Accurate Prediction of the Stock Price Using Linear Regression Model of Machine Learning

Saksham Sharma, Manthan Kulkarni Sonu Kumar Mourya,

Thadi Guna Vardhan, Nidhi Lal

Dept. of Computer Science and Engineering
IIIT Nagpur, India

amsakshamsharmaofficial@gmail.com , manthandec123@gmail.com , sonumouryachp@gmail.com , gunavardhanthadi@gmail.com , nidhi.2592@gmail.com

Abstract - If a man investor can be successful why can't a machine ?Stock prices are a function of information and rational expectations , and that newly revealed information about company's prospect is almost immediately reflected in the current price of the stock. Considering growth in stock Market, prediction of Stock Market by using Machine Learning Models had increased with the development of Machine Learning. Optimisation of Root mean squared error significantly using
cluster analysis on data that is by splitting data set into clusters depending on their type . Splitting was done on data set which contained similar linearity which resulted lowering RMSE value by preprocessing training and testing set .

I. INTRODUCTION

A stock market, equity market or share market is the aggregation of buyers and sellers of stocks (shares), which represent ownership claims on businesses; these may include *securities* listed on a public Stock exchange, as well as stock that is only traded privately.Difficulty in prediction comes from the complexities associated with market dynamics where parameters are constantly shifting and are not fully defined [1].Examples of the latter include shares of private companies which are sold to investors through equity crowdfunding platforms. Stock exchanges list shares of common equity as well as other security types, e.g. corporate bonds and convertible bonds. Vector regression approach for stock market forecasting can also be used [2].'Average' is easily one of the most common things we use in our day-to-day lives. For instance, calculating the average marks to determine overall performance, or finding the average t emperature of the past few days to get an idea about today's temperature – these all are routine tasks we do on a regular basis. So this is a good starting point to use on our dataset for making predictions.Survey by extreme machine learning[3]. The predicted closing price for each day will be the average of a set of previously observed values. Instead of using the simple average, we will be using the moving average technique which uses the latest set of values for each prediction. Stock market prediction through empirical analysis[4].In other words, for each subsequent step, the predicted values are taken into consideration while removing the oldest observed value from the set. Financial news analysis is also used [5].The machine learning algorithm that can be implemented on this data is linear regression. The linear regression model returns an equation that determines the relationship between the independent variables and the dependent variable.

- Y = M1*X1+M2*X2+ ... +Mn*Xn
- The equation of linear regression line -

$$h(x\_i) = (b\_0)+ (b\_1) *(x\_i)$$

h(x_i) represents the predicted response value for 'ith' observation. And b_0 and b_1 are regression coefficients and represent y-intercept and slope of the regression line respectively.Multivariate linear regression model is one where there is one continuous dependent variable (E.R.P in our case) which is dependent on multiple factors and predictors.These factors are independent variables in a formula to calculate the dependent variable.

$$Yi = \beta0+\beta1xi1+\beta2xi2+...+\beta pxip+\epsilon$$

where, for i=n observations:

yi = dependent variable(Estimated Relative Performance)

xi = explanatory variables

β0 = y-intercept (constant term)

βp =slope coefficients for each explanatory variableϵ=the model's error term (also known as the residuals)

II. Related work

The Data set was of Tata stocks on which stock price closing prediction was done by using linear regression with certain random mean squared value .Improving prediction by market news and stock prices[6]. Usage of discrete probability mass function[7]. Using multivariate statistical model in stock market prediction [8]. Linear regression is a simple technique and quite easy to interpret. One problem in using regression algorithms is that the model overfits to the date and month column. Instead of taking into account the previous values from the point of prediction, the model will consider the value from the same *date* a month ago, or the same *date/month* a year ago. Trusting stock market[9]. The obtained RMSE value previously was a bit higher one. Using Linear Regression with few libraries got the RMSE value to be on higher side value .Probability and statistical inference [10].Plot of closing price Vs Date/Month/Year was plotted .

