Homogenous Ensemble of Time-Series Models for Indian Stock Market

Sourabh Yadav[1](✉) and Nonita Sharma[2]

[1] Gautam Buddha University, Greater Noida, Uttar Pradesh, India
sy9643391664@gmail.com
[2] Dr B R Ambedkar National Institute of Technology, Jalandhar, Punjab, India
nonita@nitj.ac.in

Abstract. In the present era, Stock Market has become the storyteller of all the financial activity 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 place 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 prepares the best model. 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 of any country is the key factor for determining the country's growth and economy. Stock market is a place where all the public listed companies trades their shares to raise their capital. Looking at the historical trends of the stock market, predicting the stock prices will not be an easily accomplished task. Therefore, predicting the stock market prices will definitely prove to be a great helping hand for those who invest in the stock market. It will help to determine the country's growth and economy for the future, which will assist the higher officials of any country for framing their policies for the development of the nation. Moreover, it will help the general public to understand the trends of the market, and when and how much one can invest for getting the maximum returns. There are several parameters of stock market, and BSE SENSEX is one of them. The BSE SENSEX, also called BSE 30 or SENSEX is a free floated market-weighted index of 30 well established and financially sound companies, listed on Bombay Stock Exchange.

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

A. Mondal et al. (Eds.): BDA 2018, LNCS 11297, pp. 100–114, 2018.
https://doi.org/10.1007/978-3-030-04780-1_7

to analyze the stock prices, but the statistical approach for analyzing the prices is one of the most widely used approach [11]. Statistical Analysis is collecting, exploring and presenting the data for understanding the patterns and trends (if any) present in the dataset. Furthermore, if time series approach is used, it will provide the better accuracy and precise prediction model [16]. Time Series for any dataset is an existence of data over a continuous time interval. Time series analysis is analyzing the time series data for the better understanding of the trends and patterns. 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 [1]. GDP for any country is the final value of the goods and services manufactured within the geographical boundaries of any country during a particular period of time. Inflation is the measure of the increase in prices of goods and services in a country annually. Exchange Rate is defined as the price of a country's currency in terms of another country's currency. If one sees these factors mathematically, these factors are directly proportional to the increase and decrease of the values of stock market prices.

This manuscript specifically targets for predicting the BSE SENSEX depending upon the historical values [17] and factors affecting the BSE SENSEX. Performing the univariate analysis or understanding the historical trends in the dataset, provides a model for predicting the stock prices depending upon past values. The historical data of the past 18 years was analyzed and best fit model is prepared depending upon the mean error of various forecasting models. Various forecasting models when applied in combination with each other and compared simultaneously, gives the best output [3]. Depending upon the results and errors of various forecasting model, error matrix is prepared for better understanding. To increase the accuracy of the results found in univariate analysis, the next target was of multivariate analysis. Hence, the next target is to determine the correlation values among the BSE SENSEX vector and all the factor 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, one ensemble is also prepared, to improve the accuracy and precision of the model. Then, in the end, all the results are compared to find the best model for forecasting. The data used in the analysis is of 18 years for predicting forecasting model on the basis of univariate analysis and is of 15 years for performing multivariate analysis.

2 Problem Statement

The Main objective covered in this manuscript is to predict the BSE SENSEX value accurately and precisely. To achieve this objective, there are some sub-objectives. First sub-objective is to predict the BSE SENSEX value by univariate analysis or by analyzing the historical values and trends in the dataset and obtain a suitable forecasting model which give the least mean error. Second sub-objective is to improve the accuracy of the model by analyzing the factors affecting the BSE SENSEX and performing the multivariate analysis on most affecting factor in which mathematical equation comes out as an output. Third and last sub-objective targets to build an ensemble.

3 Proposed Method

Figure 1 depicts the proposed methodology for building the precise and accurate prediction model.

