Scientific Research Publishing

Journal of Computer and Communications, 2019, 7, 42-53
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Data Mining Based Approach to Customer Behaviour in an Electronic Settings

A. Tope-Oke, C. A. Afolalu, O. Omofade

Department of Mathematical and Physical Sciences, Afe Babalola University, Ado Ekiti, Nigeria
Email: topeokea@abuad.edu.ng, catherinea@abuad.edu.ng, somofade @gmail.com

How to cite this paper: Tope-Oke, A., Afolalu, C.A. and Omofade, O. (2019) A Data Mining Based Approach to Customer Behaviour in an Electronic Settings. *Journal of Computer and Communications*, **7**, 42-53.
https://doi.org/10.4236/jcc.2019.75004

Received: March 3, 2019
Accepted: May 28, 2019
Published: May 31, 2019

Abstract

The understanding of customer incidents and behaviour is crucial to the success of any organization. Evidence from literature shows a prediction pattern of products to customer. These studies predicted product characteristics leaving out the customers characteristics. To address this gap, this study aims to design datamining system and implement it on an electronic commerce organization website. The customer information and history (clickstreams) from the electronic commerce website was used to predict the customers' behaviour. This will give meaningful and usable data patterns to organizations. Python programming language was used to design the datamining system, while PHP, HTML, and JavaScript were used for the e-commerce website. A brief description of the background of e-commerce and data mining, previous work of researchers who have worked on data mining in e-commerce settings, was reviewed and the relationship between their findings and this work was established. The data mining system utilizes consensus clustering technique and the clustering algorithm with a graphical-based approach. Furthermore, the interaction between the data mining system and the customer's dataset on an ecommerce website was defined. Quantitative evidence for determining the number and membership of possible customer behavioural clusters within the dataset was generated.

Keywords

Customer Behavior, Datamining, Ecommerce, Website, Electronic

1. Introduction

Data mining is the process of discovering meaningful pattern and correlation by sifting through large quantities of data stored in repositories. There are several tools for this data generation, which include abstractions, aggregations, summa-

rization and characteristics of data [1]. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. It concentrates on discovering and identifying rules that describe specific and sequential patterns within the data. Market basket analysis was one of the first applications of data mining. This technique identifies items that typically occur together in purchase transactions [2]. However, with the rapidly increasing volume of data in modern times, more automatic and effective mining approaches are required. Early methods such as Bayes' theorem in the 1700s and regression analysis in the 1800s were some of the first techniques used to identify patterns in data. After the 1900s, the explosion, ubiquity, and continuously developing power of computer technology, data collection and data storage were remarkably enlarged. Sequel to the increasing complexity of datasets, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering, genetic algorithms in the 1950s, Decision trees in the 1960s and support vector machines in the 1980s. Data mining therefore is the process of applying these methods to data with the intention of uncovering hidden patterns [3]. Data mining has been used for many years by many fields such as businesses, scientists and governments.

Electronic presence is often used to establish an organization's image. It has been used to promote goods and services, and to provide customer support on the internet and other electronic platform. The success of the electronic presence, however, affects and reflects directly the success of the organization in the electronic settings. As the Internet works on a basis of interchangeable data, there are new data sources that companies are required to exploit to achieve an improved performance. This data enables ecommerce managers to have a grasp of the business in ways that were not previously possible. Through an electronic store, it is possible to track much more data that is a direct result of how the customer interacts with the company. The clickstream data therefore is important in understanding customer behaviour and it is also the main source of information for the companies to adapt their service according to their electronic audience [4]. In order to understand customer behaviour, this study aims to design a customer mining system and implement it on an electronic grocery website base on their customer clickstreams. Clickstream data, or clickstreams, is the common terminology for the collection of Web logs that compose the session of a specific customer on the company website. These sessions contain information regarding the path that a customer took through the website's structure. Thus, this work will provide a system that will be used to capture, and interpret the patterns associated with Customer behaviour in an E-commerce setting using the clickstream analysis of the customer.

