Performance Analysis of Regression Based Machine Learning Techniques for Prediction of Stock Market Movement

Nitin Nandkumar Sakhare, S. Sagar Imambi

Abstract— ***Prediction of stock market movement is extremely difficult due to its high mutable nature. The rapid ups and downs occur in stock market because of impact from foreign commodities like emotional behavior of investors, political, psychological and economical factors. Continuous unsettlement in the stock market is major reason why investors sell out at the wrong time and often fail to gain the benefit. While investing in stock market investors must not forget the risk of reward rule and expose their holdings to greater risks. Although it is not possible predict stock market movement with full accuracy, losses from selling stocks at wrong time and its impacts can be reduce to greater extent using prediction of stock market movement based on analysis of historical data. Investors always need accurate predictions and they should use stock information wisely. A great quantity of chronological data is available in the context of stock market behavior. For stock market movement prediction, a number of machine learning algorithms are available. Use of particular machine learning algorithm has huge impact on prediction results obtained. In this paper, we have compared three different machine learning algorithms, namely, Linear Regression, Polynomial Regression and Support Vector Regression. We have applied stated techniques on data consisted of index and stock prices of S&P 500.***

Keywords: prediction; stock market; machine learning; regression; index; stock prices

I. INTRODUCTION

Now a day's worldwide stock market is facing a lot of changes rapidly. It is very difficult to predict Changes in stock prices as it is highly influenced by political news, economic events in the global markets [1]. With the introduction of Long Term Capital Gain Taxes for equity based markets in 2018-Budget of India, there was a big fall in stock market and investors loose around 5000 billion dollars in a single day on 8th February 2018. Investors also started to withdraw their profits to larger extent causing stock market to fall further. Hike in raw material prices, worldwide crude oil prices etc., downturn in IT industry are another major reasons why global markets, especially the Indian economy is highly exposed to global markets. In 1991, Prime Minister, P.V. Narsimha Rao and Finance Minister, Manmohan Singh bought a concept of Foreign Direct Investment (FDI) concept in India so that overseas business corporations invest directly in rapidly emergent concealed Indian stock markets and they can take benefits proportional to their investments. Since then FDI is steadily increasing in India. As a result of this, in first quarter of 2015, India has clearly defeated major global economies of United States and China by extending stock market investments of $31 billion. In the same quarter US had shown investments of $27 billion and China had shown investments of $28. Last few years, a lot of overseas funds have been invested into Indian stock market. As a result of this India is experiencing fast economic growth. Most of these multinational funds are game changing players in changing gears of Indian economy and their activities are making movement Indian stock market in non-linear way. Investment decisions are driven by events in stock market. Indian stock market is highly influenced by movement is global stock markets especially American Stock markets. .

A large number of small cap, mid cap and large cap companies from different nations all across the globe primarily depend on exporting goods, raw materials, commodity products to American or US based markets. Therefore US economy is considered to be the most powerful economy in the world [2]. If there is any adverse news regarding US economy, straight away economies of other countries start collapsing. As a result of this stock analysts of different nations track down the news related to US economy like USA employment numbers, sub-prime crisis of USA, FED interest rate movement very closely. Post FDI India's influence on global market has been increased drastically. Although Indian Market has a strong cohesion with domestic market, economy is growing at faster rate. A lot of Indian companies listed in NIFTY are investing in foreign stock markets like S&P 500, NYSE, London Stock exchange and NASDAQ etc and contributing in exporting domestic products to foreign markets causing revenue percentage to be increased in global stock market by considerable margin. This is the primary reason for movement of stock price indices in Indian stock market like National Stock Exchange (NSE), Bombay Stock Exchange (BSE), Nifty (50).

Revised Version Manuscript Received on March 08, 2019.

Nitin Nandkumar Sakhare, Research Scholar, Computer Science and Engineering Department, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India

Dr. S. Sagar Imambi, Associate Professor, Computer Science and Engineering Department, Koneru Lakshmaiah Education Foundation, Guntur-, Andhra Pradesh, India

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

All these volatile factors need to be carefully analyzed by stock market analyst and investors need to follow recommendations given by these analysts. Continuous unsettlement in the stock market is major reason why investors sell out at the wrong time and often fail to gain the benefit. While investing in stock market investors must not forget the risk of reward rule and expose their holdings to greater risks [3]. Although it is not possible predict stock market movement with full accuracy, losses from selling stocks at wrong time and its impacts can be reduce to greater extent using prediction of stock market movement based on analysis of historical data. Investors always need accurate predictions and they should use stock information wisely. A huge quantity of past data is available in the context of stock market behavior. Machine learning plays extremely important role in analyzing and prediction of stock market movements based historical data available. A number of machine learning algorithms can be used for prediction of stock market movement. However there is no as such best algorithm which can predict stock market movement with high precision. Therefore this creates need for performance analysis of different machine learning algorithms to reach to best machine learning algorithm resulting most optimal and precise prediction of stock movement thereby minimizing financial loses of investors investing in stock market[18][19].

For investigation reason we have utilized dataset of American securities exchange record dependent available capitalizations of 500 extensive organizations known as standard and poor's (S&P) 500. S&P 500 have 500 noteworthy stocks which are recorded on the NYSE or NASDAQ [14].

II. LITERATURE REVIEW

Prediction of future price of stocks, shares, equity traded funds, mutual funds from past price movement over the period of time is one of the most important challenges for finance based organizations and profession traders. It is very difficult to develop an optimal strategy for stock market investment due its highly mutable nature.

Machine learning makes it possible for systems/computers to automatically learn and act like humans do without being externally programmed by consuming data and information from real world observations. In machine learning emphasis is given on developing computer based programs that can access data, learn themselves and improve the prediction experience over the period of time. Machine learning is a practice of using algorithms to read and interpret the data and interpreted data to be used for prediction of future events. It is based on algorithms that can learn from data without depending on rule based coding [8].

