Stock Market Prediction on High Frequency data

M. Sanjay Kumar Reddy
Department of Computer Science and Engineering,
Kalasalingam Academy of Research and Education,
Virudhnagar, Tamil nadu ,India
9920004280@klu.ac.in

N. Vinay Kumar Reddy
Department of Computer Science and Engineering,
Kalasalingam Academy of Research and Education,
Virudhnagar, Tamil nadu ,India
9920004374@klu.ac.in

S. Bhavani
Department of Computer Science and Engineering,
Kalasalingam Academy of Research and Education,
Virudhnagar, Tamil nadu ,India
9920004298@klu.ac.in

R. Sumathi
Department of Computer Science and Engineering,
Kalasalingam Academy of Research and Education,
Virudhnagar, Tamil nadu ,India
r.sumathi@klu.ac.in

***Abstract*—Prediction of prices in Stock Market is an complex task. It involves more contact between human and computer. We will use more efficient algorithms to get the result more accurate. The proposed methodology here is Linear Regression, Ridge Regression, Lasso Regression and Polynomial Regression. This case will provide us the accurate results and this experiment results are effective and suitable for the prediction. Firstly we will collect the data from the kaggle, then we will apply the proposed algorithms and the code is changed according to the results we get the accuracy we are getting. Finally this includesthe work flow of the prediction of share market. The results from the experiment can show that the methodology suggested is remarkably productive and also appropriate for predicting before a short period of time .**

Keywords*— *Ridge Regression, Lasso Regression, Polynomial Regression , Linear regression

I. INTRODUCTION

Stock Market is a vast area where the capitalists and dealers try to trade in and purchase, by alternating the stock prices up or down. The Stock prices are determined by principles of desire of client consumers and provision. The Ultimate destination of purchasing shares is to make more wealth by purchasing stocks in companies where we become aware of rise in the share price.
Predicting prices in Stock Market is a tough job as the prices affect on various circumstances. As part of long established process, there are mainly two preceding methods that are proposed in order to predict the prices in Stock Market of a particular organization.
These days, with the recently developed high level skillfulness in accordance with specialized or basic analyzing has helped in predicting the prices in stock market.
Specifically, in analyzing the market in stocks it is not an easy job the data here is very vast and also uncoordinated.
Here we are going to use Deep learning techniques to approach this project. The exchanges in stock market frequently tend to be regulated by any advancement in that province. As a result, the figure of changing and the body of facts provided are noteworthy to a certain extent such that in increasing financial gain as well as resolving period of time as well.

At every case of prediction or at any case firstly we will consider calculating Variance on to the other algorithms if thatwas not fitted we will move on to the another like that we areproposing four algorithms Polynomial Regression, Linear Regression, Ridge Regression and Lasso Regression from all these we will conclude which algorithms suits best for more accuracy and why that suits best. For knowing this we will follow a step by step process

Collecting data and then we will process it then testing and training and finally we will conclude the best algorithm and we will also get the more accurate results.

A. General Work flow of Stock Market Prediction

Every analytical work starts with the data collection.
Now we wish for predicting the clearing rate with the help of time-series data collected at an extremely fine scale by different types of algorithms.

This includes the 3 crucial elements:

1. Collection of Information
2. Information preprocessing
3. Prediction

we wish for predicting the clearing rate with the help of time-series data collected at an extremely fine scale combining classical financial models and different types of algorithms.

II. BACKGROUND AND RELATED WORKS

After collecting all the data as files and data bases etc. we will move on to the next step data preprocessing.
Data preprocessing

Firstly we will assemble S&P500 intraday trading data from Kaggle(Kaggle Datasets).The genuine information documents consists of 1234 statements. For each statement, each record of time or date, open , high , low , close price and last are at hand . The mean range b/w cost is 1 min. And this data preprocessing approach follows several steps :

1. Data cleaning
2. Feature creation
3. Feature scaling and normalization
4. Training Data and Testing Data
5. Implementing Algorithms

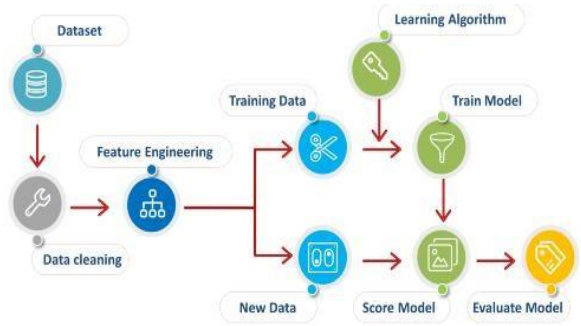

Fig 1 : The flow of Stock prediction

Data cleaning detects and removes the major errors. This is all regarding determining the baseless and absent values and managing this one. From the information set taken by us contains two types of absent information conceivably divided towards two classifications : Values listed as absent for a brief span and absent values for a protracted span . Next, we practiced two methods:
1. Value imputation: mis values for a brief span (b/w 1 and 5 in-stances) are exchanged with above values. 2. Instances discard: absent values for a protracted span are discontinued.

