International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Prediction of Market Behavior for Short Term Stock Prices using Regression Techniques

Tousif Al Rashid, Vinay Kumar Goyal

Abstract*: Stock price Prediction always been a desired area for many institutions of finance. As price prediction in finance has long been a challenging task due to volume and speed of the data, investors are always looking for good algorithm to know the future price. The various machine learning algorithms (MLR, SVM, Random Forest etc.) used to predict and make further decision on stock market. The errors of predicted prices may be minimized, if the labeled dataset is mined in a efficient way. As the technical analysis always plays a major role to put profit in a investors pocket, a very simple algorithm is proposed for short term closing price prediction after analyzing similar types of movements of last few days prices to the historical data of that stock. A novel approach using Correlation Coefficients, Euclidian Distance and machine learning techniques is proposed to forecast a meaningful price based on the SBI data, fetched from the Yahoo Finance.

Index Terms*: Multiple Linear Regression, Random Forest, Technical analysis, Short term stock prediction, Data Mining.

I. INTRODUCTION

Stock market is a place where traders can buy and sell stocks, bonds and other securities. Broadly it can be divided into two components, 1) Primary market, and 2) Secondary market. In the primary market, the trader can purchase shares directly from the company. As [10] explored that, in Secondary Market, trader buys and sells the stocks and bonds among themselves . In the era of globalization and with the advent of fast pace digital technology, the financial data is increasing at the unprecedented rate. Stock market prediction is difficult due to its unforeseeable nature and high risk . Again financial data is more complicated than that of statistical data due to irregular movements, seasonal and cyclic variations. Due to the growth of massive financial data, we require a automated process to analyze this. The ultimate goal of a trader is to achieve high profit and gaining high profit from stock market depends on analysis of movement of highly fluctuating and irregular stock price values. Keeping track on historical data of individual stocks will reduce the uncertainty associated to investment decision making. And the only way, the traders can make huge profit is by forecasting prices with low error.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Tousif Al Rashid*, Computer Science and Engineering, Chandigarh University, Mohali, India.
Vinay Kumar Goyal, Statistical and Mathematical Modeling and Machine Learning particularly in health sciences and financial prediction.

Knowledge can be extracted from the historical data. The process of extracting the hidden useful knowledge from the data is called data mining

TABLE I NOTATIONS AND DESCRIPTIONS

Notation	Description
ML	Machine learning
MLR	Multiple linear regression
RF	Random forest
NHPC	Nearest highly positive correlated
APE	Absolute percentage error
MAPE	Mean absolute percentage error
ED	Euclidian Distance
CP	Closing Price
PCI	Percentage change of inputs
PCO	Percentage change of outputs

[11] . Machine learning can help to do tasks beyond Human capabilities. Finding the useful patterns from massive data sets is a talented domain, which can be achieved by using machine learning Techniques [12] . Supervised learning, Unsupervised learning and Reinforcement learning are the three categories of machine learning [6] . Supervised learning is intended to learn from labeled dataset using various machine learning techniques and used to find interesting patterns in huge dataset. Based on the labeled values of dataset the supervised learning techniques are categorized into regression and classification. When the target values are continuous, regression comes into the frame and when the label set is comes from values of finite set, it is a classification. Unsupervised learning is unsupervised because the learning is not done under any supervision. The unsupervised learning algorithms segment the unlabeled data into clusters and produce the parameter values. Reinforcement learning is about reinforcing to take action and maximize the reward in a particular situation. To forecast the future trends of a stock, the traditional market prediction approaches usually utilize the historical price- related data of that stock [13] . Markku Karhunen [7] investigate whether a asset management can be benefitted from an expert system or not. Using statistical and machine-learning algorithms, author generate binary predictions of returns of the market. The methods used include regularized logistic regressions, logistic regressions and similarity-based classification. Because of non-stationary and noisy characteristics of samples. The stock price prediction always been a challenging task for researchers and speculators. Wen Long et al [8] proposed a multi-filters neural network (MFNN) model for feature extraction on financial time series data and price movement forecasting task.

