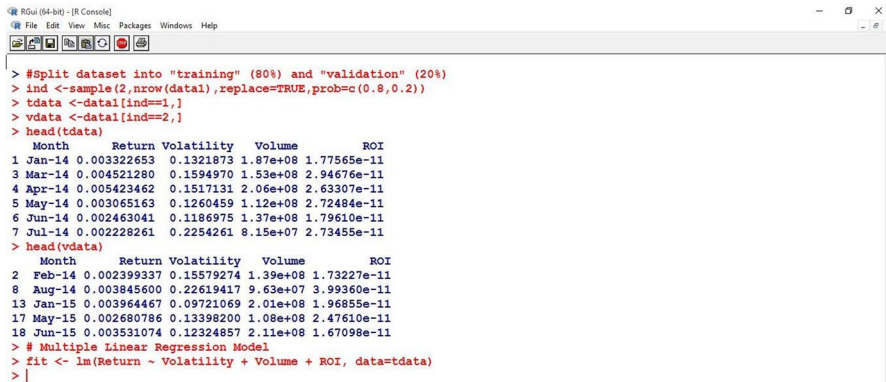


```

RStudio (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
> data1 <- read.csv (file.choose(), header=TRUE)
> data1
  Month      Return Volatility      Volume      ROI
1 Jan-14 0.003322653 0.13218729 187000000 1.77565e-11
2 Feb-14 0.002399337 0.15579274 139000000 1.73227e-11
3 Mar-14 0.004521280 0.15949703 153000000 2.94676e-11
4 Apr-14 0.005423462 0.15171312 206000000 2.63307e-11
5 May-14 0.003065163 0.12604586 112000000 2.72484e-11
6 Jun-14 0.002463041 0.11869747 137000000 1.79610e-11
7 Jul-14 0.002228261 0.22542607 815000000 2.73455e-11
8 Aug-14 0.003845600 0.22619417 963000000 3.99360e-11
9 Sep-14 0.003590472 0.11440166 895000000 4.01211e-11
10 Oct-14 0.002670452 0.11995322 943000000 2.83044e-11
11 Nov-14 0.004288527 0.15751860 157000000 2.72773e-11
12 Dec-14 0.003882371 0.12933653 170000000 2.28521e-11
13 Jan-15 0.003964467 0.09721069 201000000 1.96855e-11
14 Feb-15 0.001262033 0.12935884 152000000 8.32910e-12
15 Mar-15 0.003506501 0.31499221 117000000 3.00704e-11
16 Apr-15 0.006502657 0.17478762 196000000 3.32029e-11
17 May-15 0.002680786 0.13398200 108000000 2.47610e-11
18 Jun-15 0.003531074 0.12324857 211000000 1.67098e-11
19 Jul-15 0.003840878 0.15815056 186000000 2.06937e-11
20 Aug-15 0.003716579 0.25580851 161000000 2.30631e-11
21 Sep-15 0.001993352 0.15376615 977000000 2.04056e-11
22 Oct-15 0.003798730 0.09210921 977000000 3.88706e-11
23 Nov-15 0.001043872 0.17389262 902000000 1.15699e-11
24 Dec-15 0.003720610 0.10796769 89160627 4.17293e-11
> |

```

Fig. 4 Show the read data using R tool



```

RStudio (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
> #Split dataset into "training" (80%) and "validation" (20%)
> ind <- sample(2, nrow(data1), replace=TRUE, prob=c(0.8, 0.2))
> tdata <- data1[ind==1,]
> vdata <- data1[ind==2,]
> head(tdata)
  Month      Return Volatility      Volume      ROI
1 Jan-14 0.003322653 0.1321873 1.87e+08 1.77565e-11
3 Mar-14 0.004521280 0.1594970 1.53e+08 2.94676e-11
4 Apr-14 0.005423462 0.1517131 2.06e+08 2.63307e-11
5 May-14 0.003065163 0.1260459 1.12e+08 2.72484e-11
6 Jun-14 0.002463041 0.1186975 1.37e+08 1.79610e-11
7 Jul-14 0.002228261 0.2254261 8.15e+07 2.73455e-11
> head(vdata)
  Month      Return Volatility      Volume      ROI
2 Feb-14 0.002399337 0.15579274 1.39e+08 1.73227e-11
8 Aug-14 0.003845600 0.22619417 9.63e+07 3.99360e-11
13 Jan-15 0.003964467 0.09721069 2.01e+08 1.96855e-11
17 May-15 0.002680786 0.13398200 1.08e+08 2.47610e-11
18 Jun-15 0.003531074 0.12324857 2.11e+08 1.67098e-11
> # Multiple Linear Regression Model
> fit <- lm(Return ~ Volatility + Volume + ROI, data=tdata)
> |

