International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

A Machine Learning Approach For Stock Forecasting Using Regression Algorithm

Ganesha M[1], Pruthviraj S[2], Sharath R[3], Sushmitha N C[4], Shubha S[5]

Assistant Professor, Dept. of ISE, R.R Institute of Technology, Bengaluru, India[1]

UG Scholars, Dept. of ISE, R.R Institute of Technology, Bengaluru, India[2,3,4,5]

ABSTRACT: Stock market or Share market is one of the most complicated and sophisticated way to do business. Small ownerships, brokerage corporations, banking sector, all depend on this very body to make revenue and divide risks; a very complicated model. However, this paper proposes to use machine learning algorithm to predict the future stock price for exchange by using open source libraries and preexisting algorithms to help make this unpredictable format of business a little more predictable. We shall see how this simple implementation will bring acceptable results. The outcome is completely based on numbers and assumes a lot of axioms that may or may not follow in the real world so as the time of prediction.

KEYWORDS: Basics, Data Analysis, Fundamental,Implementation, Linear Regression, Stock Market, SupervisedMachine Learning.

I. INTRODUCTION

The Stock market is the collection of purchaser and supplier of stocks, which will represent the ownership claim on businesses. Market participant includes individual wholesale investors, organization investors such as banks, mutual funds, insurance companies and so on. Stock market participation refers to the number of agents who buy and sell equity backend securities either direct or indirect in a financial exchange. The stock market is one of the important parts for companies to raise their funds. Both wholesale and organization investors are involved in the stock market and want to know whether some stock will rise or fall over a certain period of time.

The stock market development role is played by the authority. Mainly by those particular laws that concern the transparency, the security and the equal remedy of the investors. It maintains a secret of the investors in the stock market and influences them to invest in it.

II. RELATED WORK

- **Feasibility study**

Simply putting stock market value can't be predict the accurate value. In near future, like any complex problems, has occurs and too many variables/features used to predict the value. The stock market is the place where sellers and buyers converge. When there are less sellers and more buyers, then the price increases. When there are more sellers and less buyers, then the price decreases. So, there is an aspect which causes people to sell and buy. It has less to do with logic but more with emotion. Because emotion can't be predictable, stock market price movements will also follow same thing.

- **Mat plot lib**

It will provide a selection for backend integration with the help of LaTeX. It will have multiple plots on the same axes. These multiple subplots can be obtained in a single figure. i.e., 3D plotting. This will work with labelled data sets which are similar to data Frames in pandas, but not only can it cycle but also line style and hatches.

III. PREDICTION MODEL

- ***Data Analysis Stage***

In this stage, we shall look at the raw data available to us and study it in-order to identify suitable attributes for the prediction of our selected label. Now the data that we're going to use for our program is taken from www.quandl.com, a premier dataset providing platform.

The dataset taken is for GOOGL by WIKI and can be extracted from quad using the token "WIKI/GOOGL".We have extracted and used approximately 14 years of data. The attributes of the dataset include:

Open (Opening price of Stock)

High (Highest price possible at an instance of time)

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

Low (Lowest price possible at an instance of time)
Close (Closing price of stock)
Volume (Total times traded during a day)
Split ratio
Adj. Open
Adj. High
Adj. Low Adjusted values of above attributes
Adj. Close
Adj. Volume

We select the attribute "*Close*" to be our label (The variable which we shall be predicting) and use "Adj. Open, Adj. High, Adj. Close, Adj. Low and Adj. Volume" to extract the features that will help us predict the outcomebetter.
It is to be noted that we use adjusted values over raw as these values are already processed and free from common data gathering errors.
Now we aware that the graphs made for stock analysis use the above attributes to plot them. Such graphs are called OHLCV graphs [11] and are very informative about the status of the stocks. Now we use the same graphing parameters to decide out features for the classifier. Let's define the set of features which we shall be using:
□□*Adj. Close*: This is an important source of information as this decides market opening price for the next day and volume expectancy for the day.
□□*HL_PCT*: This is a derived feature which is defined by:
We use percentage change as this helps us reduce the number of features but retain the net information involved. High-Low is a relevant feature because this helps us formulate the shape of the OHLCV graph.
□□*PCT_change*: This is also a derived feature, defined by: We do the same treatment with Open and Close as High and Low, since they both are very relevant in our prediction model and helps us reduce number of redundant features as well.
□□*Adj. Volume*: This is a very important decision parameter as the volume traded has the most direct impact on future stock price than any other feature. Therefore we shall use this as it is in our case.
We have successfully analyzed the data and extracted the useful information that we shall be needing for the classifier. This is a very crucial step and shall be treated with extreme care. A miss of information or small error in deriving useful information will lead to a fail prediction model and a very inefficient classifier.
Also, the features extracted are very specific to the subject used and will definitely vary from subject to subject. Generalization is possible if and only if, the data of the other subject is collected with the same coherence as the earlier

