International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

Data Mining of Indian Stock Market from April, 2015 to March, 2016 Using Curve Fitting Technique

Dipankar Das

Assistant Professor, BCA(H), The Heritage Academy, Kolkata, West Bengal, India

---

Abstract – *The foundation of today's world is knowledge. Every single event generates some amount of data and by analyzing these vast amounts of data a lot of useful information can be obtained which may lead us to knowledge discovery. Data analysis is an essential step of knowledge discovery. The stock market generates a huge amount of data and by analyzing these data huge amount of useful information can be collected. This paper analyzed the S&P BSE SENSEX data from April, 2015 to March, 2016 to identify the pattern or tend of the S&P BSE SENSEX during this time period by using curve fitting technique. In this paper we have observed that the Compound, Growth and Exponential curves best fit the data i.e. S&P BSE SENSEX versus time during April, 2015 to March, 2016.*

Key Words: Curve Fitting, S&P BSE SENSEX, Data Mining, Compound Model, Growth Model, Exponential Model.

1. INTRODUCTION

The data can be analyzed from different point of view and different types of useful information can be obtained from these analyses. Every day large amounts of data are generated around us and analyzing these data in a proper and meaningful way is one of the challenging tasks lying in front of us.

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information [3]. According to Elayidom, Idikkula and Alexander (2009), "Data mining is the principle of analyzing large database and picking out interesting Patterns" [1]. Silwattananusarn and Tuamsuk (2012) in their work had stated that "Data mining is one of the most important steps of the knowledge discovery in databases process and is considered as significant subfield in knowledge management" [4]. Fernandez, et. al. had pointed out that for Data Mining there are different classes of statistical methods *e.g.* (i) Classical statistics (regression, curve fitting *etc.*), (ii) Induction of symbolic rules and (iii) Neural networks [2]. Elayidom, Idikkula and Alexander (2009) stated "Curve fitting is finding a curve which has the best fit to a series of data points" [1]. According to Bari, Chaouchi and Jung (n.d.) "Curve fitting is a process used in predictive analytics in which the goal is to create a curve that depicts the mathematical function that best fits the actual (original) data points in a data series" [5].

In the present study, the S&P BSE SENSEX (BSE's popular equity index) [6] has been considered. The study focuses on finding the underlying pattern or trend of S&P BSE SENSEX data from April, 2015 to March, 2016 using curve fitting technique to discover the knowledge of "How the S&P BSE SENSEX data perform during this time period".

2. RELATED WORK

Groll (2011) had discussed about stock price prediction based on curve fitting, pointed out that "in course of its evolution, any stock price seems to follow some trend at some point of time" and also noted that "the function seems to be a good approximation to past prices, chance might be that it will still be an approximation in the future" [7]. Juarna, Kuswanto, Mukhyi and Supriyanto (2015) had used curve fitting computation to predict a stock index in Indonesia Stock Exchange (namely IDX30) [8]. Talebnia, Ebrahimi and Darvishi (2015) had used the method of curve fitting with sinusoidal functions to examine the effects of Discounted Residual Income on stock price of companies listed in Tehran Security Exchange [9].

3. OBJECTIVES OF THE STUDY

(i) To find out the mean value of S&P BSE SENSEX for each month from April, 2015 to March, 2016.
(ii) To find out the mode value of S&P BSE SENSEX for each month from April, 2015 to March, 2016.
(iii) To find out the mean value of the largest cluster for each month (from April, 2015 to March, 2016) of S&P BSE SENSEX using "Two –Step Clustering Algorithm" and applying the distance measure as Log – Likelihood.
(iv) To find out the best curve(s) that can be fitted to the data points *i.e.* mean value of S&P BSE SENSEX for each month from April, 2015 to March, 2016 *versus* time.
(v) To find out the best curve(s) that can be fitted to the data points *i.e.* mode value of S&P BSE SENSEX for each month from April, 2015 to March, 2016 *versus* time.
(vi) To find out the best curve(s) that can be fitted to the data points *i.e.* mean value of the largest cluster of S&P BSE SENSEX for each month from April, 2015 to March, 2016 *versus* time.
(vii) To find out the best curve(s) that can be fitted to the data points *i.e.* daily values (Adjusted close) of the S&P

BSE SENSEX from April, 2015 to March, 2016 *versus* time.

