Customer Segmentation Using Machine Learning

*Anuharsh Singh
**Dr P.G.Dangwal

ABSTRACT

Marketing analysis can extract useful information about individuals, trends, and segment from the mass of data. Datamining uses sophisticated statistical and mathematical techniques such as cluster analysis, automatic interaction detection, predictive modeling and neural networking. Customer Segmentation is one of the most substantialuses of unsupervised learning. Utilizing clustering techniques, organizations be able to recognize numerous segments of customers enabling to focus on the prospective clients. Customer segmentation relies on identifying key differentiators that divide customers into groups that can be targeted. The major segmentation variables are – geographic, demographic, psychographic and behavioral segmentation. In this paper, an attempt to define a model for customer segmentation has been made whilst taking into consideration the demographic and psychographic variables of the segmentation. In fulfillment of this purpose, a survey has been conducted on the frequent visitors of Malls & Retail Stores in the City of Dehradun. Optimal clusters are determined using popular methods such as Elbow Method, Silhouette Method, and Gap Statistic Method.

Keywords: Machine Learning, Customer Segmentation

1. Introduction

1.1 Customer Segmentation

Customer segmentation is the drill of dividing a customer base into groups of individuals that are akin in specific ways pertinent to marketing. Business organization engaging in customer segmentation operate under the fact that every customer is unique and the marketing efforts would be better served if they target specific, smaller groups with messages that those customers would find relevant and lead them to make purchases.

Business Organizations that deploy customer segmentation are under the notion each customer has various prerequisites and require a marketing exertion to address them suitably. Organizations aim to gain an abysmal approach of the customer they are focusing on. Along these lines, their point must be explicit and ought to be custom-made to address the prerequisites of every individual customer. In addition, through the data gathered, organizations can achieve more profound comprehension of customer inclinations as well as the requirements for discovering valuable sections that would harvest them maximum gain. Along these lines, they can strategize their marketing techniques more productively and limit the possibility of risk to their investment.

Business organizations also hope to gain a more profound comprehension of their customers' inclinations and requirements with the idea of discovering what each fragment finds most valuable to more precisely, tailor marketing materials toward that segment. Customer segmentation relies on identifying key differentiators that divide customers into groups that can be targeted. The major segmentation variables are - geographic, demographic, psychographic and behavioral segmentation.

1.2 Machine Learning

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

*Operations Executive, PayU Payments Pvt Ltd, Netherlands
**Associate Professor, IMS Unison University, Dehradun

Some machine learning methods:

- Supervised machine learning algorithms can smear what has been learned in the past to new data using labeled examples to predict future events. Initiating from the analysis of a known training data set, the learning algorithm yields an inferred function to make predictions about the output values. The system is able to deliver targets for any different input after sufficient training. The learning algorithm can also equate its output with the correct, intended output and find errors in order to transform the model accordingly.

- Unsupervised machine learning algorithms are used when the data used to train is neither classified nor categorized. Unsupervised learning studies how systems can infer a function to define a hidden structure from uncategorized data. The system doesn't figure out the exact output, however it reconnoiters the data and can portray inferences from data sets to describe hidden structures from uncategorized data.

- Semi-supervised machine learning algorithms persists somewhere in in the middle of supervised and unsupervised learning, since they utilize both categorized and uncategorized data for training - typically a small amount of categorized data and a great amount of uncategorized data. The systems that use this method are able to extensively improve learning accurateness. Usually, semi-supervised learning is preferred when the assimilated categorized data requires skilled and relevant resources in order to train it / learn from it. Otherwise, obtainingun categorized data generally doesn't require additional resources.

- Reinforcement machine learning algorithms is a learning technique that interrelates with its environment by producing actions and determines errors or rewards. Trial and error search and delayed reward are the most pertinent physiognomies of reinforcement learning. This method allows machines and software proxies to automatically determine the idyllic behavior within a explicit context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is acknowledged as the reinforcement signal.

1.3 K-Means Algorithm

K-means clustering algorithm is an unsupervised method to group data in the order of their resemblances. Envision that there are numerous points spread over an n-dimensional universe. Therefore, to classify this data based on their similitude, K-means clustering algorithm is used.

K-means clustering is an iterative algorithm that segments a group of data containing n values into k subgroups. Every one of the n value belongs to the k cluster with the adjoining mean.

The aforementioned means that given a group of objects is partitioned into numerous sub-groups. These sub-groups are formed based on their resemblance and the distance of each data-point in the sub-group with the mean of their centroid.

The aim of the K-means clustering is to abate the Euclidean distance that each point has from the centroid of the cluster. This is renowned as intra-cluster variance and can be curtailed using the following squared error function

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where,

'||xi - cj||' is the Euclidean distance between xi and cj.

'n' is the number of data points in ith cluster.

'k' is the number of cluster centers.

Whist operational with clusters, it is important to specify the amount of clusters to use in the analysis. For a precise result, it is vital to make use of the optimal number of clusters. There are three popular methods used to determine the optimal number of clusters -

- Elbow method: this technique define the clusters such that the intra-cluster variation stays minimum. minimize (sum W(Ck)), k=1…k

- Silhouette method: this technique calculates the mean of silhouette observations for different values of k.