A integrated model of technical analysis with machine learning techniques can be more efficient to make profitable trading decisions[3]. Volume also have it's impact on stock prediction. Dinesh Bhuriya et.al [2] predicted the stock market behavior on the basis of volume. Hiransha M et al [10] predicted the stock market behavior using linear and non linear models and concluded that linear model is not good as compared to non linear models for long term prediction of stock market behavior. After reviewing the above literature, it has been observed that short term closing price may be predicted using linear and non linear model. In this paper prediction of short term closing price based on a created labeled dataset is proposed using Multiple Linear Regression(MLR) and Random Forest.

A. Techniques Used

1) Correlation coefficient: The various mathematical and statistical models have proven track record in financial data prediction so far. There are various techniques available in the literature. But in the present study we have used correlation coefficient for pattern recognition and to check the similar kind of movements. In the current study we used the Pearson's correlation coefficient as mentioned explained by [4],

$$r_{xy} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sqrt[2]{\sum_{i=1}^{n}(x-\bar{x})^2}\sqrt[2]{\sum_{i=1}^{n}(y-\bar{y})^2}} \quad (1)$$

- where n is size of sample data.
- $\bar{x}, \bar{y}$ are the sample mean.

The correlation coefficients has been extensively used to analyse relationship of trends between various markets.

2) Euclidean distance: The Euclidean distance is used to measure the linear distance between two points in a Euclidean space. If $x = (x_1,x_2,...,x_n)$, $y = (y_1,y_2,...,y_n)$ are points in Euclidean n space. [5] expressed the Euclidean distance between x and y as :

$$d(x_i, y_i) = \sqrt[2]{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

The Euclidian distance has been used to calculate nearest correlated segments.

II. Mining Similar Movements

The stock prices always follows random walk but we can find out similar kind of movements of days Closing, High, Low and Opening prices and with the help of ML techniques we can predict the future prices.

The segments with similar movements of High, Low, Open and Closing prices with recent day's price movements of High, Low, Open and Closing prices as shown in Fig 1, can be extracted by using Pearson's coefficient of correlation and the Euclidian distance with specific steps,

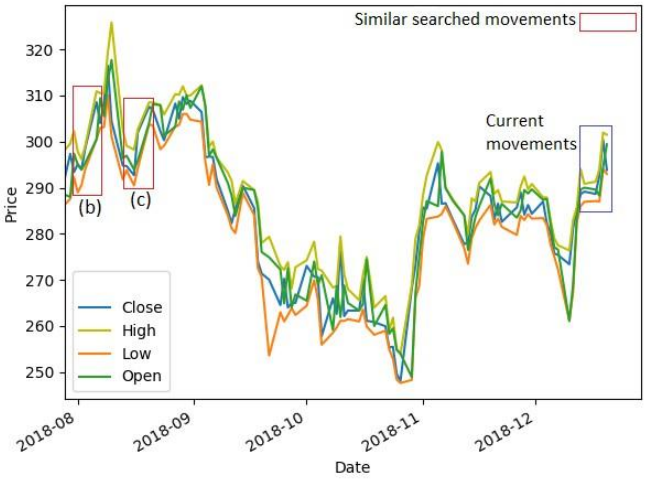


Fig. 1. Similar movements of High, Low, Open and Closing prices with recent days prices.

TABLE II
EACH DAY PRICES.

Date	High	Low	Open	Close	Day
2002-05-13	21.983	21.710	21.776	21.748	d_1
2005-07-28	22.008	21.479	21.900	21.541	d_2
2005-07-29	21.677	21.144	21.606	21.276	d_3
:	:	:	:	:	:
2018-12-20	301.50	293.00	299.39	293.85	d_t

A. Procedure:

In order to find out the similar kind of movements, we segmented the whole time series data of stock prices for t number of days $(d_1,d_2,...,d_t)$.where d is a tuple of opening, high, low and closing price values as shown in TABLE II. The task after segmentation of whole dataset, is to find out the most nearest highly positive correlated(NHPC) segments. Where the last segment represents the current day's movements as shown in Fig 1 .

