Classification of Finance Data Using Data Mining and Regression Modeling

[1]Rupali Yadavrao Charde and [2]Sumit Sharma

[1]Master of Computer Science, [2]Head of Computer Science Department
[1, 2,] Computer Science
[1, 2] Vaishnavi Institute of Technology & Science, Bhopal, India

_____

Abstract: In this paper the classification of various company data has been analyzed using a Data Mining. Data mining is the field of computer science in which sorting of important data from large data sets to identify patterns and to find some relationships to solve problems through data analysis. It also allows us to predict future trends with great accuracy. By creating associate rules by analyzing data from the relationships and from frequent patterns. We use graphs and charts to classify data sets. Some other data mining are that parameter which include sequence or path analysis. Such as classification, clustering and forecasting. Sequence or path analysis parameter works on the patterns in which one event leads to some another event. A set of order list of common data type is known as sequence while classification parameter moreover looks for new patterns and result to change in the way the data is organized. Forecasting is a method or technique which is used to predict future using regression and past data set using various mathematical models for futuristic model for accurate prediction.

Key words: Apple, Amazon, Microsoft, Facebook, Data Mining, Knowledge Discovery in Data (KDD)
_____

I. INTRODUCTION

The various application of Data Mining enriches the information with wide horizon of knowledge in data pattern or trend. In order to uncover the interesting data patterns hidden in these large data, efficient and scalable knowledge extraction methods are needed. Knowledge is valuable only when it can be used effectively to improve the decision making process of an organization. Traditionally data was turned into analytical knowledge by means of manual experiments [1].

Data mining: It is a process to extract hidden information from real world datasets. Depending on the goals of the knowledge discovery process such as prediction or description, this step applies algorithms to extract the knowledge from the transformed data.

Interpretation and evaluation: The knowledge obtained as a result of performing a data mining task must be correctly interpreted and properly evaluated to ensure that the resulting information is meaningful and accurate. Also it is to be presented to the users in an understandable manner [2]. Generally visualization is the technique used to show the complex results of a data mining task in a simplified manner to venerable data sets

II. ABOUT R

The goal of this section is to provide a introduction to the key issues of the R language. R is a functional language for statistical computation and graphics. It can be seen as a dialect of the S language (developed at AT&T) for which John Chambers was awarded the 1998 Association for Computing Machinery (ACM) Software award that mentioned that this language "forever altered how people analyze, visualize and manipulate data" [1] [3].

The interesting fact about R is that it has great automatic library to the required outcome. The R has created a bridge between user and computational outcomes. In our paper it has help us to predict the future values of stock data of Amazon, Facebook, Microsoft and Apple

III. DATA MINING

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events [1]. Data mining is also known as Knowledge Discovery in Data (KDD).
The key properties of data mining are:
3.1. Automatic discovery of patterns
3.2. Prediction of likely outcomes
3.3. Creation of actionable information
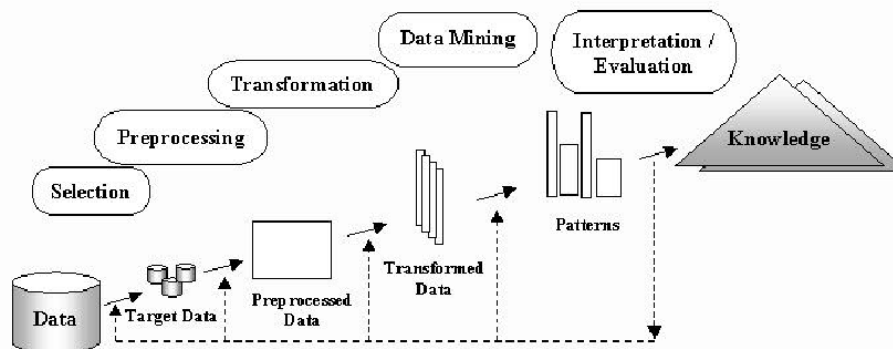3.4. Focus on large data sets and databases

Figure 1: diagram of data mining as knowledge

IV. PREDICTION MATHEMATICAL MODEL

Generally speaking, our goal is to have good forecasts of the future price of the stock, so that profitable orders can be placed on time. This general goal should allow us to easily define what to predict with our models—it should resort to forecast the future values of the price time series of stock [4] [5] [6].