```

Fig. 5 Apply lm() R function to train the Regression model

$$\text{return} = -3.527e^{-03} + 1.838e^{-03} \times \text{volatility} + 2.61e^{-11} \times \text{volume} + 1.200e^{+08} \times \text{ROI} \quad (22)$$

Equation 22 is the proposed trained model regression equation and is able to predict stock return based on the values of volatility, volume, ROI (return on investment) and others parameters like slope and regression coefficients. Further details of the proposed model are discussed in Sect. 4; (9) *Prediction*: predict() function is applied by using fit and vdata (validation data) as function parameters to test the model performance (Fig. 7); *Call Recommendation() function*: This Recommendation() function, contains different parameters, namely: fit, volatility, volume and ROI, and

```

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

> summary(fit)

Call:
lm(formula = Return ~ Volatility + Volume + ROI, data = tdata)

Residuals:
    Min       1Q   Median       3Q      Max
-0.0004187 -0.0002274 -0.0000112  0.0001585  0.0006043

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.527e-03  4.455e-04  -7.917 9.79e-07 ***
Volatility   1.838e-03  1.344e-03   1.368   0.191
Volume      2.612e-11  1.836e-12  14.232 4.07e-10 ***
ROI         1.200e+08  8.619e+06  13.921 5.54e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0003118 on 15 degrees of freedom
Multiple R-squared:  0.9539,    Adjusted R-squared:  0.9446
F-statistic: 103.4 on 3 and 15 DF,  p-value: 3.031e-10

> |

```

Fig. 6 Model statistics by applying R Summary(), function

```

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

> pred <- predict(fit,vdata)
> head(pred)
      2      8     13     17     18
0.002468911 0.004196281 0.004264366 0.002511532 0.004216405
> head(vdata)
  Month Return Volatility Volume ROI
2 Feb-14 0.002399337 0.15579274 1.39e+08 1.73227e-11
8 Aug-14 0.003845600 0.22619417 9.63e+07 3.99360e-11
13 Jan-15 0.003964467 0.09721069 2.01e+08 1.96855e-11
17 May-15 0.002680786 0.13398200 1.08e+08 2.47610e-11
18 Jun-15 0.003531074 0.12324857 2.11e+08 1.67098e-11
> |

```

Fig. 7 Apply predict() function using R software

Recommendation System for Stock Trend Prediction

Enter Volatility: Enter Volume: Enter ROI:

Recommendation: Invest Money on Stock

Fig. 8 Recommendation using R software

Return[], which are passed to the *predict()* function to return a *Predicted_Val*, if the *Predicted_Val* is greater than the actual return values, increment the *counter* variable by one; finally, if the counter value is greater or less than the *threshold-value* (70% of the total instances) then print a recommendation message, i.e. “invest on stock” or “not invest on stock” (see Figs. 8, 9).

Fig. 9 Recommendation using R software

4 Results and discussion

The earlier studies (Ladan et al. 2014; Kamley et al. 2013; Javaid 2010; Ariyo et al. 2014; Devi et al. 2013) used open, high, low, inflation rate, exchange rate, interest rate, returns, volatility, trading volume and index close price. However, they achieved low accuracy, whereas, this work uses different stock-related attributes including stock closing, stock volatility, volume, ROI (return on investment) and stock return to increase the efficiency of the proposed work by improving the accuracy. The mentioned variables have some sort of significant relationship among them. To check the significance of each parameter, T-test is applied (Ahmadifard et al. 2013). Table 11 shows the relative significance of each parameter in all of the three models.

