Customer Segmentation & Recommendations on retail business using machine learning

Students' Names and IDs:

Nafisa Islam– 17101448
Md. Toyeb – 17101399
Rezwana Mahfuza – 17301016
Md. Asaduzzaman Faisal-17301188

Supervisor:

Dr. Md. Golam Rabiul Alam
Associate Professor, Dept. of Computer Science and Engineering
BRAC University

Abstract

Retail business is no longer a new thing. Many people use this business idea to sell and promote their products as well as to pay for products. In this study, we will try to find a new recommended personalization technique that can be used to find potential customers to boost sales. We will review customer segmentation using data, methods, and processes from a customer segmentation research. The objective is to picture positive business situated results of aggregate bunching that will be exhibited on real-live marketing databases. We will try a different approach by introducing a natural model for markets that are governed by a recommendation system and this recommendation from a concerned segment of customers is more likely to be accepted than others. We will also use recommendations using text analysis and opinion mining techniques to enable the extraction of various types of review elements, such as the discussed topics, contextual information, and comparative opinions based on customer segmentation.

Keywords: segmentation, business intelligence, clustering, classification, K-means clustering, heterogeneous, hierarchical, decision tree, recommendation, customers, retail.

1.Introduction

1.1 Motivation

In the business sector, customers are the most significant group. Without satisfied customers who stay faithful and establish their association with the company, there will not be any market opportunities. Due to the CoVID-19 pandemic, many businesses are at a huge loss. The recent data of Forbes shows that the majority of businesses on this year's Global 2000 list have seen a significant drop in their stock prices since last year [9]. According to L. J. 1,2020 [8], even in the USA, one of the strongest countries in the world, small businesses have lost $255 to $431 billion per month. Customer segmentation has operated for several years and based on this segmentation recommendations are made. But we should build a system where customers are classified by their general data and behavioral patterns and after evaluating the data from this a better framework can be recommended. Efficient segmentation generally entails assessing each segment on such factors such as reliability, growth opportunities, scale, usability, responsiveness, and whether consumers in that segment and their marketing activities are compatible with the priorities and resources of the organization. We propose a general method for recommending the best potential customers to be recommended as a target for a particular business sector focused on customer segmentation.

1.2 Problem Statement

It is very traditional nowadays to find successful customers on a specific RFM market field, which is based on three dimensions (Recency, Frequency & Monetary Value). The analysis of the RFM helps to determine when and what value products should be suggested to the client.[5] In comparison, other methods, such as relationship rules and sequential rules, also help to recommend accurate values of products for segmented customers at the right time. It is possible to classify recommendation systems in many distinct ways. One perspective separates them into two categories: systems that gather data on the prior buying activity of the consumer and systems that run on the basis of actual customer purchases and their purchase process activity according to Cao and Li, 2007 [6]. The proposed solutions are divided into three key groups by a common classification: content-based, interactive filtering-based, and hybrid solutions.[7] The content-based framework gathers product specification information, then relies on current

products and buying experience of the customers. Based on the experience of prior orders by customers, mutual filtering-based systems work in the recommendation system. Hybrid solutions use applied techniques in the respective content-based and interactive filtering-based methods to define interface concerns resulting from either of these processes and frameworks and it also uses additional approaches to increase the accuracy of the recommendation framework.
In past researches, the technique of customer segmentation was mostly limited to segmenting the customer base using different clustering methods. The researchers concluded with only segmentation that usually generated some useful insights. In our paper, we intend to advance it to the next stage which will help the retail business entities to make better decisions in terms of making marketing strategies as well as recommending products to their customers in a fruitful way. In addition to clustering techniques like K-means clustering, Hierarchical clustering, we intend to find the most important variable that creates major differences in the segmentation. Furthermore, we also find out the least important variables or attributes that can be merged with another attribute. We are planning to find out these attributes by determining information gain and entropy using the decision tree algorithm. If we can find out the most important attributes, the business organizations can develop their products as well as marketing strategies by focusing on these attributes.

