A STUDY OF RELATIONSHIP BETWEEN INDIAN AND ASIAN STOCK MARKETS USING MACHINE LEARNING TECHNIQUES

[1]Manjula KA, [2]Karthikeyan P

[1]Assistant Professor, [2]Assistant Professor

[1]Department of Computer Science, University of Calicut, India.

[2]Department of Management Studies, Kannur University, India.

Abstract: Stock market indices help to identify the broad trends in the markets and is important for the investor to know the index value to earn profit. Prediction of stock prices has got a very significant role in today's economy, since a reasonably accurate prediction has the possibility to yield high financial benefits and hedge against market risks. In this work, Nifty 50, the index for the National Stock Exchange of India Ltd. (NSE) is taken to represent Indian stock market and the influence of selected Asian stock market indices on Nifty is studied. The Asian stock market indices considered for this study are Straits Times Index (FTSE) representing Singapore's stock market index, HangSeng representing Hongkong's stock market index, KOSPI representing South Korea's stock market index, Nikkei225 representing Japan's stock market index, Shanghai Composite Index (SSE) representing China's stock market index, and TSEC representing Taiwan's stock market index. The correlation and regression of Nifty with these selected asian stock indices were analysed with the help of machine learning techniques Multiple linear regression and Gradient boosting regression. It is found that there exists a relationship between NSE and the selected Asian stock markets and these Asian markets could be used for predicting changes in Indian stock market. This study suggests that the gradient boosting regression method yields better results when compared to linear regression method in prediction of Nifty with respect to these selected asian stock market indices.

I.INTRODUCTION

A strong financial market with broad participation is essential for a developed economy. A developing economy like India which has shown high growth potential in the past decades requires large capital investments by both government and corporates. India needs to raise resources for companies to satisfy the capital needs of the economy and also to ensure that the benefits of growth reach the bottom of the socio-economic pyramid. In this regard a capital market all over the world has played a big role in development of the economy of that country. This is due to the fact that capital markets bring together investors who have funds with them and corporates who need funds. This transfer of funds from those who have surplus funds to those who are deficit in funds are done efficiently by capital markets in developed countries and some of the developing countries. Capital markets consist of various markets like share market, market for corporate bonds, banking, insurance etc. Among these the most dominant market is share market.

A stock market or a share market is a designated market for trading various kinds of securities in a controlled, secure and managed environment. It is a place where buyers and sellers of stocks come together, physically or virtually. The stock market brings together hundreds of thousands of market participants who wish to buy and sell shares, and it ensures fair pricing practices and transparency in transactions. The modern day computer-aided stock markets operate electronically whereas earlier stock markets used to issue and deal with paper made share certificates. India has one of the most technically advanced stock exchanges in the world. The whole system is computerised and order-driven, the order placed by an investor is automatically matched with the best limit order. This system provides more transparency as it shows all buy and sell orders. The Indian stock market mainly functions on two major stock exchanges, the BSE (Bombay Stock Exchange) and NSE (National Stock Exchange). In terms of market capitalization, BSE and NSE have a place in top five stock exchanges of developing economies of the world. The Bombay Stock Exchange is one of the oldest exchanges across the world, while the National Stock Exchange is among the best in terms of sophistication and advancement of technology.

The Indian stock market scene really picked up after the opening up of the economy in the early nineties. The reforms in the Indian economy and the resulting globalization and free movements of capital across boundaries of nation have integrated financial market worldwide. Also the technological innovations have improved market assimilation. These developments in Indian stock market has also resulted in increased volatility in this market. The markets are seen to be turbulent and on roller coaster rides at times and these volatilities cannot at all times be attributed to events in Indian economy. So in order to understand the movements in stock markets in India it would be good if the trends in this market is compared to the trends in the other global markets and study the relationship between various markets.

This study intends to do a comparative analysis of the Indian Stock Market Nifty, the index for the National Stock Exchange of India Ltd (NSE) with respect to selected Asian counterparts - Straits Times Index(FTSE) representing Singapore's stock market index, HangSeng representing Hongkong's stock market index, KOSPI representing South Korea's stock market index, Nikkei225 representing Japan's stock market index, Shanghai Composite Index (SSE) representing China's stock market index, and TSEC representing Taiwan's stock market index. The objective of this study is to understand the relationship between NSE and these selected Asian stock markets and to find out if these Asian markets could be used for predicting changes in Indian stock market.

