ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 5, Issue 2 (Apr. - Jun. 2017)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2017)

An Analysis on Descriptive Data Mining For Finding The Prediction of Measure Actual Customer Behavior Using Big Data

[I]Sushmita Acharya, [II]Manish Sahu
[I]M.Tech. Scholar E. Security, [II]Assistant Professor
[I,II]Dept. of Information Technology Shri Shankaracharya Engineering College, Bhilai (C.G.), India

Abstract
Consumer loyalty assessment permits an organization to comprehend the issues, or key drivers, that cause fulfilment or dissatisfaction with an administration encounter. Thus it can center its time and assets all the more successfully. For associations new to this process, the primary stages needs a survey of what the administration gives, where it sits in setting with other related administrations in clients' psyches, who its clients are and what data about the client experience is as of now reachable. This, subjective research ought to be directed with clients and staff to emphasize key issues that the study should catch. Now choiced clients meeting and techniques to be utilized are selected. Customer loyalty and estimation issues have vital parts for organizations in giving and keeping up a sensible preferred standpoint. In this paper we have approached many papers and way to find the customer behavior using Big Data Technology.

Keywords
Technology of Big Data, Online Behavior of Customer, Forecasting, Social Network Site etc.

Introduction

In two ways contrasted with customary logistic model, Big Data analytics provides superior hand.

Big Data analytics draws the talented utilization of basic design connected to massive volume of information which would be too large for the customary methodical environment. Looking in to propose, a simple basic calculation with massive volume of information's is more specific in compare to complex calculations with little data information. The capability of applying it to gargantuan measures of information's devoid of trading off execution produces the focused edge. Second, Big analytics hints at the modification of the model.

Gradually in a progressive way, examination calculations are provided by database administration framework (DBMS) sellers. To pull away from the pack organizations must go well before what is provided and improve by using more up to date, more complicated accurate examination.

In general the benefits of Big Data over any other technology are faster better decision, cost reduction, new product and service.

We develop a calculation which meets both above considered focuses.

Here in figure 1 we show the concept of our research work using Big Data analytics.

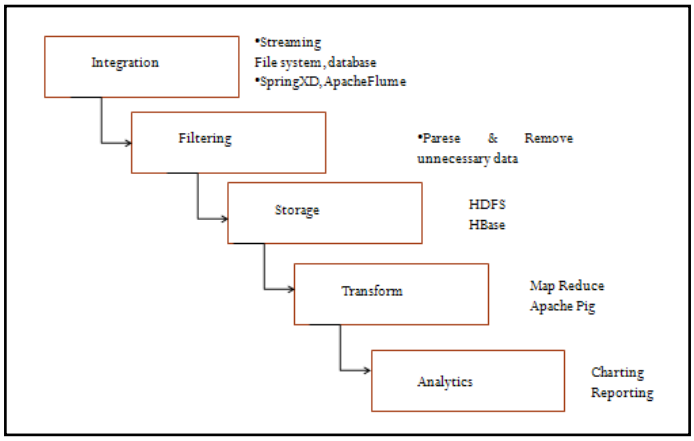

Fig.1: Our Research Work Steps

Big Data Technology

Big Data is getting to be a prominent amongst the mainly discussed innovation inclines these days. The authentic test with the vast association is to get most intense out of the information effectively accessible and prophesy what sort of information to gather later on. Instructions to take the present information and formulate it important that it gives us accurate knowledge in the past information is one of the key dialogue focuses in a huge number of the official gatherings in associations. With the flash of the information the test has gone to the following level and now a Big Data is becoming into the truth in many associations. Here below shown a hadoop big data platform to big data analysis.