Prediction is a very important task that data analyst perform using machine learning algorithms. There are large numbers of machine learning algorithms that can be used for prediction of future movements. Use of particular machine learning algorithm creates a huge impact on prediction results obtained. It is very important to compare different machine learning algorithms to be used for predictions of stock market movement [9].

Roberto Rosas-Romero, Alejandro Diaz-Torres, Gibran Etcheverry (2016) completed a huge work in determining of Stock Return Prices with Sparse Representation of Financial Time Series. They performed time arrangement investigation of stock value returns utilizing relapse based calculations [2].

Yaqing Xia, Yulong Lin, Zhiqian Chen, (2013) utilized Support Vector Regression for Prediction of Stock Trend [6].

The decision of machine learning issue to tackle specific issue dependably relies upon the size, quality, and nature of the information. It relies upon what you need to do with the appropriate response. Regression based systems are considered as best in class strategies for the expectation of time-arrangement information. A large portion of the scientists have utilized regression based learning procedures like linear regression, support vector relapse (SVR), polynomial regression for the forecast of stock qualities. It depends on number of numerical counts that specific calculation performs. It relies upon how much time modeler has. Machine learning is essentially a sub area of Artificial Intelligence (AI) and is arranged into four classifications dependent on kind of learning model: supervised/administered, unsupervised and semi-supervised and reinforcement learning.

A. Supervised Learning

Supervised learning is based on function approximation, where an algorithm is trained using a dataset called as training dataset (usually 60-80% data from original dataset)and in the end of the process we select the function that best describes the relation between input and output variables. Y=f(X). It is very difficult to find out a function that always predicts precise relation between input variable and output variable.

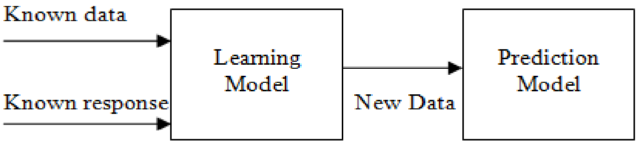


Fig.1. Supervised Learning

Supervised learning calculations try to build up a model which for the most part thinks of connection between a yield variable and at least one information factors. Supervised learning recognizes conditions between the objective variable known as expectation yield and the information factors to such an extent that one can undoubtedly foresee the yield esteems for new information regularly known as testing information dependent on the connections which it has gained from the preparation informational indexes (input factors). Yield variable or target variable is dependably an incentive to be anticipated and is known as needy variable. This yield variable shows either straightforwardly extent connection or in a roundabout way corresponding connection with info variable. Information factors are free factors as their qualities do not rely upon any substance included. As info factors are utilized to foresee yield variable, they are additionally called as indicators. Utilizing these arrangements of factors, modeler produces a capacity that outline to wanted yields. The tutoring procedure proceeds until the model demonstrates a steady execution with amplifying dimension of precision that modeler expects on the preparation. Regression, Decision Tree, Random Forest, K-nearest neighbors, Naïve Bayes, Support Vector Machine, Neural Network, and so forth are some outstanding calculations from directed learning class [2][5][6][9][16].

B. Unsupervised Learning

In unsupervised learning as the name suggests do not have any target or output variable to predict / estimate. There may not be any kind of dependency or relationship between input and output. y = f(x) functionality will not exists with unsupervised learning technique. In unsupervised learning focus is given on forming the well organized structure of data like clusters or dividing input data into groups based on some kind of similarity measure or finding association between two objects. In unsupervised learning, input data are unlabeled. It depends on scanning for various posts at complex information with the goal that it seems less difficult or increasingly composed dependent on similitude, design coordinating or affiliation measures. These calculations are especially valuable in situations where the information examiner doesn't realize what to search for in the information [23][24].

Unsupervised learning problems mainly classified into following types:

a. Association: This technique is particularly useful in to discovering the different patterns associated with the co-occurrence of items in a gathering. A well known application of association rule mining is used in market-basket analysis. Example: If a customer purchases shoes, he is 80% likely to also purchase socks [23].

b. Clustering: Clustering is very important unsupervised learning technique which forms the groups and allocates the input data samples into these groups based on similarity measures. Objects within the same clusters are more similar to each other whereas objects within different clusters exhibit dissimilar characteristics. Euclidean distance is most commonly used similarity measure in clustering. k-means, k-nearest neighbors are popular algorithms [22].

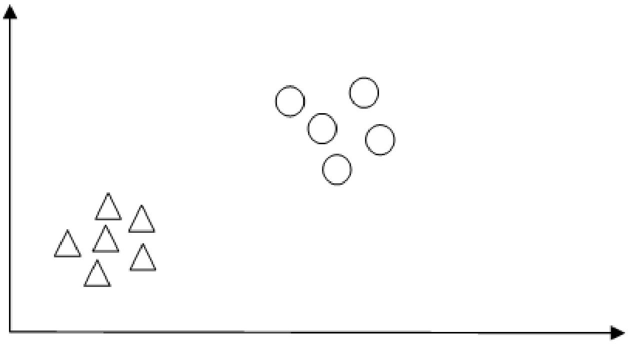


Fig.2. Unsupervised Learning

c. Dimensionality Reduction: Dimensionality Reduction is an unsupervised learning technique which focuses on reducing the number of irrelevant features of a dataset and at the same time makes sure that important information within dataset is preserved. Feature extraction methods and feature selection methods are two ways to perform dimensionality reduction. Feature Selection forms subset of features from the original set of features by selecting relevant features. Feature Extraction performs data transformation from a high-dimensional space to a low-dimensional space. Principal Component Analysis (PCA), Backward Feature Elimination, Forward Feature Construction are some of the important dimensionality reduction techniques [23].