1.1.Stock Market Index

A stock index, or stock market index, is an index that measures a stock market, or a subset of the stock market, that helps investors to compare current price levels with past prices to calculate market performance. It is computed from the prices of selected stocks by calculating a weighted arithmetic mean. Stock market indices play a very significant role nowadays. They help to identify the broad trends in the markets and is important for the investor to know the index value to earn profit. There are many global issues that affect the indices of different stock markets. Understanding the correlation between Indian stock market and other stock markets will help the investor to analyse the position of the Indian stock market. In this study Nifty, the index for NSE is taken to represent Indian stock market and the influence of selected Asian stock market indices on Nifty is studied.

1.1.1.Nifty

The NIFTY 50 is the flagship index on the National Stock Exchange of India Ltd. (NSE). The Nifty, just like BSE benchmark Sensex, is today used for benchmarking portfolios and returns of mutual fund schemes and launching index funds. The Nifty index was launched on April 22, 1996, with a base value of 1,000 counted from November 3, 1995. Live Nifty quotes are available on NSEIndia.com, ETMarkets.com and numerous other web platforms and TV channels at any point of time. It is supposed to reflect the health of the listed universe of Indian companies. Fifty leading companies from a variety of sectors constitute the Nifty index.

1.1.2.Nikkei 225

The Nikkei 225 more commonly called the Nikkei, the Nikkei index, or the Nikkei Stock Average is a stock market index for the Tokyo Stock Exchange (TSE). It has been calculated daily by the Nihon Keizai Shimbun (Nikkei) newspaper since 1950. It is a price-weighted index with the unit yen, and the components are reviewed once a year. Currently, Nikkei is the most widely quoted average of Japanese equities.

1.1.3.Shanghai Composite Index (SSE)

The SSE Composite Indexis a stock market index of all stocks (A shares and B shares) that are traded at the Shanghai Stock Exchange. SSE Indices are all calculated using a Paasche weighted composite price index formula. The base day for SSE Composite Index is December 19, 1990, and the base period is the total market capitalization of all stocks of that day.

1.1.4.HangSeng

It is a weighted stock market index in Hong Kong and is basically used to record and observe daily variation in the prices of the largest companies of the Hong Kong equity market. In Hong Kong, this is the main indicator of the overall market performance in Hong Kong. The 48 component companies of HangSeng represent about 60% of market capitalization of the Hong Kong Stock Exchange. It was started on November 24, 1969, and HangSeng Indices Company Limited is currently maintaining and compiling the index.

1.1.5.KOSPI

The Korea Composite Stock Price Index or KOSPI was launched in 1983 with the base value of 100 as of January 4, 1980. KOSPI is the major stock market index of South Korea. The index represents all common stocks traded on the Korea Exchange. The index calculation is based on market capitalization method.

1.1.6.FTSE

The FTSE Straits Times Index (STI) is a benchmark index for the Singapore equity market. It consists of 30 companies listed on the Singapore stock exchange. It is calculated by Singapore Press Holdings, FTSE and Singapore Exchange. STI has been replaced from STII (Straits Times Industrials Index) when there was a sectoral reclassification of the companies listed in the Singapore Exchange and resulted in the removal of industrial category.

1.1.7.TSEC

TSEC is the stock market index comprised of companies traded on the Taiwan Stock Exchange (TWSE). The TSEC weighted index is made up of all the stocks in the Taiwan Stock Exchange and each is given a weight based on its market capitalization.

1.2.Machine Learning

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and is considered as a subset of artificial intelligence. Machine learning algorithms works by building a mathematical model based on sample data, known as training data, in order to make predictions without being explicitly programmed for the same. Machine learning algorithms are used in a wide variety of applications like email filtering and computer vision, where it is difficult to develop conventional algorithms. Linear regression and Gradient boosting regression techniques are two popular machine learning techniques used for prediction problems.

