2014

<!ter@tors.In/>

{predicting the unpredictable}

Stock Market Prediction Web Application

NAME OF THE PROJECT:

Stock Market Prediction Based on Markov Model

PROJECT SUBMITTED BY:

Akshat Kumar (10900210041)

Rima Roy (10900210017)

Rohit Kumar Rai (10900210020)

Satrajit Das (10900210068)

UNDER THE GUIDANCE OF

Prof Chandan Banerjee

HOD of Department of Information Technology

DEPARTMENT OF INFORMATION TECHNOLOGY

NETAJI SUBHAS ENGINEERING COLLEGE.

GARIA, KOLKATA – 700152.

Certificate

This is to certify that this Project entitled

“ ------ Stock Market Prediction Based on Markov Model ------“

is a bona fide record of work done by:

Project group:

Akshat Kumar

Rima Roy

Rohit Kumar Rai

Satrajit Das

Under my guidance and supervision and submitted in partial fulfillment of the requirements for the award of B.Tech degree in Information Technology by the West Bengal University of Technology.

Signature of Project guide

Name & designation of Project guide:

Prof. Chandan Banerjee(HOD of Department of Information Technology)

Department of Information Technology

Netaji Subhas Engineering College.

Garia, Kolkata - 700152

ACKNOWLEDGEMENT

We would like to extend our heartiest thanks and gratitude to the people who played a pivotal role in making our final year project a success and make it stand where it is now.

This was a vast project and required great expertise and it wouldn't have been possible without the arduous efforts of our mentor Prof. Chandan Banerjee and Prof. Trishit Banerjee who helped us in every little way possible to implement our thoughts.

Akshat Kumar

Rima Roy

Rohit Kumar Rai

Satrajit Das

CONTENTS

Abstract

Stock market analysis and prediction is one of the interesting areas in which past data could be used to anticipate and predict data and information about future. Technically speaking, this area is of high importance for professionals in the industry of finance and stock exchange as they can lead and direct future trends or manage crises over time. Using the stochastic process called Markov Chains, we sought out to predict the immediate future stock prices for a few given companies. We found the moving averages for the data and the grouped them into ten different states of results. We then applied Markov Chain calculations to the data to create a 4x4 transitional probability matrix. Using this transition matrix we solved a system of equations and found 4 steady states that were variables that represented the probability that a stock price for a given day would fall into one of the four states. When we use this information we can apply our actual data to these equations and predict the next stock prices for the near future. We were able to successfully predict the next few days of stock prices using this method. This application is based on PHP and MySql.

Introduction

In our project, we analyzed a year's worth of stock portfolio for a selected group of companies and apply moving averages and Markov Chains to the data in hopes to predict the stock prices for the near future. The first thing we did was to apply moving averages to create an approximate evaluation of the data. In order to find moving averages, we first had to apply a moving average with an increment of three. This involves taking the sum of three days of stock and then dividing it by three. However, one can only do this starting at day three, because there was enough data to actually create a moving average. We eventually had a data set that included a moving average price and a closing price. We then needed to find a difference data set to apply Markov Chains to. We took the difference of the closing price and the moving average price. These differences were going to be what we applied Markov Chains to.

However, we first had to group the differences into four blocks. The reason why we did this was to create more accurate observations that it makes it more exact when we analyze the data. We then create a transition matrix. The entries in the matrix represented how many times the data points go from one block to another. This leads to 16 observations of data. For example, the first row of entries the matrix represents the number of times the data goes from the first block and stays in the first block, the number of times the data goes from the first block to the second block, the number of times the data goes from the first block to the third block, and finally the number of times the data goes from the first block to the fourth block. All entries needed to be in decimal form, so the total number of observation points divided each entry.

With the matrix, we could now apply Markovian properties to our data. In other words, using Markovian properties we created a system of equations with the unknown variables being our steady states that we are aiming to obtain. These equations are sums of probabilities multiplied by our unknown variables. We then aimed to solve the system of equations to find our steady state probabilities.

With our steady state probabilities we were now able to predict where each immediate stock price can fall into an interval. These probabilities are now good indicators of where the stock prices will fall. We then observed the new data and made some observations. We were able to predict what the possible price range for a given day could be. In conclusion, applying Markov Chains is an effective way to predict stock prices.

Related work

There has been much work done in the field of stock market prediction which has been based on neural network, hidden Markov model etc. To keep our project simple with a decent level of accuracy, and also owing to the time constraint we opted for Markov chain process to make the predictions based on [1]. The field of prediction is progressing each day, with constant attempts to increase the accuracy of the predictions and hence this project also leaves a lot of chance for future works and enhancements.

Proposed work

Software & Hardware requirements :-

Software :

1. XAMPP Server (Apache, Mysql, php).
2. Sublime text.
3. Google chrome for testing.

Hardware :

1. Core i3
2. 4 GB RAM
3. 500 GB Hard Disk

The web app has been developed to run across all platforms.

BACKGROUND:-

A Markov Chain is a stochastic process that has the Markovian property.

Definition 1.1: A stochastic process is defined to be an indexed collection of random variables $\{X_t\}$, where the index t runs through a given set T, generally labeled with the set of non-negative integers. The variable X_t is meant to represent a measurable characteristic, or point of interest. For example, if one were to look at a collection of muffins, some with blueberries and some without blueberries, the variable X_t can be used to label muffins with blueberries. Supposedly, if there were four blueberry muffins within a given collection, the set X_t could be designated as the set of blueberry muffins, with each muffin labeled as X_1, X_2, X_3, or X_4.Thus, it is evident from this example that stochastic processes are discrete collections of random variables.