C. Semi-supervised learning

In Supervised learning there are labels associated with all input data samples present in dataset based on which classification/prediction can be performed. In unsupervised there are no labels for any input data samples in the dataset. Semi-supervised learning is a midway approach. There exists a major class of machine learning problems where we need to form groups of data samples for which no labels are associated with them and we also need to assign labels for some of the data samples so as to perform classification or prediction. Semi-supervised learning techniques are best suited for the model building which has mix kind of input data sample as in some data samples have no labels associated with them whereas some has labels tagged with them. Semi-supervised learning techniques are most appropriate when the gathering participations of the unlabeled information are obscure and this information conveys imperative data about the gathering parameters. Semi-supervised learning regular utilized in Deep conviction systems, where a few layers are learning the structure of the information (unsupervised) and other layer is utilized to play out the grouping (prepared with directed information) [23].

D. Reinforcement Learning

Reinforcement learning is utilized to prepare the machine with the goal that it can settle on explicit choices. The machine is given an affair to a genuine domain where it trains itself ceaselessly utilizing experimentation and endeavors to settle on ideal choices/activities. This machine gains from past understanding and attempts to catch the most ideal learning to settle on exact business choices.

Reinforcement learning is an agent based machine learning technique in which based on present state and learning behavior agent makes a decision on the best next action so that choice of next action will maximize the reward [21].

Application areas of reinforcement learning include robotics, Internet of Things (IoT). In robotics a robot is exposed to real world environment and obstacles. Robot is expected to avoid collisions by receiving negative feedback after detecting the obstacles. Based on readings from infrared sensor robot must choose the next optimal action.

In reinforcement learning, an agent is supposed to respond or take decision at each data point it faces. After some time agent also receives reward points based on goodness of the decision. Based on this, the algorithm modifies its strategy in order to achieve the highest reward [21].

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

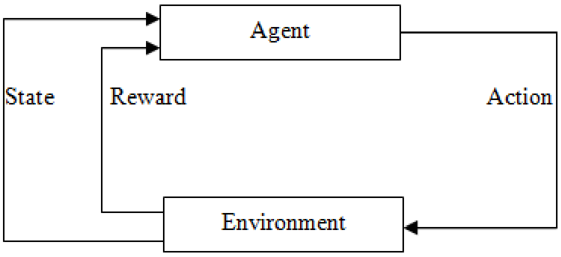


Fig.3. Reinforcement Learning

Reinforcement learning takes following approach so that agent can take sharp and precise decisions

1. Input state is observed by the agent.
2. Based on the input state observed and Decision making function with which agent is trained, it takes some action.
3. Based on the goodness of the decision taken by agent, the agent receives reward called reinforcement from the environment.
4. The input state-action pair information about the reward is stored and again fed back to agent.
5. Agent learns in iterative fashion so as to get maximum reward

As stock market movement is dynamic, nonlinear and constantly updating it is not possible to predict future movement of stock market using unsupervised, semi supervised machine learning algorithms. S&P 500 stock record has stock information of 500 vast organizations having regular stock recorded on the New York Stock Exchange (NYSE) or National Association of Securities Dealers Automated Quotation System (NASDAQ) [14]. Reinforcement Learning may give accurate predictions of stock market movement but training the model based on trial and error will take too much time. Also it is very difficult to calculate reward function for 500 stocks. As investors need real time price for investing in stocks, it is not an optimal solution to use reinforcement learning technique. Therefore this paper addresses analysis of supervised learning techniques for prediction of stock market movement. This paper aims to analyze the performance of regression based machine learning techniques like Linear Regression, Polynomial Regression and Support Vector regression for predicting stock market movement.

III. METHODOLOGY

Nature of stock data is an important factor to be considered during prediction of stock market movement [15][16]. As stock market data is mutable and numeric it is very important to use different machine algorithms for analysis purpose. This paper attempts to analyze stock market data using the regression based machine learning algorithms.

A. Linear Regression

Linear regression is a supervised machine learning technique which aims to build a learning model and tries to establish relationship between two variables by best fitting a linear line to input data. This line follows the equation of line:

$$y = mx + c \qquad (1)$$

Here y variable is considered to be a dependent or output variable explanatory variable, x is considered to be a independent variable and c represents intercept. y is a variable to be predicted and also known as criterion variable. Prediction of y is based on values of x hence x is called as predictor variable. Predictor variable could be one or more.

When there is only one predictor variable, the prediction method is called simple regression and when there are more than one predictor variable, the prediction method is called multi regression.

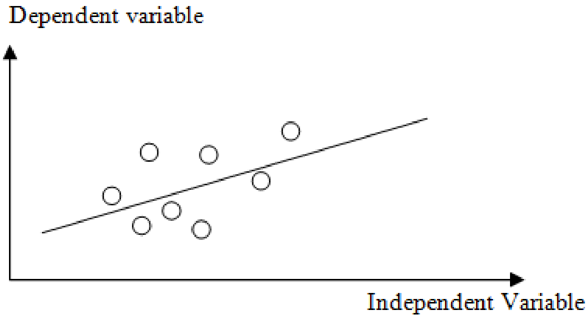


Fig.4. Linear Regression Model

As prediction is core logic of regression, we are making use of regression technique to solve the problem of stock market movement prediction.

Before using linear regression model for prediction one must determine if there exists a relationship between variable to be predicted/observed and input data [11]. There must significant amount of association between input and output variable. Strength of relation between these two variables can be identified using scatter plot. Correlation coefficient is an important numerical measure which measures how strongly two variables are associated or correlated with each other. Correlation coefficient takes value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

The dataset "Standards and Poor’s (S&P) 500" contains, among other variables, 500 stocks and movement of their value after every 60 seconds. In this paper we are attempting to predict value of stocks for further next intervals of 60 seconds [20].

