IJCSC 0973-7391 Volume 7 • Number 1 Sept 2015 - March 2016 pp. 264-269 available online at www.csjournalss.com

Comparing Multiple Linear Regression and Linear Support Vector Regression for Predicting in Stock Market

Dr. Priyank Thakkar
Associate Professor, CSE Department, Institute of Technology, Nirma University
priyank.thakkar@nirmauni.ac.in

Abstract: The paper focuses on predicting future values of stock market indices using Multiple Linear Regression (MLR) and linear Support Vector Regression (SVR). It attempts to examine performance of these two techniques for the task of predicting future values of stock market indices. Predictions are made for closing values of two stock market indices, one day ahead of time. These stock market indices are S&P BSE Sensex and CNX Nifty from Indian stock markets. A state of indices on any given day is described using ten technical indicators. Both the techniques performed almost identically but it was seen that MLR performed slightly better than linear SVR. Paper examines the prediction performance quantitatively and suggests a few directions for the future work.

Keywords: Stock Market, Multiple Linear Regression, Support Vector Regression.

1. Introduction and Literature Survey

Fundamental and technical analysis are two possible ways to analyse the stock marketswhich can help predictions of future values or prices of stocks and stock market indices. Fundamental analysis considers political scenario, performance of the industry,state of the economy, intrinsic values of stocks etc. The focus of the technical analyst is to study the statistics generated by stock market activity. In technical analysis, historical values of price and volume are used to compute various statistical indicators which reflect health of the stocks and help deciding whether to invest in a particular stock.

Artificial Neural Networks (ANN) and SVR are two most predominantly used techniques for predicting future values of stocks and stock market indices.Zhang and Wu used a backpropagation neural network with an ImprovedBacterial Chemotaxis Optimization (IBCO) for predicting future values of stock market index [1].Their focus was not only the short-term prediction (next day) but they also demonstrated efficiency of their proposal for long term prediction (15 days). A forecasting modelbased on chaotic mapping, firefly algorithm and SVR was proposed in [2] to predict stock market price.

Predicting future values of S&P BSE Sensex and CNX Nifty was also attempted in [3, 4]. The focus of the work carried out in [3] was just to predict the movement of the direction of these indices, while in [4], future values of these indices were predicted. Predictions were made for 1 to 10 days, 15 days and 30 days in advance. They proposed two stage models for prediction tasks. Their main focus was predicting closing values of far days and their model did not allow them to improve next day’s prediction performance.

This paper focuses just on predicting closing value for the next day. This studyis inspired from the work carried out in [4] and experiments are also based on the work in [4]. Study examines the prediction performance of standard algorithms such as MLR [5] and linear SVR [6]. It also attempts to suggest some useful future directions which may be helpful in improving prediction performance.

Rest of the paper is organized as follows. Section 2 discusses about data used for experimentation and how it is prepared for experimentation. MLR and linear SVR which are used as the prediction models in this paper are described in brief in section 3. Section 4 explains experimental evaluation and paper is concluded in section 5. Section 5 also presents some useful directions for future work.

2. Research Data and Data Preparation

Experiments were carried out on two stock market indices from Indian stock markets. These indices were S&P BSE Sensex and CNX Nifty. Historical values of these indices from 3rd January 2011 to 31st December 2015 were used to form the dataset for experimentation. It is assumed here that data before and beyond this date range is not available.

DOI: 10.090592/IJCSC.2016.041

Ten technical indicators as indicated in Table 1 were used to describe the stock market indices on any given day.

