THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Design and Implementation of Efficient Regression Analysis Techniques in Derivative Market

Cerene Mariam Abraham
Research Scholar, Cochin University of Science and Technology, Kochi, India
M. Sudheep Elayidom
Associate Professor, Cochin University of Science and Technology, Kochi, India
T. Santhanakrishnan
Scientist, Ministry of Defence, Kochi, India

Abstract:
The equity market is the spot where almost all major economic transactions in the world happen at an aggressive rate. The shareowners can avail the financial benefits of the companies whose stocks they hold. Among different sectors, derivatives are an emerging transaction area in financial market, which can avoid the sudden change in price, thus controlling the risk. A correct forecasting of this stock market trend beforehand can make huge profits. One of the most authentic way to predict the future is to try to interpret the present and accordingly we have assign our objective as the analysis of Indian Stock Market so as to build a better future for investment. The work till date on this problem seems mostly focused on data mining techniques. Ninety days' data of some of the popular companies such as LT, SBIN, TATAMOTORS, HINDALCO and AXISBANK are taken for this context and indicators such as open interest, number of contracts and deliverable quantity have chosen for regression analyses. This paper analyses and shows the effect of these variables on predicting the rise or fall of closing price of stock future indices.

Keywords: *Temporal data mining, open interest, deliverable quantity, regression analysis*

1. Introduction

The area of data mining and knowledge discovery deals with pulling out of interesting patterns or knowledge from huge amounts of raw data and the goal is to find beguiling patterns which are novel, valid, understandable and useful. Temporal data mining (TDM) is field by which time-related patterns are extracted and itemized from temporal data and it labels the tasks such as segmentation, clustering, classification, forecasting, and indexing of event sequences and time series or sections of time series or sequences [1]. The analysis process begins with a set of temporal data, uses a methodology to build an optimal representation of the structure of the data during which time knowledge is acquired. Stock market data, also known as equity market data is a temporal data. So TDM in stock market helps to extract relevant patterns from the equity data and thus to predict the future stock price.

The forecasting of the stock market is an enchanting task because successful prediction could yield outstanding gain [3]. The economy of a country is steadily influenced and strongly connected by their stock trade's execution. The trait that all equity sectors have in common is the uncertainty. This mark is not competent for the vendor but it is also inexorable whenever the equity market is picked as the investment tool. The foremost way that one can do is to check out to reduce this uncertainty. Stock Market prediction is one apparatus in this operation. Prediction of stock price has been at pivot for years since it can return noteworthy financial profit. Derivative market is a type of financial market whose value is derived from the underlying assets. Prices in derivatives reflect the perception of market about the future. It helps in transfer of risk.

2. Literature Survey

It seems that not too many outputs on derivative market exist. Little work has been done on verifying which all are the most suitable indicators for mining of a given stock data set. There are certain variables that can change the rise or fall of stock movement. The indicators such as Price ratio and Price comparison were used to compare two stock prices in [8].

Studying similarity measures for clustering of similar stocks is described in [9]. They cluster the stocks according to various measures and compare the results to the ground truth clustering based on the standard and poor 500 index. Association rules were used in the prediction in [10]. They construct a transactional database where each transactional record in the database represents one trading day and it contains a list of winners. The closing price is x% more than the previous day's close price where x% is the trading overhead. There they describe the dependencies between trading overhead and closing price of the day.

The effect of neural network model for time series data in forecasting is discussed in [2]. Forecasting accuracy is verified and measured with reference to an Indian stock market index such as Bombay stock exchange and NIFTY midcap 50. The model achieved the lower prediction error and it may be fit into any stock market data. Time series prediction can be done if the source data with less noise term.
The prediction of financial time series data is a highly complicated task due to following reasons; First, Financial time series often act nearly like a random-walk practices. Second, Financial time series are confined by regime shifting. Statistical properties of the time series are divergent at different points in time because the process is time dependent. Third, the financial time series are usually very noisy. There is a huge amount of biased unpredictable day-to-day changes and finally, In the long run, a new prediction technique becomes a part of the process to be predicted. From the historical data, closing price of stock is only selected one for prediction.
In this article, we describe about the problem on section III. Data analysis is performed in section IV. Results and future scope are explained in section V and section VI respectively.

3. Problem Statement

A stock market is a complex information processing system in large scale that responds to a wide range of socio-economic parameters. This paper aims to study and analyse the derivative data related to five large cap companies and proposes the most suited statistical model to predict the future price. Large cap stock refers to the stocks of well-established and large companies that are having a strong market existence and therefore considered as safe speculation. Information regarding large cap companies is readily available in magazines and newspapers. For this work the data from large cap companies such as LT, SBIN, TATAMOTORS, HINDALCO and AXISBANK are taken into account.

