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, Special Issue 3, April 2017

Finding Potential Customers using Business Analytics in E-Commerce

R. Poornima, N.Priyadharsini, D.Jayanthi

UG Scholar, Department of Information Technology, Sri Venkateswara College of Engineering, Sriperumbudur, Chennai, India

UG Scholar, Department of Information Technology, Sri Venkateswara College of Engineering, Sriperumbudur, Chennai, India

Assistant Professor, Department of Information Technology, Sri Venkateswara College of Engineering, Sriperumbudur, Chennai, India

Abstract: Online shopping has emerged as one of the most popular internet activities, providing a variety of products for consumers and a multiplicity of sales challenges for e-commerce players. This has also led to a huge increase in the e-commerce data, making it a biggest data asset that can be explored to provide key insights to the needs of the customers and help improve customer targeting. In this project, we present e-commerce customer analytics through hierarchical & K-means clustering, profiling and advanced optimization techniques like Bootstrap sampling to identify ideal customer segments for effective targeting for cross-sell/up-sell opportunities. This technique is more effective in identifying more robust customer profiles compared to traditional customer profiling and help marketers reach to the right set of customers with right offers at the right time, therebydriving more sales and engagement.

KEYWORDS: Hierarchical clustering, K-Means clustering, Boot-strap sampling, Customer profiling.

I. INTRODUCTION

E-Commerce websites need to track customer behavior to increase their sales and enhance profits by targeting the potential customers. Effective analysis techniques are needed to track the right set of customers. Hierarchical clustering is applied on the datasets for determining the number of clusters the datasets can be categorized into. Bootstrap sampling is done along with k-means clustering. Customer profiling describes the set of customers clustered and helps in personalized targeting. Ecommerce in India is still far from achieving its full potential. For an industry plagued by poor penetration of the internet, smart phones, online banking and general internet awareness, data analytics has come to the rescue! Data analyzing is a process of examining large data sets containing a variety of data types — Big Data — to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings provide for effective marketing, new revenue opportunities, better customer service, improved operational efficiency and competitive advantages. One of the major benefits for ecommerce companies is from being internet-enabled, every step of their business from customer acquisition to product delivery, can be tracked.

Data analytics is helping ecommerce companies adopt real-time pricing based on customer demand and competitive pricing. With the analyses of large amounts of product data and consumer likes, purchases and reviews, among others, the portfolio of products can be optimized for each user or group of users based on similar clusters or segments. Analytics also help predict what customer needs and assist in recommending the products. However, like anything else, data analytics comes with its own set of challenges. Since the data generated is large in terms of volume and comes in various forms like structured (e.g. Name, age, sex, address, and preferences), and unstructured (like reviews and feedback, tweets, videos, voice, clicks, etc.)

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, Special Issue 3, April 2017

Traditional techniques of analyzing the data prove to be inefficient. It is found that brand marketing, buyer perception, risk assessment, product pipeline management, backlog and fulfillment tracking, recommendation engines and pricing are some of the most popular demands on the analytics side from the ecommerce space.

A . HIERARCHICAL CLUSTERING

Hierarchical clustering, as the name suggests is an algorithm that builds hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left.The results of hierarchical clustering can be shown using Dendrogram.

B. BOOTSTRAP SAMPLING

The basic idea of bootstrapping is that inference about a population from sample data can be modelled by *resampling* the sample data and performing inference about a sample from resampled data. As the population is unknown, the true error in a sample statistic against its population value is unknowable. In bootstrap-resamples, the 'population' is in fact the sample, and this is known; hence the quality of inference of the 'true' sample from resampled data, is measurable.

C. K-MEANS CLUSTERING

K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 5 steps:

1. Specify the desired number of clusters K.
2. Randomly assign each data point to a cluster.
3. Compute cluster centroids.
4. Re-assign each point to the closest cluster centroid.
5. Re-compute cluster centroids

D. CUSTOMER PROFILING

Customer profiling is a way to create a portrait of your customers to help you make design decisions concerning your service. Customers are broken down into groups of customers sharing similar goals and characteristics and each group is given a representative with a photo, a name, and a description.

E. VISUAL REPRESENTATION AND REPORTING

Data visualization is the presentation of data in a pictorial or graphical format. The analyzed datais represented to the business in the form of graphs and charts. This provides insights for businesses to target the right group of customers for cross-selling and promotional offers.