(viii) Identifying the mathematical equation(s) of the best curve(s) that can be fitted to the data points *i.e.* mean value of S&P BSE SENSEX for each month from April, 2015 to March, 2016 *versus* time.

(ix) Identifying the mathematical equation(s) of the best curve(s) that can be fitted to the data points *i.e.* mode value of S&P BSE SENSEX for each month from April, 2015 to March, 2016 *versus* time.

(x) Identifying the mathematical equation(s) of the best curve(s) that can be fitted to the data points *i.e.* mean value of the largest cluster S&P BSE SENSEX for each month from April, 2015 to March, 2016 *versus* time.

(xi) Identifying the mathematical equation(s) of the best curve(s) that can be fitted to the data points *i.e.* daily values (Adjusted close) of the S&P BSE SENSEX from April, 2015 to March, 2016 *versus* time.

4. RESEARCH METHODOLOGY

Step 1: Collection of the historical daily dataset of S&P BSE SENSEX (*Adjusted close values*) from April, 2015 to March, 2016 from "Yahoo Finance" website [10].

Step 2: Month wise grouping of daily dataset.

Step 3: Finding the mean value for each month of S&P BSE SENSEX (from April, 2015 to March, 2016).

Step 4: Finding the mode value for each month of S&P BSE SENSEX (from April, 2015 to March, 2016).

Step 5: Finding the largest cluster using Two-Step Clustering algorithm [11] for each month of S&P BSE SENSEX (from April, 2015 to March, 2016). In this study Log – Likelihood distance measure has been used and Schwarz's Bayesian Criterion (BIC) has been chosen as clustering criterion.

Step 6: Finding the mean value of the largest cluster for each month of S&P BSE SENSEX (from April, 2015 to March, 2016) identified in the previous step (Step 5).

Step 7: Identifying the best curve(s) that can be fitted to the data points *i.e.* (a) mean value of S&P BSE SENSEX for each month from April, 2015 to March, 2016 *versus* time, (b) mode value of S&P BSE SENSEX for each month from April, 2015 to March, 2016 *versus* time, (c) mean value of the largest cluster of S&P BSE SENSEX for each month from April, 2015 to March, 2016 *versus* time and (d) daily values (Adjusted close) of the S&P BSE SENSEX from April, 2015 to March, 2016 *versus* time using curve fitting technique. The goodness of fit statistics used for identifying the best model are (i) R square, (ii) Adjusted R square and (iii) Root Mean Square Error (RMSE) where the decision rules are taken as (i) high value of R square, (ii) high value of Adjusted R square and (iii) low value of RMSE [12]. In this study, the researchers have used (i) Linear, (ii) Logarithmic, (iii) Inverse, (iv) Quadratic, (v) Cubic, (vi) Compound, (vii) Power, (viii) S, (ix) Growth and (x) Exponential types of models for curve fitting.

Step 8: Performing *F – Test* of the best model(s) to identify whether the independent variable reliably predicts the dependent variable. The decision rule is – if the significance of *F – test* is less than .05 then it can be concluded that the independent variable reliably predicts the dependent variable [13].

Software used: We have used SPSS 20 for analyzing the dataset.