Let the daily average price be approximated by

$$Pi = \frac{Ci + Hi + Li}{3}$$

(1)

Where C_i, H_i and L_i are the close, high, and low quotes for day i, respectively. Let V_i be the set of k percentage variations of today's close to the following k days average prices (often called arithmetic returns):

$$V_i = \left\{ \frac{Pi+j - Ci}{Ci} \right\}_{j=1}^{k}$$

(2)

Our indicator variable is the total sum of the variations whose absolute value is above our target margin p%:

$$Ti = \sum_v \{ v \in Vi : v > p\% \lor v < -p\% \}$$

(3)

We will center our analysis on the Close quote, as our buy/sell decisions will be made at the end of each daily session. The initial set of features will be formed by several past returns on the Close price. The h-days (arithmetic) returns, or percentage variations, can be calculated as

$$R_{i-h} = \frac{C_i - C_{i-h}}{C_{i-h}}$$

(4)

Where C_i is the Close price at session i.

V. Data Extraction

In this paper the report of four companies has been extracted of stock market of Apple, Amazon, Microsoft and Facebook. The clear objective of report is to discover the possibilities of their stock prices and giving a clear aim to work in targeted area [7].
The following data sets are arranged in following manner:-
1. Apple
2. Amazon
3. Microsoft
4. Facebook
5. All clubbed together

In build model process a regression algorithm estimates the target values of the function for prediction for each case in our model. In a model if a different data set of target values is unknown, we can apply relationships between predicators and target to summarize in our model in regression [8]. These models are tested by computing various available statistics that compare the values between expected vale and predicted value. Historically data for this kind of regression model, the project is divided into basic two models.
I. One for building the model
II. One for testing the model

The various application of regression modeling has been found. Some they are laid below [9]
5.1 Financial forecasting
5.2 Time series prediction
5.3 Business planning
5.4 Trend analysis
5.5 Environmental modeling
5.6 Biomedical modeling
There are mainly two ways to iterate the value applying regression:-
1. Linear regression
2. Non-linear regression

VI. Graphs of All Selected Company

Graph 1: The stock price of Apple, Amazon, Microsoft and Facebook


Graph 2: For scale data

At initial period of time Amazon is at peak and followed by Microsoft, Facebook and Apple. Apple rise to peak in June by leaving other behind. Where in next month Facebook grow to boom, this reason is in rainy season due increased in idealized time among the user. The Microsoft market remains constant throughout year till date. The Apple stock market decrease in February 2018 and then boosted up. In last Amazon is on top height and then followed by Apple, then by Microsoft and Facebook at the bottom.Hence we conclude that the following application has given us the forecasted value to planning for future application.

VII. Individual mining of different company

The daily stock price variation is required to clarify forward approach of programs to have high iterated values for stock of Apple, Amazon, Microsoft and Facebook Company all over the world. To have reasonable number it becomes highly demanding to have more accurate result with appropriate mathematical model to resolve it. To have it, we have following way to have it.