Customer Behaviour is the decision processes and acts of people involved in buying and using products. It is defined as the study of individuals, groups, or organizations and the processes they use to select, secure, use, and dispose of

products, services, experiences, or ideas to satisfy their needs and wants [5]. Customers who visit sites leave behind valuable information about themselves, analyzing these has the potential to improve business performance through the understanding of past and present customers to determine and identify future customers and their behaviour. These organizations also must learn to take advantage of what they have in large quantity which is the customer's data. [6] describes data mining as an approach to predict user behaviour in e-commerce sites. The core of their approach involves extracting knowledge from integrated data of purchase patterns of past users obtainable from web server logs to predict the purchase behaviour of future users. Data mining, Customer profiling and personalization allow organizations to study patterns through the mining of customer's data to plan their business activities and operations as well as develop new research on products or services for prosperous ecommerce. The outcome of data mining creates the possibilities for organizations to be able to track their customers purchasing patterns, demand trends and locations, making their tactical decision more effective for the advancement of their business [7]. The next section delves into related studies and how they informed this study.

2. Review of Related Studies

Xuesong and Kaifan [8] designed a web mining-based tourism e-commerce recommender system based on Association Rules mining and the Apriori algorithm. The study carried out an application example on five tour products, and they were required to find the frequent product set with strong relation by Association Rule mining. These products were then recommended to web Customers to increase the cross-sales of tourist website. Of all the five Products used in the study, the first, second and fourth products has strong association rules which helps the e-commerce seller recommend the products with strong relation to web customers to increase cross-sales of tourism e-commerce. This study is like their work since they both predict relationships. However, this work is focused on customer characteristics rather than product characteristics and we are making use of Cluster-based Similarity Partitioning Algorithm (CSPA) rather than apriori.

Saloni and Veenu [9] delved into the Importance of Domain Knowledge in Web Recommender Systems with the aim to offer personalized product recommendations to customers. In order to achieve their aim, the consumption demands of customers were studied by looking into the customer's intentions of purchase. The study made use of logical association and clustering algorithm. [9] revealed the basis of developing recommender systems in websites is a function of the Indicators. These Indicators were chosen for mapping the Customers to a cluster followed by some clustering algorithm. The indicators identified are Sequence of web pages viewed, time spent on web page, frequency of visits to a web page, rating of a web page. In addition to these, there were similarity indicators, similarity in sequence among different customers, similarity based on

frequency, similarity with respect to time spent. [10] worked on Item-Based Collaborative Filtering Recommendation Algorithms with the aim of recommending valuable products to target web customers. They constructed a model based on customer's behaviour such as registration information, evaluation data, and online shopping history and find out the nearest neighbour set, after which recommendation of valuable products to target web Customer was done using the nearest neighbour set.

The study made use of Collaborative Filtering to filter customers. The collection of customer's web information plays an important role and affects the recommendation results. The realization of customer based on collaborative filtering is divided into three steps. First, a matrix of evaluations on different products was constructed. It was discovered that the evaluation data can be acquired from server and expressed by a m x n matrix. Where, m stands for the number of customers and n the number of items. The items can be online products or web pages of a website. Secondly, the nearest neighbor set was calculated. Thirdly, recommendation result was generated for target customers. This study on the other is not a recommender system to customers but is designed for organization so that the customer data are translated into meaningful interpretations.

Evidence from literature shows a prediction pattern of products to customer [8] [9] [10]. These studies predicted product characteristics leaving out the customers characteristics. To address this gap, a datamining system was designed and implemented electronic commerce organization website. The customer information and history (clickstreams) from the electronic commerce website was then used to predict the customers behaviour. This will give meaningful and usable data patterns to organizations.

Consequently, this work or a follow up on it will be useful to managers who seek in-depth knowledge about their customer behaviour and how it can translate to more profit for business.

3. Objective of the Work

To address the gap identified above, the objectives of this study are:

1) To design a datamining system,

2) To implement it on an electronic commerce website by predicting customer behaviour.