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, Special Issue 3, April 2017

II. PROPOSED METHODOLOGY

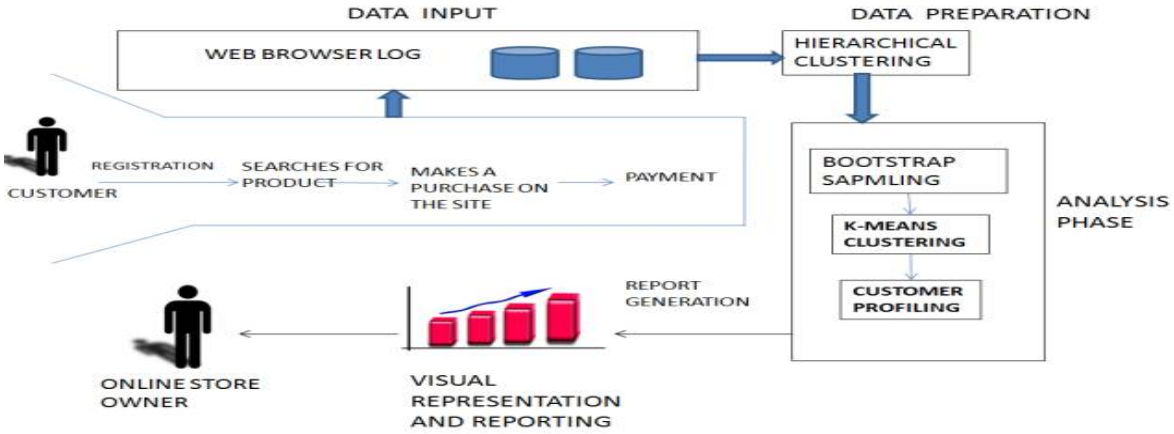


Fig. 1. Proposed Architecture of Analytics in E-Commerce

Fig. 1. shows the representation of different phases involved in the analysis of e-commerce data. When a customer registers and make a payment on the e-commerce website, the details gets stored in the weblog in unstructured form. This data is made into a structured form by data pre-processing. Hierarchical clustering is applied on the datasets to identify the category of clusters. Bootstrap sampling is done for accuracy of the clusters. K-Means clustering gives the list of cutomers in each cluster and customer profiling describes the clusters. This helps in targeting the potential customers.

III. MODULES

A. DATA PRE-PROCESSING

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Data preprocessing is used database-driven applications such as customer relationship management.

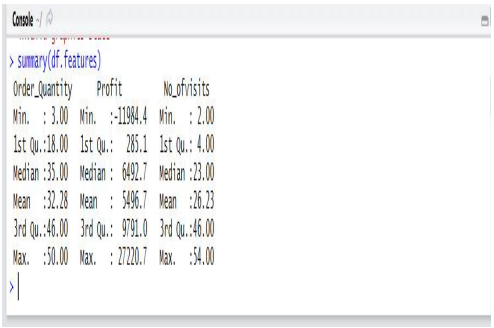


Fig. 2. Summary of the fields in the datasets

Fig. 2. Shows the summary function is used which helps in identifying the first quantile, second quantile, third quantile, mean, minimum, median and maximum value. Univariate analysis is done where the values of the datasets are scaled and centered. The values are in the range between -2 and +2. The table below shows the output of Univariate analysis.

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

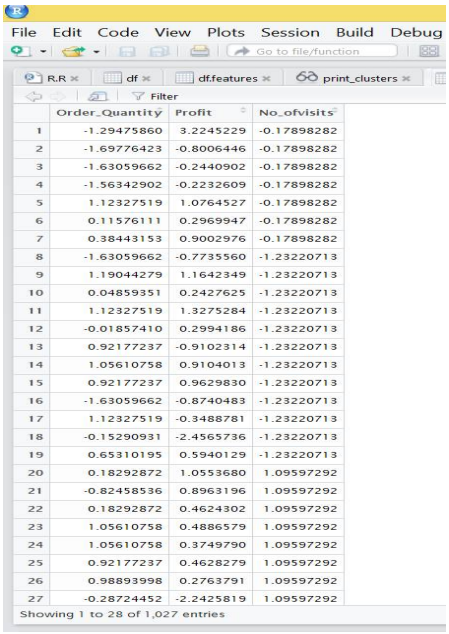
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 3, April 2017

This is basically done to identify the outliers and remove them from the dataset to be analyzed. The fields needed for analysis of the customer behavior like order quantity, profit and number of visits are alone used in a data frame named df.features.



