DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

A

PROJECT REPORT

ON

"STOCK MARKET PREDICTION USING MACHINE LEARNING"

BY

CHETAN SAH - RA1911026010035

ABSTRACT

Stock is an unpredictable curve. Prediction in stock market is covered with the 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. Stock consists of fluctuating data which makes data as an integral source of efficiency. Impact on the same chances the efficiency of the prediction. 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. The Paper focuses on the use of Regression and LSTM to predict stock values.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 INFORMATION ON STOCK

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 to 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 is as an investment plan and its something people see as a long term investment that secures and provides an 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 value for now but has to potential to grow with the time. Its an investment that provides the long time run and deals with long time goals with the fair objectives. The value of share you invest today has to give you an yield of best tomorrow but its 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. Its never been on the same level and the pattern of the same is still unpredictable till the time. Some closeness and prediction method had been derived and approximates values and the rough figures are generated hoping for the best but all of the resource can't be trusted and are still unpredictable in nature.

Knowing the market situation and researching on 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 in an instantaneous rate even in a day there are many highs and lows in the market and having said the resources and the timing the external and internal agent. Stock is a fascinating resource to start with.

Stock in other term 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 on the same and implementing on the daily basis had made a fortune out of it. There are various agents available in 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 an other way for company 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 pictures. 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 saving 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 consider 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. A investor can sell and buy the stock as per their needs. They can share their stock to their respective or the other individuals where as there are many stock brokers available out in the firm playing with the same.

1.2 PROBLEM DEFINITION

Stock is an unpredictable curve that had been in picture ever since. Its essence had been ever long living and indulging. It had grown its popularity with respect to time. People are more fascinating and interested on the same then before times. Same for the case for the organization. Organization had created it as a better source of revenue generation rather than investing and taking a loan approval from the bank It's way efficient and less hectic from the firm point of view.

Stock is unpredictable and its been the same from the start. Its way of escalating and deescalating had been phenomenon and experiencing the same is the best integral part of it. It has its upper hand and flexibility with the changes that has the chances of uprising as well as crashing the whole market. Its easily defined in few words but making an essence and understanding the same is way more hectic and time consuming.

Simpler it sound complex are its phenomenon and integrating the same. Its has its whole different sets of dependencies and integration from different agents which fluctuate the same in the market. Finding an accurate and getting the exact values out of the same is still unaligned and no particular model of the same is seen in the market value.

Finding the closest and getting an accurate proximate value out of such an unpredictability is a problem in itself. Merging of the data getting the best prediction to increase the efficiency alongside considering the different expects of the moderator is tough and we took the same in consideration and implemented with every aspect to generate the best out of the same and get a result that can be better interrupted and the efficiency remains the same with the value of different aspects of creating an impact of reducing the risk and influencing the same over the time period to gain the most out of it.

This is totally based on Machine Learning Algorithm to proceed and provide an effective result. Getting the data and processing it and generating a forecast for three days is the problem statement that we worked on.

1.3 PROJECT PURPOSE

Stock market prediction is a prediction system software that illuminate the risk that undergoes during the investment in 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 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 deals with flexible resources and the agents behind it.

Investment during a fiscal day determines the opening stock market for the next day. It has its dependencies and is total integration with the level of finances and revenue generation. The stock is tremendous and hectic in nature. The main theme of the project is to predict the turning curves and bring the predictability method and undergo the process and algorithms to conclude to a viable resource source.

Everything flows a pattern. Pattern is the way of derivation and so holds true for the stock too. Stock in day to day life follows a pattern movement. Increase in some resource can increase the price of some whereas decrease the price rate for the others, The source and the outcome are derived on the polarity basis which can either be positive, neutral or an negative flow. Correlation of the given polarity is determined and an effective source and reliability is established.

This project helps in bridging the resources and empowering the people to know and trade the most out of stock and understand the generation and the vulnerabilities that has to be seen and predicted. The enhancement of the same is done with the resource graph which makes an user or the customer to analyses the same and take the needs and important details before dealing and consider those things for the yield that the person is willing to invest on. Forecasting of the stock prediction is done by the available data source and the prediction is done for the upcoming week. The predictability itself is a challenge and that's the main purpose of the report.

1.4 PROJECT FEATURES

Features deals with the flexibilities and the top marks that one can present. The project was headed with the resource available and the most that the company demands and that is finance. Taking about finance and learning on the same gave an idea on the fiscal and stocks. So the featuring of the idea came with handling and automating the resource which other agents are making fortune out of it.

Knowledge is a bliss and learning is the curiosity where as outcome is the expectation so the resource deals with the importation and extraction of multiple machine learning algorithms to learn, process and yield the result to derive and conclude a possible outcome set that is effective and generative in nature.

There are various model that outflows in market which are trying it's best on creating a resource and give the predictability to most of it accurate but everything is not the same and the conclusion of the same are not ideal. The efficiency varies as the variation in the stock market and it's prediction.

The project was purposed with the intent sole to make and undergo the following way of computing. The first deals with the data extraction that is done with clearing of data and its chunks from the database or the dataset. The second flow is the training from the source training is done and classified. During the same supervision is done and the last part is the generation of the yield which provides the result after computation of the same.

Salient features included are the Visualization and the prediction that gives a boost. Uses of different forecasting algorithm to forecast that holds true and are suffice in nature to yield to the positive resource source. Diving and initializing the expects that needs to be considered. Mitigating the risk factors to bridge and uplift the investment.

Analyzing and utilizing the same to support the live environment. Keep a track of progressive result and it's evaluation on day to day basis to find the flows and the level of integration. Automating for the ideas and making it most by using feasible algorithms which can undergo learning and implement the updates in itself to summon the efforts that one needs to take for the best.