- Gap statistic: this technique is used to equate the total intra-cluster discrepancy for different values of k along with their anticipated values under the null reference dispersal of data.

2. Literature Review

Subsequently analyzing the literature available on Customer Segmentation Using Machine Learning, it is bring into being that limited research is done that tend towards Customer Segmentation Using Machine learning. Most of the studies conducted were in different fields and were very specialized. Researchers had focused on problems of either customer segmentation or machine learning. Whereas few researchers had focused upon streamlining the results of K-means clustering algorithm.

The paper "Monte Carlo Simulation and Clustering for Customer Segmentation in Business Organization" (Andry A. and Bellania N.,2017,) is the most relevant study to this topic. However, the researchers has utilized simulated data from Monte Carlo Simulation.

In addition, the paper "Using Data Mining Techniques In Customer Segmentation" (Ziafat And M Shakeri,2014) explores the important uses of data mining techniques that are pertinent in Customer Segmentation. It explains the effective use of Customer Relationship Management systems and ways marketers can utilize this in their businesses.

Hence, in this paper, the real time customer data has been collected using surveys, questionnaires and interviews and compiled for analysis. Moreover, a model has been developed which can be used for further study and research.

2.1 Customer Characteristics As Criteria For Market-Segmentation In Libraries (1988, M S Sridhar, E-LIS).

This study highlights the limitations of marketing approach to library and information services and Analyses the data collected for a larger study of information-behavior of the Indian space technologists. Being paternalistic systems, libraries try to create awareness among users and persuade them to use information, documents and their users need other library services. It can be concluded from the analysis of the data presented above that the IST as users of their `primary library' can be segmented by their characteristics for planning and providing information services. As the data requirements of each segment within the IST has significantly diverse from others, the `primary library' should take note of such take annotation of such segmentation within its target market.

2.2 Domain? Specific Market Segmentation (1994, Fred van Raaij, W. and Verhallen, T., European Journal of Marketing, Vol. 28, ISSN: 0309-0566).

In this paper, Domain-specific market segmentation is proposed as a promising methodology contrasted and the segmentation dependent on general and brand-specific factors. Domain-specific factors are dynamic, whereas general and brand-specific factors are latent in the formation of the segments. Product differentiation as the counterpart to display segmentation is demonstrated in a supply demand model, identified with importance structure analysis. Domain-specific market segmentation is most adequately done with canonical analysis. Segmentation results give the strategy producer a separated perspective on the consumer market. The segments found may demonstrate conceivable outcomes for new products and better approaches to convey about products. To guarantee that changes in market structure can be observed, it is fitting to build a gadget, a short survey, which may effortlessly recognize individuals as having a place with a particular segment.

2.3 A Practical Yet Meaningful Approach To Customer Segmentation (1998, C Marcus, Journal of Consumer Marketing, ISSN: 0736-3761)

This paper presents the idea of the Customer Value Matrix, a customer segmentation methodology that is particularly well?suited for small and medium retail and service administration organizations. The paper offers bits of knowledge into the explanations behind the advancement of this functional methodology, a solid approach for its execution, and vital and strategic utilizations of the idea. The material is upheld with solid evidence from "real?world" models highlighting an assortment of little retail and service organizations. The paper concludes with a discourse of the administrative ramifications for organizations that oversee chains of little retail or service organizations concerning how they can exploit local relationship marketing.

2.4 Customer Focus through Market Segmentation (2001, D Nilsson and J Olsson, Göteborg University, ISSN 1403-851X).

The purpose of this study was to conduct an empirical and theoretical study to illustrate and illuminate how a manufacturing company operating in an industrial market can become more customer focused through the implementation of market segmentation. According to the literature, the pre-eminent way to identify customers' needs and wants is to understand what the customers perceive as value and how that value is generated. The theory implies that value is the relation between the perceived benefits and perceived sacrifices of goods and services and that actual value is created when the customer utilizes these goods and services. An identification of the influencers of value will lead to understanding of customers' needs and wants. In addition, how they will change.

2.5 Natural clustering: the modularity approach (2007, L. Angelini, D. Marinazzo, M. Pellicoro, and S. Stramaglia, DipartimentoInterateneo di Fisica, Bari, Italy).

In this study, researchers had shown that modularity, a quantity introduced in the study of networked systems, could be generalized and be used in the clustering problem as an indicator for the quality of the solution. The problem of data clustering consists of grouping together items so that two points belonging to the same group (cluster) are, in some sense, more similar than two that belong to different ones; it has applications in several fields such as pattern recognition, bioinformatics, learning, astrophysics and more. We have shown that the problem of finding the optimal classification in hierarchical clustering can be turned into the problem of finding

communities in a weighted network. Results reported in this article were obtained using the CMC algorithm, but it should be clear that the modularity method could be applied to any hierarchical clustering algorithm.

2.6 Customer Segmentation In The Medical-Devices Industry (2007, P Basu And E K Kim, Massachusetts Institute Of Technology).