1.3 Research Objectives

The primary aim of our study is to build a decision tree on which recommendations can be made by using appropriate algorithms. This study is intended to create a system to help the companies to identify the diverse user segments that use their services. Through our proposed system business companies can better understand the nature of traditional customers by examining the disparity in spatial and temporal activity between them, and by then grouping or clustering together customers who display common behaviors.
Customer segmentation and recommendation according to the clustering plays a huge role in the business world. Our objective is to retain and expand the market and create the opportunity to increase the profit margin. It can be used for understanding the market situation. Suitable recommendations and offers are the main key to hold the existing customers and gain their trust.

Our research objectives are given in points:

- For better understanding of business.
- To identify least and most profitable customers more efficiently.
- To improve customer service.
- For providing better offers and recommendations for customers.
- To optimize the segmentation process.

1.4 Method and Techniques

To maintain an outstanding relationship with all the customers, nowadays, most of the e-commerce shops are investing resources in Customer Relationship Management. CRM is the best strategy for building, managing loyal, and long-lasting customer relationships. The main goal of the CRM is to understand the need of each customer individually.

Clustering is one of the most popular techniques which is used mostly in order to create segmented groups of customers effectively. In clustering, the data set is divided into groups, and the data points in each group. These data points are grouped together by identifying the correspondences according to the attributes found in raw data. The main goal to find relevant and insightful clusters that can be used for analysis purposes.

In the past, for segmentation purposes, K-means clustering and hierarchical clustering were popularly used to get an efficient result. In our paper, we will use the decision tree algorithm in addition to these two techniques. Our main purpose for using the decision tree algorithm to find out which factor impacts most in the case of customer segmentation is by calculating the information gain and entropy of each variable.

K-means clustering: According to Maimon and Rokach, (2005), "the aim of the K-Means algorithm is to divide M points in N dimensions into K clusters (assume k centroids) fixed a priori.These centroids should be placed in a way so that the results are optimal which otherwise can differ if locations of the centroids change. So, they should be placed as far as possible from each other. Each data point is then taken and associated with the nearest centroid until no data points are pending. This way an early grouping is done and at this point, new centroids have to be recalculated as these will be the centers of the clusters formed earlier. After having calculated these centroids, the data points are then allocated to the clusters to the nearest centroids. In this iteration, the centroids change their position stepwise until no further modifications have to be done and the location of the centroids remains intact" [10].

Hierarchical Clustering: According to Maimon and Rokach, (2005), "hierarchical clustering is a method of cluster analysis that creates a hierarchy of data points as they move into a cluster or out of the cluster" [10].

Decision Tree Algorithm: It is a method of supervised learning that can be used to solve regression and classification related problems. It is a tree-structured classifier, where internal nodes represent the variables of a dataset, branches represent the decision rules and each leaf node represents the final outcome.

2. Literature Review

As the retail sector is growing, the competition between the different retail entities is also growing at a higher rate. So, they are increasing their budget on marketing strategies to achieve this competitive advantage. Customer segmentation is a popular technique to group the customer base into externally distinct and internally uniform segments to create varied marketing strategies for targeting each segment according to its behaviors. Previously, various tools like surveys have been used to measure the qualitative needs as well as the quantitative needs. Though data-driven customer segmentation has some drawbacks it helps an organization to make better decisions in a cheaper way.

According to Bonomo M [1], integration of data from multiple OSNs helps in profiling and profile matching phases with a broader range of groups, implementation of big data technology for storage and control of larger input datasets is successfully seen here. It is really easy to categorize the customers through social media as millions of people depend on social media. The recommendation is centered on the contrast between the related profiles of OSNs to the prospective customers and brands. By analyzing the user generated contents and user actions a general framework for the recommendation of the most relevant users for an advertising campaign can be easily made. But in this case, many future buyers may be missed, as many individuals , particularly those around 50 + years of age, are not accustomed to too much social media usage. [20] Based on their perceptions, they mostly order separate items from the same firms, not relying on social media. Using only social media sites, it is extremely difficult to obtain data on them. Besides, many people may not want to include so much information on social media. Hence, it's a big challenge to extract information depending on On-Line Social Networks focused on profile matching and customized campaigns.