A stochastic process often has the following structure:

The current status of the system can fall into any one of a set of (M+1) mutually exclusive categories called states. For convenience, these states are labeled with integers from 0 to M. The random variable Xt represents the state of the system at time t, so it's only possible values are 0 to M. The system is observed at particular points of time, labeled t= 0 to M. Thus, the stochastic process $\{X_t\} = \{X_0, X_1, X_2, \ldots\}$ provides a mathematical representation of how the status of the physical system evolves over time. Using the previous example of a collection of muffins, the variable X_2 here would represent the number of blueberry muffins at time, t=2.

Definition 1.2: A stochastic process $\{X_t\}$ is said to have the Markovian property if $P\{X_{t+1}=j \mid X_0 = k_0, X_1 = k_1,\dots,X_{t-1} = k_{t-1}, X_t = i\} = P\{X_{t+1} = j \mid X_t = i\}$, for t = 0,1,2... and every sequence i, j, k_0, k_1,..., k_{t-1}. This is saying that the probability of X_{t+1} being equal to j is solely dependent upon the preceding event of what X_t equals.

Conditional probabilities for Markov Chains are called transition probabilities.

Definition 1.3: If Conditional probabilities are defined as $P\{X_{t+1}=j \mid X_t=i\}$ then, for each i and j, stationary one-step transition probabilities for a Markov Chain are defined as, $P\{X_{t+1}=j \mid X_t=i\}=P\{X_1=j \mid X_0=i\}$ for all t=1,2,...

Stationary transition probabilities indicate that transition probabilities do not change over time. Aside from one-step transition probabilities, Markov Chains can also have n-step transition probabilities, which is the conditional probability that the process will be in state j after n-steps provided that it starts in state i at time t.

Definition 1.4: n-step transition probabilities are defined as the conditional probability $\{X_{t+n}=j \mid X_t=i\}=P\{X_n=j \mid X_0=i\}$ for all t=0,1,...

Therefore, a Markov Chain is a stochastic process that states that the conditional probability of a future event relies on the present state of the process, rather than any past states, or events. A conventional way to note stationary transition probabilities that will be seen later in this paper is:-

$P_{ij}=\{X_{t+1}=j \mid X_t=i\}$

$^{(n)}=\{X_{t+1}=j \mid X_t=i\}$.

Chapman-Kolmogorov Equations:

We use Chapman-Kolmogorov Equations to provide a method to compute all of the n-step transition probabilities:

$^{(n)}= \sum{}^{(m)}p_{kj}{}^{(n-m)}$ For all i = 0, 1, . . ., M,

j = 0, 1, . . ., M,

And any m = 1, 2, . . ., n -1,

n = m + 1, m + 2, . . .

These equations are used to point out that when we go from one steady state to another in n steps, the process will be in some other state after exactly m (m is less than n) states. Thus the summation is just the conditional probability that, given a starting point in one state, the process goes to the other state after m steps and then to the next state in n –m steps.

Therefore, by summing up these conditional probabilities over all the possible steady states must yield

$^{(n)}=\sum p_{ik}{}^{(n-1)}$

And

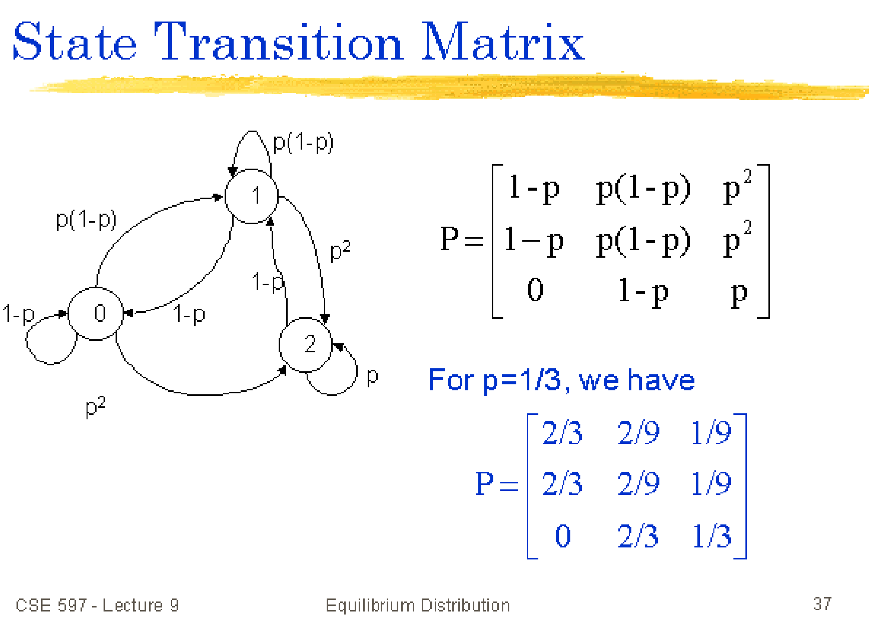
$^{(n)} = \sum {}^{(n-1)} p_{kj}$

This means that these expressions allow us to obtain the n-step probabilities from the one-step transition probabilities recursively.

Transition Matrix

The conditional probabilities for a stochastic process can be organized into an n-step transition matrix. A transition matrix shows the transition probability in a particular column and row as the transition from the row state to the column state. Since transition matrices are comprised of conditional probabilities, each entry of a transition matrix is nonnegative and less than 1. Each row of a transition matrix must also sum to the value 1 since each row signifies a state of the overall stochastic process, and each entry within each row is a conditional probability for the process to be in that state.

State Transition Diagrams: A convenient and useful method to visualize the state of Markov Chains when they have stationary transition probabilities and a finite number of states is through the use of a state transition diagram. In such diagram, each state of a Markov chain is drawn as a numbered node, and the conditional probability of moving from one state to another is drawn by connecting the nodes with an edge and labelling the edge with the numbered probability.



