A Study on State Bank of India ShareUsingRegression Analysis

M. Vijayakanth[1], V. Veeramanikandan[2]

[1]Research Scholar, Dept. of Computer and Information Science, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India
Email: vijayakanth82@gmail.com

[2]Assistant Professor, Dept. of Computer Science, ThiruKolanjiappar Govt. Arts College, Vridhachalam-606 001, Tamil Nadu, India
Email: klmvmani@gmail.com

Abstract

Data mining is defined as a process to find and extract hidden information from a larger dataset. It is used to analysing data patterns in volume of data using one or more data mining techniques. Data mining has applications in multiple fields, like science, engineering, banking, share marketing and various fields of research. A stock market is similarly called a share market. The stock market helps you trade financial instruments shares of companies and allows trading of shares. In this paper consider banking share namely state bank of India (SBI) share analysing and predicting to consider different related parameters namely trading date, price, high, low, volume and changes in share using regression approaches. Numerical illustrations also provide to prove the results and discussions using different parameters

Keywords:Data Mining; Share marketing; SBI Share; Regression model and Forecasting.

DOI Number: 10.48047/NQ.2022.20.20.NQ109309 NeuroQuantology2022;20(20):3119-3126

1. Introduction

In India BSC (Bombay Stock Exchange) is the leading stock exchange organization and was started in the year of 1875 as the Native Share and Stockbrokers Association. In total number nearly 6,000 companies listed under BSE. BSC is one of the largest stock brokage companies in the world. The Bombay Stock Exchange has contributed to the Indian capital markets. In Asia, the BSE is the premier stock exchange and also holds a stock trading platform for small and medium-sized enterprises (SMEs). BSE specializes in providing other capital markets services including clearing, settlement, and market risk.

Data mining is the practice of analysing hidden patterns utilizing existing data. Data mining is also known as KDD for surviving with advanced data analysis [Han et al. (2011)]. The main stages of a data mining procedure are data discovery, data acquisition, data cleansing, integration, data selection, data transformation, and knowledge discovery [Bocca et al. (2016)]. It is a statistical utility for exploited in the perspective of statistical models and its basic objective is either to forecast the future results and the necessitated correlation coefficients. Quantify how well-observed results are predicated by the model, based on the level of the overall change in outcomes that is described by the model. [Steel et al. (1960), Glantz et al. (1990), Draper et al.

(1998), Lima et al. (2013) and Rathod et al. (2018)].

The authors discuss that the investigation based on recent hybridization of STDL and OKELM using short- and medium-term prediction for everyday sharing likes close price of the CRUDE OIL index. The parameter study of ELM done by utilizing the Gray Wolf Optimization Algorithm (GWO) to the predictive performance. The benefits in the intended work are done through the benefit of two related quantities called MASE and SMAPE. [Veeramanikandan (2020)] and [Jeyakarthic (2020)].

The aid of advanced machine learning techniques has significantly enhanced prediction accuracies. Analysis and projection of stock markets proceed to remain one owing to dynamic and chaotic information on the most difficult research areas. The machine learning technique have been workout market analysis and prediction techniques[Rouf et al. (2021)]. The outcome of the HSPM (hybrid stock prediction model) using the accuracy parameters namely MAE the RMSE metric. The performance of the prediction model and its accuracy in DNN and ANN, with a 5% to 7% progress in the RMSE score. Indian stock price data are deemed for the work [Manujakshi et al. (2022)].

2. Experimental Methods or Methodology

2.1 Simple Linear Regression Model (SLRM)

Regression analysis is a statistical tool adopted to establish a link between two or more variables. Likewise, one of these variables, called the predictor variable, means the value is collected using experiments. Another important variable is called the response variable, which means it is extracted from the predictor. The general mathematical equation for simple linear regression model is,

$$y = a_x + b \quad \ldots (1)$$

Where y is the required response variable, x is the needed predictor variable, and a, b is called coefficients.