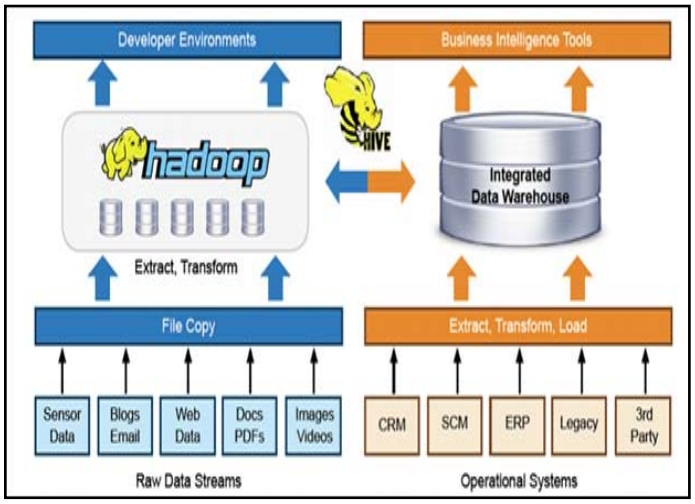

Fig. 2: Hadoop Big data Platform

We get a kick out of the chance to contrast gigantic information and the Rubik's 3D shape. we trust they have abundant likenesses. Much the same as a Rubik's solid shape it has a broad range of arrangements. Give us a chance to picture a Rubik's solid shape unravelling challenge where there are numerous specialists involvement. On the off chance that you take five Rubik's 3D square and stir up a similar way and offer it to five different master to unravel it. It is entirely imaginable that all the five individuals will

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

understand the Rubik's shape in divisions of the seconds however in the event that you pay consideration on the same, nearly you will see that in spite of the fact that the ultimate result is the same, the course taken to understand the Rubik's 3D square is not the same. Each master will begin at a better place and will attempt to resolve it with various techniques. Some may explain one shading first and others may deal with another shading first. regardless of the fact that they take after a similar kind of calculation to clarify the bewilder they will begin and end at a better place and their moves will be varied at various events

Literature Survey

1. Web Site Visit Forecasting Using Data Mining Techniques

Identifying relation between large amounts of data from various fields like scientific research, business planning, traffic analysis etc.

"Data Mining" techniques is used. This research will help to search data mining techniques applicability in web site visit prediction, domain. We concentrate here on time series regression techniques that will be used to analyse to forecast time dependant data points. At what way those techniques be applied to forecast web site visit will be explained."

In modern technologically changing world targeting the business for right customers at right time, ability of predicting web site visiting patterns have a significant value for every site owner. The available Business Intelligence (BI) tools are so expensive that small and medium size companies are unable to afford it. Available BI tools but does not support forecasting features focused only on analysing user behaviours on web site and log related result. This research is targeted on web site administrators and owners to helps future prediction of web site visit on marketing strategies. Companies need to know the visiting history for their sites to plan for new product; upgrades etc. On the other hand density of forecast visitors will help to allocate or deallocate servers. The above requirements exist in website visit prediction domain motivated me to provide a suitable solution. The aim of this research work is to apply a suitable forecasting technique to predict website thus result derived through forecasting will assist web site owner in-

- Predict total number of visitors with in next week time.
- Predict number of visitors n a given day (Sunday, Monday etc.) with in next week time

The objectives for achieving the aim are-

- Identifying and investigating data mining technique that can be used for time series data forecasting.
- Identifying and applying suitable pre-processing techniques to clean the data.
- How the identified technique can be applied in web site visit prediction domain.

2. Analyzing customer behaviour using data mining techniques: Optimizing relationships with customer

A simple listing of marketing contacts can no longer achieve customer satisfaction, but detailed information of customer such as past purchase and prediction of future purchase are required to know his behaviour.

Data mining finds hidden information in the database.

Data mining technologies and techniques for recognizing and tracking pattern within data helps business through layers seemingly unrelated data for meaningful relationship.

Here we discuss a business and technological over view of data mining and outline how we can achieve optimum customer profitability through data mining application.

The aim of this paper is to find out the role of data mining in customer focus business strategy. Today customer can compare things with a click of mouse. Thus customer relationship is most

valuable asset of companies.

3. Predicting customer behaviour: The University of Melbourne's KDD Cup report

KDD cup 2009 competition with our ideas& methodology for problem modelling is discussed here. The important stages included aggressive nonparametric feature selection, careful treatment of categorical variables and tuning a gradient boosting machine under Bernoulli loss with trees Orange Labs France organized the KDD cup 2009. The data consisted of information about telecom customers with 15000 predictor variables The competition involved producing binary classifiers for three types of consumer behaviour -

- Churn, which is whether someone ceases to be a customer.
- Appetency, being the propensity to buy a service or product, and
- upselling, where a more profitable or additional service is sold to a customer.