B. Step 1

In the first step, divide the t no of days into (t-n+1) segments, where n is the length of segments and the value of n is considered as $(5 \leq n \leq 9)$. Fig 2 is visualizing the segmentation of t days data. As the total trading days in a week is five, the minimum segment length is considered as five and the maximum length is considered is nine. After prediction of closing prices, it is observed that the best length of segment is, n=7. As the length 7 is producing the less total mean absolute percentage error (MAPE) as shown in TABLE IX. In this paper the cosidered segment length is 7.

C Step 2

The next task is to identify the segments from the set of $(s_1,s_2,,s_{(t-n)})$, which are nearest highly positive correlated (NHPC) with the last segment $s_{(t-n+1)}$. The Pearson's coefficient correlation and the Euclidian distance is used, with the specific steps as follows,

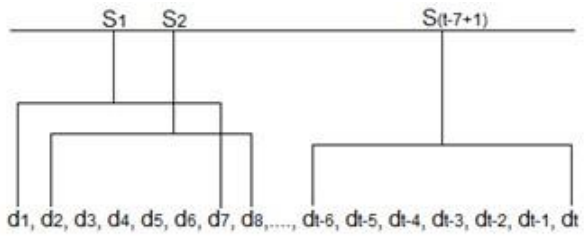


Fig. 2. The segmentation of t time series into (t-n+1) total segments , where the length of each segment is n=7.

1) : Calculate the correlation between the last segment's $((t-n+1)$'th segment) Opening, High, Low and Closing prices with $s_{(t-n+1)}$'s Opening, High, Low and Closing prices correspondingly. As corresponding prices are the same the correlation set between them will be [1.0,1.0,1.0,1.0].

2) : Now, calculate the correlation between i'th segment's High, Low, Opening and Closing prices with last segment, $s_{(t-n+1)}$'s corresponding High, Low, Opening and Closing prices. where i is lies between $(1 \le i \le (t-n+1))$.

- The correlation set will be $[-1 \le CH_{(i,t-n+1)} \le 1, -1 \le CL_{(i,t-n+1)} \le 1, -1 \le CO_{(i,t-n+1)} \le 1, -1 \le CC_{(i,t-n+1)} \le 1]$ where,

- $CH_{(i,t-n+1)}$, $CL_{(i,t-n+1)}$, $CO_{(i,t-n+1)}$ and $CC_{(i,t-n+1)}$ represents the correlations between i'th segment's High, Low, Open and Closing prices with $(t-n+1)$th segment's High , Low, Open and Closing prices respectively.

a) Step 3: Repeat, Step 2. For every i^0th segment, where i is $(1 \le i \le (t-n+1))$. TABLE III representing the correlations between i'th segment's high, low, open and closing prices with (t-n+1)'th segment's high, low, open and closing prices respectively.

b) Step 4: Calculate the Euclidean distance between each $(s_1, s_2, .., s_{(t-n)})$ segment's correlation set $[CH_{(i,t-n+1)}, CL_{(i,t-n+1)}, CO_{(i,t-n+1)}, CC_{(i,t-n+1)}]$ with the correlation set of last segment $s_{(t-n+1)}$, which is [1.0,1.0,1.0,1.0], as :

$$d(C_{(i,\alpha)}, C_{(\alpha,\alpha)}) = \sqrt{(1-CH_{i,\alpha})^2 + (1-CL_{i,\alpha})^2 + (1-CO_{i,\alpha})^2 + (1-CC_{i,\alpha})^2} \quad (3)$$

where,

- $C_{(i,\alpha)}$ represents the correlation set between i'th segment and α'th segment .
- $\alpha = t-n+1$

c) Step 5: Return all Euclidian distances excluding the Euclidian distance between $d(C_{(\alpha,\alpha)}, C_{(\alpha,\alpha)})$. TABLE IV showing the Euclidian distances between segment s_i and $s_{(t-n+1)}$. 200 most nearest highly positive correlated segments with last segment $s_{(t-n+1)}$, are collected by sorting all the Euclidean distances as shown in TABLE V. Here the Euclidean distance is used to find the nearest highly positive correlated segments to get the nearest similar trends of prices with the recent days prices.