4. Input and Output Design of the System

1) Analysis of Existing Systems

In relevance to an e-commerce setting, data mining is the analysis of historical customer activities transacted online on an e-commerce website, stored as data in a database [11]. The goal is to reveal hidden patterns and trends. Data mining software uses advanced pattern recognition algorithms to sift through large amounts of data to assist in discovering previously unknown strategic business information. In contrast to the related studies reviewed where the techniques

that were used for data mining were described by pure mathematical expressions; which are used to compute the values required to study the data for specific patterns. However, this work adopted the Object-Oriented Analysis and Design (OOADM) methodologies to design this model. Even though the existing system requires a reasonable measure of mathematical background to be deployed effectively for the mining of data, there is a requirement to reorganize the data in the database before the data mining operation commences.

2) Methodology

The proposed data mining system is a data visualization and data mining toolkit. The mining technique used is Consensus Clustering. Consensus clustering, also called aggregation of clustering (or partitions), refers to the situation in which several different (input) clustering's have been obtained for a dataset and it is desired to find a single clustering (consensus) which is a better fit in some sense than the existing clusterings. Consensus Clustering is a method that provides quantitative evidence for determining the number and membership of possible clusters within a dataset [12]. A graph-based approach, Cluster-based Similarity Partitioning Algorithm (CSPA) was adopted. In this algorithm, the similarity between two data-points is defined to be directly proportional to number of constituent clustering's of the body in which they are clustered together. The intuition is that the more similar two data-points are the higher is the chance that constituent clustering's will place them in the same cluster. The reason for choosing this algorithm is that CSPA is the simplest heuristic amongst others such as Hyper Graph Partitioning Algorithm (HGPA), and Meta-Clustering Algorithm (MCLA), its computational and storage complexity are both quadratic in n, and this algorithm is computationally less expensive. In this approach, a measure of similarity between a pair of data points can be estimated as the ratio of several clustering's in which the two data points are shared the same clusters to the total number of clustering's in the body. More precisely, the similarity between two data points x_i and x_j is defined as:

$$S_{ij} = S\left(x_i, x_j\right) = \left[cxc\right], \text{ where } c = 1, \; I\left(\pi c\left(x_i\right)\right) = \pi c\left(x_j\right) \tag{1}$$

where I is the indicator function. Thus, any similarity-based clustering algorithms can be applied to the similarity matrix S to find a consensus clustering of the body. First, the similarity matrix is computed as in Equation (1). Then, an induced similarity graph, where vertices correspond to data points and edges' weights to similarity measures, is partitioned into K clusters using METIS. This implies that the shortest the shortest distance (S_{ij}) from the data point is where a cluster lies. METIS multilevel approach has three phases with several algorithms for each phase:

1) Coarsen the graph by generating a sequence of graphs $G_0, G_1, \cdots, G_N$, where G_0 is the original graph and for each $0 \le i \le j \le N$, the number of vertices in G_i is greater than the number of vertices in G_j.

2) Compute a partition of G_N

3) Project the partition back through the sequence in the order of $G_N, \cdots, G_0$, refining it with respect to each graph.

The final partition computed during the third phase (the refined partition projected onto G_0) is a partition of the original graph. It will feature a visual programming front-end for explorative data analysis and interactive data visualization. In visualizing the data mining system result, Python open-source libraries for scientific computing such as NumPy, SciPy and Scikit-learn was used, while its graphical user interface will operate within a cross-platform Qt framework using PyQt Python bindings. The data mining system was written in python with C data types which is also known as Cython. The Object-oriented representation of the system architecture is illustrated in Figure 1, while the class diagram and sequence diagram of the system are shown in Figure 2 and Figure 3 respectively.

Figure 1 is a set of iterative processes that describe the system. The customer uses the ecommerce website to shop for products online. As the customer clicks on products, this click of the customer using the e-commerce website will be stored in a database. The data mining system based on web click stream technique is used to extract the appropriate patterns for a customer from the data base and then the company or organization uses to produce a customized experience for the customer.

Figure 2 describes the classes and attributes of the data mining system. The classes include customer, e-commerce website, database and the datamining system. The attributes on the other hand include: Sign-up(), Browse category(), Get password information(), track clicks(), Get customer information(), Access customer information from database() and a host of other attributes shown in Figure 2.