Medical device trades are non-seasonal and do not show marketing effects. The objective of this study was to validate the proposed approaches researchers reviewed as a basis for recommending ways to segment customers for enhancing service while decreasing cost. Researchers had recommended three types of segmentation: by region, by order method and by division. Segmentation by region divides the customers into 4 regions based on their location. Segmentation by ordering method divides the customers in terms of how they had placed order, for example, phone, fax or EDI whilst segmentation by division breaks up the customer base into the various divisions the company had. It had become increasingly clear to the company that such an option was not economically viable and made very less business sense. Some of the ways that the company could look at segmenting the customer profile included segmentation by regions, segmentation by divisions and segmentation by the method of placing orders.

2.7 Supervised k-Means Clustering (2008, T Finley and T Joachims, Cornell University).

In this study, researchers provided a means to parameterize the popular canonical k-means clustering algorithm based on learning a similarity measure amongst item pairs, and then delivered a supervised k-means clustering technique to learn these parameterizations using a structural SVM. The supervised k-means clustering technique learns this correspondence measure based on a training data set of item sets and complete partitioning over those sets, choosing parameterizations optimized for good performance over the training set. Then researchers theoretically characterized the learning algorithm, drawing a distinction between the iterative local search k-means clustering method and the relaxed spectral relaxation, as leading to under constrained and over-constrained supervised k-means clustering learners, respectively. Empirically, the supervised k-means clustering algorithms exhibited superior performance compared to naïve pair wise learning or unsupervised k-means. The under constrained and over constrained supervised k-means clustering learners compared to each other exhibited different performance, though neither was clearly consistently superior to the other.

2.8 How The Initialization Affects The Stability Of The K-Means Algorithm (2009, S Bubeck, M Meila, U V Luxburg, arxiv:0907.5494).

In this study, researchers had investigated the role of the initialization for the stability of the k-means clustering algorithm. As opposed to other papers, researchers had consider the actual k-means algorithm and do not ignore its property of being stuck in local optima. Researchers were interested in the actual clustering, not only in the costs of the solution. They had analyze when different initializations lead to the same local optimum, and when they lead to different local optima. This enables them to prove that it is reasonable to select the number of clusters based on stability scores.

2.9 Segmentation and Customer Insight in Contemporary Services Marketing Practice: Why Grouping Customers Is No Longer Enough (2009, C Bailey, P R Baines, H Wilson, M Clark, Journal of Marketing Management, Volume 25, Issue 3 & 4, Pages 227-252).

This study take the view on such questions is under-informed by an adequate understanding of current segmentation practice, and in particular of the role of segmentation within the wider process of customer insight Customer segmentation. The act of apportioning customers into like-minded groups, remains in use in all the case studies for marketing planning purposes: identifying potential target groups, prioritizing these, and developing propositions for them. This paper is based on five service-based multinational companies, it epitomizes only one particular set of practices and apprehensions regarding the act of market segmentation, indicating the set of techniques used to action market segmentation rather than confirming their ubiquity in commerce more generally.

2.10 Application of k-Means Clustering algorithm for prediction of Students' Academic Performance (2010, O. J. Oyelade, O. O. Oladipupo and I. C. Obagbuwa, International Journal of Computer Science and Information Security, Vol. 7 No. 1).

In this study, a system for analyzing students' results based on cluster analysis and uses standard statistical algorithms to arrange their scores data according to the level of their performance had been described. Researchers had also implemented k-mean clustering algorithm for analyzing students' result data. The model was combined with the deterministic model to evaluate the students' results. Researchers had provided a simple and qualitative methodology to compare the predictive power of clustering algorithm as well as the Euclidean distance as a extent of similarity distance. This clustering algorithm obliges as a benchmark to monitor the advancement of students' performance in higher institution. It also augments the decision-making by academician to monitor the candidates' performance semester by semester by improving on the future academic results.

2.11 Clustering Processes (2010, D Ryabko, 27th International Conference on Machine Learning, Haifa, Israel. pp.919-926).

The problem of clustering is measured, for the case when each data point is a sample generated by immobile ergodic process. Researchers proposed a very natural asymptotic notion of consistency, and show that simple consistent algorithms exist, under most general non-parametric assumption. In this study, researchers had proposed a framework for defining consistency of clustering algorithms, when the data comes as a set of samples drawn from stationary processes. The main advantage of this framework is its generality: no assumptions have to be made on the distribution of the data, beyond stationarity and ergodicity. The proposed notion of consistency is so simple and natural, that it may be suggested to be used as a basic sanity check for all clustering algorithms that are used on sequence-like data.

2.12 Clustering Stability: An Overview (2010, U V Luxburg, Max Planck Institute for Biological Cybernetics, T ?ubingen, Germany).

In this paper, researchers had given a high-level overview about the existing literature on clustering stability. In addition to presenting the results in a slightly informal but accessible way, researchers related them to each other and discussed their different implications. Stability can discriminate between different values of K, and the values of K that lead to stable results have desirable properties. If the data set contains a few well-separated clusters that can be represented by a center-based clustering, then stability has the potential to discover the correct number of clusters. While stability is relatively well studied for the K-means algorithm, there does not exist much work on the stability of completely different clustering mechanisms.