Table 11 gives experiment-wise detail of different predictors, their t-value, and associated p-value. The T-value column (T-statistics) shows the significance of the relationship between dependent and independent variables. In experiment no. 1, stock volatility reflects a positive but insignificant relationship with stock return; stock volume has a positive and significant relationship with the stock return, which means that increase in stock volume will bring an increase in stock return; and ROI

Table 11 Parameters relationship and significance

Experiment #	Dataset used	Parameters used	T-value	P-value	Pr(> t)
1	KSE 100 Index	Stock volatility	1.368	0.191	Insignificant
		Stock volume	14.232	$4.07e^{-10}$	Significant
		Return on investment (ROI)	13.921	$5.54e^{-10}$	Significant
		Stock return			
2	Lucky Cement Stock	Stock volatility	2.873	0.05	Insignificant
		Stock volume	5.824	$3.35e^{-05}$	Significant
		Return on Investment (ROI)	9.356	$1.19e^{-07}$	Significant
		Stock return			
3	Engro Fertilizer Limited	Stock volatility	2.873	0.01	Significant
		Stock volume	4.840	0.0002	Significant
		Return on Investment (ROI)	17.29	$7.62e^{-11}$	Significant
		Stock return			

also has a positive and significant relationship with stock return. Table 11 shows that experiment no. 2 also yields the same result as experiment no. 1. The experiment no. 3 shows that the stock volatility establishes a positive and significant relationship with the stock return. Similarly, stock volume has a positive and significant relationship and finally, the ROI has a positive but significant relationship with stock return. The proposed regression-based prediction model provides an efficient prediction of the stock trend.

The results reported in Table 11 gives following findings:

On KSE 100 Index, there is a significant positive relationship between the response variable “stock return” with predictor variables: (a) ROI and (b) stock volume. Therefore, the model has a strong tendency of predicting the stock trend efficiently.

On Lucky Cement Stock, there is a significant positive relationship between the response variable “stock return” with predictor variables: (a) ROI and (b) stock volume. Therefore, there the model has a strong tendency of predicting the stock trend efficiently.

On Engro Fertilizer Limited, there is a significant positive relationship between the response variable “stock return” with predictor variables: (a) stock volatility (b) Stock Volume and (c) ROI. Therefore, our model shows a strong tendency of predicting the monthly stock trend.

4.1 Coefficient of determination (multiple R-square)

Prediction accuracy shows that how efficient the model is with prediction ability with respect to forecasting the future stock market up and down. To determine the model accuracy, multiple r-square is used, which is also called the coefficient of determination. It shows that how much variation in stock return is explained by predictor variables (Nagar and Hahsler 2012). The Table 12 shows that, in experiment no. 1, we used KSE 100 index as an input dataset to the multiple regression model and obtained 95% prediction accuracy, which indicates that 95% variance in stock return is explained by predictors. Similarly, in experiment no. 2 Lucky Cement Stock dataset is used to train the model by receiving an 89% prediction accuracy. Finally, in experiment no. 3, Engro Fertilizer Limited stock data is used by achieving 97% prediction accuracy.

Value of multiple R-squared in Table 12 is inspected. Using “KSE 100 Index” dataset, the value of multiple R-squared is 0.95, which depicts that the predictor

Table 12 Model accuracy on three datasets

Experiment no.	Dataset used	Prediction accuracy (%)
1	KSE 100 Index	95
2	Lucky Cement Stock	89
3	Engro Fertilizer Limited	97

can measure the variance in stock return up to 95%. While applying our model on “Lucky Cement” dataset, the value of multiple R-squared is 0.89, which depicts that the predictor can measure the variance in stock return up to 89%. Finally, the “Engro Fertilizer Limited” dataset yields a 0.97 value of multiple R-squared, which reflects that our model has the 97% prediction accuracy.