5. DATA ANALYSIS & FINDINGS

The (i) month wise mean value of S&P BSE SENSEX (Mean), (ii) month wise mode value of S&P BSE SENSEX (Mode) and (iii) mean value of the largest cluster for each month (Mean of Largest Cluster) of S&P BSE SENSEX from April, 2015 to March, 2016 is given in the following table (Table -1):

Table -1: Mean, Mode and Mean of Largest Cluster

Month, Year	Mean	Mode	Mean of Largest Cluster
April, 2015	28164.11	28260.14	28667.53
May, 2015	27406.75	26599.11	27650
June, 2015	27137.88	26370.98	26728.47
July, 2015	28015.60	27459.23	28196.66
August, 2015	27386.86	25714.66	27915.26
September, 2015	25705.37	24893.81	25827.52
October, 2015	27011.80	26220.95	26844.29
November, 2015	26013.65	25482.52	25781.09
December, 2015	25658.19	25036.05	25963.07
January, 2016	24706.95	23962.21	24551.38
February, 2016	23688.30	22951.83	23328.81
March, 2016	24811.95	23779.35	24640.35

The goodness of fit statistics of the Mean *versus* time is given in the following table (Table -2):

Table -2: Goodness of fit statistics of the Mean *versus* time

Model Name	R Square	Adjusted R Square	RMSE
Linear	.809	.790	654.166
Logarithmic	.664	.630	868.779
Inverse	.423	.365	1138.525
Quadratic	.824	.785	662.161
Cubic	.831	.767	689.165
Compound	.804	.785	.025
Power	.653	.618	.034
S	.410	.351	.044
Growth	.804	.785	.025
Exponential	.804	.785	.025

Findings: From the above table we observe that Compound, Growth and Exponential models are having moderately high R Square values (*.804*), moderately high Adjusted R Square values (*.785*) and very low RMSE values (*.025*). Therefore, these three (*3*) models have been selected as candidate models (candidate to be the best curve) in this study.

The goodness of fit statistics of the Mode *versus* time is given in the following table (Table -3):

Table -3: Goodness of fit statistics of the Mode *versus* time

Model Name	R Square	Adjusted R Square	RMSE
Linear	.810	.791	708.176
Logarithmic	.745	.719	820.382
Inverse	.563	.519	1072.951
Quadratic	.810	.768	745.110
Cubic	.817	.749	775.753
Compound	.810	.791	.028
Power	.731	.704	.033
S	.541	.495	.043
Growth	.810	.791	.028
Exponential	.810	.791	.028

Findings: From the above table we observe that Compound, Growth and Exponential models are having moderately high R Square values (*.810*), moderately high Adjusted R Square values (*.791*) and very low RMSE values (*.028*). Therefore, these three (*3*) models have been selected as candidate models (candidate to be the best curve) in this study.

The goodness of fit statistics of the Mean of Largest Cluster *versus* time is given in the following table (Table -4):

Table -4: Goodness of fit statistics of the Mean of Largest Cluster *versus* time

Model Name	R Square	Adjusted R Square	RMSE
Linear	.788	.767	789.041
Logarithmic	.665	.631	991.946
Inverse	.449	.394	1271.846
Quadratic	.799	.755	808.815
Cubic	.801	.726	854.663
Compound	.783	.761	.031
Power	.652	.617	.039
S	.432	.375	.050
Growth	.783	.761	.031
Exponential	.783	.761	.031

Findings: From the above table we observe that Compound, Growth and Exponential models are having moderately high R Square values (*.783*), moderately high Adjusted R Square values (*.761*) and very low RMSE values (*.031*). Therefore, these three (*3*) models have been selected as candidate models (candidate to be the best curve) in this study.

The goodness of fit statistics of the daily values (Adjusted close) of the S&P BSE SENSEX *versus* time is given in the following table (Table -5):

Table -5: Goodness of fit statistics of the daily values (Adjusted close) of the S&P BSE SENSEX *versus* time

Model Name	R Square	Adjusted R Square	RMSE
Linear	.717	.716	778.581
Logarithmic	.508	.506	1026.188
Inverse	.089	.086	1395.926
Quadratic	.735	.732	755.186
Cubic	.740	.737	748.921
Compound	.713	.712	.030
Power	.498	.495	.040
S	.085	.081	.054
Growth	.713	.712	.030
Exponential	.713	.712	.030

Findings: From the above table we observe that Compound, Growth and Exponential models are having moderately high R Square values (*.713*), moderately high Adjusted R Square values (*.712*) and very low RMSE values (*.030*). Therefore, these three (*3*) models have been selected as candidate models (candidate to be the best curve) in this study.

