Term V – (Batch 2013 – 15)

Forecasting TATA STEEL’s stock prices

Business Analytics – Final Project

Objective

The objective of this project is to compare four different forecasting methodologies to see which one is the best to forecast TATA STEEL's stock prices.

Data & Methodology

1. Data for stock prices were downloaded from **www.nseindia.com**, for the past 1 year, starting from 16th December 2013 to 12th December 2014.
2. Four different forecasting techniques were chosen for this study:
    a. Multiple Linear Regression
    b. Artificial Neural Network (ANN)
    c. Winter-Holt's Method
    d. ARIMA Modeling
3. In each of these methods, we used Mean Absolute Percentage Error (MAPE) as a measure of forecast accuracy.
4. For estimating the best architecture for the neural network, many trials were done taking different values for the number of epochs, weight decay constant, number of hidden layers and the number of nodes in each hidden layer. The data set was split into training and validation sets (60:40 ratio) & last 12 days was taken as a hold-out sample to estimate the network's prediction accuracy.
5. Once a good fit model was developed on the development set, the model was used to forecast for the hold-out sample.
6. The model that gave use the least MAPE was categorized as the best model.
7. Analysis was done using MS Excel 2007 & EViews 8.0.

Analysis & Interpretation

We have analyzed four different forecasting procedures to understand the underlying issues in using a particular method for forecasting stock prices.

Once we have found that a method gives us a superior or inferior forecast accuracy when compared to other methods, we have tried to explain the reasons behind it & then, have clearly stated the method's limitations. Final recommendations were made based on a consolidated analysis on all the four methods.

I. Multiple Linear Regression Model

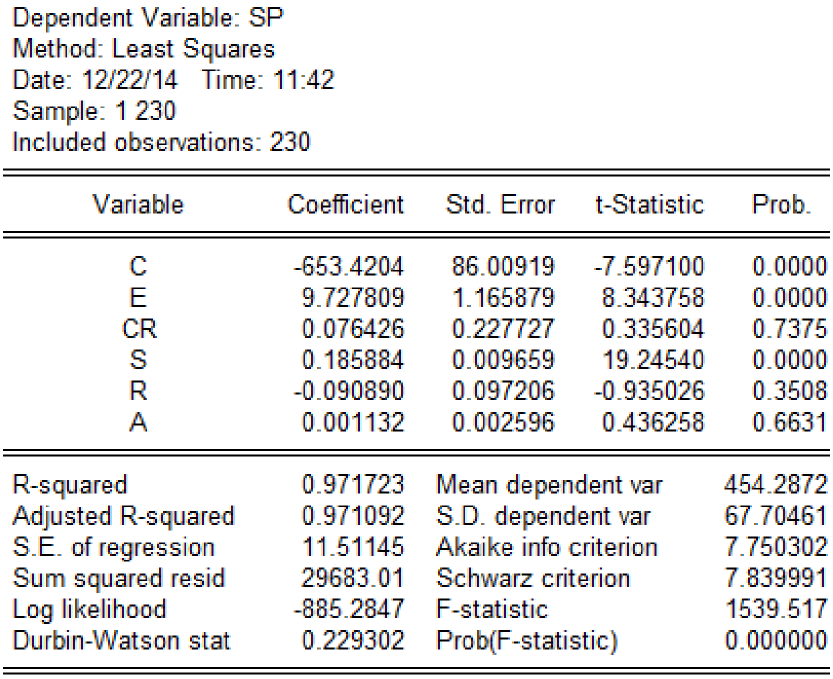
We have tried to forecast TATA Steel's stock price (denoted as **SP** in the regression model) using 5 independent variables as listed below.

S	Steel Index
R	Realty Index
E	Exchange Rate
CR	Crude Oil Price
A	Auto Index

For this purpose, we have taken last 1 year's data of all the variables (data source mentioned in appendix). The total number of observations are for 242 days, starting from 16th Dec 2013 to 12th Dec 2014. For model development, we have used the first 230 cases & reserved the 12 cases for model validation. Regression analysis was done using EViews 8.0

1. Regression Model 1

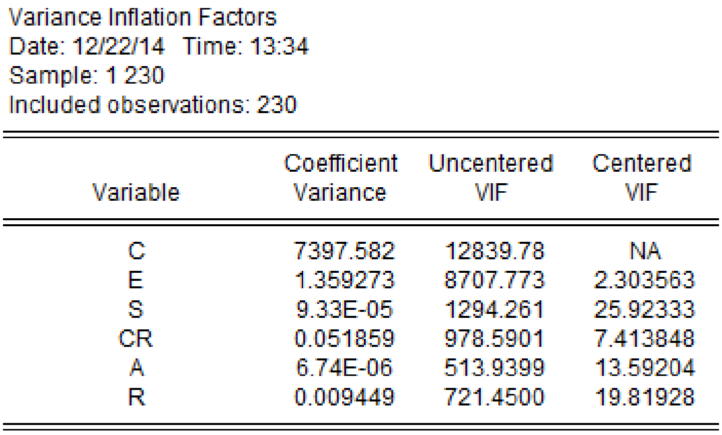
A simple linear model was estimated using EViews 8.0. The estimated equation is given below.



```
Dependent Variable: SP
Method: Least Squares
Date: 12/22/14   Time: 11:42
Sample: 1 230
Included observations: 230
==========================================================================
       Variable        Coefficient    Std. Error    t-Statistic    Prob.
==========================================================================
          C            -653.4204      86.00919      -7.597100     0.0000
          E             9.727809      1.165879       8.343758     0.0000
          CR            0.076426      0.227727       0.335604     0.7375
          S             0.185884      0.009659       19.24540     0.0000
          R            -0.090890      0.097206      -0.935026     0.3508
          A             0.001132      0.002596       0.436258     0.6631
==========================================================================
R-squared               0.971723    Mean dependent var         454.2872
Adjusted R-squared      0.971092    S.D. dependent var         67.70461
S.E. of regression      11.51145    Akaike info criterion      7.750302
Sum squared resid       29683.01    Schwarz criterion          7.839991
Log likelihood         -885.2847    F-statistic                1539.517
Durbin-Watson stat      0.229302    Prob(F-statistic)          0.000000
==========================================================================
```

We can see that the coefficients of **CR**, **A** & **R** are insignificant. To analyze this further, variance inflation factor (**VIF**) for each of the variables were estimated to see if there is a problem of

multi-collinearity. We can also see that the Durbin-Watson statistics is close to “**0**”. Thus, there is problem of **auto-correlation** also.



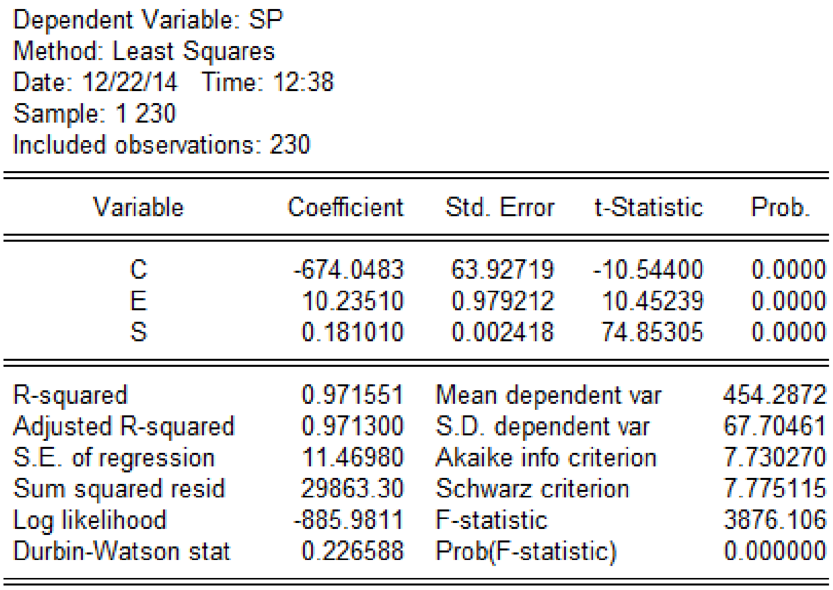
```
Variance Inflation Factors
Date: 12/22/14   Time: 13:34
Sample: 1 230
Included observations: 230
==========================================================
                    Coefficient   Uncentered    Centered
     Variable        Variance        VIF          VIF
==========================================================
        C            7397.582     12839.78        NA
        E            1.359273     8707.773     2.303563
        S            9.33E-05     1294.261     25.92333
        CR           0.051859     978.5901     7.413848
        A            6.74E-06     513.9399     13.59204
        R            0.009449     721.4500     19.81928
==========================================================
```

From the VIF table we see that S, A & R are causing the multi-collinearity problem (VIF > 10).