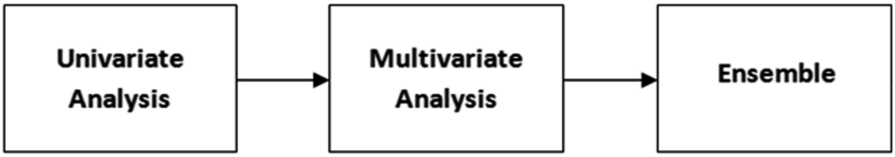

Fig. 1. Proposed methodology

3.1 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. Univariate analysis is one of the simplest forms, for analyzing the dataset in which previous values or historical values of the dataset is used for performing the analysis. ‘Uni’ means one, ‘variate’ here means variable, so one variable analysis is known as univariate analysis. For performing analysis on any dataset, univariate analysis acts as a basic step. 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 [7]. After collecting the data, data cleaning is the next step. Data Cleaning refers to the process of removing invalid data points from the dataset [14]. Cleaning is the process of removing the data points which are disconnected from the effect and assumption which are needed to be isolated. In this process, these particular data points are ignored, and analysis has been conducted on the remaining data. 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 in 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 [2]. The next step involves plotting the time series object for analyzing the components of the time series data i.e. trend, seasonality, stationarity, and heteroskedasticity. Among these components, stationarity is most important. When the mean and variance are constant for a particular dataset, it is said that dataset holds the stationarity. (i.e. their joint distribution does not change over time).

The Plot of the time series will suggest that whether the data is stationary or not, which further suggests that data is volatile or not. If the data is not stationary, then it means there is a large deviation from the mean of the dataset. The data which not stationary, it will be quite unpredictable. Further, for testing the stationarity, different tests are performed like the Ljung-Box Test and Augmented Dickey-Fuller Test.

Next Step involves Testing for Stationarity under which two test are performed on the dataset i.e.

Ljung-Box Test: The Ljung–Box test is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero.

Augmented Dickey-Fuller (ADF) Test: The ADF test is unit root test for stationarity. Unit roots can cause unpredictable results in your time series analysis. The Augmented Dickey-Fuller test can be used with serial correlation. In this lag length is the parameter which is important in finding the meaningful results. In this lag length is the parameter which is important in finding the meaningful results [9].

Moreover, the Null Hypothesis states that large the p-value indicates non-stationarity and smaller p values indicate stationarity [8]. If in ADF test results are not in favor i.e. p-value comes out to be relatively high, then there is a need to do some further visual inspection, otherwise, next step i.e. Decomposition of the dataset can be skipped. So, if test depicts the high p-value then next step is decomposition of the dataset. This involves breaking down the dataset into parameters that are, Observed, Trend, Seasonal, Random [5]. When the Seasonal vector is plotted, it gives us indication towards stationarity or non-stationarity. If the data is stationary, the first phase of model estimation can be skipped. Basically, Model estimation comprises of two phases i.e. in the first phase, non-stationary data is transformed into stationary data and second phase, building a model. So the next step is 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. Moreover, it helps to judge Co-variance of the dataset. When there is large autocorrelation within our lagged values, then, in that case, there is a need to take the difference of time series object in order to transform the series into a stationary series. The Difference of the series means calculating the differences between all consecutive values of a vector. This will helps to stabilize the mean, thereby making the time-series object stationary. Next step is to plot the transformed time series. This plot will suggest that whether the series is now stationary or not. To confirm the stationarity, ACF and PACF plots are again plotted for the differenced time series which clears the doubt about the stationarity. Further stationarity can be tested by different tests like the Ljung-Box Test, Augmented Dickey-Fuller Test, which will give the p-value very small in comparison to the previous p-values, which again proves the stationarity. Next job is the Building of 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 upon its own inertia. Its main application is in the area of short-term forecasting requiring at least 40 historical data points. It works best when the data exhibits a stable or consistent pattern over time with a minimum amount of outliers. It is the preferred choice because of its simplicity and wide acceptability. It offers great flexibility to work upon univariate time series [12].

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 [15].

The model which have the least error or have 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 [6].

