International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 8 (2019) pp. 1931-1934

Stock Market Prediction using Linear Regression and Support Vector Machines

Vaishnavi Gururaj[#1], Shriya V R[#2] and Dr. Ashwini K[#3]

[#123] *CSE Department, Global Academy of Technology, Bengaluru, India.*

Abstract
Machine learning (ML) is a technology that gives the systems the ability to learn on its own through real-world interactions and generalizing from examples without being explicitly programmed as in the case of rule-based programming.
Machine Learning can play a key role in a wide range of critical applications. In machine learning, Linear Regression (LR) is a basic technique by which a linear trend can be obtained. But Support Vector Machines (SVMs) have advanced features such as high accuracy and predictability.

In this paper we survey the pros and cons of using both these techniques to predict values and compare both algorithms.

Keywords: Prediction, Datasets, Linear Regression, Support Vector Machines, Machine Learning.

INTRODUCTION

One of the most important tasks in ML is to predict, with high accuracy and speed, the trend and the results for any given dataset. Before the era of Artificial Intelligence (AI) and ML, predictions were done manually by a statistician who would plot graphs and use mathematical methods and models to observe trends.

One of these methods was to fit a straight line of the form $y = mx + c$ to a graph such that the line passes through the maximum number of data points of the given dataset. Mathematically speaking, on plotting the values of the dataset on a graph, fit a straight line through the points such that the square of the distance between each point and the line is minimum. This line, called the *hypothesis* is used to predict the *y* value for any given *x*. This prediction technique is called Linear Regression and the formula used is called the Least Squares method. This technique is widely known to statisticians and has also been used as one of the basic concepts of ML.

The hypothesis function of Linear Regression has the general form,

$$y = h_\theta(x) = \theta_0 + \theta_1 x \qquad (1)$$

Note that this is like the equation of a straight line. The values of θ_0 and θ_1 is given to $h_\theta(x)$ to get the estimated output *y*. Note that the effort is to get various values of θ_0 and θ_1 to try to find values which provide the best possible "fit" or the most representative "straight line" through the data points mapped on the x-y plane. The accuracy of our hypothesis is measured using a cost function, which takes an average of all the results of the hypothesis with inputs from x's compared to the actual output y's.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \qquad (2)$$

This function is otherwise called the "Squared error function" or "Mean squared error". The mean is halved (½m) as a convenience for the computation of the gradient descent, as the derivative term of the square function will cancel out the ½ term.

The other advanced technique is the Support Vector Machine which was invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963 [1]. The SVMs were initially developed as Classification algorithms. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. Figure 1. shows the visual detail of an SVM. The points to the left and right of the support vectors are two classes that have been classified as being separate, by the SVM.

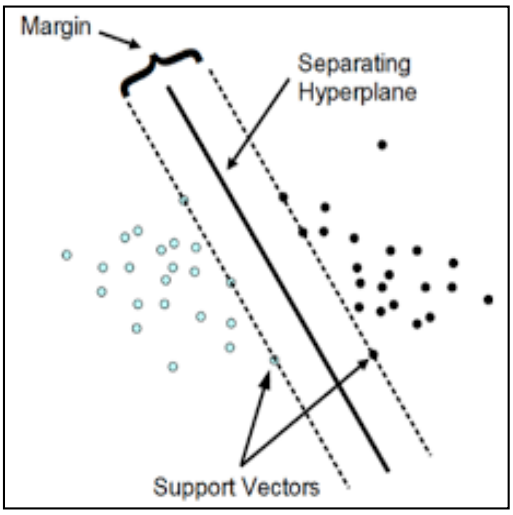

Fig. 1. Two classes separated by the largest gap

A version of SVM for regression was proposed in 1996 by Vladimir N. Vapnik, Harris Drucker, Christopher J. C. Burges,

Linda Kaufman and Alexander J. Smola [1]. This method is called the Support Vector Regression. The model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction. Another SVM version known as Least Squares Support Vector Machine (LS-SVM) has been proposed by Suykens and Vandewalle [5].

Now there are two ML techniques that can be used to perform predictions. Using the two techniques which have their own advantages and disadvantages in the current literature. On applying Linear Regression to a sample data that is easily available and perform observations followed by performing more observations using the Support Vector Regression. The observations and results are plotted on a graph and the two techniques are compared.