1.5 MODULES DESCRIPTION

1.5.1 DATA SET

This is the fundamental module before starting of the project. The dataset is a group of data that are mended together to show the data variations in a time span to undergo further estimation and the source of the resources and its outcome for the later time of evaluation. It generates the result optimization and gives a feasible time period to customize and get the flow to the derivation.

This increases and are used in the level of research and finding the best suitable resource out of the same the resources has to be finely estimated and derived for the best possible outcome and the finest the value become the better is the level of extraction and closure is the best yield values that needs to be considered.

1.5.2 DATA ABSTRACTION

Abstraction is the finding of the resource to its best to categorized the above dataset and learning the best out of it. Abstraction of the data is the integral part to the flow. All the data are a huge set of chunks which on processing can limitize the yield result and the computational mean too. Thus with the available resources the data yield had to be derivative.

Abstraction of the dataset is to customize the data set and finding the best suitable constraints to take into consideration and the unwanted resources are the dump which will be dumped and the supreme cluster is created with the valuable constrains and a pattern is needed to be derived from the same.

Data are cleared on this level for the beginning of the process. The valuable data are the set that brings the value to the data set for a better understanding and giving a better yield and production by evaluating the same.

This is a feature abstraction module to extract the featuring of the dataset. This is a feature model process where all the feasible resources are categorized and the same will be in use for the featuring.

1.5.3 TRAINING DATASET

After the abstraction of the data and clustering of the same. The machine had to be trained for which the training data plays the important role. There are thousands of machine learning algorithms that are into place and evolving with the same. The best to the practice of machine learning is to yield the result and the content to derive what's needed with the time frame.

This is a supervised learning form where the input are passed so that the system learns from the same. Various variants of inputs are passed which were stored in the dataset. Every resource is considered and taken into consideration. After considering the whole set of information and the resource the machine tries to learn from the passed dataset. The dataset has to be wide and versatile. After considering the learning it tries to integrate with the same type and flow like the same as the human mind and creates a pattern and the links between the same.

1.5.4 TEST DATASET

These are the sets of data that gives the result after learning from the data. This is the test generation with the output result. Results are generated in each phase of testing. This is also termed as the testing phase. Now a new set of datasets are passed which are deliberately like the training dataset and the efficiency of the same is calculated.

Over-Fitting of the dataset. Validation of the same with the effective constraints and hyper parameters are checked. This phase is training and the output is evaluated with the set of training. After each process of computation the set of data are trained and efficiency of the same is measured and is evaluated with the others.

Various batches of the test is implemented to get to the level of accuracy and derive result to fetch and yield for the best performance and to be true to the effectiveness of the data which is not biased with any constrains available. This determines the efficiency of the system which is must for the predictions.

1.5.5 RESULT EVALUATION

This is the main part for any implementation of the project. Evaluation of the key point to the success. All the categorization of the work and the best to know the resource fundaments and again establishing the same to check the validity and the work flow and check on the output is must. The evaluation, utilization and implementation undergoes a various level of extraction and evaluation.

The main theme is to provide and come up with the output with an accuracy that can be used and implemented. From the starting to the final the process is categorized, supervised and efficiency is check and the working is undergone. Testing is tested and it's evaluation are mended.

The process undergoes the same for various time and phase. Testing of the same undergoes sequential iteration for many more to meet up to the constituency. The remarks are to be noted and further work is done on the same with the implementation of the different aligned resources that are integrated with the available resources and its outcome.

After the evaluation and customization of the same the result is to be potted in a visible form and the best form of visibility is the graph. The Graph visualization is the best way of visualization that keeps the audience engaged for a long time. Derivation of the outcome is easily accessible and interpreted and the flow diagram is shown with the stock prediction that gives an upper hold to the appearance and shows the best level of the content.

After establishing a graph connectivity the customer or the user takes time to process the data and take that picture into consideration and can avail for the upcoming stock by investing in the same.

CHAPTER 2 LITERATURE SURVEY

One of the integral part to maintain the consistency is the literature survey. It's the crucial steps to be followed in the development process. The Software Development needs authenticity of the resources and the availability of the same. This part helps in discovering the content that been worked on and find the utilization and the implementation of the same in today's time. The key factor to the development is the economy and the strength of the product. Once the innovation of the same undergoes through the building phase the support and the resource flow is to be monitored and computed. This is also known as the Research phase where all the research is embedded and done to carry the flow.

2.1 MACHINE LEARNING

One of the finest word heard in today time is Machine Learning. Either it be at work or different places the machine learning has been an integral part of todays technology. Though its evolutionalizing and developing in a rapid rate and development and deployment of the same is still in progress. The machine learning itself had brought a random changes in today worlds because of which automation is in frame which was a mare existence in the past.

It's an aspiring term in todays time. One of the move that all the firm are interested into. It's a leading pillar for tomorrow leading the world to a better future of evolution where the customization and labor work can be reduce to half and the safety of the survival can be with held to stand tall for the better utilization of human mind. Keeping that in picture it's been a hazard to many more in terms of irrespective field of interest. Since Machine is considered most efficient and the level of mistakes are kept at the minimum the level of work flow can be a work of hazard and further improvement on the same may create a thousands sitting idle in home creating a larger impact on unemployment and livelihood. Which in other way is a threat to the society too.

ML is the abbreviation for Machine Learning. In other word it is making a human mind fitting inside a machine which uses the same to perform the task of thousands. Machine Learning deals with the higher aspects of learning techniques and algorithm which are highly aligned to make the work flow seamlessly effortless with the human tendency of doing work.