Categorizing States of Markov Chains

Since Markov chains are long run stochastic processes that include transitional probabilities which indicate the likelihood the process will move from one state to another, it is often necessary to categorize, or classify, the varying types of states.

Definition 1.5: A state k is said to be accessible from a state j if$^{(n)}$>0 for some $n\geq 0$, or simply stated, the system can eventually move from state j to state k.

Definition 1.6: If a state j is accessible from a state k and k is accessible from state j then states j and k are said to communicate with one another.

In a Markov chain, every state communicates with itself, since$^{(0)}$=$P\{X_0=j\,|\,X_0=j\}=1$, and if a state j is said to communicate with another state k then k communicates with j. If a state j communicates with k and k communicated with l then state j communicates with state l.

Since different states can communicate with one another within the same system, Markov Chains can be placed into classes, which are groupings of states that only communicate with one another. If every state of a Markov chain communicates with every other state within the chain, that is if the entire Markov chain is in itself one class, then the chain is said to be irreducible.

Transient States

When studying a Markov chain, certain states may only be accessible from others and often times whether or not the system will move from one state to another is convenient to know.

Definition 1.7: If some state j is accessible from state k, but k is not accessible from j (provided that $j\neq k$), then j is considered a transient state. Therefore a transient state is one where once the process enters the state, the process can never return to the state, and once the process enters the state, there exists a positive probability that the process will move to another state and will never return to the original state. Thus, the chain will only enter transient states a finite number of times.

Recurrent States and Absorbing States

Aside from transient states where once the process enters, it can leave, but will never return, other states have the characteristic that the process will most certainly return to it after it has entered it once, and certain others have the characteristic that once the process enters the state, it will never leave the state.

Definition 1.8: If a stochastic process, such as a Markov chain, enters a state, and will definitely return to it, the state is said to be recurrent. Hence, recurrent states cannot be transient; however, they can be absorbing.

Definition 1.9: A state is considered to be absorbent, if after entering the state, the process will never leave the state. If for example, the state j is an absorbing state, then $P_{jj}=1$.

From the above definitions, it is apparent that when grouping the states of a Markov chain into classes, each state belonging to a class is either transient or recurrent. For an irreducible finite-state Markov chain, every state is recurrent, and for any finite-state Markov chain, all the states cannot be transient.

Periodicity and Ergodicity

Periodicity is defined as the following:

Definition 1.10: For a state j in a Markov chain, the period is the largest integer t (where $t>0$) such that $P_{jj}^{(n)}=0$ for all values of n other than t, 2t, 3t, …

If a process can be in a state j at times m and m+1, the state of the period is 1, and is called aperiodic. Every state within a class of a Markov chain shares the same period. For finite-state Markov chains, aperiodicity can lead to ergodicity.

Definition 1.11: In a finite-state Markov chain, recurrent states that are aperiodic are called ergodic states, and a Markov chain is called ergodic if all of its states are ergodic (or, aperiodic) states.

Steady State Probabilities

After the n-step transition probabilities for a Markov chain have been calculated, the Markov chain will display the characteristic of a steady state. Meaning, that if the value of n is large enough, every row of the matrix will be the same, and such, the probability that the process is in each state does not depend on the initial state of the process. Therefore, the probability that the process will be in each state k after a certain number of transitions is a limiting probability that exists independently of the initial state. This can be defined as:

For any irreducible ergodic Markov chain, $lim_{n\to\infty}{}^{(n)}$ exists and is independent of i. Furthermore, $\lim_{n\to\infty}{}^{(n)}=\pi_j>0$

where the π_j uniquely satisfy the following steady-state equations:

$\pi j=\sum\pi_j p_{ij}$ for $j=0,1,\ldots,M,$

$\sum\pi_j=1.$

The steady state probabilities of the Markov chain are π_j. These values indicate that after a large number of transitions the probability of finding the process in a particular state such as j tends to the value of π_j which is independent of the initial state. The π_j are also known as stationary probabilities, when if the initial probability of being in state j is given by π_j for all j, then the probability of finding the process in state j at time n=1,2,… is also given by π_j, or $P\{X_n=j\}=\pi_j$.

Applying Steady State Probabilities to the Transition Matrix

In order to solve for the steady state probabilities discussed above, the aforementioned formulas must be applied to the transition matrix, and the linear system needs to be solved.

Stock Terminology

In order to proceed with the application of Markov chains to the prediction of stock prices, a set of terminology regarding stocks is also useful to have. This project in particular uses opening and closing prices.

Definition: A stock market is an exchange where security trading is conducted by professional stockbrokers. It is a public market in which shares of different companies are bought and sold. The technical analysis for it is anticipating future price movements using historical prices, trading volume, open interest, and other trading data to study price patterns.

Definition: The opening price refers to the price of each individual share on the beginning of that trading day

Once a list of stock prices has been found, calculating a moving average for the prices provides a method for forecasting the stock prices. Moving averages show the general tendency of the stock prices over the long run, and therefore provide a simple and useful way to predict the future of the prices.