	Order_Quantity	Profit	No_ofvisits
1	-1.29475860	3.2245229	-0.17898282
2	-1.69776423	-0.8006446	-0.17898282
3	-1.63059662	-0.2440902	-0.17898282
4	-1.56342902	-0.2232609	-0.17898282
5	1.12327519	1.0764527	-0.17898282
6	0.11576111	0.2969947	-0.17898282
7	0.38443153	0.9002976	-0.17898282
8	-1.63059662	-0.7735560	-1.23220713
9	1.19044279	1.1642349	-1.23220713
10	0.04859351	0.2427625	-1.23220713
11	1.12327519	1.3275284	-1.23220713
12	-0.01857410	0.2994186	-1.23220713
13	0.92177237	-0.9102314	-1.23220713
14	1.05610758	0.9104013	-1.23220713
15	0.92177237	0.9629830	-1.23220713
16	-1.63059662	-0.8740483	-1.23220713
17	1.12327519	-0.3488781	-1.23220713
18	-0.15290931	-2.4565736	-1.23220713
19	0.65310195	0.5940129	-1.23220713
20	0.18292872	1.0553680	1.09597292
21	-0.82458536	0.8963196	1.09597292
22	0.18292872	0.4624302	1.09597292
23	1.05610758	0.4886579	1.09597292
24	1.05610758	0.3749790	1.09597292
25	0.92177237	0.4628279	1.09597292
26	0.98893998	0.2763791	1.09597292
27	-0.28724452	-2.2425819	1.09597292

Showing 1 to 28 of 1,027 entries

Fig. 3. P-matrix table

Fig.3. Shows the representation of p-matrix table which contains the scaled data for analysis.

B. CLUSTER IDENTIFICATION USING HIERARCHICAL CLUSTERING

In hierarchical clustering method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation. Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each point using a distance function. Then, the matrix is updated to display the distance between each cluster. This is exploratory approach to identify the ideal number of clusters. This process involves finding distance matrix using Euclidean distance method. The resultant matrix consists of distances. Hierarchical clustering is done using the ward method on this matrix. Cluster Dendrogram is plotted. From this, the ideal number of clusters is identified. The extracted clusters are represented for further analysis.

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, Special Issue 3, April 2017

