Data Mining Methods & Implementation of Predictive Data Mining Architecture

Mr. Venkata Ramana Kathi[1], Dr.GeetanjaliAmarawat[2]
[1]Research Scholar in CSE at MADHAV University, Abu road, sirohi, Rajasthan.
[2]Head & Professor in the Department of CSE at MADHAV University, Abu road, Sirohi, Rajasthan.

Abstract—Predicting the performance of a Stock Market prediction Analysis. The scope of this paper is to investigate the accuracy of data mining techniques in such an environment. The first step of the study is to stock market predction data. We collected records of more than the 30days of Stock Market data , from a National Stock Exchange(NSE) Mumbai. The second step is to clean the data and choose the relevant attributes. In the third step, NaiveBayesSimple, MultiLayerPerception, Query Trending System, Time Series Analysis, Effective Market Hypoheis(EMH) algorithms were constructed and their performances were evaluated. The study revealed that the MultiLayerPerception is more accurate than the other algorithms. This work will help the institute to accurately predict the Stock Market Analysis.

Keywords:-Data Mining, Classification, Decision Tree Algorithm, placement Prediction.

I Introduction

This thesis manages the development of prescient data mining procedures which help the speculators to find concealed examples from the memorable financial time arrangement data to conjecture the stock market. The significant commitments of this thesis are to augment the forecast capacity of financial stock market utilizing time arrangement quantitative technical investigation, especially by defining varieties of existing technical pointers; to devise combinational algorithms by joining different technical markers and the varieties of the markers figured; to devise a remarkable metric in light of "Rectify Profitable Trade Signal-CPTS" for assessment of the stock market foreseeing algorithms; to examine the execution of the Moving Average Crossover (MAC) Algorithm on the noteworthy time arrangement data in the viewpoint of rate of CPTS created and the rate of yearly Returns on Investment (ROI) produced, which is taken as a benchmark to contrast the execution of MAC and the proposed combinational algorithms.

1.1 DEVELOPMENT OF DATA MINING

The present development of data mining capacities and items is the consequence of years of impact from many controls, including databases, data recovery, measurements, algorithms and machine learning. Another computer science territory that has majorly affected the KDD procedure is interactive media and designs. Acceptance is utilized to continue from particular learning to more broad data. This sort of system is frequently found in AI applications.Since the essential goal of data mining is to depict a few attributes of an arrangement of data by a general model, this approach can be seen as a sort of pressure. Here the point by point data inside the database are preoccupied and packed to a littler depiction of the data attributes that are found in the model.

1.3 A STATICTICAL PERSPECTIVE OF DATA MINING BAYES THEOREM

With factual surmising, data about a data circulation is deduced by examining data that take after that appropriation. Given an arrangement of data X={x1,…,…,xn}, a data mining issue is to reveal properties of the circulation from which the set comes. Bayes lead is a system to gauge the probability of a property given the arrangement of data as proof or info. Assume that either hypothesis h1 or hypothesis h2 must happen, yet not both. Additionally assume that xi is a recognizable occasion.

DEFINITION 1.3: Bayes Rule or Bayes Theorem is

$$P(H_1/x_i)= \frac{P(xi\,)/\,P(h1)}{P\left(\frac{xi}{h1}\right)P(h1)+P\left(\frac{xi}{h1}\right)P(h1)}$$

Here P (h1 | xi) is known as the back likelihood, while P (h1) is the earlier likelihood related with hypothesis h1. P (xi) is the likelihood of the event of data esteem xi and P (xi | h1) is the restrictive likelihood that, given a hypothesis, the tuple full fills it.

Where there are m diverse hypothesis we have:
Bayes control enable us to relegate probabilities of hypothesis given a data esteem, P (hj | xi). Here we examine tuples when in reality every xi might be a characteristic esteem or other data name. Every hello there might be a quality esteem, set of property estimations, or even a of characteristic esteems.
Hypothesis testing endeavors to locate a model that clarifies the watched data by first making a hypothesis and then testing the hypothesis against the data. The hypothesis more often than not is checked by examining a data test. In the event that the hypothesis holds for the specimen, it is expected to hold for the populace as a rule. Given a populace, the underlying hypothesis to be tried, H0, is known as the invalid hypothesis. Dismissal of the invalid hypothesis causes another hypothesis, H1, called the option hypothesis, to be made.