Definition: Given a sequence, $\{a_i\}_{=1}^{N}$, an n-moving average is a new sequence $\{s_i\}_{i=1}^{N-n+1}$ defined from the a_i by taking the average of the subsequence of n terms15,

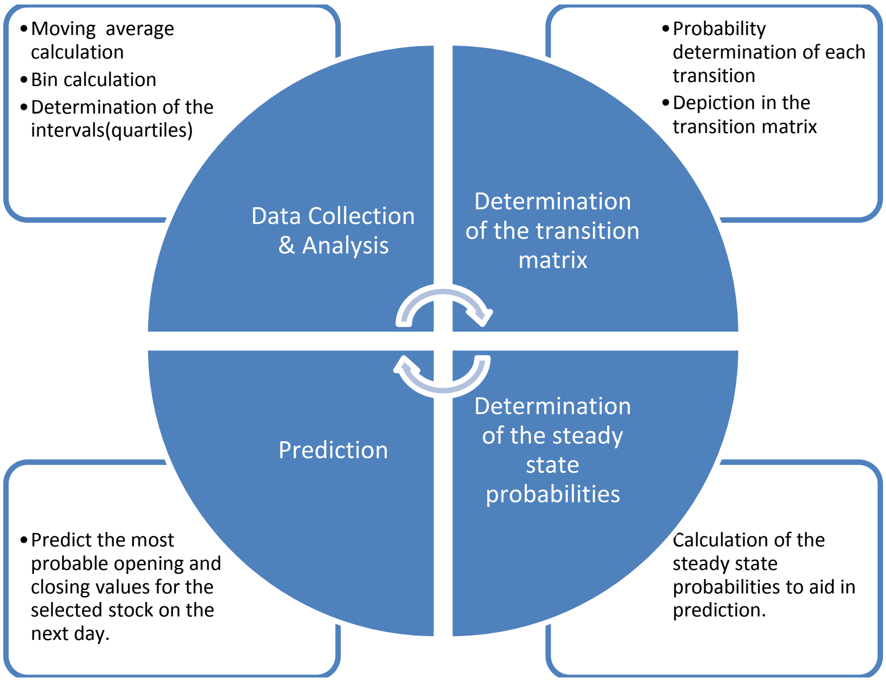
$s_i = 1/(\sum a_j)_{j=1}^{i+n-1}$.

METHODOLOGY:-

As mentioned earlier, the final objective of this project is to examine a set of stock prices and use the probability method of Markov chains to predict the values of the stock prices in their immediate future. Thus to conduct this work, a data set of such prices was first collected, examined, and then the probability method was applied.

System Design

PROCESS OVERVIEW:-



We have selected one year's worth data for five different stocks and have applied Markov chain calculations on this data in order to make the predictions.

Once the stock prices were found, the first step towards applying Markov chains to the data set began with the calculation of moving averages. Moving averages provide a forecast for future prices and therefore are crucial to our work here. Using the difference between the forecasted and actual prices enabled us to make our predictions for the possibility of where future prices may lie. These moving averages were calculated for both opening and closing prices for an interval (called *i*) of 3 days.

After the moving averages were calculated for the set of stock prices, the difference between each actual price and the moving average of each individual day was calculated. This information is what we would use to predict future stock prices. Once the difference between each day's price was calculated, we then focused on binning each of the difference prices into four intervals set within the larger interval from the lowest difference price to the highest difference price. The bins were calculated using the following formula:

K = sqrt(N),

where K is the total number of bins and N stands for the total number of readings taken into consideration. The width of each bin is calculated using the following formula, w = (max value – min value) / K,

where 'max value' and 'min value' refers to the max and min of the difference (between the actual value and the moving average). The intervals were then calculated based on quartile calculations (i.e at intervals of N/4, 2N/4 and 3N/4). Each of the intervals was labeled P1, P2, P3, P4 respectively. After the intervals were established for each data set of difference prices, each individual difference price was labeled as to which interval it fell in.

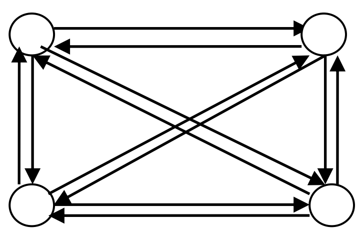
Once each difference price was labeled with its corresponding interval, the number of transitions for each individual difference price interval to the next difference price interval was counted. For example, if Day 115's difference price belonged to interval P2, and Day 116's difference price belonged to interval P3, then a one count was added to the transition from P2 to P3 (labeled for convenience as P23). Every such transition from each interval, or state, was counted and recorded. The number of points belonging to each interval was also recorded.

Once all the above information was recorded, a one-step transition matrix was ready to be prepared. Each entry of the matrix is supposed to be the probability of the data points moving from, or transitioning from, one state to another, with the states corresponding to the appropriate rows and columns. In order to calculate each entry of the matrix, the values of p_{ij} were divided by the total number of difference prices in the interval p_i, which corresponds to the aforementioned P_i.

After this one-step transition matrix is built for each interval on which the moving averages were created, the steady state probabilities can be found. The steady state probabilities are found solving the linear systems with the transition matrix multiplied with the vector πj. The steady states indicate the probability that the difference of the prices will be within the aforementioned intervals. This provides a percentage of where future difference prices may fall, and thus provides for a prediction of what the future holds for these stocks.

Our transition matrix is of the following form:

$$\begin{bmatrix} \mathbf{P11} & \mathbf{P12} & \mathbf{P13} & \mathbf{P14} \\ \mathbf{P21} & \mathbf{P22} & \mathbf{P23} & \mathbf{P24} \\ \mathbf{P31} & \mathbf{P32} & \mathbf{P33} & \mathbf{P34} \\ \mathbf{P41} & \mathbf{P42} & \mathbf{P43} & \mathbf{P44} \end{bmatrix}$$



In this transition matrix each column represents the probability of being in a particular interval (calculated based upon the bin). For example, the first column shows the probability of the price lying in interval P1, the second column depicts the probability of the price lying in interval P2 and so on. Let us say, that our transition matrix is represented by Q.