Table 1: Technical Indicators

Simple n-day MovingAverage (SMA) (n = 10, here)	$\frac{C_t + C_{t-1} + C_{t-2} + \cdots + C_{t-9}}{n}$
n-day Exponential Moving Average (EMA) (n = 10, here)	$EMA(k)_t = EMA(k)_{t-1} + \alpha \times (C_t - EMA(k)_{t-1})$
Momentum (MOM)	$C_t - C_{t-9}$
Stochastic K% (STCK%)	$\frac{C_t - LL}{HH - LL} \times 100\ \%$
Stochastic D% (STCD%)	$\frac{\sum_{i=0}^{n-1} K_{t-i}}{10}\ \%$
Relative Strength Index (RSI)	$100 - \frac{100}{1 + \frac{\sum_{i=0}^{n-1} \frac{UP_{t-i}}{n}}{\sum_{i=0}^{n-1} \frac{DW_{t-i}}{n}}}$
Moving Average ConvergenceDivergence(MACD)	$MACD(n)_{t-1} + \frac{2}{n+1} \times (DIFF_t - MACD(n)_{t-1})$
Larry William's R% (LWR)	$\frac{HH - C_t}{HH - LL} \times -100$
Accumulation/Distribution(A/D) Oscillator (ADO)	$\frac{[(H_t - O_t) + (C_t - L_t)]}{[2 \times (H_t - L_t)]} \times 100$
Commodity Channel Index(CCI)	$\frac{M_t - SM_t}{0.015\ D_t}$

Where,
$C_t$is the closing price of tth day,
$O_t$is the opening price of tth day,
$L_t$is the low price of tth day,
$H_t$is the high price of tth day,
$DIFF_t = EMA(12)_t - EMA(26)_t$,
$\alpha = \frac{2}{k+1}$, where k designates how many days moving average,
LL = Lowest Low in n days,
HH = Highest High in n days,
$M_t = \frac{H_t + L_t + C_t}{3}$,
$SM_t = \frac{(\sum_{i=1}^{n} M_{t-i+1})}{n}$,
$D_t = \frac{(\sum_{i=1}^{n} |M_{t-i+1} - SM_t|)}{n}$,
UP_tmeans upward price change,
DW_tmeans downward price change

These technical indicators were calculated for both the indices for all the days in the datasets except for first 25 days. It can be seen from the formula of MACD that MACD requires historical values which are as many as 25 days old in addition to the current day's values. This means that to compute MACD for first 25 days, the data before the considered date range is required which is against the

DOI: 10.090592/IJCSC.2016.041

initial assumption and therefore technical indicatorswere not calculated for first 25 days of both the datasets.

It is further to notice that task of this paper is to predict tomorrow's closing value of these two indices. This means that next day's closing value is essential either for training or verifying the forecast of the prediction models. It is easy to understand from the initial assumption that for the last day of the datasets, next day's closing value is not available. As a result of this, technical indicators were not calculated for the last day and it was neither used for training nor testing.

Once these indicators were calculated, they were used to describe indices on any given day. This led to a situation where the health of an index on any given day was described by a vector of length 10 comprising of these technical indicators.

Both the datasets were divided into training and testing sets. First 80% days of the datasets were used to form the training sets while the remaining days were used to form the testing sets.

Z-score normalization was used to normalize the data.

3. Prediction Models

Multiple Linear Regression (MLR) and Support Vector Regression (SVR) were used as the prediction models.

3.1 Multiple Linear Regression

Multiple linear regression is an extension of simple linear regression. Simple linear regression involves only one independent variable and the idea is to find the best line which minimizes the squared error.

In this study, each day is described by means of 10 technical indicators resulting in 10 independent variables. Closing price continues to represent a dependent variable. This leads to the scenario where there are multiple independent variables and one dependent variable. If we assume that there exists a linear relationship between independent and dependent variables then multiple linear regression can be used and it can find the best hyperplane which minimizes the error. More details on MLR can be found in [5].

3.2 Linear Support Vector Regression

SVR is based on principles similar to those of SVM. SVM is typically used for classification while SVR is used for regression and therefore output variable is continuous in nature. This makes prediction tasks even more difficult as there are infinite number of possibilities for the output variable. SVR uses margin of tolerance ϵ and tolerates the error up to this tolerance. The main principle of minimizing error and maximizing margin while tolerating part of the error remains same.

In this paper, linear support vector regression wasused and inputs were the 10-dimensional vectors in the training set. Target variable was a column vector comprising of closing value of next dayfor each of the days in the training set.More details of SVR can be found in [6].