4.2 Evaluating the fitness of proposed model

To evaluate the fitness of proposed regression model, F-test is applied. The F-test evaluates the overall relationship between the target and the set of predictors (Simon and Raoot 2012). Results reported in Table 13 show the F-value along with probability values obtained from the three experiments. The “F-statistics” column of the experiment on “KSE 100 index” dataset reports a value of 103.4, whereas its corresponding “probability value/significance level” is 0.0001, showing that the target/response variable has a significant relationship with the set of predictors. Conclusively, our model is good-fit for the aforementioned dataset. The “F-statistics” column of the experiment on “Lucky Cement” dataset reports a value of 42.95, whereas its corresponding “probability value/significance level” is 0.001, showing that the target/response variable has a significant relationship with the set of predictors. Conclusively, our model is good-fit for the aforementioned dataset. The “F-statistics” column of the experiment on “Engro Fertilizer Limited” dataset reports a value of 191.7, whereas its corresponding “probability value/significance level” is 0.01, showing that the target/response variable has a significant relationship with the set of predictors. Conclusively, our model is good-fit for the aforementioned dataset.

4.3 Proposed model assist individuals and companies to invest or not in the stock market by providing user friendly interface

To assist the investors and companies in taking a decision regarding making investments in the stock market, we provide the user-friendly interface using R libraries and functions. Different textboxes, labels and button controls are created by using R, gWidgets_0.0-54¹, digest_0.6.10² and gWidgetstcltk_0.0-55³ library. Users can enter data: volatility, volume and ROI in the textboxes, after pressing the button, suggestions are displayed, such as “invest money on stock” or “not invest money on stock”, by propagating the data into the fitted model (see Figs. 8 and 9).

Table 13 F-statistics and model fitness

Dataset	F-statistics	Probability value	Model fitness
KSE 100 Index	F-statistic: 80.68 on 3 and 15 DF	p-value: $3.031e^{-10}$ or 0.0001	Good fit
Lucky Cement	F-statistic: 31.39 on 3 and 15 DF	p-value: $1.336e^{-07}$ or 0.001	Good fit
Abbot Company	F-statistic: 14.43 on 3 and 13 DF	p-value: $1.333e^{-11}$ or 0.01	Good fit

4.4 Comparison of proposed model with state-of-the-art methods

Finally, we compare our proposed stock prediction model with the state-of-the-art methods (Ladan et al. 2014; Kamley et al. 2013; Javaid 2010).

In their work on stock prediction, Ladan et al. (Ladan et al. 2014), used multiple regression model to predict stock prices, by taking Inflation rate, exchange rate and interest rate as independent variables and all share index as a dependent variable. The used dataset obtained from CBN Statistical Bulletin (2010), by achieving 54% prediction accuracy, it means that 54% variability in all share index is explained by predictors.

Kamley et al. (2013) used multiple regression model to predict stock prices, by taking open, high and close as independent variables. They used “Infosys Company” dataset obtained from “Yahoo finance” by achieving 89% prediction accuracy.

Javaid (2010) used multiple regression model to predict market shares of companies on different parameters, namely: KIBOR, dividend, earning per share, gross domestic product, and inflation. They used Karachi-Stock-Exchange (KSE) 30-Index dataset and achieved up to 62% prediction accuracy.

As compared to the aforementioned baseline studies, we used four variables including *Stock-return*, *volatility*, *volume*, and *return-on-investment*, which are derived using a computation and pre-processing algorithm (Algorithm 1). The computation and pre-processing algorithm is used to normalize the sparse nature of data for achieving better prediction results. Table 14 shows a comparison of our method with the baseline methods.

We applied different machine learning techniques, such as Decision Tree, Naïve Bayes, Neural Network, SVM, and our model on Apple Inc Dataset listed on NASDAQ (<https://www.nasdaq.com/symbol/aapl/historical>). Table 15 shows that the proposed model has achieved best results in terms of average accuracy as compared to other models.

