Forecasting of Indian Stock Market Using Time-Series Models

Sourabh Yadav and Nonita Sharma

Abstract In the present era, stock market has become the storyteller of all the financial activities of any country. Therefore, stock market has become the place of high risks, but even then it is attracting the mass because of its high return value. Stock market tells about the economy of any country and has become one of the biggest investment places for the general public. In this manuscript, we present the various forecasting approaches and linear regression algorithm to successfully predict the Bombay Stock Exchange (BSE) SENSEX value with high accuracy. Depending upon the analysis performed, it can be said successfully that linear regression in combination with different mathematical functions produces the best results. This model gives the best output with BSE SENSEX values and gross domestic product (GDP) values as it shows the least *p*-value as 5.382e−10 when compared with other model's *p*-values.

Keywords Stock market · Forecasting · Time series · Univariate analysis · Multivariate analysis · Regression · Linear regression

1 Introduction

Stock market prediction has been an active area of research as it has huge applications in financial domain. Stock market trend analysis is one of the difficult tasks because of the daily ups and downs in prices of the stock. Hence, it is important to build an accurate and precise prediction model for predicting the stock prices. Further, there are various approaches to analyze the stock prices, but the statistical approach for analyzing the prices is one of the most widely used approaches [1]. Furthermore, if time-series approach is used, it will provide the better accuracy and precise prediction

S. Yadav (✉)
Gautam Buddha University, Greater Noida, Uttar Pradesh, India
e-mail: sy9643391664@gmail.com

N. Sharma
Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, Punjab, India
e-mail: nonita@nitj.ac.in

S.-L. Peng et al. (eds.), *Computing and Network Sustainability*, Lecture Notes in Networks and Systems 75, https://doi.org/10.1007/978-981-13-7150-9_43

model [2]. There are various parameters of stock market, and BSE SENSEX is one of them. Moreover, there are many other additional factors which affect the BSE SENSEX, like gross domestic product (GDP), inflation, exchange rates like the value of US dollar in Indian rupee, and many other [3].

This manuscript specifically targets for predicting the BSE SENSEX depending upon the historical values [4] and factors affecting the BSE SENSEX. Initially, univariate analysis or understanding the historical trends in the dataset is performed to provide a model for predicting the stock prices depending upon past values [5]. To increase the accuracy of the results found in univariate analysis, the next step performed is of multivariate analysis. Multivariate analysis involves determining the correlation values among the BSE SENSEX vector and all the factors affecting BSE SENSEX. Depending upon the correlation values, the correlation matrix is prepared to judge highly affecting factor. Moreover, multivariate analysis of the dataset provides a mathematical relation between highly affecting factors and prices of stock. Hence, the next target was to create a mathematical relation between BSE SENSEX values and additional factors affecting the BSE SENSEX. Further, ensemble is created to improve the accuracy and precision of the model.

2 Proposed Method

Proposed method involves creating the ensemble of the various regression techniques applied in the dataset. Ensemble, also known as data combiner, is a data mining approach that converges the strength of multiple models to achieve better accuracy in prediction.

2.1 Univariate Analysis

Step 1 for preparing the prediction model is to create a forecasting model depending upon the previous trends in the dataset or univariate analysis. For performing analysis on any dataset, univariate analysis acts as a basic step. 'Uni' means one, 'variate' here means variable, so one variable analysis is known as univariate analysis. Under this step, the first step is data collection. Data collection is the process of collection of data from all the relevant sources in a systematic fashion that enables one to answer the relevant questions and evaluate outcomes [6]. After collecting the data, data cleaning is the next step. Data cleaning refers to the process of removing invalid data points from the dataset [7]. After data cleaning, the next step is an exploratory analysis of the dataset. For exploratory analysis, data is loaded in the statistical environment for performing the different statistical functions on the dataset. Further, the dataset is converted into time series. This means that data exists over a continuous time interval with equal spacing between every two consecutive measurements. Converting the dataset into time series always proves to be an effective method for the analysis of

any dataset, especially in the stock analysis [8]. The next step involves checking the time series for stationarity which can be done by performing Ljung–Box test and augmented Dickey–Fuller test.