Linear regression is a technique which consists of searching for the best-fitting straight line through the input data points. The best-fitting line is called a regression line. Once the regression line is modeled for group of data, data values which lie away from line are called as outliers. Outliers represent erroneous data and may indicate a poorly fitting regression line. With such erroneous data values accuracy of prediction can reduce to great extent. Outliers can have huge effects on the linear regression. Linear regression always searches for the linear relationship in the form of straight line between output variable and input variable/variables. Linear regression is always based on the assumption that there exists a linear relationship between output variable and input variable, which is not always true. Linear regression does not provide best fit for nonlinear data. This makes sense to consider Polynomial regression for prediction of stock movement.

B. Polynomial Regression

Sometimes, a graph of the independent versus a dependent variable may suggest there is a nonlinear relationship. Such kind of relationship can be understood with the help of a polynomial regression model. Polynomial regression model for a single variable can be given as:

$$y = a_0 + a_1 x + a_2 x^2 + \cdots\cdots + a_m x^m, m < n \quad (2)$$

where m is called the degree of the polynomial. Although polynomial regression represents nonlinear behavior, still it is considered as linear regression with regression coefficients , $a_o, a_1, \ldots\ldots\ldots\ldots\ldots a_m$.

Great thing about polynomial regression is that there could be more than one independent variable and these variables may need to have interaction with them in order to predict dependent variable Y.

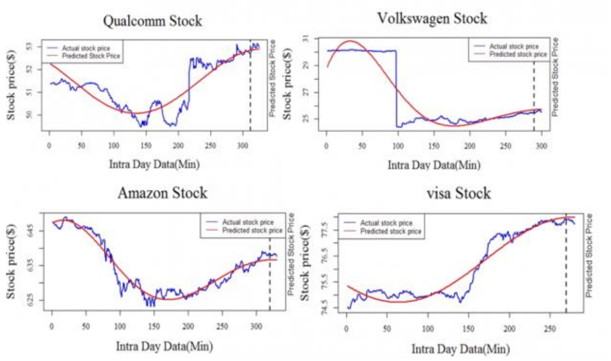


Fig.5. Polynomial Regression for prediction of Stock Price Movement

With experminetal work performed here in the context of polynomial regression, modeler has to keep mind general principles otherwise polynomial regression may produce meaningless results beyond the scope of the model. These general principles are given here.

- The fitted curve is more reliable only if model is trained with data of size much larger.
- Do not extrapolate beyond confines of observed values otherwise due to nonlinear nature, polynomial regression may produce meaningless results beyond the scope of the model.
- Values of independent variables must not be too large causing overflow for high degree polynomials. Therefore input variables X need to scale down.

C. Support Vector Regression (SVR)

A lot of study has been done over applying support vector machine for stock market analysis. However support vector machine solves binary classification problem whereas SVR acknowledges the presence of non-linearity in the data and provides a proficient prediction model. SVR uses the same basic idea as Support Vector Machine (SVM), a classification algorithm, but applies it to predict real values rather than a class.

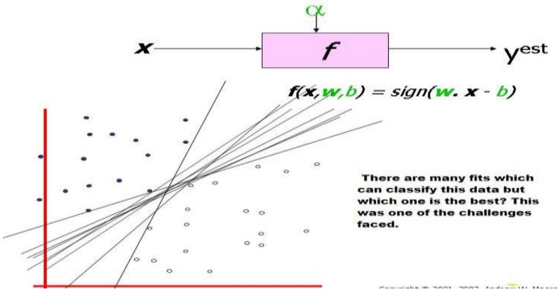

α
x
f
y^est
f(x,w,b) = sign(w. x - b)
There are many fits which can classify this data but which one is the best? This was one of the challenges faced.

Fig. 6. Number of hyperplanes possible for classification

Support vector machine is specialized supervised machine learning technique which not only forms a hyper-plane separating two classes but it identifies a just right hyper-plane which segregates two classes better. SVM identifies only one hyper-plane that maximizes the margin.

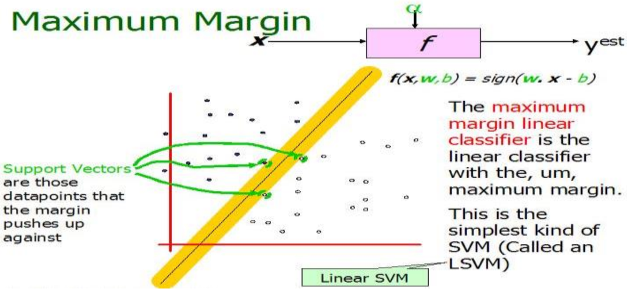

Maximum Margin
α
x
f
y^est
f(x,w,b) = sign(w. x - b)
The maximum margin linear classifier is the linear classifier with the, um, maximum margin.
This is the simplest kind of SVM (Called an LSVM)
Support Vectors are those datapoints that the margin pushes up against
Linear SVM

Fig. 7. Support Vector Machine

In SVM, kernel functions are set of mathematical functions that take data as input and transform this input from one form to the required form. Kernel function basically performs mapping of low dimensional input space to a higher dimensional. There are different kernel functions like linear, nonlinear, polynomial, Radial Basis Function (RBF) which can be used with different SVM algorithms [5]. Kernel function transforms extremely complex data and finds out the process to classify the data based on the predefined labels or classes [12]. Following are some commonly used kernel functions.

1. *Polynomial:* A polynomial kernel is a popular function for non-linear modeling. Polynomial kernels are typically used in image processing applications.

$$k(x, x^{'}) = (x, x^{'})^d \quad (3)$$

$$k(x, x^{'}) = ((x, x^{'}) + 1)^d \quad (4)$$

Here d is known as degree of polynomial function.