Competitors were given a training set of 50,000 observations with additional 50,000 in test set, which organisers used for model evaluation. The predictive accuracy was measured by the area under the ROC curve (AUC) which integrates sensitivity over all possible specificities of the model. The average of the AUC of three separate classifications tasks was used to rank competitors. A reduced data of 230 variables was also available which is not taken into consideration by our team for primary entry. There was a fast component with predictions for the test data due within 5 days of the full data being released, and a slow component to submit prediction in 5 weeks.IBM Research produced the best model for both fast and slow components. As per rules of the competition no team could win both parts, hence university of Melbourne team won first prize in the slow component, as it was second best.

Table1 : Shows final results of both IBM and University of Melbourne. Our model based entirely on large data set and smaller data set provided has not been taken in to consideration

Table 1: Final model performance for IBM research and the University of Melbourne

Team	Model			
	Churn	appetency	upselling	average
IBM Research	0.7651	0.8819	0.9092	0.85206
Univ. Melbourne	0.7570	0.8836	0.9048	0.84847

The data set provided in KDD cup 2009 is typical of many existing data mining problems There are large number of observation which enables many signals to be resolved through the noise allowing complex models to be fit.

As companies and other organisation are able to collect huge information's regarding customers so the number of predictors are large in the KDD cup 2009.However many of these predictors hold little or no useful information's , so the ability to keep out superfluous variables from final analysis is important. Out of

the predictors many have missing values, some are continuous and some are categorical. Out of the categorical predictors some have large number of levels with little exposure . For continuous variables the distribution among the observation can have extreme values. There is a potential for significant interaction between different predictors. Finaly responses are highly unbalanced. Only 7% of the upselling observations were labelled "1" All these factors need to be taken in to account to produce a satisfactory model. Section 2 to 4 details the stages of our modelling for the KDD cup while section 5 makes some comments on the computational resources used.

4. A Study of Customer Behaviour Through Web Mining

Achieving customer satisfaction is no longer satisfied, with a simple analysis of marketing contact. Web mining is extraction of remarkable and potentially useful patterns and latent information's from web activities and web documents by using data mining technology.

In the modern E-commerce era the irrevocable most important challenge is to comprehend to a maximum possible of the customers' desires, wants, purchasing patterns to ensure competitiveness in the E-commerce market. Any information related to customer behaviour is of vital importance in the highly competitive nature of E- commerce market. Hence to reduce competition and simultaneously to increase business profit, web mining can be used to know those obvious data which have vital value. This paper aims to study classification of web mining to extract customer behaviour through the used techniques and processes & explore the application of web mining in E-commerce to increase profit.

5. Analyzing Customer Behaviour through Data Mining

In India retail industry is yet to grow. To attract customer's monetary promotion such as discount, coupons, rebates etc. in retail sector are important.

With the optimum spread use of sales promotions it has become important for the organization to know well such practices and understand the challenges. `

This research help organization in identifying the set of customers who are attracted when sales promotion kinds are issued this will also enable the organization to formulate new policies to attract new customers.

Large number of retail centers are opened to provide quality products at low price thus high competition and narrow profit margins have pushed retailers implementing data warehouse. In retail chains stores huge amount of data collected over years. It enables an industry to focus on the most important information in the database and allows retailers to make more knowledgeable decisions by predicting future trends and behaviour.

Data mining is used here to improve the sales in the business.. Data mining is the extraction of hidden information from huge databases to predict customer behaviour and help organization to have large market share and cultivate loyal customers.

This research paper presents decision tree algorithm as a simple and efficient tool for analyzing the customers 'details and distribution of coupons for higher sales and larger market share

Proposed methodology

Big Analytics gives competitive advantage in two ways compared to the conventional analytical model. First, Big Analytics describes the efficient use of a simple model applied to huge data that would be too large for the traditional analytical situation. Research suggests that a simple algorithm with a huge data is more accurate than a complicated algorithm with little data. The capability to apply it to huge amounts of data without compromising performance generates the competitive age. Here Figure 1 shows the actual ideology of our research work. Analysis algorithms are provided straight way by database management system (DBMS) vendors. Companies must go well ahead of what is provided and innovate by using newer, more complicated statistical analysis so as to pull away from the pack. To satisfy both above defined points we have developed an algorithm.

