Predicting Financial Markets using Machine Learning Algorithms

Rohith Pattathil

***Abstract*— One of the financial applications of time series forecasting techniques is predicting stock prices. The aim of this application is to predict the future value of a company's stock. Stock market prediction has received a lot of focus from active traders, investors and researchers. Predicting stock prices can be done using multiple Machine Learning algorithms including but not limited to Moving Average, Linear Regression, ,Long Short Term Memory and Support Vector Regression. Most of these algorithms provide a decent rate of accuracy for predicting stock prices. In this paper, we will use the Investpy python API (Link to API `https://pypi.org/project/investpy/`), to import the data-set of any user specified stock and the data-set will include the open, low, high, close and volume of the company for a specific date range. We will compare and contrast multiple Machine Learning and Neural Network techniques.**

I. INTRODUCTION

There are two basic approaches for predicting how a stock would perform on any given day. The first one being using fundamental analysis. This approach is used by long term investors, who find companies with very strong foundation and current financial condition of the company. What we mean by strong foundation is, the company's current Debts, Assets, past and recent mergers or acquisitions , market capitalisation, Cash flow, and Profit and Loss. The second being Technical analysis, in which historical prices and trends are analysed to predict how the stock would perform on a future date or when the trend would repeat. Technical analysis in most cases is performed by aggressive traders who buy and sell in short term. Short term could mean a quarter or even mean day trading. For using Machine Leaning in predicting stock prices, in this paper we will use Technical Analysis since understanding financial condition of a company requires human intervention and is a more robust topic, which would be considered as a future work for this project

Before diving into Technical analysis and how machine learning algorithms are being applied for it, let us first understand the basic terminologies being used in this paper

- *Open Price :* The price with which the stock opened on a given date
- *Close Price :* The price with which the stock closed on a given date
- *High :* The highest price the stock went to on a given date
- *Low :* The lowest price the stock had on a given date

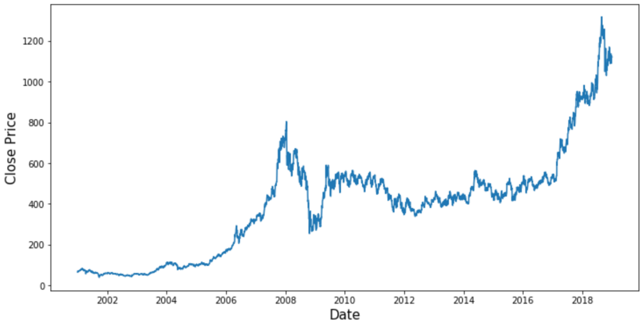


Fig. 1: Reliance close price graph from 2001 to 2019

To understand how stock market prices are plotted against dates, let us consider Figure 1, which is plot of close price for the company "Reliance India Pvt. Ltd" against the dates in range 2001 to 2019. We will be using this company for further analysis in this paper, although the project has been coded to perform all algorithms on a company that is specified by the user.

The most important part of Machine Learning is the dataset being used. In this project, all algorithms are being applied on dataset extracted from the python API Investpy. The API is real time and provides the most concrete dataset. It includes the Open, High, Low and Close price sorted by dates of the stock being searched for. It also provides the volume of trade performed on a given date. We divide the dataset obtained into test and train data and check the results after different machine learning algorithms are applied. Moving Average, Linear Regression, Long Short Term Memory and Support Vector Regression are performed on the selected stock.

The remainder of this paper has the following parts: Section II provides inputs on the prior work in this research area and how each paper, journal or book influenced this project. Section III describes the methodology being used for the project. Section IV provides detailed explanation on each algorithm used and their advantages and disadvantages. The section V contains graphs plots, confidence and the root mean squared error for each algorithm. We plot graphs on the actual vs predicted price over a period of time to understand which algorithm performs better in the given scenario. The Section VI contains the conclusions of the project and how it can be further improved. The last Section contains the list of all references that were used in order to complete this project.