One system to perform hypothesis testing depends on the utilization of the chi-squared measurement. Really, there is an arrangement of methodology alluded to as chi squared. These techniques can be utilized to test the relationship between two watched variable esteems and to decide whether an arrangement of watched variable esteems is measurably critical. A hypothesis is first made, and then they watched esteems are looked at in view of this hypothesis. Accepting that O speaks to the watched data and E is the normal esteems on the hypothesis, the chi-squared measurement, X2 , is characterized as:

$$X^2 = \frac{\Sigma(O-E^2)}{E}$$

Direct relapse accept that a straight relationship exists between the info data and the yield data. The basic recipe for a straight relationship is utilized as a part of this model:

$$Y = c0 + c_1x_1 + \ldots\ldots + c_nx_n$$

Here there are n input factors, which are called indicators or regressors; one yield variable (the variable being anticipated), which is known as the reaction; and n + 1 constants, which are picked amid the demonstrating procedure to coordinate the info illustrations. This is called various direct relapse on the grounds that there is more than one indicator.

IIFINANCIAL MARKET AND DATA MINING

2.1 Introduction

In economics, a financial market is a mechanism that facilitates individuals to easily purchase and offer financial securities, for example, stocks and securities, wares, for example, valuable metals or agricultural merchandise, and other fungible things of value at low transaction costs and at costs that mirror the effective market theory.

Trade or business is the ready exchange of products, administrations, or both. A mechanism that licenses trade is called a market. The original type of trade was barter, the immediate exchange of products and enterprises. Modem traders negotiate through a medium of exchange, for example, cash. Thus, purchasing can be separated from offering, or earning.

The creation of cash and later credit, paper cash and non-physical cash, greatly disentangled and promoted trade.

2.2 Types of Financial Market

An economy which relies primarily on interactions between buyers and sellers toallocate resources is known as a market economy. The financial markets can be dividedinto different subtypes:

1. Capital markets which consist of:
a. Stock markets, which provide financing through the issuance of shares orcommon stock, and enable the subsequent trading thereof
b. Bond markets, which provide financing through the issuance of bonds, andenable the subsequent trading thereof
2. Commodity markets, which facilitate the trading of commodities.
3. Money markets, which provide short term debt financing and investment.
4. Derivatives markets, which provide instruments for the management offinancial risk.
5. Insurance markets, which facilitate the redistribution of various risks.
6. Futures markets, which provide standardized forward contracts for tradingproducts at some future date. It is a transaction in which delivery of thecommodity is deferred until after the contract has been made. Althoughthe delivery is made in the future, the price is determined on the initialtrade date. The Futures Market is a market of contracts to buy and sellgoods at specified prices and times.

All these markets are very unsafe and require considerable information and experience to counteract substantial misfortunes. They also require close attention to market developments. Stocks, then again, are less hazardous because developments of the market are usually gradual. Although here and now speculation strategies are conceivable, most view stocks as long haul ventures Trading is the activity of purchasing and offering financial instruments with the end goal of gaining a benefit. Securities are traded in two sorts of markets: primary and secondary market. Recently framed or issued securities are purchased or sold in primary markets.
Secondary markets allow financial specialists to offer securities that they hold or purchase existing securities. It is a market where financial specialists purchase securities or assets from different speculators, rather than from issuing companies themselves.

2.3 Data Mining

Data mining is the procedure of extraction of intriguing, non-trivial, verifiable, beforehand obscure and potentially helpful patterns or information from tremendous amount of data.