From the correlation matrix below, we can see that Realty Index (**R**) & Auto Index (**A**) are highly correlated with Steel Index (**S**). This is because **R** & **A** represent the steel demand data which has been already captured by the Steel Index data (**S**), as it represents steel prices (prices are fixed based on demand). Thus, having **R** & **A** makes it redundant as we can have just the steel index data (**S)**, instead of these 2 factors.

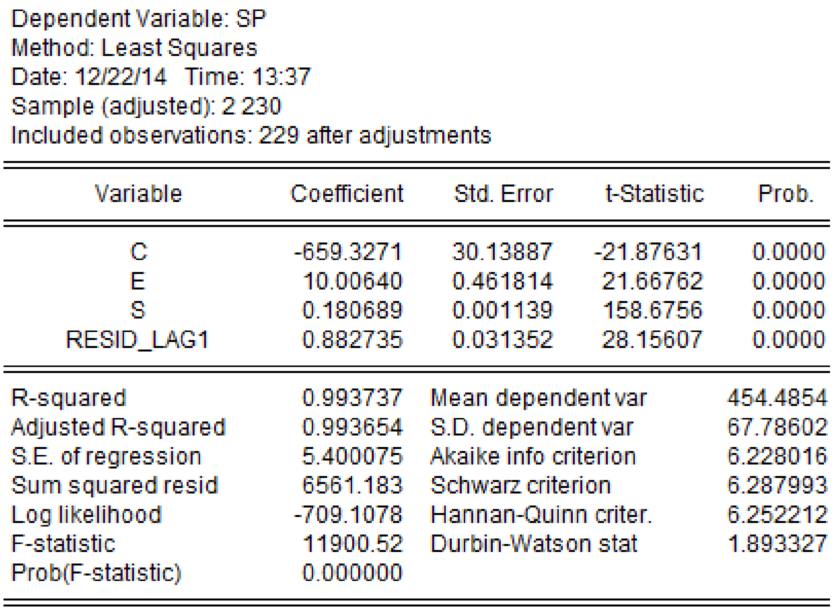
	SP	E	CR	S	R	A
SP	1	-0.50703	-0.08513	0.978208	0.925094	0.656554
E	-0.50703	1	-0.38005	-0.60971	-0.65204	-0.09938
CR	-0.08513	-0.38005	1	-0.04303	-0.02395	-0.74494
S	0.978208	-0.60971	-0.04303	1	0.961904	0.63417
R	0.925094	-0.65204	-0.02395	0.961904	1	0.574067
A	0.656554	-0.09938	-0.74494	0.63417	0.574067	1

We can also see that CR has a very weak relation with SP & thus, it can also be removed. After removing R, A & CR, our estimated equation is as follows.



```
Dependent Variable: SP
Method: Least Squares
Date: 12/22/14   Time: 12:38
Sample: 1 230
Included observations: 230
==================================================================
        Variable       Coefficient   Std. Error   t-Statistic   Prob.
==================================================================
           C            -674.0483    63.92719    -10.54400    0.0000
           E             10.23510    0.979212     10.45239    0.0000
           S             0.181010    0.002418     74.85305    0.0000
==================================================================
R-squared                0.971551   Mean dependent var      454.2872
Adjusted R-squared       0.971300   S.D. dependent var      67.70461
S.E. of regression       11.46980   Akaike info criterion   7.730270
Sum squared resid        29863.30   Schwarz criterion       7.775115
Log likelihood          -885.9811   F-statistic             3876.106
Durbin-Watson stat       0.226588   Prob(F-statistic)       0.000000
==================================================================
```

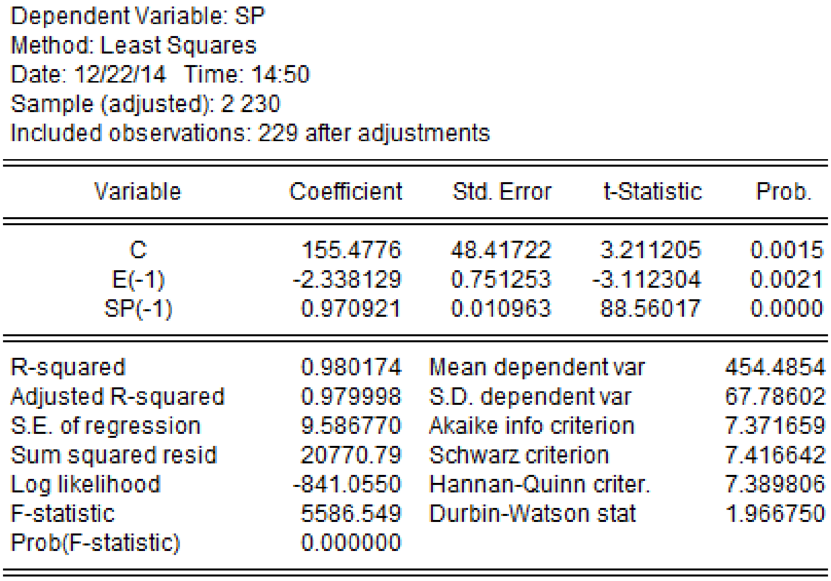
Even in our new model, the Durbin-Watson stat is close to "**0**". To remove this problem of auto correlation, we include the residuals of lag 1. After including the previous residual value as on of the regressors, we get the following model.



```
Dependent Variable: SP
Method: Least Squares
Date: 12/22/14   Time: 13:37
Sample (adjusted): 2 230
Included observations: 229 after adjustments
==================================================================
        Variable       Coefficient   Std. Error   t-Statistic   Prob.
==================================================================
           C            -659.3271    30.13887    -21.87631    0.0000
           E             10.00640    0.461814     21.66762    0.0000
           S             0.180689    0.001139     158.6756    0.0000
      RESID_LAG1         0.882735    0.031352     28.15607    0.0000
==================================================================
R-squared                0.993737   Mean dependent var      454.4854
Adjusted R-squared       0.993654   S.D. dependent var      67.78602
S.E. of regression       5.400075   Akaike info criterion   6.228016
Sum squared resid        6561.183   Schwarz criterion       6.287993
Log likelihood          -709.1078   Hannan-Quinn criter.    6.252212
F-statistic              11900.52   Durbin-Watson stat      1.893327
Prob(F-statistic)        0.000000
==================================================================
```

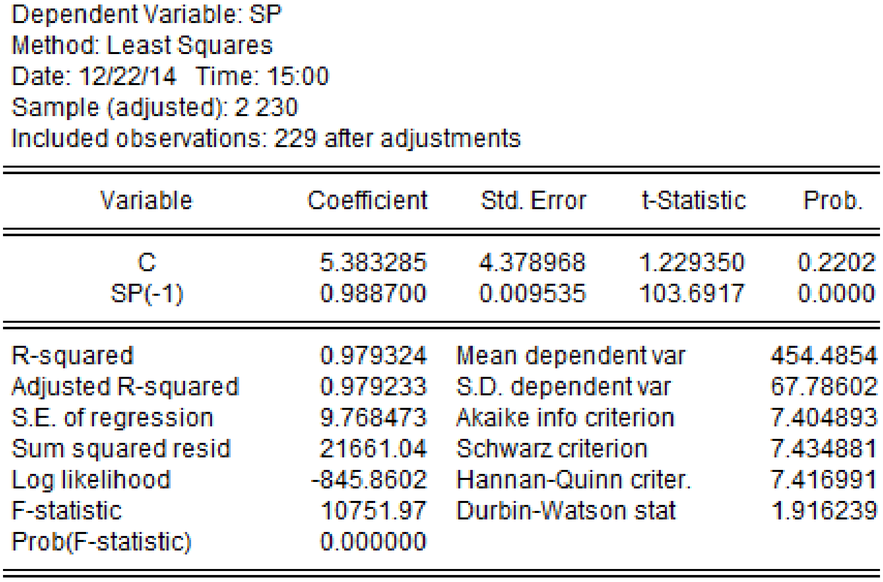
Even if we remove the auto-correlation problem, this, model will be of seldom use as we are using the closing values of Exchange-Rate & Steel Index for each day. Thus, to forecast tomorrow's stock price, we would need the Exchange-Rate & Steel Index for tomorrow (which is not available). We can instead use the previous day's value for these 2 variables to forecast for today. Or else, we have to forecast the two variables independently & then estimate the stock price using the forecasted variables in the above equation (which is a tedious task to perform).