Figure 3, however, describes operations of the classes of the data mining system and their interrelationships. Section V describes the input and output data outlook for the system.

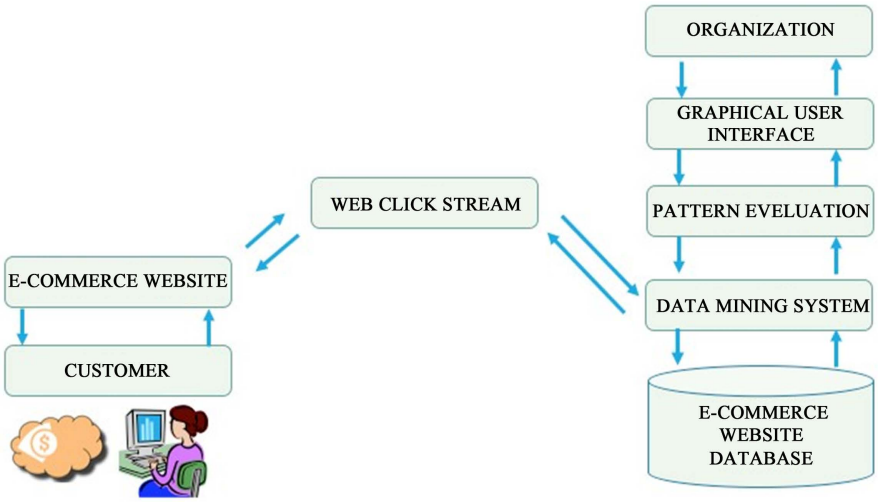

Figure 1. Architecture of the system.