To elaborate the models type 1 (precision) and type 2 errors (recall), and to compare with other models, we present the commentary as follows: the recommendation module of our system assists in making decision about whether to purchase or sell the stock. To evaluate the robustness of the model, type 1 (precision) and type 2 (recall) errors are elaborated by running the leading models and comparing the results obtained from our proposed system. Following are the formulas for computing precision, recall and accuracy.

$$Precision(p) = \frac{tp}{tp + fp}$$

$$Recall(r) = \frac{tp}{tp + fn}$$

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn}$$

Table 14 Comparison with the state-of-the-art stock prediction methods table

Study	Objective	Data sets	Methods	Parameters	Prediction accuracy (%)
Ladan et al. (2014)	Stock prediction	CBN Statistical Bulletin (2010)	Multiple regression R square	Inflation rate, exchange rate, interest rate	54
Kamley et al. (2013)	Stock prediction	Yahoo finance, Infosys Company	Multiple regression R square	Open, close and high	89
Javaid (2010)	Stock prediction	Karachi stock exchange 30-Index	Multiple regression analysis	KIBOR, dividend, earning per share, gross domestic product, and inflation	62
Proposed	Stock prediction	KSE stock (2014–2015)	Multiple regression analysis, data pre-processing, correlation, model fitness, variant nature	Stock return, volatility, volume, return on investment	97 (Engro) 95 (KSE 100 index) 89% (Lucky Cement)

Table 15 Comparison with the existing stock prediction methods on Apple Inc Dataset listed on NASDAQ

Model	Avg. accuracy (%)
Decision tree	67.481
Naïve Bayes	66.516
Neural network	68.441
SVM	69.643
Proposed (our model)	76.414

Table 16 Comparison with leading models

Study	Model	Accuracy (%)	Precision	Recall
Khadim et al. (2016)	Random forest	83.52	0.82	0.81
Ariyo et al. (2014)	ARIMA	0.90	0.91	0.92
Proposed (our work)	Multiple regression	0.94	0.95	0.93

Table 17 Efficiency of the proposed model across samsung electronics dataset listed on NASDAQ

Trading time frame (months)	Precision	Recall	Accuracy
1	0.831	0.841	0.850
2	0.884	0.862	0.856
3	0.924	0.918	0.896
4	0.947	0.950	0.963
5	0.951	0.934	0.942
6	0.960	0.942	0.955

where tp is the number of true +ive values, fp is the number false +ive values, tn is the number of true -ive values, and fn is the number of false -ive values.

We compute aforementioned metrics using Samsung Electronic (<https://www.nasdaq.com/symbol/ssnlf/historical>) listed on NASDAQ (Qin et al. 2017). The results are listed in Table 16, showing effectiveness of the proposed system with respect to comparing models.

We discuss the generality of your model across the different stock markets. The performance evaluation results shown in Tables 17, 18 and 19 across different datasets (Samsung Electronic (<https://www.nasdaq.com/symbol/ssnlf/historical>), General Electric (<https://www.nasdaq.com/symbol/ge/historical>), and Baba (Alibaba) (<https://www.nasdaq.com/symbol/baba/historical>)) for a trading time frame of 1 to 6 months, show that the proposed system achieved promising results in terms of precision, recall and accuracy.

Table 18 Efficiency of the proposed model across Baba (Alibaba) Dataset listed on NASDAQ

Trading time frame (months)	Precision	Recall	Accuracy
1	0.822	0.835	0.841
2	0.846	0.864	0.866
3	0.870	0.891	0.901
4	0.922	0.940	0.912
5	0.944	0.956	0.947
6	0.951	0.922	0.953

Table 19 Efficiency of the model across general electric dataset listed on NASDAQ

Trading time frame (months)	Precision	Recall	Accuracy
1	0.811	0.832	0.803
2	0.834	0.851	0.823
3	0.854	0.873	0.841
4	0.911	0.934	0.872
5	0.937	0.948	0.894
6	0.958	0.935	0.916

