

A Comparative Study of Different Statistical Techniques Applied to Predict Share Value of State Bank of India (SBI)

Hota H.S., Sahu Pushpanjali

Abstract —. Prediction of share value is one of the critical job and is necessary for the current financial scenario, due to the high uncertainty prediction system can not predict the share value with high accuracy. In this piece of research work an attempt is made to analyze the prediction based on statistical techniques with special reference to the share value of State Bank of India (SBI). The data that is downloaded consists share value for open, close, volume, high, and low in equal interval of time from Jan-2003 to May-2011. Two different techniques ARIMA and Exponential Smoothing is used to compare the accuracy. Statistical measure are carried out and it is found that expert modeler is working well for the prediction of share value of SBI. The future value for the next 5 months from May-2011 from both the models are also evaluated
Keywords- Expert modeler ,Exponential Smoothing , Auto Regressive Integrated Moving Average (ARIMA).

I. INTRODUCTION

Prediction or forecasting [1] in financial scenario specially in share market is full of uncertainty. Various factors are involve which can affect the forecasting and may cause financial loss to individual , financial institution or industrial organization, in order to minimize this, a good and reliable system must be developed.

Statistical techniques are the one through which one can predict the share value, which is based on the historical data. In this paper the historical data that is collected for training and testing the model is taken from yahoo financial [2] which contains data from Jan-2003 to May-2011. There are total 7 field and 101 records. Many works have been done and different techniques have been applied for forecasting share value by the different researchers ,this piece of research work confined on two well known techniques ARIMA and Exponential Smoothing , ARIMA is applied for electricity price forecasting [7] by Javier Contreras et al and found good result.

H.S.Hota , Assistant Professor , Dept of CSIT ,GGV Bilaspur ,India M.No -9425222658 (e_mail:hota_hari@rediffmail.com)
Miss Pushpanjali Sahu ,Assistant Professor , Dept of CS ,Govt. College Korba ,M.No -9755270280 (e_mail:sahupushpanjali@gmail.com)

II. DESCRIPTION OF DATA

As mentioned ,data is downloaded from yahoo finance[2]containing historical data of SBI share value with 7 fields of which only 6 fields are taken:

Open:The value of share that is opened on a particular date.

High: The value of share that is highest on a particular date.

Low: The value of share that is lowest on a particular date.

Close: The value of share that is closed on a particular date.

Volume: The volume field shows the number of share sold during a particular date.

Date: The date field shows the time interval at which the various records are taken.

The data set contains total of 101 records which is collected on monthly basis from Jan-2003 to May-2011 and the nature of data is time-series.

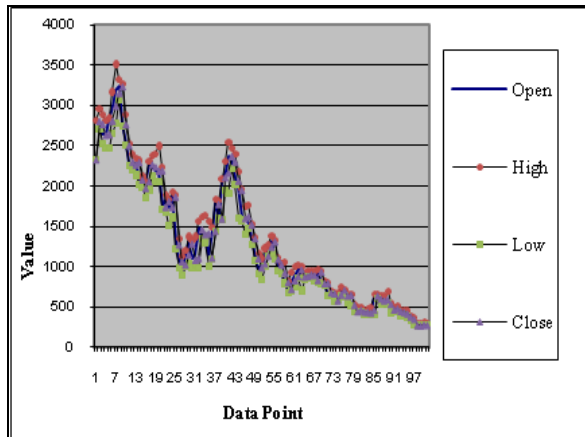
A time series data [1],[9] is an ordered collection of measurements taken at regular intervals. According to Werner Z.Hirsch, "A Time series is a sequence of values of some variate corresponding to successive points in time". "Time series" are usually related to economic data and the economist are generally responsible for the development of time series analysis technique, yet, they are applicable to all other phenomena that are related to time. Time series data are used to help in understanding past behavior and helps in forecasting and planning.

The past behavior of a series helps us to identify patterns and make better forecasts. By plotting a graph a time series exhibit one or more of the following features:

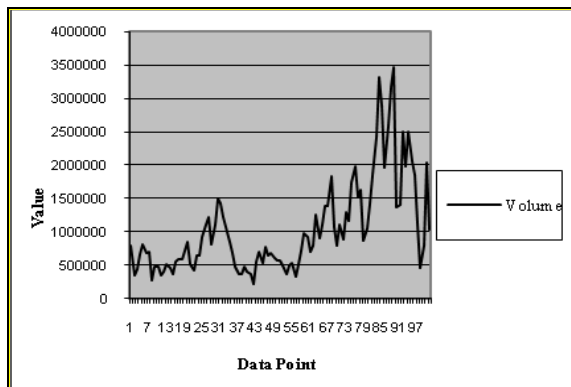
Trends, seasonal , non seasonal cycles, pulses, steps and outliers. The historical data set for share market gives the trends and seasonality patterns that helps us to decide the accurate model for forecasting the future values and thus helps the investors to make better decision to buy ,sell or to retain the share to gain profit in their business.