2. *Gaussian Radial Basis Function (GRBF)*: GRBF are general purpose kernel functions and most commonly used with a Gaussian form. GRBP are particularly used when there no prior knowledge about the data.

$$k(x, x^{'}) = \exp\left(-\frac{\|(x-x^{'})^2\|}{2\sigma^2}\right) \quad (5)$$

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

3. *Exponential Radial Basis Function*: When there are discontinuities in the input data space or when data points are discrete in nature, a radial basis function produces a separate linear solution and makes these discontinuities to be acceptable.

$$k(x,x^{'}) = \exp\left(- \frac{\|(x-x^{'})\|}{2\sigma^{2}}\right) \qquad (6)$$

4. *Multi-Layer Perceptron (MLP)*: MLP is based on Neural Networks with a single hidden layer which can also be used to represent kernel function.

$$k(x,x^{'}) = \tanh(p(x,x^{'})+\sigma) \qquad (7)$$

Support Vector Machine is one of the most popular supervised learning techniques which can be applied not only to classification problems but also to the problems with time series analysis and regression. In case of regression, the dependent variable is numerical rather than categorical. Unlike linear or polynomial regression models output model from SVR does not depend on distributions of the underlying dependent and independent variables. SVR also allows construction of non-linear output model without changing any independent variable. Most important feature of SVR is principle of maximum margin algorithm. With SVR each data item is plotted as a point in n-dimensional space. Size of the dimensional space is number of features present within dataset, with the value of each feature being the value of a particular coordinate. In SVR a non-linear kernel function is leaned by SVR model. Using this nonlinear kernel function mapping of low dimensional feature space into high dimensional feature space is performed. [6]. This transformation is done using the parameters which are completely independent of the dimensionality of feature space. Another advantage with SVR is that it does not care about the prediction results as long as the value of standard deviation is less than the predefined threshold value. Considering the fact that output in stock value prediction is a real number it is difficult to predict as we could have infinite possibilities regarding stock value movement.

IV. EXPERIMENTAL RESULTS

For experimental analysis purpose we have used dataset of S&P 500. S&P 500 stock index has stock data of 500 large companies like Apple, Bank of America, Accenture, AT&T etc. on the New York Stock Exchange (NYSE) or National Association of Securities Dealers Automated Quotation System (NASDAQ) [14]. This dataset has stock value of 500 stocks and value of stock updated every 60 seconds. Considering mutable and non-linear movement of stock values, 80% of data is used as training data and 20% is used for testing. A tool used for analysis of this dataset is R programming. R is the most comprehensive tool available for extensive data analysis which includes all standard statistical tests, models and analysis components.

Parameters used for measuring the prediction accuracy are standard deviation and Mean Square Error (MSE) [7]. As S&P 500 has 500 stocks listed on its index, we have measured prediction accuracy for all 500 stocks and prediction accuracy of each machine learning algorithms considered in this paper for entire dataset is calculated as average of standard deviation for all 500 stocks. S&P 500 stock record has stock information of 500 vast organizations having regular stock recorded on the New York Stock Exchange (NYSE) or National Association of Securities Dealers Automated Quotation System (NASDAQ) [14]. Movement is indicated after periodic interval of 60 seconds. Dataset used for experimentation has 41267 instances. 33013 instances are used as training instances whereas 8254 instances are used as testing instances.

Experimental steps taken for analysis are as follows:

1. S&P 500 dataset is imported in CSV format.

2. Dataset is divides as 80% data for training and 20% for testing.

3. Feature scale values of stocks within column

4. Create the trainer model

5. Trainer model learns the pattern within the elements for the training set

6. Values in test dataset are predicted.

7. Standard deviation is calculated as:

a. Find the absolute difference between predicted values and actual values.

$$\varepsilon \;\; =| \, y_{predicted} \text{ - } y_{actual} \;\; | \qquad (8)$$

b. Find the addition of the differences.

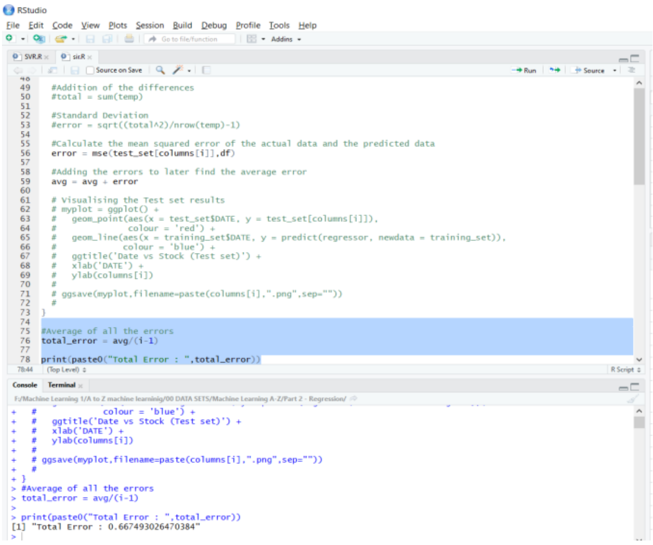
$$\sum_{i=1}^{500} \varepsilon_i \qquad (9)$$

c. Calculate Standard Deviation

$$\sigma = \sqrt{\frac{(total\_error)^2}{(n-1)}} \qquad (10)$$

d. Find the average standard deviation

$$\text{Average Standard Deviation} = \frac{\sigma}{500} \quad (11)$$



```
  #Addition of the differences
  #total = sum(temp)

  #Standard Deviation
  #error = sqrt((total^2)/nrow(temp)-1)

  #Calculate the mean squared error of the actual data and the predicted data
  error = mse(test_set[columns[i]],df)

  #Adding the errors to later find the average error
  avg = avg + error

  # Visualising the Test set results
  # myplot = ggplot() +
  #   geom_point(aes(x = test_set$DATE, y = test_set[columns[i]]),
  #              colour = 'red') +
  #   geom_line(aes(x = training_set$DATE, y = predict(regressor, newdata = training_set)),
  #             colour = 'blue') +
  #   ggtitle('Date vs Stock (Test set)') +
  #   xlab('DATE') +
  #   ylab(columns[i])
  #
  # ggsave(myplot,filename=paste(columns[i],".png",sep=""))
  #
}

#Average of all the errors
total_error = avg/(i-1)

print(paste0("Total Error : ",total_error))
```