3.1. Stock Data

Market data is time related (temporal) and trade related data for a financial gadgets described by a trading venue such as stock exchange. It allows investors and traders to understand the latest price and see historical trends for tools such as derivatives, currencies and fixed income products.
Traditional theory tells that up and down trends in equity market is due to the factors such as earnings per share, inflation, book value, economic strength etc… In this work we are taking the following parameters to analyze the market price of particular stocks.

- Deliverable quantity
- Open interest
- Number of contracts

By taking a period of three months' stock data of five companies, we are evaluating the effect of these predictors on the response variable, price of the stock. Open interest refers to the total number of open contracts on a security. Number of contracts is the total number of contracts on a security. Deliverable volume is the quantity of stocks which actually pass from one group to another.

3.2. Coefficient of Determination

The coefficient of determination, denoted R^2, is a numerical value that stipulates how well data fit a statistical model [4]. An R^2 of value 1 reveals that the regression line fits the data perfectly, while zero value of R^2 shows that the line does not fit the data at all, it is a curve.
It is the ratio of regression sum of squares to total sum of squares. It will give goodness of fit of a model. If the value of R squared is above 0.3, it is said that the model is good to fit.

3.3. Significance of Regression

The test for the significance of regression is carried out using the analysis of variance (ANOVA) [7]. This test is used to check if a linear statistical relationship exists between the response variable and at least one of the predictor variables. The statements of hypotheses are:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ and $H_1 : \beta_j \neq 0$ (1) $\alpha + \beta = \chi.$ (1) (1)

for atleast one j. The significance of regression can be tested using the F-statistic given by

F = MSR/MSE, where MSR = SSR/(p-1) and MSE = SSE/(n-p). From the ANOVA table we can obtain the value of F. We reject the null hypotheses H_0 if the calculated value, F, is greater than the tabled value, F_0.

3.4. Autocorrelation

An Autocorrelation is defined as the relationship between values divided from each other by a given time lag. Durbin-Watson test is used to identify the presence of autocorrelation in the residuals from a regression analysis. The value given by autocorrelation always lies between 0 and 4. If the Durbin-Watson statistic is appreciably less than 2, there is an affirmation of positive serial correlation. If the value is nearly 2, this means the errors are independent, which is primary step in regression analysis.

4. Data Analysis

For the data analysis, three months' stock data of five companies such as LT, SBIN, TATAMOTORS, HINDALCO and AXISBANK are taken into consideration. Data is available from nseindia website and we have used three parameters as regressors initially to check effect of these parameters on the variable, closing price of the stock. We try to fit a linear regression model.

The general linear model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \epsilon \qquad (2)$$

This connects a response variable y to the predictors $x_1, x_2 \ldots . x_{p-1}$. Let n observations be available for the response variable as well as for the predictors.

Regression coefficient of three predictors for five companies are shown below. Here X denotes open interest, Y denotes Number of contracts and Z denotes deliverable quantity or deliverable volume. Null hypotheses states that these three parameters are insignificant.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	1811.082	15.873		114.101	.000
	Z	7.946E-006	.000	.031	.494	.623
	X	-1.196E-005	.000	-.283	-3.037	.004
	Y	-.007	.001	-.698	-6.735	.000

a. Dependent Variable: P

Figure 1: Regression analysis of LT

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	269.586	2.648		101.810	.000
	X	-3.670E-007	.000	-.521	-4.495	.000
	Y	.000	.000	-.406	-3.498	.001
	Z	8.506E-007	.000	.137	2.259	.028

a. Dependent Variable: P

Figure 2: Regression analysis of SBIN

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	382.663	7.935		48.225	.000
	X	-7.214E-007	.000	-.382	-2.541	.014
	Y	-.001	.000	-.463	-3.113	.003
	Z	-4.565E-007	.000	-.032	-.320	.751

a. Dependent Variable: P

Figure 3: Regression analysis of TATAMOTORS

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	106.361	1.675		63.496	.000
	X	-2.994E-007	.000	-.514	-6.516	.000
	Y	-.001	.000	-.417	-5.213	.000
	Z	-1.205E-006	.000	-.124	-1.988	.052

a. Dependent Variable: P

Figure 4: Regression analysis of HINDALCO

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	585.116	4.803		121.819	.000
	X	-2.372E-006	.000	-.629	-5.557	.000
	Y	-.001	.000	-.298	-2.560	.013
	Z	-1.360E-006	.000	-.064	-1.131	.263

a. Dependent Variable: P

Figure 5: Regression analysis of AXISBANK

5. Results and Discussions

We can see that p-value corresponding regressors Z, that is deliverable volume in figures are greater than 0.05. Hence we do not reject the hypothesis that the parameters are zero, that is, any of the parameters is significant and the variables significantly affect the closing price of the stock.