- **Proposed system**
  Our proposed system will allow investors to predict the next-day movement for a particular stock or index, which is important for daily traders. Three machine learning models, decision trees, neural networks and support vector machines, serve as the basis for our "inference engine". To build these models, a list of potential predictors from these data-sources has been generated. The list includes both variables (i.e. data in raw form) and features (e.g. summary statistics and predictors combining multiple variables) selected from these

IV. IMPLEMENTATION

This phase is initiated after the system has been tested and accepted by the user. In this stage, the system is installed to support the intended business functions. System performance is compared to performance objectives established during the planning phase. Implementation includes user notification, user training, installation of hardware, installation of software onto production computers, and integration of the system into daily work processes.

- **Least square Support Vector Machine**

Least squares support vector machines (LS-SVM) are least squares versions of support vector machines (SVM), which are a set of related supervised learning methods that analyze data and recognize patterns, and which are used for classification and regression analysis. In this version one finds the solution by solving a set of linear equations instead of a convex quadratic programming (QP) problem for classical SVMs. Least squares SVM classifiers, were proposed by Suykens and Vandewalle [28].

Let X is $n \times p$ input data matrix and y is $n \times 1$output vector. Given the n

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

i ii x y 1 { , } □□training data set, where *p xi*□*R*and *y R i* □□, the LS-SVM goal is to construct the function *f* (*x*) □□*y* , which represents the dependence of the output *i y* on the input *i x* . This function is formulated as:

$$f(x)- WT \square(x) \square\square\square$$

Where *W* and □(*x*) :*p n R* □*R*are *n*□1column vectors, and *b*□*R*. LS-SVM algorithm [5] computes the function (1) from a similar minimization problem found in the SVM method [3]. However the main difference is that LSSVM involves equality constraints instead of inequalities, and it is based on a least square cost function. Furthermore, the LS-SVM method solves a linear problem while conventional SVM solves a quadratic one.

.

V. TRAINING AND TESTING

Testing: is carried out for testing modules constructed form the system design. Each part is compiled using inputs for a specific part. Every modules are grouped into a large unit during unit testing.

- **Unit Testing**: would be done in each and every phase of project design and coding. The testing of each modules interface will be carried out to make the actual flow of information into and out of the programming unit while testing. The temporary generated output data is to maintain securely its integrity throughout the algorithm execution take place by checking the local data structure. At last, all error-handling functional paths are also been tested.
- **Integration testing**: We usually perform system testing to find out the error result from unanticipated interactions between the sub-system and system features. Software must be used to detect and find out all possible errors, once the source code is generated before delivering to customers/users. To finding the errors, series of test cases must be performed which ultimately uncover all the possible existing errors. Different software techniques is used for this process. This techniques will provides systematic guidelines for designing test of the software features that will be performed by internal logic and train input and output domains of a programming to discover errors in programming function, behaviour and performance
- **Validation and Verification**
  The testing process is one part of boarder subject which is referring to verification and validation. We have to confess that the system specifications and trying to meet the customer's/users requirements and for this under processing purpose, we have to verify and validate the dataset/details of product to make assured everything is in proper place. The two different things are verification and validation. The first one is performed to finding out that the software is correctly implemented as a specific functionality and other one is done to finding it the customer's requirements are properly covered or not by end of the task/project. Verification of the project was taken place to make assured that the project covered with all the requirements and specifications of our project. We made sure that, our project is up to the standard mark as we planned at the begin of our project development.