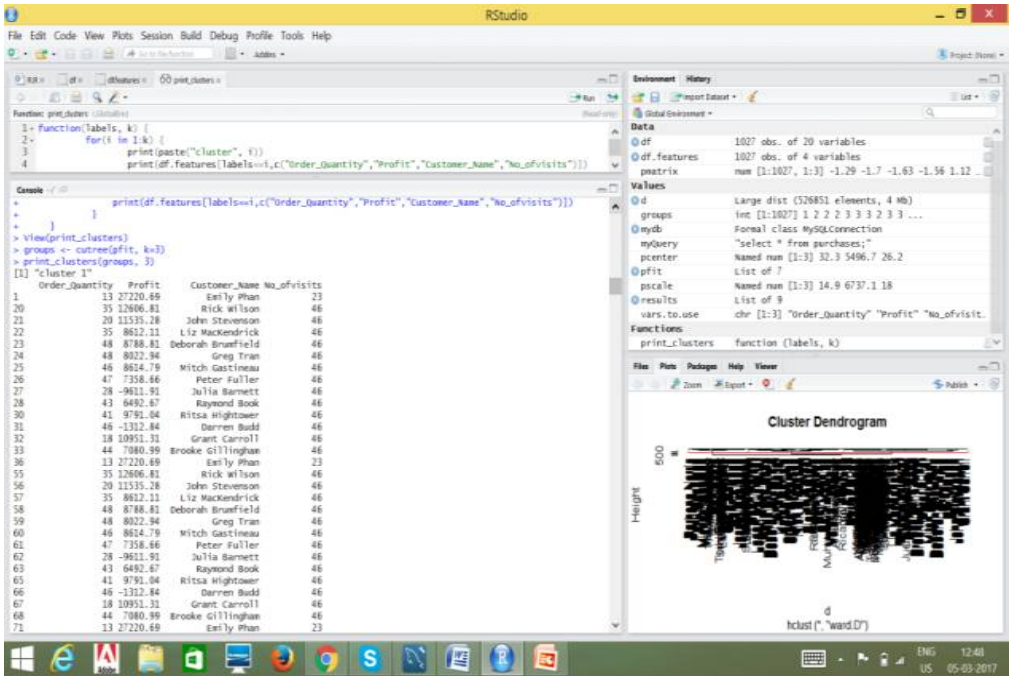


Fig. 4. Customers in each cluster printed.

Fig.4. Shows the number and the list of customers.

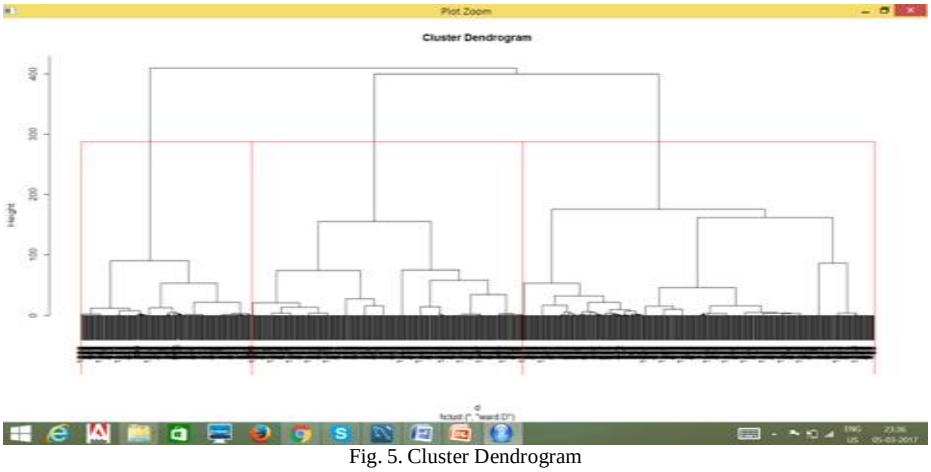


Fig. 5. Cluster Dendrogram

Fig. 5. shows the category of clusters as three. Hierarchical clustering helps in identifying the number of clusters.

C. CLUSTER VALIDATION USING K-MEANS CLUSTERING AND BOOTSTRAP SAMPLING

Clusterboot function is used for resampling the data. K-Means clustering is done on the samples of data iteratively. The stability of the clusters are determined by finding out the stability vectors and the number of dissolved clusters. The lesser the number of dissolved clusters, higher is the stability of the clusters. The value of the stability vectors close to one indicate stable clusters. A bootstrap sample is a smaller sample that is "bootstrapped" from a larger sample. Bootstrapping is a type of re-sampling where large numbers of smaller samples of the same size are repeatedly drawn, with replacement, from a single original sample.

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, Special Issue 3, April 2017

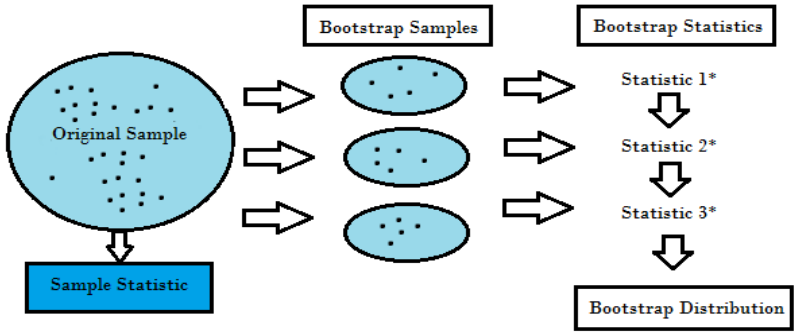


Fig. 6.Bootstrap sampling

Fig.6. shows the iterative sampling of fixed number of samples.