TABLE IV
EUCLIDEAN DISTANCES BETWEEN CORRELATION SET OF SEGMENT s_i (EXCLUDING $t-n+1^0 th$ SEGMENT)AND $s_{(t-n+1)}$

s_i	Euclidean distance
s_1	3.5393
s_2	3.4523
s_3	3.2118
:	:
$s_{(t-n)}$	0.52904

III. CREATING LABELED DATA FOR PREDICTION

In this paper as we are trying to predict upcoming two days closing prices, the labeled data need to create very carefully. Using Pearson's correlation coefficient and Euclidean distance, two hundred most nearest highly positive correlated segments are collected. These segments follow similar price movements with the last segment $s_{(t-n+1)}$.

d) Step 1: To create labeled dataset, only the closing prices are considered from the 200 most nearest highly positive correlated segments. Then the values of $CC_{(i,t-n+1)}$ of 200 collected segments are sorted. From sorted correlations of closing prices, 100 segments with more than 85 percent positive correlation coefficient values are considered. The detailed process is visualized in Fig 3.

e) Step 2: calculate the average closing prices for each 100 segments (TABLE VI).

TABLE VI
AVERAGE CLOSING PRICES OF 100 CONSIDERED SEGMENTS.

s_i	s_{875}	s_{360}	s_{1979}	.. .	s_{259}
$CC_{(i,t-n+1)}$	0.9829	0.9578	0.9473	.. .	0.8642
Average close	88.863	42.635	209.908	.. .	28.492

f) Step 3: calculate the percentage changes of closing prices of each 100 segments, from its corresponding average closing prices and consider it as input dataset of the label dataset. The percentage changes of closing prices from it's corresponding mean closing prices are calculated in Fig 4, where in $X_{i,j}$, i represents the segment number and j represents the entry number to a segment.

g) Step 4: fetch the adjacent next two day's closing prices of each 100 segments to create the target values.

d) Step 5: calculated the percentage changes of adjacent next two days closing prices of each 100 segments, from its corresponding average closing prices calculated in step 2 and consider it as two days target values (Fig 5), where in $Y_{i,j}$, i represents the segment number and j represents adjacent next 1'st and 2'nd day's closing prices of i th segment.

e) Step 6: return all the input data and their corresponding target values(Fig 6). As the goal is to predict the upcoming two days closing prices, we need to find adjacent next to days closing prices of last segment $s_{(t-n+1)}$, as the Target values of $s_{(t-n+1)}$ are unknown in TABLE VIII. But before predicting adjacent next two day's closing prices of $(t-n+1)$'th segment , calculate the percentage changes of closing prices of segment $s_{(t-n+1)}$ from it's average closing price as shown in TABLE

f)

TABLE III

CORRELATION COLLECTION BETWEEN SEGMENT s_i AND $s_{(t-n+1)}$

s_i	$CH_{(i,t-n+1)}$	$CL_{(i,t-n+1)}$	$CO_{(i,t-n+1)}$	$CC_{(i,t-n+1)}$
s_1	-0.80798	-0.85735	-0.64069	-0.76545
s_2	-0.78463	-0.79783	-0.63213	-0.68461
s_3	-0.76459	-0.46044	-0.55521	-0.62815
:	:	:	:	:
$s_{(t-n)}$	0.75004	0.76665	0.77012	0.66815
$s_{(t-n+1)}$	1.0	1.0	1.0	1.0

TABLE V

TWO-HUNDRED NEAREST HIGHLY POSITIVE CORRELATED SEGMENTS WITH EUCLIDEAN DISTANCES.