Data mining is a solitary stride in a larger procedure of Knowledge Discovery in Databases (KDD). KDD is thought to be an all the more encompassing procedure that incorporates data warehousing, target data choice, data cleaning, pre-processing, transformation and lessening, data mining, show determination, evaluation and interpretation, and finally consolidation and utilization of the extracted "learning".

2.3.1 Descriptive Data Mining

It depicts the data set in a brief manner and presents fascinating general properties of the data. This enables to see sets of summarized data in brief, distinct terms. Such data depictions may give an overall photo of a class of data or recognize it from an arrangement of comparative classes.

2.3.2 Predictive Mining

It is an analytic procedure intended to investigate large amounts of data in search of predictable patterns and/or systematic relationships amongst variables, and then to validate the discoveries by applying the identified patterns to new subsets of data. The ultimate goal is expectation - and prescient data mining is the most widely recognized kind of data mining and one that has the most direct business applications.
This is finished utilizing large noteworthy market data to speak to varying conditions and affirming that the time arrangement patterns have statistically significant prescient power for the accompanying two potential reasons.

1. High probability of profitable trades and

2. High profitable returns for the investment.

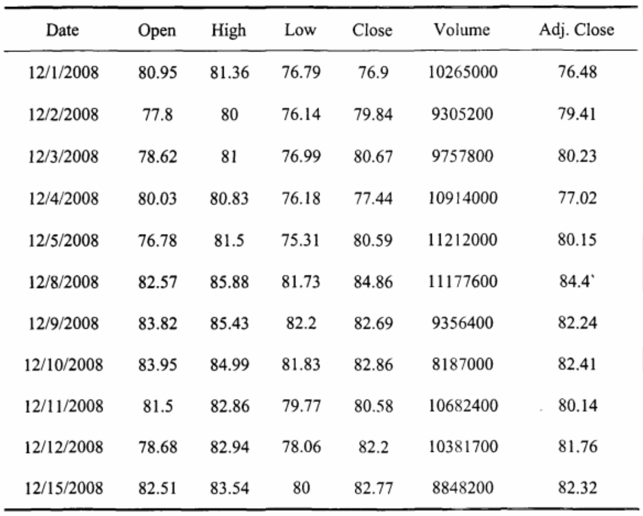
The temporal nature of the financial stock market data resembles the evolvingempirical and evidential data. This has lead to the design of a model for organizingthe large volume of evidence that an intelligence analyst may have at her / hisdisposition

2.4 Time Series

Time series are temporal sequences of measures that can be mined for information. Time series data is a sequence of data points, measured typically at successive times,spaced at (often) uniform time intervals. Time series analysis comprises methods thatattempt to understand such time series, or to understand the underlying context of thedata points like, where did they come from? And what generated them? or to makeforecasts or predictions. The analysis of evolving time series data lead to the discovery ofmany interesting financial market predictive patterns. Evolving time series dataanalysis maximizes the profit-loss value prediction of financial stock using time seriesquantitative analysis.

2.4.1 Financial Time Series Data

The stock market data are stored in bar format, which is a time series data. Atypical time series historic data for IBM stock in bar format for the period from 01-12-2008 to 15-12-2008, downloaded from yahoo finance web site is shown in Table 2.2. TheBar has six fields and they are open, low, high and close prices, volume, and starting timestamp. Open is the open price of the stock in the given time period. Low is the lowestprice of the stock, high is the highest price of the stock and close is the closing price ofthe stock. Volume represents the total number of stocks traded during the period.



Date	Open	High	Low	Close	Volume	Adj. Close
12/1/2008	80.95	81.36	76.79	76.9	10265000	76.48
12/2/2008	77.8	80	76.14	79.84	9305200	79.41
12/3/2008	78.62	81	76.99	80.67	9757800	80.23
12/4/2008	80.03	80.83	76.18	77.44	10914000	77.02
12/5/2008	76.78	81.5	75.31	80.59	11212000	80.15
12/8/2008	82.57	85.88	81.73	84.86	11177600	84.4`
12/9/2008	83.82	85.43	82.2	82.69	9356400	82.24
12/10/2008	83.95	84.99	81.83	82.86	8187000	82.41
12/11/2008	81.5	82.86	79.77	80.58	10682400	80.14
12/12/2008	78.68	82.94	78.06	82.2	10381700	81.76
12/15/2008	82.51	83.54	80	82.77	8848200	82.32