II. Prior Work

Stock market prediction is exceptionally complex and challenging because prices vary randomly and sometime only because of a market sentiment. After the literature survey, It was easy to conclude that using Time Series Machine Learning techniques, to make stock market predictions were being accepted positively across the globe. Machine Learning methods are demonstrating to be much more precise and quicker as compared to contemporary prediction methods.

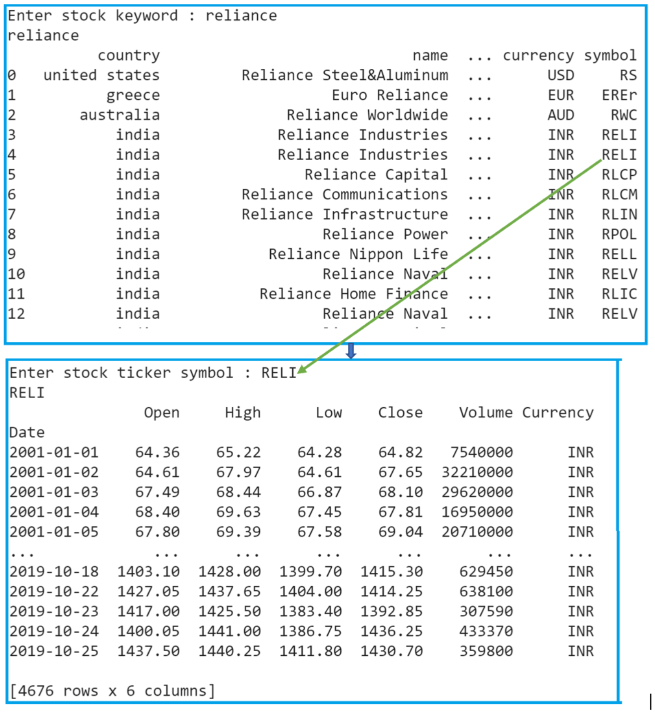
As a part of the literature survey, some of the papers and journals referred are discussed in this section. Firstly,"Developing a Prediction Model for Stock Analysis" by R. Yamini Nivetha and Dr. C. Dhaya[3]. This paper helped understand Support Vector Machine. Secondly, the paper "Stock Market Prediction Using Machine Learning" by Ishita Parmar, Navanshu Agarwal and others [6], This paper provided a lot of critical fundamental knowledge on Linear Regression and Long Short Term Memory models.

Thirdly, "Survey of Stock Market Prediction Using Machine Learning Approach" [7] by Ashish Sharma, Dinesh Bhuriya and Upendra Singh. This paper was referred to understand the different types of regression and how each type can be compared with the others. Han Lock Siew and Md Jan Nordin's paper "Regression Techniques for the Prediction of Stock" [8] provided insightful research on applying regression on stock predictions. The book "Introduction to Data Mining" by Pang-Ning Tan , Michael Steinbach, and Vipin Kumar was also referred as a comprehensive guide to time series data analysis.[5]

III. Methodology

Stock market prediction is an intricate problem due to the varying number of variables that cause changes to the stock price. Stock prices may seem very ambiguous at start, requires no definitive reason for a fluctuation and it may even look like gambling but there have been concrete evidences where humans were able to predict stock prices using financial numbers and graphs such as candle stick graphs. As per popular belief, machines can have better accuracy on comparison to humans. This was the initial train of thought because of which Machine Learning algorithms were used for stock price prediction. By legitimate utilization of Machine Learning techniques, one can relate historical data to future data, train the algorithm to learn from it and make required presumptions.

