Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 4, April : 2024

# CUSTOMER SEGMENTATION USING CLUSTERING AND DATA MINING TECHNIQUES

**Mr. Y V NageshMeesala[1] P Viswa Sai Praveen[2], R Haritha[3], P Charvik[4,] P Phaneendra[5]**

[1] Associate Professor, Department of Computer Science & Engineering, Raghu Engineering College, Vishakhapatnam, Andhra Pradesh

[2,3,4,5] Student of B-TECH, Raghu Institute Of Technology , Vishakhapatnam, Andhra Pradesh

Email:- praveenvishwa25@gmail.com , rongaliharitha2@gmail.com , pentakota.charvik@gmail.com , Phaneewhat@gmail.com

## ABSTRACT

Understanding consumer behavior and classifying customers based on their demographics and purchase patterns is essential in today's competitive industry. This is a crucial component of customer segmentation since it helps marketers better target various audience segments with promotional, marketing, and product development strategies. Customer segmentation is the process of grouping customers according to shared traits, spending preferences, and shopping trends. The process of figuring out how to engage with clients in different categories to optimize each client's value to the business is known as customer segmentation. Marketers may connect with each consumer in the most efficient manner by using customer segmentation. Utilizing a vast amount of client data, customer segmentation studies accurately identify discrete customer groups according to behavioral, demographic, and other factors. Unlike supervised machine learning techniques, K-means clustering is an unsupervised approach. This approach is used when the dataset contains unlabeled data. Information that hasn't been categorized or grouped into any of the accessible groups is known as unlabeled data. In customer segmentation, various methods are used to find the ideal number of clusters; however, each method has limitations of its own. For example, the Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm fails when the density of the clusters changes. Research on Recency, Frequency, Monetary Value (RFM) is based on past data rather than forecasts for the future. The usage of labeled data is necessary since the Hierarchical Clustering approach cannot undo previous work. On the other hand, the K-means approach guarantees convergence, initiates the centroid's positions, and promptly adjusts to novel and ideal cluster numbers.

**Keywords:-**
Clustering, dbscan, sptial clustering, Frequency, Monetary Value, Recency

## 1. INTRODUCTION

Using clustering and data mining techniques for customer segmentation has multiple benefits for a project:

Targeted Marketing: Companies can focus their marketing efforts on particular groups by segmenting their client base according to shared demographics, tastes, or behaviors. Higher conversion rates and better return on marketing investment are frequently the results of this focused strategy.

**1.1 Product Customization**: Businesses can tailor their goods and services to better fit the demands and preferences of various client segments by having a thorough understanding of these segments. Increased client happiness and loyalty may result from this.

**1.2 Resource Allocation**: Companies can more efficiently distribute their resources by recognizing high-value consumer categories. They could concentrate their sales efforts, for instance, on market niches having the biggest potential for generating income.

**1.3 client Retention**: Businesses can proactively apply retention measures for at-risk categories by segmenting their client base depending on likelihood of churn. By doing this, you may lower customer attrition and raise client lifetime value.

**1.4 Market analysis**: Understanding market dynamics and trends can be gained through the use of customer segmentation. Businesses can spot new market possibilities and dangers by examining the traits of various market segments.

**1.5 Personalization**: Segmentation enables businesses to deliver personalized experiences to their customers. Whether through targeted promotions, product recommendations, or customer service interactions, personalization can enhance the overall customer experience and drive loyalty.

**1.6 Risk Management**: Segmenting customers based on risk factors such as creditworthiness or likelihood of default can help financial institutions manage their risk exposure more effectively.

**1.7 Overall**, customer segmentation using clustering and data mining techniques is a powerful strategy for businesses to better understand their customer base, optimize their marketing efforts, and improve overall business performance.