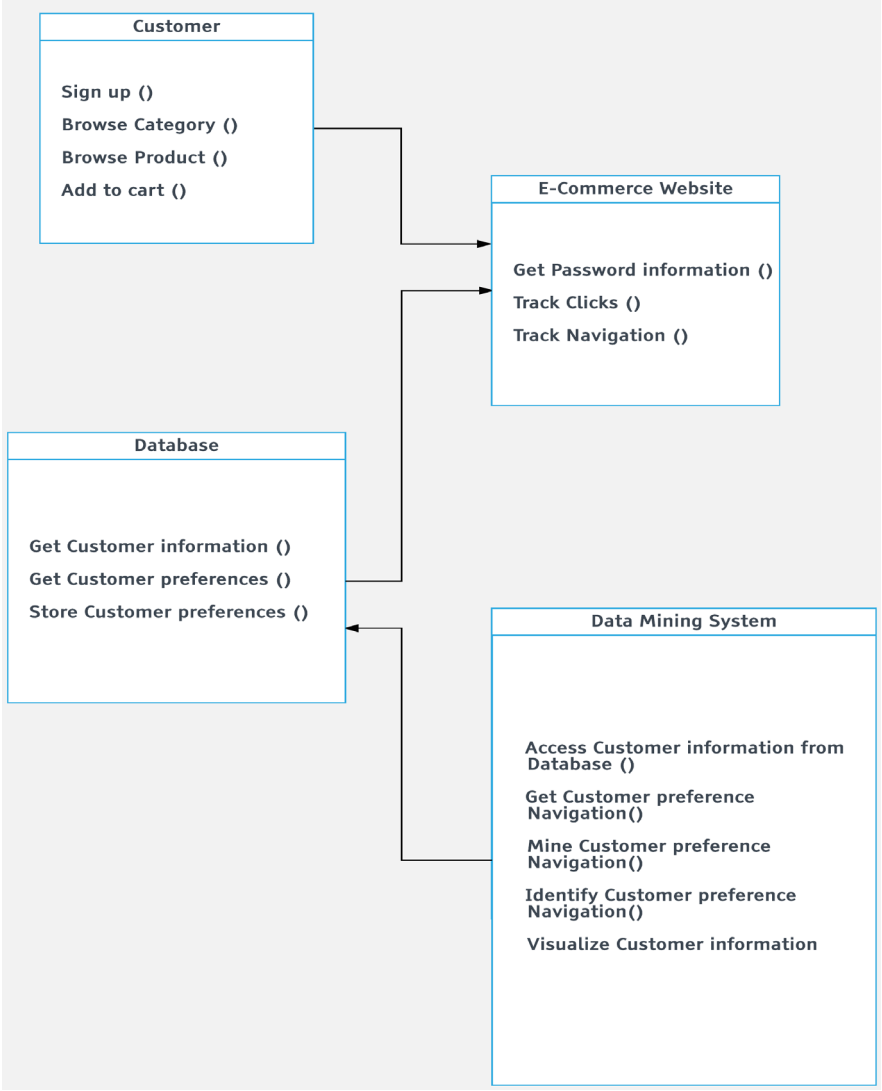

Figure 2. Class diagram for the system.

5. System Interface

1) Input Data

The information collected from customers on the e-commerce website is the input for the data mining system to clean and preprocess the data. **Figure 4** shows the dialog box for the user to select the data file to be analyzed.

2) Output Result

The results obtained are easy to understand and is interpretable for an e-commerce company or data analyst user as they are represented using an object-oriented approach.

a) Geographic quantitative message:

i) **Figure 5** displays the frequency of purchases of certain products by the regions. This shows that Rice is the most purchased product from the ecommerce website in the South-South region of Nigeria, the customers in the South-East region preferably purchase vegetables and fruits and millet products over other

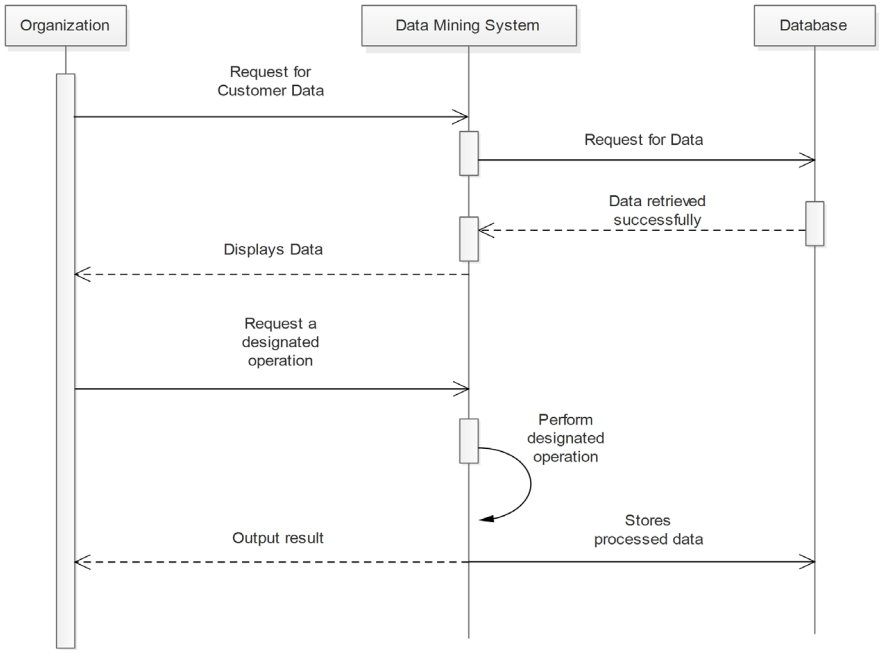

Figure 3. Sequence diagram for the datamining system.