Fig1 Graph between closing price and date(previous)

RMSE value was betterly founded with Linear Regression to be around Ton ! Which was found to be little with other methods . But with Linear Regression the efficiency of the model was less .Linear Regression applied on plot didn't turn out to be promising as it was expected.

III. Proposed Work

As the main theme is prediction of stock price using linear regression model . The given data set was broken into clusters according to the type to which they belong with respect to the whole data set.After analyzing the data set clusters were created which contained three clusters in total. Analyzation turned to promising as linear regression for minimising RMSE got better . With Linear regression as Model technique training and testing part was done differently on each three of the clusters and the RMSE obtained was very optimistic from all three .Now the RMSE for the entire Data Set was mean of the three RMSE obtained from this clusters.

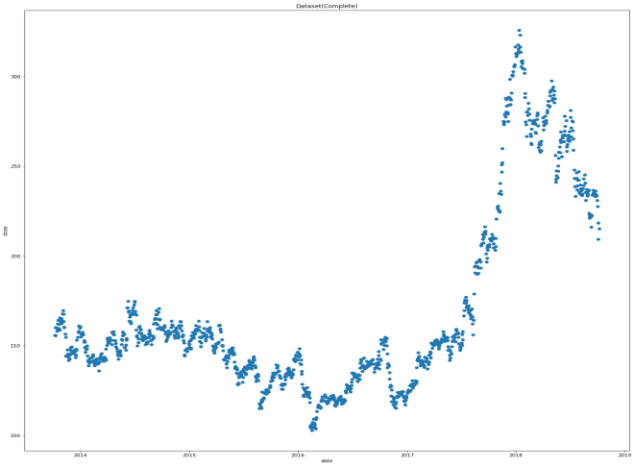

Fig 2 Total Dataset of closing price Vs days from 2014 to 2019.

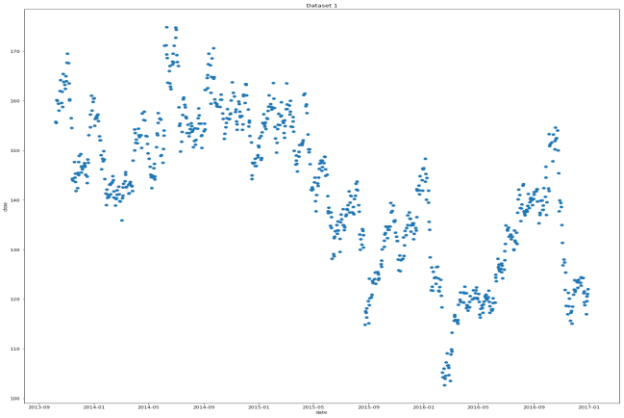

Fig3 Dataset of closing price Vs days from 2014 to 2017.

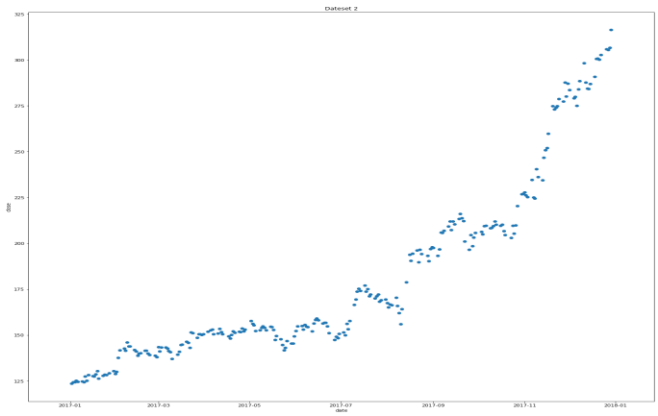

Fig 4 Dataset of closing price Vs days from 2017 to 2018.