An important question when evaluating clusters is whether a given cluster is "real"-does the cluster represent actual structure in the data, or is it an artifact of the clustering algorithm? This is especially important with clustering algorithms like k-means, where the user has to specify the number of clusters a priori. It's been our experience that clustering algorithms will often produce several clusters that represent actual structure or relationships in the data, and then one or two clusters that are buckets that represent "other" or "miscellaneous." Clusters of "other" tend to be made up of data points that have no real relationship to each other; they just don't fit anywhere else. One way to assess whether a cluster represents true structure is to see if the cluster holds up under plausible variations in the dataset. The fpc package has a function called clusterboot()that uses bootstrap resampling to evaluate how stable a given cluster is clusterboot() is an integrated function that both performs the clustering and evaluates the final produced clusters. It has interfaces to a number of R clustering algorithms, including both hclust and kmeans. Clusterboot's algorithm uses the Jaccard coefficient, a similarity measure between sets. The Jaccard similarity between two sets A and B is the ratio of the number of elements in the intersection of A and B over the number of elements in the union of A and B. The basic general strategy is as follows:

1. Cluster the data as usual.
2. Draw a new dataset (of the same size as the original) by resampling the original dataset with replacement (meaning that some of the data points may show up more than once, and others not at all). Cluster the new dataset.
3. For every cluster in the original clustering, find the most similar cluster in the new clustering (the one that gives the maximum Jaccard coefficient) and record that value. If this maximum Jaccard coefficient is less than 0.5, the original cluster is considered to be dissolved-it didn't show up in the new clustering. A cluster that's dissolved too often is probably not a "real" cluster.
4. Repeat steps 2-3 several times.

Different clustering algorithms can give different stability values, even when the algorithms produce highly similar clusterings, so clusterboot() is also measuring how stable the clustering algorithm Is.

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, Special Issue 3, April 2017

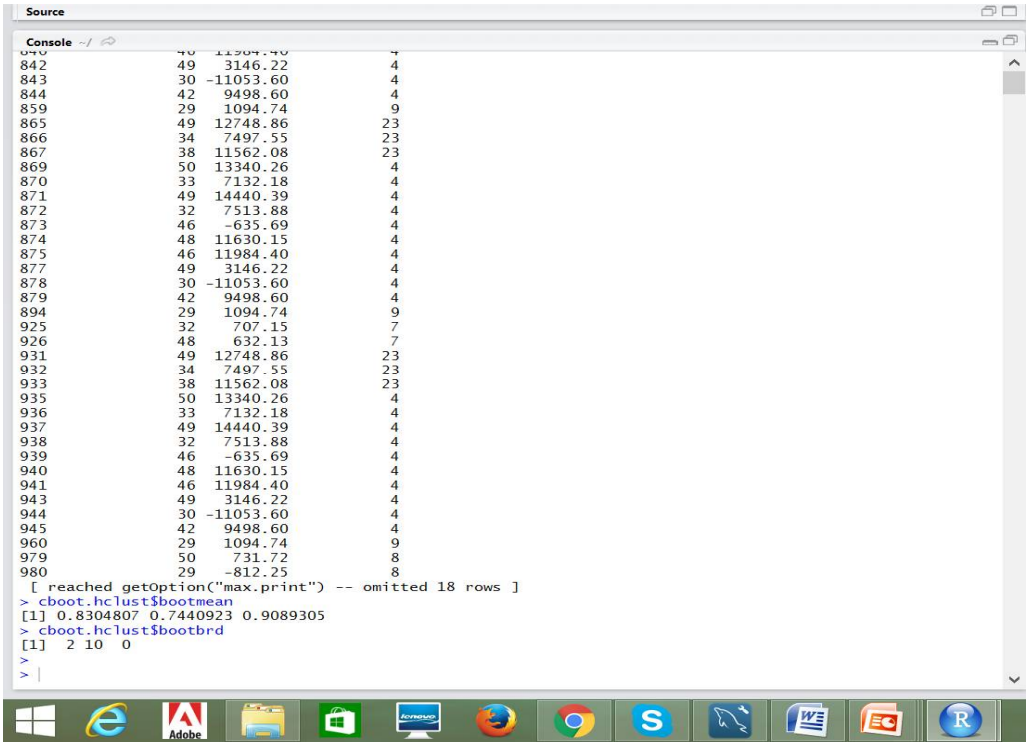


Fig.7.The mean value close to one represents stability of clusters. The bootbrd value represents number of dissolved clusters.