We have based our calculations to obtain the steady state probabilities by considering various methods before we came up with the following technique. We iteratively compute Q^2, Q^3, Q^4, ……., Q^8. We stop at Q^8 because we observe that the probabilities take a constant value after 5 or 6 iterations. We have still considered till 8 iterations for issues of accuracy.

Then we have considered the steady state probability of being in a particular state as 1, that is, we form 4 steady state probability matrices to tell us the probability or possibility of the stock price lying in a particular range. The matrices are of the form:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

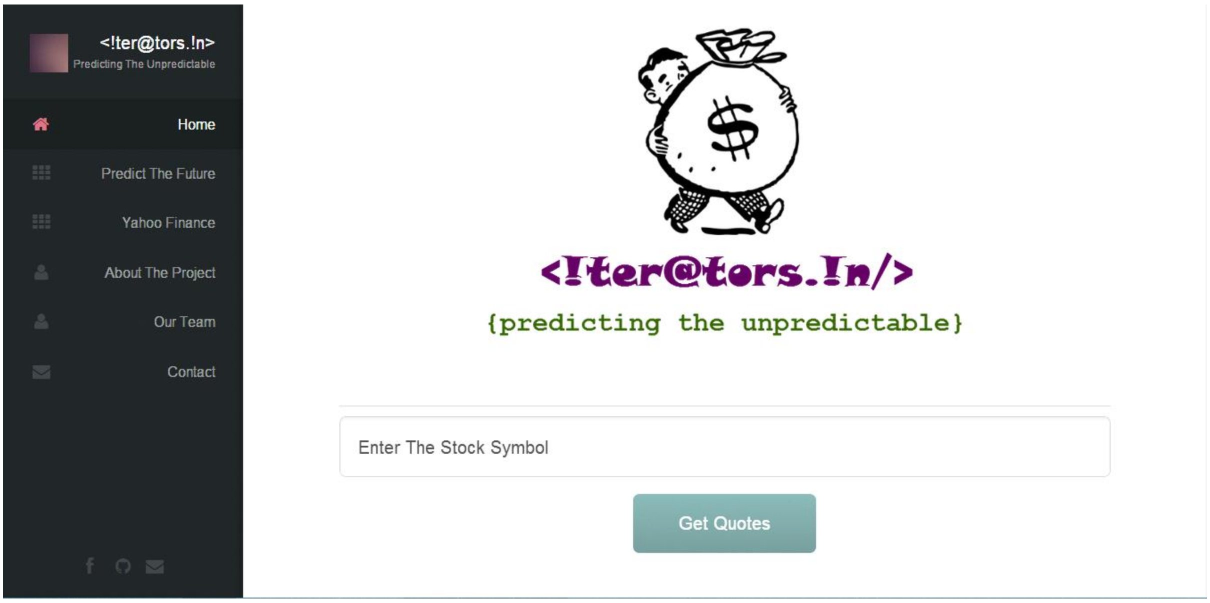
$$B = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$$

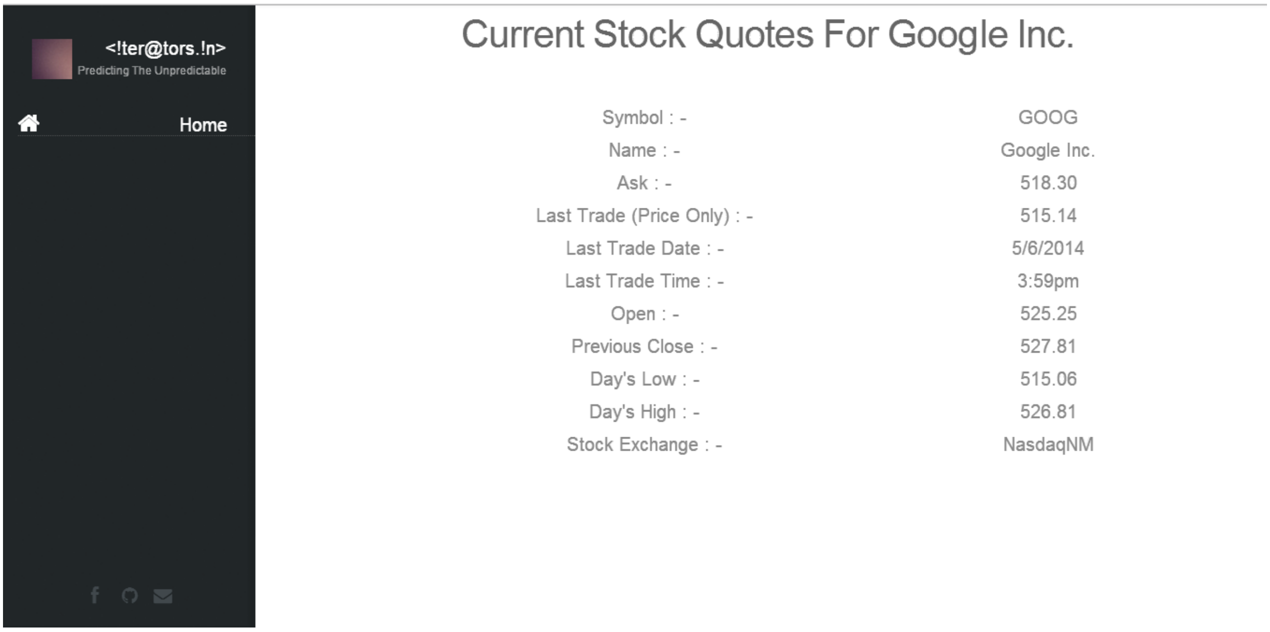
$$D = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$$

Matrix A when multiplied with Q^2 , Q^3 , ……, Q^8 iteratively, finally give us the probability of the stock price lying in interval P1. Similarly when matrix B is multiplied iteratively, it gives the probability of the stock price lying in interval P2, matrix C gives us the probability for interval P3 and D for interval P4 respectively.