Algorithm of such are improvising in nature which learns by themselves and fit themselves in the world of impairment by getting the required data and adjusting with the same giving the effective results out of the same. ML is a subsidiary or the subset of an AI(Artificial Intelligence). It is a mathematical model where computation of the testcases plays the major role in driving of the results.

A wide level of machine learning architecture are implemented today to turn on the yield factor and make people life more efficient in terms of livelihood. Various use of such in Message Filtering like spams, Trash automation are automated and carried out by the same. Since the efficiency is way more than a human tendency. Multi-tasking and processing is also initiated by the same giving a dual output which a human can never ever possibly be able to.

Statistics is the major key role in driving the machine learning in figure. It deals with computation of statistics in a wide range view and processing the same to give an data driven output causing it more sensible and resourcesable. Not only to the same it optimizes the resources and the efficiency is unbitable and reliable in terms of any means.

Though its being evolutional in nature but it has integrated itself well with the terms of computational and digitalization. Various computational fields like Data Mining, Statistical Analysis, Optimization of resources, Automation are a major part of it. Here the machine has the capacity to process the result on its own as same as the human bring. This process can be initiator as well as the derivable. The statistical flow is mainly reasonable with data driven pattern even the unstructured or the semi-structured data can be processed and approximate answer to the same can be derived. All the equations

are derived and the closest value to it's aligned field is found and the proximity is determined.

The classification of the same can be listed as follows:

2.1.1 SUPERVISED LEARNING

Supervised Learning deals with the supervision of the machine to derive the necessity input required. It's a mathematical model where the inputs and output of the same is already known and its passed to the machine to get the expected output so that the efficiency is determined and this is the learning phase for the machine. Here the feeding and derivation of the same is measured.

Here the machines filters the inputs learns from the functional unit. Compute it and stores it into its memory for further process and if found a matching pattern it uses the same and learns from it and plot a result out of the same.

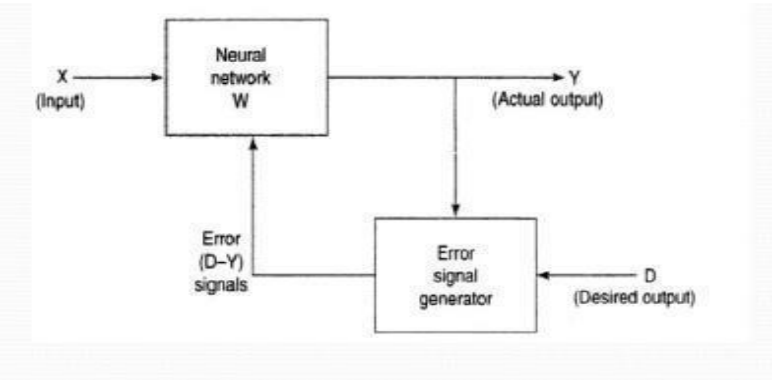

Fig 2.1: Supervised learning

This is a dependent process. The machine totally depends on the user who has to feed the inputs and has to check the efficiency of the same and correct it with the flow of iteration. It's an ANN network. During the training phase vectors are taken into consideration.

Up in the above figure There's an input vector and the output vector. The input vector derives and gives an output flow of the output vector. If the error signal is generated then the iteration is undergone where as lacking of the same means the

output field is derived and the output result is accurate and no modification needs to be undergone for same.

2.1.2 UNSUPERVISED LEARNING

Unsupervised learning deals with learning by itself. It is also known as self learning algorithm. Here only the input vector is known and passed. So the variance of the result deals with the input factors. Here the input factors are grouped and clustered. Cluster is the main essence of this technique.

Test Data are passed and with the iteration of the same it learns from it derives itself more closer to the conclusion part. Labelled is missed in the data set and classification and categorization of the same had to be done my the machine itself. Cluster and Communalization is the main essence of it.

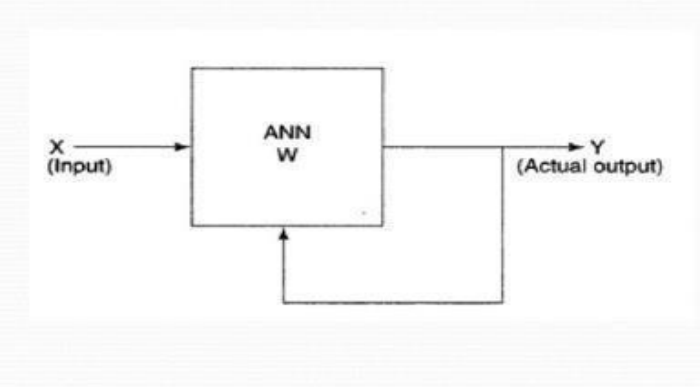

Fig 2.2: Unsupervised Learning

As described in the above figure, In this ANN network when the input is processed by the function the output had to be self derived and to be matched with the cluster set to provide the result. If the result lacks the interpretation then it undergoes the iteration. All the data sets are formed and combined in a cluster set for the effective uses of the same in further cases.

Feedbacks are not reciprocated incase of such it responds to commonalities. If the commonalities is found between the dataset then it applies the previous functionalities and derive the data. If not set then it learns and identifies for the others.

2.1.3 REINFORCEMENT LEARNING

In this type of learning a reinforced strategy is used. Its deals with blooming of the knowledge. It's neither Supervised nor Unsupervised form of learning. They use dynamic techniques for letting the user know the output and the derivation of the same.