Feature creation: Feature manufacturing, aka feature design, is the action of assembling latest characteristics from genuine information to upskill a machine learning model. Dealers are habitually challenged with distinctive tactics. We can then difference between specialized examination and basic examination. Specialized examination depend at most on the stock market, considering past value and few measures

By implementing more features we will get the more accurate results and more exact prediction can be done with the increasing of more features the efficiency of the work will be more and also increased.

Feature scaling: Upskilling and experimenting information are systematized where ˜x correspondent to x after standardization and Min is the minimal value of x after a period of time and over the examinations, Max is the maximal value of x. After attaining the estimated output, de-systematized is tested that one may obtain the genuine beneficial value.

$$\bar{x} = x - \text{Min} / \text{Max} - \text{Min} \quad (1)$$

This part includes the training and testing of the data .

III. Proposed Methodology

We suggest an web application in studying algorithm for estimating the rate of a specified stock by virtue of various algorithms mentioned below.
This section of the paper deals with the algorithms and the working of various types of classifiers. The basic input we need the algorithm is a pre processed raw text (from dataset). Here we are using three algorithms mainly named :

1. Linear Regression
2. Polynomial Regression
3. Ridge Regression
4. Lasso Regression

A. Linear Regression

Linear regression is conceivably one of the major acclaimed and prominent algorithm in census and machine learning.
Linear regression does endeavor predict tendency and future values.It is fundamentally, still not with precise definiteness. This is in regard to future prices in stock market or the economic marketing go. It certainly is significantly advantageous and considerably habituated apart from the economical markets too.
It can reveal you on moderate basis, what outmoded the previous tendency, and like quoted primitively, in case that the future tendency occur in the course of the time the synonymous just as previous tendency then your estimation will be sufficiently precise utilizing simple linear regression model which is nothing much of effort

Utilizing the percent error, a span perhaps be concluded and it perhaps be presumed the future value could reside in reach of the scope.
The linear regression model recur an equation that resolves the connection b/w the independent variables and the dependent variable.

The equation for linear regression perhaps authored as:

$$Y = B0 + B1 * x \quad (2)$$

In elevated proportions when we acquire multiple input (x), the line is called a plane or a hyper-plane. The characterization accordingly is the configuration of the equation and the distinct values accustomed for the coefficients (e.g. B0 and B1 in the above specimen).
A linear regression approach can carry out well for issues as like Big Mart trading where the self-reliant factors are effective for deciding the aim value.

Fig 2: Representation of the Linear Regression line

B. Polynomial Regression

Polynomial Regression is a regression methodology which sculpts the relation b/w a dependent variable (y) and an independent variable (x) as an nth degree polynomial. The following is the Polynomial Regression equation:

$$y= b0+b1x1+ b2x12+ b2x13+..... bnx1n \quad (3)$$

In ML, it is often known as the unique case of Multiple Linear Regression. Since we append a few more polynomial terms to the Multiple Linear Regression equation to make it Polynomial Regression.

When we apply a linear model to a linear dataset, we get such a good result, which we saw in Simple Linear Regression, although when we apply the same model even without changes to a non-linear dataset, we get a rapid outcome. As a result, the loss function will keep increasing, the error rate goes up, and the accuracy will actually reduce.

In such cases, under which data points are positioned non-linearly, the Polynomial Regression model is necessary. We can understand better it by comparing the linear dataset to the non-linear dataset as in illustration below.