2.13 Judging The Quality Of Customer Segments: Segmentation Effectiveness (2010, Dibb, Sally and Simkin, Journal of Strategic Marketing, 18(2) pp. 113-131).

Despite widespread use, developing and implementing segmentation schemes is rarely problem free. Testing the quality and robustness of segments is one of the difficulties, which marketers face. This study uses a longitudinal case study from the Eastern European mobile phone market, the practical application; influence and effectiveness of these segment eminence criteria are scrutinized. The outcomes reveal the value of combining 'hard' statistical and 'soft' segment quality criteria to test the cogency and stoutness of segmentation outputs prior to implementing the segmentation. Implications from this case for managers seeking to select and deploy suitable quality criteria are considered.

2.14 An Approach To Optimized Customer Segmentation And Profiling Using RFM, LTV, And Demographic Feature (2011, Int. J. Electronic Customer Relationship Management, Vol 5).

Customer segmentation and profiling, as the primary stage of CRM (Hruschka, 1986), is an increasingly pressing issue in today's over-competitive commercial area. It uncovers that various groupings of including RFM and demographic factors in clustering ought to be inspected to accomplish the most appropriate one for the segmentation process. Because of significance of LTV in all phases of CRM, in this examination the best system was picked dependent on the LTV scattering quality. At the end of the day, the chose structure had the option to separate groups better among different systems with respect to their LTVs. Lastly, the creative position based visualization strategy in this investigation brought about progressively significant profiles. Planning distinctive marketing plans would be encouraged when the correlation depends on the proposed rank based profiling.

2.15 Mining Customer Knowledge For Channel And Product Segmentation (2013, S H Liao, Y J Chen, And H W Yang, Applied Artificial Intelligence, ISNN: 0883-9514).

This study examines the concept of multichannel service output with regard to different needs and buying behaviors for differentiate channels and products. It takes Computers, Communications and Consumer (3C) products as an example and uses a two-step data mining approach to the cluster analysis and association rules to analyze customer channels and product segmentation. It uses cluster analysis and association rules to analyze the entire database and then divides the customers into two groups with small differences within the groups and large differences between clusters. Thus, this study finds 3C product-buying behavior patterns, as well as customer purchase preferences and customer purchase demands, in order to engender different 3C segmentation marketing alternative. The desired outcome is to reap the benefits of competitive advantage. Working People (Cluster-1) and Students (Cluster-2) both use the Internet community as a source of information. Therefore, researchers proposed that manufacturers should provide information in the Internet community when new products enter the market. In addition, manufacturers can target customers in the working people cluster by advertising in special magazine issues that market to that group.

2.16 Market Research About Customer Segmentation (2013, J Xing, HAMK University Of Applied Sciences).

The purpose of this was to resolve two complications: One is to find out what customer segmentation the researchers

has and to utilize customer segmentation and the 20/80 rule both analysis methods to recognize diverse groups of customers, and provide the finest products to meet individual necessities. Another delinquent is to find out who Tarjoustalo's competitors are and what products do customers buy from its competitors. As a result, the main finding the key customers are over 40 years' old customers with 2-person household size. The number of female customer more than male customer. These customers mainly lives and shops in Forssa city once a week regularly. Some of the customers even visit store almost every day.

2.17 Customer Segmentation Using Unobserved Heterogeneity In The Perceived-Value - Loyalty-Intentions Link (2013, Floh, Arne And Zauner, Alexander And Koller, Monika And Rusch, Thomas, Journal Of Business Research, ISSN 0148-2963).

This study identifies three different segments that are internally consistent and stable across different service industries, using two data sets: the wireless telecommunication industry and the financial services industry. The three segments found were characterized as "rationalists", "functionalists" and "value maximizers". These results point the way for value-based segmentation in faithfulness initiatives and reflect the significance of a multidimensional conceptualization of perceived value, involving psychological and affective components. The outcome validate the fact that assuming a homogeneous value-faithfulness link provides a deceptive view of the market. The paper determines suggestions for promoting examination and practice regarding segmentation, situating, reliability programs and vital collusions. The results of this paper strongly support the argument that perceived value influences behavioral intentions, but also that the effects differ in magnitude depending on the consumer segment. Hence, the basic model, assuming a homogeneous sample, provides a misleading view of consumer evaluations, with regression coefficients reflecting merely the 'midpoints' of given perceptions.

2.18 K-means vs Mini Batch K-means: A comparison (2013, J Béjar, Universit at Politècnica de Catalunya).

Mini Batch K-means has been proposed as an alternative to the K-means algorithm for clustering massive datasets. The benefit of this algorithm is to decrease the computational cost by not using all the dataset each iteration but a subsample of a fixed size. The tenacity of this paper is to perform realistic experiments using artificial datasets with controlled characteristics to evaluate how much cluster quality is lost when applying this algorithm. From the experiments, a first conclusion is that to use the sum of square Euclidean distances to the centroids to compare the quality of partitions from k-means and mini batch k-means as proposed is not adequate. Inconsistent conclusions can be drawn for the influence of the number of attributes from this measure. The adjusted rand index seems a better choice because is independent of the representation of the data and clusters. Mini batch k-means has the main advantage of reducing of computational cost of finding a partition. This cost is proportional to the size of the sample batch used and this difference is more evident when the number of clusters is larger.