F:/Machine Learning 1/A to Z machine learninig/00 DATA SETS/Machine Learning A-Z/Part 2 - Regression/

```
+   #             colour = 'blue') +
+   #   ggtitle('Date vs Stock (Test set)') +
+   #   xlab('DATE') +
+   #   ylab(columns[i])
+   #
+   # ggsave(myplot,filename=paste(columns[i],".png",sep=""))
+   #
+ }
> #Average of all the errors
> total_error = avg/(i-1)
>
> print(paste0("Total Error : ",total_error))
[1] "Total Error : 0.667493026470384"
> 
```

Fig. 8. Analysis with Linear Regression using R programming tool

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

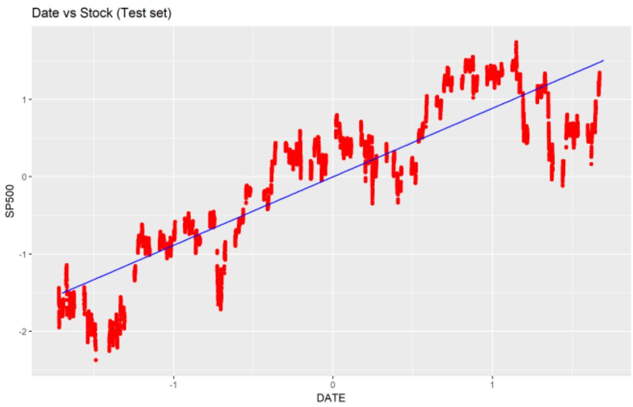
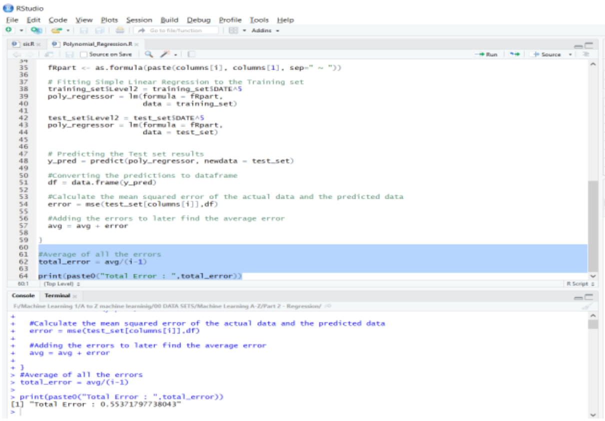


Fig 9. Linear Regression


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
sicR
Polynomial_Regression.R
Source on Save
Run
Source
60:1
(Top Level)
R Script
Console
Terminal
F:/Machine Learning 1/A to Z machine learninig/00 DATA SETS/Machine Learning A-Z/Part 2 - Regression/

```
  fRpart <- as.formula(paste(columns[i], columns[1], sep=" ~ "))

  # Fitting Simple Linear Regression to the Training set
  training_set$Level2 = training_set$DATE^5
  poly_regressor = lm(formula = fRpart,
                      data = training_set)

  test_set$Level2 = test_set$DATE^5
  poly_regressor = lm(formula = fRpart,
                      data = test_set)


  # Predicting the Test set results
  y_pred = predict(poly_regressor, newdata = test_set)

  #Converting the predictions to dataframe
  df = data.frame(y_pred)

  #Calculate the mean squared error of the actual data and the predicted data
  error = mse(test_set[columns[i]],df)

  #Adding the errors to later find the average error
  avg = avg + error

}

#Average of all the errors
total_error = avg/(i-1)

print(paste0("Total Error : ",total_error))
```

```
+
+   #Calculate the mean squared error of the actual data and the predicted data
+   error = mse(test_set[columns[i]],df)
+
+   #Adding the errors to later find the average error
+   avg = avg + error
+
+ }
> #Average of all the errors
> total_error = avg/(i-1)
>
> print(paste0("Total Error : ",total_error))
[1] "Total Error : 0.55371797738043"
>
```

Fig. 10. Analysis with Plynomial Regression using R programming tool

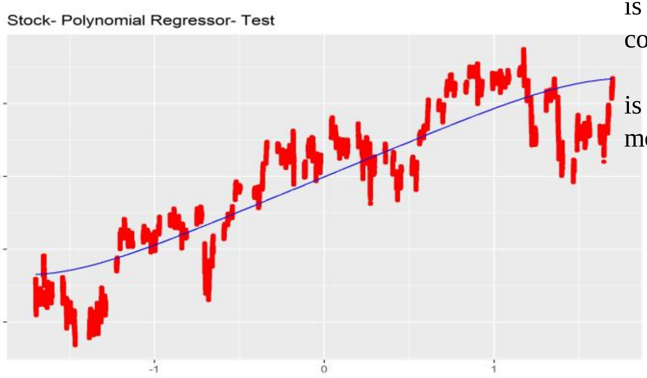
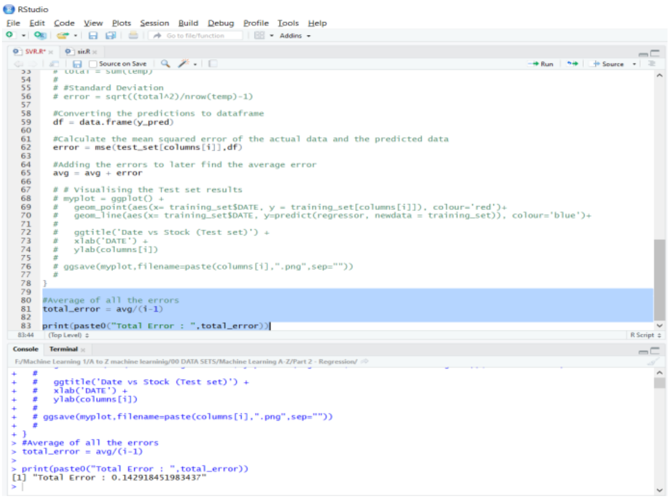


