ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Analysis of E-Customers Behavior Using Naïve Bayes Algorithm

Prof. Gajanan P. Arsalwad[1], Angela A. Dhanawade[2], Rupali J. More[2], Pranali A. Kulkarni[2].

Assistant Professor, Department of Information Technology, Trinity College of Engineering and Research, Pune, India[1]

Under Graduate Student, Department of Information Technology, Trinity College of Engineering and Research, Pune, India[2]

ABSTRACT: Data Mining is where the available data is tested so that a new information is obtained. Clickstream data is a users activity on Internet and lets you know about the behavior of user. In this paper, we make use of clickstream data. Taking the users session history generation and the user activity Naïve Bayes algorithm is applied. This algorithm considers a few parameters and based on that the analysis is performed. Result of this analysis can be used in Customer Relationship Management and Business Intelligence.

KEYWORDS: Data mining, clickstream, Web mining.

I. INTRODUCTION

The marketing of products or services can be done using digital technology so that it will reach out to consumers. The advantages of using the digital market is that it offers more choices, lower prices, easy search and access to online customers. Thus, this is why the digital market is expanding day by day. As a result customers behavior patterns are gaining more importance in buying the things online[2]. From the 1990's the development of electronics market entirely changed the way customers perform online transactions[3]. Traditional markets have become an alternative source because of this digital market. Behavior of customer is the study of when, why, how and where people do or do not buy a products. Understanding the customer's needs is essential while considering for an online e-commerce application. Another application for data mining is Web mining which helps in discovering usage patterns and behaviors from the web data[5]. Clickstream data is an example of web mining. This data will help in serving the clients requests and also improve the sales of the business.

Web usage mining has gained much attention from research and e-business professionals and it offers many benefits to an e-commerce web site such as:
• Targeting customers based on usage behavior or profile (personalization)
• Adjusting web content and structure dynamically based on page access pattern of users (adaptive web site).[5]

With the increase in all of thesei.e the web usage, clickstream data and the mouse movements of the online customers, we have implemented this model to analyse the behavior of the customers.
In our model we will be dynamically generating the web data of customers and analysis will be performed based on some attributes that are defined in the dataset used section. According the analysis will be performed.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

II. OVERVIEW

1. Existing System

From the existing system [1], the statistics are as follows: A confusion matrix scorer is applied to calculate accuracy. Table 1 presents the confusion matrixes and accuracy statistics for both decision tree and artificial neural network analysis. Table 1 reads that overall accuracy for prediction is 90.42%. Accuracy at predicting whether customer will buying is 96.3%. When it comes to predict whether customer will pay and leave then accuracy is 40.2%. Since, F measure is 0.947 for predicting whether a customer will leave without paying; the rules generated by decision tree analysis may be used for some business purpose.

TABLE 1. CONFUSION MATRIX AND ACCURACY STATISTICS FOR DECISION TREE ANALYSIS.

Left Before Paying \ Prediction (Decision Tree)	yes	no	Accuracy	Precision	Specifity	F-Measure
yes	2690	103	96,30%	93.2%	40.2%	0.947
no	196	132	40,20%	56.2%	96.3%	0.469
Correct classified:2822						
Wrong classified:299						
Accuracy:90.42%						
Error:9.58%						
Left Before Paying \ Prediction (Neural Network)	yes	no	Accuracy	Precision	Specifity	F-Measure
yes	2720	104	96,30%	96%	62%	96.2%
no	113	184	62,00%	63.9%	96.3	62.9%
Correct classified: 2.904						
Wrong classified: 217						
Accuracy: 93.047%						
Error :6.953%						

2. Drawbacks of Existing System.

In the existing system, it is only practicability of the work done. Although the system in existing have used only some data features described in the dataset.Using the neutral network analysis and the decision tree analysis together can become a complicated task.Also offers for the interested customers are not given instantly.[1]

3. Need of Naïve Bayes Algorithm

Using this algorithm, we can use multiple features as described in the dataset at once and we can predict whether a person will purchase the thing or not. Naive Bayes gives accurate results for large datasets. Offers to interested customers are given instantly and can also provide analysis on a particular area.