7.1.   Apple
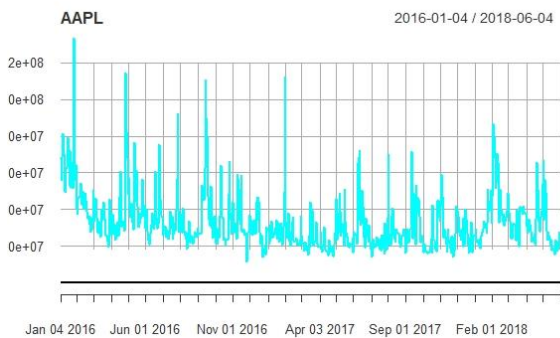
Graph 3: Daily Rise and Fall
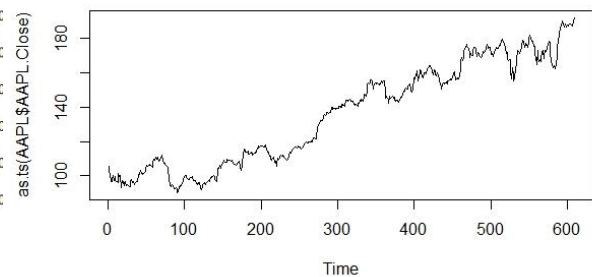
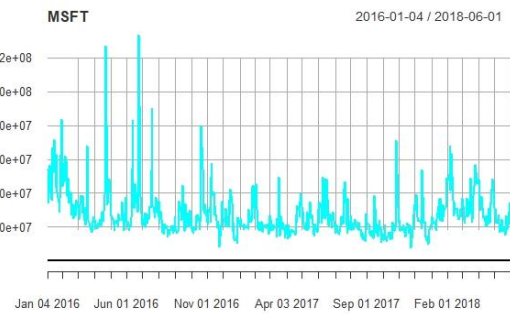Graph 4: Total variation in 609 days

7.2.   Amazon
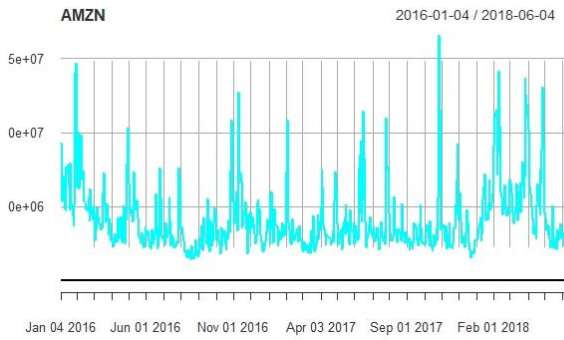
Graph 5: Daily Rise and Fall
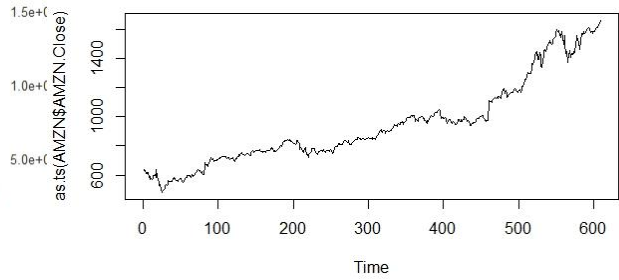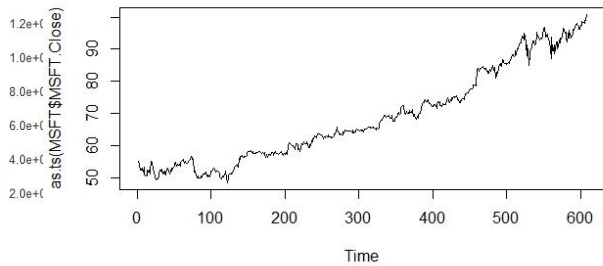