METHODOLOGY

A. Environment

This survey has been performed using a statistical language such as R, with the RStudio development environment. It is therefore easier to view the observations later by plotting them using functions predefined in R. Since Linear Regression and SVMs are standard algorithms used in almost all Data Science fields, the built-in functions in R packages are being used for this purpose.

B. Time-Series Forecasting

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus, it is a sequence of discrete-time data. Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. In statistics, prediction is a part of statistical inference. When information is transferred across time, often to specific points in time, the process is known as forecasting [2].

C. Sliding-Window Method

In time series prediction, the time series are typically expanded into three or higher-dimensional space to exploit the information that is implicit in them. Given a sequence of numbers for a time series dataset, the data can be restructured to look like a supervised learning problem. This can be done by using previous time steps as input variables and using the next time step as the output variable [2].

Consider the following time series:

Time	Price
1	100
2	110
3	108
4	115
5	120

Re-organizing the time series dataset this way, the data would look as follows:

X	Y
?	100
100	110
110	108
108	115
115	120
120	?

It is seen that the previous time step is the input (X) and the next time step is the output (y) in this supervised learning problem. It can be observed that there is no previous value that can be used to predict the first value in the sequence. This row is deleted as it cannot be used. Additionally, there is no known next value for the prediction of the last value in the sequence. This value is deleted while training the supervised model [2].

The use of prior time steps to predict the next time step is called the sliding window method. For short, it may be called the window method in some literature. In statistics and time series analysis, this is called a lag or lag method. This sliding window is the basis for how we can turn any time series dataset into a supervised learning problem. It can be seen how this can work to turn a time series into either a regression or a classification supervised learning problem for real-valued or labelled time series values.

SYSTEM SETUP

A. Data Collection

Historical daily prices were taken from Quandl website. Quandl provides an R API package that we can make use of to fetch the stock price of companies for any time range in just a single line of code. This eliminates the need to mine the data manually through other means. A typical internet connection is usually required, but Quandl allows the download of offline CSV file containing the stock data. In this survey, we make use of exactly 1 year of stock data of The Coca-Cola Company, from January 2017 to 2018.

B. Analysis Method

In order to evaluate the two techniques of ML, it is sufficient to show that the predicted model fits the data as accurately as possible. Actual prediction is not performed, but rather proof how well Linear Regression and SVMs fit the data using the training data itself as the test data set is given. Showing how well the model fits can therefore demonstrate that the method can be extended to predict actual future value. In this survey a Simple Linear Regression in One Variable is considered, namely the Closing Stock price or the End of Day price for the prediction.

C. Performance Evaluation Methods

The performance measures that were used to assess the predictive accuracy of the proposed system included the root mean square error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), the mean square error (MSE), the correlation coefficient (R), the non-linear regression multiple correlation coefficient. These indexes are used to

measure whether the predicted values are close to the actual values.

D. Functions Used

The Linear Regression Analysis was performed by using R-language's **lm()** command. The SVM was made use of by the **svm()** command in liquidSVM package, which is available in CRAN. The Quandl APIs make use of the **quandl()** command. To evaluate the performance and results, we use other inbuilt functions to calculate MSE, MAE, and R2.

EXPERIMENTAL RESULTS

A. Linear Regression

The first step that was performed was to fetch the values or to download them as a CSV file. In this literature, the stock prices of the Coca-Cola company within the date range 1-Jan-2017 to 1-Jan-2018 were obtained. The obtained data-frame had two columns namely, Date and Close which were initially plotted onto the graph using the *plot()* functions.

A linear model was later fit to this graph and displayed and observations were made. The resultant graph and the fitted model are as shown in Figure 2. The values of MSE, MSE, MAE, MAPE and R were obtained and are shown below.

Method	Result
RMSE	3.22
MAE	2.53
MSE	10.37
R-Squared	0.73

B. Support Vector Machine

The exact same steps were performed for Prediction using SVM with the only change being the calling of the *svm()* function in the e1071 package.