The F – test and the significance of F – test of the candidate models for (i) Mean *versus* time (ii) Mode *versus* time, (iii) Mean of largest cluster *versus* time and (iv) Daily values (Adjusted close) of the S&P BSE SENSEX *versus* time is given in the following table (Table -6).

Table -6: F – test and significance of F – test of the candidate models for all the four cases

Case	Model Name	F – Test Value	Sig.
Mean *versus* time	Compound	41.051	.000078
	Growth	41.051	.000078
	Exponential	41.051	.000078
Mode *versus* time	Compound	42.625	.000066
	Growth	42.625	.000066
	Exponential	42.625	.000066
Mean of largest cluster versus time	Compound	36.090	.000131
	Growth	36.090	.000131
	Exponential	36.090	.000131
Daily values (Adjusted close) of the S&P BSE SENSEX *versus* time	Compound	615.003	.000000
	Growth	615.003	.000000
	Exponential	615.003	.000000

Findings: The significance of *F – test* of all the candidate models (for all the four cases) is less than .05 and therefore we may conclude that the independent variable reliably predicts the dependent variable and the models are good fit for the data.

From the above five (5) tables (Table -2 to Table -6) we may conclude that in all the four cases *i.e.* (a) Mean *versus* time (b) Mode *versus* time, (c) Mean of largest cluster *versus* time and (d) Daily values (Adjusted close) of the S&P BSE SENSEX *versus* time the (i) Compound, (ii) Growth and (iii) Exponential models are the best models amongst the ten (*10*) types of models tested in this research paper.

(a) The graphical representations and the mathematical equations of the best models for Mean *versus* time are given below:

(i) Compound model:

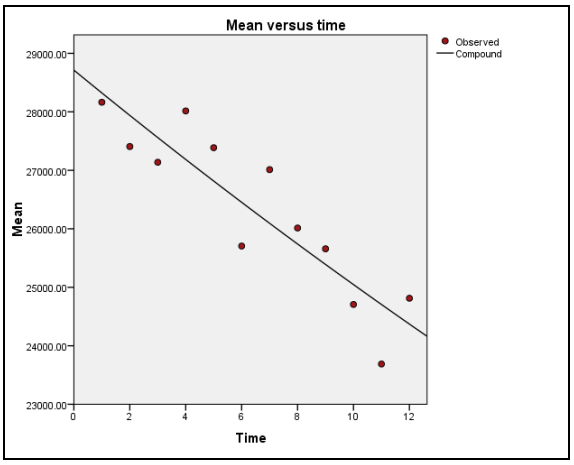

Chart -1: Compound model of Mean *versus* time

Mathematical equation of Compound model:
Mean = 28712.7507 * 0.9864**time

(ii) Growth model:

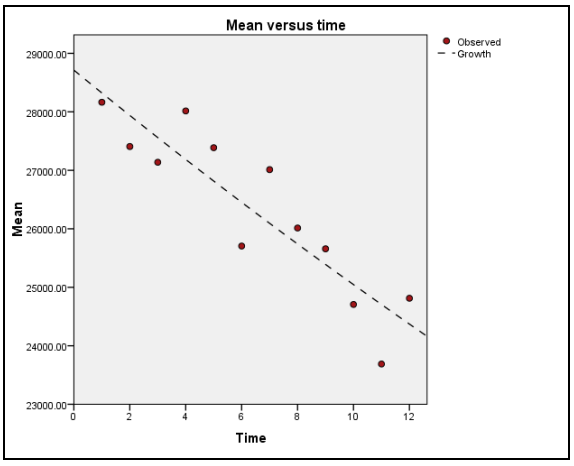

Chart -2: Growth model of Mean *versus* time

Mathematical equation of Growth model:
Mean = exp(10.2651 + -0.0137*time)

(iii) Exponential model:

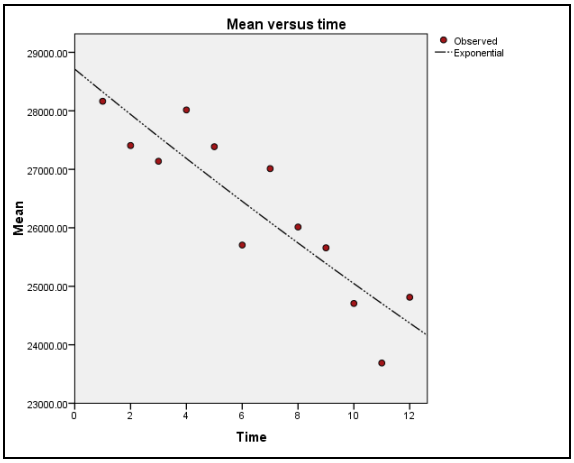

Chart -3: Exponential model of Mean *versus* time

Mathematical equation of Exponential model:
Mean = 28712.7507*exp(-0.0137*time)

(b) The graphical representations and the mathematical equations of the best models for Mode *versus* time are given below:

(i) Compound model:

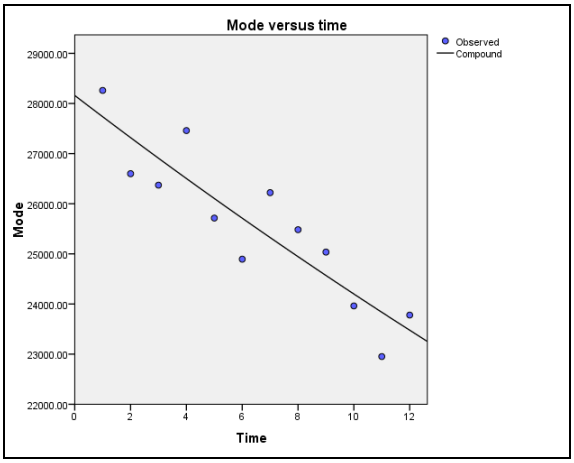

Chart -4: Compound model of Mode *versus* time

Mathematical equation of Compound model:
Mode = 28159.3155 * 0.9850**time

(ii) Growth model:

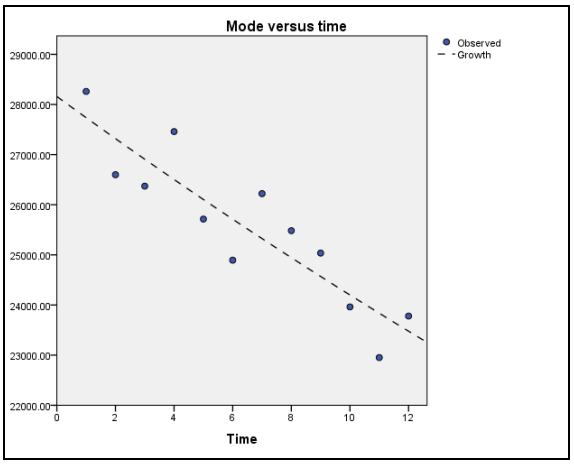

Chart -5: Growth model of Mode *versus* time

Mathematical equation of Growth model:
Mode = exp(10.2456 + -0.0152*time)

(iii) Exponential model:

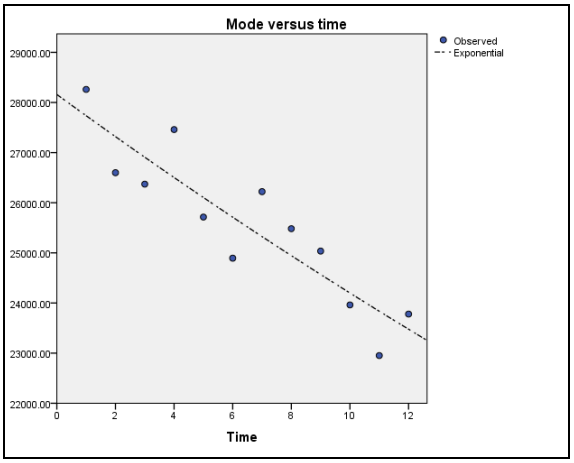

Chart -6: Exponential model of Mode *versus* time

Mathematical equation of Growth model:
Mode = 28159.3155*exp(-0.0152*time)