Table 2.2 A typical Financial Times Series Historic Data of IBM Stock

III FINANCIAL ANALYSIS

3.1 An Efficient Market

The idea of productivity in financial matters is a general time for the esteem allotted to a circumstance by some measure intended to catch the measure of waste or "grinding" or other undesirable monetary elements introduce. Inside this specific situation, it has a few very particular implications.

An "efficient" market is characterized as a market where there are huge quantities of objective, "benefit maximizes" effectively contending, with each attempting to anticipate future market estimations of individual securities, and where imperative current data is unreservedly accessible to all members. An opposite point of view of this view is introduced beneath.

3.1.1 The Efficient Market Hypothesis (EMH)

The Efficient Market Hypothesis (EMH) was created by Professor Fama, who is frequently thought of as the father of EMH. The EMH proposes two urgent ideas that have characterized the discussion on efficient markets from that point.

1. Types of informational efficiency.
2. Joint Hypothesis problem.

3.1.2 Joint Hypothesis Problem

The Joint Hypothesis Problem expresses that trial of the EMH are joint trial of market productivity and the specific resource estimating model used to lead the test. The idea of "Joint Hypothesis Problem" in EMH is shown as the thought of market effectiveness, which couldn't be rejected without a going with dismissal of the model of market harmony.

The advantage evaluating model depends on the model of market harmony and it is a circumstance where the supply of anything is precisely equivalent to its demand. Since neither there is surplus nor deficiency in the market and consequently, there is no natural inclination at the cost of the thing to change. This is the general price setting instrument. This idea has as far back as vexed scientists

3.2 Fundamental Analysis

Essential analysis is the investigation of an organization's "basics" with the point of ascertaining precisely what a recorded organization is worth. In view of that valuation, financial specialists will either purchase or offer its offer. The Fundamental Analysis depends on the macroeconomic data and the essential financial status of organizations like money supply, loan fee, inflationary rates, profit yields, income yield, and income yield, book to market proportion, price-income proportion and slacked returns.

3.2.1 Top-Down Approach

The best down contributing takes a gander at an organization's working condition notwithstanding its own techniques and likely future execution. The systems can be compared to a rearranged pyramid. Adopting a best down strategy to an organization's profit's prospects includes first taking a gander at the expansive macroeconomic, social and political condition in which the organization works.

3.2.2 Bottom-Up Approach

The Bottom-up essential analysis channels the diverse organizations in a segment by taking a gander at the individual "venture story" of each, and breaking down a significant number of various numbers. These numbers incorporate both the financial articulations distributed by each organization and also particular proportions that speculators ascertain from these in a procedure that is known as "calculating". It additionally incorporates an appraisal of an organization's administration: its believability, encounter and vital knowledge. From that the expert makes determinations concerning the engaging quality of that particular organization as a speculation and assessments it's actual, or characteristic esteem.

3.3 Types of Trading Systems

Many different kinds of trading systems used by professional practitioners to meetout various application needs are described below.

3.3.1 Chart Trading System

This is the first level of trading system and is based on plotting the data as charts.

The data having a place with a solitary time interim is utilized as a part of the diagram trading frameworks. There are no near examinations between different time interims. Consequently, outline trading frameworks are exceptionally principal in nature. Be that as it may, turning out with the best diagram trading framework is the initial phase during the time spent building complex trading frameworks. Along these lines, the diagram trading frameworks hold the key for future achievement. The exactness of these frameworks must be high and the arrival from them ought to likewise be high.