Complete data set is shown in appendix-A and the trend of the data is shown graphically in Fig.1(a) for open, high, low and close and Fig.1(b) for volume .The trend estimates were considered for the data sets of continuous months from year 2003 to 2011.The X-axis of the graph represents the data points (Month wise starting from May 2011 from origin) for the corresponding months. The Y-axis of the graph represents the

corresponding values of share market for open, high, low, close, and volume. Fig. 1(a) shows an upward pattern from Jan 2003 to May 2011 for the given series. The series values tend to increase over time while Fig. 1(b) shows a downward pattern from Jan 2003 to May 2011 for the given series. The series values tend to decrease over time. From the figure it is clear that data set has trends and seasonality characteristics. The volume decreases down as the price goes up with time



(a)



(b)

Figure 1. Trend of data related to SBI Share Market based on (a) For Open,High,Low and Close (b) For Volume

model is statistical. Using the direct statistical techniques different models are developed to forecast SBI share

III. METHODOLOGY USED

Various methods have been applied to predict share value of SBI. The Methodology used here to build the

- Time series data in which there is no trend and no seasonality .
- Time series data in which there is trend and no seasonality.
- Time series data in which there is seasonality and no trend.
- Time series data in which there is trend and seasonality.

value using the data explained in table -1. The three Methods used for forecasting are explained below -

A. *Exponential Smoothing*: [8] Exponential smoothing is a method of forecasting that uses weighted values of previous series observations to predict future values. It forecasts one point at a time, adjusting its forecasts as new data come in.

There are different models in Exponential Smoothing based on trends and seasonality of time series data which are shown clearly in table -2.

TABLE 2. DIFFERENT EXPONENTIAL MODELS

S No	Model	Nature of time series data	
		Trend	Seasonality
1	Simple	NO	No
2	Holt's Linear Trend	Yes	No
3	Brown's Linear Trend	Yes	No
4	Damped Trend	Yes	No
5	Simple seasonal	No	Yes
6	Winter's Additive	Yes	Yes
7	Winter's Multiplicative	Yes	Yes

From the above table it is clear that exponential smoothing method can be divided into four different categories according to the nature of time series data :

Appropriate model of exponential smoothing method is applied from table according to the nature of time series data ,in our case time series data has both trends and seasonality characteristics hence winter's model will be the best suitable.

Mathematical expression for this model is explained below-

$$\begin{aligned}
 L(t) &= \alpha(Y(t) - S(t-s)) + (1-\alpha)(L(t-1) + T(t-1)) \\
 T(t) &= \gamma(L(t) - L(t-1)) + (1-\gamma)T(t-1) \\
 S(t) &= \delta(Y(t) - L(t)) + (1-\delta)S(t-s) \\
 \hat{Y}_t(k) &= L(t) + kT(t) + S(t+k-s)
 \end{aligned}$$

Where,

Y_t ($t=1,2,\dots,n$) Univariate time series under investigation.

S The seasonal length.

α Level smoothing weight
 γ Trend smoothing weight
 δ Season smoothing weight

For the additive winter's model, fit $y = \alpha t + \sum_{i=1}^s \beta_i I_i(t)$ to the data where t is time and $I_i(t)$ are seasonal dummies. This model does not have an intercept and $T=\alpha$, $S=\beta-\text{mean}(\beta)$. Here, $L=y_n$ for the above model and β is the vector of slope(of length s).

B. Auto Regressive Integrated Moving Average (ARIMA): [8] ARIMA processes are a class of stochastic processes used to analyze time series. The application of the ARIMA method for the study of time series analysis is due to Box and Jenkins [1].

n Total number of investigations.
 $\hat{Y}_t(k)$ Model-estimated k -step ahead at time t for series Y .

processes used to analyze time series. The application of the ARIMA method for the study of time series analysis is due to Box and Jenkins [1].

C. Expert Modeler: [8] Expert Modeler automatically identifies and estimates the best-fitting ARIMA or Exponential Smoothing model for one or more target variables, thus eliminating the need to identify an appropriate model through trial and error. The overall working process of expert modeler is shown in Fig. 2.