Fig. 11. Polynomial regression


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
SVR.R*
sir.R
Source on Save
Run
Source
83:44
(Top Level)
R Script
Console
Terminal
F:/Machine Learning 1/A to Z machine learninig/00 DATA SETS/Machine Learning A-Z/Part 2 - Regression/

```
  # total = sum(temp)
  #
  # #Standard Deviation
  # error = sqrt((total^2)/nrow(temp)-1)

  #Converting the predictions to dataframe
  df = data.frame(y_pred)

  #Calculate the mean squared error of the actual data and the predicted data
  error = mse(test_set[columns[i]],df)

  #Adding the errors to later find the average error
  avg = avg + error

  # # Visualising the Test set results
  # myplot = ggplot() +
  #   geom_point(aes(x= training_set$DATE, y = training_set[columns[i]]), colour='red')+
  #   geom_line(aes(x= training_set$DATE, y=predict(regressor, newdata = training_set)), colour='blue')+
  #
  #   ggtitle('Date vs Stock (Test set)') +
  #   xlab('DATE') +
  #   ylab(columns[i])
  #
  # ggsave(myplot,filename=paste(columns[i],".png",sep=""))
  #
}

#Average of all the errors
total_error = avg/(i-1)

print(paste0("Total Error : ",total_error))
```

```
+   #
+   #   ggtitle('Date vs Stock (Test set)') +
+   #   xlab('DATE') +
+   #   ylab(columns[i])
+   #
+   # ggsave(myplot,filename=paste(columns[i],".png",sep=""))
+   #
+ }
> #Average of all the errors
> total_error = avg/(i-1)
>
> print(paste0("Total Error : ",total_error))
[1] "Total Error : 0.142918451983437"
>
```

Fig. 12. Analysis with Support Vector Regression using R programming tool

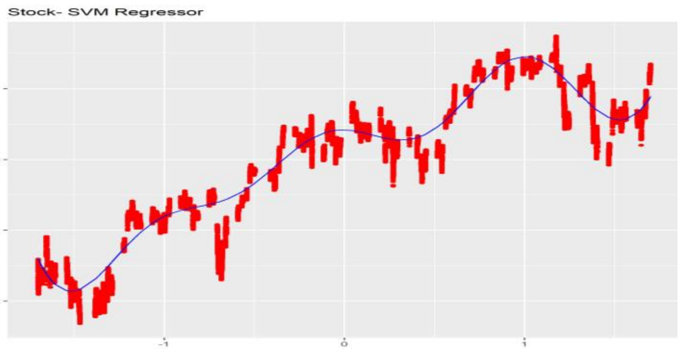


Fig. 13. Support Vector regression

TABLE I. STOCK MARKET MOVEMENT PREDICTION ACCURACY

Name of algorithm	Average Standard Deviation	Mean Squared Error (MSE)
Linear Regression	29.70	0.6675
Polynomial Regression	25.59	0.5537
Support vector regression	5.694	0.1429

V.PERFORMANCE COMPARISON

Performance of linear regression, polynomial regression and Support Vector Regression algorithm is measured using average standard deviation calculated in equation (11).

Prediction accuracy is inversely proportional to the average standard deviation. Hence less is the standard deviation more is the number of instances correctly predicted by the corresponding machine learning algorithm.

From the graph it is clear that Support Vector Regression is the best technique for prediction of stock market movement.

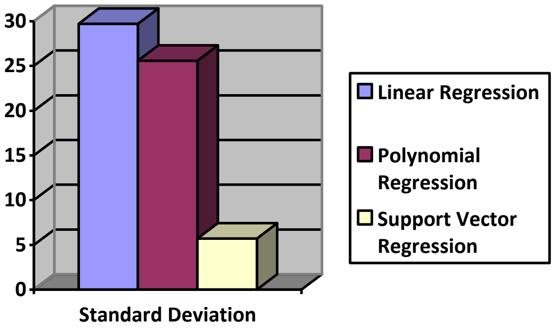


Fig.14. Performance Comparison of Linear Regression, Polynomial Regression, Support Vector Regression for prediction of S&P 500 Stock Market Movement using Standard Deviation

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

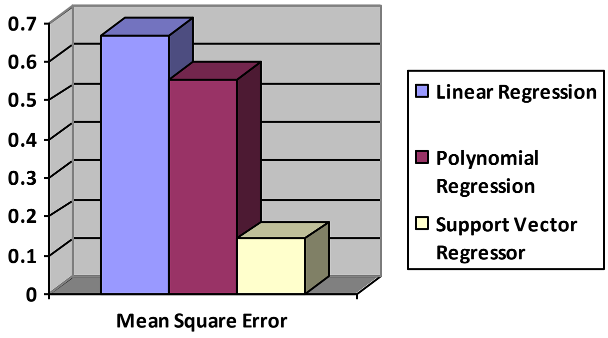


Fig.15. Performance Comparison of Linear Regression, Polynomial Regression, Support Vector Regression for prediction of S&P 500 Stock Market Movement using Mean Square Error

VI. CONCLUSION

Classification, clustering, pattern matching, association rule mining, regression, prediction, reinforcement learning are most popular machine learning techniques useful for data analysis. However choice of the technique will greatly influence the results obtained. It is observed that supervised learning techniques are much better and extensively used in stock market prediction. In supervised learning large number of algorithms like Regression, Decision Tree, Artificial Neural Networks, Random Forest, K- nearest neighbors, Naïve Bayes, Support Vector Machine are available. Considering the mutable and non-linear characteristics of stock market data, we have shown performance analysis of linear regression, Polynomial regression, Support Vector regression for prediction of stock market movement. From the analysis it is clear that Support vector regression is outstanding in comparison of linear regression, Polynomial regression,

Training the prediction model is most important step for any regression based supervised learning technique. In SVR training is relatively easy as compared to other regression techniques. Scalability is another important feature of SVR and SVR performs well irrespective of size of feature dimension space. SVR also handles complexity levels of different kernel functions and error can be unambiguously managed.