## 2. LITERATURE SURVEY AND RELATED WORK

### 2.1 Explore Patterns of Customer Segments

Customer behavior evolves and changes over time in real-world settings. In order to establish efficient marketing strategies, this dynamism must be taken into account while conducting consumer segmentation analyses and other business-related tasks. The study's major goal is to look at the patterns of structural changes in consumer groups. There hasn't been any study done on this subject yet. This is the first research to look at the influence of consumer dynamics on structural changes in segments. The goal of this study is to create a way for describing and explaining this problem. Using clustering and sequential rule mining approaches, a novel strategy is suggested. A newconcept and methodology for identifying distinct sequential rules are also established. The proposed strategy is tested using data from customers.[1]

### 2.2 Discovering New Business Opportunities

For sales and marketing departments inside major organizations, identifying and understanding new markets, clients, and partners is a vital task. Intel's Sales and Marketing Group (SMG) is experiencing similar challenges as it expands into new markets and industries and evolves its present business. To aid SMGs in sorting through millions of organizations across several locations and languages to find relevant routes, sophisticated automation that enables a fine-grained knowledge of enterprises is essential in today's challenging technological and commercial context. We show a system developed in our company that mines millions of public businesses

web pages to provide a faceted client representation. We focus on two key characteristics of the customer that are important for discovering suitable opportunities:Industry sectors (which range from huge verticals like banking and insurance to smaller niches like manufacturing).

### 2.3 RFM Customer Segmentation

The RFM model, which is utilized in the traditional retail business for consumer segmentation, is not ideal for an industry with different social group qualities, hence the RFMC model is established by adding the social relations parameter C. For this empirical investigation, educational e-commerce firm M was chosen, and the k-means algorithm was utilized to cluster legitimate clients of enterprise M, resulting in five unique customer groups and proving the model's usefulness.

### 2.4 High-Dimensional Customer Segmentation

The omnichannel become a hot topic as a result of the rapid rise of e-commerce and clients' growing familiarity with multichannel shopping. To satisfy the present trend ofclient demand, several firm organizations to work on the omnichannel business issue and are progressively devoting their efforts to both online and offline business. As a result, there's no doubting that understanding online customers' purchasing patterns is critical to omnichannel success. Using the RFM (recency, frequency, monetary) model and the k-means clustering technique, consumers' information is retrieved and customers are segmented. To expand the RFM model, we divide total frequency and monetary data into weekly level data, resulting in a reduction in the number ofvariables associated with one week.

### 2.5 Churn Prediction & Customer Segmentation

In the telecom industry, the cost of keeping existing customers is much lower, therefore recruiting new customers is no longer a viable option. Churn control is critical in the telecommunications industry. The purpose of this paper is to give a unified customer analytics method for churn management because there has been limited research merging churn prediction with customer segmentation. The methodology has six components, including data pretreatment, exploratory data analysis (EDA), churn prediction, factor analysis, customer segmentation, and customer behaviorists approach combine churn prediction with customer segmentation to provide telecom operators with a detailed churn analysis that will help them better manage customer churn. Three datasets were used to test six machine learning classifiers. The turnover rate of your customers is the most important factor to consider.

## 3. Implementation Study

1) Outliers that are isolated in low-density regions are identified as outliers byDBSCAN, which merges points that are close together. Two important factors create the model's 'density': the minimum number of points required to generate dense regions min samples and the distance required to establish a neighborhood ep. Larger min samples or lower eps demand a higher density to construct a cluster.

2) RFM Analysis is a marketing method that combines the three aspects of recency,and monetary value to analyze and understand customer behavior. The RFM frequency Analysis will help businesses divide their customer base into a variety of homogenous groups, allowing them to connect with each group using a variety of targeted marketing strategies.

3) A way of grouping comparable components is hierarchical clustering, often known as hierarchical cluster analysis. The endpoint is made up of several clusters, each of which is unique yet shares a lot of similarities

### 3.1 Proposed Methodology

The number of segmentation choices is practically limitless, and they are mostly determined by the quantity of consumer data we have available. It starts with the basics, such as gender, hobby, or age, and progresses to elements such as the amountof time since the user last visited the business or the amount of time spent on website.

i)Geographic: The concept of geographic consumer segmentation is straightforward;it all boils down to the user's geographic location. This may be done in several different ways. You may sort your results by city, zip code, nation, or state.

ii)Demographics: Demographic segmentation takes into account the structure, size, and movement patterns of consumers over time and space. Many businesses create and market products that are based on gender differences. Another important factor is the status of the parents. This type of information may be obtained through customer surveys.

iii) Behavioral: Customer segmentation based on past behavior can be used to forecast future behavior. Customers' favorite brands, for example, or the times of the year when they make the most purchases. The behavioral component of consumer segmentation seeks to understand not just why people buy things, but also how those reasons change over time.