Case ID	Description	Input Data	Expected Output	Actual Output	Comments
1	Verify that user is able to register with valid credentials	User Details	Successful Registered	Successful Registered	As Expected
2	Verify that Enter/Tab key works as a substitute for the Register in button	User Details	If Enter, Register button should work and if Tab, cursor should go from one textbox to another	Both button work perfectly	As Expected
3	Verify that User is not able to Register with blank field	User Details	Should display a message fields can't be empty	Displays a message fields can't be empty	As Expected
4	Verify that User is able to Login with Valid Credentials	Login Details	Should lead you to Dashboard page	Leads the user to the analysis and prediction page	As Expected
5	Verify that User is unable to Login with Valid Credentials	Login Details	Should display a message user detail doesn't exists	User details doesn't exists	As Expected
6	Verify that User is not able to	password:	Should display a message	Display	As

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

	login with blank Username	XXXX	username can't be empty	username can't be empty	Expected
7	Verify that User is not able to Login with blank Password	Username: XXXX	Should display a message password can't be empty	Display password can't be empty	As Expected
8	Verify that Enter/Tab key works as substitute for the Login in button	Login Details	If Enter, Login button should work and if Tab, cursor should go from one textbox to another	Both button works perfectly	As Expected

VI. RESULT AND DISCUSSION

All datasets are available in [13]. All datasets are divided into training part (70%) and testing part (30%). Fig. 3 to Fig. 14 outline the application of Proposed LS-SVM-PSO model compared with LS-SVM and ANN-BP algorithms at different data set with different sectors of the market. In Fig. 3, Fig. 4, and Fig. 5, which present results of three companies in information technology sector (Adobe, Oracle and HP), results show that LS-SVM optimized with PSO is the best one with lowest error value followed by LS-SVM algorithm. Fig. 6 and Fig. 7 represent results of financial sector (American Express, and Bank of New york), we can remark that the predicted curve using the proposed LS-SVM-PSO algorithm is most close to the real curve which achieves best accuracy, followed by LS-SVM, while ANN-BP is the worst one. Fig. 8 represents results of using PSO-LS-SVM model in Honeywell company which represent industrials stock sector, proposed model still achieves best performance.Fig. 9 and Fig. 10 outline the application of the proposed algorithm to hospera and life technologies companies in health stock sector. From figures one can remark the enhancement in the error rate achieved by the proposed model. Fig. 11 and Fig. 12 outline the results of testing proposed model on Exxon-mobile and duke energy companies which represent energy stock sector. PSO-LS-SVM also the best especially in fluctuation cases. Fig. 13 represents results for FMC Corporation in materials stock sector. The achievements of proposed model is very promising compared with SVM and ANN Fig. 14 outlines results for AT&T from communication stock sector. We can notice from figure the role of proposed model in reducing the error rate and overcoming local minima problems which found in ANN results.

Table 1 outlines Mean Square Error (MSE) performance function for proposed algorithm. It can be remarked that the LS-SVM optimized with PSO always gives an advance over LS-SVM and ANN trained with LM algorithms in all performance functions and in all trends and sectors. Proposed model performs better than other algorithms especially in cases with fluctuations in the time series function.