In these sort of algorithm set they don't assume the environmental set. These are even used in higher and complex mechanism finding likes genetic algorithm. They are widely in progress and implemented most in automation for the better efficiency of the establishment. These algorithms are used in Games and Automation of the vehicle resources.

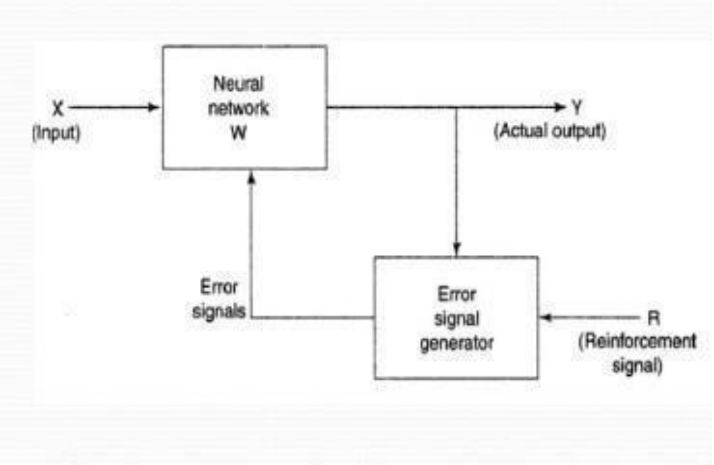

Fig 2.3: Reinforcement Learning

As described in the figure the input vector is passed to a ANN model where the functionalities of the same are stored. If the accurate output is derived then a reward is given to the user making it go to the next level for further task of completion. If not then the Error signal is generated for the same. The accuracy level is calculated and passed down to the user stating the same.

The user sees the percentage of match and pass down and tries other keys of iteration to get the most out of it and complete the task to carry on the ladder of success. This is the same with the machine. Machine iterates the same and to the error signal an add on of reinforced signal is passed which the machine learn and iterates on the same to get closer to the actual results.

2.2 TECHNICAL SURVEY

2.2.1 SURVEY – I

Historic data are of great values and that been proved by Sathik and Sekhar[1]. They derived a hidden patterns from the dataset and have out generated a investment decision plan using different data mining technologies. They used the same output to invest on the stocks. The efficiency of the same was found to be 84.26% which was consider to be a higher hit rate.

2.2.2 SURVEY – II

ANN or Artificial neural networks was discovered later Liam and Jing[2]. They used the ANN techniques to classify, predict and recognize the data sets. In neural network the brain phenomenon is studied and the implementation of brain neurons are tried to be practiced. Output generated from the same were used in trading prediction and stability. In the research pages they have mentioned a seven prediction models in neural network for the higher efficiency yield. Sampling. Training and recommending are one of it's features mentioned.

2.2.3 SURVEY – III

Neural Network was found well integrating with Linear Equation and it's relation. Kun Huang and Tiffany [3] used the same to implement a time series fuzzy network model to improvise and predict the forecasting. The efficiency of the model was found to be deliberate but the computational time was higher than the expected causing it slow for prediction.

2.2.4 SURVEY IV

Bajkunthu and Md. Rafiul [4] approached interrelated market forecasting. The approach for the same was initiated with the help of HMM (Hidden Markov Models). Hmm is used for classification of the item set in bulk and can be even help in pattern matching, It's a hybrid model implemented for efficiency in forecasting of stock market.

2.3 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 make their lifestyle more deliberate than the previous. In the past few years various strategies and the plans had been derived and deployed ever since it's 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 machine and integrating on the same is the hot key of research. There are various people contributing on the same research. ASheta tried its invention on two nonlinear process and had came up with TS which is used as a model for fuzzy sets.

All the learning system 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 human brain itself. The main motto of learning was limitized and 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 market. Monitoring of the same is not possible in the existing system.

The existing system in stock market predictions are apparently biased because it consider 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 days and the loss margin can be higher with respect to time. An initial instance is taken for prediction.

2.4 PROPOSED SYSTEM

Stock is unpredicted and liberal in nature. The follow of the same 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.

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

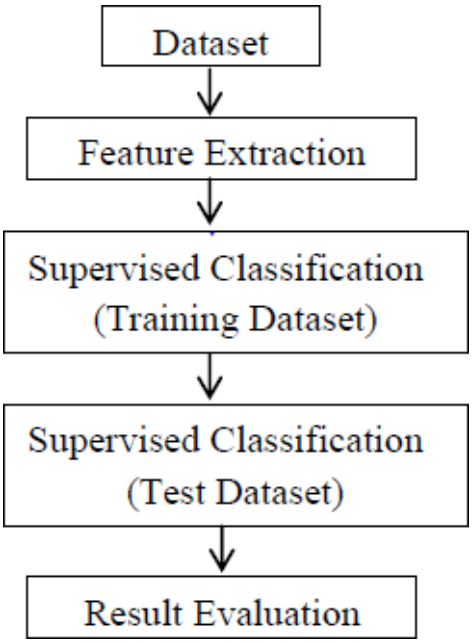

Fig 2.4: System Flow

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's. The result generated are passed and visualized in the graphical form.

2.5 SOFTWARE DESCRIPTION

2.5.1 JUPYTER NOTEBOOK

Jupiter Notebook or so called IPython Notebook is an interactive web based computational mean for starting with Jupiter Notebook documents. The term notebook itself is a huge entity to represent the integration with different entity sets. JSON is the main document form from the same for the execution which follows the brief on the schema and the input and output means. It has high integration with several language set and has various flexibilities with the choices.