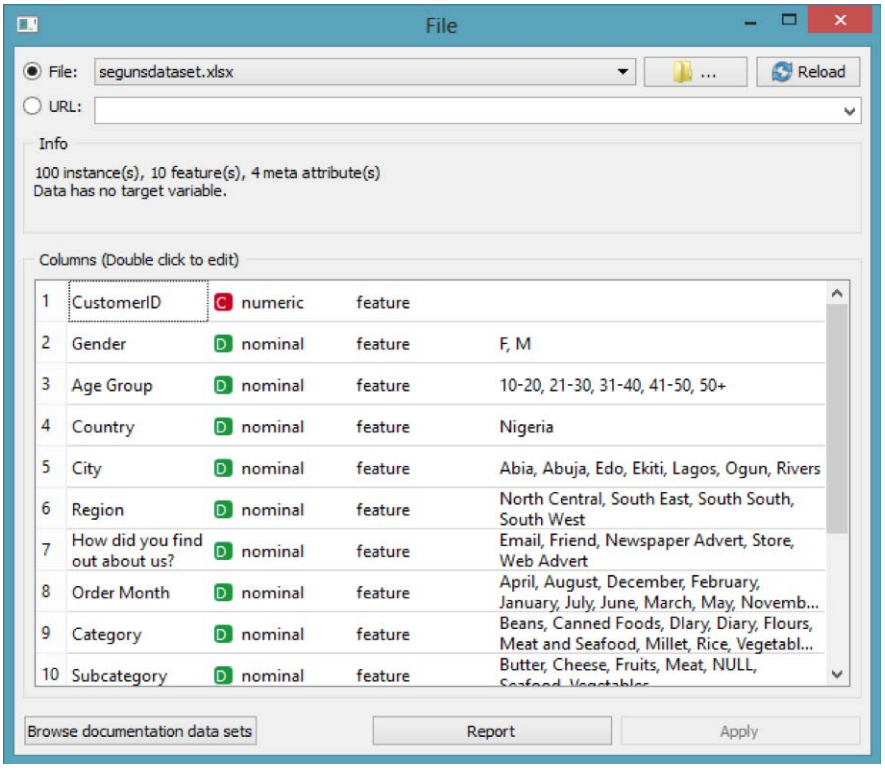

Figure 4. Input data for the data mining system.

products on the website, Dairy products and meat and seafood are the most sought after products in the North-Central region, and South-West region are not frequent buyers of vegetables and fruits who also dislike canned foods.

ii) The distribution in **Figure 6** shows the customers between the age of 21

and 30 who frequently visit the website in respect to the state they reside in. This can be used to create campaigns for the youths on certain states. This figure displays an outlier in Ogun state, which implies an absent population of customers who visit the website from Ogun state. It also explains that Lagos state is the highest state with customers between the age of 21 and 30 visiting the website and performing transactions.

b) Correlation Quantitative Message:

The scatter plot in Figure 7 is used to represent the month's customers make purchases in year against the age group of the customers in correlation to the gender of customers who visit the e-commerce website.

c) Month sequence quantitative message

The distribution plot in Figure 8 displays the month of the year with the most order of products, which is May, followed by April, slightly followed by the

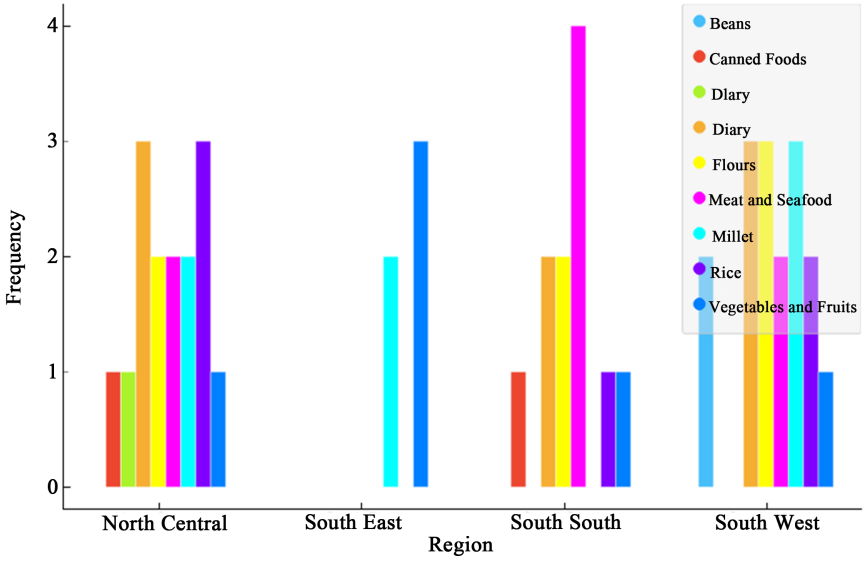

Figure 5. Frequency of purchases of certain products by the regions.