The dataset being used is being imported from the python API Investpy. Investpy API is available to import from the library and is developed for research purposes. Once imported, the dataset requested contains the stock prices and other relevant information that are required for the predictions. The data contains the stock price at regular intervals over the period that is being requested during import. It includes the Open, Close, High , Low prices , Date, Volume and Currency. All the data obtained is converted into dataframe using the Python's Pandas library. After which, normalisation of the data is done using the sklearn python library. We then divide the data into train and test in the ratio 80:20 respectively. Although Machine Learning has tons of models, we will only compare and contrast Moving Average, Linear Regression, Support Vector Regression and Long Short Term Memory in this project. All the analysis is being performed for one stock, namely "Reliance Industries" in this report although the Python code developed requests the user for input on the "Company Ticker Symbol" they would like predictions for. Figure 2 shows in the input screen for user to select company for the predictions.



```
Enter stock keyword : reliance
reliance
          country                        name  ... currency symbol
0   united states      Reliance Steel&Aluminum  ...      USD     RS
1          greece               Euro Reliance  ...      EUR   EREr
2       australia          Reliance Worldwide  ...      AUD    RWC
3           india         Reliance Industries  ...      INR   RELI
4           india         Reliance Industries  ...      INR   RELI
5           india            Reliance Capital  ...      INR   RLCP
6           india     Reliance Communications  ...      INR   RLCM
7           india     Reliance Infrastructure  ...      INR   RLIN
8           india              Reliance Power  ...      INR   RPOL
9           india         Reliance Nippon Life  ...      INR   RELL
10          india              Reliance Naval  ...      INR   RELV
11          india       Reliance Home Finance  ...      INR   RLIC
12          india              Reliance Naval  ...      INR   RELV
```

```
Enter stock ticker symbol : RELI
RELI
               Open     High      Low    Close    Volume Currency
Date
2001-01-01    64.36    65.22    64.28    64.82   7540000      INR
2001-01-02    64.61    67.97    64.61    67.65  32210000      INR
2001-01-03    67.49    68.44    66.87    68.10  29620000      INR
2001-01-04    68.40    69.63    67.45    67.81  16950000      INR
2001-01-05    67.80    69.39    67.58    69.04  20710000      INR
...             ...      ...      ...      ...       ...      ...
2019-10-18  1403.10  1428.00  1399.70  1415.30    629450      INR
2019-10-22  1427.05  1437.65  1404.00  1414.25    638100      INR
2019-10-23  1417.00  1425.50  1383.40  1392.85    307590      INR
2019-10-24  1400.05  1441.00  1386.75  1436.25    433370      INR
2019-10-25  1437.50  1440.25  1411.80  1430.70    359800      INR

[4676 rows x 6 columns]
```

Fig. 2: Selection screen for user to chose stock for predictions.

IV. Algorithms

Let us now take a deep dive into each algorithm,

A. Moving Average

"Average" is one of the foremost concepts we use in our daily lives frequently. Simple Moving Average is one of the most primitive time series prediction algorithms. In our application, Moving Average displays the stock trend by calculating average stock price in a periodic manner.[2]The prices used for this average calculations are the close price on each day. The formula for moving average is given by,

$$\mathbf{MA}_n = \sum_{i=n-T}^{n-1} \frac{x_i}{T} \tag{1}$$

Where $\mathrm{MA_n}$ is the moving average price of the $\mathrm{n^{th}}$, day $\mathrm{x_i}$ is the stock closing price on $\mathrm{i^{th}}$day and T is the duration for the moving average.

This can be better understood by the illustration shown in Figure 3.

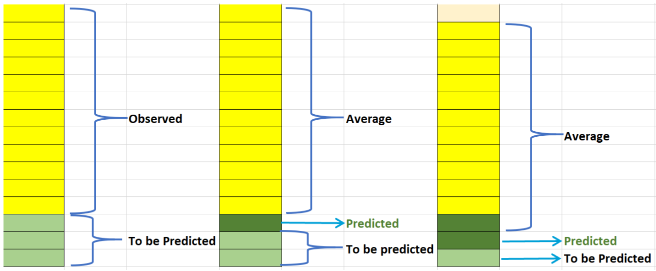
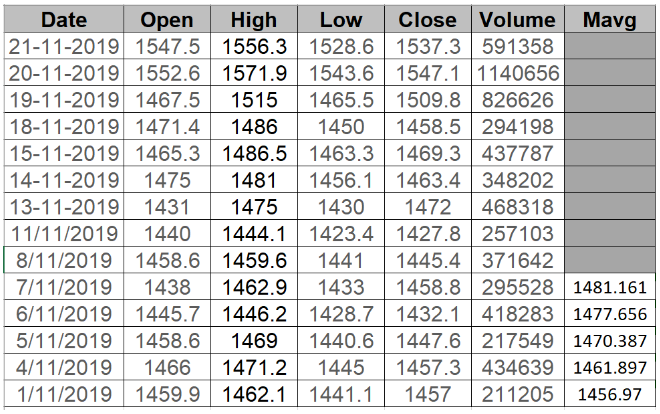