The extension used for the same is ".ipynb" which runs in this platform. It's an open-source software package with interactive communication means. It has it's open standards for the same. It's an open community best for budding programmers . The flexibility of the same is phenomenon and splendidly done the configuration and integration of the same is simplest and easy on hold so that no prior distortion is generated and the efficiency of the same is measured through out any system of choice. It's the best software sets that been used across cross for designing and developing of the products and support wide help support.

Not only to that, it provides scalability in the code and the deployment of the same. Various Language can be changed and the project can be undertaken on the same. The created notebook files can be shared and stored in various means for further utilization. It supports cultivated and interactive output sets. Easily crossed over for graphing, plotting and visualizing of the elements.

Data Integration of the same is to it's best. The integration of big data and it can process chunks of values in an approx. time which gives a better performance and the higher computational means. Various works on data like cleaning, cleansing, transforming modeling and visualizing can be done by the same.

CHAPTER 3 REQUIREMENT ANALYSIS

3.1 FUNCTIONAL REQUIREMENTS

Functional requirements deals with the functionality of the software in the engineering view. The component flow and the structural flow of the same is enhanced and described by it.

The functional statement deals with the raw datasets that are categorized and learning from the same dataset. Later the datasets are categorized into clusters and the impairment of the same is checked for the efficiency purpose. After the dataset cleaning the data are cleansed and the machine learns and finds the pattern set for the same it undergoes various iteration and produce output.

3.2 NON-FUNCTIONAL REQUIREMENTS

Non functional requirement deals with the external factors which are non-functional in nature It is used for analysis purpose. Under the same the judgment of the operations are carried out for its performance. Stock is feasible and is ever changing so these extra effects and the requirements helps it to get the latest updates and integrate in a one go where the technicians can work on and solve a bug or a draft if any.

The non-functional requirements followed are it’s efficiency and hit gain ratio. The usability of the code for the further effectiveness and to implement and look for the security console. The System is reliable and the performance is maintained with the support of integration and portability of the same.

3.3 HARDWARE REQUIREMENTS

Processor	: Intel i5 or above
RAM	: Minimum 225MB or more.
Hard Disk	: Minimum 2 GB of space
Input Device	: Keyboard
Output Device	: Screens of Monitor or a Laptop

3.4 SOFTWARE REQUIREMENTS

- Operating system : Windows & Linux
- IDE : Jupiter Notebook
- Data Set : .csv file
- Visualization : mat plot lib, pandas.
- Server : Web Server with HTTP process.

CHAPTER 4

DESIGN

4.1 DESIGN GOALS

To make the project runs smoothly it's required that we make plan and design some accepts like flowcharts and system architecture which are defined below.

4.1.1 Data Collection

Data collection is one of the important and basic thing in our project. The right dataset must be provided to get robust results. Our data mainly consists of previous year or weeks stock prices. We will be taking and analyzing data from Kaggle. After that seeing the accuracy we will use the data in our model.

4.1.2 Data Preprocessing

Human can understand any type of data but machine can't our model will also learn from scratch so it's better to make the data more machine readable. Raw data is usually inconsistent or incomplete .Data preprocessing involves checking missing values, splitting the dataset and training the machine etc.

4.1.3 Training Model

Similar to feeding somethings, machine/model should also learn by feeding and learning on data. The data set extracted from Kaggle will be used to train the model. The training model uses a raw set of data as the undefined dataset which is collected from the previous fiscal year and from the same dataset a refine view is presented which is seen as the desired output. For the refining of the dataset various algorithms are implemented to show the desired output.

4.2 SYSTEM ARCHITECTURE

The dataset we use for the proposed project is been taken from Kaggle. But, this data set is in raw format. The data set is a collection of valuation of stock market information about some companies. The initial step is to convert raw data into processed data. Which is done by feature extraction, since the raw data collected have multiple attributes but only some of those attributes are needed for the prediction. Feature extraction is a reduction process.

The structure, behavior and views of a system is given by structural model.