3.2 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. Many times, univariate analysis is preferred because multivariate analysis results are difficult to interpret. For multivariate analysis, there are enough number of techniques, so depending upon the type of datasets, a particular technique is followed. Multivariate analysis can be performed by just analyzing the trend of all the factors which can have a great impact on the main dataset. Multivariate analysis is performed with the factors which have great influence on the dependent variable. The Dependent variable is a variable which needs to be detected after the analysis. Principle Component Regression (PCR) technique is the most commonly followed technique for the multivariate analysis. This technique is simply based upon Principle Component Analysis. PCR focuses to reduce the Dimensionality of the datasets. Moreover, it avoids the multicollinearity among the predictor variables. Results from Step 1 can be improved if a relationship analysis is carried out among the dataset vectors 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 Predictor variable whose value is to carried out by performing different experiments, and another variable is Response Variable. Response Variable is a variable whose value is procured by Predictor Variable. Generally, there are seven types of Regression available which are listed below:

Linear Regression: It is the most commonly know regression modeling technique. In this type of modeling technique, there can more than one independent variable which can be either discrete or continuous, but dependent variable must be continuous and the nature of the line of regression should be linear [13].
Logistic Regression: It is the type of regression which focuses to determine the probability of the event i.e. either success or failure [10]. Logistic Regression must be used or preferred when the dependent variable is in binary form i.e. 0 or 1, True or False, Yes or No.
Polynomial Regression: It is the type of regression model in which regression equation is of the form polynomial i.e. independent variable has the power more than 1. In this best-fit regression line is not particularly a straight line.
Stepwise Regression: It is the type of regression in which multiple independent variables are required. The selection of the predictor variable is done automatically. There is no intervention of humans. Its basic aim is to produce a best fit model with the highest possible accuracy.
Ridge Regression: It is the type of regression model which applied when independent variables a high absolute correlation value or have multiple collinearities. In this alpha value is set to be a 0.
Lasso Regression: It is the type of regression model which similar to ridge regression model but just uses absolute values instead of squares in penalty function. Moreover, the alpha value is set to be 1 in this model.
ElasticNet Regression Model: It is the hybrid model of ridge and lasso regression model. In this alpha value is set as 0.5.

The Linear Regression approach is preferred over other regression approaches as all other regression approaches are build by understanding the working of linear regression [21]. 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 models 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.

3.3 Ensemble Technique

Step 3 involves the building of ensemble for 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. Basically Ensemble means combining multiple algorithms for improving the accuracy of the model. Ensemble method is one of the most influential developments in the field of data mining. They combine the multiple models into one, by extracting most accurate models from all those multiple models. Ensembling is performed depending upon the dataset. The necessary steps to perform the technique is outlined in below:

Algorithm: Ensemble of various Regression Methods

```
1.  Perform univariate analysis and find the model that best fits the data.
2.  Determine the predictor variables $L_1, L_2, \dots\dots.. L_k$ and response variable $R_i$
3.  Evaluate the correlation coefficient $Cor(L_{j(j\in 1..k)}, R_i)$ between the all the predictor variables and response variables
4.  Decrease the residuals by taking the logarithm.
5.  If $Cor\left(L_{j(j\in 1..k)}, R_i\right) > 0.5$, the variable will be included in the ensemble otherwise it is not included in the ensemble.
6.  Find the maximum value of the $\max(Cor(L_{j(j\in 1..k)}, R_i)$ and find the mathematical equation of response variable $R_i$, with the predictor variable depending upon maximum correlation value, $\max(Cor(L_{j(j\in 1..k)}, R_i)$, as E1.
7.  Using the variables obtained in step 5, calculate the numeric sum of all the variables with each other and store it in another variable as $B = \sum L_{j(j\in 1..k)}$
8.  Using the variable B, obtained in step 7, calculate the mathematical equation between variable B and $R_i$. model obtained in step 1, perform multivariate analysis between $L_{j(j\in 1..k)}$ and $R_i$ as E2.
9.  Compare the equations E1 and E2 obtained from step 6 and step 8.
10. If E1 fits to the data more accurately and precisely, then model obtained in step 6 is best fit else model obtained in step 8 best fit
```

All the prediction models can be analyzed depending on its accuracy and precision. The model with more accuracy and precision will be the best-fitted model for our dataset.

4 Results and Discussions

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, tseries package. Datasets used in our analysis are BSE SENSEX, GDP of India, USD Prices in Rupee, and Inflation and the sources for these datasets are mentioned in Table 1.