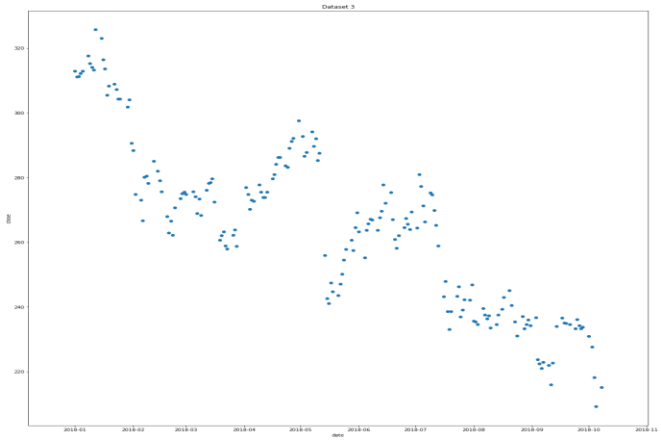

Fig 5 Dataset of closing price Vs days from 2018 to 2019.

IV. Result and Discussion

Most of data set exhibit linear behavior with some extent of non linearity and because of this RMSE value increases and reliability of the regression model decreases and so tackle such situation data set can be split based on the parts where it exhibit non linearity in order to get multiple data sets or derived data sets having increased linearity thus lowered RMSE which in turn lower the overall RMSE for the whole dataset.The above applied technique resulted in reducing the RMSE value for the given dataset significantly and also gives a method to resolve non-linearity in values by splitting the dataset and applying linear regression individually which in turn result in improving the RMSE value and more accurate result in our case the stock market prediction.The Fig 2 shows the data before splitting and it is not linear thus when linear regression RMSE value of 104 is observed but when graph is clearly observed it is clear if graph is split into three sections linearity in each increases as depicted in Fig 3,4 and 5 and thus in our work we came reduced RMSE value from 104 to 15 by splitting dataset method.

V. Conclusions

Another way to lower the RMSE value further training set and testing can be resetted for finding the minimum RMSE

possible and increase accuracy.Future work include effective method of splitting the dataset for more efficient analysing.
The prediction of closing stock with achieved RMSE would be better one but it can't be exact accurate everytime as stock price is affected by news about the company and other factors about demonetization or merger /demerger of the companies .By splitting the dataset according to the non linearity observed we can reduce the RMSE value significantly.

REFERENCES

[1]. Schumaker RP, Chen H (2010) A discrete stock price prediction engine based on financial news. Computer 43(1):51–56

[2] Yeh C-Y, Huang C-W, Lee S-J (2011) A multiple-kernel support vector regression approach for stock market price forecasting. Expert Syst Appl 38:2177–2186

[3] Huang G-B, Wang DH, Lan Y (2011) Extreme learning machines: a survey. Int J Mach Learn Cybernet 2(2):107–122

[4] Li, X., Xie, H., Wang, R., Cai, Y., Cao, J., Wang, F., … Deng, X. (2014). Empirical analysis: stock market prediction via extreme learning machine. Neural Computing and Applications, 27(1), 67–78. doi:10.1007/s00521-014-1550-z

[5] Seo Y-W, Giampapa J, Sycara K (2004) Financial news analysis for intelligent portfolio management. PhD thesis, Robotics Institute, Carnegie Mellon University

[6] Li X, Wang C, Dong J, Wang F, Deng X, Shanfeng Z (2011) Improving stock market prediction by integrating both market news and stock prices. In: Hameurlain A, Liddle S, Schewe K-D, Zhou X (eds) Database and expert systems applications. Lecture notes in computer science, volume 6861. Springer, Berlin, pp 279–293

[7] Aitchison, J., 1955. On the distribution of a positive random variable having a discrete probability mass at the origin. Journal of the American Statistical Assocation 50, 901–990.

[8] Fahrmeir, L., Tutz, G., 1996. Multivariate Statistical Modelling Based on Generalized Linear Model. Springer-Verlag, New York, N.Y.

[9] Guiso, L., Sapienza, P., Zingales, L., 2007. Trusting the stock market. NBER working paper.

[10] Spanos, A., 1999. Probability Theory and Statistical Inference: Econometric Modeling with Observational Data. Cambridge University Press, NY