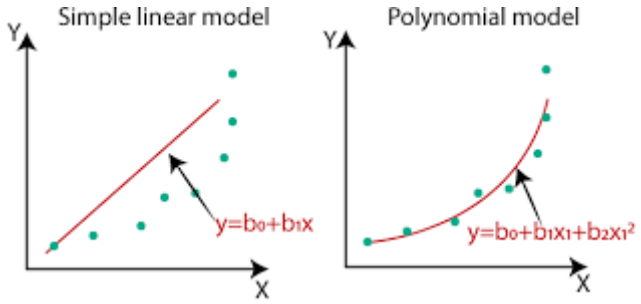

Fig 3 : Representation of the Polynomial Regression

C. Lasso Regression

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model makes use of deformation. Deformation is the process by which data values are narrowed towards a focal place known as the mean. The lasso methodology emphasizes basic, discrete models (i.e. models with fewer parameters). This kind of regression is better suited for models with high stages of multi co linearity or whenever you want to digitalize definite aspects of model evaluation, such as variable selection/parameter eradication.

Residual Sum of Squares + λ * (Sum of the absolute value of the magnitude of coefficients) (4)

I. λ reflects the quantity of deformation
II. λ = 0 denotes that all attributes are taken into account, and it is roughly comparable to linear regression, which uses just the residual sum of squares to establish a model.

The main objective of lasso regression is to find the statistical technique with the lowest standard error for a statistical regression analysis. The lasso accomplishes all that by enforcing a restriction on the model parameters, causing regression coefficients for certain variables to decrease further towards 0.

Lasso is indeed a regression analysis methodology that utilizes attribute selection and batch normalization to improve the predictability and understandability of the resultant mathematical approach.

D. Ridge Regression

Ridge regression is a prototype optimization method of analyzing data that has multiple collinearities. L2 function is performed by this method. When there is a problem with multicollinearity, least-squares are objective, and deviations are substantial, resulting in estimated parameters that are far from the absolute measurements.
Ridge Regression Example: Ridge regression could be used to evaluate prostate-specific antibodies and diagnostic steps in men who are about to possess one's prostates eliminated. Ridge regression performs well if a tiny proportion of genuine coefficients is minimal and perhaps even zero.

Ridge regression is a type of regression tool that uses L2 regularization. The other type of regularization, L1 regularization, reduces the impact of the coefficients by applying an L1 penalty equivalent to absolute number of the enormity of the coefficients. It could result in a complete eradication of some coefficients, resulting in sparse models. L2 regularization includes an L2 penalty equivalent to the square of the enormity of the coefficients. The same feature reduces the size of all coefficients (so none are eliminated). L2 regularization, unlike L1, does not produce lacking in detail models.

Ridge Regression is a linear regression continuation which includes a batch normalization penalty in the gradient descent all through learning.

VI. Experimentation and Results

The extracted data contains 5 attributes :

1. Date value: Date of the observation on share market
2. Open value: Opening Price of the share market
3. High value: Highest Price of the share market
4. Low value: Lowest Reached by the share market
5. Last Value: Final Value Reached by the share market
6. Close value: Closing price of the share market

We can get the accuracy of the predicted observation using the variance values. We can observe that the accuracy is proportional to the variance value (almost equal to 1). From this prediction we can conclude that applying the various regression models to the observation can increase the accuracy.