3.3.2 Query Trading System

Question Trading System is intended for near examination. Examination is done inside a stock crosswise over various time interims of bar data. Correlation is done between various stocks also. Inquiry gives the office to contrast a stock and any file image. For instance, one can contrast a stock and the list of the business gathering to understand how a stock is performing contrasted with the record. As it were, gold stocks having images "GOLD" – Randgold Resources Ltd, "NEM"-Newmont Mining Corporation, "AU" - AngloGold Ashanti Ltd, and "KGC" - Kinross Gold Corporation can be contrasted and the gold list "$XAU.X". Note that each record is created by consolidating weighted estimations of individual stocks inside the business gathering.

IV RESULTS AND ANALYSIS

The goal of this proposition, "Advancement of Predictive Data Mining Architecture", is to plan and create calculations to gauge the stock market incline in light of the entomb day monetary time series data. The examination center is bound to the push range of combinational quantitative specialized calculations defeating the constraints of different techniques talked about in the section IV. All the monetary time series anticipating calculations revealed till date utilize "Era of Correct Trade Signal - GCTS" metric mirroring the Market Trend to demonstrate the effectiveness of the particular calculations. 'Rates of return - ROT, to be specific, ensured profit is a 'Prime Metric' to quantify the "Effectiveness" of any monetary time series gauging calculation. Along these lines, straightforward era of correct trade signals demonstrating stock market inclines alone can't be considered as the "Effectiveness" measure of any monetary time series anticipating calculation as they don't promise themselves on 'Degrees of profitability - ROT of money to be contributed, i.e, profit.

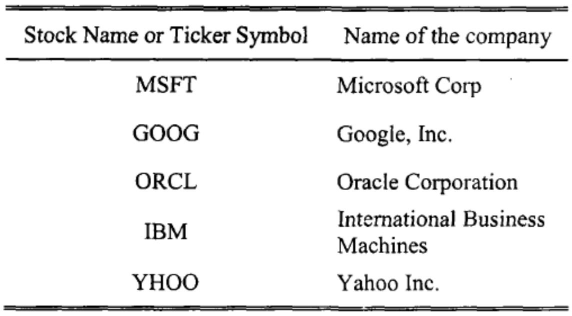
4.1 The Performance Evaluation Metrics

The key metric, CPTS and other metrics used to articulate the performance evaluation of the developed algorithms discussed in Chapter V are built over the concept of CTT and are listed below. The CPTS is discussed in the rest of this chapter. This is needed to measure the effectiveness, efficiency and quality of the techniques used for forecasting the stock market.

- Number of profitable trade signals generated,
- Percentage of CPTS generated
- Percentage of false or non-productive signals generated and
- Percentage of annual ROI of money.

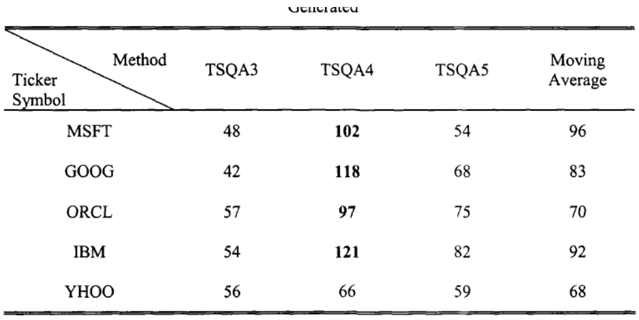
4.2 Results and Analysis

The execution as far as aggregate number of predictions made by TSQA3, TSQA4, TSQA5 and the seat stamp specialized marker Moving Average Crossover calculation for the stocks considered for back testing is given in the Table 6.2. A calculation which produces most extreme number of trade signals is ordinarily great. In this viewpoint TSQA4 is doing admirably and the same can be seen from the graphical introduction appeared in Figure 6.1. The best esteems are highlighted in flagrant characters.



Stock Name or Ticker Symbol	Name of the company
MSFT	Microsoft Corp
GOOG	Google, Inc.
ORCL	Oracle Corporation
IBM	International Business Machines
YHOO	Yahoo Inc.