This study is aimed to understand the influence of Asian stock markets on Indian stock markets and also looks at the methods for predicting the value of Nifty based on these asian market indices. Prediction is always a hot research topic in the domain of stock marks and for this, regression is one of the widely used process. Regression analysis is used extensively in trading, risk management and economics. Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor) and is used for forecasting, time series modelling and finding the causal effect relationship between the variables[1]. Regression helps to predict continuous value outputs and identify the relationship between the given input and output due to this predictive capability. Its importance rises every day with the availability of large amounts of data and increased awareness of the practical value of data.

Linear regression is considered to be a simple regression technique and it was developed in the field of statistics as a model for understanding the relationship between input and output numerical variables, and has been borrowed by machine learning for its prediction tasks. In simple linear regression, to predict the value of the dependent variable "y", the value of a single independent variable "x" is used. Multiple regression is an extension of simple linear regression in which more than one independent variable (x) is used to predict a single dependent variable (y). The predicted value of y is a linear transformation of the x variables such that the sum of squared deviations of the observed and predicted y is a minimum. The difference is that in multiple linear regression, multiple independent variables (x_1, x_2, …, x_n) are used to predict "y" instead of just one independent variable "x". For n independent variables, the data points (x_1, x_2, x_3 …, x_n, y) exist in a n + 1dimensional space. The linear model, which is n dimensional can be represented by the n + 1 coefficients β_0, β_1, …, β_n so that y is approximated by the equation $y = \beta_0 + \beta_1*x_1 + \beta_2*x_2 + \ldots + \beta_n*x_n$. Linear regression is the simplest machine learning algorithm available and most popular technique for predicting a continuous variable and this is based on the assumption that there exists a linear relationship between the outcome and the predictor variables. It is both a statistical algorithm and a machine learning algorithm.

Gradient boosting is another machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees[2]. The common ensemble techniques like random forests rely on simple averaging of models in the ensemble. The family of boosting methods is based on a different, constructive strategy of ensemble formation. The main idea of boosting is to add new models to the ensemble sequentially. At each particular iteration, a new weak, base-learner model is trained with respect to the error of the whole ensemble learnt so far[3]. Gradient boosting is considered as an effective machine learning algorithm.

Since python is a powerful programming platform with capabilities for providing machine learning libraries for regression, python is used for programming in this study. We have applied machine learning algorithms Multiple Linear regression and Gradient boosting algorithms implemented in python language for analyzing the data of stock market indices corresponding to Nifty, FTSE, HangSeng, Kospi, Nikkei, SSE and TSEC during the period 2007-2016.The correlation of Nifty with above indices is studied here and efficient prediction method for Nifty with respect to these six indices is developed here.

In this research paper, the whole work is arranged in different sections as follows. Section II discusses the research works related to this domain. Section III discusses the proposed methods for analysing the relationships and prediction of stock indices. Section IV discusses results of experiments and the final section presents conclusions.

II.REVIEW OF LITERATURE

In economic literature, the issues related to stock market integration and co-movements of stock prices across countries have received considerable attention. Vigilant assessment of global stock market movements in recent years suggests that there exists a substantial degree of interdependence among national stock markets [4].As a fundamental part of capital market, stock investment has crucial importance in optimizing capital allocation, funding as well as increasing the value of assets. The prediction of stock price is of high practical significance for investors due to the high income and high-risk characteristics of stock investments. But it is to be noted that stock price is usually a random fluctuation which is affected by speculative factors and often it is almost impossible to make accurate predictions[5]. Stock prediction is a very complex task and various factors should be considered for forecasting the market accurately and efficiently [6]. Forecasting the direction of future stock prices is a widely studied topic in many fields including trading, finance, statistics and computer science. The motivation for which is naturally to predict the direction of future prices such that stocks can be bought and sold at profitable positions. Simple sentiment analysis is found to be useful in predicting the Indonesian stock market and linear regression method is found to be effective in predicting the company stock price[7]. It is also found that linear regression can be applied to predict stock market indicators using Twitter data as exogenous input and the results shows that daily number of tweets is correlated with certain stock market indicators at three different levels, from the stock market to industry sector and individual company stocks[8]. Bankruptcy prediction has been a subject of interest for almost a century and it still ranks high among hottest topics in economics. The difficulty in predicting financial distress is to develop a predictive model that combines various econometric measures to foresee a financial condition of a firm. It is seen that a prediction model that utilizes Extreme Gradient Boosting for learning an ensemble of decision trees is efficient in bankruptcy prediction[9].In another research work, it is shown that by applying linear regression for predicting behavior of S&P 500 index, the prediction method has a good performance in comparison to real volumes and the stockholders can invest confidentially based on the same[10].