Experiment Results

Our research will focus on the analysis of customer behaviour of the web site in this our main approach is to find out the maximum shares among the different parameter like business, technology, life style, positive key words, negative keywords etc. i.e. shows the user responsiveness towards the above parameter through this we can conclude our research that the maximum share gain parameter will be the most popular keyword among the online news website for the popularity. With the help of the above research we can help many of the marketing enterprises, advertisement agency etc for marketing and advertisement purpose. Here below the experimental setup and the result are given.

Sample Data and Test Data

We are thanks full to mashable.com who gave us precise data set for our research. Mashable is a global, multi-platform media and leisure business enterprise. Powered by means of its very own proprietary generation, Mashable is the move-to source for tech, virtual culture and entertainment content material for its committed and influential audience around the world. At Mashable we're seeking out proficient and bold employees prompted by the digital revolution. Click this button to view and practice to open positions right here on Mashable's activity Board.

Fig. 3: mashable.com for our sample data set

Mashable hosts numerous industry-main events in the course of the year, bringing together social influencers to percentage thoughts, make connections & have fun. Mashable is a worldwide, multi-platform media and leisure company.

Powered with the aid of its own proprietary era, Mashable is the go-to supply for tech, digital tradition and entertainment content for its devoted and influential target market around the world.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Mashable is deeply devoted to Social suitable. From day by day insurance about how human beings are changing the sector via technology to our annual Social properly Summit, we consider we've got a duty with our community to make the world a higher area. Our undertaking is to offer voice to hundreds of thousands who are looking to make their lives better through digital innovation. Mashable is social. Like, really social. With 27 million followers and counting, the subjects we cowl and the approaches readers can access our content are constantly growing. Want to realize how you can connect with Mashable? Take a look at out all the locations you may find us below, and click right here to enroll in our newsletters.

Experiment Setup

For our research we're going to be used check pattern statistics as mashable on-line news records to be had in mashable.Com. It is freely to be had for check and studies. For writing java program we're using notepad++ v6.Nine, Java improvement package version is JDK 1.7 for java environment and hadoop 2.3 for windows, and windows eight.1 operating system. Right here in this web website "http://www.Codeproject.Com/Articles/757934/Apache-Hadoop-for-home windows-Platform" the installation manner is given, observe the ones steps for setting up hadoop environment.

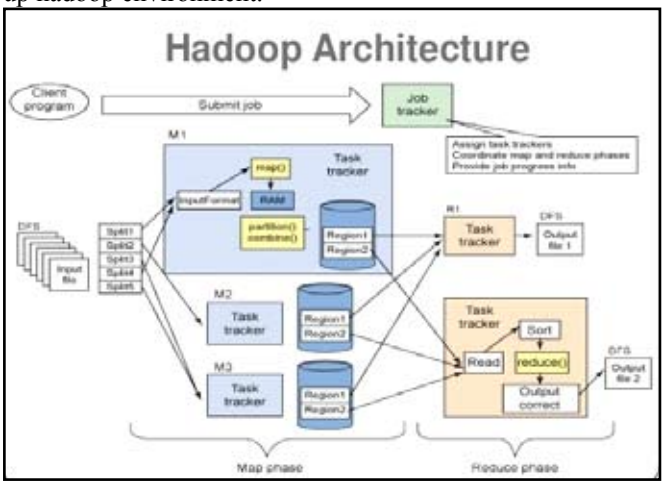

Fig.4: Hadoop Architecture

Open cmd prompt in admin mode and start hadoop demon using C:\Hadoop-2.3-master\sbin\start-yarn command snap shot is follows. After this start dfs by using C:\Hadoop-2.3-master\sbin\ start-dfs command the snapshot in figure

