International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

-

STOCK MARKET PREDICTION AND EFFICIENCY ANALYSIS

Cherupally Shivani[1], G. Pragna[2], K. Rohitha[3]

Nadia Anjum[4]

Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana, India

Abstract: Stock is an unpredictable curve. Prediction in the stock market is covered with complexity and instability. The main aim for the persuasion of the topic is to predict the stability in the future market stocks. Many researchers have performed their research on the movement of future market evolution. In the recent trend of Stock Market Prediction Technologies machine learning has integrated itself in the picture for deployment and prediction of training sets and data models. Machine Learning employs different predictive models and algorithms to predict and automate things of requirement. We are using the Nifty 50 dataset for our project. It consists of six attributes and 4000 records spanning from the year 2004 till 2020. For the implementation of our project we are utilizing the machine learning techniques such as Linear Regression, Polynomial Regression, Knn algorithm and Decision tree. Performance of the algorithms will be measured using precision accuracy.

Keywords: Prediction, precision, accuracy, performance, linear regression, polynomial regression, decision tree, algorithm.

1. Introduction

1.1 About Project

We all have heard the word stock one way or the other. Particularly stock is related with the associates and companies which are commercialized and are settling in the world of marketization. The other word used for stock is share which is prominently used in day to day life. People even term it as an investment plan and it's something people see as a long term investment that secures and provides abundant funds during the retirement age.

Buying a company stock is purchasing a small share of it. People invest on the same to get a long term benefit which they think is less valuable for now but has the potential to grow with the time. It's an investment that provides the long term run and deals with long term goals with fair objectives. The value of the share you invest today has to give you a yield of best tomorrow but it's not the same.

-

Market is unpredictable so are the resources and the factors that are taken to drive it off or on on the set. It's never been on the same level and the pattern of the same is still unpredictable till the time. Some closeness and prediction methods have been derived and approximate values and the rough figures are generated hoping for the best but all of the resources can't be trusted and are still unpredictable in nature.

Knowing the market situation and researching the same is the best way to find the reliability for which there are many agents who have taken the same as a profession and are making a fortune out of it. They predict and advise but the advisory cost and the charge is higher and the stock evaluation is never less the same.

Market is changing at an instantaneous rate even in a day. There are many highs and lows in the market and having said the resources and the timing of the external and internal agent. Stock is a fascinating resource to start with.

Stock in other terms is defined as the fair share or the ownership representation explaining the security measures and the agreement between two parties which are an individual and the company. Stock is there from the start and due to its tendency of uncertainty it has been a word of fancy. People researching the same and implementing on a daily basis had made a fortune out of it. There are various agents available in the market for making you understand and invest on the same and the charges of the same are hectic and insanely expensive.

The main resources for the company is the fund to carry out the daily work and create a profit out of it. In time of need for an higher budget estimation and to overgrow from the resources they need the finance and undergoing a finance loan for approval, passing and having one is hectic and the banks are vultures for which the interest rate is higher than the other form of investment hence limiting the margin of the product.

Stock is another way for companies to collect revenue and boost up the production for the upper yield and to gain the most out of the business plan for the bigger picture. This is found to be an effective way to invest and grow in the commercial field and a better alternative to tackle the financial crisis during the requirement.

For an investor its a risk phenomenon where they invest their savings and hope it brings back the return in higher yield. If the evaluation of the same increases then the stock evaluation and its price increases causing the financial gain to both the parties. In Indian Society it is even considered as a side point business and people believe it as a hand of luck.

When an individual purchases a company stock then they're referred as a shareholder and they will get a share out of the same as they have invested in their profit or the gain. An investor can sell and buy the stock as per their needs. They can share their stock to their respective or the other individuals whereas there are many stock brokers available out in the firm playing with the same.

-

1.2 Objectives of the Project

A stock in general terms is holding a particular company's share makes you a shareholder.

Stock Price Prediction using machine learning helps you discover the future value of company stock and other financial assets traded on an exchange. The entire idea of predicting stock prices is to gain significant profits. Predicting how the stock market will perform is a hard task to do.