III.DATA AND METHODOLOGY

In this research paper, two regression models for prediction of Indian stock market index Nifty are built using machine learning techniques and a comparison is performed. The two machine learning methods considered here for performing regression are linear regression and gradient boosting regression. The data for this study was taken from the website of yahoo finance. The data set is composed of different stock market indices. The stock market indices we have considered are Indian stock market index– Nifty 50,

Singapore stock market index - FTSE Straits Times Index (STI), Hongkong stock market index - Hang Seng , South Korean stock market index - KOSPI,Japan stock market index - Nikkei225,China stock market index - Shanghai Composite Index (SSE) and Taiwan stock market index(TSEC) - Taiwan Capitalization Weighted Stock Index (TAIEX). The daily values of all these stock market indices for the period January 2007 to December 2016 are collected and necessary pre-processing is done on the data to ensure the accuracy of data. The dataset with 2064 records with respect to each stock market index is used for analysis with machine learning algorithms.

Linear regression is one of the most widely known modelling techniques used in predictive modelling and the most famous regression technique in Machine Learning. This machine learning algorithm is based on supervised learning and it performs a regression task. Linear regression analysis fits a straight line to given data in order to capture the linear relationship between the data elements. This regression technique models the relationship between two variables and the linear regression equation to compute regression line has the form $Y = a + bX$, where Y is the dependent variable, X is the independent variable, b is the slope of the line and a is the y-intercept ie., the value of y when x = 0. The following equations can be used to find a and b.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Multiple regression, or multiple linear regression, is an extension of linear regression which explains the relationship between multiple independent or predictor variables and one dependent or criterion variable. In simple linear regression, there is one independent and one dependent variable, but in multiple regression there is more than one independent variable and one dependent variable. The general mathematical equation for multiple regression can be represented as follows.
$y = a + b_1x_1 + b_2x_2 + ...b_nx_n$. where y is the dependent variable, a, b_1, b_2...b_n are the coefficients and x_1, x_2, ...x_n are independent variables.

The merit of Linear regression is that it is simple and easy to understand. Linear Regression is simple to implement and it is easier to interpret the output coefficients. When it is known that the relationship between the independent and dependent variable is a linear relationship, this algorithm is the best to use as it is less complex when compared to other algorithms. Linear regression assumes a linear relationship between dependent and independent variables and also independence between predictor attributes. One demerit of Linear Regression is that it is susceptible to over-fitting. Another issue is that the Linear regression's performance is poor when there are non-linear relationships. Though simple to operate, linear regression is not able to accommodate more complex patterns.

Because of the limitations of the linear regression technique, another technique for implementing regression by using regression trees is also included in this study. Regression trees or decision trees learn in a hierarchical fashion by repeatedly splitting the dataset into separate branches that maximize the information gain of each split. This branching structure makes it possible for the regression trees to naturally learn non-linear relationships. These are also known as ensembles. Gradient Boosted Trees (GBM), the second technique considered in this study, is one such ensemble method of combining predictions from many individual trees. The merit of decision trees is that they can learn non-linear relationships, and are more robust to outliers. Also, the performance of Ensembles is observed to be good in many prediction studies. Gradient boosting decision tree is considered to be one of the most powerful techniques for building predictive models. This is a useful machine learning technique for regression and classification problems. Gradient boosting regression is used to predict continuous values like car price, while Gradient Boosting Classification is used for predicting classes like whether a plant belongs to a particular species. Gradient boosting regression produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Adding up a number of sub functions to create a composite function that models given data points is called additive modelling. Gradient boosting technique applies additive modelling to result in a good model starting from an approximate model. When more simple models are introduced, the overall model becomes stronger eventually and these simple models are called weak models or weak learners in boosting. Boosting constructs and adds weak models in a stage-wise fashion, one after the other, each one chosen to improve the overall performance of model.