Fig. 3: Basic principle of Moving Average Calculations [9]

The durations in moving average are usually multiples of 5 or 10. Figure 4 [9] shows a sample of the dataset obtained. ”Close” column in the figure represents the closing price of the stock on a particular day.



Date	Open	High	Low	Close	Volume	Mavg
21-11-2019	1547.5	1556.3	1528.6	1537.3	591358	
20-11-2019	1552.6	1571.9	1543.6	1547.1	1140656	
19-11-2019	1467.5	1515	1465.5	1509.8	826626	
18-11-2019	1471.4	1486	1450	1458.5	294198	
15-11-2019	1465.3	1486.5	1463.3	1469.3	437787	
14-11-2019	1475	1481	1456.1	1463.4	348202	
13-11-2019	1431	1475	1430	1472	468318	
11/11/2019	1440	1444.1	1423.4	1427.8	257103	
8/11/2019	1458.6	1459.6	1441	1445.4	371642	
7/11/2019	1438	1462.9	1433	1458.8	295528	1481.161
6/11/2019	1445.7	1446.2	1428.7	1432.1	418283	1477.656
5/11/2019	1458.6	1469	1440.6	1447.6	217549	1470.387
4/11/2019	1466	1471.2	1445	1457.3	434639	1461.897
1/11/2019	1459.9	1462.1	1441.1	1457	211205	1456.97

Fig. 4: Example ”Reliance” stock price obtained

If we set T=10 and would like to calculate the moving average for the next 5 days, we average closing price from 21-11-2019 to 08-11-2019 and use that as the price for the next day.This procedure is followed for the next 4 days. To understand the trend, the best approach is to identify two moving averages. One long-term with a larger value for T and another short-term with a smaller value for T. If the short-term trend is larger than long-term it shows the trend will go up and vice-versa.

B. Regression

Regression is a statistical procedure for identifying relationship between variables. For better understanding, it is the study how dependent variable changes when an independent variable is changed while the other independent variable is held constant.[7] In most cases, we calculate the mean value of the dependent variable while fixing the independent variables at a constant value. Also, regression function is the estimation target is a function of the dependent variable. Regression analysis can also be looked at as probability function, i.e understanding variation of dependent variable around the regression function.

There are multiple techniques developed for regression analysis and they are broadly categorized into parametric and non-parametric regression analysis. Parametric regression analysis has a finite number of unknown parameters where as, in the case of non-parametric regression number of parameters is infinite, which can also be rephrased as infinite dimensions. Regression analysis is broadly used for prediction and forecasting. This brings us to its application in stock market prediction. Linear regression and polynomial regression are types of parametric regression and we will be using these for our analysis.

1) Linear Regression: One of the most basic Machine Learning Algorithm is the Linear Regression. This model returns a condition that identifies the relationship between the independent variables and the dependent variables.[7] The equation for linear regression can be written as follows

$$Y = \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + .. + \delta_n x_n + \epsilon \qquad (2)$$

where, Y is the dependent variable, $\mathrm{x_1}$,$\mathrm{x_1}$,$\mathrm{x_1}$ ..$\mathrm{x_n}$ represents the independent variables, $\delta_1, \delta_2, \delta_3..\delta_\mathrm{n}$ represent the weights and ϵ is the unobserved random error. Since the degree of this equation is 1, linear regression always plots a straight line. In many cases including stock market predictions linear regression may not hold, which is why we improve this condition by increasing the degree of the polynomial.