Stock market prediction is a prediction system software that illuminates the risk that undergoes during the investment in the stock market. It predicts the stock rates and its rate of exchange acknowledging the basic understanding and the statistical analysis in front of users.

Data is considered as the digital fuel that gives the possibilities of higher yearn and gives the upcoming terms. Knowledge is power and the same holds correct with the stock. Stock is unpredictable and over-changing its dynamic in nature. The rise and fall of the same is uneven and can't be classified so easily. Dependencies of the same deal with flexible resources and the agents behind it.

1.3 Scope of the Project

Analysis of stocks using Machine Learning will be useful for new investors to invest in the stock market based on the various factors considered by the software.

Stock market includes daily activities like sensex calculation, exchange of shares. The exchange provides an efficient and transparent market for trading in equity, debt instruments and derivatives.Our software will be analyzing sensex based on the company's stock value. The stock values of company depend on many factors, some of them are:

Demand and Supply: Demand and Supply of shares of a company is a major reason for price change in stocks. When Demand Increases and Supply is less, price rises. and vice versa.

Corporate results: This will be regarding the profits or progress of the company over a span of time say 3 months.

Popularity: Main Strength in hands of share buyer. Popularity of a company can affect buyers. Like if any good news of a company, may result in rise of stock price. And bad news may break dreams.

The stock value depends on other factors as well, but we are taking into consideration only these main factors

2. Literature Survey

-

2.1 Existing System

As many have invested their time and effort in this world trade for getting it closer and more reliable to the people for carrying out the resources and making their lifestyle more deliberate than the previous. In the past few years various strategies and the plans had been derived and deployed ever since its continuation and the topic is still a point of research where people are coming up with ideas to solve.

Intelligence fascinates mankind and having one in a machine and integrating on the same is the hotkey of research. There are various people contributing to the same research. ASheta tried its invention on two nonlinear processes and came up with TS which is used as a model for fuzzy sets.

All the learning systems from the past are limited and are simplest in nature where learning of the simple algorithm for a computational mean is not enough which can even be done by the human brain itself. The main motto of learning was limitized and the learning model was not efficient.

The existing models can't cope up with the vulnerabilities and remove the rarest information that they can't process causing it a major data loss which creates a problem in forecasting.

Observation is the integral part in the resource and prediction management. If the outcome can't be observed it's point of time estimation is compromised causing it less liable in the market. Monitoring of the same is not possible in the existing system.

The existing system in stock market predictions is apparently biased because it considers a only source point for data source. Before the prediction of the data set a simple data retrieval should be generated and tested on the training data set which are more flexible and versatile in nature.

Loss of sights is a major problem in the existing system as the stock varies each day and the loss margin can be higher with respect to time. An initial instance is taken for prediction.

2.2 Proposed System

Stock is unpredicted and liberal in nature. The following is impressive and reluctant in nature. Finding the predictability and getting the nearest is the best hit goal for the same. The exact and accurate estimation of the same is never-less possible.

3. Proposed Architecture

There are various constraints that in-fluctuate the pricing and the rate of stock. Those constraints had to be taken in consideration before jumping to the conclusion and report derivation.