III. THE DATASET USED IN THE STUDY

Dataset used in the existing system will be same[1]. Dataset has a server side program to collect clickstream data from the company's web server; at the same time, another java script program has been used to collect data from client side. Data attributes we have collected and used in the study are as follows:

Special day: If it is one week or earlier than an official or religious day such as Christmas, Independence Day etc. this parameter takes true value as 1, otherwise it is 0.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Day: represents the day of week: Sunday, Monday, and Tuesday etc.
Period of day: We have four periods for this variable; morning, afternoon, evening and after midnight. So, morning: 1; afternoon: 2; evening: 3 and after midnight: 4.
Time spent on the side: This variable includes total time spent on the site. It is calculated as seconds.
Search: If the customer searches a certain product on the site by entering keywords, this variable takes true value labelled with 1; otherwise it is labelled 0 (false) by default.
Category of search: Products have been categorized into 4. Skirt, jeans, shorts and pants are one group labelled as 1; shoes, boots, sandals are grouped and labelled as 2; dress, jacket cardigan, overcoat, sweater etc. are 3 and underwear products are 4.
Number of items in basket: It shows the number of different items in the online shopping basket. If the customer has two identical products this is counted as one. There must be at least one difference, such as color or size between two products, to count them two separate items.
Discounted Item in the Basket: The e-commerce company makes promotion campaigns or applies discounts. This variable identifies if the item in the basket is a promotional or discounted one. If there is at least one discounted item in the basket this variable takes 1(True), otherwise 0 (False).
Product category of the item in the basket: There are five categories in this field: Female, male, unisex, child (girl), child (boy).
Item add time: It shows the time (in seconds) of the first item added to basket. If the basket is empty it takes 0 value.
Amount of clicks: Number of the items clicked. Menu item clicks are not counted. We counted only the click made on products.
Click No: This shows the order of the click made by the customer. It takes values as 1st, 2nd, 3rd, etc.
Clicked item: Items are labelled as in category of search: Skirt, jeans, shorts and pants are one group labelled as 1; shoes, boots, sandals are grouped and labelled as 2; dress, jacket cardigan, overcoat, sweater etc. are 3 and underwear products are 4.
Click time: It shows the time (in seconds) when the product is clicked. For example if the user clicked his/her first item to examine on the 100th seconds on his/her visit, 100 is attained to this variable.
Source: It represents the source where e-customers come to the site from. This may be a search engine, another site or promotional mails sent by the company. If the customer is coming from a search engine it takes 1, if s/he is coming from a promotional mailing, it takes 2 and all other sides or sources are labelled as 3.
Left without purchase: If the customer checks out properly by making payment this variable takes false value labelled with 0; otherwise it is 1 (true) by default.

IV. MATHAMATICAL MODEL

In our model, we have used set theory as mathematical model.
Let s (be a main set of) ≡ {SDB, LDB, C, A, S, MR, AO}
where,

- SDB is the copy of the server database. This database is responsible for storing user information related to cloud interactions. Database will contain the session history of the user, the number of click count, number of items added to cart etc.
- LDB is a set of local database that a user owns. It consists of data tables having data items related to the products and their sales transactions.
- C is a set of all clients using the server database and mining services from the server. And (c_1 , c_2 , c_3, ...c_n) Є C.
- A is a set of algorithms applied on the input data to get mining results. In our project , we have used Naïve Bayes as the algorithm. As there are number of data which has to be used for analysis from dataset. Naïve Bayes enables you to used a number of parameters so that max probability results can be shown.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

- S is the server component of the system. The server is responsible for registering, authenticating and providing associations to the end user.
- MR is a set of mining rules that are applied on the input dataset provided by the client from his LDB. And (mr_1 , mr_2 , mr_3, ...mr_n) Є MR
- AO is a set of associations that are extracted from the input and a form the output of the system.

V. PURPOSED SYSTEM

- **Architectural Diagram.**

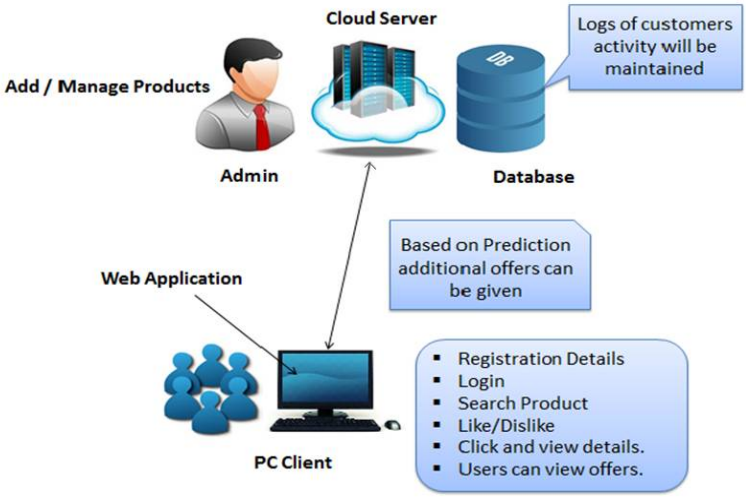


Fig. Architectural Diagram

From the above figure, you can illustrate that there are three models i.e the admin module, the client module and the cloud server. The clients will log in to the site and perform search product, like/dislike products, check and view details of products etc. Based on the clients browsing history, the cloud model performs analysis on it. The admin will give offers based on the prediction. Logs of customers activity will be maintained in the database.