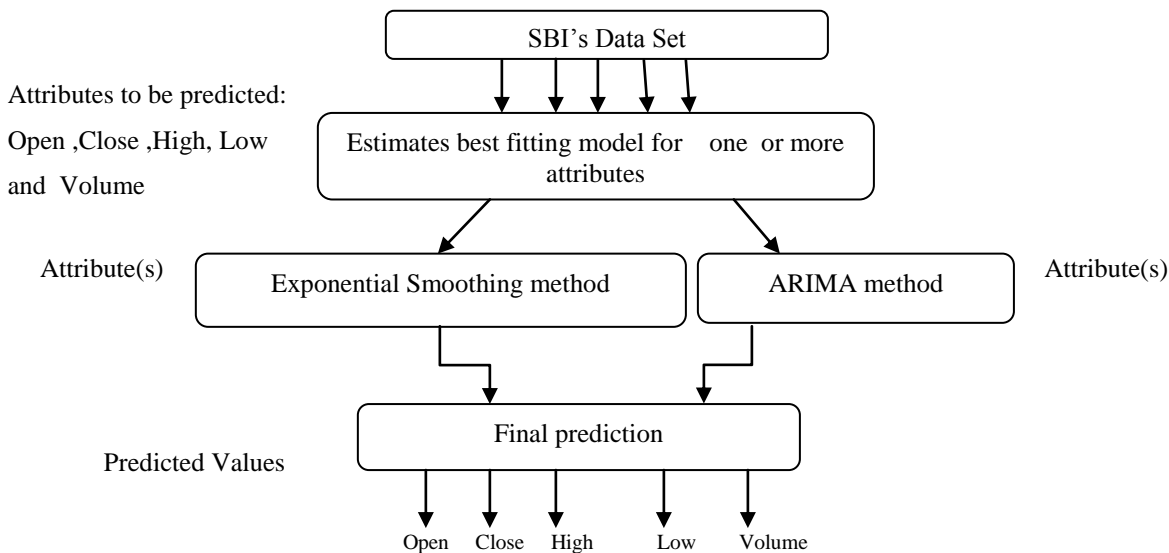


Figure2. Process of Expert modeler

IV. MODEL SIMULATION

To simulate the model a stream is designed using SPSS Clementine Software [8] for forecasting as shown in Fig.3. The source node consists of monthly data. A filter node is used to filter the Adj Close field which is not taken for forecasting. The filter node is

connected to a filler node which is used to change the default date format to match the format of the date field, this is necessary for the conversion of the date field to work as expected. The type node is connected with a filler node which specifies the field metadata and properties that are important for modeling. The direction is set to none for date field, and to out for the other specified fields in the type node. A time interval node is followed by the type node is used to set the time interval.

Above mentioned techniques are used in time series node to train and test the data using ARIMA and winters additive models. The nugget named as Exponential and Expert shown in the figure 3 are the

generated executable nodes for both the methods respectively.

Entire data set is divided into two parts training and testing. The training data consists 96 records which is used to train the model, and forecasted share value for all the fields are shown in table-3. From the table it is clear that actual value is closer to the predicted value for the corresponding data field. The model is tested with remaining data i.e. record no. 97 to 101 (Only 5 records) from table- 1 using both the model and result is tabulated in table- 3(a) and 3(b) for both the models. Fig. 4(a),(b),(c),(d),(e) and Fig. 5(a),(b),(c),(d),(e) shows the comparative time plot graph of actual and predicted value of exponential smoothing method and expert modeler method respectively. One can see from the figure that predicted value is very much closer to the actual value for all the predicted field. Predicted value for corresponding field is represented with OPEN(P), HIGH(P), LOW(P), CLOSE(P) and VOLUME(P). Similarly table- 4(a) and 4 (b) shows the predicted value for all the fields for the testing data.

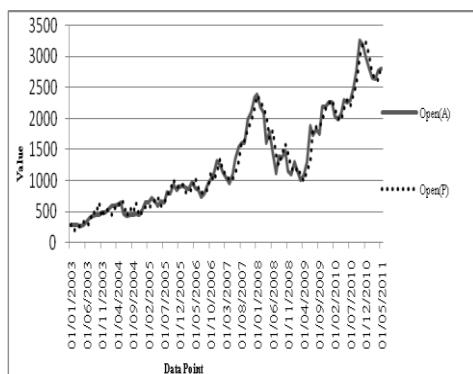
TABLE 4. TABLE SHOWING FORECASTED FOR TESTING DATA SET

Record #	Date	OPEN(P)	HIGH(P)	LOW(P)	CLOSE(P)	VOLUME(P)
97.	Jan 2011	2990.1	3264.4	2748.4	2653.8	79351.8
98.	Feb 2011	2909.8	3359.5	2844.4	2573.6	68762.7
99.	Mar 2011	2869.5	3457.3	2943.8	2520.6	550683.1
100.	Apr 2011	2825.2	3558.0	3046.6	2718.0	389000.0
101.	May 2011	2906.7	3661.7	3153.0	2801.5	812814.1