2.19 Using Data Mining Techniques In Customer Segmentation (2014,H Ziafat And M Shakeri,Int. Journal of Engineering Research and Applications, ISSN: 2248-9622).

The amalgamation of business domain expertise with the power of data mining practices can benefit organizations gain a competitive advantage in their efforts to enhance customer management. Researchers theoretically discuss about customer relationship management and then utilize couple of data mining algorithm specially clustering techniques for customer segmentation. The researchers concentrated on behavioral segmentation. In this paper, researchers had outlined how data mining can help an organization to better address the CRM objectives and achieve personalized and operative customer management through customer insight.

2.20 Customer Segmentation Research: Case Study ToRegon Ay (2014,V Hyväri Centria University Of Applied Sciences).

This study was made for Regon Ay, a company based in Haapajärvi, Finland. The Purpose of this study was to identify the segments of potential customers for firewood processors. The study aims to explain the most effective combination of marketing methods in order to reach potential customers. Consumers read magazines, use internet and other media to glean more information but above all else, people turn to the people around them for seeking advice and guidance. According to the results in questionnaires, Regon Ay should focus on Push strategy in their marketing where a lot of personal selling occurs and the product is literally pushed to the customer either by retailers or by the company itself.

2.21 Customer Segmentation In Retail Facility Location Planning (2014, Müller, Sven; Haase, Knut, Business Research, ISSN 2198-2627).

In this study, researchers had discussed a facility location model to maximize firms' patronage, while a multinomial logit model (MNL) determines demand. Researchers account for customer segmentation based on customer characteristics. Hence, they were able to reduce the bias to the objective, which is due to constant substitution patterns of the MNL. By an intelligible example, they demonstrate

that the independence from IIA of the MNL may yield false predictions. Assuming, if the customers of a demand point are homogenous, i.e., they exhibit the same observable characteristics, and then there is no need for segmentation. If the customers to be heterogeneous then segmentation of the customers according to their characteristics (income and age, for example) should be employed. By proper segmentation, we are able to reduce the predictive bias of the MNL in terms of market shares.

2.22 Customer Segmentation By Factors Influencing Brand Loyalty And Customer Involvement (2016, T Vebrová1, K Venclová1 And S Rojík, Department Of Management, Czech University Of Life Sciences, Kamýcká 129, 165 21 Prague 6).

The aim of the study is to categorize factors influencing brand loyalty and customer involvement. The researchers' aim is to consider consequent segmentation of customers with respect to various degrees of brand loyalty as well as customer involvement. As a basis for K?means clustering, were used components extracted by the exploratory factor analysis. Using cluster analysis, researchers identified four sections of Czech users of mobile phones with a several degree of brand loyalty and four segments of Czech users of mobile phones with a various degree of customer involvement. These segments were defined according to characteristics describing components of brand loyalty and customer involvement. The results of dispersive analysis in the following table show that the Fisher test for all variables of construct Brand loyalty is much higher than one and is significant. It points to the appropriateness of using cluster analysis. The biggest influence on the formation of clusters by brand loyalty have variables Trust and Commitment. In the segmentation by level of customer involvement has been identified these four segments: Highly involved customers, Medium involved customers, Low involved customers and not involved customers.

2.23 The New Model Of Customer Segmentation In Postal Enterprises (2016, P Kolarovszki, J Tengler, and M Majerþáková, Procedia - Social and Behavioral Sciences 230, ISNN 121 - 127).

This study deals with the new approach to customer segmentation in postal services. This study describes the basic approaches to customer segmentation and designing an advanced CRM model based on the multidimensional matrix. Th i s study proposes a new model of segmentation of the customers in the postal service area by using multidimensional segmentation. It contains a theory, which is comparing and finding connections between theoretical information, traditional view of customer segmentation and between new sights of building individual approach to the customer.

2.24 Monte Carlo Simulation and Clustering for Customer Segmentation in Business Organization (2017, Andry A. and Bellania N., Third International Conference on Science and Technology).

This paper tells the best way to produce customer's pay information and make information cluster to upgrading customer potential by using data. Besides, the outcome brings us understanding into which segments of the customer may unserved appropriately considering about their average income with their spending conduct. In this paper, Clustering technique was applied in one of the branches of Indonesia Telecommunication Company, PT. Telkom Indonesia Makassar city (TWM). TWM objective is to upgrade their advertising systems for home segment customers. TWM system is to break down customer segments in the city dependent on first, customer's salary data and second, customer's payment conduct. The two estimations to help TWM market purchasing power insights.