2) Polynomial Regression: Polynomial regression is another kind of regression where the degree of the equation is greater than 1 in contrast to linear regression. Polynomial regression fits a nonlinear relationship between the value of x and corresponding value of conditional value of y,denoted by $E(y|x)$, has been used to describe nonlinear phenomena.[7] Although polynomial regression fits a nonlinear model to the data, as an estimation problem it is linear, in the sense that the regression function $E(y|x)$ is linear in the unknown parameters that are estimated from the time series data. This the reason why polynomial regression is considered a exclusive case of multiple linear regression

In most applications of Polynomial regression we model the value of y, the dependent variable as an n_{th} degree polynomial, which gives as the basic equation for Polynomial regression model[7],

$$Y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + .. + a_n x^n + \epsilon \qquad (3)$$

This equation is also considered linear from the estimation point of view since the regression function is linear considering unknown variables a_1,a_2,..a_n. We also consider x^1, x^2, ..., x^n as independent variables in this type of regression.

C. Support Vector Regression

To understand Support Vector Regression, It is important to first understand what Support Vector Machines do

1) Support Vector Machine: In machine learning, support vector machines (SVM) are supervised learning models which has an associated learning algorithms that analyze data used for classification or regression analysis.[4] It creates a hyperplane or group of hyperplanes in large dimensional spaces. These can be used for classification or regression. A functional margin is created with the largest distance to the nearest training data of the dataset. The formula derived for this is

$$Y = w_0 + w_1 x_1 + w_2 x_2 \tag{4}$$

where, w_0,w_1 and w_2 are learned by the SVM and x_1 and x_2 are the inputs. [4] Figure 5 illustrates Support Vector Machine diagrammatically for better understanding.

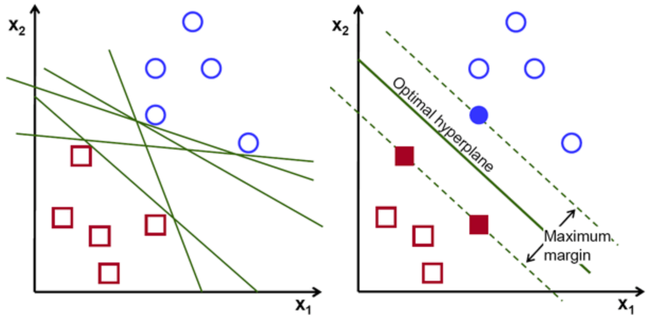


Fig. 5: Illustration on SVM [1]

2) Support Vector Regression: Support Vector Regression tries to minimize complexity by reducing $||\omega||^2$ and accomplishes linear regression in high dimentional space using ϵ, which is the loss. Also, Slack variances ξ_i,ξ_i^* are introduced, where i =1,2...n, which measures the abnormal training data outside ϵ in the dataset.[4] The function that is minimized by SVR is :

$$\frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{5}$$

$$min = \begin{cases} y_i - f(x_i, \omega) \leq \epsilon + \xi_i^* \\ f(x_i, \omega) - y_i \geq \epsilon + \xi_i^* \\ \xi_i \xi_i^* \geq 0, i = 1, 2..n \end{cases} \tag{6}$$

This optimization can be changed into a dual problem ,which can be given by

$$f(x) = \sum_{i=1}^{n_{sv}}(\alpha_i - \alpha_i^*)K(x_i, x) \tag{7}$$

where , $0 \leq \alpha_i^* \leq C$ and $0 \leq \alpha_i \leq C$ and n_{sv} is the number of support vectors used. [4]

D. Long Short Term Memory

Long Short Term Memory is an advanced part of Artificial Neural Networks, which is also recurring. In this type of Nueral Network the previous state is preserved. The main difference between recurring Neural Network and LSTM is that, RNNs have involvement with short term dependencies while LSTM is long term. This is the reason LSTM was chosen for stock market prediction in multiple researches and also is a part of this project. The stock market prediction depends upon large amount of data and is dependent on long term history of the company. Therefore, LSTM calculates error by using RNNs with long term memory which is the reason for its higher accuracy rates.[6] Long Short Term Memory can be understood by the below illustration