In conjunction with operator data, adopting the k-means clustering algorithm and SVD algorithm to perform customer clustering to achieve the best customer classification in the clustering process by analyzing the behavioral characteristics of different classes as described in the paper [2]. By analyzing the behavioral patterns of customers, we can specify the customers in various classes and a highly efficient suggestion framework can be created by evaluating the trends. The e-commerce company can be defined in a definitive way by evaluating the behavioral characteristics by this and hence stronger outcomes can come from it. The clustering approach described here is undeniably useful, but it is really difficult to collect necessary data to evaluate customer value and the reduction of the data dimension with utmost accuracy.

Again Ramaraju C and Savarimuthu N [3] said by using CPV value the actual dataset consists of both consumer demographic profiles as well as transaction information can be used to improve

predictive accuracy. The classifier model is developed using the market basket data SPSS algorithm. Along with the test data collection, the classifier model is checked and used for customer segment prediction. One of the benefits of this study is specific, non-trivial, accessible information can be derived based on voluminous retail data , allowing decision-makers to make decisions on important business activities. The prediction accuracy is really noteworthy here. One of the major drawbacks is it only depends on the retail data of the customers but customers' preferences may be altered from time to time.

The paper [4] deals with the Upsaily scheme which uses clustering as one of the strategies for evaluating consumer actions in order to promote creation of purchasing recommendations. To optimize the feedback presented to the clients, the Upsaily method also uses classification algorithms. The consumer is characterized by the following characteristics: the frequency of their transaction (dimension of frequency), the number of days after the last order (dimension of recurrence) and the average order value (value of currency). The customer definition is reinforced by data on the number of orders. Based on the assumptions raised, the Upsaily solution will help to gain useful information from your online and offline business.

In another study [5], customer clustering was combined with CF and its nearest K neighbors in this process to search the transactions of sales centers. The CF recommender method has active implementations today. Collaborative filtering (CF) is a widely used approach to recommendation that produces recommendations based on user interest correlates. A new approach to improve the precision of CF was introduced in this report. Customer clustering was combined with CF and its nearest K neighbors in this process to search the transactions of one of the sales centers. K-means has difficulty clustering knowledge when there are different sizes and intensity of clusters.One need to generalize k-means to cluster certain data.

We looked into a paper[18] where we communicate more efficiently with consumers by using market segmentation, giving the company a marketing edge over rivals becomes quite easy. In order to do customer segmentation, many things need to be examined, such as business purpose, data collection, performance assessment which helps one to retrieve a wholesome idea of a business. The more contemporary and new clients are known, the more we can segment them. Similarities can be large variables such as the age or income of customers; they can also be as granular as where customers shop, where they study data, and how they find out about your business.Nowadays,Customers are more likely to be impacted by the ratings for various attitudinal measurements and bank service benefits. According to the writer of this paper, cooperation with banks and influencers is more relevant.We can say that they failed to evaluate customers' satisfaction as they could not pay much attention to the given feedback.

In the paper[15],The importance of behaving with customers according to their background and category has evolved significantly in recent years. Customer segmentation has been used as the basis for customer understanding and classification, and as segmenting customers is the RFM model, the most successful approach has been considered.One of the drawbacks of the previous studies is segmentation is not universal or constant. Geographic segmentation is a portion that competently complements a marketing strategy to target goods or services on the basis of where their consumers live. For example, people living in colder continents such as Europe are more interested in hot clothes, heating devices, whereas people living in hotter continents such as Australia riveted on beach holidays, breezy outfits, and cold drinks. Segmentation needs to be flexible and adaptive to ensure good recommendations and customer care. The company could introduce various goods for that specific market by using geographic segmentation or could also use different marketing tactics to attract the said geography. If segmentation fails to meet these criteria then product inevitability may fall down. Segmentation is not universal because with time buying habits and taste changes which refers that the segmentation needs to keep updated. Not all of your consumers are identical, but there are subsets of clients that share similar characteristics within the uniform collection of your customer base. The practice of grouping customers by those feature vectors is prospect segmentation, also called market segmentation and customer segmentation.