Next step involves testing for stationarity under which ADF test is performed on the dataset. Moreover, the null hypothesis states that large *p*-value indicates non-stationarity and smaller *p*-values indicate stationarity [9]. The next step is the decomposition of the dataset which involves breaking down the dataset into parameters that are observed, trend, seasonal, random [10]. The next step is performing the model estimation. Firstly, auto-correlation function (ACF) plot and partial auto-correlation function (PACF) plot are prepared. These ACF and PACF plots tell about the correlation factor of statistical analysis and in turn help to judge covariance of the dataset. Next step is to build the model, which means deducing that which particular model applies best on our dataset depending upon our statistical results. Different models are:

Autoregressive Integrated Moving Average (ARIMA) Model: ARIMA is a forecasting technique that projects the future values of a series entirely based on its own inertia. Its main application is in the area of short-term forecasting requiring at least 40 historical data points. It offers great flexibility to work upon univariate time series [11].
BoxCox Transformation: BoxCox transformations are generally used to transform non-normally distributed data to become approximately normal.
Exponential Smoothing Forecast: This forecasting method relies on weighted averages of past observations where the most recent observations hold higher weight. This method is suitable for forecasting data with no trend or seasonal pattern.
Mean Forecast: This forecasting method relies on the mean of the historical data.
Naive Forecast: The naive forecasting method which gives an output as ARIMA (0, 1, 0) with a random walk model that is applied to time-series object.
Seasonal Naive Forecast: This forecasting method works almost on the same principles as the naive method, but works better when the data is seasonal.
Neural Network: Neural networks are forecasting methods that are based on simple mathematical models of the brain. They allow complex nonlinear relationships between the response variable and its predictors. This model is very helpful when combined with the statistical computational approach for forecasting of stock market [12].

The model which has the least error or has the higher accuracy will be the best fit model for the dataset. Moreover, error analysis suggests the improvements that can be made in the results in the future [13].

2.2 Multivariate Analysis

Step 2 involves multivariate analysis for improving the results of Step 1. Multivariate analysis is a statistical approach in which dataset is analyzed on the basis of different

factors and the main objective is to prepare a combined model for better performance, analysis, and accuracy. Results from Step 1 can be improved if a relationship analysis is carried out among the dataset vector and factors affecting the dataset. For relationship analysis, the statistical approach used is regression. Regression is the statistical approach which is used to build a model in terms of mathematical equations for determining the relationship among the different factors with the main variable. In regression, one of the variables is known as a predictor variable whose value is carried out by performing different experiments, and another variable is response variable.

The linear regression approach is preferred over other regression approaches as all other regression approaches are built by understanding the working of linear regression [14]. A key requirement for linear regression is linearity among the variables. Moreover, correlation values also help to judge the dependability of any response variable upon the predictor variable. The correlation values have the range of -1 to 1. So, larger the absolute value of the correlation coefficient, more the dependability of variables upon each other and more is the linearity among them. After determining the correlation value, the most influencing factor will be extracted. Furthermore, model fitting is done by applying different mathematical functions like logarithmic function or exponential function, on both response variable and predictor variable, for making model estimation simple and easier. Moreover, instead of passing a single factor, i.e., most influencing factor, one can pass all the factors at the same time as an argument to the regression algorithm. Then, whichever model performs better will be the best fit model. For determining the accuracy, steps will be the same, i.e., summarization of regression model.

2.3 Ensemble Technique

Step 3 involves the building of ensemble for the dataset which involves combining the result from Step 2, i.e., after multivariate analysis, and hence creating an equation which provides the best accuracy. The Ensemble can be built by taking the numerical total of all the factors affecting BSE SENSEX, i.e., GDP, inflation, USD value. In this firstly, the absolute value of the correlation coefficients needs to be more than 0.5 for effective linearity. The correlation coefficient suggests that a linear regression model can be constructed as an experiment to improve the accuracy. For applying regression, open vector is used as the response variable and the total vector as a predictor variable. As a result, there will be a linear equation between the open vector and the total vector which is represented by Eq. 1.