2.25 Segmentation and Efficient Marketing, Case: Teboil Simpele (2018, R Harmanen, Saimaa University of Applied Sciences)

The aim of this paper is to study customer base of TeboilSimpele; of whom the clientele consists of, have any changes occurred during the two and a half five years and to examine if customer segments can be recognized. In addition, marketing related issues are studied as well, e.g. customers' interest and attitudes towards marketing and effective marketing channels. An ideal result of the research is to identify different customer segments important for the company as well as to find the most suitable marketing networks for each segment in order to focus on correct marketing activities on the correct segment. The results showed that the number of the regular customers had increased significantly whereas the share of random customers had decreased. These results can either indicate satisfied customers and steady base of regular customers affecting on the shares, on the other hand there is also a chance that there is a lack of pass byers and tourists thus affecting on the results.?

3. Research Methodology

Several amounts of data has been gathered to run Machine Learning algorithms for Clustering Analysis. The customer data was collected via surveying customers visiting Malls in Dehradun City based on their age, gender, average annual income and their spending score. The workflow of Customer Segmentation using Machine Learning shown in Figure 1 below:

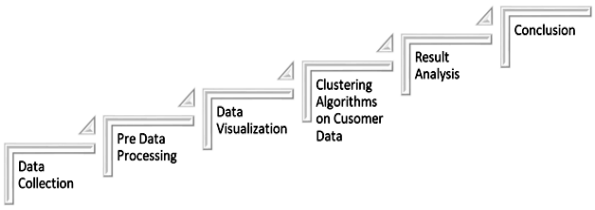

Figure 1: Research Work Flow

The first process is collection of data for the research. For the purpose of this research, data of 200 respondents has been collected. After gathering the data, the noise in data must be reduced by conducting preprocessing activity. The pre-processing step contains various activities for instance data cleaning, data integration, and data reduction to get a better data form. The third process is visualizing the data that is the graphical representation of the data. Data visualization is an open way to perceive and understand trends, outliers, and patterns in the data. After visualization, the next step is to perform clustering process using K-Means algorithm, for customers' data. The fifth process is to analyze the insight from visualization and clustering until we find the pattern or model. Finally, the conclusions can be drawn based on the study. To conduct the analysis, "R language" has been used.

4. Data Analysis & Data Interpretation

4.1 Demographic Characteristics

Demographics	Details	Frequencies	Percentages
Gender	Male	88	44%
	Female	112	56%
Age	Less than 25 years	35	18%
	25 to 35 years	63	32%
	36 to 45 years	36	18%
	46 to 55 years	37	19%
	Above 55 years	29	15%
Annual Income	Less than 250K	20	10%
	250K to 500K	38	19%
	500K to 1000K	108	54%
	1000K to 1500K	32	16%
	Above 1500K	2	1%
Spending Score	Less than or equal to 25	39	20%
	26 to 50	64	32%
	51 to 75	59	30%
	76 to 100	38	19%

Table 1: Summary of Demographic Characteristics

Table 1 shows the gender wise distribution of respondents. Overall, the final sample comprised more female respondents 112 (56%), than males 88 (44%). With this information, it has been presumed that lower percentage of males than females are frequent visitors of shopping malls, convenience stores, supermarkets and grocery shops. The age wise distribution of respondents illustrates the highest number of respondents belongs to the age group of 25 to 35 years (32%) as opposed to less respondents belonging to age group of Above 55 years (15%). The average annual income distribution shows that majority of the sample (108) respondents earns an average annual income of 500K to 1000K (54%).

The present study uses demographics such as age, gender, annual income and average spending score for the basis of customer segmentation.

4.2 Optimal Number of Clusters

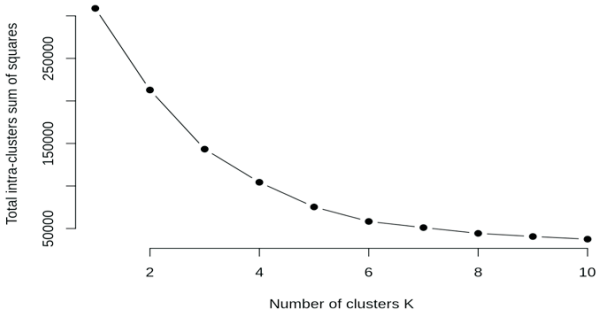

Figure 2: Elbow Method

From the above graph, it has been concluded that four is the appropriate number of clusters since it appears at the bend in the elbow plot.

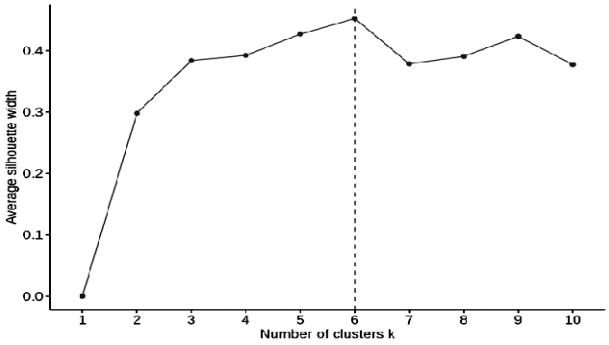

Figure 3: Average Silhouette Method

From the graph (Figure 3), it has been concluded that six is the appropriate number of clusters since it appears at the vertical line showing highest average in the plot.