It is known that SVR involves a constant "c" which controls a trade-off between an approximation error and the weight vector norm $\|w\|$. To decide the best value of "c", 5-fold cross validation of training set was carried out. Four different values of "c" 0.1, 1, 10 & 100 were checked and the value for which minimum cross validation error was achieved was selected as the final value of "c". SVR was then trained using the entire training set and the best value of "c" that was found.

4. Experimental Evaluation

As mentioned earlier, both the datasets were divided into training and test sets. Training set was used to train prediction models and test set was used to measure prediction performance. Prediction performance was measured in terms of Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|A_i - F_i|$$

DOI: 10.090592/IJCSC.2016.041

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \times 100$$

Table 2 shows results of MLR and linear SVR for both the stock market indices. It can be seen that both the techniques perform almost identically but MLR is marginally better than linear SVR for both the stock market indices.

Table 2: Experimental Results

Stock Market Index	Technique			
	Multiple Linear Regression (MLR)		Linear Support Vector Regression (Linear SVR)	
	MAE	MAPE (%)	MAE	MAPE (%)
CNX Nifty	69.05	0.83	70.29	0.85
S&P BSE Sensex	217.83	0.80	219.18	0.80

Figure 1, 2, 3 & 4 depict actual and predicted values of these stock market indices for the testing days.

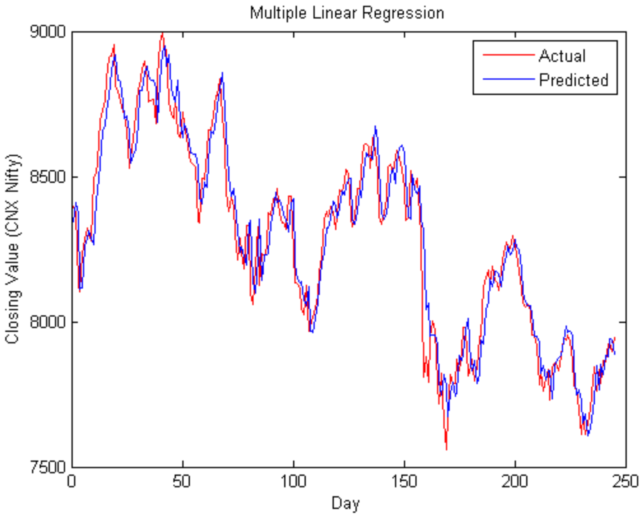

Figure 1: Performance of MLR (CNX Nifty)

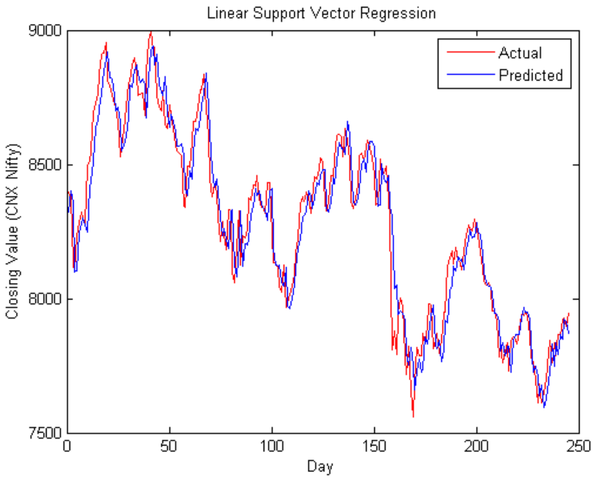

Figure 2: Performance of Linear SVR (CNX Nifty)