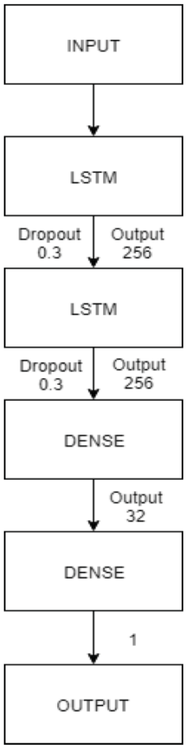


Fig. 6: Illustration on Long Short Term Memory [6]

LSTM has an remembering cell, an input gate, an output gate as well as an forget gate. The cell remembers the long term propagation and the gates are used to regulate the cell.

V. Experiments and Results

The proposed Machine Learning algorithms are implemented on the dataset obtained from Investpy API for the company Reliance Industries and the below results were obtained.

A. Moving Average

The plot in figure 7 portrays the application of moving average on the dataset. The blue , green and red graphs represents the actual data, the rolling mean for 100 days, and the rolling mean for 200 days respectively.

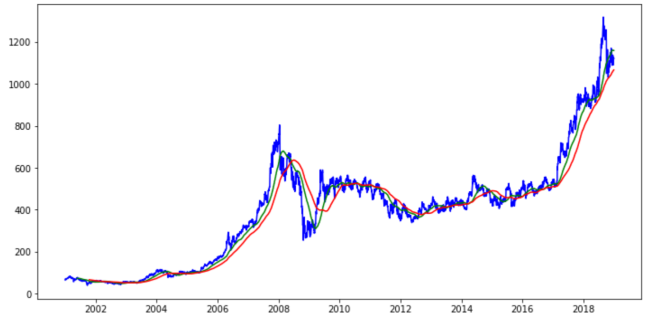


Fig. 7: Illustration of Moving Average (100 and 200 rolling mean)

The Root Mean Squared Error obtained for the moving average of rolling mean for 100 days is 76.62 and rolling mean for 200 days is 51.67. Although the moving average does not provide promising results based on the graph and root mean squared error, moving average can be used to predict the trend i.e, if the company's stock price is likely to go higher or lower in the near future as explained in the algorithm section.

B. Linear and Polynomial regression

The graph plotted in figure 8 represents the application of Linear regression(degree 1) and Polynomial regression of degree 5. The Linear regression confidence is 70.22% while the polynomial regression confidence is 77.84%. The Root mean squared error obtained for Linear regression and Polynomial regression is of 345.07 and 337.64 respectively. This proves Polynomial Regression tries to fit the polynomial as close as the actual graph as possible and a polynomial of greater degree can produce better results. We can also use the obtained results to predict trend of stock price for the company as in the case of moving average.

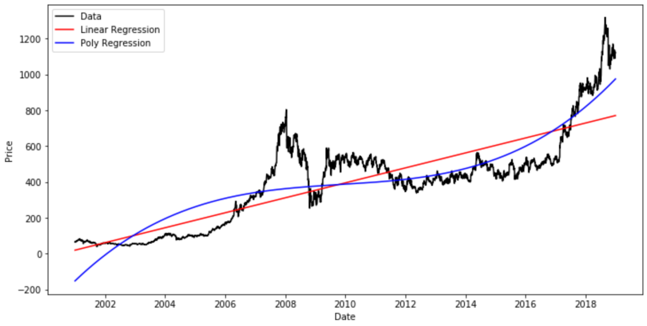


Fig. 8: Illustration of Linear and Polynomial Regression

C. Long Short Term Memory

The graph in Figure 9 shows the results of application of Long Short Term Memory on the dataset. The long Short Term Memory provides better results on comparison with other Machine Learning Algorithms. The Root Mean Squared Error that was obtained is of 49.22. The Confidence is 99.9%.