A support vector machine as stated in this literature plots points on a hyperplane such that data points belonging to two different classes are separated by Support Vectors by the largest gap possible. But this is defined for Classification problems which can be extended for Regression [1].

Method	Result
RMSE	1.58
MAE	1.33
MSE	2.51
R-Squared	0.93

It is therefore, conclusive that SVM performs better than LR. The resultant plots are shown below in figures Fig. 2 and Fig. 3. The figures Fig. 2 and Fig. 3 have been placed at the end of the paper.

ACKNOWLEDGMENT

We are immensely grateful to Dr. Ashwini K, Associate Professor, Dept. of Computer Science and Engineering, Global Academy of Technology, Bengaluru for her constant guidance and unending support.

REFERENCES

[1] "Support-Vector Networks" Corinna Cortes, Vladimir Vapnik

[2] "Time Series Forecasting as Supervised Learning" *https://machinelearningmastery.com/time-series-forecasting-supervised-learning/*

[3] "Forward Forecast of Stock Price Using Sliding-window Metaheuristic-optimized Machine Learning Regression" *Jui-Sheng Chou and Thi-Kha Nguyen*

[4] "Stock Market Predication Using A Linear Regression" Dinesh Bhuriya et al.

[5] "Least Squares Support Vector Machine Classifiers" J.A.K. Suykens And J. Vandewalle

[6] J.-S. Chou, and N.-T. Ngo, "Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns," Applied Energy, vol. 177, pp. 751-770, 2016.

[7] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, Least squares support vector machines: World Scientific, Singapore, 2002.

[8] Shunrong Shen, Haomiao Jiang, and T. Zhang, "Stock Market Forecasting Using Machine Learning Algorithms," IEEE Transactions on Neural Networks, vol. 84, no. 4, pp. 21, 2015

[9] J. A. K. Suykens, "Nonlinear modelling and support vector machines." pp. 287-294.

[10] C.-J. Lu, T.-S. Lee, and C.-C. Chiu, "Financial time series forecasting using independent component analysis and support vector regression," Decision Support Systems, vol. 47, no. 2, pp. 115-125, 2009.