Sr.No.	NHPC segments	$CH_{(i,t-n+1)}$	$CL_{(i,t-n+1)}$	$CO_{(i,t-n+1)}$	$CC_{(i,t-n+1)}$	ED
1	s_{360}	0.9413	0.9837	0.9777	0.9578	0.0773
2	s_{3909}	0.9552	0.9702	0.9501	0.9236	0.1058
3	s_{1979}	0.9217	0.9773	0.9499	0.9473	0.1091
:	:	:	:	:	:	:
200	s_{2429}	0.8966	0.8325	0.7828	0.8753	0.3184

TABLE VII

THE LAST SEGMENT, $s_{(t-n+1)}$'S CLOSING PRICE'S PERCENTAGE DISTRIBUTIONS FROM ITS MEAN CLOSING PRICE.

Date	Closing price	Mean CP (X_{t-n+1})	PCI($_{t-n+1,j}$)
2018-12-12	285.250	291.207	-2.046
2018-12-13	288.600	291.207	-0.895
2018-12-14	289.150	291.207	-0.706
2018-12-17	288.650	291.207	-0.878
2018-12-18	292.550	291.207	-0.461
2018-12-19	300.400	291.207	3.157
2018-12-20	293.850	291.207	0.908

VII, where percentage change of each entry from the mean closing price of $(t-n+1)$'th segment can be can be calculated as:

$$\mathrm{PCI}(_{t-n+1,j}) = \frac{X_{t-n+1,j} - \overline{X_{t-n+1}}}{\overline{X_{t-n+1}}} * 100$$

- where, X_{t-n+1} denotes the mean value of $(t-n+1)$'th segments's closing prices.

IV. PREDICTION METHODS

To predict upcoming two day's closing prices from the created labeled data set, the multiple linear regression model and Random Forest model are used. Regression is very much ideal for Predicting continuous values. From a given labeled dataset, the regression establishes a relationship between predicted and the target values and makes a pattern to predict values for datasets whose target values are unknown. Regression model is defined using labeled dataset and based on that model, target value is predicted [1]. Random Forest is an ensemble learning technique for regression and classification. As Random Forest is forest of decision trees, It can be said that Random Forest is a advanced model of decission tree [9] . In this section the multiple linear regression and Random Forest is used to know the future prices. First the model trained with the labeled data and then two days closing prices are predicted. The predicted

TABLE IX

TOTAL MAPE FOR EACH VALUE OF n

value of n	MAPE(MLR)	MAPE(Random Forest)	total MAPE
5	1.679	1.685	3.364
6	1.218	0.658	1.876
7	0.680	1.146	1.826
8	0.673	1.847	2.520
9	0.577	2.720	3.297

results are compared with the actual results using MAPE to find the error.

V. RESULT ANALYSIS

The data of SBI stock dataset of Bombay Stock Exchange has been fetched from yahoo finance. After collecting 100 segment's closing price's percentage changes from their corresponding mean closing prices, considered it as input data and adjacent next two day's closing price's percentage changes as target values and this labeled data is putted into the multiple linear regression and random forest to get trained.
The multiple linear regression model worked very well on this created dataset. The predicted values we got are the values of percentage changes of adjacent next two day's closing prices, from the mean closing price of last segment $s_{(t-n+1)}$. After calculating the percentage changes from mean value, we got the predicted closing prices.To check whether the multiple linear regression model or the Random forest model is performing better , Mean absolute percentage error(MAPE) is calculated (TABLE X)between actual closing prices and predicted closing prices .

