Volume 8, No. 3, March – April 2017

ISSN No. 0976-5697

ernational Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Comparative Study of Stock Prediction System using Regression Techniques

UtkarshSinha*, NihalVidyala,ShraeySaxena
Student, VIT University,Vellore, India

Prof. Sathyaraj R
Professor, VIT University,Vellore, India

***Abstract*:** There is a creating enthusiasm for the prediction and analysis of stock prices to augment the profits for the end client. This figuring will give an inexact estimation whether the stock price will increment or abatement sooner rather than later. To play out this prediction we will utilizeinformation mining algorithms like linear regression, support vector regression(SVR) and ridge regression. The method will likewise give a comparative study of these three algorithms on the premise of their precision in prediction of the stocks. The prediction system alongside the information of fundamental and technical analysis can prompt an extremely precise prediction, which will bring about exceptional yields. The system can be utilized by traders, typical clients, brokers and so on.

***Keywords*:** Stocks, regression, support vector machines, fundamental analysis, technical analysis, behavioral finance, API

I. INTRODUCTION

A stock market, equity market or share market is the aggregate of purchasers and dealers of stocks(also called offers), which address proprietorship ensures on organizations; these may consolidate securities recorded on an open stock exchange and what's more those solitary traded covertly. Trade in stock markets infers the exchange for money of a stock or security from a vender to a purchaser. This requires these two social affairs to yield to a price. Their buy or offer solicitations may be executed for their advantage by a stock exchange trader.

Stock market figure is the exhibit of endeavoring to choose the future estimation of an organization stock or other budgetary instrument exchanged on an exchange. Figure methods of insight fall into three general characterizations, which can (and routinely do) cover. They are key fundamental analysis, technical examination (charting) and technological methods using data mining.

i.Fundamental Analysts are concerned with the organization that underlies the stock itself. They survey an organization's past execution and also the legitimacy of its records.

ii.Technical investigators or chartists attempt to choose the future cost of a stock develop only in light of the (potential) examples of the past cost .

iii.The most unmistakable creative methodologies incorporate the usage of artificial neural networks(ANNs) and Genetic Algorithms close by different regression systems and SVM.Scholars discovered bacterial chemotaxis improvement technique may perform superior to GA.[1] Tobias Preis et al. acquainted a technique with distinguish online antecedents for stock market moves, utilising exchanging methodologies in light of pursuit volume data gave by Google Trends.[2] Their analysis of Google scan volume for 98 terms of differing financial importance, distributed in Scientific Reports,[3] proposes that increments in look volume for financially applicable inquiry terms have a tendency to go before substantial misfortunes in financial markets.[4][5][6][7]

II. LITERATURE REVIEW

Nicholas I. Sapankevych, Ravi Sankar give a review of time series prediction applications in their paper[8]utilising a novel machine learning approach: support vector machines (SVM). In their paper, they give a concise instructional exercise on SVMs for time series prediction. It gives a knowledge into the applications utilising SVM for time series prediction. It plots a portion of the points of interest and difficulties in utilising SVMs for time series prediction. SVM's are utilised to precisely figure time series data when the fundamental framework procedures are regularly nonlinear, non-stationary and not characterised from the earlier. It demonstrates to utilise RBF networks with versatile focuses and widths. It gives a strategy on the best way to predict utilising these two methods. It then gives some trial data and the outcomes alongside a conclusion.

Vatsal H. Shah in his Paper[9] breaks down different machine learning systems to predict the future stock qualities. Highlights two types of stock prediction philosophies: fundamental and technical analysis. Fundamental analysis manages the organisation whose stock is being exchanged and settles on choice in light of the organisation's past execution, and so forth. Technical analysis manages constant assessment of data series, and decides future stock cost in view of past stock cost. The conditions utilised as a part of this paper took after a similar general model: data pre-processing, learning algorithms implemented, training, testing and re-implementation.Uses high, low, open, close and volume values for each stock.Learning algorithms implemented are decision stump, linear regression, boosting and support

vector machines.Support Vector Machine consolidated with boosting gave best outcomes.

K. Ashwin Kumar, T. NiranjanBabu, NitishVaishy, K. Lavanya in their paper review[10]utilises four sorts of linear regression models and four sorts of neural system algorithms to remove linear and non-linear elements of the market, individually. SVM reliably shows bring down RMS mistake when contrasted with NEAT. Finishes up by saying nonlinear mix procedure proposed by this paper has much potential in time-series estimating.