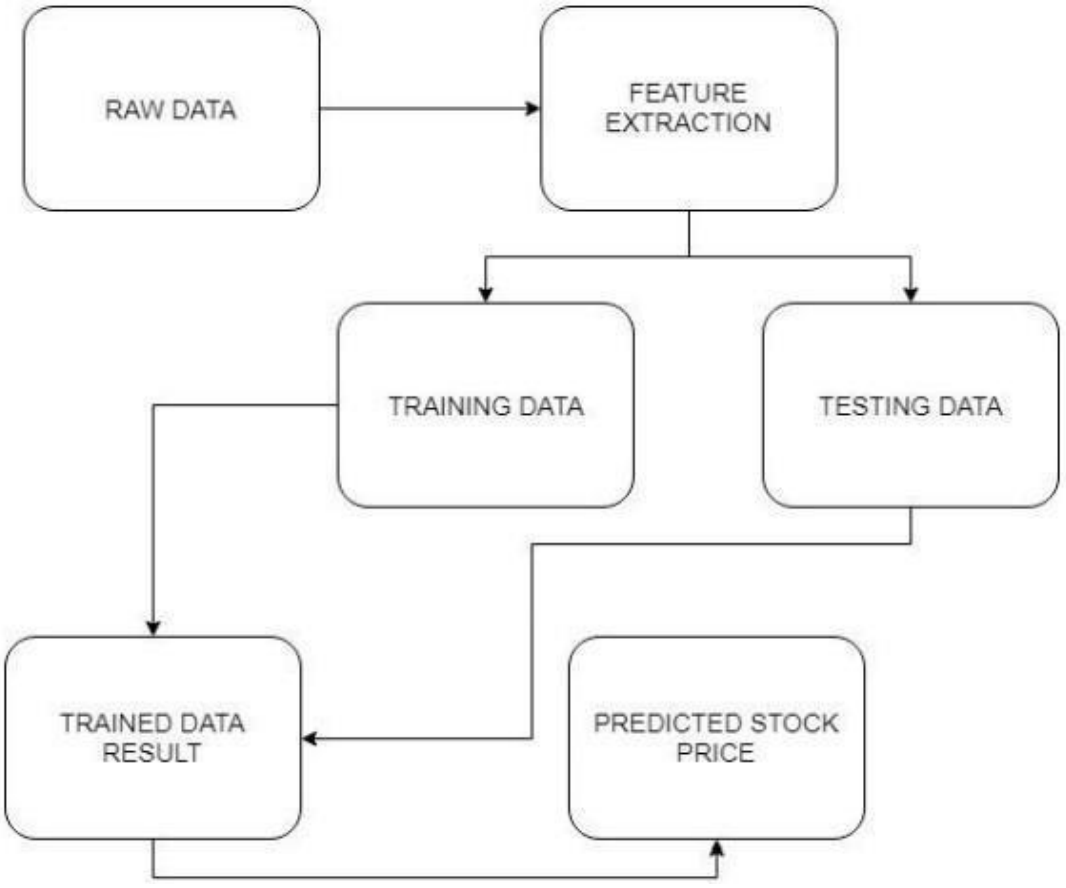

Fig 4.1: System Architecture

The above figure 4.1 gives the demonstration on the dataset extraction and refining the raw dataset by categorizing into two phase of training and testing data.

From the given dataset a well modified categorization is extracted and a graph set is plotted to gain the required output which gives the stock prediction range.

4.3 Use case Diagram

A dynamic and behavioral diagram in UML is use case diagram. Use cases are basically set of actions, services which are used by system. To visualize the functionality requirement of the system this use case diagram are used. The internal and external events or party that may influence the system are also picturized. Use case diagram specify how the system acts on any action without worrying to know about the details how that functionality is achieved.

For the project we have created the below mentioned use case diagram.

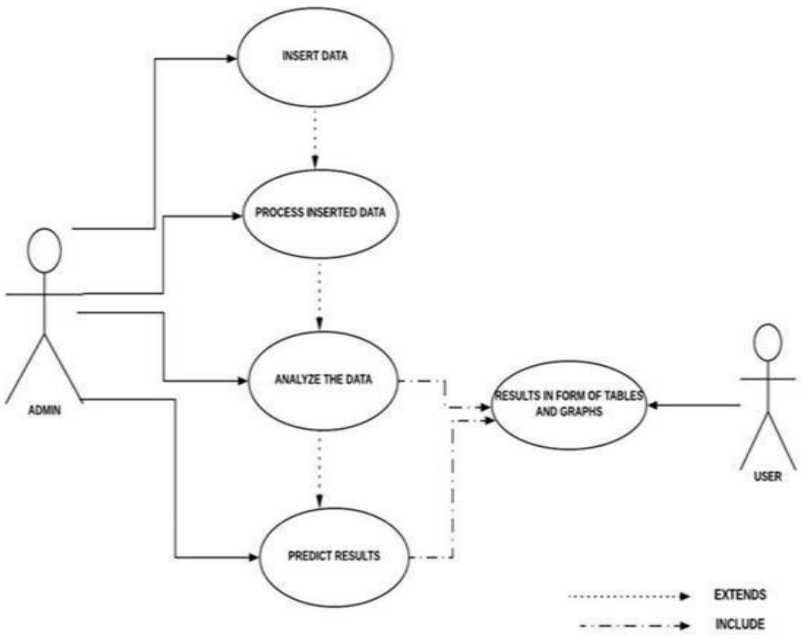

Fig 4.2: Use Case Diagram

The above figure 4.2 shows the use-case diagram of the entitled project and it's flow. From the diagram it's seen that the user gives the raw dataset as input and with the flow of the input in the system.

The system evaluates and process the dataset train itself with the provided dataset and extract the meaningful dataset to process and refine the cluster data and from the given cluster of the data, the plotting of the data values are shown and with

the given range the system plots the data gives a figurative output as prediction and display the same as the refined output in the display screen.

4.4 Data Flow Diagram

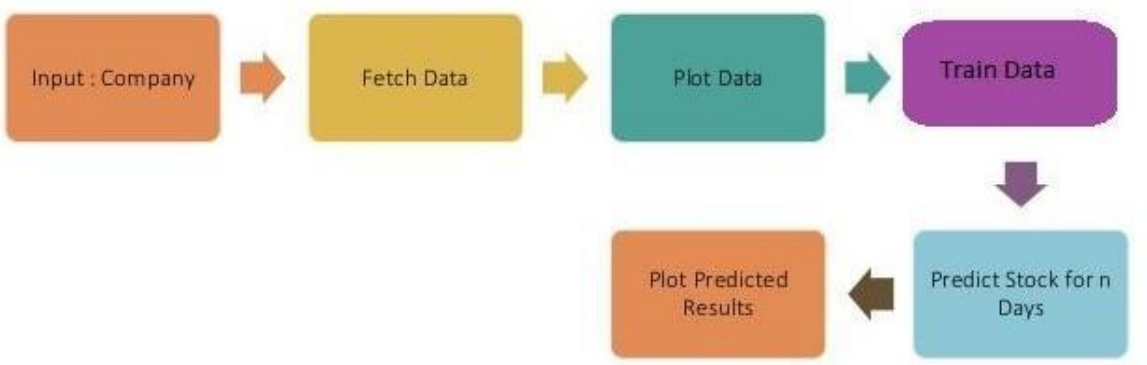

Fig 4.3 Data Flow Diagram

In the above fig 4.3 we are taking a company fetching the data of the company from the panda's data-reader library then we are plotting the data, then we train the data to predict the stock for certain number of days. In this way data is flowing in our system.