REFERENCES

1. Jing Zhang, Shicheng Cui, Yan Xu, Qianmu Li, Tao Li, “A novel data-driven stock price trend prediction system”, ScienceDirect, Expert Systems With Applications 97 ,2018, pp.60–69,
2. Roberto Rosas-Romero, Alejandro Diaz-Torres, Gibran Etcheverry, “Forecasting of Stock Return Prices with Sparse Representation of Financial Time Series Over Redudant Dictionaries”, ScienceDirect, Expert Systems With Applications, 2016, pp.1-12.
3. Rodolfo C. Cavalcante, Rodrigo C. Brasileiro, Victor L.F. Souza, Jarley P. Nobrega, Adriano L.I. Oliveira, “Computational Intelligence and Financial Markets: A Survey and Future Directions”, ScienceDirect, Expert Systems With Applications”, 2016, pp. 1-12.
4. Yauheniya Shynkevich, T.M.McGinnity, Sonya A. Coleman, Ammar Belatreche, Yuhua Li, “Forecasting price movements using technical indicators: Investing the impact of varying input window length”, ScienceDirect, Expert Systems with Applications, 2017, pp.1-18.
5. Lean Yu, Huanhuan Chen, Shouyang Wang, Kin Keung Lai, “Evolving Least Squares Support Vector Machines for Stock Market Trend Mining”, IEEE Transaction on Evolutionary Computation, vol. 13, issue 1, Feburary 2009, pp.87-102.
6. Yaqing Xia, Yulong Lin, Zhiqian Chen, “Support Vector Regression for Prediction of Stock Trend”, 6th IEEE International Conference on Information Management, Innovation Management and Industrial Engineering, 2013, pp.123-126.
7. Halit Alper Tayali, Seda Tolun, “Dimension reduction in mean-variance portfolio optimization”, ScienceDirect, Expert Systems with Applications, 2018, pp. 161-169.
8. Jigar Patel, Sahil Shah, Priyank Thakkar, K. Kotecha, “Predicting stock market index using fusion of machine learning techniques”, ScienceDirect, Expert Systems with Applications, 2014, pp.1-11.
9. Jigar Patel, Sahil Shah, Priyank Thakkar, K. Kotecha, “Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques”, ScienceDirect, Expert Systems with Applications, 2014.
10. Jothimani, Dhanya, Shankar, Ravi, Surendra S.,”Ensemble of non-classical decomposition models and machine learning models for stock index prediction”, 12th MWAIS Proceedings, 2017.
11. Abel Rubio, Jose D. Bermudez, Enriqueta Vercher, “ Forecasting portfolio returns using weighted fuzzy time series methods”, ScienceDirect, International Journal of Approximate Reasoning, 2016, pp.1-12.
12. Feng Wang, Yongquan Zhong, Qi Rao, Kongsham Li, Hao Zhang, “Exploring mutual information information based sentimental analysis with kernel based extreme learning machine for stock prediction”, Soft Computing, Springer-Verlang, Berlin, Heidelberg 2016.
13. PELLAKURI, V., Rao, D. R., Prasanna, P. L., & Santhi, M. V. B. T. (2015). “A Conceptual Framework For Approaching Predictive Modeling Using Multivariate Regression Analysis Vs Artificial Neural Network.” Journal of Theoretical & Applied Information Technology, 2015,Vol.77,issue 2.
14. Prodromos E. Tsinaslandis, “Subsequence dynamic time wrapping for charting: Bullish and bearish class predictions for NYSE stocks”, ScienceDirect, Expert Systems with Applications 94(2018) pp.193-204.
15. D. Diaz, B.Theodoulidis, P.Sampaio, “Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices”, ScienceDirect, Expert Systems with Applications 2011, vol.28, issue 10, pp.12757-12771.
16. Mingyue Qiu, Yu Song, Fumio Akagi, “Application of artificial neural network fot the prediction of stock market returns: The case of Japanese Stock Market”, ScienceDirect, Chaos, Solitons and Fractals 85, 2016, pp.1-7.
17. Rajashree Dash, Pradipta Dash, “A hybrid stock trading framework integrating technical analysis with machine learning techniques”, ScienceDirect, The journal of finance and data science 2, 2016 pp.42-57.
18. Ashish Sharma, Dinesh Bhuriya, Upendra Singh, “Survey of Stock Market Prediction Using Machine Learning Approach”, IEEE ICECAT 2017 proceedings, pp.506-510.
19. Zahid Iqbal, R.Ilyas, W. Shahzad, Z.Mahmood, J.Anjum, “Efficient Machine Learning Techniques for Stock Market Prediction”, IJERA, Vol.3, issue 6, Nov-2013, pp.855-867.
20. https://www.kaggle.com/camnugent/sandp500
21. Zhiyong Tan, Chai Quek, Philip Y.K. Cheng,”Stock Trading with Cycles: A financial application of ANFIS and reinforcement learning”, ScienceDirect, Expert Systems With Applications, Vol. 38, Issue 5, May 2011, pp.4741-4755.
22. Martin Langkvist, Lars Karlsson, Ami Loutfi, “A Review of Unsupervised Feature Learning and Deep Learning for Time Series Modeling”, ScienceDirect, Pattern Recognition Letters, Vol. 42, June 2014, pp. 11-24.
23. R. Sathya, Annamma Abraham, “Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification”, International Journal of Advanced Research in Artificial Intelligence, vol.2, issue 2 2013, pp.34-38