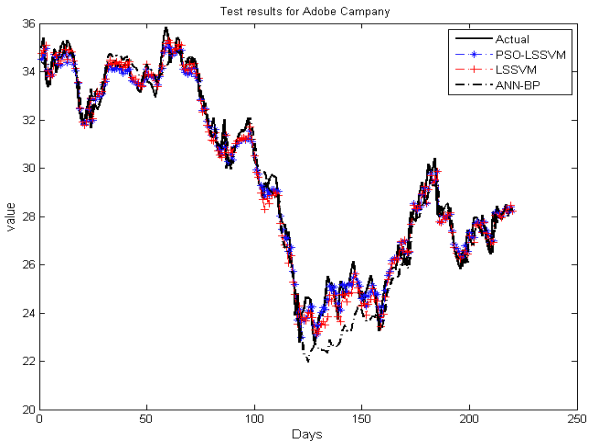

Fig. 3: Results for Adobe Company

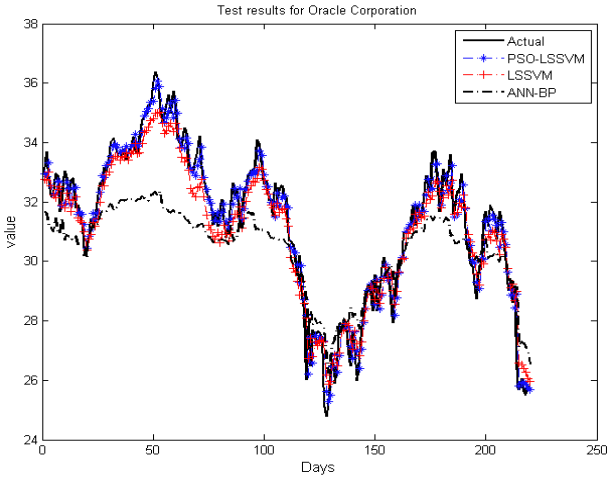

Fig. 4: Results for Oracle Company

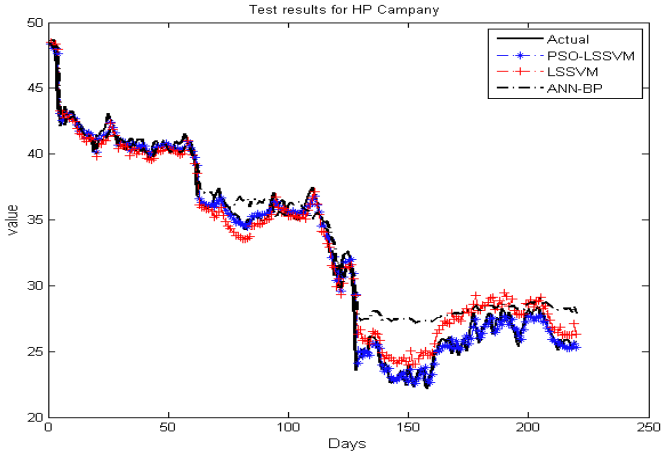

Fig. 5: Results for HP Company

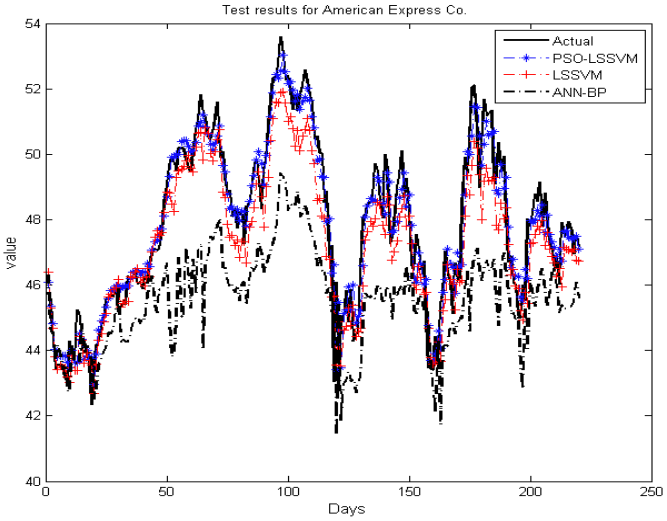

Fig. 6: Results for American Express Co

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

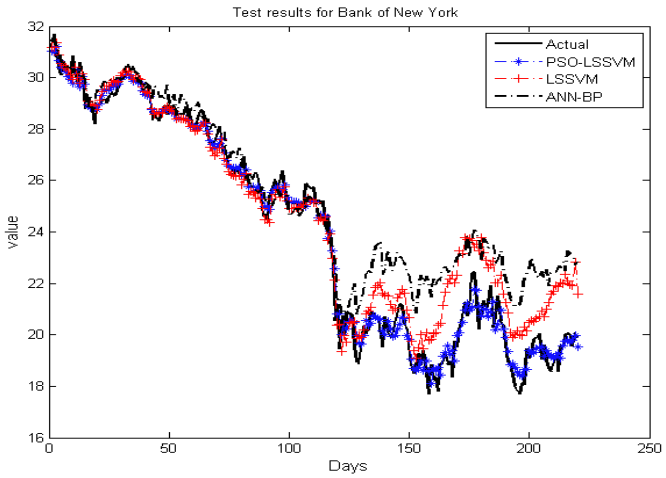

Fig. 7: Results for Bank of New York

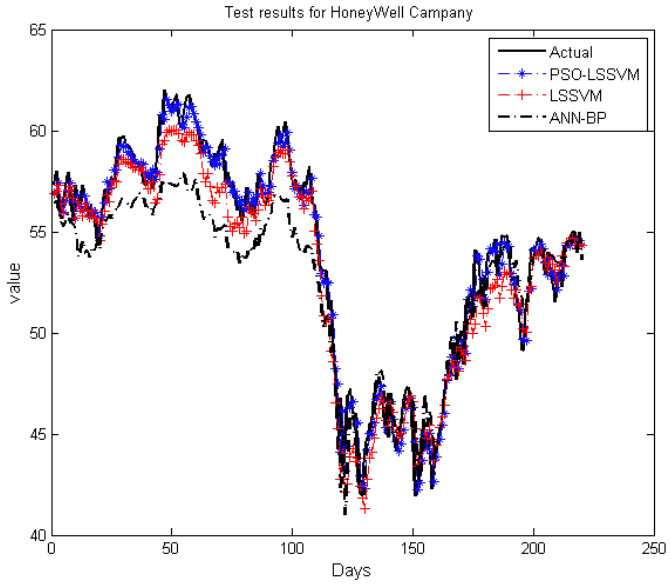

Fig. 8: Results for Honeywell Company

VII. CONCLUSION

Machine learning as we have seen till now, is a very powerful tool and as evitable, it has some great application. We have seen till now that machine learning is very much dependent upon data. Thus it is important understand that data is quite invaluable and as simple is it may sound, data analysis is not an easy task.

Machine learning have found tremendous application and has evolved further into deep learning and neural networks, but the core idea is more or less the same for all of them. This paper delivers a smooth insight of how to implement machine learning. There are various ways, methods and techniques available to handle and solve various problems, in different situations imaginable. This paper is limited to only supervised machine learning, and tries to explain only the fundamentals of this complex process.

ACKNOWLEDGMENT

This paper is written based upon the final year project of the author during his degree programme in the year 2020 after the invaluable guidance by Prof. Ganesha M (Asst. Prof.) RRIT, India.The author of this paper, does not claim rights to any of thealgorithm, code, data, formula used, definitions, problemsolving approach, as his property. He has only used his brain in compiling it all together and made efforts in obtaining results and putting it together in the format of an IJIRCCE paper.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