To construct a boosted regression model, initially a simple model $f_0(x)$ is considered to predict an initial approximation of y given feature vector x. Then the overall model $F_M(x)$ is reached   by adding one or more tweaks $\Delta_m(x)$. When F_M is found to be good enough, no more weak models are added.

$$\hat{y} = f_0(\mathbf{x}) + \Delta_1(\mathbf{x}) + \Delta_2(\mathbf{x}) + ... + \Delta_M(\mathbf{x})$$
$$= f_0(\mathbf{x}) + \sum_{m=1}^{M} \Delta_m(\mathbf{x})$$
$$= F_M(\mathbf{x})$$

The demerit of these types of regression trees is that they may be prone to overfitting because of the reason that they can keep branching until they memorize the training data.

Since, Regression analysis helps us as a technique for using data to identify relationships among variables and to use these relationships to make predictions. We have applied regression in this study. Based on the assumption that the outcome we are predicting depends linearly on the information used to make the prediction we have used linear regression first. We have multiple variables to predict the value of Nifty and hence we have made use of   linear regression of multiple variables. The limitation of linear regression is that it assumes a linear relationship and this cannot be true always with respect to a dataset on stock indices. So we need

to include some different approach also. If the predictors are chosen from a range of different models like regressors and decision trees and new predictors are learning from mistakes committed by previous predictors, there is more chance to come close to actual predictions. Hence we have considered a boosting algorithm namely gradient boosting regression also to ensure prediction accuracy.

We have implemented multivariate analysis of the stock market index data in python with the help of Scikit-Learn which is a very popular machine learning library very suitable for implementing regression with two variables as well as multiple variables. Apart from the Linear regression technique, Scikit-Learn of python is also suitable for implementing gradient boosting regressionand it is known to be with better accuracy than many other regression methods[11].

The accuracy of both methods is tested using the regression evaluation metrics R^2, MAE, MSE and RMSE. These are commonly used metrics to evaluate the prediction error rates and model performance in regression analysis. R-squared is a statistical measure of how close the data are to the fitted regression line and is considered to be goodness-of-fit measure for regression models. This measures the strength of the relationship between the model and the dependent variable. MAE (Mean absolute error) measures the average magnitude of the errors in a set of forecasts, without considering their direction. MSE (Mean Squared Error) measures the mean of squares of differences between the original and predicted values in a set of forecasts. RMSE (Root Mean Squared Error) is calculated as the square root of MSE. These evaluation metrics can be expressed by the following equations,

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \qquad MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \qquad R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,
$\hat{y}$ − predicted value of y
$\bar{y}$ − mean value of y

A comparison of performance of linear regression and gradient boosting regression methods on the given dataset with respect to these metrics is also performed.

IV. RESULTS AND DISCUSSION

The dataset for Nifty, FTSE, HangSeng, KOPSI, Nikkei, SSE and TSEC stock indices were analysed for identifying their correlation using python library and the following table (Table 1) depicts the correlation of the different stock market indices. This table shows that with respect to the data taken here, Nifty to FTSE correlation is 0.481984, Nifty to HangSeng correlation is 0.581101, Nifty to Kospi correlation is 0.745837, Nifty to Nikkei correlation is 0.650583, Nifty to SSE correlation is 0.142948 and Nifty to TSEC correlation is 0.732842. The highest correlation is shown for Nifty to Kospi (0.745837) and the lowest(0.14248) is shown for Nifty to SSE. Thus it can be seen that Indian market represented by Nifty shows moderate to high level of correlation with other Asian markets considered for the study except Shanghai Composite Index (SSE).

Table 1: Correlation matrix for selected stock market indices.