2.2 Correlation Coefficient (CC)

The CC or coefficient of determination denoted R2 or r2 score which is used to moderation in the dependent variable means predicted from the independent variables. In this case, the r (CC) returns nearly 1.0 means strong positive correlation. If the value of r returns nearly -1 means strong negative correlation and return 0 means no correlation between all the variables.

$$\boldsymbol{r} = \frac{\boldsymbol{n}(\sum \boldsymbol{xy}) - (\sum \boldsymbol{x})(\sum \boldsymbol{y})}{\sqrt{[\boldsymbol{n}\sum \boldsymbol{x^2} - (\sum \boldsymbol{x})^2] - [\boldsymbol{n}\sum \boldsymbol{y^2} - (\sum \boldsymbol{y})^2]}} \quad \ldots (2)$$

2.3 Mean Absolute Error (MAE)

In machine learning approach, MAE means the average of absolute error in future prediction which means error range between prediction and observations. In data mining and ML research, the MAE denoted as loss function. The given accuracy formula is:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad \ldots (3)$$

where n is called the no. of elements in the iterations, Σ which is used to add them all up and $|y_i - x_i|$ called absolute error between actual and predicted.

2.4 Root Mean Square Error (RMSE)

The RMSE is one of the familiar accuracy finding methods in data analysis, which is used to compute test the quality of prediction or forecasting. RMSE sometime

name as root mean square deviation which is used to find the residuals between prediction and truth for all data points. The RMSE calculated using the following formula.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \|y(i) - \hat{y}(i)\|^2}{n}} \quad \ldots (4)$$

wheren is called the number of elements in the iterations, y(i) meansi[th] measurement, and y^(i) called the prediction.

2.5 Relative Absolute Error (RAE)

The RAE is used to compute the accuracy for relatively comparison of each and every performance of a pred If RAE=0, the model behavior or accuracy is perfect.

$$\mathrm{RMSE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i - \bar{y}|} \quad \ldots (5)$$

wheren is called the number of elements in the observations, y(i) called the realized value and y^(i) called the prediction and $\bar{y}$ means the mean values of corresponding variables.

2.5 Root Relative Squared Error (RRSE)

RRSEis one of the accuracy metrics for predictive models called regression. It's anaccuracy parameter which is used to compute the first result and behavior ofmodel is performing. It is also an inheritance from RSE. The RRSE parameter for finding the process of square root for sum of squared errors for the corresponding predictive model with sum of squared errors.

$$\mathrm{RRSE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad \ldots (6)$$

wheren is called the number of elements in the observations, y(i) called the realized value and y^(i) called the prediction and $\bar{y}$ means the mean values of corresponding variables.

Numerical Illustrations

The secondary dataset (table 1) is taken from https://in.investing.com/. In table 1, indicate State Bank of India (SBI) BSE share details which is include monthly average for share price, opening price, month wise high, low price and changes (%). Based on table 1, find the future growth in banking sectors of SBI dataset. Regression model is one of the best statistical techniques which is used to find the future forecasting.

Table 1. State Bank of India Share Dataset

Date	Price	Open	High	Low	Changes %
Aug-21	456.95	433.7	467.25	432.05	5.82
Jul-21	431.8	420.3	444.4	417.15	3.01
Jun-21	419.2	426.05	441.95	400.5	-1.21
May-21	424.35	349.6	433.65	341.4	20.04
Apr-21	353.5	367.5	371.75	321.5	-2.96
Mar-21	364.3	395.1	408.9	345.2	-6.63
Feb-21	390.15	285.1	427.7	282.75	38.30
Jan-21	282.1	274.9	310.9	269.5	2.60

Dec-20	274.95	245	279.9	244.1	12.57
Nov-20	244.25	192.2	253	190.05	29.06
Oct-20	189.25	187	207.3	185.9	2.08
Sep-20	185.4	213.1	219	175.5	-12.55
Aug-20	212	192	231.55	189.55	10.73
Jul-20	191.45	179.45	202.5	178.6	7.28
Jun-20	178.45	164	197.5	163.35	10.63
May-20	161.3	182.45	183.5	149.45	-15.33
Apr-20	190.5	194	198	175	-3.23
Mar-20	196.85	311	312	173.55	-35.03
Feb-20	303	317.95	331.9	295.35	-4.85
Jan-20	318.45	334.7	339.85	305.65	-4.58
Dec-19	333.75	343.9	344.6	308	-2.37
Nov-19	341.85	312.4	351	299.7	9.43
Oct-19	312.4	272.75	317.8	244.35	15.36
Sep-19	270.8	272	316	266.95	-1.11