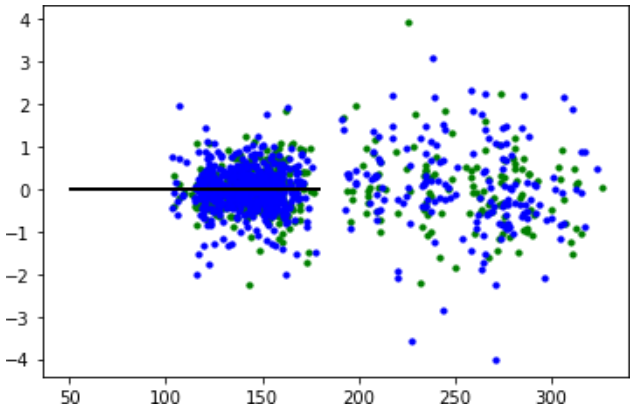

Fig 4: Graph showing the Lasso regression

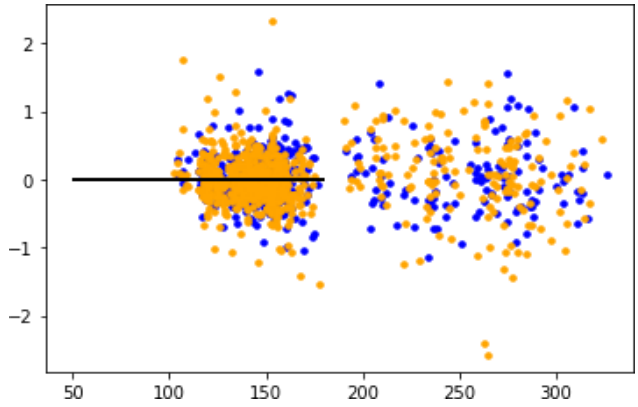

Fig 5: Graph showing the Ridge regression

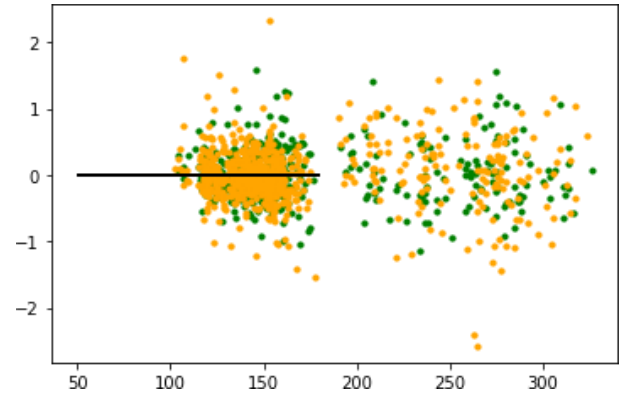

Fig 6: Graph showing the Linear regression

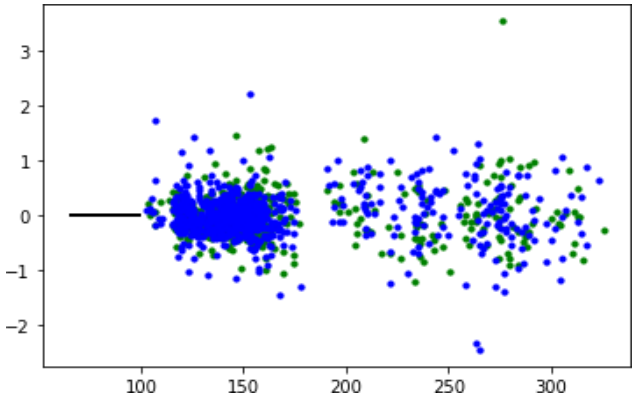

Fig 7: Graph showing the Polynomial regression

Data set : Tata company

We have studied the effectiveness of the Algorithms we have introduced in our Research by taking the dataset of the Tata Company which contains the Close, Low, High, Last, Total Turnover attributes with these we compared the variance value to get the accurate prediction algorithm.

The below table shows the comparison between the various algorithms used on basis of Variance value.

TABLE I Variations in Variance Values

Algorithm	Variance
Linear Regression	0.9999444971551508
Polynomial Regression	0.9999365423529375
Lasso Regression	0.9998683155736493
Ridge Regression	0.9999444972164427

V. CONCLUSION

The results of comparison among various algorithms shows that Ridge Regression has a better prediction accuracy than the other three algorithms we used

This model is based on the Various Regression algorithms using the data from kaggle. We can predict the closing values accurately within 5 min , 1 min , 10 min ahead using Ridge Regression, Linear Regression Polynomial Regression and Lasso Regression algorithms . we conclude that from the above data and results more the variance value the more is the accuracy.

So from the above information Ridge Regression Algorithm suits more good for prediction of share market

Among the four algorithms the accuracy from the Ridge Regression was more and that algorithm has given more accurate predictions.

For the Future work we will focus more on deep learning and multiple supervised learning techniques like regressions in order to achieve more accurate predictions on share market.

VI. ACKNOWLEDGMENT

We are immensely grateful to Mrs. R .Sumathi , Professor, Dept. of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Virudhnagar, Madhurai, Tamil Nadu, India for her constant guidance and unending support.

VII. REFERENCES

[1] International Workshop on Artificial Intelligence & Internet of Things (A2IoT) August 9-12, 2020, Leuven, Belgium Stock Marketprediction on High frequency data using Long-Short Term Memory Zineb Lanbouri , Said Achchab.

[2] International Conference on Computational Intelligence and Data Science (ICCIDS 2019) Stock Closing Price Prediction using Machine Learning Techniques Mehar Vijh, Deeksha Chandola,Vinay Anand Tikkiwal, Arun Kumar.

[3] Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. “On the difficulty of training recurrent neural networks.” International Conference on Machine Learning. 2013.

[4] Stock Price Prediction Using Long Short Term Memory Raghav Nandakumar, Uttamraj K R, Vishal R, Y V Lokeswar

[5] Parmar, Ishita, et al. "Stock market prediction using machinelearning." 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC). IEEE, 2018.