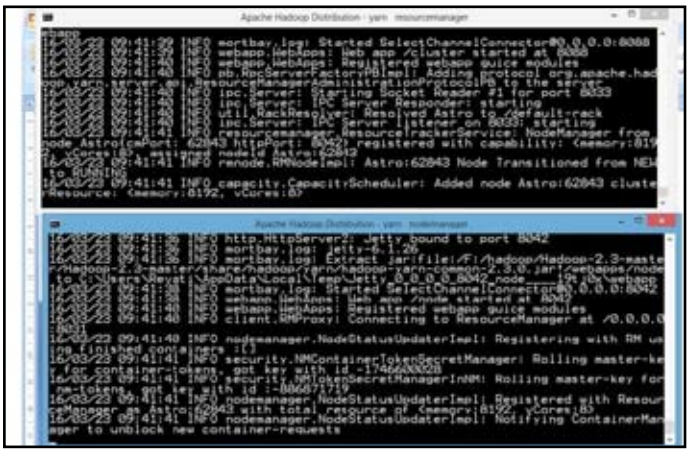

Fig.5: Yarn Hadoop

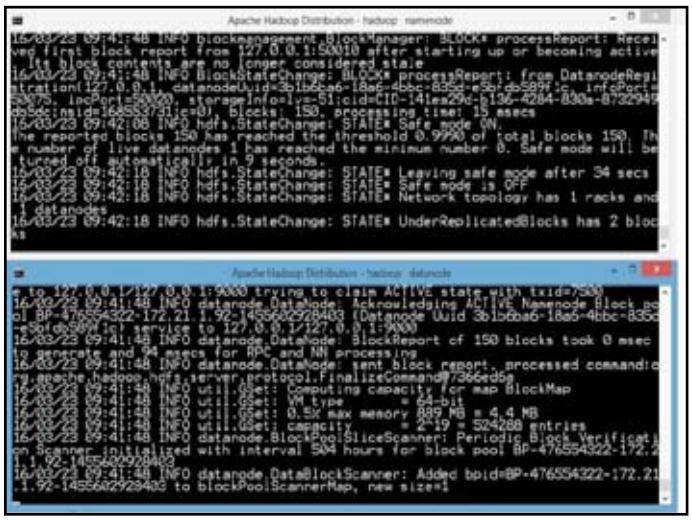

Fig.6: DFS Of Hadoop

Experiment Output

After running the above program using different parameter such as life, technology, business, world and entertainment with comparatively shares, we find the result with the table given below And the graph can be find in given below figure

Table Output result

Parameter	Shares
Life	8156765403
Entertainment	6574836627
Business	8787877832
Technology	8988778786
World	6775565665

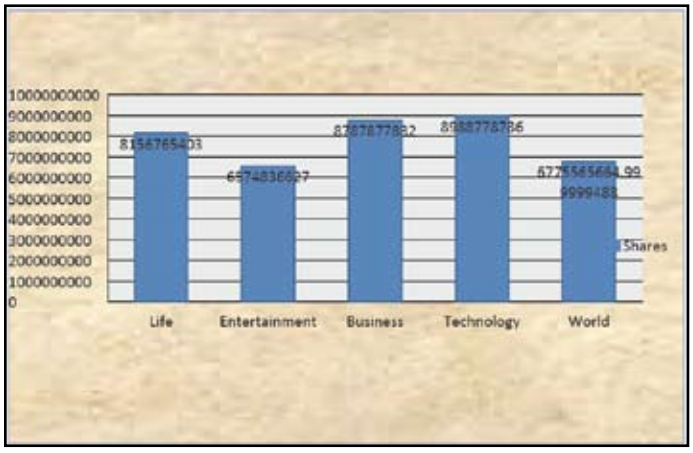

Fig. 6: Study of different parameter with their shares in website (mashable.com)

References

[1] *A Study of Customer Behaviour Through Web Mining, Gan Teck Wei1, Shirly Kho2, Wahidah Husain3, Zurinahni Zainol4, 1,2,3,4 School of Computer Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia, Journal of Information Sciences and Computing Technologies ISSN 2394-9066, SCITECH RESEARCH ORGANISATION Volume 2, Issue 1, February, 2015.*

[2] *Predicting customer behaviour: The University of Melbourne's KDD Cup report , Hugh Miller, Sandy Clarke,*

Stephen Lane, Andrew Lonie, David Lazaridis, Slave Petrovski, Owen Jones, JMLR: Workshop and Conference Proceedings 7: 45-55 KDD cup, 2009.