In the study conducted by Jun Xing[21], he said that a wide range of products have been introduced among customers due to segmentation because all companies aim to satisfy their customers' needs and strive heart and soul to support the various customer segments. The main objective is to gain customer loyalty and increase sales for higher revenue and to introduce a new variety of products.The weakness of this study is customers' refusal for sharing information. In this rapidly changing world, people are very concise with time and not interested in sharing information as they are continuously being chased. They feel insecure as chances of misinterpreting and misusing their data are very high and they are less likely to give macroscale personal information that consumes time and energy from a very hectic schedule. It creates hurdles to gather customer information.

This paper[17] deals with a range of challenges in the study. One of them is selecting a suitable model for segmentation. Customer segmentation can be categorized into various techniques, such as RFM technique, target technique which works on a single attribute, unsupervised technique refers to the clustering process, etc. For a particular category of product market, choosing a certain technique for different businesses should be wise and effective. We will use unsupervised methods as we are functioning with a large volume of information with distinct types of attributes in our paper.This paper also taught us a lot of prominent items with challenges, such as how to research rivals and their market strategies, how to increase profits with efficient customer segmentation, how segmenting customers can have social effects, etc. To

segment customers with knowledge of a specific society and their purchasing patterns, this paper allows us to understand how to integrate all these things effectively.

In another study conducted by Abhishek Arunasis Basu[22], he tried to create a framework for customer segmentation that could help the transit agency. Moreover, he added a predictive attrition model to determine the factor which may reduce the customers. In this research[22], he divided all the customers into two main subsections, short-term segment and long-term segment. This type of grouping has a possibility to provide more genuine insight as there is a scope of working further in these subgroups. On the other hand, this work only explores with only k-means clustering which may produce inefficient insights.

As every person is heterogeneous in nature, it is challenging to make some homogenous groups from different types of variables. Main aim of customer segmentation is to determine the most potential customers and to increase the profit margin by targeting the most profitable group of customers.[11] According to K. Tsiptsis[13], to create an optimal customer segmentation and clustering homogenous groups of people have to be made with similar aspects. On the other hand, every person is heterogeneous in nature. Everyone has his/her own different choice, psychology, lifestyle, income, personality, educational background, race, geographic location etc. Some of these traits are quantifiable and some are not. There is a negligible chance of possibility that one person is totally similar with another person. Specially, if segmentation is chosen to create with psychographic details, the data representation is a little bit tough. As a result, it is very hard to create homogenous groups with heterogeneous people. In Ziafat Hasan's words[19], suitable attributes have to be chosen to create these groups or segments properly. One attribute can change the result drastically.

One of the most difficult parts of segmentation is collecting and working with a large number of data with many attributes. In H., Jung's opinion[16], more data is equivalent to more precision. As a result, it is very difficult to collect these data. Customers won't be comfortable to disclose so much information or fill up forms to buy a product. Many retail shops use membership cards or cell phone numbers to track these records. Still, it is very hard to collect these amounts of data. According to Sağlam, B. and Salman's Hwang[14], data filtering is a must in data segmenting. After collecting the data, the data have to be filtered to avoid data-puke. Customer data has to be captured and stored, then deduplication and merging different attributes with common identifiers have to be done. Then corrupted or null values must be checked to drop those data. After that, the dataset must be reformatted by adding some newly generated attributes and dropping unused attributes.