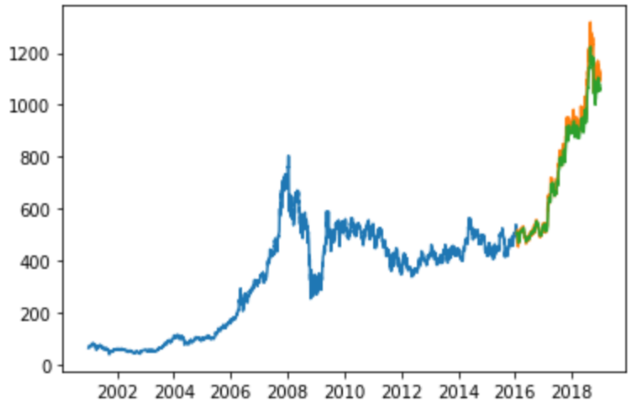


Fig. 9: Illustration of LSTM against actual data

D. Support Vector Regression

The graph in Figure 10 helps us understand the application of Support Vector Regression. The Support Vector Regression provides a very high confidence of 99.99%, Which is a very high confidence on comparison to the other approaches. The RMSE obtained for SVR is 2.84, which again proves its efficiency and accuracy on comparison with other approaches.

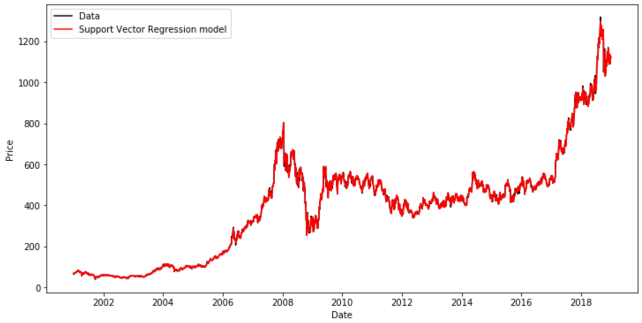


Fig. 10: Illustration of Support Vector Regression against actual data

VI. Conclusions

This project was an experiment to predict stock prices and identify the best machine Learning algorithm among the 4 methods used in the project. All models have given reasonable results among which Support Vector regression and Long Short Term Memory gave the best results, Although the other approaches can be used to predict the stock trend but not the actual price.

If this project was to be used in real life, Best recommendation would be to use all four approaches and buy or sell stocks based on all the results. Also, it is critical to note that stock prices are also dependent on the company's financial fundamentals and financial changes that happen over night. These changes can be mergers, acquisitions , quarterly profits, to name a few. Such changes can substantially impact the stock price on the next day and

Machine Learning algorithms may not be able to predict them.

As a future scope for this project, I would like to use a web crawler to look up websites like Yahoo Money, MoneyControl, news channels and other social media websites to identify market sentiment and recent company changes. These can be used to further enhance the existing Machine Learning Algorithms implemented in the current project.

REFERENCES

[1] Rohith Gandhi in website Towards Data Science. for svm. https://towardsdatascience.com/support-vector-machine-introduction-to -machine-learning-algorithms-934a444fca47.

[2] S. Lauren and S. D. Harlili. Stock trend prediction using simple moving average supported by news classification. In *2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 135–139, Aug 2014.

[3] R. Y. Nivetha and C. Dhaya. Developing a prediction model for stock analysis. In *2017 International Conference on Technical Advancements in Computers and Communications (ICTACC)*, pages 1–3, April 2017.

[4] M. Ouahilal, M. El Mohajir, M. Chahhou, and B. E. El Mohajir. Optimizing stock market price prediction using a hybrid approach based on hp filter and support vector regression. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pages 290–294, Oct 2016.

[5] Michael Steinbach Pang-Ning Tan and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

[6] I. Parmar, N. Agarwal, S. Saxena, R. Arora, S. Gupta, H. Dhiman, and L. Chouhan. Stock market prediction using machine learning. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, pages 574–576, Dec 2018.

[7] A. Sharma, D. Bhuriya, and U. Singh. Survey of stock market prediction using machine learning approach. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 2, pages 506–509, April 2017.

[8] H. L. Siew and M. J. Nordin. Regression techniques for the prediction of stock price trend. In *2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE)*, pages 1–5, Sep. 2012.

[9] Aishwarya Singh's website. for moving average. https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/?.