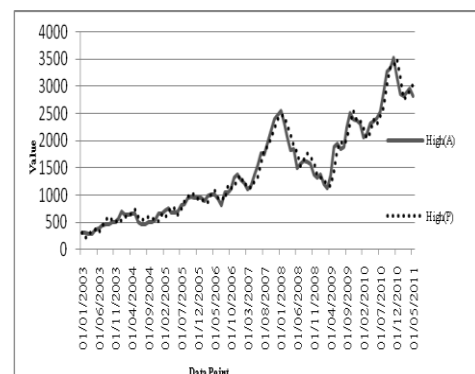
(a) Using Expert modeler method

Record#	Date	OPEN(P)	HIGH(P)	LOW(P)	CLOSE(P)	VOLUME(P)
97.	Jan 2011	2990.1	3134.5	2572.4	2699.0	665905.1
98.	Feb 2011	2909.8	3043.3	2558.1	2661.2	769676.0
99.	Mar 2011	2869.5	3012.2	2470.6	2614.3	598313.2
100.	Apr 2011	2825.2	3036.0	2516.8	2690.4	390201.2
101.	May 2011	2906.7	3142.0	2526.2	2741.4	819812.4

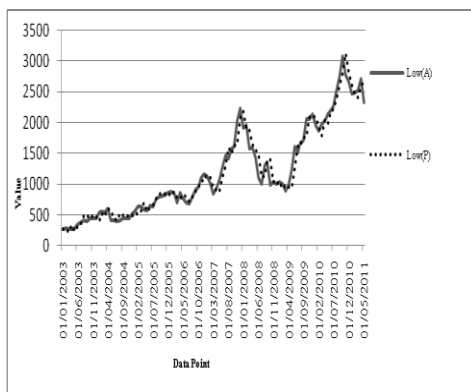
(b)Using Exponential Smoothing method



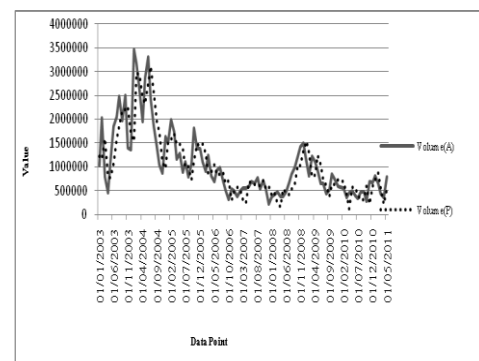
(a)



(b)

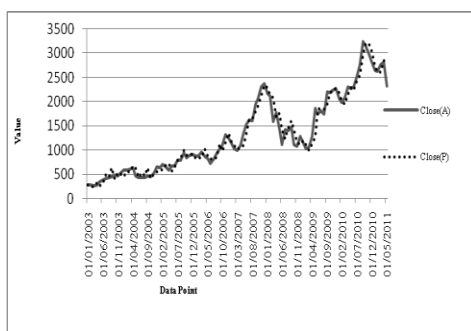


(c)

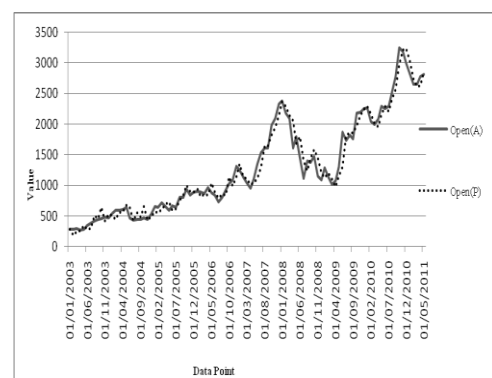


(e)

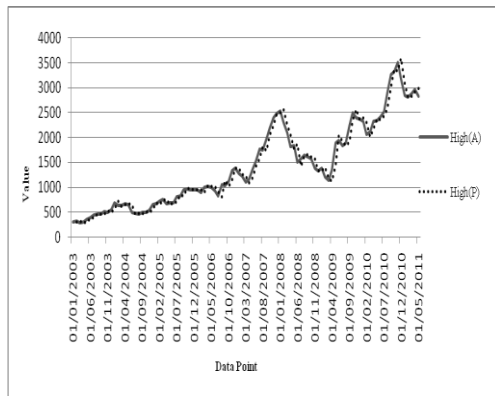
Figure 4. Actual and predicted time plot graph of Exponential Smoothing method after training (a)For Open (b)For High (c) For Low (d)For Close (e)For Volume



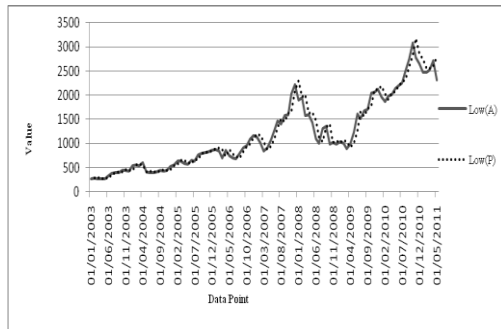
(d)