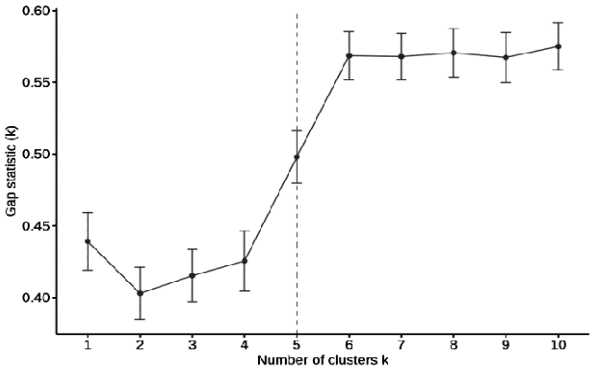

Figure 4: Gap Statistic Method

From the above graph, it has been concluded that five is the appropriate number of clusters since it appears at the vertical line showing highest average in the plot.

5. Testing and Evaluation of Model

Envisioning the clustering results with the two principle components: Annual Income of the sample and Average Spending Score of the sample.

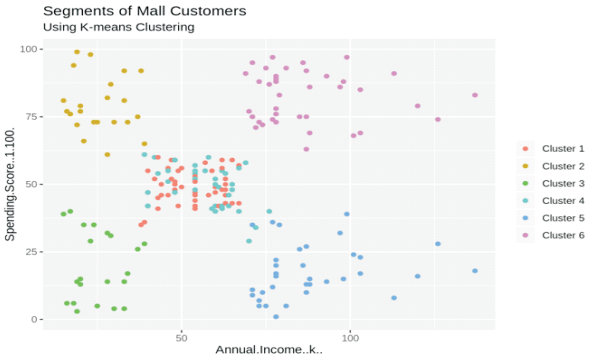

Figure 5: Clusters of Annual Income & Spending Score

From the above figure, it has been observed that there is a distribution of six clusters as follows:

Cluster one and cluster four clusters represent the respondents with the mediocre income salary as well as the mediocre annual spend of salary.

Cluster 6 represents the respondents having a superior annual income as well as a superior annual spend.

Cluster 3 denotes the respondents with a diminutive annual income as well as adiminutive yearly spend of income.

Cluster 5 denotes the respondents with a superior annual income and a diminutive yearly spend.

Cluster 2 represents the respondents with a diminutive annual income but asuperior yearly expenditure.

Envisioning the clustering results using other two principle components: Annual Income of the sample and Age of the sample.

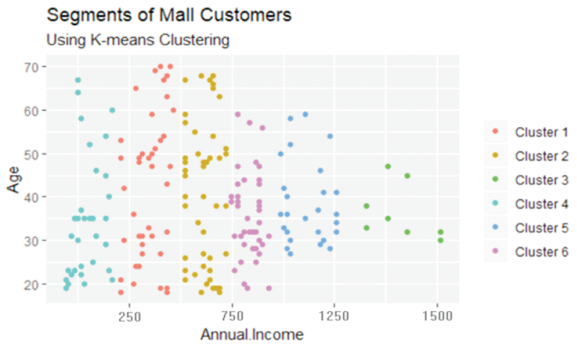

Figure 6: Clusters of Annual Income % Age

From the above visualization, we observe that there is a distribution of six clusters as follows:

Cluster 4 - This clusters represent the respondents with the income salary less than INR 250K and of various age groups.

Cluster 1 - This clusters represent the respondents with the income salary between INR 250K to INR 500K and of various age groups.

Cluster 2 - This clusters represent the respondents with the income salary between INR 500K to INR 750K and of various age groups.

Cluster 6 - This clusters represent the respondents with the income salary between INR 750K to INR 1000K and of various age groups.

Cluster 5 - This clusters represent the respondents with the income salary between INR 1000K to INR 1250K and of various age groups.

Cluster 3 - This clusters represent the respondents with the income salary between INR 1250K to INR 1500K and of age groups from 25 to 50.

6. Conclusion

Utilizing clustering techniques it is easier to understand the variables much better, prompting one to take careful decisions. By identifying of customers, companies can announce products and services that focuses customers based on numerous strictures like income, age, spending patterns, etc. Moreover, intricate patterns such as product reviews could be taken into contemplation for better segmentation.

To become more lucrative, businesses need to be able to distinguish among customers in order to efficiently satisfy the needs of the diverse segments.

It is no secret that some customers are more profitable than others are. Nevertheless, to be profitable over the long period, leaders must have a clear understanding of how profitability links with customer segmentation. Even if one a highly targeted customer demographic in businesses, there are yet variations amongst individual customers. Recognizing these variances will allow businesses to tailor their methodology to the needs of varying customer segments and allow businesses to effectually serve wider group of people.

Customer segmentation model allows for the effective distribution of marketing assets and the maximization of cross and up-selling opportunities. When groups of customers are sent an email that is precise to their needs, it is easier for business organizations to send those customers special offers.

Additional welfares of customer segmentation include staying a step ahead of the competition and ascertaining new products that existing or potential customers could be interested in.