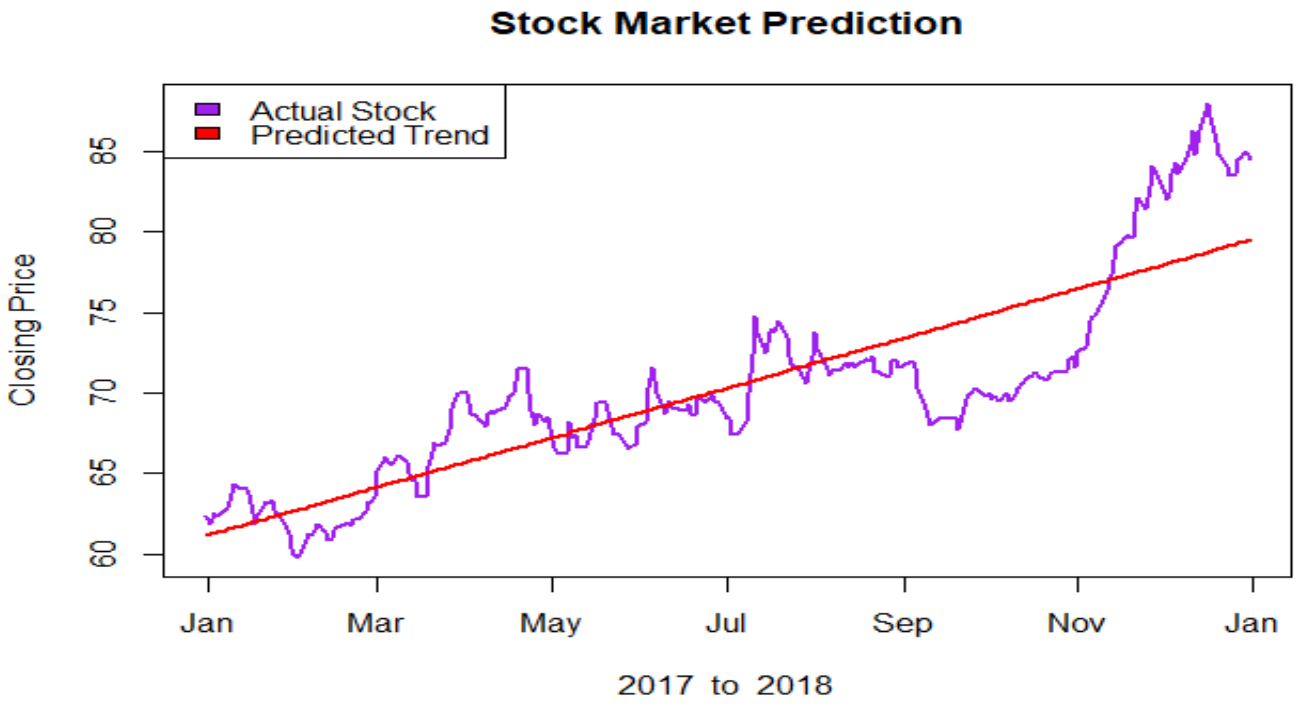

Fig. 2. Stock Prediction of Coca-Cola using Linear Regression

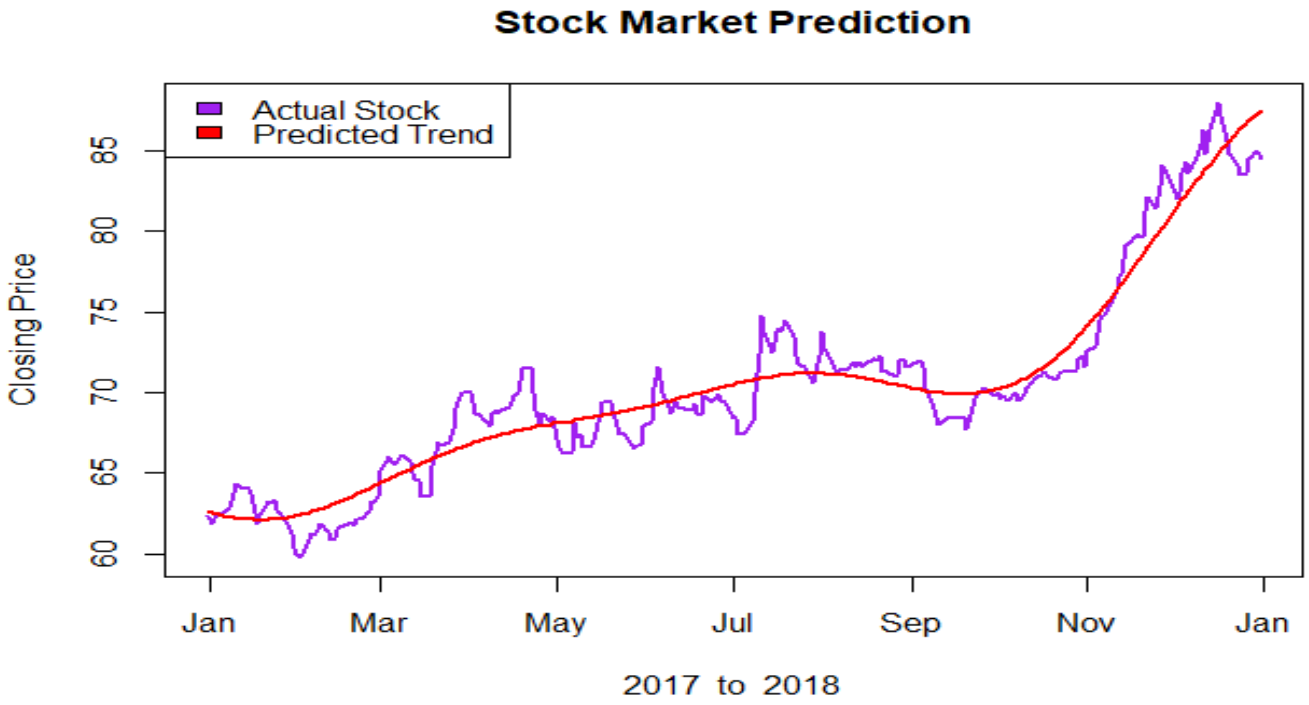

Fig. 3. Stock Prediction of Coca-Cola using SVM