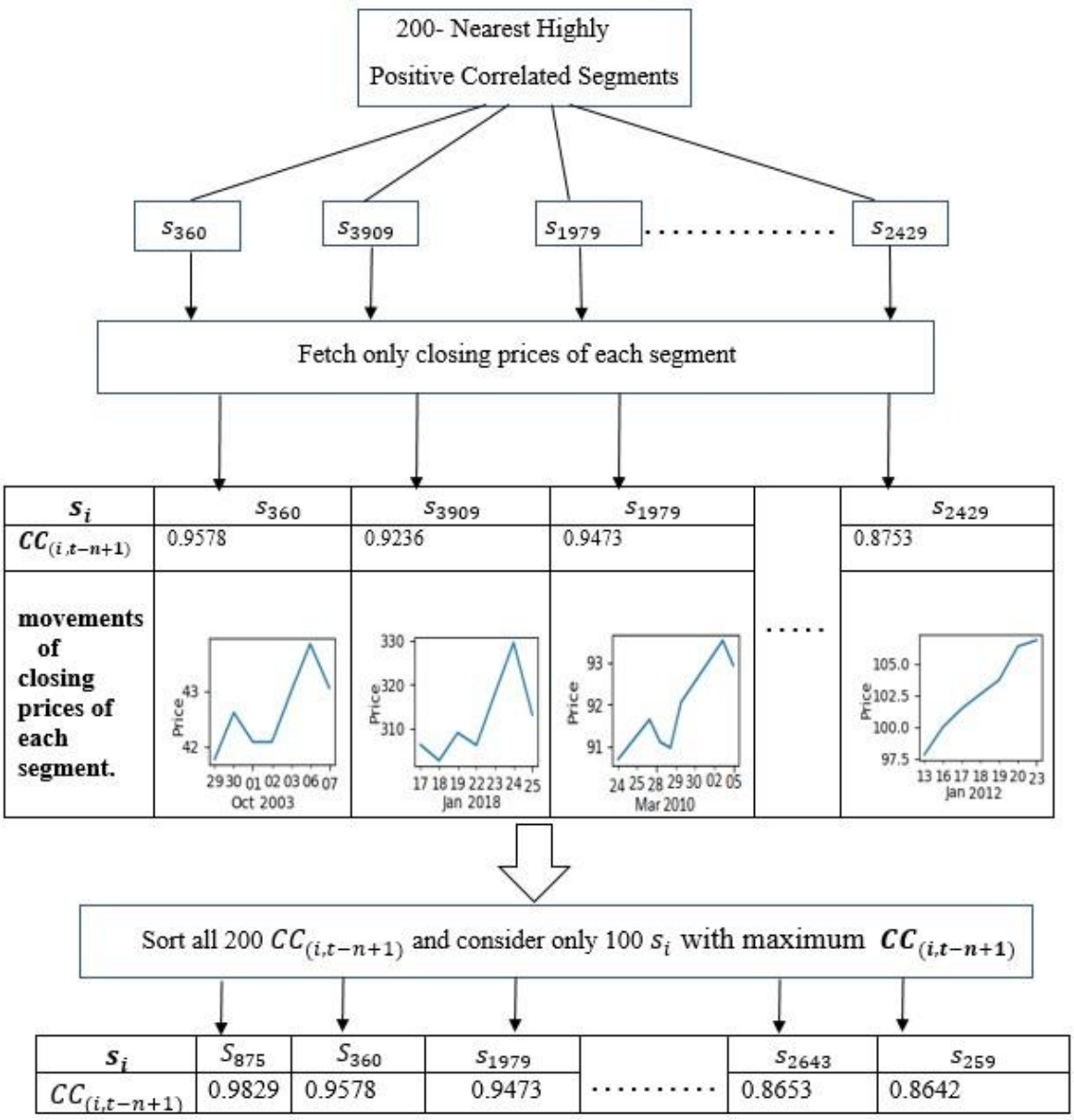


Fig. 3. collecting 100 segment's closing prices with maximum $CC_{(i,t-n+1}$

s_i				Input Values				Target Values	
	$PCI_{(i,1)}$	$PCI_{(i,2)}$	$PCI_{(i,3)}$	$PCI_{(i,4)}$	$PCI_{(i,5)}$	$PCI_{(i,6)}$	$PCI_{(i,7)}$	$PCO_{(i,1)}$	$PCO_{(i,2)}$
$St-n+1$	-2.046	-0.895	-0.706	-0.878	-0.461	3.157	0.908	- -	- -

Table VIII: Target values to be predicted

Date	Actual Closing Price	Predicted Closing Price (MLR)	APE(MLR)	Predicted Closing Price(Random Forest)	APE(Random Forest)
2018-12-21	291.650	292.998	0.462	294.325	0.917
2018-12-24	292.800	295.434	0.899	288.775	1.375
MAPE			0.680		1.146