-

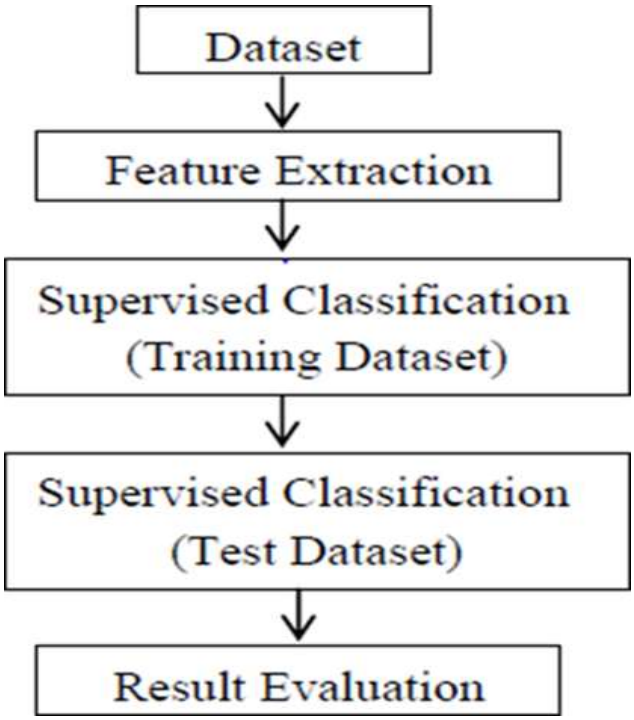

Fig- System flow

Here as described in the figure above, the proposed system will have an input from the dataset which will be extracted featured wise and Classified underneath. The classification technique used is supervised and the various techniques of machine level algorithms are implemented on the same.

Training Dataset are created for training the machine and the test cases are derived and implemented to carry out the visualization and the plotting. The results generated are passed and visualized in the graphical form.

4. Implementation

4.1 Algorithm

4.1.1 LINEAR REGRESSION:

One of the well known algorithms used in machine learning is linear regression. It is covered under both statistical as well as in machine learning. It is used for analyzing the dependency between two variables one is known dependency whose value is known and the other is unknown. The value of the unknown

-

dependency is checked with the known dependencies and the result is found and derived on its basis.

The dependency of the variable chances and are categorized into two types. Positive Linear Regression is the regression flow when both the dependencies show the growth rate and both are totally dependent and supportive with the changes flow. Negative Regression is the regression flow where one dependency cancels the growth of the other. If one dependency shows the tendency to grow whereas the other one is decreasing then this graph flow comes in picture.

They are Single Linear Regression (SLR), it's the fundamental block of linear regression. It assumes that the two dependencies are linearly aligned and changing the values on the same will effect the other equally.

Multi Linear Regression is an extension of the SLR algorithm where different fundamentals are considered with regards to the dependencies. It even deals with residual errors.

4.1.2 POLYNOMIAL REGRESSION:

It is a form of Non-Linear Regression. In this form of regression the two constraints one having known dependencies whereas the other part is unknown and is generalized with the help of n^{th} polynomial value. Research is a wide level of scope that one's involved in. There are various curves and lines estimation which can't be normally fixed and plotted with the limitation of linear regression. If trying to do so there will be a higher error ratio which will bring down the integrity and the reliability of the system in itself.

Thus to overcome this barrier and to represent the most of the curve in every way possible either it be a straight line, or hysteresis curves. This regression helps to analyze the curve in every possible format and helps to reduce the redundancy and the trial points of errors alongside with optimized cost factor which is a great boost to the algorithm itself.

It is widely used for the complexity to solve and takes the particular values which are unique in nature and peculiar values that needs to be considered before to set the outcome. The natural uses of the same is found in epidemics growth and to see the growth ratio of the tissue.

Including all the values of peculiarity increases the effectiveness and the efficiency of the same so it is more reliable than a Linear Regression. It has a wide range of coverage so no distortion of information. No data is lost during the processing and cleansing of the dataset.

A prediction model is generated from the high dependencies set that increases the expectancy and gets ones closer to the proximate values. It provides the best representation of constraint dependencies with one another making it easy for the user to understand and see the conversion of the same.

4.1.3 K-NEAREST NEIGHBORS (KNN):

One of the Machine Learning Algorithms which is classified both under regression and classification . This is a supervised learning module. It's an essential module in Machine Learning. It is commonly used in the Data Mining process.

This machine learning algorithm is used to solve the regression and classification of the datasets along

-

with its highly demanded on pattern machine as well as detection of intrusion. As the name suggests it deals with the neighboring dataset closer datasets are assumed as a proximity.

Similarity of the dataset with reference to data modules, distance and vector modules are calculated and plotted on the same. The nearest point is calculated among the given dataset which is defined by a constant 'k' which can be an integer value. Individual distance between the data is plotted and calculated. Euclidean is the most appropriately used for the same.