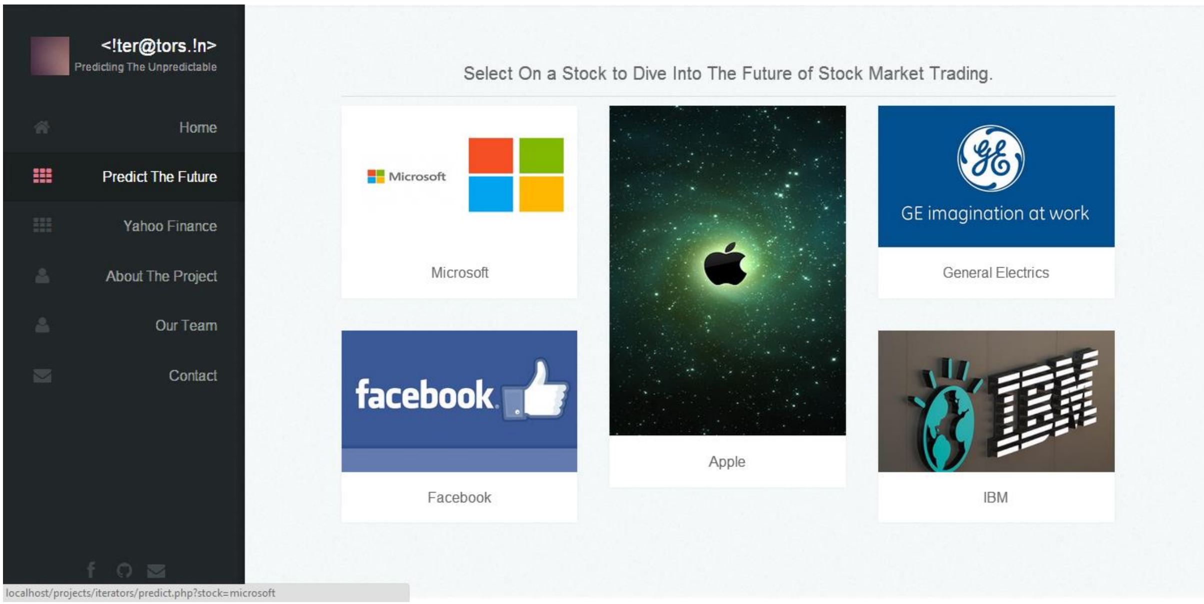
SNAPSHOTS



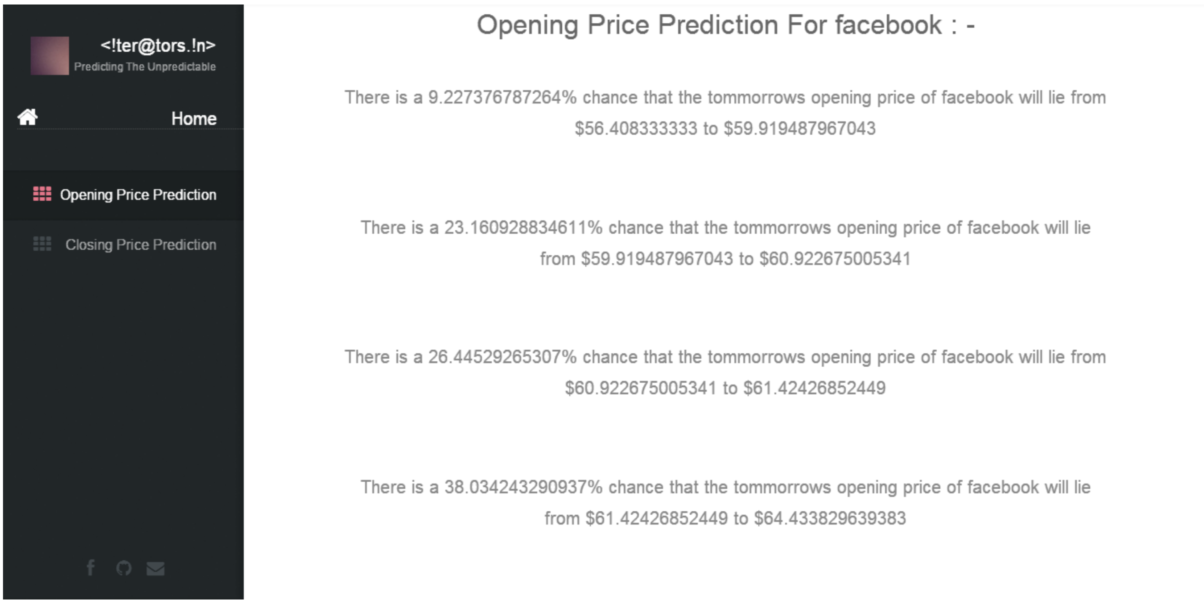
Pic. 1: Homepage



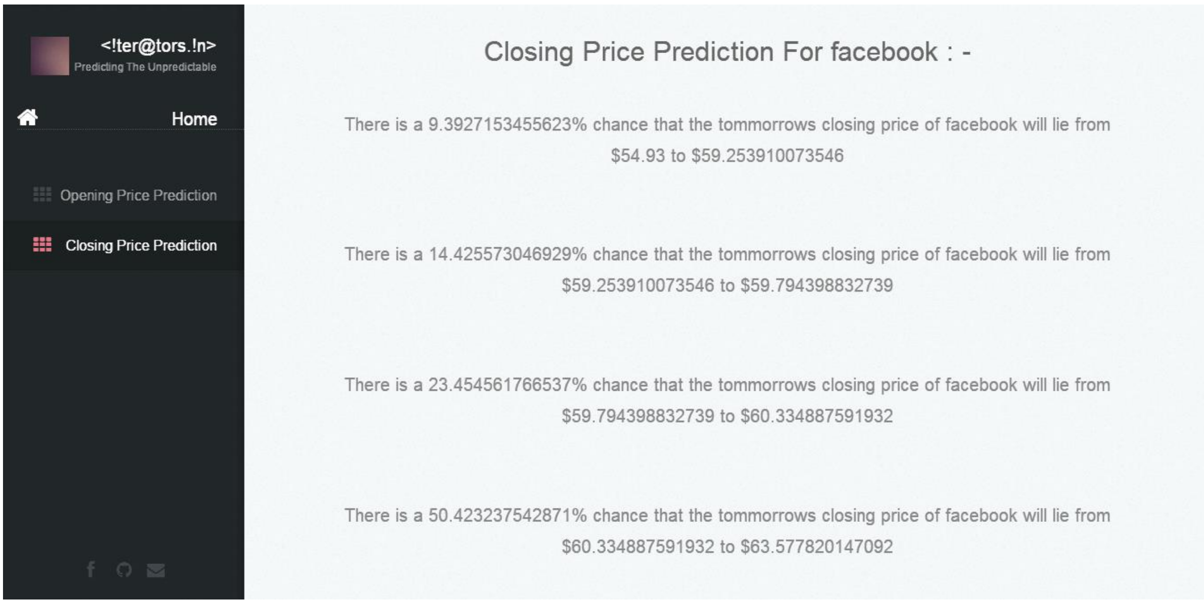
Pic 2: Current Stock Data



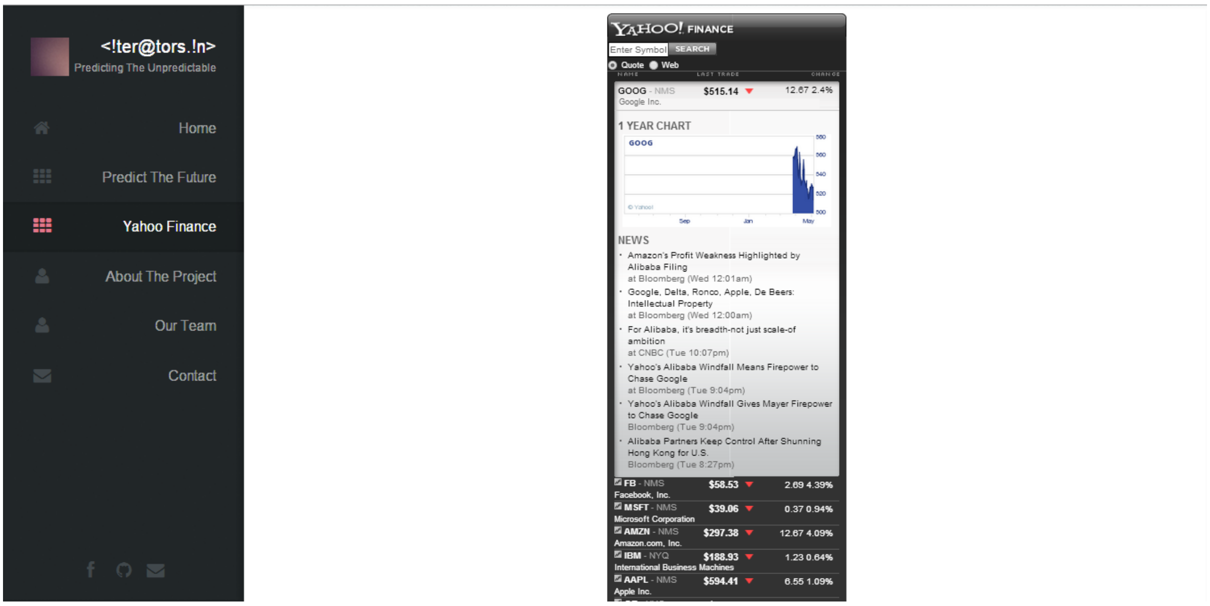
Pic 2: Prediction Options



Pic 4: Opening Price Prediction

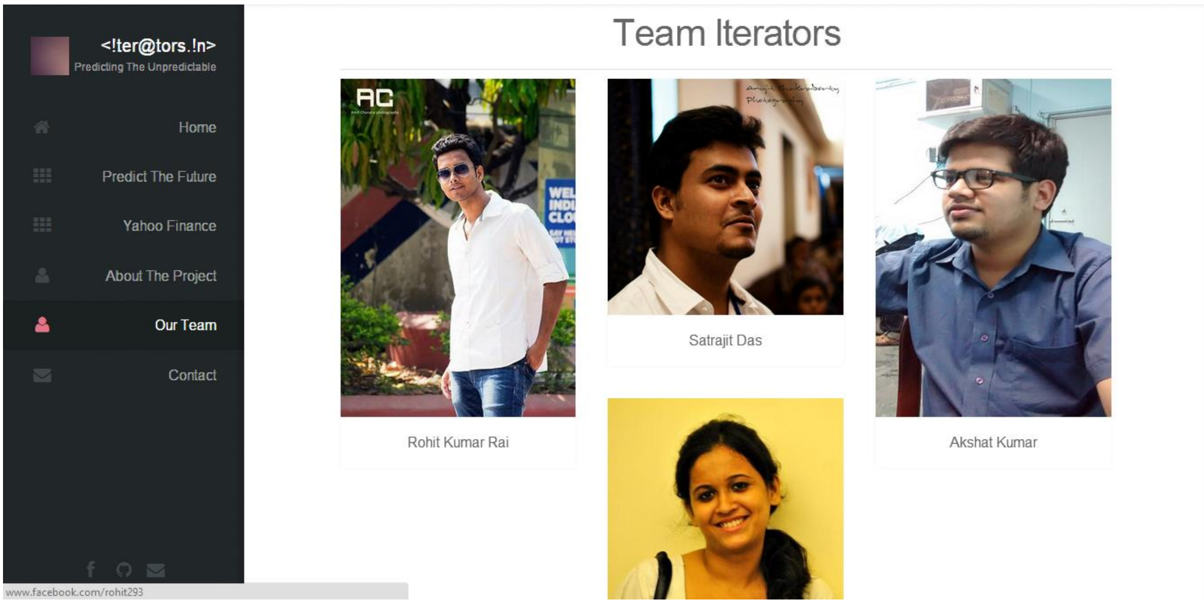


Pic 5: Closing Price Prediction

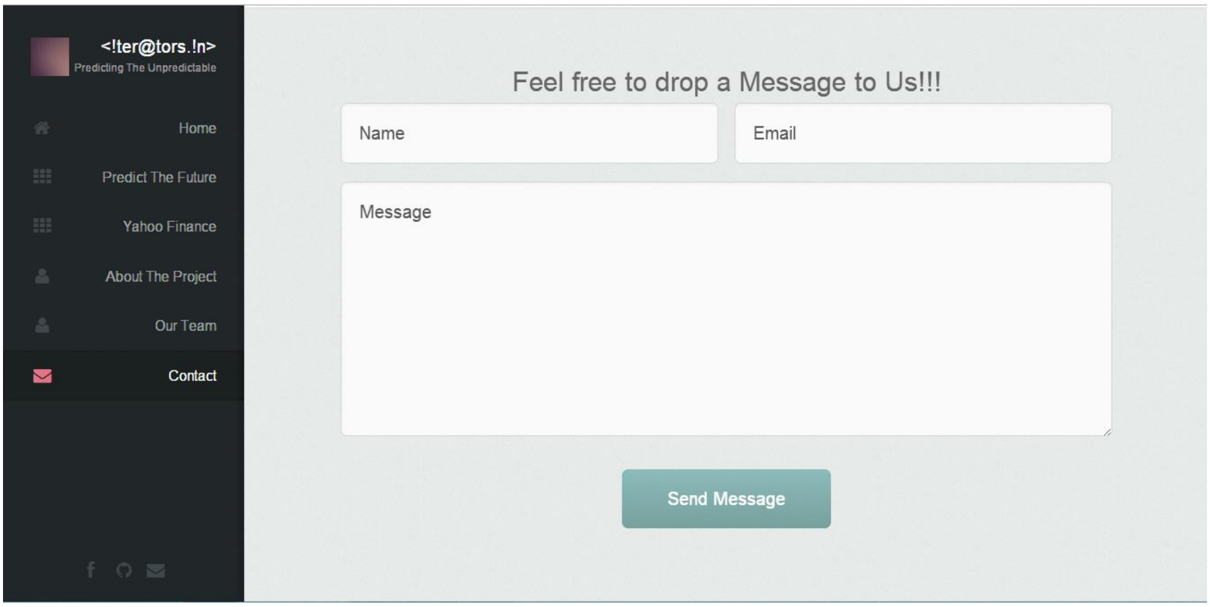


Pic 6: Yahoo Finance Badge

Pic 7: About The Project



Pic 8: Our Team



Pic 9: Contact

Conclusion

Thus we have built this web app in view of all those investors who are at a loss as to whether or not they should make any further investments in a particular share. Since the working algorithm of this app is a mathematical model, the accuracy of this app will be greatly enhanced. This app will be providing mathematical figures which will make it convenient for an investor to come to a decision and that too without the hassle of focusing all their grey cells in an attempt to take note of the expert