[3] Analyzing Customer Behaviour through Data Mining, Sandeep Kumar, Rakesh Kumar Arora, International Journal of Computer Applications Technology and Research Volume 4– Issue 12, 884 - 888, 2015, ISSN: 2319–8656.

[4] Analyzing customer behaviour using data mining techniques: Optimizing relationships with customer, Aditya Kumar Gupta, Chakit Gupta**, * Lecturer, School of Management Sciences, Varanasi**Research Scholar, Indian Institute of Professional Studies, Lucknow, Vol. VI, No. 1; June, 2010.*

[5] Web Site Visit Forecasting Using Data Mining Techniques, Chandana Napagoda, Department of Computer Science, University of Moratuwa Katubedda, Sri Lanka, IJSTR, VOLUME 2, ISSUE 12, DECEMBER 2013, ISSN 2277-8616.

[6] Data mining techniques: A source for consumer behaviour analysis, Abhijit Raorane1 & R.V.Kulkarni2 1Department of computer science, Vivekanand College, Tarabai park Kolhapur, 2Head of the Department, Chh. Shahu Institute of business Education and Research Centre Kolhapur.

[7] Data Mining as a tool to Predict the Churn Behaviour among Indian bank Customers, Manjit Kaur, Dr. Kawaljeet Singh, Dr. Neeraj Sharma, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 1 Issue: 9 ISSN: 2321-8169, 720 – 725.

[8] Prediction of churn behavior of bank customers using data mining tools, Dr. U. Devi Prasad, S. Madhavi, Business Intelligence Journal - January, 2012 Vol.5 No.1

[9] Customer Data Clustering Using Data Mining Tecnique, Dr. Sankar Rajagopal, Enterprise DW/BI Consultant Tata Consultancy Services, Newark, DE, USA, JDMS, Vol.3, No.4, November 2011

[10] Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade, M'hammed Abdous1, Wu He2, Cherng-Jyh Yen ,1Center for Learning and Teaching ,2Information Technology Department of Information Technology & DecisionSciences, 3Educational Research and Statistics Department of Educational Foundations and Leadership, Old Dominion University, Norfolk, Virginia, 23529, USA,*Corresponding Author, C.-J. (2012), Educational Technology & Society, 15 (3), 77–88.*

[11] Gauging Heterogeneity in Online Consumer Behaviour Data: A Proximity Graph Approach Natalie Jane de Vries1,2, Ahmed Shamsul Arefin1,2, Pablo Moscato1,2, 1 School of Electrical Engineering and Computer Science, The University of Newcastle, Australia , 2 The Priority Research Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine, 2014 IEEE Fourth International Conference on Big Data and Cloud Computing.*

[12] A Data Mining Based Approach to a Firm's Marketing Channel, Esra Kahya Ozyirmidokuza,, Kumru Uyarb, Mustafa Hakan Ozyirmidokuzc, aErciyes University, Computer Technologies Department, Melikgazi, Kayseri, 38039, Turkey, bNuh Naci Yazgan University, Production and Marketing Department, Kayseri, 38170, Turkey, cBosch Thermotechnic, Ankara, 06810, Turkey, ScienceDirect, 22nd International Economic Conference – IECS 2015 "Economic Prospects in the context of Growing Global and Regional Interdependencies", IECS 2015*

[13] Sushmita Acharya, Manish Sahu, "A Survey Paper on Descriptive Data Mining for Finding The Prediction of Measure Actual Customer Behavior Using Big Data" M.Tech. Scholar E. Security, Dept. of Information Technology Shri Shankaracharya Engineering College, Bhilai (C.G.), India, Assistant Professor, Dept. of Information Technology Shri Shankaracharya Engineering College, Bhilai (C.G.), India, sushmitaacharya.83@gmail.com, manishsahu1@gmail.com, IJARCST , Vol. 4, Issue 4 (Oct. - Dec. 2016).