To make this model useful, we take 1 lag of both E & S and use them to forecast for the current stock price. After estimating the equation, it was found to have auto-correlation problem & thus to remove it, an auto-regressive model was used along with the lagged Exchange rate as a factor (even lagged residuals did not help in removing the problem).



```
Dependent Variable: SP
Method: Least Squares
Date: 12/22/14   Time: 14:50
Sample (adjusted): 2 230
Included observations: 229 after adjustments

     Variable          Coefficient    Std. Error    t-Statistic    Prob.

        C               155.4776      48.41722      3.211205      0.0015
      E(-1)            -2.338129      0.751253     -3.112304      0.0021
      SP(-1)            0.970921      0.010963      88.56017      0.0000

R-squared               0.980174    Mean dependent var      454.4854
Adjusted R-squared      0.979998    S.D. dependent var      67.78602
S.E. of regression      9.586770    Akaike info criterion   7.371659
Sum squared resid       20770.79    Schwarz criterion       7.416642
Log likelihood         -841.0550    Hannan-Quinn criter.    7.389806
F-statistic             5586.549    Durbin-Watson stat      1.966750
Prob(F-statistic)       0.000000
```

Even after removing the lagged Exchange rates, we did not see a significant decrease in the R^2 value. This suggests that SP is heavily dependent on its past values.



```
Dependent Variable: SP
Method: Least Squares
Date: 12/22/14   Time: 15:00
Sample (adjusted): 2 230
Included observations: 229 after adjustments

     Variable          Coefficient    Std. Error    t-Statistic    Prob.

        C               5.383285      4.378968      1.229350      0.2202
      SP(-1)            0.988700      0.009535      103.6917      0.0000

R-squared               0.979324    Mean dependent var      454.4854
Adjusted R-squared      0.979233    S.D. dependent var      67.78602
S.E. of regression      9.768473    Akaike info criterion   7.404893
Sum squared resid       21661.04    Schwarz criterion       7.434881
Log likelihood         -845.8602    Hannan-Quinn criter.    7.416991
F-statistic             10751.97    Durbin-Watson stat      1.916239
Prob(F-statistic)       0.000000
```

Thus, in predicting stock prices using a multiple-linear-regression, our chosen set of factors was not very helpful. There are 2 interpretations from this analysis:

a. Majority of the useful information is lost in the errors or the lagged value of the dependent variable (stock price) itself. Thus, a model that uses just the stock price data (like ARIMA or Holt-Winter's method) will be much more helpful.
b. The relationship established by our multiple-linear-regression model was not complex enough to establish the right relationship between the chosen set of independent variables & stock price. Thus, a model like Artificial Neural Networks would be helpful in predicting the stock price.

II. Artificial Neural Network

Artificial Neural nets are known to have a high predictive accuracy, but the quality of output depends heavily on the input variables we choose for modeling. From our previous analysis, we know that Exchange rate & Steel Index have a very good relation with Stock prices. Therefore, we only use these 2 factors in constructing our neural network.

The following trends were observed during the hit-and-trial approach to arrive at the best model:

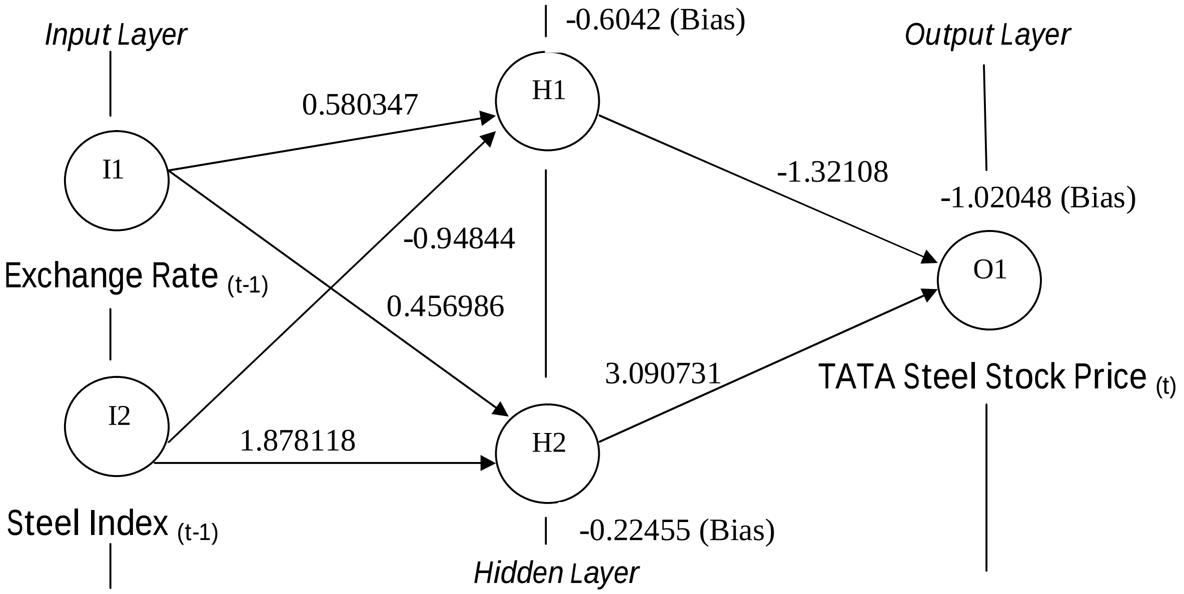
1. The validation error was the least when the weight decay factor was "0". Validation error (RMS) increases exponentially when the weight decay factor increases from 0 to 0.5.
2. As the number of units increased from 2 to 5 in hidden layer 1, the validation error reached a minimum level & started increasing when the number of units increased from 6 & beyond.
3. As the number of epochs increased from 30 to 40, the least value of validation error occurred when the model had 35 epochs (for 1 hidden layer & 5 neurons).
4. Number of epochs that give us a lowest validation error (RMS) is dependent on the number of neurons in the hidden layer. Lower the number neurons, higher the number of epochs required.

This way of reducing the number of neurons & increasing the number of epochs gave us a simpler and a good-fit model.

According to theory, **learning rate** controls the rate at which the weights are adjusted based on the errors generated from the output node. Whereas, **weight decay factor** is a form of regularizing the magnitudes of weights so as to keep them at a low value. This way we are controlling high variations in the output variable because of the un-controlled magnitude of weights in the intermediate layers.

Weight decay adds a penalty term to the error function that is being generated. The usual penalty is the sum of squared weights times a **decay constant**. This modified error function, is then used to update the weights depending on the **learning rate**. Thus, **weight decay constant** & **learning rate** are 2 different constants, playing 2 different roles in updating errors & weights. This was confirmed by taking different epochs too see if the weights updated or not, if weight decay is 0. We saw that the weights kept updating even though the weight decay is 0, which confirms that learning rate & weight decay is not the same.

Our final model can be visualized as follows:



Parameters/Options	
Input variables normalized	Yes
Network Architecture	Manual
Seed: Initial Weights	12345
# Hidden Layers	1
# Nodes in Hidden Layer 1	2
# of Epochs	125
Step size for gradient descent	0.1
Weight change momentum	0.6
Error tolerance	0.01
Weight decay	0
Hidden layer activation function	Standard
Output layer activation function	Standard

Inter-Layer Connections Weights

	Input Layer		
Hidden Layer 1	**E_Lag1**	**S_Lag1**	**Bias**
Neuron 1	0.580347	-0.94844	-0.6042
Neuron 2	0.456986	1.878118	-0.22455

	Hidden Layer 1		
Output Layer	**Neuron 1**	**Neuron 2**	**Bias**
Response	-1.32108	3.090731	-1.02048

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
20047.04914	12.05274	-0.06466

Validation Data Scoring - Summary Report

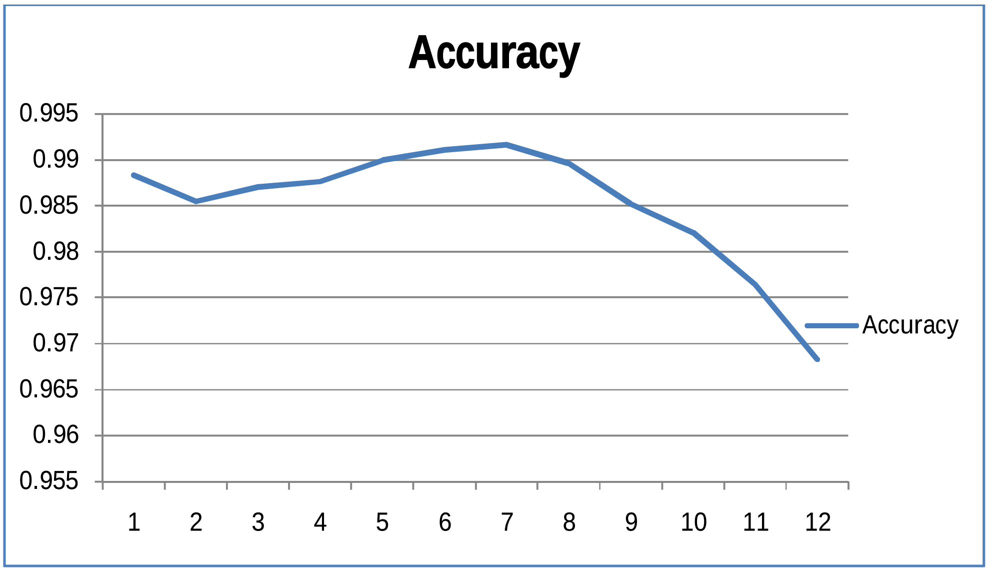
Total sum of squared errors	RMS Error	Average Error
13039.25334	11.97032	-1.98789

Input variables were normalized as from experience and theory it is known that neural nets give best results when the variables are lying in the range of 0 to 1 or similarly smaller values. We have taken 229 cases & split them into training and validation sets to build and evaluate the neural network (results shown above). We have reserved 12 days data as test data to check for the forecast accuracy of the model (27th Nov’14 to 12th Dec’14).