The coefficient of determination R squared is useful in checking the effectiveness of a given model. R squared is defined as the proportion of the total response variation that is explained by the model. For this data, the R squared value is given in the following Table 1. If the value of R squared is greater than 0.3, we can say that the model is good.

The value of R squared under LT suggests that 86 percent of the total response variation is explained in this linear model. The value of R squared under company SBIN suggests that 80.2 percent of the total response variation is explained in this linear model.

The value of R squared under TATAMOTORS suggests that 60.6 percent of the total response variation is explained in this linear model. The value of R squared under HINDALCO suggests that 82.6 percent of the total response variation is explained in this linear model. The value of R squared under AXISBANK suggests that 85.2 percent of the total response variation is explained in this linear model.

Company	R square	Adjusted R^2	Standard Error of Estimation
LT	.867	.860	49.18218
SBIN	.813	.802	8.82758
TATAMOTORS	.628	.606	19.35097
HINDALCO	.835	.826	6.11781
AXISBANK	.860	.852	17.82672

Table 1: R squared value of the five companies

The value of R squared of TATAMOTORS suggests that 60.6 percent of the total response variation is explained in this linear model. The value of R squared of HINDALCO suggests that 82.6 percent of the total response variation is explained in this linear model. The value of R squared under AXISBANK suggests that 85.2 percent of the total response variation is explained in this linear model.
Even though R squared is an important measure for checking the effectiveness of a model, one cannot say the model is effective solely based on the value of R squared. Other factors also need to be taken into consideration.

The basic assumptions on regression are [5]:

- Errors should be independent.
- The Means should be zero and variance should be constant.

The Durbin- Watson test is used to check the dependency. If the value is nearly 2, the errors are independent [6]. The Durbin – Watson values of five companies is as shown in Table 2. From the table, it is clear that the values are not nearly 2, which means that the errors are dependent. This violates the primary step of regression.

Company	R square	Adjusted R^2	Standard Error of Estimation	Durbin-Watson
LT	.867	.860	49.18218	.732
SBIN	.813	.802	8.82758	.694
TATAMOTORS	.628	.606	19.35097	.494
HINDALCO	.835	.826	6.11781	.617
AXISBANK	.860	.852	17.82672	.815

Table 2: Durbin-Watson values of the five companies

6. Results and Discussions

Prediction of stock price has been at pivot for years since it can return noteworthy financial profit. Therefore, forecasting of equity data is a thrilling task. In this paper we have analyzed the stock data of five companies over a period of three months. Through statistical regression techniques, we tried to fit this stock data. The high value of R squared tells that prediction is possible on this data. The main steps for doing regression are to check whether the errors are independent and means should be of value zero and variance should be constant. Autocorrelation is checked by Durbin-Watson statistical test and result on these data shows that the errors are dependent. Thus it violates the regression and in order to make predictions on these data, we move to time series analysis to fit the data. Thus prediction of stock data can be done using time series analysis.

7. Acknowledgment

The author wishes to express her gratitude to Mr. Manoj P. Michel, Cochin Stock Exchange, who introduced to equity domain and for his useful discussions.

8. References

i. Weiqiang Lin, Mehmet A. Orgun, Graham J. Williams. An overview of temporal data mining. The Australasian Data Mining Workshop. 2002 p. 83-89.
ii. Ashok kumar, S. Murugan. Performance Analysis of Indian Stock Market Index using Neural Network Time Series Model. IEEE; 2013.
iii. Aditya Gupta, Bhuwan Dhingra. Stock Market Prediction Using Hidden Markov Models. IEEE; 2012.
iv. Edward R. Dougherty, Seungchan Kim, Yidong Chen. Coefficient of determination in nonlinear signal processing. Signal Processing 2000; p. 2219–2235.
v. Simon G. Thompson, Julian P. T. Higgins. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine 2002; p. 1559–1573.
vi. Mei-Yu Lee. On the Durbin-Watson statistic based on a Z-test in large samples. International Journal Of Computational Economics and Econometrics. 2016.
vii. Samprit Chatterjee, Ali S.Hadi. Regression Analysis by Example. 5th ed. Wiley Series in Probability and Statistics; 2015.
viii. http://www.incrediblecharts.com
ix. Gavrilov M, Anguelov D, Indyk P and Motwani R, "Mining the Stock Market: Which Measure is Best?", Proc. Of 6th ACM Int'l Conference on Knowledge Discovery and Data Mining 2000, pp 487-496.
x. Lu H, Han J and Feng L, "Stock Movement Prediction and N-dimensional Inter Transaction Association Rules", ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1998, pp 12.1-12.7.