Table X: Predicted closing prices with calculated MAPE

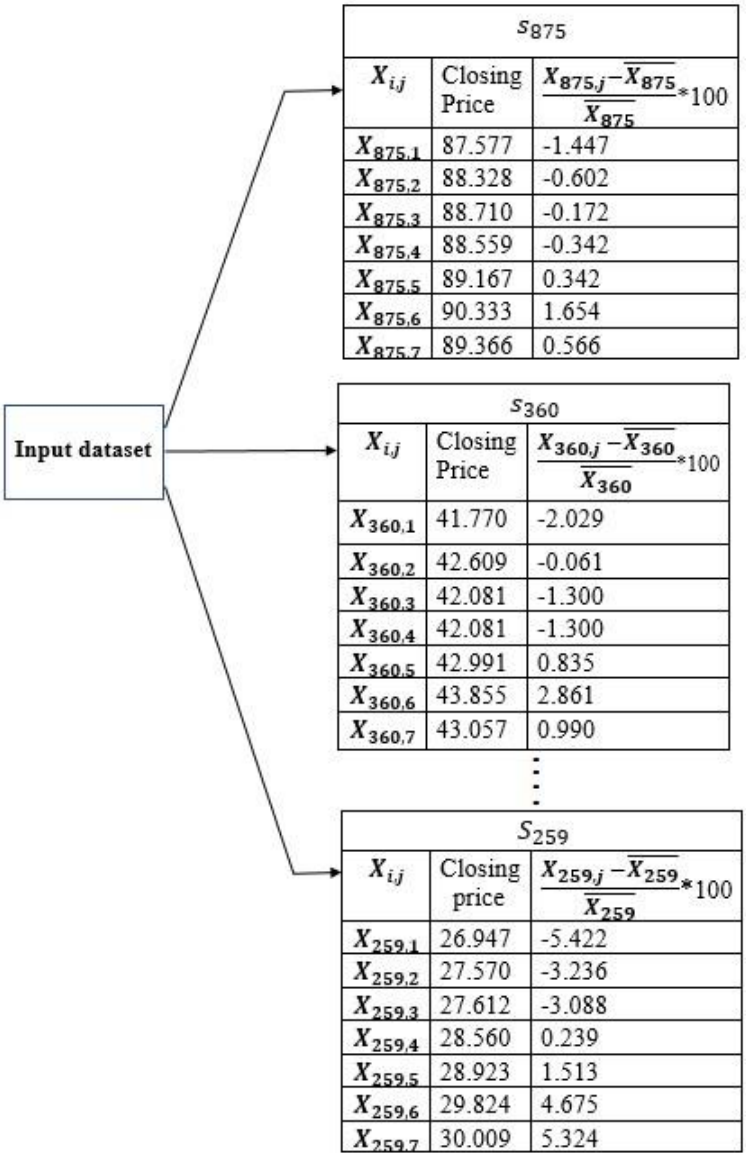
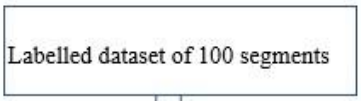
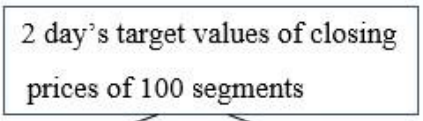


Fig. 4. Change in percentage of closing prices for every 100 segments with respect to their corresponding mean closing price



S_i	Input Values							Target values	
	$PCI_{(i,1)}$	$PCI_{(i,2)}$	$PCI_{(i,3)}$	$PCI_{(i,4)}$	$PCI_{(i,5)}$	$PCI_{(i,6)}$	$PCI_{(i,7)}$	$PCO_{(i,1)}$	$PCO_{(i,2)}$
S_{875}	-1.447	-0.602	-0.172	-0.342	0.342	1.654	0.566	-1.760	-2.731
S_{360}	-2.029	-0.061	-1.300	-1.300	0.835	2.861	0.990	4.897	6.323
:	:	:	:	:	:	:	:	:	:
S_{259}	-5.422	-3.236	-3.088	0.239	1.513	4.675	5.324	6.387	7.554