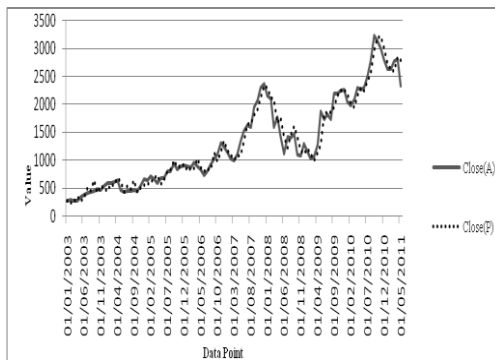
(a)



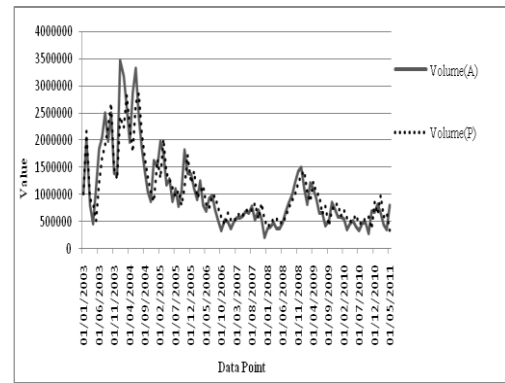
(b)



(c)



(d)



(e)

Figure 5. Actual and predicted time plot graph of Expert modeler method after training (a)For Open (b)For High (c) For Low (d)For Close (e)For Volume

V. RESULT ANALYSIS AND MODEL VALIDATION

To compare two different methods following statistical error measurement techniques are being used -

A. Mean Absolute Percentage Error (MAPE) – MAPE can be calculated according to the following formula:

$$MAPE = \frac{\text{Observed value} - \text{Predicted value}}{\text{Predicted value}} \times 100$$

When MAPE is less ,accuracy of the model will be more.

B. Stationary R – Square (R2)** – which provides an estimate of the proportion of the total variation in the series. The higher the value (Max 10) the better the fit of the model.

C .Significance Level – Provides an indication of whether the model is correctly specified or not.

Using above statistical measure both the models are observed for its accuracy to predict SBI's share value ,we have compared the forecasted data in case of training and testing for both the models and the results are shown in table-5(a) and table-5(b) respectively.

TABLE 5. STATISTICAL MEASURES

(a) Statistical Measures for training data set

Model	Attributes	Stationary R Square	Significance	MAPE
Expert Modeler	Open	0.555	0.112	9.94
	High	0.202	0.352	8.15
	Low	0.13	0.396	9.17
	Close	0.633	0.5	9.74
	Volume	0.761	0.135	12.32
Exponential Smoothing	Open	0.555	0.112	9.94
	High	0.527	0.297	8.45
	Low	0.536	0.017	9.24
	Close	0.562	0.146	9.74
	Volume	0.677	0.024	15.77

(b) Statistical Measures for testing data set

Model	Attributes	Stationary R Square	Significance	MAPE
Expert Modeler	Open	7.5	0.112	10.23
	High	0.202	0.352	8.48
	Low	0.13	0.396	9.72
	Close	0.633	0.5	10.02
	Volume	0.761	0.135	11.29
Exponential Smoothing	Open	0.555	0.112	10.23
	High	0.527	0.297	8.89
	Low	0.536	0.017	9.35
	Close	0.562	0.146	9.90
	Volume	0.677	0.024	12.57

The Significance value of expert modeler for open, high, low, close and volume are more than 0.05 while for exponential smoothing, it is more for open, high and close but less for low and volume. Taking all the statistical measures, stationary R-squared, significance value and MAPE into account, the expert modeler gives better result. The fact that training error is less then testing error which is obvious ,because testing data are unseen which is not previously known or seen by the model although testing error other than volume are in acceptable range specially in context of forecasting.

The model validation is carried out with expert modeler because it is better then other model explained in this paper, for the validation purpose data is again downloaded from the same site [2] and compared with the predicted data of the model for the month of Jun, July, Aug, Sep, and Oct, for the year 2011. The Expert Modeler method used here for forecasting, automatically selects Winters Additive model for Open and Close attributes and ARIMA for High, Low, and Volume attributes.

Table- 6 gives the time series forecasted values of share market for open, high, low, close and volume for next 5 months .

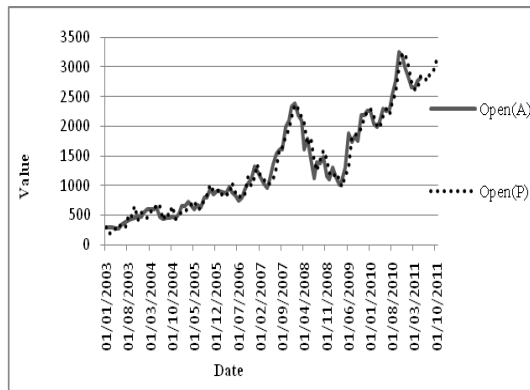
Data of table- 6 have been matched with the recently downloaded data and the result was found satisfactory ,say for example predicted value for open ,close ,low ,high and volume for the month of July 2011 are 2668.3,2445,2170.3,2959.65 and 369841.9 respectively while corresponding data available in the site for the same duration are 2419.8 , 2342, 1973.15, 2959.65 and 338900 respectively.