Table 2. Machine learning approaches and its performance metric

Machine Learning Approaches	Linear Regression	MAE	RMSE	RAE	RRSE
	-0.6048	11.1083	15.0684	100.0000	100.0000
CC					

Table 3. Machine learning approaches with price, open, high, low and its performance metric

Machine Learning Approaches	CC	MAE	RMSE	RAE	RRSE
Linear Regression	0.9689	14.2750	20.8333	19.7649	24.3449

Table 4. Machine learning approaches with open, high, low and its performance metric

Supervised Machine Learning	CC	MAE	RMSE	RAE	RRSE
Linear Regression	0.9217	19.5866	32.5647	27.1192	38.0537

Table 5. Prediction for open, high, and low using multiple linear regression and its performance metric

Prediction Parameter	Linear Regression Model	CC	MAE	RMSE	RAE	RRSE
Open	Open = 0.4308* High+0.5266*Low+ 10.3543	0.8983	23.7751	37.8621	32.1753	43.6728
High	High = 0.5281 * Open + 0.4928 *	0.9182	24.4902	35.4980	31.9745	38.4713

	Low + 34.7627					
Low	Low = 0.5284 * Open + 0.4033 * High +(-13.9538)	0.9217	19.5866	32.5647	27.1192	38.0537

Table 7. Prediction with high and low using linear regression

Date	Open	Actual High	Predicted High	Actual Low	Predicted Low
Aug-21	433.70	467.25	461.56	432.05	401.39
Jul-21	420.30	444.40	448.38	417.15	388.99
Jun-21	426.05	441.95	454.04	400.50	394.31
May-21	349.60	433.65	378.81	341.40	323.57
Apr-21	367.50	371.75	396.42	321.50	340.13
Mar-21	395.10	408.90	423.58	345.20	365.67
Feb-21	285.10	427.70	315.34	282.75	263.89
Jan-21	274.90	310.90	305.31	269.50	254.45
Dec-20	245.00	279.90	275.88	244.10	226.78
Nov-20	192.20	253.00	223.93	190.05	177.93
Oct-20	187.00	207.30	218.81	185.90	173.11
Sep-20	213.10	219.00	244.49	175.50	197.26
Aug-20	192.00	231.55	223.73	189.55	177.74
Jul-20	179.45	202.50	211.38	178.60	166.13
Jun-20	164.00	197.50	196.18	163.35	151.83
May-20	182.45	183.50	214.33	149.45	168.90
Apr-20	194.00	198.00	225.70	175.00	179.59
Mar-20	311.00	312.00	340.83	173.55	287.85
Feb-20	317.95	331.90	347.67	295.35	294.28
Jan-20	334.70	339.85	364.15	305.65	309.78
Dec-19	343.90	344.60	373.20	308.00	318.29
Nov-19	312.40	351.00	342.21	299.70	289.15
Oct-19	272.75	317.80	303.19	244.35	252.46
Sep-19	272.00	316.00	302.45	266.95	251.76