1. Admin Module

In the admin application, there are mainly two phases. Firstly, the add and manage products where admin is able to add, delete and manage products. Secondly, the analysis part where the actual algorithm will be implemented. He will also be able to view the prediction results. There is connectivity between the admin application and the database. Based on the prediction results offers will be given to the individual interested customers. Admin will be able to execute queries on the database for managing the products such as add, delete, update.

2. Client Application Module

This module will be developed with the help of AWT/Swing. Swing is a part of Oracle's Java Foundation Class(JFC) which provides API for graphical user interface. Swing provides more sophisticated set of GUI components and also provides a native look and feel that emulates them across several platforms. The client module will request for the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

services from the server. Clients will get registered to the application. After that when he gets logged in into the application, he will start searching for some products. He will search the products based on the category, name, like/dislike, rating of a particular product. If the customer is interested in buying any product, he will view some more details about the product. Instantly the customers will get offers on the product if he views the same category of products again and again.

3. Cloud Server Module

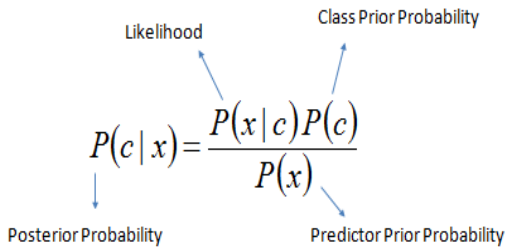
The cloud server will be having GlassFish server. GlassFish server is an open source server and also based on server. GlassFish uses a derivative of Apache Tomcat as the servlet container for serving Web content. Server is responsible for users authentication and also provide the services requested by the users. Maintaining the users logs based on the clicks and the activity. JDBC connectivity will be used for connecting the database where the database will be MySQL. Server will apply prediction on logs to analyse the User behaviour and will prediction if item can be purchased by the customer or not.

VI. ALGORITHM

Naive Bayes is a simple technique for constructing classifier models that assign class labels to probleminstances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. Naive Bayes models uses the method of maximum likelihood.

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Where,
-$P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
-$P(c)$ is the prior probability of *class*.
-$P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
-$P(x)$ is the prior probability of *predictor*.

VII. CONCLUSION

In this study, application makes analysis of customers behavior based on different attributes. Naive Bayes algorithm helps for considering different attributes at a single time. Also it produces accurate results for large datasets. This analysis will directly produce increase in sells. Users will get more and more offers with this. So this application helps shopping effective and easy.

VIII. FUTURE WORK

This implementation can be done on Android platform as well for most users these days carry out their online tasks on smart phones. Also the algorithm for carrying out the same task can be done using a hybrid approach. If the number of users for such application goes on increasing this also can be implemented on Hadoop platform.

REFERENCES

[1] GokhanSilahtaroglu, Hale Donertasli,2015. Analysis and Prediction of E-Customers Behavior by Mining Clickstream Data.
[2] Thompson S., H. Teo, 2006. To buy or not to buy online: adopters and non-adopters of online shopping in Singapore. Behaviour and Information Technology, 25(6), 497-509.
[3] Aikaterini C. et.al., 2013. Online and mobile customer behaviour: a critical evaluation of Grounded Theory studies , Behaviour and Information Technology, 655-667.
[4] Schaefer, Kerstin; Kummer, Tyge-F.,2013.Determining The Performance Of Website Based Relationship Marketing ,Expert Systems With Applications,7571-7578.
[5] Hu, XH; Cercone, N, A., 2004. Data Warehouse/Online Analytic Processing Framework For Web Usage Mining And Business Intelligence Reporting, International Journal Of Intelligent Systems,585-606.
[6]Al-Zaidy, Rabeah; Fung, Benjamin C. M., Youssef, Amr M., 2012. Mining criminal networks from unstructured text documents, Digital Investigation ,147-160.
[7] Domingues, Marcos Aurelio; Soares, Carlos; Jorge, Alipio Mario,2013.Using Statistics, Visualization And Data Mining For Monitoring The Quality Of Meta-Data In Webportals., Information Systems And E-Business Management,569-595.
[8] Vicari,Donatella;2014.Alfo, Marco, Model based clustering of customer choice data, Computational Statistics and Data Analysis,71 Special Issue,3-13.
[9] Berthold MR,2008.KNIME: the Konstanz information miner In Data analysis, Machine Learning And Applications, Springer-Verlag,319-326.
[10] Yan LI Bo-qinFENG,Yan LI Feng WANG, 2009. Page Interest Estimation Based on the User's Browsing Behavior, Second International Conference on Information and Computing Science, 258-261.
[11] VinayGautam, Vivek Gautam,2013.UserBehavior Based Enhanced Protocol (UBEP) for Secure Near Field Communication, World Academy of Science, Engineering and Technology International Journal of Computer, Information, Systems and Control Engineering Vol:7 No:11, 853-863.