Since the neural net uses a logistic sigmoid function to estimate the output from each node from each layer, we use the same method to estimate the forecasted values of Stock price (SP) for the above mentioned period. Variables were normalized and then processed through to the model. The output variable (forecasted stock price) was also in the normalized form& then was converted back to the actual price level. Results of the forecast are tabulated below.

Date	SP	SP Forecasted	APE	MAPE	Accuracy
27-Nov-14	460.2	465.576871	0.011684	0.011684	0.988316
28-Nov-14	473.35	465.1438653	0.017336	0.01451	0.98549
1-Dec-14	461.15	465.7173899	0.009904	0.012975	0.987025
2-Dec-14	465.2	460.3583086	0.010408	0.012333	0.987667
3-Dec-14	463.75	464.1117392	0.00078	0.010022	0.989978
4-Dec-14	461.85	463.3676627	0.003286	0.0089	0.9911
5-Dec-14	461.15	463.4811134	0.005055	0.00835	0.99165
8-Dec-14	451.85	463.0720588	0.024836	0.010411	0.989589
9-Dec-14	436.3	458.1050813	0.049977	0.014807	0.985193
10-Dec-14	432.5	452.2342132	0.045628	0.017889	0.982111
11-Dec-14	418.75	452.1946178	0.079868	0.023524	0.976476
12-Dec-14	402.7	451.6738245	0.121614	0.031698	0.968302

We can see from the table that the **Accuracy** (1-MAPE) drops after 7 days. The accuracy for the model when forecasted for the first 7 days is **99.165%** & for the whole of 12 days, it is **96.83%**. The accuracy pattern is presented in the graph below.



We have thus found a really good forecasting model (an artificial neural network) which uses the previous day's Exchange rate & Steel Index to forecast today's stock price.

III. Winter-Holt's Method

Every time series data will have either one or more of the following components:

a. Trend (Change in the mean level of data)
b. Seasonality (predictable seasonal variations)
c. Cyclicity (cyclic changes that cannot be predicted. For example, recessions)
d. Irregular Fluctuations

The Winter-Holt method gives a mathematical approach for us to model the level (average value), trend (rate at which the level grows) & seasonality in the data, to effectively forecast for the future. The mathematical methodology is shown by the following set of equations.

$$\begin{aligned} s_0 &= x_0 \\ s_t &= \alpha \frac{x_t}{c_{t-L}} + (1-\alpha)(s_{t-1} + b_{t-1}) \\ b_t &= \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1} \\ c_t &= \gamma \frac{x_t}{s_t} + (1-\gamma)c_{t-L} \end{aligned}$$

The terms "s", "b" & "c" represent the **level**, **trend** & **seasonality** values at time period "t". **α** is the *data smoothing factor*, $0 < \alpha < 1$, **β** is the *trend smoothing factor*, $0 < \beta < 1$, and **γ** is the *seasonal change smoothing factor*, $0 < \gamma < 1$.

The forecasted value for period "t" is given by:

$$F_{t+m} = (s_t + mb_t)c_{t-L+1+(m-1)}$$

Where, "L" is the periodicity or the seasonal length in the data. For example, if the annual automobile sales peaks every January, then the seasonal length is 12 months or L = 12 (we are forecasting for monthly sales in this case).

Our goal is to find the best values of **α, β & γ** that will give us the least error in the training set, using XL Solver Add-in. We then use these constants to forecast for the hold-out sample (27th Nov'14 to 12th Dec'14) & find out the validation error.

Stock price data usually has a seasonality of 5 days, i.e. every day in a week has some seasonal component to it & this seasonal effect repeats itself for the same day, in the coming week (as the market is open only for 5 days a week).

The first 5 cases in the training set are taken for initialization i.e. to set the initial values of levels, trends & seasonal indices. The analysis is presented in the table below.

Using the Solver Add-in, the best value of the 3 constants that give us the least MSE (Mean squared error) were found as:

Alpha	Beta	Gamma
0.947007	0.01558	1

The values of Alpha & Gamma are close to 1, which means that lesser emphasis is given on data far away in the past to estimate the seasonality & level for the period “t”. Instead, the immediate past has a higher weight in estimating them. In other words, seasonality is not constant in the data, it keeps changing every week. The same applies for the **level** part. **Level** can be imagined as the actual value after removing any seasonal effects from it.

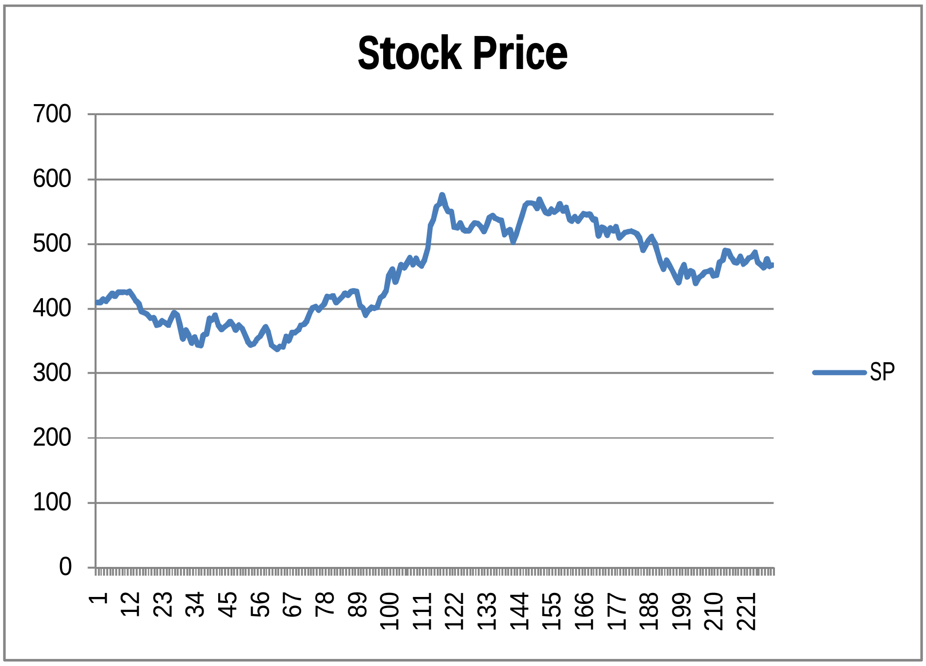
Moreover, the seasonal indices for every day in the training set was close to 1, which means that there is very less seasonal effect on the data or the data has very less **seasonality**.

The beta value is close to 0, which tells us that the data has a more or less a **constant trend**.

One limitation of this method is that we can forecast just for a limited time period. This is because we have the seasonal indices only for the next 5 days. Our forecast analysis on the hold out sample is given below:

Date	SP	Ft	Error	APE	MAPE	Accuracy
27-Nov-14	460.2	463.4883	-3.28825	0.007145	0.007145	0.992855
28-Nov-14	473.35	464.3984	8.951624	0.018911	0.013028	0.986972
1-Dec-14	461.15	464.4611	-3.3111	0.00718	0.011079	0.988921
2-Dec-14	465.2	466.1289	-0.92894	0.001997	0.008808	0.991192
3-Dec-14	463.75	465.3698	-1.61978	0.003493	0.007745	0.992255

We can see that the Accuracy (1-MAPE) is a whopping **99.225 %** for the first 5 days of the forecast. But we cannot say that this method is the best because the seasonal effects or the change in trend, are not that great to capture the high volatility in the stock prices. By our model, the data just has a more or less constant trend. So our forecast line will fluctuate far less than the actual fluctuations in stock price. To illustrate this, let us observe the stock price pattern in the graph below:

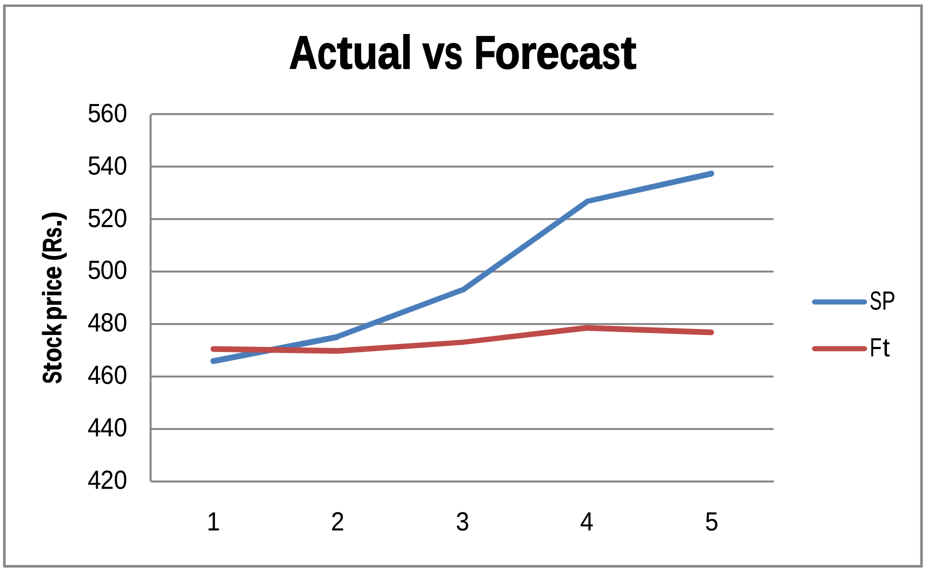


If we see the graph above, from observations 111 to 115 show a steep increase in stock price. If we use the same methodology to forecast for this period (using the previous dates as test data), we get the following results:

Alpha	Beta	Gamma
0.925488	0.018994	1

Date	SP	Ft	APE	MAPE	Accuracy
29-May-14	465.95	470.2988	0.009333	0.009333	0.990667
30-May-14	475.1	469.7105	0.011344	0.010339	0.989661
2-Jun-14	492.9	472.9928	0.040388	0.020355	0.979645
3-Jun-14	526.8	478.3882	0.091898	0.038241	0.961759
4-Jun-14	537.4	476.7867	0.11279	0.053151	0.946849

We can now see that the accuracy drops to **94.68 %**. Therefore, it is not a highly reliable model for forecasting during the periods of high volatility in stock prices. Our theory that the forecast will maintain a more or less constant trend can be confirmed from the graph below (for the forecast of this volatile period):

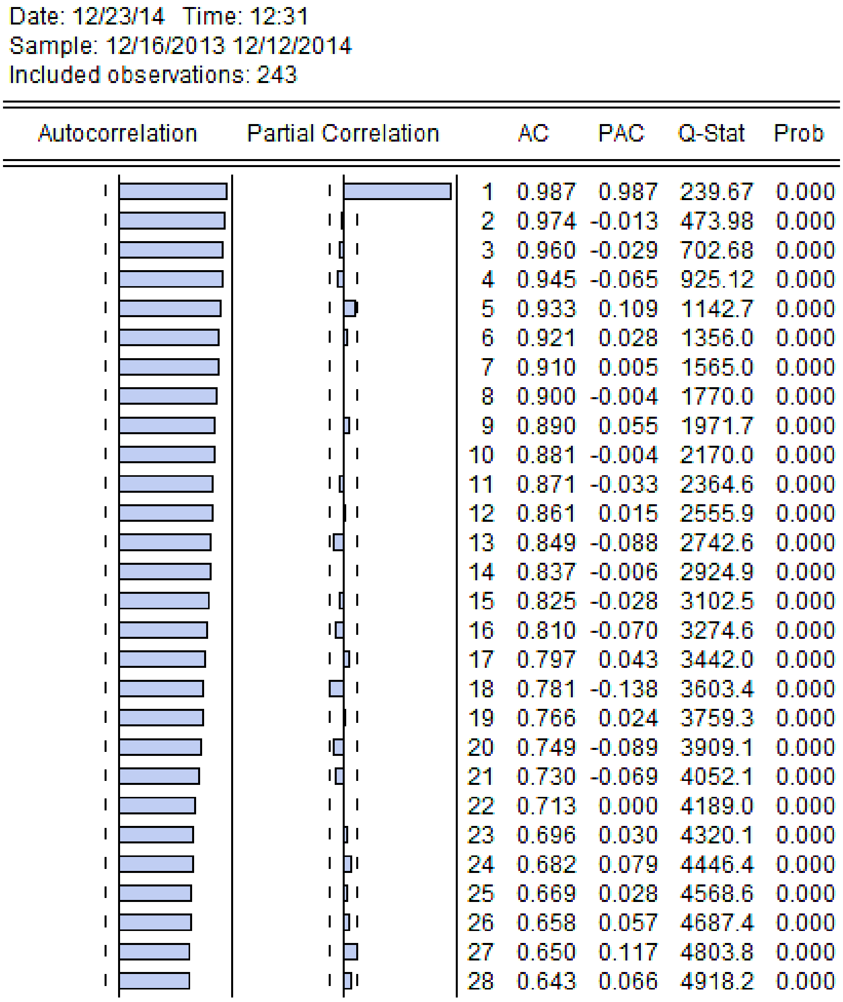


IV. ARIMA / SARIMA Modeling

The Auto-Regressive Integrated Moving Average (ARIMA) methodology gives us a way to forecast a time series data by taking into account, the dependency it has on its past data. It assumes that most of the variation in the data can be captured through the information stored in the data pattern itself. The model can only be applied for a stationary data set. Thus, if our data is non-stationary, we make it stationary & then apply the model.

1. Checking the stock price data for stationarity

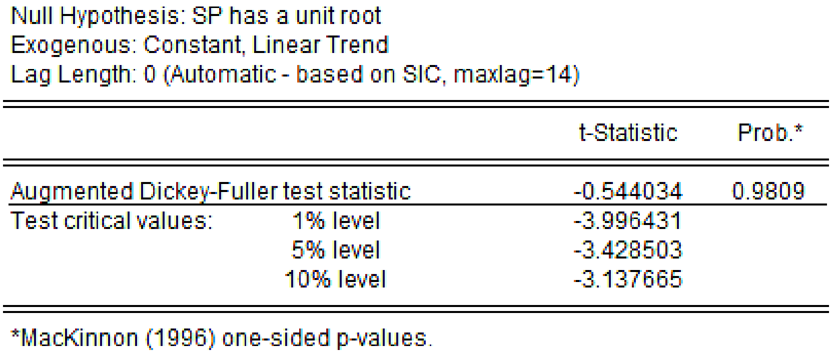
From the **correlogram** of the stock price data, we see that the data is non-stationary. The **ACF** function is highly significant at most of the lags & the significance is decreasing very slowly.



Date: 12/23/14 Time: 12:31
Sample: 12/16/2013 12/12/2014
Included observations: 243

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.987	0.987	239.67	0.000
		2	0.974	-0.013	473.98	0.000
		3	0.960	-0.029	702.68	0.000
		4	0.945	-0.065	925.12	0.000
		5	0.933	0.109	1142.7	0.000
		6	0.921	0.028	1356.0	0.000
		7	0.910	0.005	1565.0	0.000
		8	0.900	-0.004	1770.0	0.000
		9	0.890	0.055	1971.7	0.000
		10	0.881	-0.004	2170.0	0.000
		11	0.871	-0.033	2364.6	0.000
		12	0.861	0.015	2555.9	0.000
		13	0.849	-0.088	2742.6	0.000
		14	0.837	-0.006	2924.9	0.000
		15	0.825	-0.028	3102.5	0.000
		16	0.810	-0.070	3274.6	0.000
		17	0.797	0.043	3442.0	0.000
		18	0.781	-0.138	3603.4	0.000
		19	0.766	0.024	3759.3	0.000
		20	0.749	-0.089	3909.1	0.000
		21	0.730	-0.069	4052.1	0.000
		22	0.713	0.000	4189.0	0.000
		23	0.696	0.030	4320.1	0.000
		24	0.682	0.079	4446.4	0.000
		25	0.669	0.028	4568.6	0.000
		26	0.658	0.057	4687.4	0.000
		27	0.650	0.117	4803.8	0.000
		28	0.643	0.066	4918.2	0.000

To confirm this further, we conduct a Unit-Root test to check for stationarity.