(c) The graphical representations and the mathematical equations of the best models for Mean of largest cluster *versus* time are given below:

(i) Compound model:

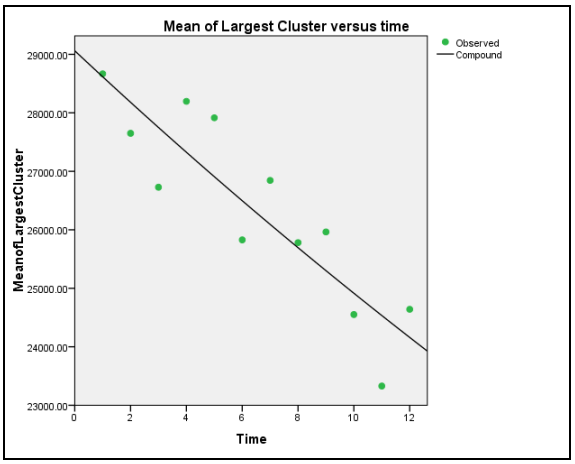

Chart -7: Compound model of Mean of largest cluster *versus* time

Mathematical equation of Compound model:
Mean of largest cluster = 29060.8064*0.9847**time

(ii) Growth model:

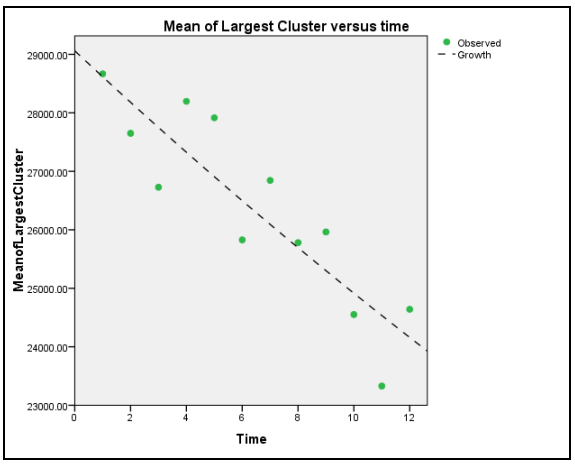

Chart -8: Growth model of Mean of largest cluster *versus* time

Mathematical equation of Growth model:
Mean of largest cluster =
exp(10.2771 + -0.0154*time)

(iii) Exponential model:

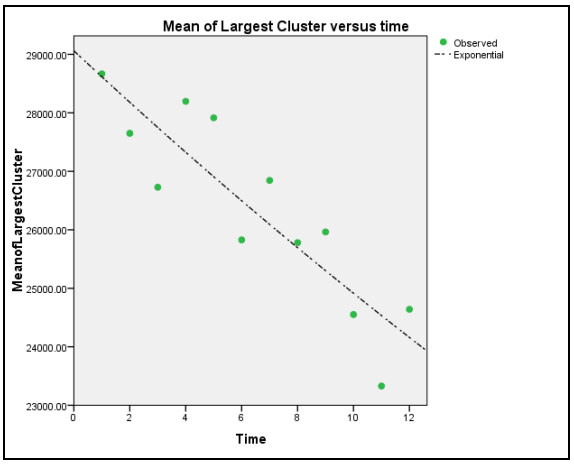

Chart -9: Exponential model of Mean of largest cluster *versus* time

Mathematical equation of Growth model:
Mean of largest cluster =
29060.8064 * exp(-0.0154 * time)

(d) The graphical representations and the mathematical equations of the best models for Daily values (Adjusted close) of the S&P BSE SENSEX *versus* time are given below:

(i) Compound model:

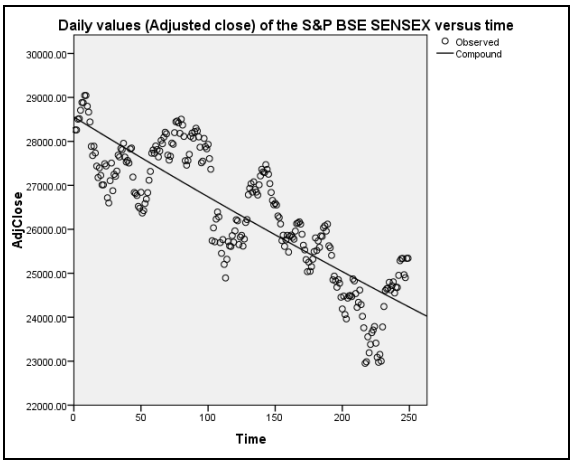

Chart -10: Compound model of Daily values (Adjusted close) of the S&P BSE SENSEX *versus* time

Mathematical equation of Compound model:
Daily values (Adjusted close) of the S&P BSE SENSEX =
28555.5859 * 0.9993**x

(ii) Growth model:

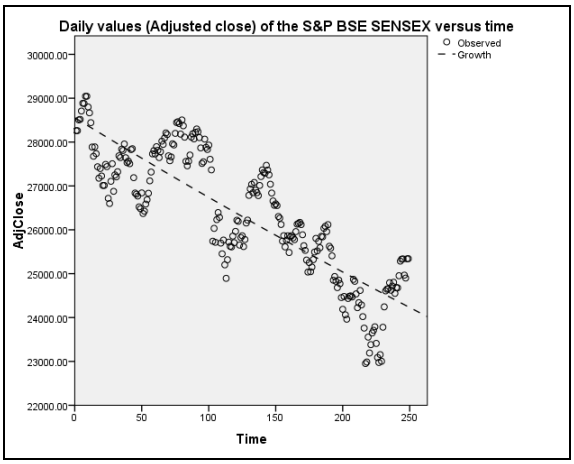

Chart -11: Growth model of Daily values (Adjusted close) of the S&P BSE SENSEX *versus* time

Mathematical equation of Growth model:
Daily values (Adjusted close) of the S&P BSE SENSEX =
exp(10.2596 + -0.0007 * time)

(iii) Exponential model:

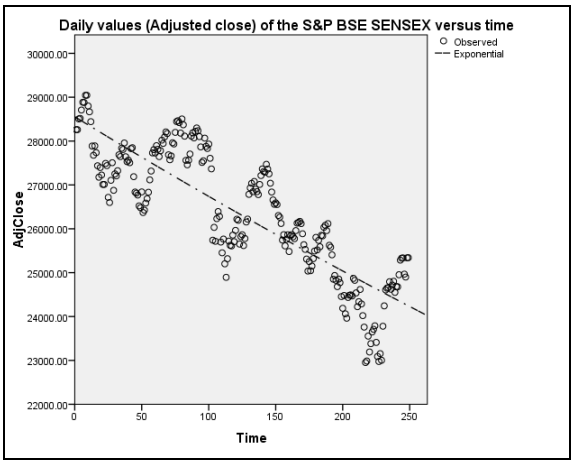

Chart -12: Exponential model of Daily values (Adjusted close) of the S&P BSE SENSEX *versus* time

Mathematical equation of Growth model:
Daily values (Adjusted close) of the S&P BSE SENSEX =
28555.5859 * exp(-0.0007 * x)

6. CONCLUSION

The main objective of the present study is to find out the hidden pattern or trend of S&P BSE SENSEX data from April, 2015 to March, 2016 by identifying the best fit curves for the dataset (using curve fitting technique). For doing the analyses we have considered four (*4*) cases *i.e.* (i) Mean value for each month of S&P BSE SENSEX *versus* time, (ii) Mode value for

each month of S&P BSE SENSEX *versus* time, (iii) Mean value of the largest cluster for each month of S&P BSE SENSEX *versus* time and (iv) daily values (Adjusted close) of the S&P BSE SENSEX *versus* time. In the present study, we have observed that in all the four (*4*) cases the Compound, Growth and Exponential curves fit the data best amongst the ten (*10*) tested models. Hence, we may conclude that the S&P BSE SENSEX shows three (*3*) types of patterns or trends (Compound, Growth and Exponential) for the time period April, 2015 to March, 2016.