5 Conclusions and future work

This work presents the results of stock prediction by applying a multiple regression model using R software. It explores the stock prediction model using historical dataset. Different experiments are performed on different datasets to evaluate the accuracy of the proposed prediction model and measure the parameter relationship. The proposed technique consists of the following steps: (i) historical dataset about different stocks are obtained from a financial resource in “data acquisition” module, (ii) pre-processing is performed on the acquired dataset, by using “*Computation and Pre-Processing of Stock Market Indicators*” to prepare it for further processing. Various formulas are applied to compute stock change, volatility and return. To minimize data inconsistency, data cleaning is applied by fill-in missing values (handling divide by zero) technique; moreover, volume reduction is performed by computing an average of each parameter. (iii) the next step is “data splitting”, which splits the data into training and validation sets with 80:20 ratio, (iv) the multiple regression model is trained on the training data by using an “*Algorithm 1*”, and then the model performance is checked on the basis of validation data; model summary shows information about the proposed model, such as parameter significance, prediction accuracy, and overall model fitness, the obtained results show that when KSE 100-index dataset is used to train the model, it gives 95% prediction accuracy, and Lucky Cement yields 89% prediction accuracy, and the prediction accuracy of Engro Fertilizer Limited dataset is 97%, and (v) finally, user-friendly interface is provided to assist individuals and companies to invest or not in the specific stock.

The proposed stock market prediction model give better prediction accuracy as compared to the state-of-the-art methods, however, different limitations are identified as follows: (i) prediction accuracy is comparatively low by using the three predictors, (ii) only monthly stock return is predicted by the proposed model, and (iii) the existing regression-based analytical approach for stock prediction has performance issues due to incorporation of only quantitative nature of data.

There are different directions for future extensions, such as (i) prediction accuracy can be increased further, if more predictors are added in the multiple regression model, (ii) the weekly and daily stock return can also be predicted instead of merely predicting the monthly stock return, (iii) to make the prediction model more robust, lagging of variables before their use in the function is required to be incorporated, and (iv) hybrid model in conjunction with quantitative and qualitative variables can be used to predict stock market trend more efficiently.

Author contributions MZA and FR conceived and designed the experiments; FR and FMK performed the experiments; FR and SA analyzed the data; SA contributed reagents/materials/analysis tools; MZA wrote the paper.

Data availability All of the data and materials are uploaded through the editorial manager.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

- Ahmadifard M, Sadenejad F, Mohammadi I, Aramesh K (2013) Forecasting stock market return using ANFIS: the case of Tehran Stock Exchange. *Int J Adv Stud Hum Soc Sci* 1(5):452–459
- Ariyo AA, Adewumi AO, Ayo CK (2014) Stock price prediction using the ARIMA model. In: Proceedings of the 16th international conference on computer modelling and simulation (UKSim), 2014 UKSim-AMSS, pp 106–112. IEEE
- Ariyo AA, Adewumi AO, Ayo CK (2014). Stock price prediction using the ARIMA model. In: Proceedings of the 16th international conference on computer modelling and simulation (UKSim), 2014 UKSim-AMSS, pp 106–112. IEEE
- Devi BU, Sundar D, Alli P (2013) An effective time series analysis for stock trend prediction using ARIMA model for nifty midcap-50. *Int J Data Min Knowl Manag Process* 3(1):65
- Enke D, Grauer M, Mehdiyev N (2011) Stock market prediction with multiple regression, fuzzy type-2 clustering and neural networks. *Procedia Comput Sci* 6:201–206
- Feinberg A, Genethliou D (2005) Load forecasting. In: Chow, J., Wu, F., Momoh, J. (eds) Applied mathematics for restructured electric power systems. Springer, Berlin, pp 269–285
- Javaid U (2010) Determinants of equity prices in the stock market. *Res J Commerce Econ Soc Sci* 4(1):98–114
- Kabacoff R (2015) *R in action: data analysis and graphics with R*. Manning Publications Co., Shelter Island
- Kamley S, Jaloree S, Thaku RS (2013) Multiple regression: a data mining approach for predicting the stock market trends. *Int J Comput Sci Eng Inf Technol Res* 3(4):173–180
- Khaidem L, Saha S, Dey SR (2016) Predicting the direction of stock market prices using random forest. arXiv preprint [arXiv:1605.00003](https://arxiv.org/abs/1605.00003)