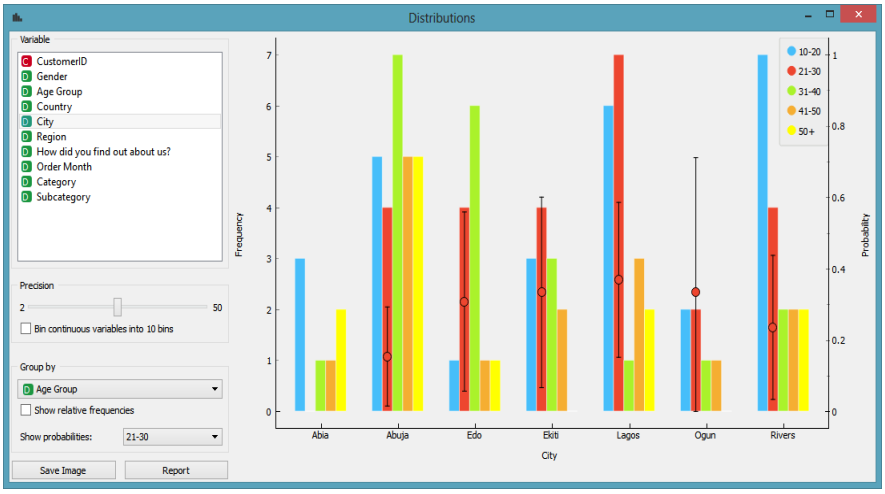

Figure 6. Geographic location of customers by age group.

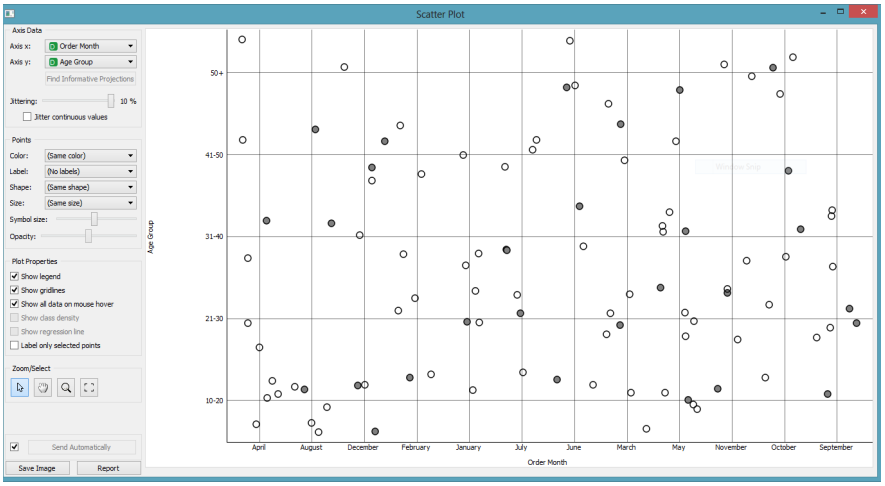

Figure 7. Months of the year and customer age group in correlation with males and females.

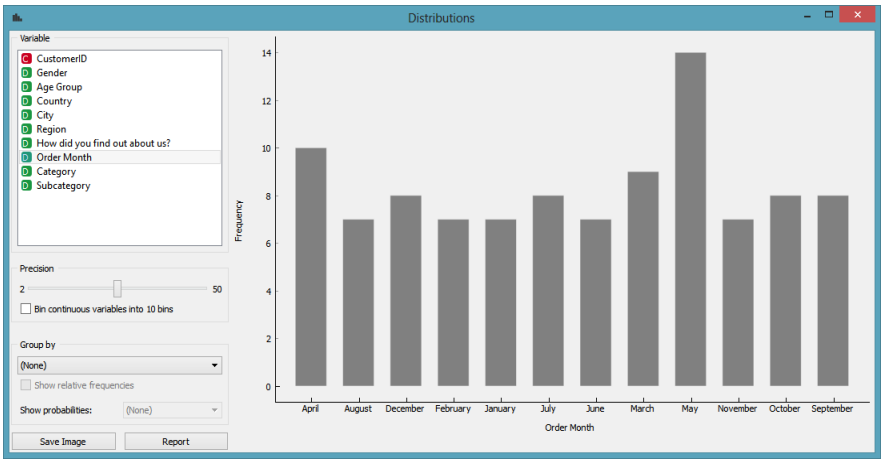

Figure 8. Plot displaying the month with the highest purchase.

month of March. This can be interpreted as customers frequently visit the website to purchase food products.