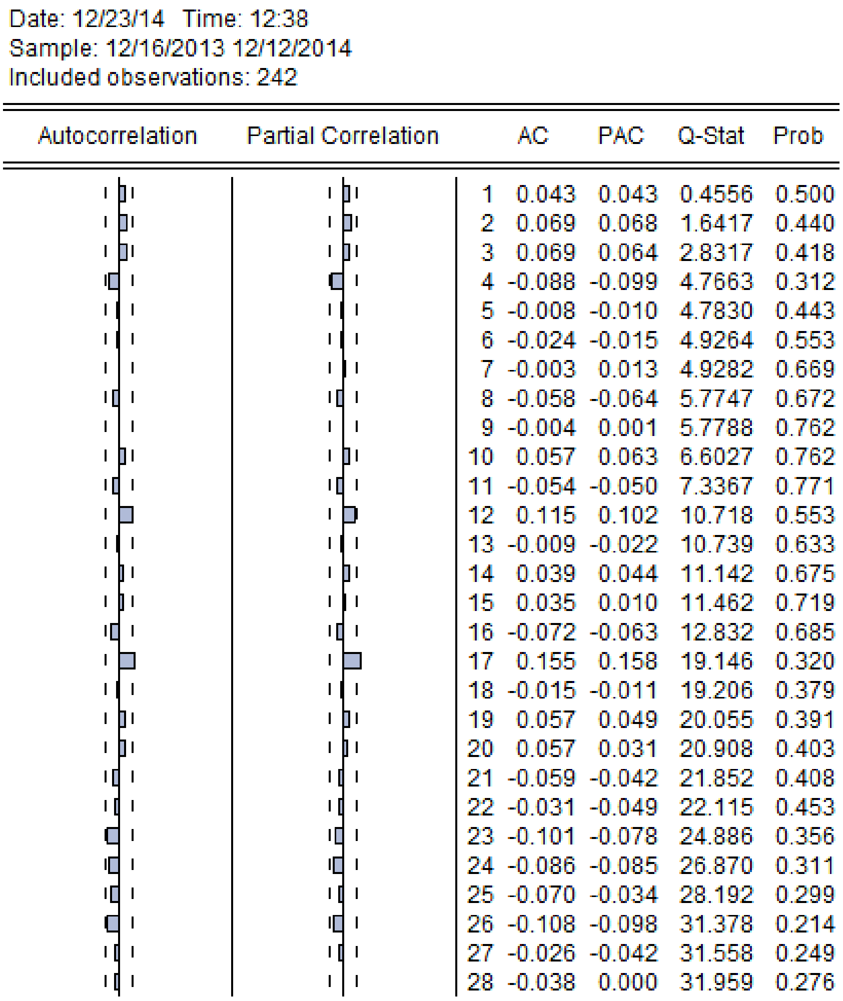
Null Hypothesis: SP has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-0.544034	0.9809
Test critical values:	1% level	-3.996431	
	5% level	-3.428503	
	10% level	-3.137665	

*MacKinnon (1996) one-sided p-values.

We can see that the absolute value of the **ADF test statistic** is less than the **t-statistic** at **5%** level. Thus, we accept the null hypothesis that the data is non-stationary.

To make the data stationary, we take the first difference of the stock price data & then observe the correlogram for stationarity.

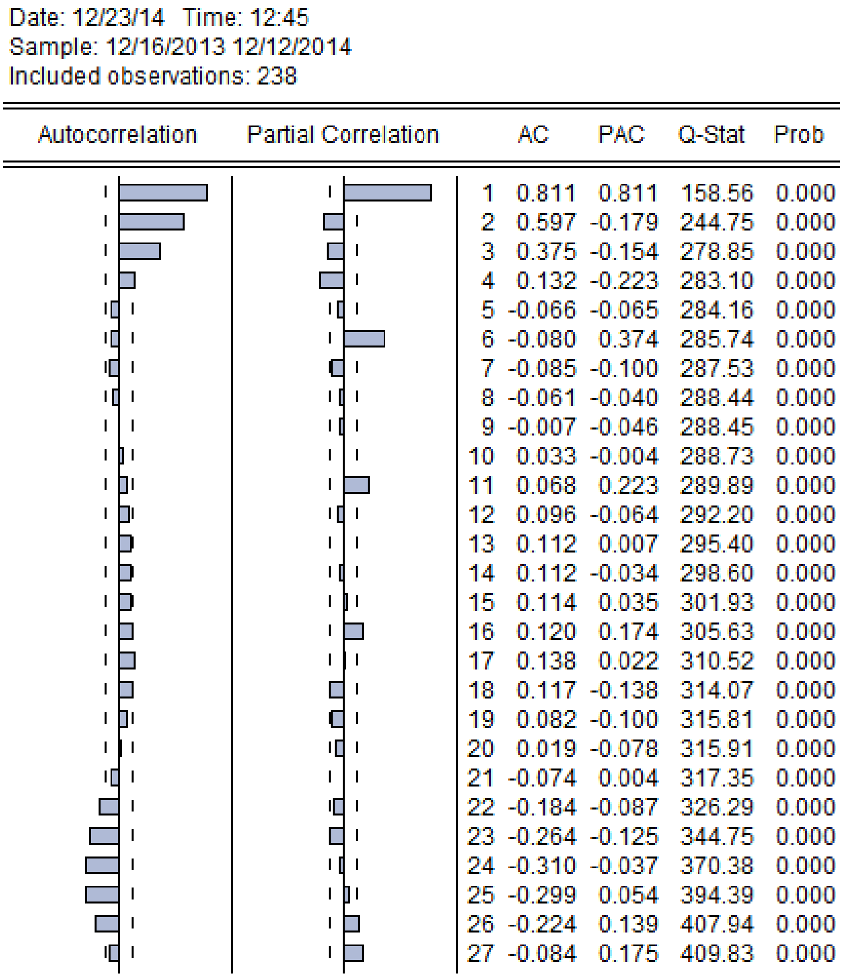


```
Date: 12/23/14   Time: 12:38
Sample: 12/16/2013 12/12/2014
Included observations: 242

Autocorrelation   Partial Correlation       AC      PAC     Q-Stat   Prob

                                        1   0.043   0.043   0.4556   0.500
                                        2   0.069   0.068   1.6417   0.440
                                        3   0.069   0.064   2.8317   0.418
                                        4  -0.088  -0.099   4.7663   0.312
                                        5  -0.008  -0.010   4.7830   0.443
                                        6  -0.024  -0.015   4.9264   0.553
                                        7  -0.003   0.013   4.9282   0.669
                                        8  -0.058  -0.064   5.7747   0.672
                                        9  -0.004   0.001   5.7788   0.762
                                       10   0.057   0.063   6.6027   0.762
                                       11  -0.054  -0.050   7.3367   0.771
                                       12   0.115   0.102   10.718   0.553
                                       13  -0.009  -0.022   10.739   0.633
                                       14   0.039   0.044   11.142   0.675
                                       15   0.035   0.010   11.462   0.719
                                       16  -0.072  -0.063   12.832   0.685
                                       17   0.155   0.158   19.146   0.320
                                       18  -0.015  -0.011   19.206   0.379
                                       19   0.057   0.049   20.055   0.391
                                       20   0.057   0.031   20.908   0.403
                                       21  -0.059  -0.042   21.852   0.408
                                       22  -0.031  -0.049   22.115   0.453
                                       23  -0.101  -0.078   24.886   0.356
                                       24  -0.086  -0.085   26.870   0.311
                                       25  -0.070  -0.034   28.192   0.299
                                       26  -0.108  -0.098   31.378   0.214
                                       27  -0.026  -0.042   31.558   0.249
                                       28  -0.038   0.000   31.959   0.276
```

From the ACF function, we see that the data follows a **purely random process** i.e. there are no interdependencies in the data set now & thus we cannot apply an ARIMA model on such a data.

Stock price data usually has a seasonality of 5 days, i.e. every day in a week has some seasonal component to it & this seasonal effect repeats itself for the same day, in the coming week (as the market is open only for 5 days a week). Thus, we try a seasonal differencing method to remove the stationarity in the data (with a lag of 5 days).

On observing the correlogram of the seasonally differenced data, we see that the data has now become stationary (from the **ACF** function).



```
Date: 12/23/14   Time: 12:45
Sample: 12/16/2013 12/12/2014
Included observations: 238

Autocorrelation   Partial Correlation        AC     PAC    Q-Stat  Prob

                                         1  0.811   0.811  158.56  0.000
                                         2  0.597  -0.179  244.75  0.000
                                         3  0.375  -0.154  278.85  0.000
                                         4  0.132  -0.223  283.10  0.000
                                         5 -0.066  -0.065  284.16  0.000
                                         6 -0.080   0.374  285.74  0.000
                                         7 -0.085  -0.100  287.53  0.000
                                         8 -0.061  -0.040  288.44  0.000
                                         9 -0.007  -0.046  288.45  0.000
                                        10  0.033  -0.004  288.73  0.000
                                        11  0.068   0.223  289.89  0.000
                                        12  0.096  -0.064  292.20  0.000
                                        13  0.112   0.007  295.40  0.000
                                        14  0.112  -0.034  298.60  0.000
                                        15  0.114   0.035  301.93  0.000
                                        16  0.120   0.174  305.63  0.000
                                        17  0.138   0.022  310.52  0.000
                                        18  0.117  -0.138  314.07  0.000
                                        19  0.082  -0.100  315.81  0.000
                                        20  0.019  -0.078  315.91  0.000
                                        21 -0.074   0.004  317.35  0.000
                                        22 -0.184  -0.087  326.29  0.000
                                        23 -0.264  -0.125  344.75  0.000
                                        24 -0.310  -0.037  370.38  0.000
                                        25 -0.299   0.054  394.39  0.000
                                        26 -0.224   0.139  407.94  0.000
                                        27 -0.084   0.175  409.83  0.000
```