Distance values are aligned and are sorted in ascending form. The closest distance index 'k' is selected and the array is sorted with the same index. Here the dataset deals with the wide range of values and the proximity of the same, The datasets are widely categorized and are distributed in nature. The distribution of the same makes it more feasible. It deals with the closeness of data.

Every division is divided into chunks of small dataset that finds the closeness proximity and derives the result on the same. It's a basic algorithm and the working of the same is easily understandable. During the starting it doesn't assume anything with regard to the dataset hence known as non-linear datasets.

It's a feasible and versatile algorithm which can both be used for classification as well as regression of the data sets. The best is the yield factor which gives a positive result set and is highly accurate and found efficient.

4.1.4 DECISION TREE:

In general, Decision tree analysis is a predictive modeling tool that can be applied across many areas. Decision trees can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions. Decision trees are the most powerful algorithms that fall under the category of supervised algorithms.

They can be used for both classification and regression tasks. The two main entities of a tree are decision nodes, where the data is split and leaves, where we got outcome. The example of a binary tree for predicting whether a person is fit or unfit provides various information like age, eating habits and exercise habits. We have the following two types of decision trees. Classification decision trees: In this kind of decision trees, the decision variable is categorical. The above decision tree is an example of a classification decision tree. Regression decision trees : In this kind of decision trees, the decision variable is continuous. Gini Index: It is the name of the cost function that is used to evaluate the binary splits in the dataset and works with the categorical target variable "Success" or "Failure".

Split Creation: A split is basically including an attribute in the dataset and a value. We can create a split in the dataset with the help of the following three parts. **Calculating Gini Score:** We have just discussed this part in the previous section.

Splitting a dataset: It may be defined as separating a dataset into two lists of rows having an index of an attribute and a split value of that attribute. After getting the two groups right and left, from the dataset, we can calculate the value of split by using the Gini score calculated in the first part. Split value will decide in which group the attribute will reside.

Evaluating all splits: Next part after finding Gini score and splitting dataset is the evaluation of all splits.

-

For this purpose, first, we must check every value associated with each attribute as a candidate split. Then we need to find the best possible split by evaluating the cost of the split. The best split will be used as a node in the decision tree.

4.1.5 LONG SHORT TERM MEMORY(LSTM):

Sequence prediction problems have been around for a long time. They are considered as one of the hardest problems to solve in the data science industry. These include a wide range of problems; from predicting sales to finding patterns in stock markets' data, from understanding movie plots to recognizing your way of speech, from language translations to predicting your next word on your iPhone's keyboard.

With the recent breakthroughs that have been happening in data science, it is found that for almost all of these sequence prediction problems, Long short Term Memory networks, LSTMs have been observed as the most effective solution.

LSTMs have an edge over conventional feed-forward neural networks and RNN in many ways. This is because of their property of selectively remembering patterns for long durations of time. The purpose of this article is to explain LSTM and enable us to use it in real life problems.

LSTMs on the other hand, make small modifications to the information by multiplications and additions. With LSTMs, the information flows through a mechanism known as cell states. This way, LSTMs can selectively remember or forget things. The information at a particular cell state has three different dependencies. Industries use them to move products around for different processes. LSTMs use this mechanism to move information around.

We may have some addition, modification or removal of information as it flows through the different layers, just like a product may be molded, painted or packed while it is on a conveyor belt.

4.2 Code Implementation

JUPYTER NOTEBOOK

Jupyter Notebook is a web-based interactive computational environment for creating Jupyter notebook documents. The "notebook" term can colloquially make reference to many different entities, mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context. A Jupyter Notebook document is a JSON document, following a versioned schema, containing an ordered list of input/output

-
cells which can contain code, text (using Markdown), mathematics, plots and rich media, usually ending with the ".ipynb" extension.