	Nifty	FTSE	HangSeng	Kospi	Nikkei	SSE	TSEC
Nifty	1.0000	0.4820	0.5811	0.7459	0.6506	0.1430	0.7328
FTSE	0.4820	1.0000	0.8634	0.7473	0.51416	0.3789	0.8507
HangSeng	0.5811	0.8634	1.0000	0.7218	0.54326	0.5517	0.8525
Kospi	0.7458	0.7473	0.7218	1.0000	0.38176	0.1253	0.8280
Nikkei	0.6506	0.5141	0.5432	0.3817	1.00006	0.4924	0.6502
SSE	0.1429	0.3789	0.5517	0.1253	0.49246	1.0000	0.4158
TSEC	0.7328	0.8507	0.8525	0.8280	0.6502	0.4158	1.0000

The data were subjected to regression analysis using Nifty as dependent variable and the others as independent variables using two methods namely Multiple Linear regression and Gradient boosting. As seen from Table 2, the gradient boosting regression results show higher R^2 value and lower MAE, MSE and RMSE values when compared to the multiple linear regression results. Hence, it can be seen that Gradient boosting method performs better than multiple linear regression in this case.

Table 2: Metrics with respect to multiple linear regression and gradient boosting regression of Nifty with indices FTSE, Hangseng, Kospi, Nikkei, SSE and TSEC.

	Multiple Linear regression	Gradient boosting regression
R^2	0.8249	0.9672
MAE	510.4993	193.9462
MSE	409874.9015	76849.0808
RMSE	640.2147	277.2166

On plotting the actual and predicted values of Nifty obtained from regression with other six indices considered for the study, it is found that Gradient boosting technique has a better fit and more predictive power (Figure 1 and Figure 2) than multiple linear regression technique.

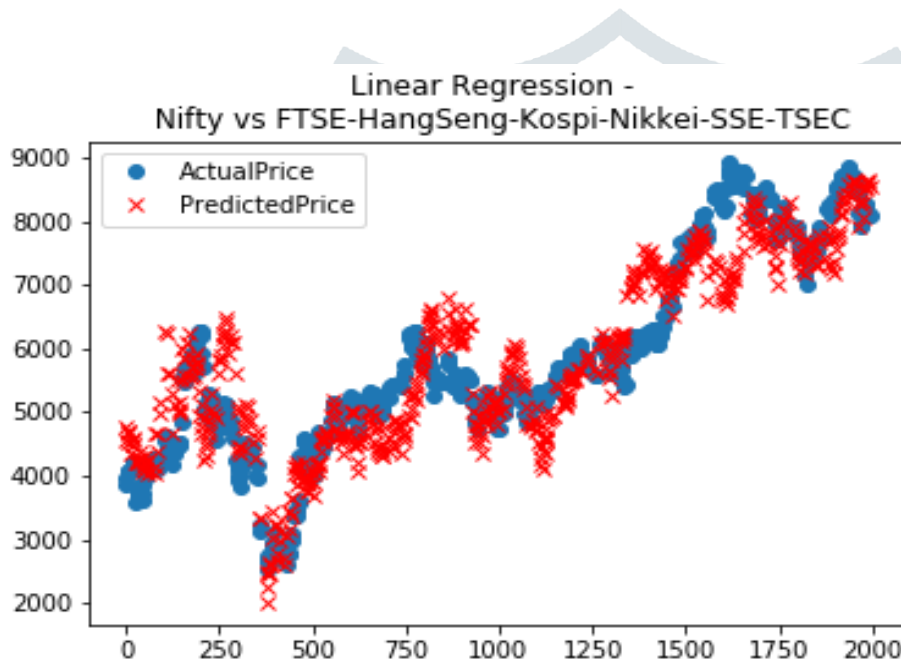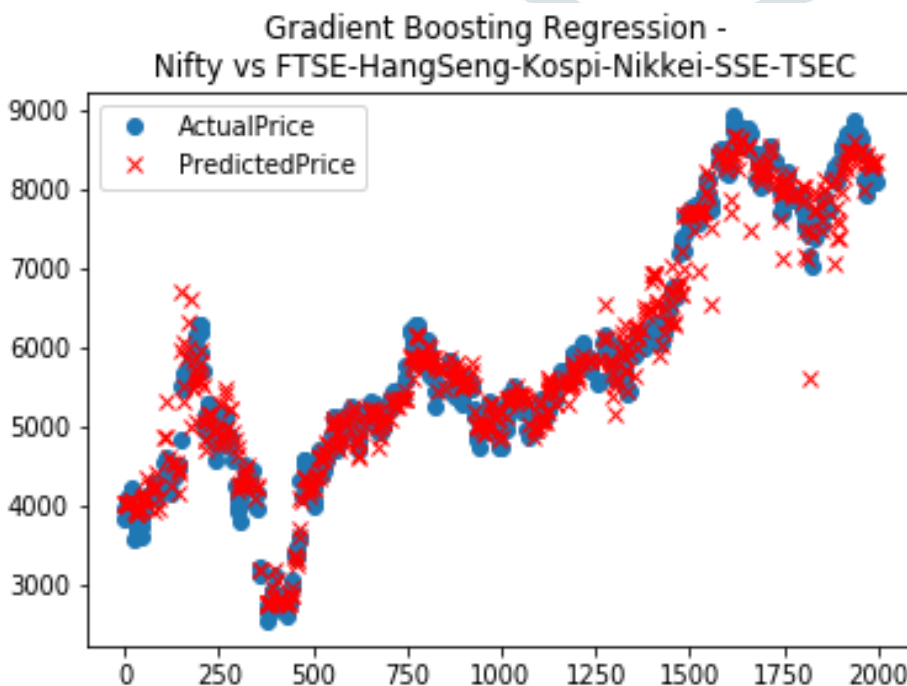Figure 1 – Multiple Linear regression – Nifty vs FTSE, HangSeng, Kospi, Nikkei, SSE, TSEC.