6. Conclusions

In conclusion, the project started out using Object oriented design methodology model, using several consensus clustering algorithms to test the data set. The final model could be delivered and has been implemented for use for not only retail companies or organizations but also similar datasets and settings. It is recommended that machine learning should also be explored. Research should be done on the application in order to implement a distinctive feature that offers the user an option to choose a preferred data visualization method. Also, a different setting such as Census data for a region, voters' polls during election should be explored. In developing a data mining system, other data mining techniques and algorithms can be tested on a dataset to give an efficient and maximal result.

During the advancement of the project, difficulties were experienced dealing with large number of dimensions and large number of data items which were problematic because of time complexity of the project's completion.

Request for customer's data from an existing e-commerce organization was not attainable due to time constraints. Another challenge was the slow computation and processing of the data mining system while testing data provided by a large population of customers stored in the database. Also, there was a limit to the amount of data size required to integrate the data mining system, because, the system can only handle less than a million customers data at once. Finally, the ecommerce website was not hosted on the internet as it was locally hosted with the use of XAMPP.

Acknowledgements

We hereby acknowledge the constructive criticism by all lecturers in the Department of Mathematical and Physical Sciences, Afe Babalola University. Thank you all.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Ismail, M., Ibrahim, M., Sanusi, Z. and Nat, M. (2015) Data Mining in Electronic Commerce: Benefits and Challenges. *International Journal of Communications, Network and System Sciences*, **8**, 501-509. https://doi.org/10.4236/ijcns.2015.812045

[2] Wang, J.C., David, C.Y. and Chris, R. (2002) Data Mining Techniques for Customer Relationship Management. *Technology in Society*, **24**, 483-502. https://doi.org/10.1016/S0160-791X(02)00038-6

[3] Kantardzic, M. (2003) Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons Inc., New York.

[4] Bucklin, R.E. and Sismeiro, C. (2009) Advances in Click-Stream Data Analysis in Marketing. *Journal of Interactive Marketing*, **23**, 35-48. https://doi.org/10.1016/j.intmar.2008.10.004

[5] Loudon, D.L. (1979) Customer Behaviour: Concepts and Applications. McGraw-Hill, New York.

[6] Vallamkondu, S. and Gruenwald, L. (2003) Integrating Purchase Patterns and Traversal Patterns to Predict Http Requests in E-Commerce Sites. *IEEE International Conference on E-Commerce*, Newport Beach, CA, 24-27 June 2003, 256-263.

[7] Michael, J.A.B. and Gordon, S.L. (1997) Data Mining Techniques: For Marketing and Sales, and Customer Relationship Management. 3rd Ed., Wiley Publishing Inc., Canada.

[8] Zhao, X.S. and Ji, K.F. (2013) Tourism E-Commerce Recommender System Based on Web Data Mining. *International Conference on Computer Science & Education*, **2**, 30-33.

[9] Saloni, A. and Veenu, M. (2013) Importance of Domain Knowledge in Web Re-

commender Systems. International *Journal of Computer Applications*, **127**, 10-12. https://doi.org/10.5120/ijca2015906643

[10] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2001) Item-Based Collaborative Filtering Recommendation Algorithms. *International Conference on World Wide Web*, Chiba, 10-14 May 2001, 285-295.

[11] O'Brien, J.A. (2011) Introduction to Information Systems. The McGraw-Hill Companies, New York.

[12] Monti, S., Tamayo, P., Mesirov, J. and Golub, T. (2003) Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Micro Array Data. *Machine Learning*, **52**, 91-118.