S.K LEARN

Scikit-learn (also known as **sklearn**) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

PYTHON

Python is an interpreted high-level general-purpose programming language. Python's
design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

5. Result

Precision: Precision is one indicator of a machine learning model's performance – the quality of a positive prediction made by the model. Precision refers to the number of true positives divided by the total number of positive predictions (i.e., the number of true positives plus the number of false positives).
Accuracy: Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.
Accuracy = Number of correct predictions Total number of predictions.
Recall: Recall literally is how many of the true positives were recalled (found), i.e. how many of the correct hits were also found. Precision (your formula is incorrect) is how many of the returned hits were true positive i.e. how many of the found were correct hits.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

-

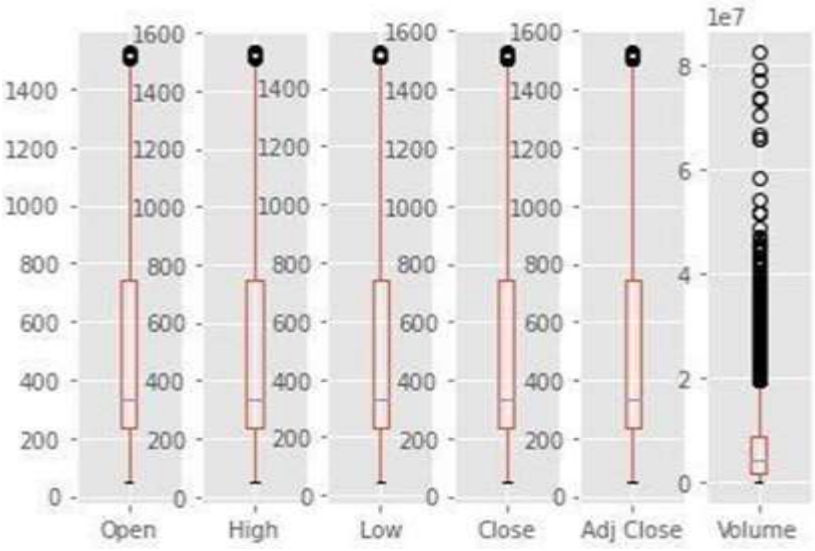

Fig. 5.1 Data Extraction and plot

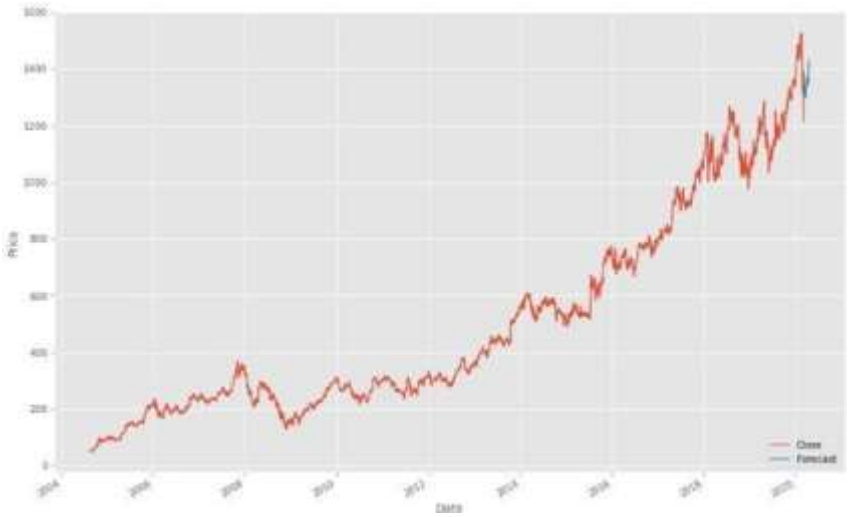

Fig. 5.2 Linear Regression

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

-

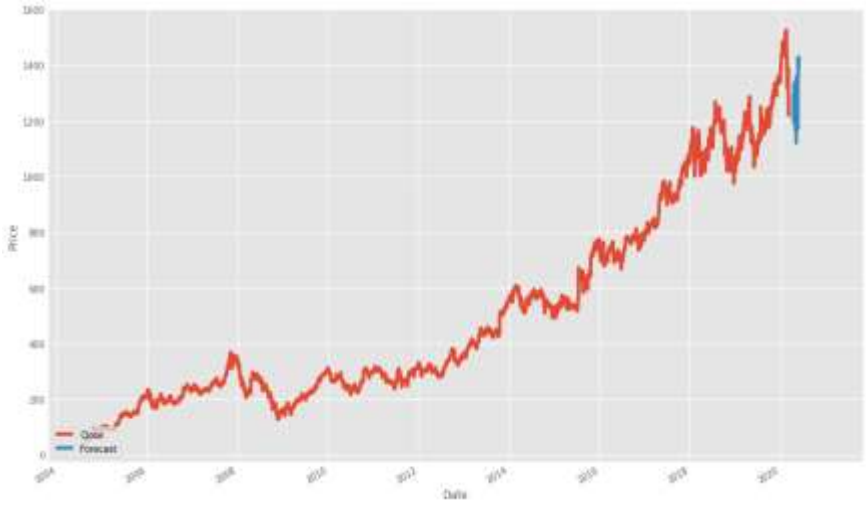

Fig. 5.3 Decision Tree

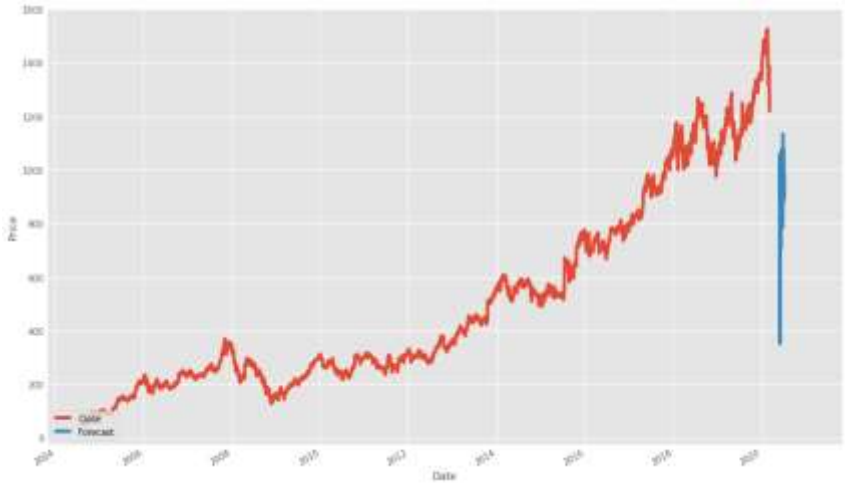

Fig. 5.4 KNN

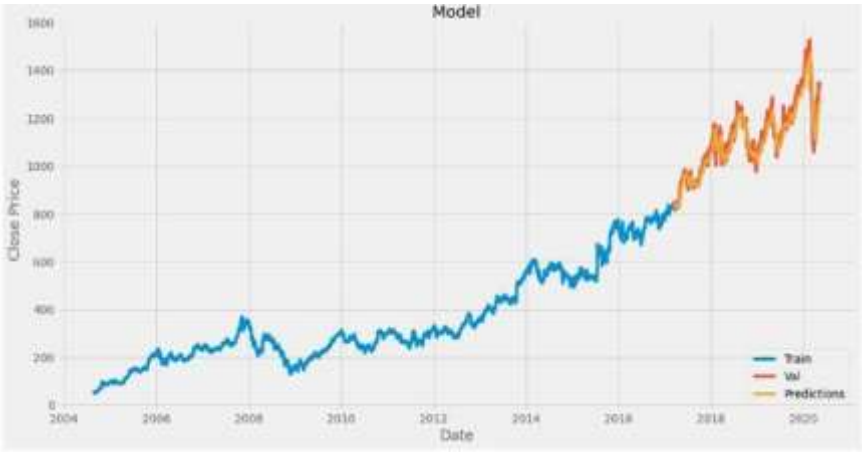

Fig 5.5 LSTM

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

-

Fig 5.6 CLOSE PRICE HISTORY

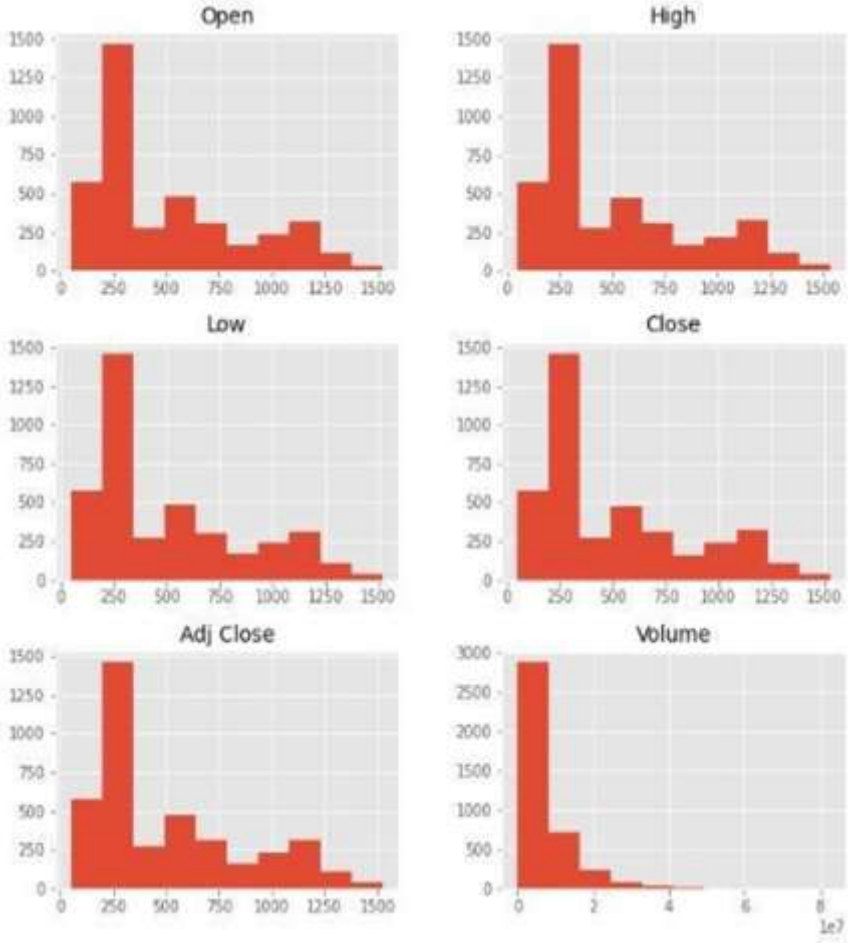

Fig 5.7 Sample data

-

6. Conclusion

To conclude stock is an unpredictable mechanism which follows the segments of the chain and the dependencies of the same are unpredictable. It is defined to be a curve which keeps on changing and turning the price from low to high and vice-versa.

As the integration of the same is higher with other dependencies so leaving one dependency compromises the level of accuracy. Accuracy is not the term used over in stock as the actual prediction is not possible for any fiscal days it keeps on changing and turning the tables day and night. Having higher component assets and the dependencies makes it more feasible and flexible in nature causing it even harder to predict. The approx value is taken into consideration and the hit or profit or the gain rate is calculated for the same.

In the project various high level machine learning algorithms are implemented and integrated and the output is generated from the same making a user visible with the outputs in the form of graph which makes it easier for them to see and interpret what's the scenario and they can decide on the same to invest and get the benefit out of it,

The proposed software takes the raw set of data from the dataset or the .csv file and processes it. The cleaning and cleansing of data is done and then further processed to gain the effective outcomes. After the computational mean the output is displayed on the screen in the form of a graph .