7. Recommendation & Suggestions

In the course of the preparation of this report and the analysis of data, it is asserted that the optimal number of cluster plays a vital role in the Modeling of K-means Algorithm. Hence, it is sturdily recommended to use any of the three aforementioned methods to attain the optimal number of clusters.

It has been realized earlier that segmentation should be used by the business organizations to gain competitive advantage and provide customized products and services to the customers to serve them better. Not all of the customers have the same backgrounds, goals, or buying patterns. Moreover, beyond simple curiosity, there are many good reasons to drill into how and why they differ. By grouping your customers into segments that share certain characteristics, businesses' will not only gain a better understanding of current customer demographics, and discover hitherto-untapped opportunities for better marketing and product development.

8. Annexures

8.1 Annexures - I:R Script

The R Script used for this analysis is accessible from the google drive.

https://drive.google.com/open?id=13D4F6a13jQNDO00DyyQ0FXi9wmy1YN4R

8.2 Annexure - II References

1. Andre H (2011), An Approach To Optimized Customer Segmentation And Profiling Using RFM, LTV, And Demographic Feature, Int. J. Electronic Customer Relationship Management, Vol 5.
2. Andry A and Bellania N (2017), Monte Carlo Simulation and Clustering for Customer Segmentation in Business Organization, Third International Conference on Science and Technology).
3. C Bailey, P R Baines, H Wilson, M Clark (2009), Segmentation and Customer Insight in Contemporary Services Marketing Practice: Why Grouping Customers Is No Longer Enough, Journal of Marketing Management, Volume 25, Issue 3 & 4, Pages 227-252
4. C Marcus (1998), A Practical Yet Meaningful Approach To Customer Segmentation, Journal of Consumer Marketing, ISSN: 0736-3761
5. D Nilsson and J Olsson (2001), Customer Focus through Market Segmentation, Göteborg University, ISSN 1403-851X
6. D Ryabko (2010), Clustering Processes, 27th International Conference on Machine Learning, Haifa, Israel. pp.919-926
7. Dibb, Sally and Simkin (2010), Judging The Quality Of Customer Segments: Segmentation Effectiveness, Journal of Strategic Marketing, 18(2) pp. 113-131).
8. F Van, Raaij, W. and Verhallen (1994), Domain? Specific Market Segmentation, European Journal of Marketing, Vol. 28, ISSN: 0309-0566
9. Floh, Arne And Zauner, Alexander And Koller, Monika And Rusch, Thomas (2013), Customer Segmentation Using Unobserved Heterogeneity In The Perceived-Value - Loyalty-Intentions Link, Journal Of Business Research, ISSN 0148-2963).
10. H Ziafat And M Shakeri (2014), Using Data Mining Techniques In Customer Segmentation, Int. Journal of Engineering Research and Applications, ISSN: 2248-9622).
11. J Béjar (2013), K-means vs Mini Batch K-means: A comparison, Universitat Politècnica de Catalunya).
12. J Xing (2013), Market Research About Customer Segmentation, HAMK University Of Applied Sciences)
13. L. Angelini, D. Marinazzo, M. Pellicoro, and S. Stramaglia (2007), Natural clustering: the modularity approach, Dipartimento Interateneo di Fisica, Bari, Italy
14. M S Sridhar (1988), Customer Characteristics As Criteria For Market-Segmentation In Libraries, E-LIS
15. Müller, Sven; Haase, Knut (2014), Customer Segmentation In Retail Facility Location Planning, Business Research, ISSN 2198-2627).
16. O. J. Oyelade, O. O. Oladipupo and I. C. Obagbuwa (2010), Application of k-Means Clustering algorithm for prediction of Students' Academic Performance, International Journal of Computer Science and Information Security, Vol. 7 No. 117.P Basu And E K Kim (2007), Customer Segmentation In The Medical-Devices Industry, Massachusetts Institute Of Technology
18. P Kolarovszki, J Tengler, and M Majerþáková (2016), The New Model Of Customer Segmentation In Postal Enterprises, Procedia - Social and Behavioral Sciences 230, ISNN 121 - 127).
19. R Harmanen (2018), Customer Segmentation and Efficient Marketing, Case: Teboil Simpele, Saimaa University of Applied Sciences)

20. S Bubeck, M Meila, U V Luxburg (2009), How The Initialization Affects The Stability Of The K-Means Algorithm, arxiv:0907.5494

21. S H Liao, Y J Chen, And H W Yang (2013), Mining Customer Knowledge For Channel And Product Segmentation, Applied Artificial Intelligence, ISNN: 0883-9514).

22. T Finley and T Joachims (2008), Supervised k-Means Clustering, Cornell University

23. T Vebrová1, K Venclová1 And S Rojík (2016), Customer Segmentation By Factors Influencing Brand Loyalty And Customer Involvement, Department Of Management, Czech University Of Life Sciences, Kamýcká 129, 165 21 Prague 6

24. U V Luxburg (2010), Clustering Stability: An Overview, Max Planck Institute for Biological Cybernetics, T ?ubingen, Germany).

25. V Hyväri (2014), Customer Segmentation Research: Case Study ToRegon Ay, Centria University Of Applied Sciences).

8.3 Annexure III - Citation

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.