Fig. 6 (a),(b),(c) ,(d) and (e) shows the comparative time plot graph in between actual (A) values and the predicted (P) values from Jan 2003 to May 2011 along with the forecasted value for next five months i.e., for June, July, Aug, Sep, and Oct 2011 for all the attributes open, close, low ,high and volume respectively, in case of expert modeler. In these figures X-axis represents time interval in years which shows monthly values while the Y-axis represents price in unit where 1 unit = 500 for Fig. 6(a),(b),(c), and (d) while for Fig.6 (e) 1 unit = 5,00,000.

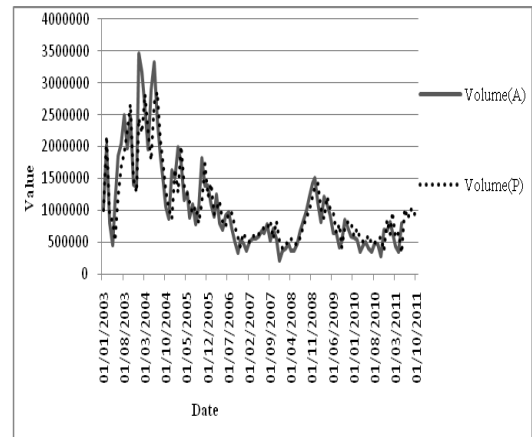
TABLE 6. FORECASTED VALUES FOR NEXT 5 MONTHS

S No	Model	Date	Forecasted values				
			Open	High	Low	Close	Volume
1	Expert Modeler	Jun2011	2516.9	2692.5	2394.1	2305.7	516277.1
2		July 2011	2668.3	2959.65	2170.3	2445	369841.9
3		Aug 2011	2687.3	2535.6	2249.6	2240.5	590800
4		Sep2011	2190	2456.6	2461.0	2451.5	877800
5		Oct 2011	2800.5	2556.6	2654.2	2607.7	896858.6

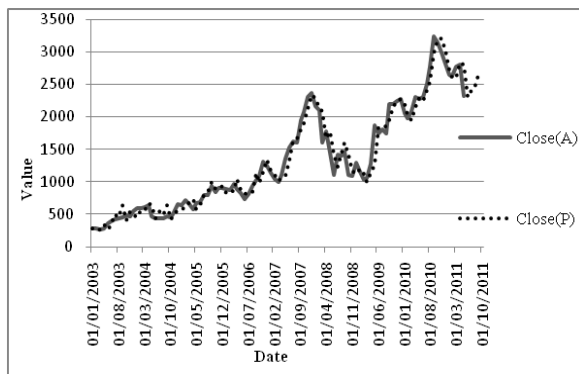
The graph in Fig. 6(a) shows that the values of the forecasted period goes down in the first two months of the forecasted period but it turns to grow up in the next three months. The time plot shows similar moves for all the other three fields close, low and high (Fig. 6(b), (c) and (d) respectively) while the graph in Fig. 6(e) shows downward move for the forecasted months .From all these graphs we can conclude or say that with the increase of share values for open, high ,low, and close, their is a decrease in the volume of the share.



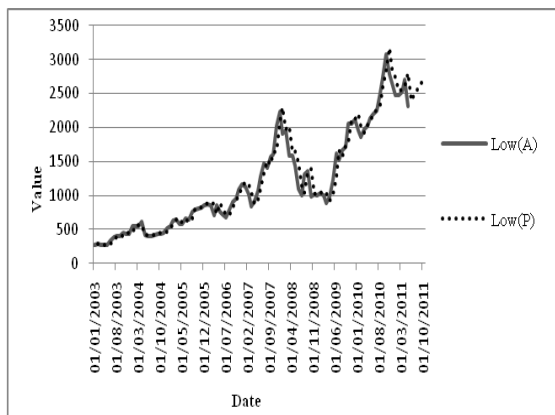
(a)



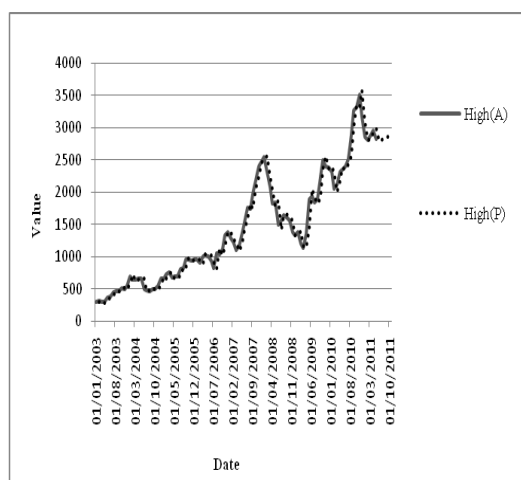
(e)