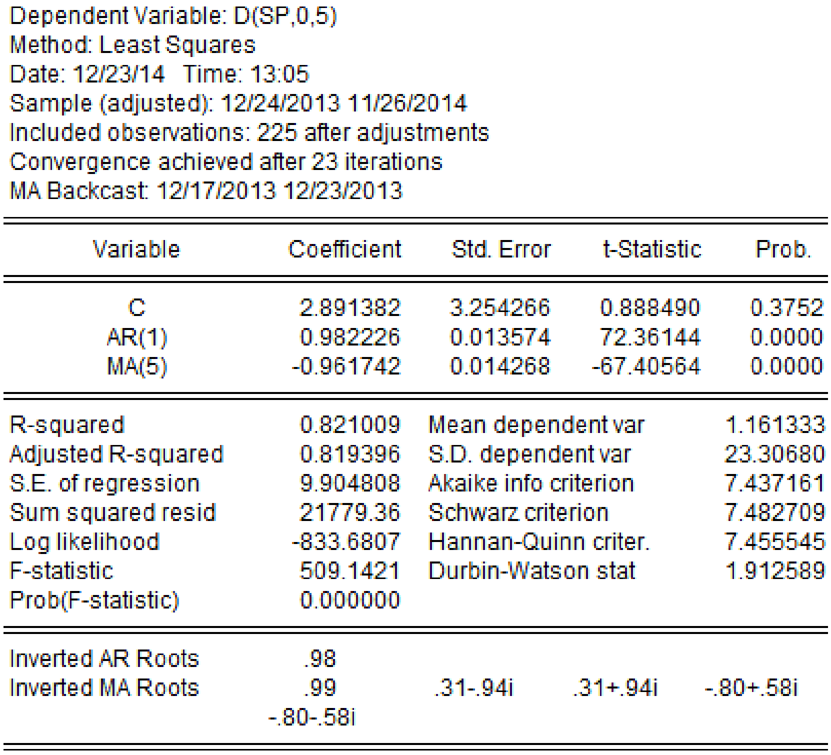
The **ACF function** is significant for the first few lags & then exponentially reduces and becomes insignificant. This pattern tells us that the data has become stationary. We cannot conduct a Unit-Root test on a seasonally differenced data in EVeiws 8.0 thus, we confirm the stationarity just from the correlogram.

2. Estimating the ARIMA equation

We now see the **PACF function** to estimate the best equation for ARIMA modeling.

We see that **Auto-Regressors (AR)** of lag 1, 6, 11, 16 are significant. To approximate thi seasonal effect in AR, using a **seasonal moving average term of lag 5 (SMA(5))**.

Moving averages (MA) of lag 2, 3 & 4 are significant. Thus, we take different combinations of ARs & MAs to get the best equation for our model. Since we are performing an ARIMA on a seasonally differenced time series data & that we have a seasonal term in the model (SMA(5)), the model is now called **SARIMA** (Seasonal ARIMA). The best-fit equation is presented below.



```
Dependent Variable: D(SP,0,5)
Method: Least Squares
Date: 12/23/14   Time: 13:05
Sample (adjusted): 12/24/2013 11/26/2014
Included observations: 225 after adjustments
Convergence achieved after 23 iterations
MA Backcast: 12/17/2013 12/23/2013

          Variable        Coefficient    Std. Error    t-Statistic     Prob.

              C             2.891382      3.254266      0.888490      0.3752
            AR(1)           0.982226      0.013574      72.36144      0.0000
            MA(5)          -0.961742      0.014268     -67.40564      0.0000

R-squared                   0.821009   Mean dependent var        1.161333
Adjusted R-squared          0.819396   S.D. dependent var        23.30680
S.E. of regression          9.904808   Akaike info criterion     7.437161
Sum squared resid           21779.36   Schwarz criterion         7.482709
Log likelihood             -833.6807   Hannan-Quinn criter.      7.455545
F-statistic                 509.1421   Durbin-Watson stat        1.912589
Prob(F-statistic)           0.000000

Inverted AR Roots             .98
Inverted MA Roots             .99         .31-.94i      .31+.94i     -.80+.58i
                            -.80-.58i
```

We have taken the first 225 observations for developing the model (till 26th Nov'14).

As we have no other MA terms other than **SMA(5)**, the final equation looks similar to an ARIMA model. The model can be represented as:

$$\textbf{ARIMA (1,0,0) x (0,1,1)}^{5}$$

The final equation can be derived as follows:

$$Y(t) = SP(t) – SP(t-5)$$

$$Y(t) = 2.891382 + 0.982226 * Y(t-1) -0.961742 * e(t-5)$$

Substituting SP instead of Y(t) in the above equation we get the following equation

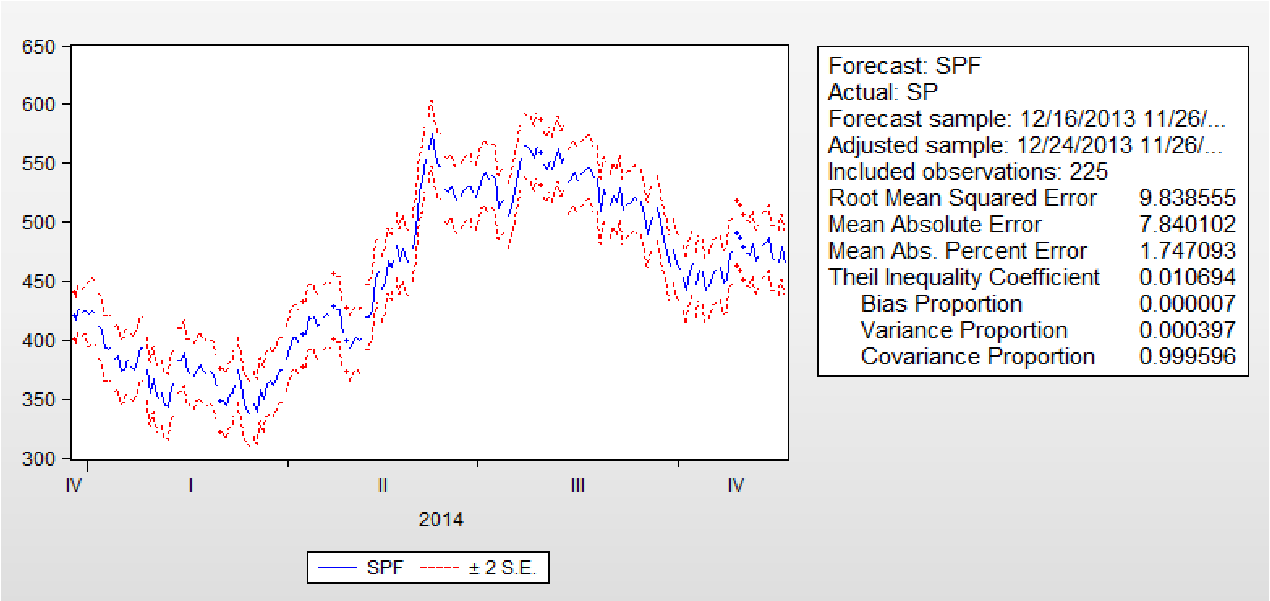
$$SP(t) – SP(t-5) = 2.891382 + 0.982226 * (S(t-1) – S(t-6)) -0.961742 * e(t-5)$$

Simplifying the above equation, we get the final equation as:

$$\boxed{SP(t) = 2.891382 + 0.982226 * S(t-1) + SP(t-5) + 0.982226 * S(t-6) -0.961742 * e(t-5)}$$

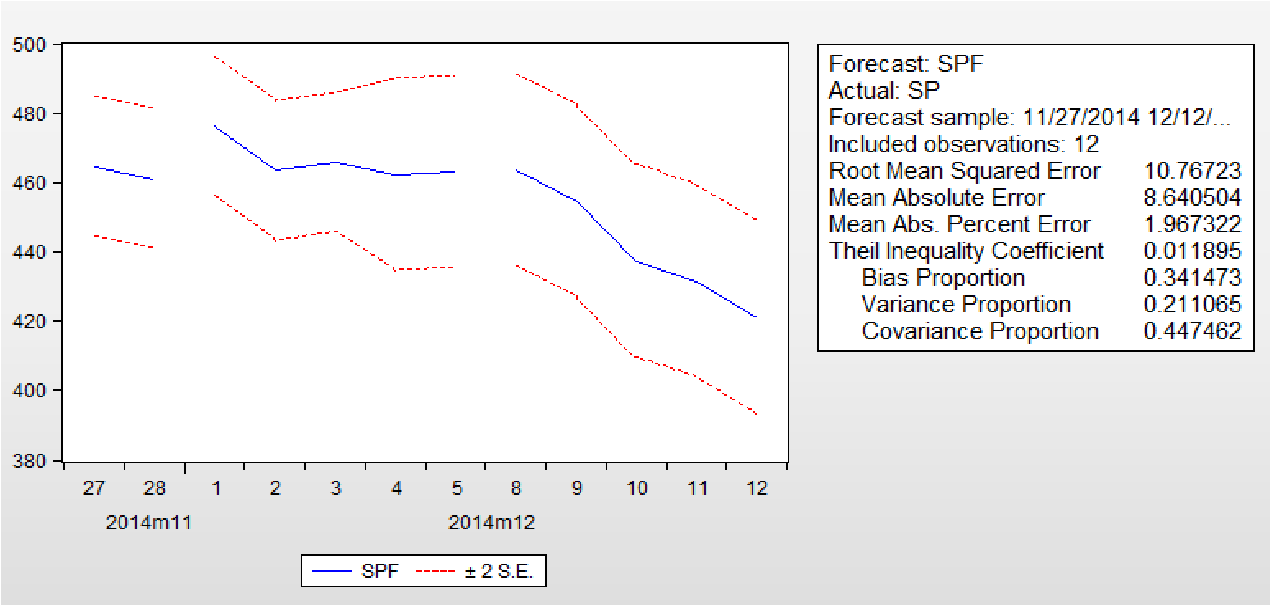
Where **e (t-5)** represents the model residual at time period "t-5" & **SP(t)** represents the stock price at time period "t".