7. Future Scope

Stock Markets are the best alternative for business to grow and it's a sideway income for the individuals who are ready to invest and earn from the same. The term stock has been in picture ever since and it's growing in bulk everyday. There are thousands of investors investing on the same and making the fortune out of it.There are middle level agents and stock vendors who learn and invest on the same. The cost for the consultation on the stock is bulky and expensive. So when it comes to people they think a lot and invest and there's no chance and certainty for the same to produce a yieldful result.So stock being unpredictable and the tendency of its growth is higher than ever. If the stock market and its prediction can be done accurately then it's going to be a gain for both the individuals and the organization. The risk factor has to be mitigated so the efficiency of the system should be high and people can be certain about their investment in time. The project can be further continued to gain the effectiveness of the prediction with additional implementations of the content that can involve real time scenarios and the way of executing and processing the real time scenario. Various constraints have to be added and performance of the same can be calculated in the future time for the effective results. The expected form of the display is a graph , whereas from the same the appearance and setting of the display can be integrated and a pie-chart and a custom graph can further be implemented on the same.

-

1. *8. References*

2. Bonde, Ganesh, and Rasheed Khaled. “Extracting the best features for predicting stock prices using machine learn- ing.” Proceedings on the International Conference on Artificial Intelligence (ICAI). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.
3. Hagenau, Michael, Michael Liebmann, Markus Hedwig, and Dirk Neumann. "Automated news reading: Stock price prediction based on financial news using context- specific features." InSystem Science (HICSS), 2012 45th Hawaii International Conference on, pp. 1040-1049. IEEE, 2012.
4. Kyoung-jae Kim, Ingoo Han. “Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index”. Expert Systems with Applications, Volume 19, Issue 2, 2000, Pages 125-132, ISSN 0957-4174.
5. Leung, Carson Kai-Sang, Richard Kyle MacKinnon, and Yang Wang. “A machine learning approach for stock price prediction.” Proceedings of the 18th International Database Engineering & Applications Symposium. ACM, 2014.
6. Md. Rafiul Hassan and Baikunth Nath, “Stock Market forecasting using Hidden Markov Model: A New Approach”, Proceedings of the 2005 5th International conference on intelligent Systems Design and Application 0-7695-2286-06/05, IEEE 2005.
7. P. Hajek, Forecasting Stock Market Trend using Prototype Generation Classifiers, WSEAS Transactions on Systems, Vol.11, No. 12, pp. 671-80, 2012.
8. Tiffany Hui-Kuang Yu and Kun-Huang Huarng, “A Neural network-based fuzzy time series model to improve forecasting”, Elsevier, 2010, pp: 3366-3372.
9. Kishor Kumar Reddy C and Vijaya Babu B, “ISPM: Improved Snow Prediction Model to Nowcast the Presence of Snow/No-Snow”, International Review on Computers and Software, 2015.
10. (http://www.praiseworthyprize.org/jsm/index.php?journal=irecos&page=article&op=view&path%5B%5D=17055)

-

11. Kishor Kumar Reddy C, Rupa C H and Vijaya Babu B, “SLGAS: Supervised Learning using Gain Ratio as Attribute Selection Measure to Nowcast Snow/No-Snow”, International Review on Computers and Software, 2015.
12. (http://www.praiseworthyprize.org/jsm/index.php?journal=irecos&page=article&op=view&path%5B%5D=16706)
13. Kishor Kumar Reddy C, Vijaya Babu B, Rupa C H, “SLEAS: Supervised Learning using Entropy as Attribute Selection Measure”, International Journal of Engineering and Technology, 2014.
14. (http://www.enggjournals.com/ijet/docs/IJET14-06-05-210.pdf)
15. Kishor Kumar Reddy C, Rupa C H and Vijaya Babu B, “A Pragmatic Methodology to Predict the Presence of Snow/No-Snow using Supervised Learning Methodologies”, International Journal of Applied Engineering Research, 2014.
16. (http://www.ripublication.com/Volume/ijaerv9n21.htm)
17. Kishor Kumar Reddy C, Rupa C H and Vijaya Babu, “SPM: A Fast and Scalable Model for Predicting Snow/No-Snow”, World Applied Sciences Journal, 2014