BSE SENSEX dataset contains four different variables that are open, high, low, and close. The open variable represents the opening price of the stock market, high variable

Table 1. Sources of dataset

Dataset	Source
BSE SENSEX	Official website of BSE India [4]
GDP	Official website World Bank [20]
USD Prices in Rupee	Official website of Reserve Bank of India [18]
Inflation	Official website of European Union [19]

represents the highest price of the stock prices, low variable represents the lowest price of the stock prices, and close variable represents the closing price of stock prices. Results after applying the procedure mentioned are detailed below: In step 1 where univariate analysis been performed, after cleaning the data, which was the initial step, time series object was created for a vector of the dataset.

After the time series object was plotted in Fig. 2, it suggests to work upon different component like Trend, Seasonality, Stationarity, Heteroskedasticity. For testing the stationarity of the dataset, different tests were performed upon dataset like Augmented Dickey-Fuller Test and Ljung-Box Test.

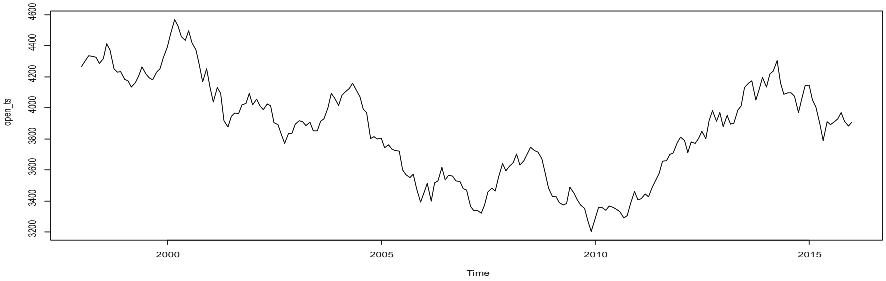

Fig. 2. Plot of time series object

The values obtained from different stationarity test are mentioned in Tables 2 and 3 which further, suggests decomposing the dataset. After analyzing the plot of decomposition, nonstationarity of data is confirmed by analyzing the ACF and PACF plots. Analyzing the ACF and PACF plot suggests that, the difference of data points in the dataset is required to transform the series into stationary series. After taking the difference of the series again above quoted tests are applied on the series to cross-check its stationarity whose results are quoted in Tables 4 and 5.

From both the above Tables 4 and 5, it is clear that series has been transformed into a stationary series successfully as p-value is less 0.05. Once the series has become the stationary series, different model functions are applied on the series and depending upon their errors, best-fit model is selected. In the dataset, 4 vectors are separately analyzed by applying different model functions depending upon which error matrix is prepared.

Table 2. Augmented Dickey-Fuller Test

Parameters	Values
Data	open_ts
Dickey-Fuller	−1.265
Lag order	5
p-value	0.8841
Alternative hypothesis	Stationary

Table 3. Ljung-Box Test

Parameters	Values
Data	open_ts
X-Squared	2580.9
df	20
p-value	2.2e−16

Table 4. Augmented Dickey-Fuller Test

Data	z
Dickey-Fuller	−6.369
Lag order	5
p-value	0.01
Alternative hypothesis	Stationary

Table 5. Ljung-Box Test

Parameters	Values
Data	z
X-Squared	2580.9
df	20
p-value	<2.2e−16

As seen from the Tables 6, 7, 8, and 9 all the models have performed differently. Considering every model, best results are given by Exponential Smoothing Model as its Mean Error value is consistent for all 4 vectors i.e. Open, High, Low, Close. Moreover Neural Network model also works good but with the exception that the Low vector of dataset is showing relatively high Mean Error compared to other 3 vectors.