$$\text{Open} = 714.2 * \text{Total} - 27557.8 \tag{1}$$

Table 1 Model estimation for open vector

	ME	RMSE	MAE	MASE	ACF1
ARIMA	−2.808	59.618	48.379	0.7	0.05
BoxCox	0.834	59.938	48.878	0.707	0.06
ETS	0.068	59.931	48.867	0.707	0.06
Meanf	0	59.928	48.865	0.707	0.06
Naïve	−0.18	82.312	64.539	0.934	−0.457
Snaive	0.468	86.49	69.132	1	0.13
Neural network	**−0.022**	16.769	12.939	0.187	−0.034

Table 2 Correlation coefficients

Vector	Correlation coefficients
Inflation	0.4155946
GDP	0.9675431
Exchange rates	0.6650287

3 Results and Discussion

The tool which is used for forecasting is *R*. Various packages related to the various functionalities described in Sect. 3 are included as: forecast package and tseries package. Datasets used in our analysis are BSE SENSEX collected from official Web site of BSE India [15], GDP of India collected from official Web site of World Bank [16], USD prices in rupee collected from official Web site of Reserve Bank of India [17], and inflation collected from official Web site of European Union [18]. BSE SENSEX dataset contains variable that is open. The open variable represents the opening price of the stock market. Results after applying the procedure mentioned are detailed below.

In Step 1, open vector is analyzed by applying different model functions depending upon which error matrix is prepared.

As seen from Table 1, all the models have performed differently. Considering every model, best results are given by neural network model as its mean error value is comparatively low.

Further in Step 2, multivariate analysis is performed to improve the accuracy. For performing multivariate analysis, linearity among the BSE SENSEX value and GDP values, inflation, and exchange rates is determined, respectively. So for that purpose, correlation coefficients are determined among the different vectors with the open vector of the dataset. Table 2 depicts the different correlation values of different vectors with the open vector.

After interpreting Table 2, it is clearly observable that GDP vector is highly correlated with BSE SENSEX open feature.

Table 3 Summary of linear regression model object with log (GDP) vector

Parameters	Values
p-value for intercept	<2e−16
p-value for logarithmic GDP coefficient	5.38e−10
Net *p*-value	5.382e−10
Multiple *R*-squared values	0.9527
Adjusted *R*-squared value	0.6650287

Table 4 Summary of a combined linear regression model with improvements

Parameters	Values
p-value for intercept	2.61e−08
p-value for exp (exp(inflation)) coefficient	0.399
p-value for log (GDP) coefficient	4.11e−06
p-value for USD value coefficient	0.472
Net *p*-value	5.213e−08
Multiple *R*-squared values	0.9606
Adjusted *R*-squared value	0.9499

So, next step is to build a linear regression model between open vector and GDP vector. The linear equation obtained by decreasing the residuals value is given in Eq. 2.

$$\log(\text{Open}) = 1.35712 \log(\text{GDP}) + 9.15080 \tag{2}$$

To check the accuracy of the equation, the regression object can be summarized as given in Table 3.

Then for better comparison among the models, next step is to build a combined model, i.e., using all the factors that affect the BSE SENSEX, i.e., inflation, exchange rates, GDP value, as a predictor variable and response variable will remain the same, i.e., open vector. Further to increase the accuracy, logarithm of open and GDP, exponential of reciprocal of inflation can be used for determining the equation. As a result, the linear equation between open as a response variable and GDP, inflation, USD value as a predictor variable is constructed, which is quoted in Eq. 3.

$$\log(\text{Open}) = 1.318037 \log(\text{GDP}) - 0.208758 e^{e^{\frac{1}{(\text{Inflation})}}} - 0.006184 \text{ USD Value} \tag{3}$$

For checking the accuracy, the regression model object is summarized, and Table 4 depicts the different *p*-values and *R*-squared values.