3.Working Plan

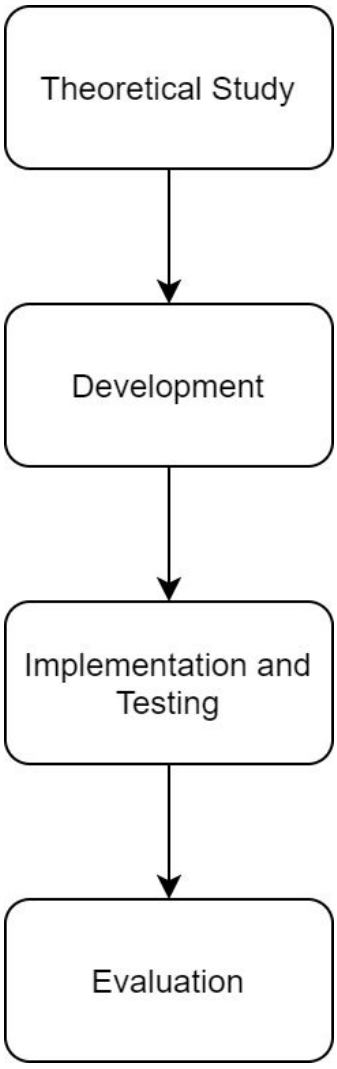

Fig 3.1 : Phases of our whole work

First phase of our work is to study the existing materials about our topic. In the second phase we are going to develop our work. After the development we will implement and test our work. At the last phase, we will come to a conclusion and our work will go through an evaluation process.