The **in-sample** accuracy measures for the model are shown below.



```
Forecast: SPF
Actual: SP
Forecast sample: 12/16/2013 11/26/...
Adjusted sample: 12/24/2013 11/26/...
Included observations: 225
Root Mean Squared Error        9.838555
Mean Absolute Error            7.840102
Mean Abs. Percent Error        1.747093
Theil Inequality Coefficient   0.010694
     Bias Proportion           0.000007
     Variance Proportion       0.000397
     Covariance Proportion     0.999596
```

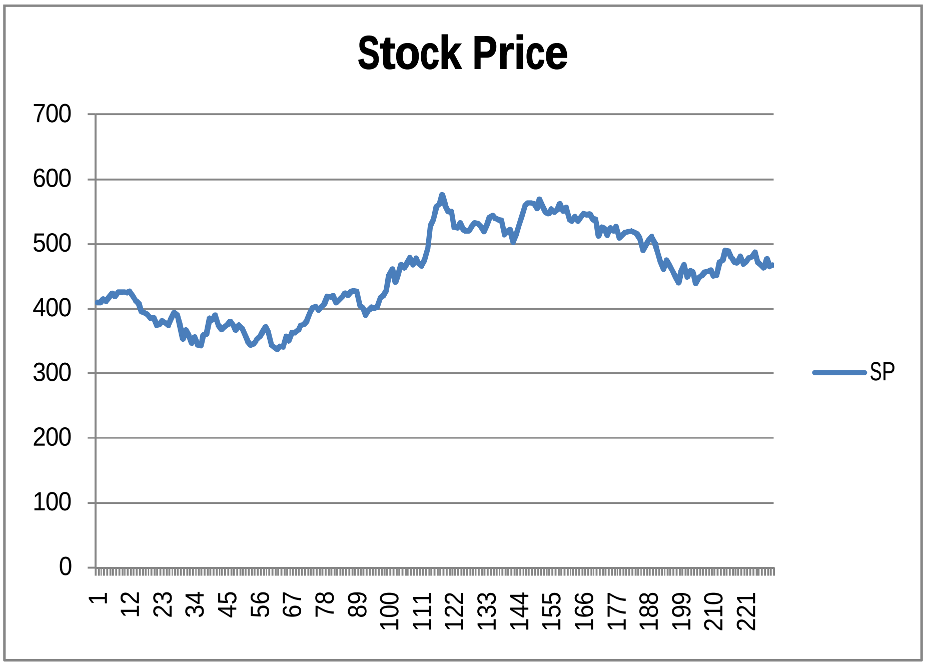
The accuracy measures, for the **hold-out sample**, are shown below.



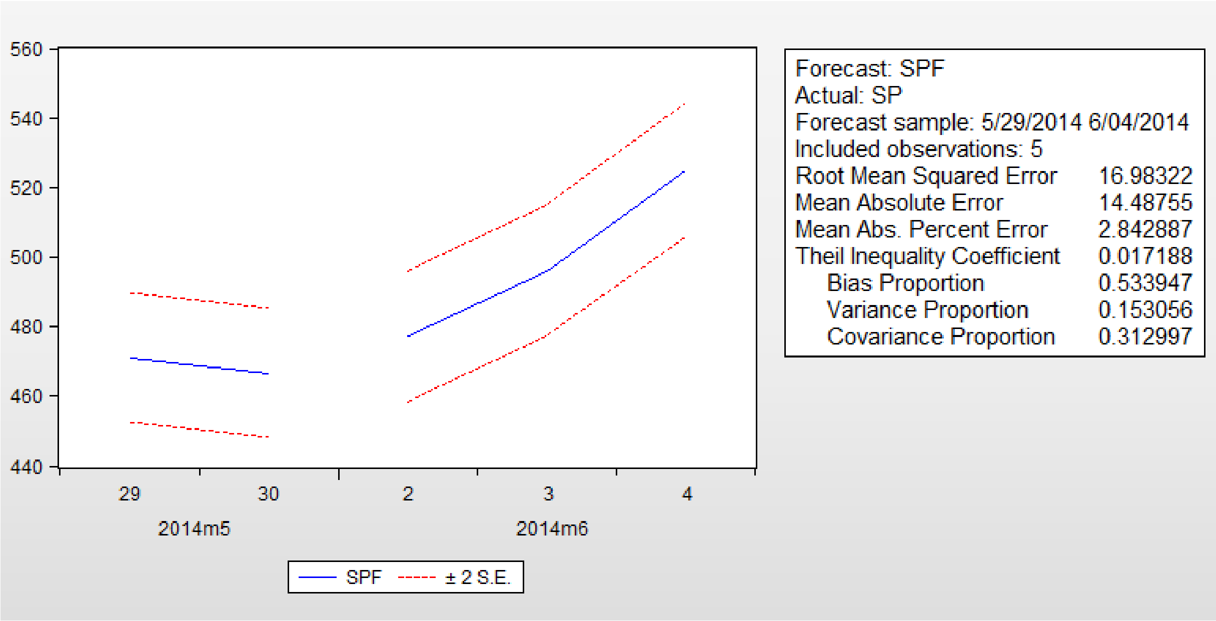
```
Forecast: SPF
Actual: SP
Forecast sample: 11/27/2014 12/12/...
Included observations: 12
Root Mean Squared Error        10.76723
Mean Absolute Error            8.640504
Mean Abs. Percent Error        1.967322
Theil Inequality Coefficient   0.011895
     Bias Proportion           0.341473
     Variance Proportion       0.211065
     Covariance Proportion     0.447462
```

We can see that the hold-out sample Accuracy (100-MAPE) is **98.04%** when the next 12 days out-of-sample data are taken.

If we see the Stock price fluctuations, from observations 111 to 115 show a steep increase in stock price.



If we use the same methodology to forecast for this period (using the previous dates as test data), we get the following results:



```
Forecast: SPF
Actual: SP
Forecast sample: 5/29/2014 6/04/2014
Included observations: 5
Root Mean Squared Error        16.98322
Mean Absolute Error            14.48755
Mean Abs. Percent Error        2.842887
Theil Inequality Coefficient   0.017188
     Bias Proportion           0.533947
     Variance Proportion       0.153056
     Covariance Proportion     0.312997
```

The pattern in the above graph shows that the model predictions also follow the same level of fluctuations as seen in the actual data set. This sample had an accuracy of **97.16%** (100 – 2.84), which was higher than what we got in the Winter-Holt’s model. Thus, the model has a good reliability, along with good prediction accuracy.

Recommendations

Among the four models analyzed, we found that Artificial Neural Networks (ANN) & ARIMA modeling gave us accurate and reliable stock price predictions. Among these 2 models, ARIMA model had a highest forecasting accuracy. Models like ARIMA take into account the information from the past data of stock prices itself. ARIMA gave us the highest forecast accuracy which in-turn tells us that most of the information about the data is stored in the data itself.

Though Artificial neural networks (ANN) captures highly complicated relationships between the factor variables and the output variable (Stock prices), factors like "market sentiment" cannot be captured by a single quantitative measures & are thus, tough to model using the ANN method. Model construction is also relatively tougher in ANNs and the forecasts can also be done reliably only for a short period in time (as the weights need to be updated to capture the evolving relationship between variables). ARIMA modeling method circumvents these limitations & give us more accurate predictions and therefore, is the most preferable method for forecasting stock prices.

Data Sources

http://www.nseindia.com/products/content/equities/equities/eq_security.htm

http://www.investing.com/commodities/brent-oil-historical-data

http://www.oanda.com/currency/historical-rates/

References

ftp://ftp.sas.com/pub/neural/FAQ3.html#A_decay