Graph 6: Total variation in 609 days

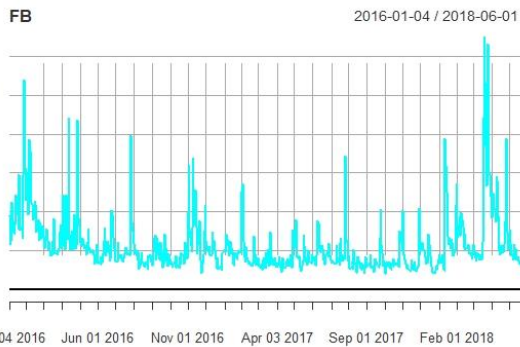### 7.3.   Microsoft

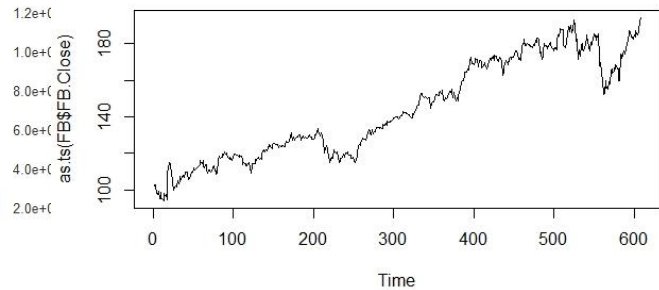Graph 7: Daily Rise and Fall

Graph 8: Total variation in 609 days
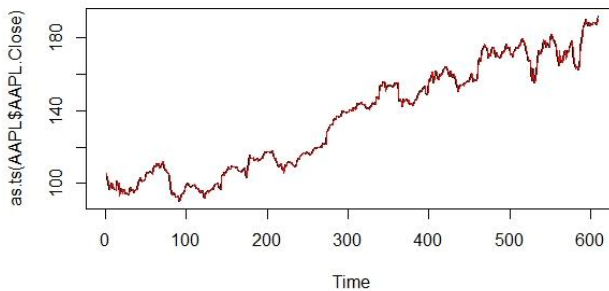
7.4.   Facebook

Graph 9: Daily Rise and Fall

Graph 10: Total variation in 609 days

VIII. The Iteration on Given Data Sets

To trace out the forecasted value we need to put the iteration values to create a mathematical model to have better understanding of   forecasted outcome. In following company graph the the original value is depicted by black colour and regression by red colour line.
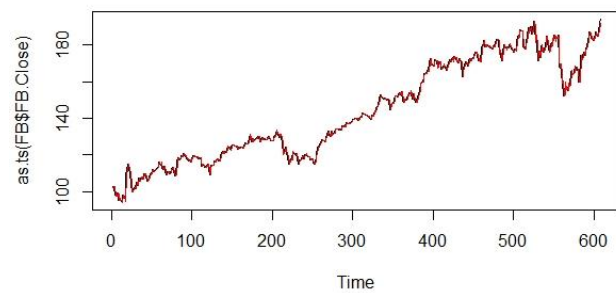

Graph 11: Iteration of mined data sets of APPLE
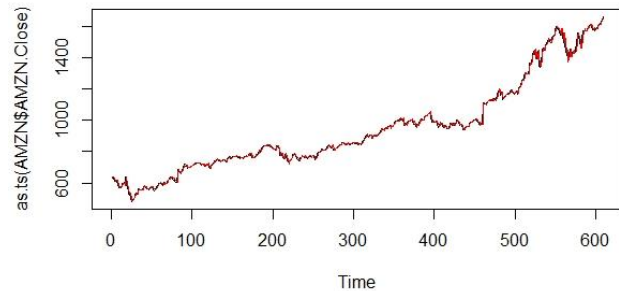
Graph 12: Iteration of mined data sets of AMAZON
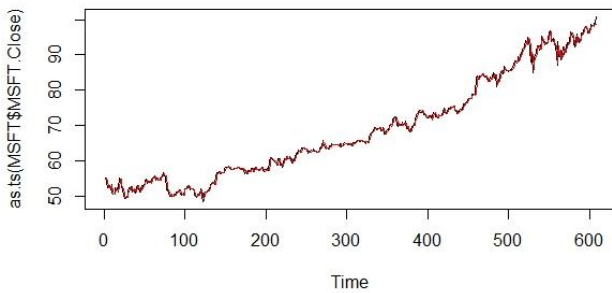
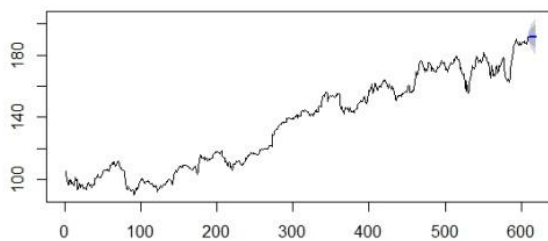Graph 13: Iteration of mined data sets of MICROSOFT

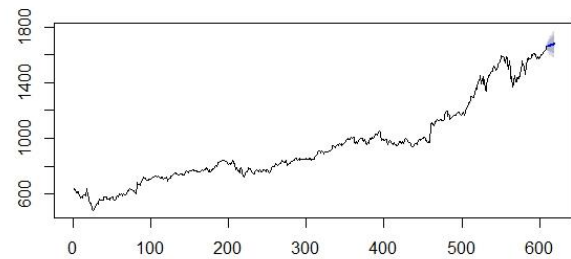Graph 14: Iteration of mined data sets of FACEBOOK