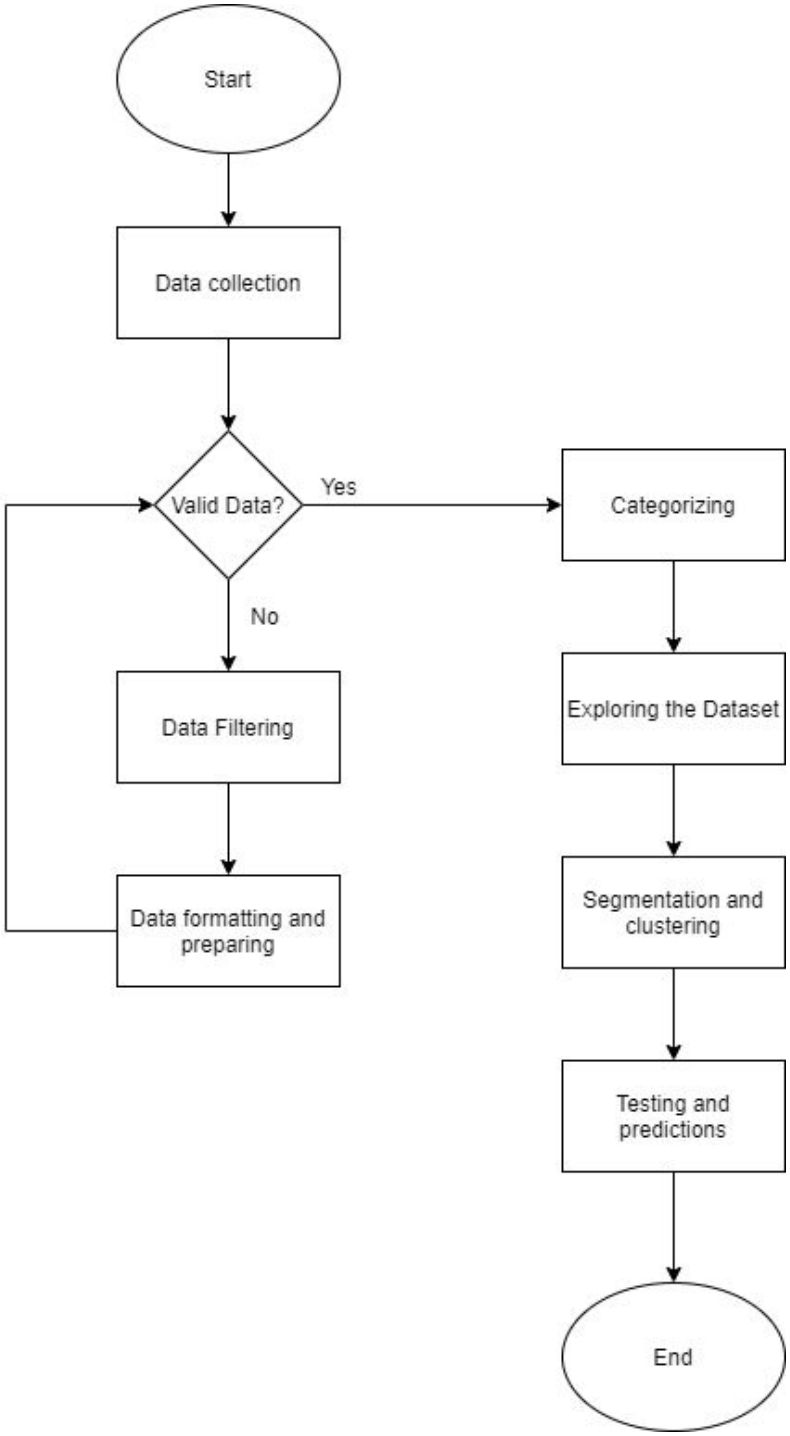

Fig 3.2 :Flowchart of working procedure of proposed system

First, we have to collect our suitable dataset and we will have to prepare it by filtering the dataset to avoid error and invalid data. Then we have to explore the dataset for better understanding and have to categorize the dataset according to our working plan. Then we have to perform segmentation and clustering and test the result.

Conclusion

As a data science tool, we use clustering as a use case, which is a good fit for market segmentation. In our paper, we looked at variables to use for B2C segmentation. When a factor has clean values, when a factor has a long tail of unstructured values, and when a factor has a limited collection of ordinal values, we have analyzed issues that occur. For companies, it is useful to develop marketing campaigns using customer segmentation to increase company sales. In order to recognize trends and classify customer groups, clustering algorithms may work with broad data sets[17]. Various clustering algorithms, specifically centroid based K-Means clustering, RFM technique decides which value products should be suggested to the client's algorithm to perform customer clustering to achieve the best customer classification in the clustering process by analyzing the behavioral characteristics of different classes. The problem with the acquisition of new customers has always been the expense. But that has been updated. Recent developments in prospect segmentation have also made segmentation an important method for attracting new customers through the judicious use of Artificial Intelligence and Machine Learning. Now, with greater cost efficiency than ever before, corporations and mid-market corporations can recognize and meet possibly new consumers. The primary purpose of our study is to create a decision tree on which suitable algorithms can be used to make recommendations. The aim of this research is to build a framework to help businesses recognize the different segments of users who use their services.

Businesses can better understand the essence of typical customers through our proposed framework by analyzing the difference between them in spatial and temporal operation and by clustering together customers that demonstrate similar behaviors. It can provide tremendous discovery of consumer data and insights at one point in time. However, the evolutionary clustering of the can be discussed as well. One strategy is to see how over time the size and an optimum number of clusters grow. In particular, the analysis of the production and optimal clusters may suggest a deeper understanding of how consumers are naturally segmented within a retail environment. This optimality on attributes and segmentation can be archived by the decision tree. Another appropriate implementation will be to add to the existing project Additional cluster reliability and validity measures. Clustering, in its very own way, Essentially, data is a way of exploring; naturally, clustering projects prefer to concentrate on Instead of assessing rigorous loss or gain metrics such as supervised machine learning, it is all about investigating the trends that appear in the results. Many data scientists have proposed a clustering analysis that continues to grow, A selection of steps or instruments to tackle common clustering issues. Computing the silhouette coefficient, constructing a proximity matrix, and converting the clustering results into a decision are traditional ways of assessing clustering. Tree and entropy computation, and inertia measurement. Although these steps do not tell the whole story, the

strengths or weaknesses of the new clustering architecture can be illuminated, and we are also meant to discover these sides.[12]