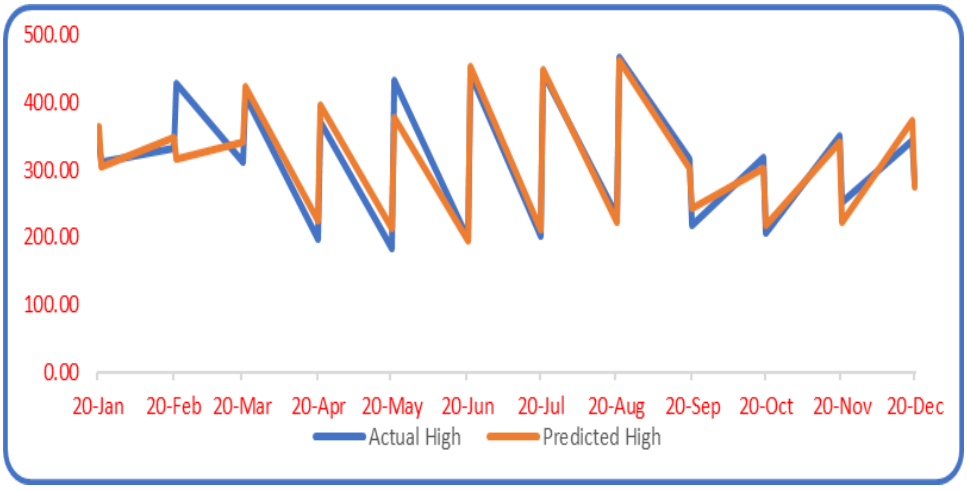

Fig. 16. Actual high and predicted high

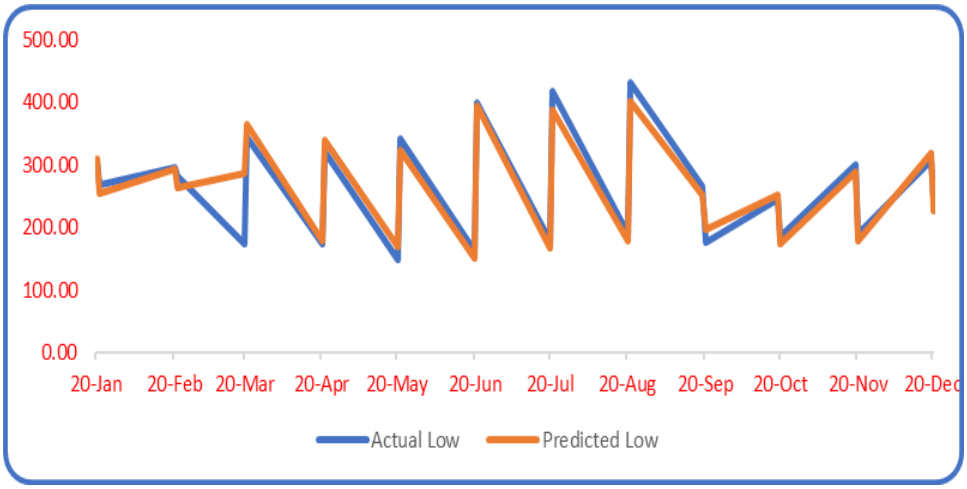

Fig. 17. Actual low and predicted low

3. Results and Discussions

In this paper, consider five different parameters namely date, price, open, high, low and changes in percentage, the related dataset shows in Table 1. Results based on Table 2, the supervised machine learning approaches consider various banking related share parameters and its performance metrics namely of correlation coefficient and MAE indicates negative values. In this case, the combination for these parameters namely price, open, high, low and changes is not suitable for future prediction. The results are shown in Table 2 and Fig. 1 to Fig. 3.

Supervised machine learning approaches consider others SBI share related parameters namely price, open, high, and low. Based on the performance metric like linear regression, random forest, random tree, and REP tree indicates strong positive correlation and MAE, RMSE, RAE and RRSE indicates best performance for using Eq. (1) to Eq. (7). In this case, the combination of these parameters like price, open, high, and low is suitable for future prediction. The numerical illustrations are shown in Table 3 and Fig. 4, to Fig. 6.

Based on Table 4, the supervised machine learning approaches consider other SBI share combination of the parameters namely open, high, and low. In this case, indicate strong positive correlations and other performance metrics also indicate best performance. Table 4 indicates best performance metrics compared to previous metrices. Numerical illustrations are shown in Fig. 7 to Fig. 9.

Every share market prediction mainly considers share open, low, and high. In this case, consider for find the future prediction for consider high and low with the known parameter share open price. Table 5, Fig. 10 to Fig. 12, shows the performance metrices using multiple linear regression. In this case high and low have strong correlations and other performance metrices have the best performance for predicting the future.