REFERENCES

[1] Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore "A Machine learning approach to Building domain-specific Search engine", IJCAI, 1999 – Citeseer
[2] Yadav, Sameer. (2017). STOCK MARKET VOLATILITY – A STUDY OF INDIAN STOCK MARKET.Global Journal for Research Analysis. 6. 629-632.
[3] Montgomery, D.C., Peck, E.A. and Vining, G.G., 2012. Introduction to linear regression analysis (Vol. 821).John Wiley & Sons.
[4] Draper, N.R.; Smith, H. (1998). Applied Regression Analysis (3rd ed.). John Wiley.ISBN 0-471-17082-8.
[5] Robert S. Pindyck and Daniel L. Rubinfeld (1998, 4h ed.). Econometric Models and Economic Forecasts
[6] "Linear Regression", 1997-1998, Yale University http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
[7] Agarwal (July 14, 2017). "Introduction to the Stock Market".Intelligent Economist.Retrieved December 18, 2017.
[8] Jason Brownlee, March 2016, "Linear Regression for machine learning", Machine learning mastery, viewed on December 2018, https://machinelearningmastery.com/linear-regression-for-machinelearning/
[9] Google Developers, Oct 2018, "Descending into ML: Linear Regression", Google LLC, https://developers.google.com/machinelearning/
crash-course/descending-into-ml/linear-regression
[10] Fiess, N.M. and MacDonald, R., 2002. Towards the fundamentals of technical analysis: analyzing the information content of High, Low and Close prices. Economic Modeling, 19(3), pp.353-374.
[11] Hurwitz, E. and Marwala, T., 2012. Common mistakes when applying computational intelligence and machine learning to stock market modeling.arXiv preprintarXiv:1208.4429.