advices available on various forms of media. This app will be providing predictions at the mere click of a finger. All the user needs to be concerned about is the stock that he/she wants to invest in. Although no prediction can be considered to be perfect, this app strives to achieve near perfection while making predictions thereby reducing the chances of incurring loss for an investor.

Our work thus far has implemented the use of Markov chains to predicting stock prices. Using the difference between forecast prices and actual prices, we have calculated the possible steady state, or probability of the future of the difference price. We checked our app to make predictions for selected stocks and the predictions were found to be nearly accurate. Our app also provides current news feed related to selected stocks, in addition to the prediction, as we have linked it to Yahoo Finance. We are using historical data, from Yahoo Finance, of the companies which are listed in NASDAQ. Our web app is scheduled for a CRON job to run a PHP script in order to update the data (based on which the predictions are made). Thus, our app is ready to enter the market in order to help all the investors.

Future Work

Our project leaves a lot of scope for future enhancements. We can improve accuracy by implementing Hidden Markov model in conjunction with the Markov model. If we are to take our work forward, we might also consider basing our calculations on deciles instead of quartiles (i.e 10 intervals instead of 4)

Although, nothing can be conclusively said about any prediction, it can be concluded that a lot of scope lies in the field of prediction based on such mathematical models.

References

- http://en.wikipedia.org/wiki/Markov_chain
- http://www.math.uah.edu/stat/markov/
- Predicting Stock Prices—(Approved Paper) by Shuchi S.Mitra and Michael J.Riggieri[1]
- http://ichart.finance.yahoo.com/table.csv?s=YHOO&a=04&b=2&c=2014&d=04&e=3&f=2014&g=d&ignore=.csv
- http://finance.yahoo.com/d/quotes.csv?s=fb&f=snal1d1t1opghx