Figure 2 – Gradient Boosting Regression – Nifty vs FTSE, HangSeng, Kospi, Nikkei, SSE, TSEC.

V.CONCLUSION

In this study, the correlation and regression of Nifty with other asian stock indices (FTSE, Hangseng, Kospi, Nikkei, SSE and TSEC) were analysed by making use of python machine learning algorithms. It is found that there exists a relationship between NSE and the selected Asian stock markets and these Asian markets could be used for predicting changes in Indian stock market. The highest correlation is shown for Nifty to Kospi (0.745837) and the lowest (0.14248) is shown for Nifty to SSE. Two machine learning techniques namely multiple linear regression and gradient boosting regression were applied for studying the regression of Nifty with respect to the other six indices. The results suggest that the gradient boost regression method yields more accurate results in terms of regression evaluation metrics R^2, MAE, MSE and RMSE when compared to linear regression method. So, the gradient boost regression is found to be useful for prediction of Nifty value with respect to these selected indices of asian stock markets in this study. There is scope for further research in exploring the suitability of other machine learning techniques to improve prediction of stock market index.

REFERENCES

[1]Davim, J.P., Oliveira, C. and Cardoso, A., 2008. Predicting the geometric form of clad in laser cladding by powder using multiple regression analysis (MRA). Materials & Design, 29(2), pp.554-557.

[2]Müller, A.C. and Guido, S., 2016. Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc.

[3]Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. Frontiers inneurorobotics, 7, p.21.

[4] Joshi, P. 2011. Market Integration of Indian Stock Market- A Study of NSE, International Journal of Research in Commerce and Management, 1(6), 36-41.

[5]Jin, X., Guo, D. and Liu, H., 2014, September. Enhanced stock prediction using social network and statistical model. In 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA) (pp. 1199-1203). IEEE.

[6]Rajput, V. and Bobde, S., 2016. Stock market forecasting techniques: literature survey. International Journal of Computer Science and Mobile Computing, 5(6), pp.500-506.

[7]Cakra, Y.E. and Trisedya, B.D., 2015, October. Stock price prediction using linear regression based on sentiment analysis. In 2015 international conference on advanced computer science and information systems (ICACSIS) (pp.147-154). IEEE.

[8]Mao, Y., Wei, W., Wang, B. and Liu, B., 2012, August. Correlating S&P 500 stocks with Twitter data. In Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research (pp. 69-72).

[9]Zięba, M., Tomczak, S.K. and Tomczak, J.M., 2016. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert systems with applications, 58, pp.93-101.

[10]Gharehchopogh, F.S., Bonab, T.H. and Khaze, S.R., 2013. A linear regression approach to prediction of stock market trading volume: a case study. International Journal of Managing Value and Supply Chains, 4(3), p.25.

[11]https://scikit-learn.org