Fig. 7. Shows the mean value of the clusters. The mean value close to one represents stability of clusters. Bootbrd value represents the number of dissolved clusters.

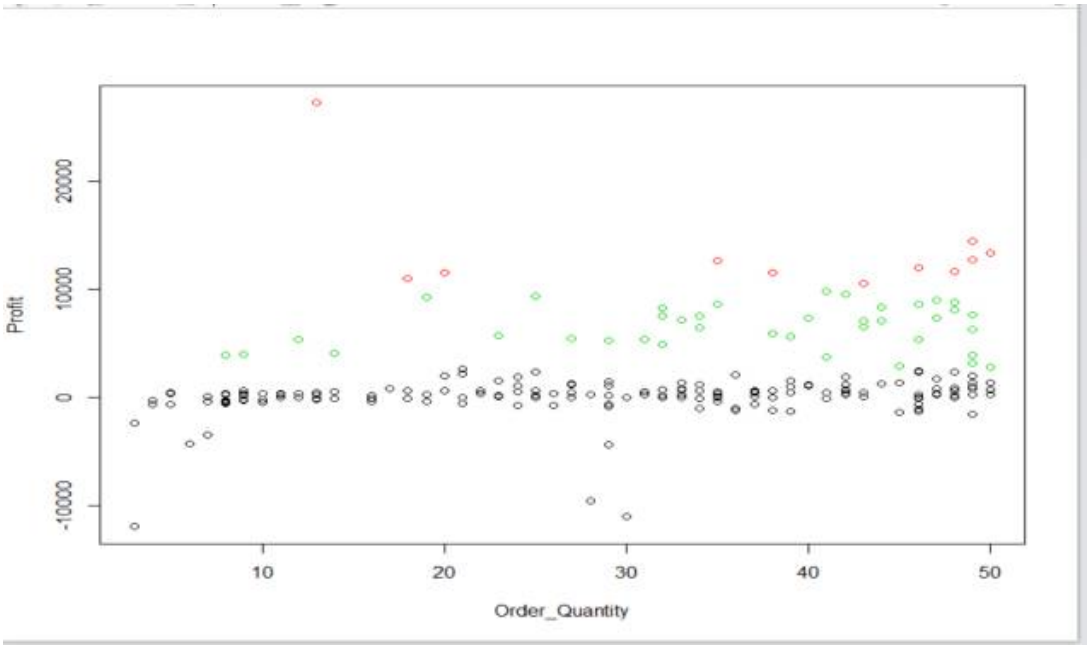


Fig. 8. Clusters of datasets

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, Special Issue 3, April 2017

Fig. 8. Shows the cluster of datasets. The black dots represent low profitable customers, the green dots represent medium profitable customers and the red dots represent high profitable customers. If the customers in green dots are given promotional offers, there are chances for increasing the number of high profitable customers.

D. CUSTOMER PROFILING

Customer profiling is a way of describing a customer categorically so that they can be grouped for marketing and advertising purposes. The clusters identified are described and similarities within the datasets in the clusters are defined for targeting potential group of customers. Personalization for ecommerce stores picks up on the site where targeting leaves off. While it's not easy to adjust site content based on specific targeting segments once the visitor is there, it is possible to further personalize the shopping experience for visitors who have already provided some clues as to what they're personally interested in.

One method that many online retailers have found to be extremely effective is installing a recommendation engine on their site. This software constantly takes in information about the visitor as he or she navigates the site. Based on where they click, how they engage, and where they spend their time, the engine will present personalized product recommendations that are most likely to appeal to that individual.

These recommendations can be combined with targeting information and data from previous visits by the same individual to further tailor what products or offers are suggested.

For example, a site visitor who previously purchased a pair of designer boots and is now hunting around the site looking at a dozen different types of sweaters can receive a recommendation to check out a designer brand sweater that is currently on sale. Based on their past purchase and current behavior, there's a good chance they'll appreciate that recommendation.

So, in a nutshell, targeting is driven by the retailer and personalization is driven by the visitor. Both are powerful tools in the hands of an online retailer, and should be considered indispensable.

The most effective online stores will take advantage of both targeting and personalization to enhance their customers' shopping experience, improve engagement, and boost sales by ensuring the right personsgets the right message at the right time.