Table 4.1 List of Ticker Symbols Used for Back Testing



Generated

Ticker Symbol \ Method	TSQA3	TSQA4	TSQA5	Moving Average
MSFT	48	**102**	54	96
GOOG	42	**118**	68	83
ORCL	57	**97**	75	70
IBM	54	**121**	82	92
YHOO	56	66	59	68

Table 4.2 Performance Comparison In Terms of Total Number of Trade Signals

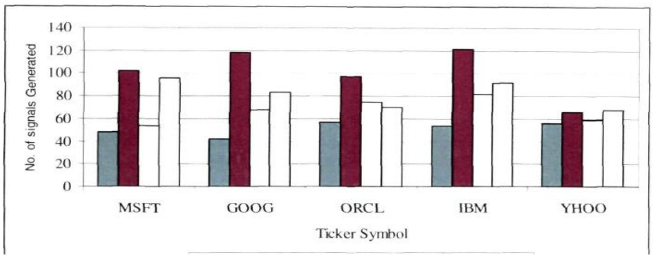


Figure 4.1 Graphical view of the comparative performance in terms of totalnumber of trade signals generated

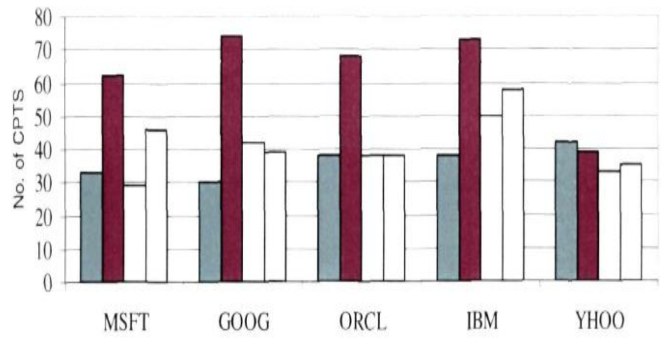


Figure 4.2 Graphical view of the comparative performance in terms of total number of CPTS generated

The execution as far as rate of CPTS produced by TSQA3, TSQA4, TSQA5 and the seat stamp specialized pointer Moving Average Crossover (MAC) calculation for the stocks considered for back testing is given in the Table 6.3. A calculation which creates greatest rate of CPTS is ordinarily good. The TSQA3 is creating a greatest of 75.00% and at least 68.75% CPTS. TSQA4 is creating a most extreme of 70.10% and at least 59.09% CPTS. The most extreme rate of CPTS produced by MAC is, obviously, not as much as the base rate of 68.75% CPTS created by TSQA3. In this point of view both TSQA3 is beating over the various analytics. In the mean while, TSQA4 is likewise doing admirably by creating over 60% CPTS in the vast majority of the stocks under examination and the same can be seen fi-om the graphical introduction appeared in Figure 6.3.

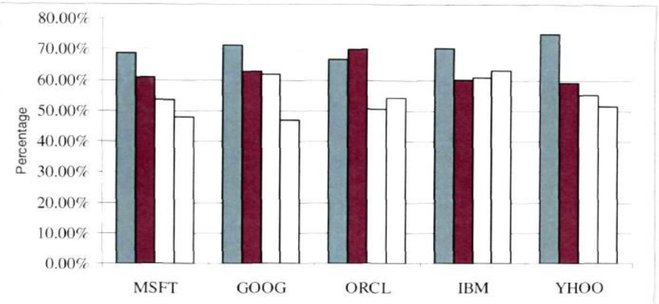


Figure 4.3 Graphical view of the comparative performance in terms of percentageCPTS generated

CONCLUSION

This is not to guarantee that this calculation would make anybody rich, yet it might be helpful for academicians, professional specialists and individual traders for exchanging examination. This may bring up the accompanying issue.

- How much would one be able to lose before the real pick up?

The appropriate response is that it is conceivable to pick up to 70.44% (by utilizing TSQA3) of the normal time, yet at the same time the odds to lose remain a reality. The fundamental commitment of this postulation was accomplished by showing that it is conceivable to extricate significant and solid models of money related time arrangement from an accumulation of well known specialized markers by methods for these combinational calculations.