Further in step 2 multivariate analysis is performed to improve the accuracy. For performing multivariate analysis datasets like GDP dataset, Inflation dataset, Exchange Rates dataset (that can be the values of USD prices in terms of Indian Rupee) and open vector of BSE SENSEX are collected annually. Then the next step is to determine the linearity among the BSE SENSEX value and GDP values, Inflation, and Exchange

Table 6. 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
Naive	−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 7. Model estimation for high vector

	ME	RMSE	MAE	MASE	ACF1
ARIMA	−2.212	56.043	44.505	0.675	0.132
BoxCox	−0.788	55.759	44.003	0.667	0.02
ETS	**0.027**	56.162	44.752	0.679	0.127
Meanf	0	56.16	44.751	0.679	0.127
Naive	−0.521	74.099	60.386	0.916	−0.43
Snaive	−0.641	83.238	65.926	1	0.204
Neural network	**−0.097**	13.049	13.765	0.209	−0.033

Table 8. Model estimation for low vector

	ME	RMSE	MAE	MASE	ACF1
ARIMA	−1.73	55.929	44.503	0.693	0.218
BoxCox	−0.053	53.792	42.937	0.669	−0.015
ETS	**0.022**	56.07	44.522	0.694	0.22
Meanf	0	56.067	44.52	0.694	0.22
Naive	−0.256	70.114	56.161	0.875	−0.309
Snaive	−0.391	79.903	64.19	1	0.272
Neural network	**0.27**	48.532	38.182	0.595	0.024

Table 9. Model estimation for close vector

	ME	RMSE	MAE	MASE	ACF1
ARIMA	−1.298	57.414	46.472	0.709	0.126
BoxCox	0.188	54.931	43.683	0.666	−0.002
ETS	**0.025**	57.547	46.314	0.706	0.129
Meanf	0	57.544	46.311	0.706	0.129
Naive	−0.307	76.075	58.987	0.9	−0.42
Snaive	−0.429	82.836	65.571	1	0.23
Neural network	**−0.031**	8.715	6.272	0.096	−0.167

Rates, respectively. So for that purpose, correlation coefficients are determined among the different vectors with the open vector of the dataset. Table 9 depicts the different correlation values of different vectors with the open vector:

After interpreting the Table 10, it is clearly observable that GDP vector is highly correlated with BSE SENSEX Open feature. So next step is to build a linear regression model between Open vector and GDP vector. Here predictor variable is the GDP vector and response variable is the open vector. As a result, there will be a linear equation between GDP and Open vector.

$$Open = 12793.6 * GDP - 2599 \tag{1}$$

To check the accuracy of the equation, the regression object can be summarized. Now p-values will help to judge whether the model is accurately fitted or not. The accuracy will be high if each p-value in the summary is less than or equals to 0.05 approximately and R-squared values are also above 0.9 (Table 11).

Table 10. Correlation coefficients

Vector	Correlation coefficients
Inflation	0.4155946
GDP	0.9675431
Exchange rates	0.6650287

Table 11. Summary of linear regression model object with GDP vector

Parameters	Values
p-value for intercept	0.064
p-value for GDP coefficients	3.83e−09
Net p-value	3.828e−09
Multiple R-squared value	0.9361
Adjusted R-squared value	0.9312

In order to increase the accuracy of the linear regression model, some mathematical functions are implemented like logarithm is applied on both the vector i.e. GDP and Open vector. Applying mathematical function will decrease the residuals value and will also rectify the other problems which would be decreasing the accuracy of the model.

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

Summary of Regression model is in Table 12.

Table 12. 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 value	0.9527
Adjusted R-squared value	0.6650287

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 same i.e. Open vector. 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.

$$Open = 13049.797 * GDP - 126.588 * Inflation + 9.607 * USD\ Value - 2497.615 \tag{3}$$

For checking the accuracy regression object is summarized, and p-values and R-squared values can be extracted which help to interpret the accuracy. p-values and R-squared values are quoted in Table 13.

Table 13. Summary of combined linear regression model object

Parameters	Values
p-value for intercept	0.640
p-value for Inflation coefficient	0.567
p-value for GDP coefficient	5.18e−06
p-value for USD value coefficient	0.937
Net p-value	5.966e−07
Multiple R-Squared Value	0.9386
Adjusted R-Squared Value	0.9218

Further to increase the accuracy, logarithm of open and GDP, exponential of exponential of reciprocal of Inflation can be used for determining the equation.

$$\log(Open) = 1.318037\log(GDP) - 0.208758e^{e^{\frac{1}{(Inflation)}}} - 0.006184USD\ Value \tag{4}$$

For checking the accuracy, the regression model object is summarized, and Table 14 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 need to be more than 0.5 for effective linearity. Correlation coefficients when calculated between open vector and total vector, it comes out to be 0.7751132. The correlation coefficient

Table 14. Summary of 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 value	0.9606
Adjusted R-squared value	0.9499

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 predictor variable. As a result, there will be a linear equation between the open vector and the total vector which is represented by Eq. 5.

$$Open = 714.2 * Total - 27557.8 \tag{5}$$

For checking the accuracy of the result, summarization of regression object is required and Table 15 shows the summary of the model.