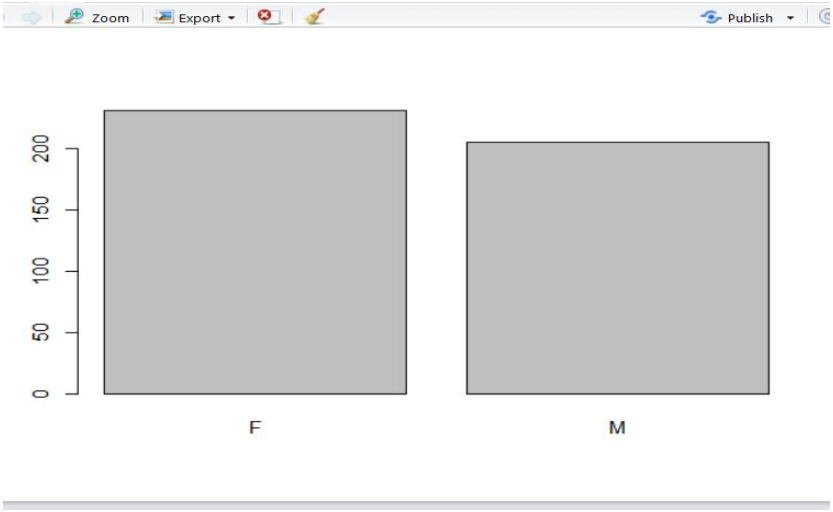


Fig. 9. This is the representation of the gender distribution in cluster 1.

Fig. 9. Shows the representation of gender distribution in cluster1.

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, Special Issue 3, April 2017

Fig. 10. This is the representation of the customer segment distribution in cluster 1.

Fig. 10. Shows the representation of the customer segment distribution

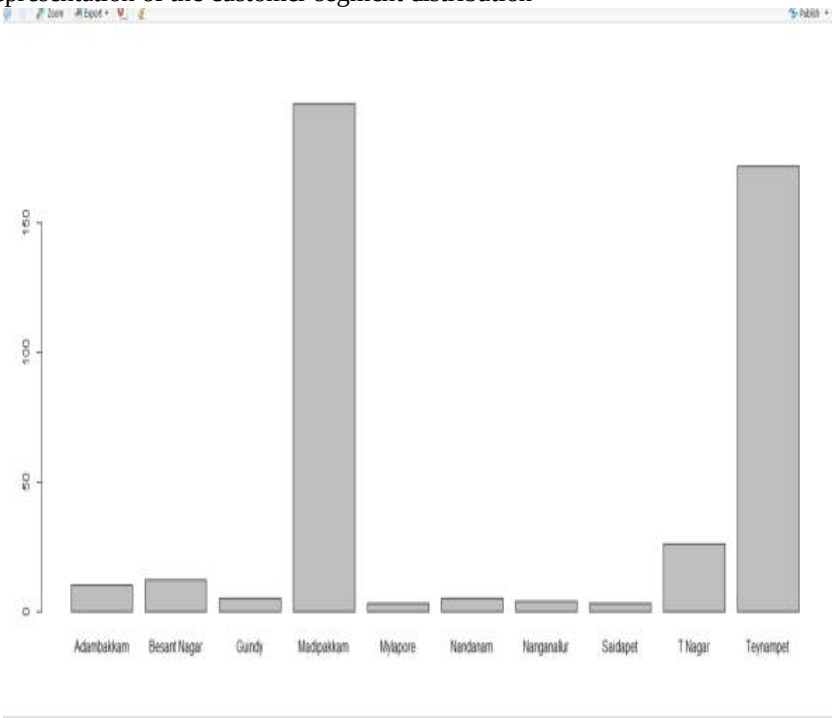


Fig. 11. This is the representation of the region distribution in cluster 1.

Fig. 11. Shows the representation of the region distribution in cluster 1.