In this present study, we have only used only ten (*10*) different types of models. There exist many more popular models which we have not explored in this case. At the same time, we have only studied S&P BSE SENSEX for a particular time period. Doing the same study on other stock indices of India for this particular time period and/or for a longer time period by including other models along with the models studied in this paper may enlighten us with more in-depth knowledge about the pattern or trend of the stocks in India.

REFERENCES

[1] Elayidom, S., Idikkula, S. M., Dr., & Alexander, J. (May, 2009). Applying Data mining using Statistical Techniques for Career Selection. *International Journal of Recent Trends in Engineering, 1*(1), 446-449. Retrieved May 11, 2016, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.379.2351&rep=rep1&type=pdf
SHORT PAPER

[2] Becerra-Fernandez., et al. (2004). Knowledge Management 1/e, Retrieved May 11, 2016, from https://home.cse.ust.hk/~dekai/523/notes/KM_Slides_Ch12.pdf

[3] Data Mining: What is Data Mining? (n.d.). Retrieved May 11, 2016, from http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[4] Silwattananusarn, T., & Tuamsuk, K., Dr. (September 2012). Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process, 2*(5), 13-24. doi:10.5121/ijdkp.2012.2502

[5] Bari, A., Chaouchi, M., & Jung, T. (n.d.). How to Use Curve Fitting in Predictive Analytics. Retrieved May 11, 2016, from http://www.dummies.com/how-to/content/how-to-use-curve-fitting-in-predictive-analytics.html

[6] BSE SENSEX. (n.d.). Retrieved May 11, 2016, from https://en.wikipedia.org/wiki/BSE_SENSEX

[7] Groll, C. (2011, August). *Working with financial data: Regression analysis and curve fitting* [PDF]. Retrieved May 11, 2016, from http://www.finmetrics.statistik.uni-muenchen.de/download/matlab_for_finance/part2.pdf

[8] Juarna, A., Kuswanto, A., Mukhyi, M. A., & Supriyanto, R. (2015, May 5-6). *Curve Fitting and Stock Price Prediction Using Least Square Method* [PDF]. Bali (Indonesia): 3rd International Conference on Business, Law and Corporate Social Responsibility (ICBLCSR'15). 107-108, http://dx.doi.org/10.15242/ICEHM.ED0515052

[9] Talebnia, G., Ebrahimi, M., & Darvishi, A. (Winter, 2015). An Application of Discounted Residual Income for Capital Assets Pricing by Method Curve Fitti ng with Sinusoidal Functions. *International Journal of Management and Business Research, 5*(1), 1-7. Retrieved May 11, 2016, from http://ijmbr.srbiau.ac.ir/pdf_5686_9fbce4cd4165de369e7ae3a4a82cb92b.html

[10] ^BSESN Historical Prices | S&P BSE SENSEX Stock - Yahoo! India Finance. (n.d.). Retrieved May 11, 2016, from https://in.finance.yahoo.com/q/hp?s=%5EBSESN

[11] Conduct and Interpret a Cluster Analysis - Statistics Solutions. (n.d.). Retrieved May 11, 2016, from http://www.statisticssolutions.com/cluster-analysis-2/

[12] Evaluating Goodness of Fit. (n.d.). Retrieved May 11, 2016, from http://in.mathworks.com/help/curvefit/evaluating-goodness-of-fit.html?requestedDomain=in.mathworks.com

[13] Annotated SPSS Output. (n.d.). Retrieved May 11, 2016, from http://www.ats.ucla.edu/stat/spss/output/reg_spss_long.htm

BIOGRAPHY

The area of interest of Mr. Dipankar Das is Data Analytics, Curve Fitting, Experimental Algorithmics etc. He is currently working as an Assistant Professor in The Heritage Academy, Kolkata, India.