CHAPTER 5 IMPLEMENTATION

These are the Machine Learning Algorithms implemented during the building of the project.

5.1 LINEAR REGRESSION

One of the well known algorithm used in machine learning is the 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 which 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 ad 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 shows the growth rate and both are totally depended 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 where as 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 here different fundaments are considered with regards to the dependencies. It even deals with residual errors

5.2 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. 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 the 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 along side with optimized cost factor which is a great boost to the algorithm itself.

Its 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 an Linear Regression. It has it's wide range coverage so no distortion of information. No data is loss during the processing and cleansing of the dataset.

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

5.3 K-NEAREST NEIGHBORS (KNN)

One of the Machine Learning Algorithm 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 Data Mining process.

To the aspect this machine learning algorithm is used to solve the regression and classification of the datasets along with it's highly demanded on pattern machine as well as detection of intrusion. As the name suggest 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 are divided into chunks of small dataset that finds the closeness proximity and derive 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.

5.4 DECISION TREE

In general, Decision tree analysis is a predictive modelling 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. Decisions trees are the most powerful algorithms that falls 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 providing 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 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 categorial 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 dataset with the help of 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 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 Gini score calculated in 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.

5.5 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 moulded, painted or packed while it is on a conveyor belt.

CHAPTER 6 TESTING

The purpose of testing is to get errors. Testing is that the process of trying to get every conceivable fault or weakness during a work product. It provides how to see the functionality of components, sub assemblies, assemblies and/or a finished product it's the method of exercising software with the intent of ensuring that the software meets its requirements and user expectations and doesn't fail in an unacceptable manner. There are various sorts of test. Each test type addresses a selected testing requirement. The various types of testing that follows are listed as below.

6.1 UNIT TESTING

Unit testing involves the planning of test cases that validate that the interior program logic is functioning properly, which program inputs produce valid outputs. All decision branches and internal code flow should be validated. it's the testing of individual software units of the appliance.

It is done after the completion of a private unit before integration. this is often a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a selected business process, application, and/or system configuration.

Unit tests make sure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

6.2 INTEGRATION TESTING

Integration tests are designed to check integrated software components to work out if they really run together program. Testing is event driven and is more concerned with the essential outcome of screens or fields.

Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the mixture of components is correct and consistent. Integration testing is specifically aimed toward exposing the issues that arise from the mixture of components.

6.3 VALIDATION TESTING

Validation testing is that the process of ensuring if the tested and developed software satisfies the client /user needs. The business requirement logic or scenarios need to be tested intimately . All the critical functionalities of an application must be tested here.

As a tester, it's always important to understand the way to verify the business logic or scenarios that are given to you. One such method that helps intimately evaluation of the functionalities is that the Validation Process**.**

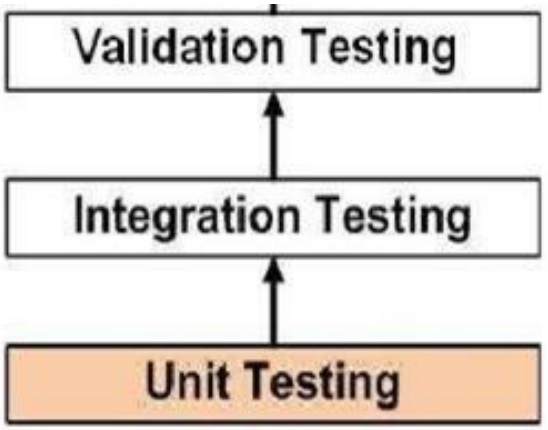

Figure 6.1: The testing process

6.4 SYSTEM TESTING

System testing of software or hardware is testing conducted on an entire , integrated system to guage the system's compliance with its specified requirements. System testing falls within the scope of recorder testing, and intrinsically , should require no knowledge of the inner design of the code or logic.

As a rule, system testing takes, as its input, all of the "integrated" software components that have successfully passed integration testing and also the software itself integrated with any applicable hardware system(s).

System testing may be a more limited sort of testing; it seeks to detect defects both within the "inter-assemblages" and also within the system as an entire .

System testing is performed on the whole system within the context of a Functional Requirement Specification(s) (FRS) and/or a System Requirement Specification (SRS).

System testing tests not only the planning , but also the behavior and even the believed expectations of the customer. it's also intended to check up to and beyond the bounds defined within the software/hardware requirements specification(s).

Test Case -1

Test Case number	TC_01
Module Under Test	Data Extraction
Description	When the program is executed, it tries to connect to YAHOO server for using the data source of the demanded company.
Output	If the connection details are correct, data is extracted . If the details are incorrect, an error is shown.
Remarks	Test Successful.

Test case-2

Test Case Number	TC_02
Module Under Test	Creating Test and Train data
Description	We divide the extracted data into 2 parts test and train data.
Input	As per the requirement the data is given.
Output	Inbuilt data dividers will distribute the data.
Remarks	Test Successful.

Test Case -3

Test Case Number	TC_05
Module Under Test	Plotting of the data
Description	When the data is ready then we plot it using the matplotlib
Input	Data
Output	If the data given is valid then it is plotted.
Remarks	Test Successful.