References

1. Mariella Bonomo, Gaspare Ciaccio, Andrea De Salve, Simona E. Rombo, Customer Recommendation Based on Profile Matching and Customized Campaigns in On-Line Social Networks, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019
2. Bin Luo, Shaofu Lin, Research on The Anonymous Customer Segmentation Model of Telecom, 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC 2020)
3. Chithra Ramaraju and Nickolas Savarimuthu, A Classification Model for Customer Segmentation, pp 3-8
4. Maciej Pondel and Jerzy Korczak, Recommendations Based on Collective Intelligence – Case of Customer Segmentation, pp 74-77
5. Seyed Mahdi Rezaeinia Rouhollah Rahmani, (2016), "Recommender system based on customer segmentation (RSCS)", Kybernetes, Vol. 45 Iss 6 pp
6. Cao, Y. and Li, Y. (2007), "An intelligent fuzzy-based recommendation system for consumer electronic products", Expert Systems with Applications, Vol. 33, pp. 230-240
7. Chen, Y.L. and Cheng, L.C. (2008), "A novel collaborative filtering approach for recomm ranked items". Expert Systems with Applications, Vol. 34, pp. 2396-2405
8. 27 Conn. Ins. L. J. 1, Penn State Law Research Paper No. 14-2020
9. https://www.forbes.com/global2000/#48e3edd335d8
10. O. Maimon and L. Rokach, “Clustering methods”, in Data Mining and Knowledge Discovery Handbook. Boston: Springer US, 2005, pp. 321-352.
11. Bilgic, Emrah & Kantardzic, Mehmed & Cakir, Ozgur. (2015). Retail Store Segmentation for Target Marketing. 32-44. 10.1007/978-3-319-20910-4_3.
12. Ryan Henry Papetti(Fall 2019),University of Arizona,CUSTOMER SEGMENTATION ANALYSIS OF CANNABIS RETAIL DATA: A MACHINE LEARNING APPROACH
13. K. Tsiptsis and A. Chorianopoulos. Data Mining Techniques in CRM : Inside Customer Segmentation. Wiley, 2010.
14. Sağlam, B., Salman, F. S., Sayın, S., & Türkay, M. (2006). A mixed-integer programming approach to the clustering problem with an application in customer segmentation. *European Journal of Operational Research*, *173*(3), 866–879. https://doi.org/10.1016/j.ejor.2005.04.048
15. Mohammadreza Tavakoli,Majid Mobini,Vahid Masoumi,Mohammadreza Molavi,Rouhollah Rahmani,Customer Segmentation and Strategy Development
    based on User Behavior Analysis, RFM model and Data Mining Techniques: A Case Study retrieved from https://www.researchgate.net/publication/330027350_Customer_Segmentation_and_Strategy_Development_Based_on_User_Behavior_Analysis_RFM_Model_and_Data_Mining_Techniques_A_Case_Study
16. Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert Systems with Applications*, *26*(2), 181–188. https://doi.org/10.1016/s0957-4174(03)00133-7
17. Arun Jagota(8 December,2019),Customer Segmentation Data Science retrieved from https://towardsdatascience.com/customer-segmentation-data-science-modeling-210fb36c90bb
18. Balmeet Kaur,Kumar Sharma(April,2019),Implementation of Customer Segmentation using Integrated Approach, International Journal of Innovative Technology and Exploring Engineering (IJITEE)
19. Ziafat, H. (2014). Customer Segmentation. *Data Mining Techniques in CRM*, 70–79. https://doi.org/10.1002/9780470685815.ch5
20. https://www.pewresearch.org/internet/fact-sheet/social-media/

21. Jun Xing(2013),Market Research about customer segmentation,HAMK University of Applied Sciences
22. A.A. Basu (2018), Data-Driven Customer Segmentation and Personalized Information Provision in Public Transit, Massachusetts Institute of Technology.