III. METHODOLOGY

The algorithms used are linear regression, ridge regression and Support Vector Regression with both a linear kernel and a polynomial kernel of degree 2. The algorithms implemented in this project are available through the open-source Python library Scikit-Learn.] These algorithms form the equations of their respective regression models by calculating coefficients using the training data provided. After the models have been mathematically derived, data is fitted using each model.

A. Algorithms Used: Linear Regression is a factual system to decide the linear connection between at least two factors. This regression is essentially utilised for prediction and causal induction. Linear Regression fits a linear model with coefficients w=(w1…wp) to limit the remaining entirety of squares between the watched reactions in the dataset, and the reactions predicted by the linear estimation. Scientifically it takes care of an issue of the shape:

$$\min_{w} ||Xw - y||_2{}^2$$

Regression demonstrates to us how variety in one variable co-happens with variety in another. What regression can't show is causation; causation is just shown logically, through substantive hypothesis. For instance, a regression with shoe measure as an autonomous variable and foot estimate as a reliant variable would demonstrate a high regression coefficient and profoundly critical parameter gauges, yet we ought not presume that higher shoe estimate causes higher foot measure. All that the arithmetic can let us know is regardless of whether they are related, and assuming this is the case, by how much.

Support Vector Machines(SVMs) are an arrangement of supervised learning methods utilised for classification, regression and outliers detection. SVMs don't specifically give likelihood evaluates, these are figured utilising a costly five-crease cross validation.The support vector machines in scikit-learn support both thick and scanty specimen vectors as information. Be that as it may, to utilise a SVM to make predictions for sparse data, it probably been fit on such data.

The model delivered by support vector classification depends just on a subset of the preparation data, in light of the fact that the cost work for building the model does not think about preparing focuses that lie past the edge. Similarly, the model created by Support Vector Regression depends just on a subset of the preparation data, in light of the fact that the cost work for building the model disregards any preparation data near the model prediction.

There are three distinct executions of Support Vector Regression: SVR, nuSVR, LinearSVR.

Ridge Regression addresses a portion of the issues of Ordinary Least Squares by forcing a penalty on the measure of coefficients. The ridge coefficients limit a penalised remaining total of squares,

$$\min_{w} ||Xw - y||_2{}^2 + \alpha||w||_2{}^2$$

Here, $\alpha \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity.

B. Experiment Design and Dataset:

The data for training and testing was gathered through the Quandl API. The API call retrieves all the financial data for nine companies from the time that the companies stocks were publicly traded. This is stored in a Pandas data frame. The data frame contained thousands of rows and columns representing all the historical data of a particular company. To define features and labels for processing by the algorithms said above, preprocessing was done on the data. Initially, each date was furnished with 5 new segments D1 to D5. These segments each contained the Adjusted Close estimation of the organisation's stock from one of the past 5 days. For instance, D5 contains the balanced close of the earlier day's stock, while D1 contains the balanced close of stock 5 days preceding the present day. These 5 sections bargain the features used to gauge the present label, which is the present day's balanced close. The dataset is then trained so as to make the last dataset that goes about as the contribution for the calculation. This dataset contains the sections D1 to D5, Adjusted Close and an essential key for reference. The following stride includes fitting every regression model. To do this, the dataset is separated into training and testing data in a 80-20 proportion. The greater part of the models are fitted with a similar training data and are tried with a similar testing data, to empower us to think about the distinction in results got for a similar info data.

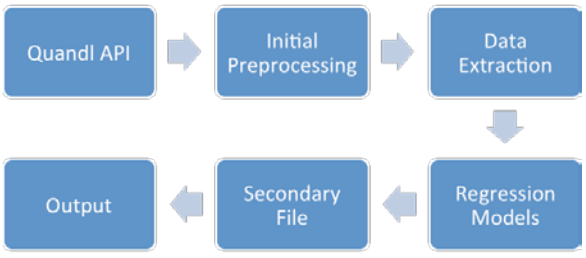

Fig 1.1 Process Flow Diagram

IV. RESULTS AND DISCUSSIONS

We tried the algorithms on the stock costs of nine distinct organisations. We think about the exactness scores utilizing R^2 and RMSE values as correlations between actual adjusted Close value and the predicted value from the regression model. RMSE is a broadly utilised metric to assess model prediction mistake. Moreover, since our goal is to precisely model the real close cost of a stock on a given day and not only the course in which the stock will move, it additionally benefits us to utilise a metric, for example, RMSE which is not interested toward the leftover. RMSE is used rather than MAE (Mean Average Error) in view of the nearness of the squaring component. This expands the heaviness of especially substantial residuals, making the blunder more observable; this element is alluring as bigger safety buffers are not satisfactory in stock prediction.