Chapter 7 SNAPSHOT

Fig 7. 1:Data extraction and plot

Fig 7.2: Linear Regression

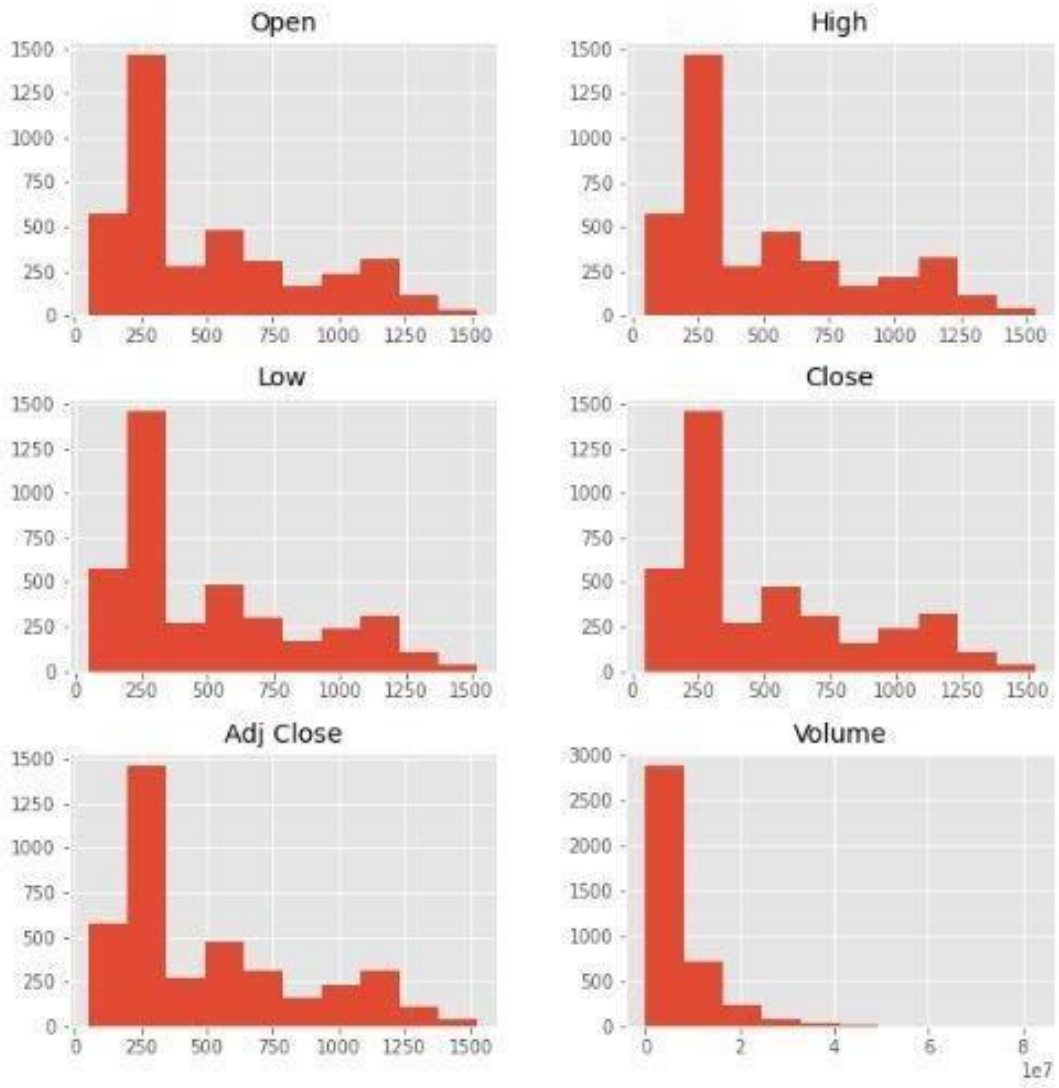

Fig 7.5: Sample data

Fig 7.6: Comparison of different companies

CHAPTER 8

CONCLUSION AND FUTURE ENHANCEMENT

8.1 CONCLUSION

To conclude stock is an unpredictable mechanism which follows the segments of chain and the dependencies of the same are unpredictable. It is defined to be an 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 dependencies 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 are 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 process 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 in the screen in the form of graph.

8.2 FUTURE ENHANCEMENT

Stock Market are the best alternative for business to grow and it's a side way income for the individuals who are ready to invest and earn from the same. The term stock had 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 accurate than it's going to be a gain for both the individuals and the organization. The risk factor have 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 addition implementations of the content that can involve real time scenario and the way of executing and processing the real time scenario. Various constrains has to be added and performance of the same can be acylated in the future time for the effective results. The expected form of the display is graph where as from the same the more appearance and setting of the display can be integrated and a pie-chart and a custom graph can further me implemented on the same.

REFERENCES

[1] K. Senthamarai Kannan, P. Sailapathi Sekar, M.Mohamed Sathik and P. Arumugam, “Financial stock market forecastusing data mining Techniques", 2010, Proceedings of the international multiconference of engineers and computerscientists.

[2] Tiffany Hui-Kuang yu and Kun-Huang Huarng, “A Neural network-based fuzzy time series model to improveforecasting”, Elsevier, 2010, pp: 3366-3372.

[3] Md. Rafiul Hassan and Baikunth Nath, “Stock Market forecasting using Hidden Markov Model: A New Approach”, Proceeding of the 2005 5th International conference on intelligent Systems Design and Application 0-7695-2286-06/05, IEEE 2005.

[4] 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.

[5] P. Hajek, Forecasting Stock Market Trend using Prototype Generation Classifiers, WSEAS Transactions on Systems, Vol.11, No. 12, pp. 671-80, 2012.

[6] 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.

[7] 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.

[8] 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.

[9] 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.