- Ladan MS, Karim AM, Adekeye KS (2014) Multiple regressions on some selected macroeconomic variables on stock market returns from 1986–2010. *Adv Econ Int Financ* 1(1):1–11
- Mondal P, Shit L, Goswami S (2014) Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. *Int J Comput Sci Eng Appl* 4(2):13
- Nagar A, Hahsler M (2012) News sentiment analysis using R to predict stock market trends, vol 2. pp 1–20
- Nasseri A, Tucker A, Cesare S (2015) Quantifying StockTwits semantic terms' trading behavior in financial markets: an effective application of decision tree algorithms. *Expert Syst Appl* 42(23):192–210
- Park NJ, George KM, Park NA (2010) A multiple regression model for trend change prediction. In: Proceedings of international conference on financial theory and engineering (ICFTE), 2009, pp 1–5
- Qasem M, Thulasiram R, Thulasiram P (2015) Twitter sentiment classification using machine learning techniques for stock markets. In: Proceedings of IEEE international conference on advances in computing, communications and informatics (ICACCI), pp 834–840
- Qin Y, Song D, Cheng H, Cheng W, Jiang G, Cottrell G (2017) In: Proceedings of the international joint conference on artificial intelligence (IJCAI)
- Rao T, Srivastava S (2012) Analyzing stock market movements using twitter sentiment analysis. In Proceedings of the IEEE international conference on advances in social networks analysis and mining (ASONAM 2012), pp 1–5
- Reddy BS (2010) Prediction of stock market indices—using SAS. In: Proceedings of 2nd IEEE international conference on information and financial engineering (ICIFE), pp 112–116
- Shen S, Jiang H, Zhang T (2012) Stock market forecasting using machine learning algorithms. Stanford University, Stanford, pp 1–5
- Simon S, Raoot A (2012) Accuracy driven artificial neural networks in stock market prediction. *Int J Soft Comput* 3(2):35–44
- Soni S, Shrivastava S (2010) Classification of Indian stock market data using machine learning algorithms. *Int J Comput Sci Eng* 2(9):2942–2946
- Yuan J, Luo Y (2014) Test on the validity of futures market's high frequency volume and price on forecast. In: Proceedings of IEEE international conference on management of e-Commerce and e-Government (ICMeCG), pp 28–32
- Zhang L, Zhang L, Teng W, Chen Y (2013) Based on information fusion technique with data mining in the application of finance early-warning. *Procedia Comput Sci* 17:695–703

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Muhammad Zubair Asghar is HEC Approved PhD supervisor, recognized by Higher Education Commission of Pakistan, and working as Assistant Professor, in the Institute of Computing and Information Technology (ICIT) at Gomal University, Pakistan. His research interests includes Text Mining, Opinion Mining and Sentiment Analysis, Natural Language Processing and Intelligent Systems. He is currently acting as Reviewer and Academic Editor of different top-tier journals, such as Plos One, IEEE ACCESS, and others. Furthermore, he is acting as Special Session Chair (Social Computing) at BESC 2018 International Conference (Taiwan), and a Guest Editor of special issue on Health Informatics, Journal of Medical Imaging and Health Informatics (impact factor 0.549).

Fazal Rahman completed his MS (Text Mining) under the supervision of Dr. Muhammad Zubair Asghar from Institute of Computing and Information Technology (ICIT) at Gomal University. His area of interest includes data mining, text mining, and opinion mining.

Fazal Masud Kundi is an Assistant Professor in the Institute of Computing and Information Technology (ICIT) at Gomal University, Pakistan. His research interests includes Text Mining, Opinion Mining and Sentiment Analysis, Natural Language Processing and Intelligent Systems.

Shakeel Ahmad is full Professor in Faculty of Computing and Information Technology in Rabigh (FCITR), King Abdul Aziz University (KAU). His area of interest includes data mining, text mining, and opinion mining.