Fig. 6. The labeled dataset



S_{875}

$Y_{i,j}$	Closing Prices	$\frac{Y_{875,j}-\overline{X_{875}}}{\overline{X_{875}}}*100$	Dates
$Y_{875,1}$	87.299	-1.760	2005-10-06
$Y_{875,2}$	86.436	-2.731	2005-10-07

.....

S_{259}

$Y_{i,j}$	Closing Prices	$\frac{Y_{259,j}-\overline{X_{259}}}{\overline{X_{259}}}*100$	Dates
$Y_{259,1}$	87.299	-1.760	2005-10-06
$Y_{259,2}$	86.436	-2.731	2005-10-07

Fig. 5. the percentage changes of adjacent next two day's closing prices for each 100 segments.

VI. CONCLUSION

SBI dataset of BSE is predicted using machine learning techniques. The model has been trained using multiple linear regression(MLR) and random forest after creating a labeled dataset by fetching the similar movements from historical dataset. It has been observed that percentage of average prediction error for two days closing prices are 0.68 and 1.146 for MLR and random forest respectively. It is observed that MLR model is performing better than Random Forest for the short term closing price prediction. So, it is concluded that linear model is more efficient than non linear model for predicting the short term closing price.

REFERENCES

1. Jennifer Hill Andrew Gelman. Data Analysis Using Regression and Multilevel/Hierarchical Models. 2006.
2. Dinesh Bhuriya, Girish Kaushal, Ashish Sharma, and Upendra Singh. Stock market predication using a linear regression. pages 510–513, 04 2017.
3. Rajashree Dash and Pradipta Kishore Dash. A hybrid stock trading framework integrating technical analysis with machine learning techniques. The Journal of Finance and Data Science, 2, 03 2016.
4. SC Gupta and VK Kapoor. Fundamental of Mathematical Statistics.
5. Jian Pei Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques. 2012.
6. Judith Hurwitz ,Daniel Kirsch. Machine Learning For Dummies. 2018.
7. Markku Karhunen. Algorithmic sign prediction and covariate selection across eleven international stock markets. Expert Systems with Applications, 115, 07 2018.
8. Wen Long, Zhichen Lu, and Lingxiao Cui. Deep learning-based feature engineering for stock price movement prediction. Knowledge-Based Systems, 164, 11 2018.
9. Gilles Louppe. Understanding Random Forests: From Theory to Practice. PhD thesis, 10 2014.
10. M, Hiransha and Gopalakrishnan, E.A. and Menon, Vijay and Kp, Soman.
11. Nse stock market prediction using deep-learning models. Procedia Computer Science, 132:1351–1362, 01 2018.

12. Gregory Piatetsky-Shapiro Usama Fayyad Ramasamy Uthurusamy, Padhraic Smyth, editor. Advances in Knowledge Discovery and Data Mining. 1996.
13. Shai Ben-David Shai Shalev-Shwartz. Understanding Machine Learning: From Theory to Algorithms. 2014.
14. Xi Zhang, Yunjia Zhang, Senzhang Wang, Yuntao Yao, Binxing Fang, and Philip Yu. Improving stock market prediction via heterogeneous information fusion. Knowledge-Based Systems, 12 2017.

AUTHORS PROFILE

Tousif Al Rashid is a ME scholar at Chandigarh University and currently he is working in financial prediction using Machine Learning and statistical techniques.

Dr. Vinay Kumar Goyal is currently working on Statistical and Mathematical Modeling and Machine Learning particularly in health sciences and financial prediction. He has is good number of publication in the related area. He is having more than 22 Years of experience in Research, Teaching and Administration. Currently, he is working as Deputy Dean and Head of the Computer Science and Engineering Department at Chandigarh University, Chandigarh.