(b)



(c)



(d)

Figure 6: Time Plot graph for future value for Expert Modeler (a)Open(A) Vs. Open(P) (b) Close(A) Vs.Close (P) (c) Low(A) Vs.Low (P) (d) High(A) Vs.High(P) (e)Volume(A) Vs.Volume (P)

Also the lines for actual and forecast data over the entire time series are very close together in the graph, indicating that this is a reliable model for the prediction of time series data of SBI share value. Hence, we are able to predict the future share values for the SBI financial.

VI. CONCLUSION

Although predicting the share value is a tedious and complicated task and also full of uncertainty, but the traditional statistical techniques can be able to predict the share value up to some extend. The two different models expert modeler and exponential smoothing are tested with SBI financial data, and the errors are measured with the help of different statistical measurement techniques like MAPE, stationary square etc. and it was found that the expert modeler result are better then that of exponential smoothing. If we will compare the MAPE in case of training and testing, we found that MAPE is less in case of expert modeler while it is little bit high for exponential smoothing. More records in the data set can give better result and predicted value will be closer to the actual value and it helps the investors to make better decision to buy, sell or to retain the share to gain profit in their business.

(e)

REFERENCES :

- [1] G.E.P. Box ,G.M.Jenkins and G.C. Reinsel "Time series analysis forecasting and control " Third edition Englewood cliffs NJ prentice hall 1994.
- [2] Web Source <http://www.finance.yahoo.com> last accessed on January 2012..
- [3] Vatsal H. Shah,"Machine Learning Techniques for Stock Prediction".
- [4] Binoy.B.Nair, V.P Mohandas, N. R. Sakthivel, " A Decision Tree- Rough Set Hybrid System for Stock Market Trend Prediction".
- [5] Dr.B.N.Gupta(1995), "Statistics", Sahitya Bhawan Publishers
- [6] R.J.Frank, N.Davey, S.P.Hunt Department of Computer Science, University of Hertfordshire,
- [7] Javier Contreras ,Francisco J. Nogales and Antonio J.Conejo " ARIMA models to prtedict next-day electricity prices " IEEE transction on power systems vol 18 ,No 3 august 2003.
- [8] SPSS Clementine Release 12.0 help fille
- [9] Box, G.E.P, Jenkins ,G.M. "Time series analysis forecasting and control ",San Francisco ,CA:Holden

Appendix -A
TABLE 1. UP TO DATE DATA OF SBI SHARE MARKET

S No	Date	Open	High	Low	Close	Volume
1	02/05/2011	2811.5	2819.55	2320	2327.6	798900
2	01/04/2011	2772	2959.9	2707	2805.6	349000
3	01/03/2011	2651	2888	2523.55	2767.9	445000
4	01/02/2011	2651.9	2813.4	2478.6	2632	662100
5	03/01/2011	2830.05	2852.45	2468.8	2641.05	819700
6	01/12/2010	2998	3172	2655.7	2811.05	682500
7	01/11/2010	3187	3515	2777	2994.1	694200
8	01/10/2010	3250	3322	3077	3151.2	280700
9	01/09/2010	2772	3268	2738.75	3233.2	478900
10	02/08/2010	2520	2884	2511	2764.85	482200
11	01/07/2010	2290	2519.9	2254.4	2503.8	343500
12	01/06/2010	2260	2402.5	2201	2302.1	403300
13	03/05/2010	2291	2348.8	2138	2268.35	505800
14	01/04/2010	2085	2318.8	2015	2297.95	452200
15	02/03/2010	1990	2120.05	1978	2079	357400
16	01/02/2010	2045	2059.95	1863	1975.85	550300
17	04/01/2010	2265	2315.25	1957	2058	583500
18	01/12/2009	2253.05	2374.75	2126.2	2269.45	585800
19	03/11/2009	2190	2394	2059.1	2238.15	725500
20	01/10/2009	2180.1	2500	2048.2	2191	855900
21	01/09/2009	1760	2235	1710.1	2195.7	508000
22	03/08/2009	1825	1886.9	1670	1743.05	428700
23	01/07/2009	1737.9	1840	1512	1814	650300
24	01/06/2009	1875	1935	1612	1742.05	651200
25	04/05/2009	1300	1891	1225	1869.1	924500
26	01/04/2009	1079.7	1355	980	1277.7	1100200
27	02/03/2009	1010	1132.25	894	1066.55	1214400
28	02/02/2009	1141.1	1205.9	1008.3	1027.1	814100
29	01/01/2009	1294.45	1376.4	1031.05	1152.2	1106800
30	01/12/2008	1095	1325	995.05	1288.25	1504000
31	03/11/2008	1155	1375	1025	1086.85	1426500
32	01/10/2008	1480	1569.9	991.1	1109.5	1220600