We likewise use a residuals versus fits plot for every regression model utilised as a part of the framework .While the translation of this plot is rather subjective, it provides information regarding the fitting of the regression model, the presence of outliers and variance of error terms. The best RMSE value is the lowest one whereas the best R^2 value is the highest one. After training of the data when it is input in the model of our system, the output are the different RMSE and R^2 values for each company concluding towards which regression is the best.

From the acquired qualities, it is exceptionally apparent that Support Vector Regression model is the most exact relapse show for anticipating stocks.The predicted close values of the stocks using these regression model is near the actual close values of the stocks which was the point of this framework. The proposed model predicts the close prices of the stock using these regression model very accurately along with indicating the direction of the stock in the near future.

Company	Linear Regression		Ridge Regression		Support Vector Regression	
Ticker	RMSE	R^2	RMSE	R^2	RMSE	R^2
FB	1.4932	0.1079	1.4932	0.108	1.4833	0.1065
YHOO	0.5573	0.1089	0.5571	0.109	0.5571	0.1091
XOM	0.8291	0.1008	0.8293	0.1036	0.8256	0.1040
GE	0.2957	0.0938	0.2936	0.0995	0.2928	0.1039
MSFT	0.6847	0.1091	0.6838	0.1082	0.6787	0.1097
BP	0.4771	0.1077	0.4773	0.1034	0.4708	0.1079
C	0.652	0.1098	0.6519	0.1099	0.6416	0.1101
PG	0.7058	0.1057	0.7056	0.1023	0.6998	0.1089
WMT	0.8146	0.0985	0.8144	0.0852	0.7925	0.1003

Table 1.1 The predicted close output values(rmse and $r^{2)}$ for all the companies using the regression model.

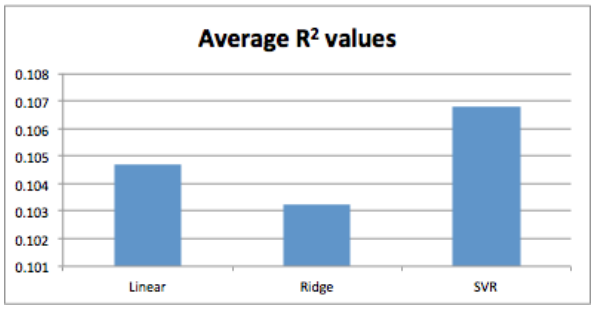

Fig 1.2 Average R^2 values for all the techniques.