Table 15. 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 value	0.6008
Adjusted R-squared value	0.5701

The next and final step is to analyze the results of all the above linear regression model and find the best and suitable model for forecasting which can closely predict the value of BSE SENSEX. Table 16 show all the net p-values of all the above regression models, through which we can easily compare the results.

Table 16. 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 values Model	5.966e−07
Net p-value for log(Open) - log(GDP) + exp(exp(1/Inflation)) + USD values Model	5.213e−08
Net p-value for ensemble	0.0006876

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

5 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 error of forecasting model in the beginning, it has been concluded that Exponential Smoothing and Neural Network gives 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 on 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. Alam, P.: Factors affecting stock market in India. Splint Int. J. Prof. **3**(9), 7 (2016)
2. Angadi, M.C., Kulkarni, A.P.: Time series data analysis for stock market prediction using data mining techniques with R. Int. J. Adv. Res. Comput. Sci. **6**(6) (2015)
3. Armstrong, J.S.: Combining Forecasts. In: Armstrong, J.S. (ed.) Principles of Forecasting. International Series in Operations Research & Management Science, vol. 30. Springer, Boston (2001)
4. BSE Homepage. http://www.bseindia.com. Accessed 05 July 2018
5. Cleveland, W.P., Tiao, G.C.: Decomposition of seasonal time series: a model for the Census X-11 program. J. Am. Stat. Assoc. **71**(355), 581–587 (1976)
6. Cole, R.: Data errors and forecasting accuracy. In: Economic forecasts and expectations: analysis of forecasting behavior and performance, pp. 47–82. NBER (1969)
7. Devers, K.J., Frankel, R.M.: Study design in qualitative research–2: sampling and data collection strategies. Educ. Health **13**(2), 263 (2000)
8. Frick, R.W.: The appropriate use of null hypothesis testing. Psychol. Methods **1**(4), 379 (1996)
9. Hall, A.: Testing for a unit root in time series with pretest data-based model selection. J. Bus. Econ. Stat. **12**(4), 461–470 (1994)
10. Larsen, K., et al.: Interpreting parameters in the logistic regression model with random effects. Biometrics **56**(3), 909–914 (2000)
11. Litterman, R.B.: A statistical approach to economic forecasting. J. Bus. Econ. Stat. **4**(1), 1–4 (1986)
12. Mondal, P., Shit, L., Goswami, S.: Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. Int. J. Comput. Sci. Eng. Appl. **4**(2), 13 (2014)
13. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to Linear Regression Analysis, vol. 821. Wiley, New York (2012)

14. Rahm, E., Do, H.H.: Data cleaning: problems and current approaches. IEEE Data Eng. Bull. **23**(4), 3–13 (2000)
15. Rao, A., et al. Survey: Stock Market Prediction Using Statistical Computational Methodologies and Artificial Neural Networks (2015)
16. Sapankevych, N.I., Sankar, R.: Time series prediction using support vector machines: a survey. IEEE Comput. Intell. Mag. **4**(2), 24–38 (2009)
17. Sharma, N., Juneja, A.: Combining of random forest estimates using LSboost for stock market index prediction. In: 2017 2nd International Conference for Convergence in Technology (I2CT). IEEE (2017)
18. The Reserve Bank of India Homepage. https://www.rbi.org.in. Accessed 10 July 2018
19. The Worldwide Inflation Data Homepage. http://www.inflation.eu. Accessed 10 July 2018
20. The World Bank Homepage. http://www.worldbank.org. Accessed 10 July 2018
21. Weisberg, S.: Applied Linear Regression, vol. 528. Wiley, New York (2005)