33	01/09/2008	1376	1618	1353	1465.65	993100
34	01/08/2008	1396	1638.9	1302	1403.6	840500
35	01/07/2008	1120	1567.5	1007	1414.75	674000
36	02/06/2008	1450	1496.7	1101.15	1111.45	474100
37	02/05/2008	1796	1840	1438.2	1443.35	368500
38	01/04/2008	1611	1819.95	1592	1776.35	369400
39	03/03/2008	2089	2089.5	1582.25	1598.85	479100
40	01/02/2008	2186	2310	1953	2109.7	409100
41	01/01/2008	2381	2540	1905.6	2162.25	374000
42	03/12/2007	2330	2475.25	2225.35	2371	216500
43	01/11/2007	2100	2400	2025	2300.3	554300
44	01/10/2007	1984	2179.7	1601	2068.15	707100
45	03/09/2007	1614.5	1969.8	1575.1	1950.7	529000
46	01/08/2007	1610	1745	1408	1599.5	777900
47	02/07/2007	1531.85	1760	1470.05	1624.5	650900
48	01/06/2007	1352.4	1531	1278.25	1525.3	688300
49	01/05/2007	1105.25	1362	1068.8	1352.4	601700
50	02/04/2007	958.65	1165	915.1	1105.25	562900
51	01/03/2007	1039	1104.9	845	992.9	562000
52	01/02/2007	1140	1229	1009.9	1039.15	501700
53	01/01/2007	1245.9	1280	1126.5	1138.05	373600
54	01/12/2006	1314	1378.7	1165.05	1245.9	492800
55	01/11/2006	1095	1324.7	1088.5	1314	526800
56	02/10/2006	1028.3	1114	949.9	1095.5	328300
57	01/09/2006	928	1043	911	1028.3	512400
58	01/08/2006	800	1053.45	795	930	708800
59	03/07/2006	734	821.45	684.15	810.05	978900
60	01/06/2006	839	931	705.25	727.4	926100
61	01/05/2006	882.35	1009	755	831	700800
62	03/04/2006	968.5	1015	856	882.35	801200
63	01/03/2006	873	998.05	704	968.05	1250700
64	01/02/2006	890	899.4	849.55	877.2	907900
65	02/01/2006	909.8	950.8	878.15	886.8	1100300
66	01/12/2005	897	948	862.15	907.45	1379900
67	01/11/2005	846.7	942.7	820.7	896.25	1390800
68	03/10/2005	940.15	961.95	805.25	838.25	1823400
69	01/09/2005	799	950	792	938.6	1063900
70	01/08/2005	801	828	764.1	796.65	784800
71	01/07/2005	641	808.8	641	800.8	1102600
72	01/06/2005	670	695.2	653.1	681.55	880200
73	02/05/2005	592	684.9	576.65	670.7	1289900
74	01/04/2005	660	677	582.2	584.8	1167400
75	01/03/2005	715.55	750.7	637	656.95	1743900
76	01/02/2005	645	719.5	631	714.4	1988800
77	03/01/2005	655	662.45	552.5	642.8	1508800
78	01/12/2004	531.9	654.9	521	652.45	1628400
79	01/11/2004	449.9	536.9	447.4	529.7	866200
80	01/10/2004	470	498.45	435.5	447.35	1052700
81	01/09/2004	445.5	497.4	441.55	468.2	1427200

82	02/08/2004	446.7	463.6	415.15	442.85	1847000
83	01/07/2004	432.35	473.3	399.95	441.95	2421300
84	01/06/2004	471	498	404.25	430.65	3315900
85	03/05/2004	638.5	658	412	465	2884700
86	01/04/2004	605	665.95	605	642.6	1954800
87	01/03/2004	590	637.4	540.65	605.7	2602300
88	02/02/2004	595.8	640.85	550.5	585.3	3174000
89	01/01/2004	540.9	689.2	540.5	595.8	3463800
90	01/12/2003	475	552.25	434	538.5	1368000
91	03/11/2003	486.5	500	438.1	470.75	1398600
92	01/10/2003	452	508.5	444.1	484.2	2498000
93	01/09/2003	441.1	457.7	395.5	451.6	1973500
94	01/08/2003	423.5	460	403.65	439.25	2492400
95	01/07/2003	385	450	379.1	421.85	2046600
96	02/06/2003	345	388	340.05	384.2	1845800
97	01/05/2003	278.55	361.85	275.5	352.3	1130900
98	01/04/2003	270	296.1	269.95	278.55	452900
99	03/03/2003	292	294.55	266.65	269.9	795400
100	03/02/2003	284	316.65	284	285.75	2030200
101	01/01/2003	283	302.4	273.9	282.25	1035100