Result and discussions based on prediction for high and low using linear regression and its performance metric indicate the best performance compared to other parameters. Numerical illustrations indicate Table 6, Fig. 13 to Fig.15. Table 7 shows actual high and predicted high subsequently actual low and predicted low using linear regression model Eq. (1) and Eq. (2). Numerical illustrations show in Fig. 16 and Fig. 17.

4. Conclusion

In banking related share market prediction having more accuracy using high price and low-price combination only. Supervised machine learning approaches and its performance metrices were also proved in this combination of future prediction,

particularly banking sectors. The same methods are taking consideration into other share sectors like software industry, oil and gas and pharma sectors in further studies.

5. References

1. Han, J.; Pei, J.; Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
2. Bocca, F. F.; Rodrigues, L. H. A. (2016). The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. Computers and electronics in agriculture, **128**, pp. 67-76.
3. Rathod, S. A.; Singh, K. N.; Patil, S. G.; Naik, R. H.; Ray; Meena, V. S. (2018). Modeling and forecasting of oilseed production of India through artificial intelligence techniques. Indian J. Agric. Sci, **88**(1), pp. 22-27.
4. De Lima, G. R. T.; Stephany, S. (2013). A new classification approach for detecting severe weather patterns. Computers & geosciences, **57**, pp. 158-165.
5. Steel, R. G. D.; Torrie, J. H. (1960). Principles and procedures of statistics. Principles and procedures of statistics.
6. Glantz, S. A.; Slinker, B. K.; Neilands, T. B. (1990). Primer of Applied Regression and Analysis of Variance. McGraw-Hill. Inc., New York.
7. Draper, N. R.; Smith, H. (1998). Applied regression analysis. **326**, John Wiley & Sons.
8. Veeramanikandan, V.; Jeyakarthic. M. (2020). Hybridization Of Stdl with Optimal Kernel Extreme Learning Machine (Okelm) Based Short Term Crude Oil Price Forecasting In Commodity Futures Market. International Journal of Scientific & Technology Research, **9**(2), pp. 4029- 4036.
9. Jeyakarthic. M.; Veeramanikandan, V. (2020). Forecasting of commodity future index a hybrid regression model based on support vector machine and grey wolf optimization algorithm. International Journal of Innovative Technology and Exploring Engineering, **9**(2), pp. 2856-2862.
10. Rouf, N.; Malik, M. B.; Arif, T.; Sharma, S.; Singh, S.; Aich, S.; Kim, H. C. (2021). Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions. Electronics, **10**(21), 2717.
11. Manujakshi, B. C.; Kabadi, M. G.; Naik, N. (2022). A Hybrid Stock Price Prediction Model Based on PRE and Deep Neural Network. Data, **7**(5), 51.
12. Rajesh, P.; Karthikeyan, M. (2017). A comparative study of data mining algorithms for decision tree approaches using weka tool. Advances in Natural and Applied Sciences, **11**(9), pp. 230-243.
13. Breiman, L. (2001). Random forests. Machine learning, **45**(1), pp. 5-32.
14. Denil, M.; Matheson, D.; De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. International conference on machine learning, pp. 665-673.
15. Koulinas, G.; Paraschos, P.; Koulouriotis, D. (2021). A machine learning-based framework for data mining and optimization of a production system. Procedia Manufacturing, **55**, pp. 431-438.

Authors Profile

	**M. Vijayakanth,** Started Academic Career at Annamalai University at the year of 2007, presently working as Assistant Professor in Department of Computer Science at ThiruKolanjiappar Government Arts College, Vridhachalam, Tamilnadu-606001. Data Mining are my area specializations.
	**Dr. V. Veeramanikandan,** Started Academic Career at Annamalai University at the year of 2003, presently working as Assistant Professor in Department of Computer Science at ThiruKolanjiappar Government Arts College, Vridhachalam, Tamilnadu-606001. Published 1 Book, 2 Chapters and several Articles at reputed Journals. General Relativity, Prediction Analysis, Business Intelligence, and Intrusion Detection in Ad-hoc Networks are few of my area specializations.