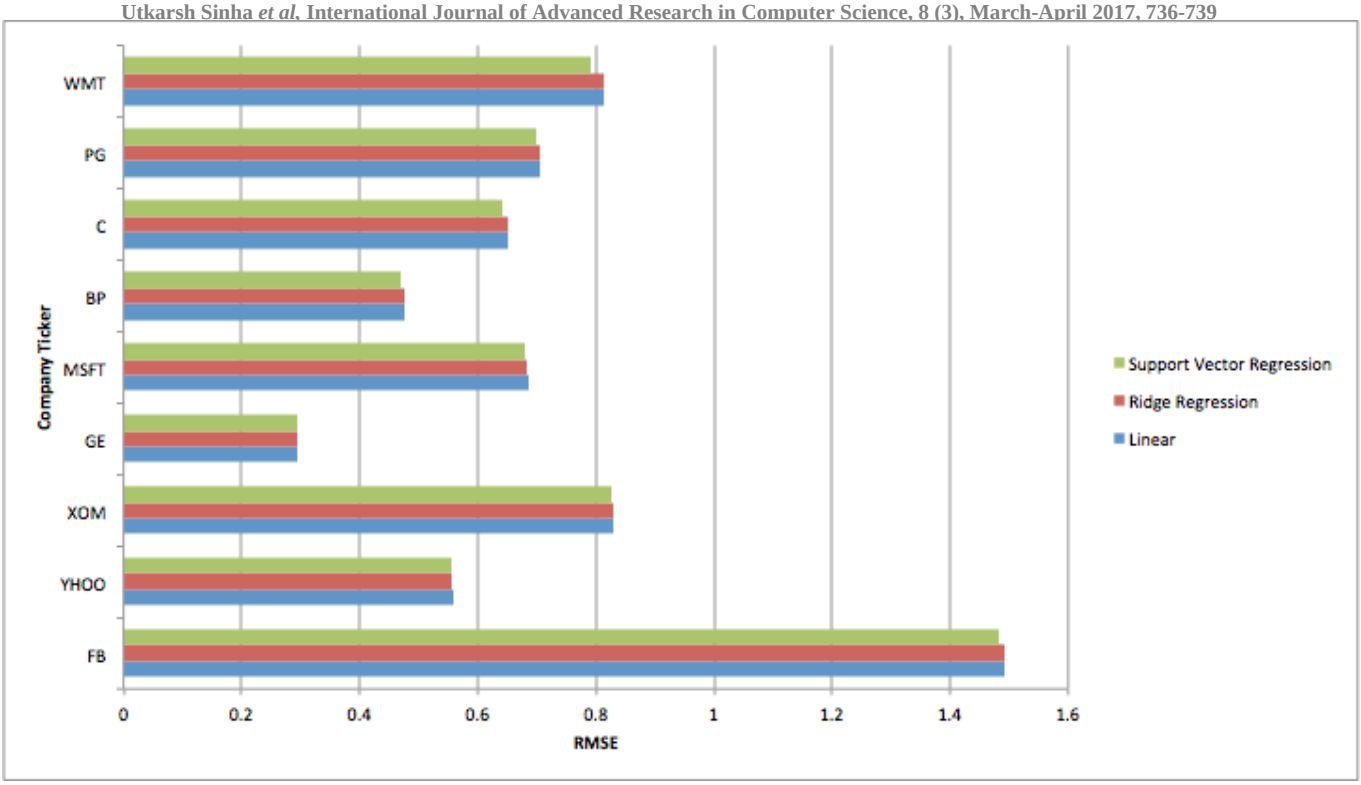

Fig1.3 The output RMSE Values for the all the companies.

V. CONCLUSION

This paper proposes a by and large nice structure for stock estimate which can satisfy three trademark requirements (with any dataset). Regardless, the most basic thought about the paper is to interface conventional (target) machine learning counts to (subjective) customer contributions to learn convoluted social structures. The framework could predict the close prices of the stock precisely utilising the regression methods. As it can be seen from the charts, Support Vector Regression is the best among the three.

VI. ACKNOWLEDGEMENT

We would like to thank Prof. Sathyaraj R., SCOPE, VIT University-Vellore, for his constant support and encouragement for this work.

VII. REFERENCES

[1] Zhang, Y.; Wu, L. (2009). "Stock Market Prediction of S&P 500 via combination of improved BCO Approach and BP Neural Network". Expert systems with applications.
[2] Philip Ball (April 26, 2013). "Counting Google searches predicts market movements". Nature.
[3] Tobias Preis, Helen Susannah Moatand H. Eugene Stanley (2013). "Quantifying Trading Behavior in Financial MarketsUsing Google Trends". Scientific Reports.
[4] Nick Bilton (April 26, 2013). "Google Search Terms Can Predict Stock Market, Study Finds". New York Times.
[5] Christopher Matthews (April 26, 2013). "Trouble With Your Investment Portfolio? Google It!".TIME Magazine.
[6] Philip Ball (April 26, 2013). "Counting Google searches predicts market movements". Nature.
[7] Bernhard Warner (April 25, 2013)."'Big Data' Researchers Turn to Google to Beat the Markets".BloombergBusinessweek.
[8] Sapankevych, N. I., &Sankar, R. (2009). Time series prediction using support vector machines: a survey. IEEE Computational Intelligence Magazine, 4(2).
[9] Shah, V. H. (2007). Machine learning techniques for stock prediction.Foundations of Machine Learning| Spring.
[10] Sendhilvel, L., Kumar, A., Babu, N., &Vaishy, N. V. N. (2016). Stock Market Prediction by Non-Linear Combination based on Support Vector Machine Regression Model. International Journal of Advanced Research in Computer Science, 7(7).