IX. RESULTS

The forecasted values are represented at the last for 10 days in advance by considering the entire variable responsible for the outcome in limit values. It had been observed that 90 percent of time the outcome lies in our predictable range.
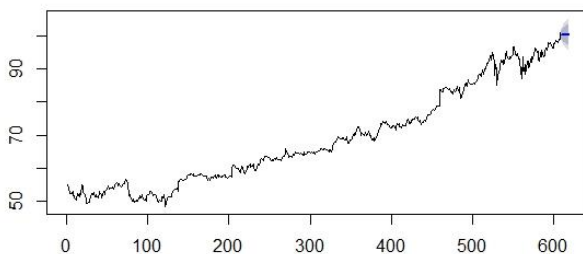

Graph 15: Prediction of data sets of APPLE

Graph 16: Prediction of data sets of AMAZON

Graph 15: Prediction of data sets of MICROSOFT
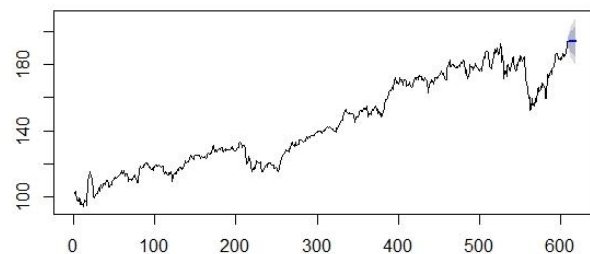
Graph 16: Prediction of data sets of FACEBOOK

From following graph we can conclude that Apple had too much of unstable values. Amazon had a very stable stock and a rapid growth in stock has been discovered, while Microsoft also had some up and down but stable in most of the time. Where as in Facebook stock values has see-saw values and highly unstable due dynamic use of it by users. Our aim is to establish software to know the future result by forecasting in permissible range of our mathematical model has been achieved successfully. This kind of model will help us to understand stock market, about disease growth, product sales etc. Hence Data Mining become a important tool for better future ahead and answer to problem of future.

ACKNOWLEDGEMENT

REFERENCES

[1]. Sachin Bhaskart, Vijay Bahadur Singh, A. K. Nayak-" Managing Data in SVM Supervised Algorithm for Data Mining Technology "- 978-1-4799-3064-7/14/©20 14 IEEE

[2]. Z. Pawlak, "Rough sets and intelligent data analysis," *Inf. Sci. (Ny).*, vol. 147, no. 1, pp. 1–12, 2002.

[3]. Weina Ma "Synthesizing Scenario-based Dataset for User Behavior Pattern Mining " International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 04 – Issue 06, November 2015

[4]. Dr.A.Bharathi , E.Deepankumar "Survey on Classification Techniques in Data Mining "International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 7 1983 – 1986

[5]. Kumaran U. and Neelu Khare "An Empirical Study of Security in Text Mining For Large Datasets"- I J C T A, 9(6), 2016, pp. 2737-2743, ISSN: 0974-5572

[6]. Parkavi A, K. Lakshmi and K.G. Srinivasa "Predicting effective course conduction strategy using Datamining techniques"- Vol. 12(24), pp. 1188-1198, 23 December, 2017, ISSN 1990-3839

[7]. Pavel Cichosz "Data Mining Algorithms Explained Using R" John Wiley & Sons, Chichester, 2015. ISBN 978-1-118-33258-0.

[8]. Kumaran U.* and Neelu Khare "An Empirical Study of Security in Text Mining For Large Datasets"I J C T A, 9(6), 2016, pp. 2737-2743ISSN: 0974-5572

[9]. Peter J. ROUSSEEUW University of Fribourg, ISES, CH-I 700 Fribourg Switzerland Received 13 June 1986 Revised 27 November 1986Journal of Computational and Applied Mathematics 20 (1987) 53-65