Next step is to build an ensemble. The ensemble can be built by taking the numerical total of all the factors affecting BSE SENSEX, i.e., GDP, inflation, USD value. In this firstly, the absolute value of the correlation coefficients needs to be more than 0.5

Table 5 Summary of ensemble

Parameters	Values
p-value for intercept	0.011661
p-value for total coefficient	0.000688
Net p-value	0.0006876
Multiple R-squared values	0.6008
Adjusted R-squared value	0.5701

Table 6 Net p-values for all the above regression models

Parameters	Values
Net p-value for open − GDP model	3.828e−09
Net p-value for log (open) − log (GDP) model	5.382e−10
Net p-value for open − GDP + inflation + USD value model	5.966e−07
Net p-value for log (open) − log (GDP) + exp (exp(1/inflation)) + USD value model	5.213e−08
Net p-value for ensemble	0.0006876

for effective linearity. Correlation coefficients when calculated between open vector and total vector, it comes out to be 0.7751132. The correlation coefficient suggests that a linear regression model can be constructed as an experiment to improve the accuracy. For applying regression, open vector is used as the response variable and the total vector as a predictor variable. As a result, there will be a linear equation between the open vector and the total vector which is represented by Eq. 4.

$$\text{Open} = 714.2 * \text{Total} - 27557.8 \tag{4}$$

For checking the accuracy of the result, summarization of regression object is required and Table 5 shows the summary of the model.

The next and final step is to analyze the results of all the above linear regression models and find the best and suitable model for forecasting which can closely predict the value of BSE SENSEX. Table 6 shows all the net p-values of all the above regression models, through which we can easily compare the results.

It is clearly observable that open-GDP model with a mathematical function gives the least p-value when compared with all other models.

4 Conclusion

In this manuscript, there is a research performed on the dataset of BSE SENSEX from January 1997 to January 2016, and the dataset of GDP of India (in trillions), inflation (in %), USD values (in rupees), which is interpreted annually from the year 2001 to

2015. On applying different forecasting models in the beginning, then applying linear regression techniques, it has been found that each and every model results differently and can be analyzed on the basis of mean error and the net p-values of the models. After analyzing the different mean errors of forecasting model in the beginning, it has been concluded that exponential smoothing and neural network give the consistently less mean error, with the small exception in the low vector where the mean error of neural network is comparatively high. Moreover, when linear regression algorithm is applied in the datasets for the improvements, it has been concluded that linear regression model of logarithmic open values of BSE SENSEX and logarithmic GDP values of India gives the best accuracy and precision, from all other quoted models.

References

1. Armstrong JS (2001) Combining forecasts. In: Principles of forecasting. Springer, Boston, MA, pp 417–439
2. Frick RW (1996) The appropriate use of null hypothesis testing. Psychol Method 1(4):379
3. Cleveland WP, Tiao GC (1976) Decomposition of seasonal time series: a model for the census X-11 program. J Am Stat Assoc 71(355):581–587
4. Sharma N, Juneja A (2017) Combining of random forest estimates using LSboost for stock market index prediction. In: 2nd international conference for convergence in technology (I2CT). IEEE
5. 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
6. Rao A, et al (2015) Survey: stock market prediction using statistical computational methodologies and artificial neural networks
7. Cole R (1969) Data errors and forecasting accuracy. In: Economic forecasts and expectations: analysis of forecasting behavior and performance. NBER, pp 47–82
8. Litterman RB (1986) A statistical approach to economic forecasting. J Bus Econ Stat 4(1):1–4
9. Alam P (2016) Factors affecting stock market in India. Spl Int J Prof 3(9):7
10. Montgomery DC, Peck EA, Vining GG (2012) Introduction to linear regression analysis vol. 821. Wiley
11. Sapankevych NI, Sankar R (2009) Time series prediction using support vector machines: a survey. IEEE Comput Intell Mag 4(2)
12. Devers KJ, Frankel RM (2000) Study design in qualitative research–2: sampling and data collection strategies. Educ Health 13(2):263
13. Rahm E, Do HH (2000) Data cleaning: problems and current approaches. IEEE Data Eng Bull 23(4):3–13
14. Angadi MC, Kulkarni AP (2015) Time series data analysis for stock market prediction using data mining techniques with R. Int J Adv Res Comput Sci 6(6)
15. BSE Homepage, http://www.bseindia.com. Last accessed 2018/07/05
16. The Worldwide Inflation Data Homepage, http://www.inflation.eu. Last accessed 2018/07/10
17. The World Bank Homepage, http://www.worldbank.org. Last accessed 2018/07/10
18. The Reserve Bank of India Homepage, https://www.rbi.org.in. Last accessed 2018/07/10