E. VISUAL REPRESENTATION AND GRAPHING

Data visualization is the presentation of data in a pictorial or graphical format. The analysed data is represented to the business in the form of graphs and charts. This provides insights for businesses to target the right group of customers for cross-selling and promotional offers. A primary goal of data visualization is to communicate information clearly

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, Special Issue 3, April 2017

and efficiently via statistical graphics , plots and information graphics. Numerical data may be encoded using dots , lines , or bars , to visually communicate a quantitative message.
Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

Data visualization is both an art and a science. It is viewed as a branch of descriptive statistics by some, but also as a grounded theory development tool by others. The rate at which data is generated has increased. Data created by internet activity and an expanding number of sensors in the environment, such as satellites, are referred to as "Big Data". Processing, analyzing and communicating this data present a variety of ethical and analytical challenges for data visualization. The field of data science and practitioners called data scientists have emerged to help address this challenge. Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics. The goal is to communicate information clearly and efficiently to users. It is one of the steps in data analysis or data science.

According to Friedman the "main goal of data visualization is to communicate information clearly and effectively through graphical means. It doesn't mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key-aspects in a more intuitive way. Yet designers often fail to achieve a balance between form and function, creating gorgeous data visualizations which fail to serve their main purpose — to communicate information".

IV. CONCLUSION AND FUTURE ENHANCEMENT

The Statistical techniques like Hierarchical and K-means clustering come handy in identifying and uncovering the latent customer groups among online shoppers. Bootstrap Sampling complements the analysis by improving the accuracy of clustering and making the conclusions of the analysis more reliable. Customer profiling is done for customized targeting of customers. This helps in making businesses understand the characteristics of loyal shoppers and do cross-selling, up-selling and provide promotional offers to the customers. Promotions are a must-have for a retail business to succeed but it’s not easy to get them right. Customer profiling helps in offering the right promotions to right people. This also helps in look-alike modeling to identify new prospects and enhance their business.

The insights from the analysis can be leveraged by the marketers to devise personalized marketing campaigns for each customer segment based on their behavior and acquire new customers based on look-alike attributes from the top performing loyal spender segment of existing customers. Different customers engage with a retail site in different ways. Predictive analytics helps look at all the different variables to generate the desired engagement from the customer. This could mean signing up for a newsletter, clicking on a promotion or some other form of engagement. Offering convenience and ease in the digital shopping experience is pretty much table stakes for ecommerce platforms.

To stand out from the crowd and resonate in the consumer’s heart as well as mind, retailers should appeal to customers’ underlying motivations, values, and sensibilities that kindle memorable and meaningful purchase experiences. Specifically, let consumer control, success opportunities, and feelings of enhanced self-worth become an integral part of your messaging and the shopping experience. In digital realms, store associates can play a vital role, making the customer feel valued. Incorporating these hooks may not only help increase the likelihood of bringing in customers, further increasing the likelihood of positive recommendations and future patronage.

REFERENCES

[1] Y. Frishman and A. Tal. Online dynamic graph drawing. IEEE Transactions on Knowledge and Data Engineering, 14(4):727–740, 2015.
[2] H. Guo, S. R. Gomez, C. Ziemkiewicz, and D. H. Laidlaw. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. IEEE Transactions on Visualization and Computer Graphics, 22(1):51–60, 2016.
[3] J. Heer and G. Robertson. Animated transitions in statistical data graphics. IEEE Transactions on, Visualization and Computer Graphics, 13(6):1240–1247, 2015.

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, Special Issue 3, April 2017

[4] T.Muhlbacher and H.Piringer. A partition-based framework for building and validating regression models. IEEE Tansactions on Visualization and Computer Graphics, 19(12):1962–1971, 2016.
[5] D. Archambault, H. Purchase, and B. Pinaud. Animation, small multiples, and the effect of mental map preservation in dynamic graphs. IEEE Transactions on Visualization and Computer Graphics, 17(4):539–552, 2016.
[6] N. Elmqvist, P. Dragicevic, and J. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. IEEE Transactions on Visualization and Computer Graphics, 14(6):1539–1148, 2017.
[7] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. IEEE Transactions on Visualization and Computer Graphics, 17(12):2203–2212, 2016.
[8] J.Blomberg, J.Giacomi, A.Mosher, and P.Swenton-Wall. Ethnographic field methods and their relation to design. Participatory design: Principles and practices, pages 123–155, 2016.
[9] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko. Effectiveness of animation in trend visualization., 14(6):1325 –1332, nov.-dec. 2016.
[10] E. Catmull. The problems of computer-assisted animation. IEEE Transactions on Visualization and Computer Graphics, 12(3):348–353, 2016.