DOI: 10.090592/IJCSC.2016.041

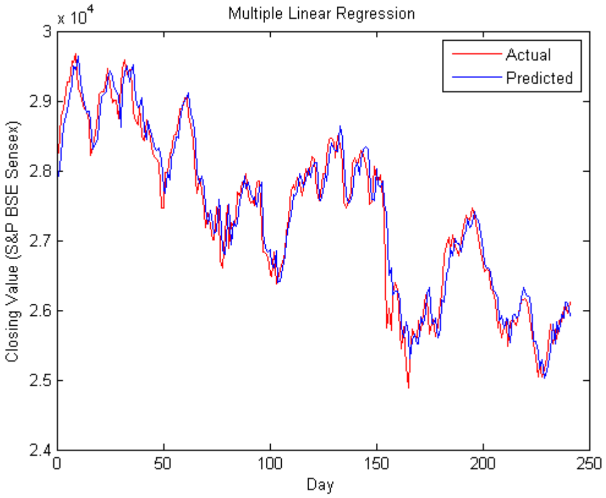

Figure 3: Performance of MLR (S&P BSE Sensex)

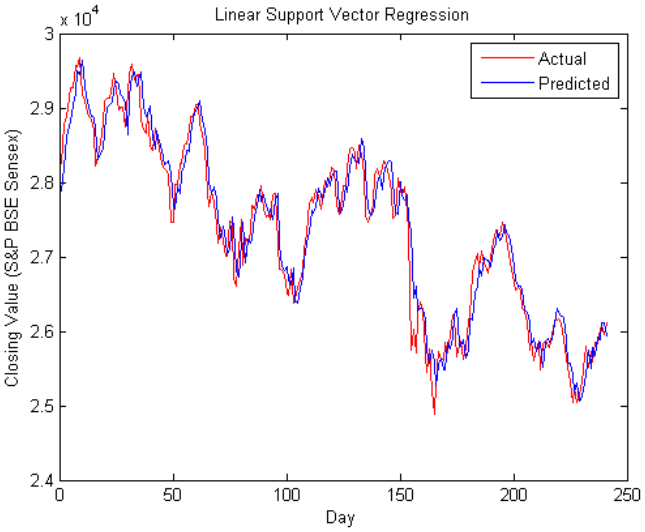

Figure 4: Performance of Linear SVR (S&P BSE Sensex)

It can be visualized that for both the stock market indices, error is approximately 0.8% and therefore there is a definite scope of improvement for the given task. Certain ways which may be helpful are discussed in the conclusion and future work section.

5. Conclusion and Future Work

The task of predicting next day's closing value for two stock market indices is addressed in this paper through Multiple Linear Regression (MLR) and Linear Support Vector Regression (SVR) techniques. Performance of both the techniques is almost identical with MLR having a slight edge over SVR. However, it will be interesting to see if this is statistically significant. One can use appropriate hypothesis testing method to verify the significance of the evidence.

It is important to notice that error value irrespective of stock market indices and techniques is approximately about 0.8%. This is less than 1% error but still not small enough and there exists a definite scope to address the problem and minimize the error for the given task.

DOI: 10.090592/IJCSC.2016.041

One can attempt to fuse knowledge which can be derived from fundamental and technical analysis. A systematic fusion of the knowledge derived through technical and fundamental analysis can work as the enhanced and improved input information to prediction models.

Another important direction for future work can be incorporation of signals derived from social media platforms.There are various social networking sites on which a lot of information is available which one can focus on.

It is further to bring to notice that different economic and political scenarios have different levels of impact. Certain scenarios have global impact while others may affect locally. They may also affect with different magnitudes to different countries and markets. These should also be considered to improve the performance of the prediction models and presents a healthy direction for the future work.

References

1. Zhang, Yudong, and Lenan Wu. "Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network." Expert systems with applications 36.5 (2009): 8849-8854.
2.Kazem, Ahmad, et al. "Support vector regression with chaos-based firefly algorithm for stock market price forecasting." Applied soft computing 13.2 (2013): 947-958.
3. Patel, Jigar, et al. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques." Expert Systems with Applications 42.1 (2015): 259-268.
4. Patel, Jigar, et al. "Predicting stock market index using fusion of machine learning techniques." Expert Systems with Applications 42.4 (2015): 2162-2172.
5.Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
6. Vapnik, Vladimir N. "An overview of statistical learning theory." IEEE transactions on neural networks 10.5 (1999): 988-999.