REFERENCES

[1] Abe, M. “Theories and model building for CRM data analysis RF analysis based on consumer behavior theories.” Ryutsu oho (Distribution Information), December, 2017. (in Japanese)

[2] Agrawal, R. Imielinski, T. and Swami, A. “Mining Association Rules Between Sets of Items in large Databases.” Proceedings of the 1998 ACM- SIGMODInternational Conference on Management of Data, Washington DC, USA, May 1998. 207-216.

[3] Ailawadi, Kusum L., Scott A. Neslin, amd Karen Gedenk. “Pursuing the Value-Conscious Consumer: Store Brands Versus National Brand Promotions.” JM, 65.1

[4] Anand, S. S. Bell, D. A. Hughes, J.G. “EDM: A General Framework for Data Mining Based on Evidence Theory.” *Data and Knowledge Engineering Journal*18.3 (1996): 189- 223.

[5] Anand, S.S. Patrick, A.R. Hughes, J.G and Bell, D.A. “A Data Mining Methodology for Cross Sales.” *Journal of Knowledge-Based Systems* 10.7 (1998): 449-461.

[6] Anon. “Data mining efforts increase business productivity and efficiency.“2008.www.bettermanagement.com/library/library.aspx? libraryid=5664.

[7] Anon. *Introduction to data mining and knowledge discovery*. Third edition. Potomac, MD: Two Crows Corporation, 1999. Print.

[8] Apte C. Liu B. Pednault E.P.D and Smyth P. “Business Applications of Data Mining.” *Communications of the ACM, Special Issue: Evolving data mining intosolutions for insights* 45.8 (2002): 49-53.

[9] Bart, Lariviere. Dirk, Van. and Den, Poel. “Predicting Customer Retention and Profitability by Using Random Forests and Regression Forests techniques.” *ExpertSystems with Applications* 29 (2005): 472-494.

[10] Bartlett M.S., “Test of Significance for Factor Analysis” 103-375-765 (Last edited on 2002/04/17 10:36:30 GMT-6), 1950.

[11] De Wulf, Kristof, Gaby Odekerken-Schröder, and Dawn Iacobucci. “Investments in consumer relationships: a cross-country and cross-industry exploration.” Journal of Marketing, 65.October (2001): 33-50.

[12] Dennis Charles, Marsland David and Cockett Tony. “Data Mining for Shopping Centers- Customer Knowledge Management Framework.” *Journal of KnowledgeManagement*5.4 (2001): 368-374.

[13] Dhanpal R. Subramanian S. Gayathri and Jobin M Scaria. “Customer Retention using Data Mining Techniques.” *International Journal of Computer Applications*11.5 (2016): 32-34.

[14] Dick, A.S. &Basu, K. “Customer Loyalty: toward an integrated conceptualFramework.” Journal of the academy of Marketing Science 22.2 (1994): 99-113. Print.

[15] Dickson, John, &Albaum, Gerald. “A method for developing tailor-made semantic differentials for specific marketing content areas.” Journal of Marketing Research. 14 (1997): 87–91.

[16] Doran Patrick. “Information Architecture for Irish Grocery Retailers using Business Intelligence Tools.” M.Sc. in Computing (Knowledge Management) Dissertation, September 2007.

[17] Dunham, M.H. *Data mining introductory and advanced topics.* Upper Saddle River, NJ: Pearson Education, New Delhi, 2003. Print. ISBN: 81-7758-785-4, 2006.

[18] DzulijanaPopovic, BojanaDalbela “Churn Prediction Model in Retail Banking Using Fuzzy C-Means Algorithm.” *Informatica* 33 (2009): 243-247.Elovici Yuval and Braha Dan. “A Decision-Theoretic Approach to Data Mining.”*IEEE Transactions on Systems, Man and Cybernetics- Part A: Systems and Humans* 33.1 (2003).

[19] Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. “From data mining to Knowledge discovery in databases.” AI Magazine, 1996: 7(3): 37-54. Print.