Psychological: The psychological segmentation of customers includes personal traits, attitudes, and beliefs. Consumer surveys are used collect this information,which may then be used to assess customer sentiment

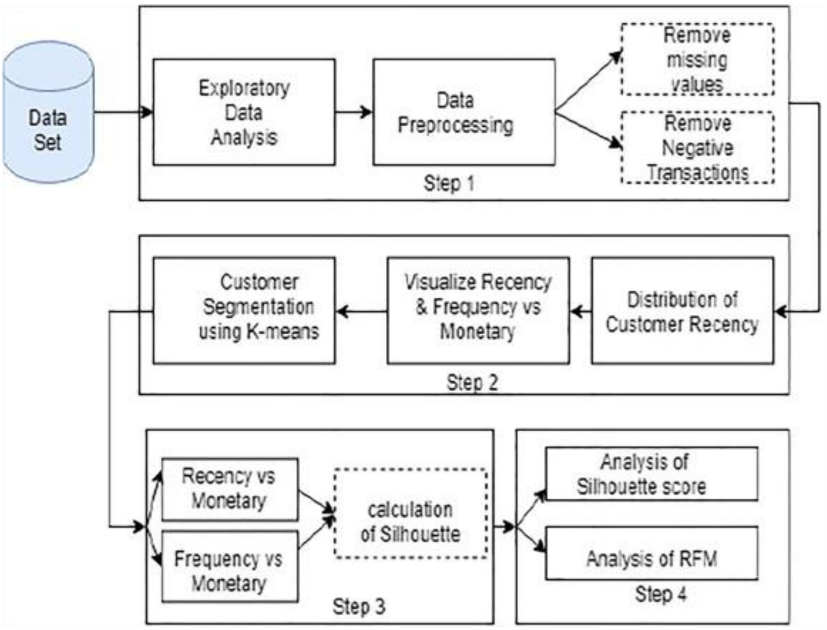

Data Set
Exploratory Data Analysis
Data Preprocessing
Remove missing values
Remove Negative Transactions
Step 1
Customer Segmentation using K-means
Visualize Recency & Frequency vs Monetary
Distribution of Customer Recency
Step 2
Recency vs Monetary
Frequency vs Monetary
calculation of Silhouette
Step 3
Analysis of Silhouette score
Analysis of RFM
Step 4

**Fig 1:- Proposed SYSTEM ARCHITECTURE**

## 4. METHODOLOGY & ALOGRITHAM

i. There must be an aim for everything. You don't want to become involved with this. Blindly. Otherwise, the result will be unorganized and untidy. You'll need a business case instead. It's identifying the most lucrative consumer groupingswithin the total pool of customers in this scenario.

ii. **Prepare the data:** The data must then be prepared in the second stage. What is the size of your data set? In this instance, a hundred, thousand or ten thousand client data is preferable. This is due to the fact that you will be able to see more patternsand trends. You'll also need a set of features based on the most relevant business indicators inside the data collection. The data is then preprocessed to reduce discrepancies, which helps in better data analysis.

iii. **Data analysis and exploration:** Data analysis and investigation is the third stage. This is an important stage since it will allow you to discover some intriguing relationships and trends in your data. You will be able to better comprehend the client's interests and purchase patterns as a result of this, and you will be able to determine which traits are most directly associated with the consumer and, obviously, the business.

iv. **Clustering analysis:** Clustering analysis in the context of a client is the fourth phase in the process. The use of a mathematical model to uncover groupings of similar consumers based on the tiniest differences between customers is known assegmentation clustering analysis. The purpose of cluster analysis within each group is to precisely categorize consumers so that personalization may be used to create more successful customer marketing. A mathematical approach called k- means clustering analysis is a popular cluster analysis tool. The resulting clusters aid in improved consumer modeling and analysis. There are no pre-set thresholds or standards in place for this procedure. The data, on the other hand, exposes the customer prototypes that exist intrinsically inside the consumer base.

v. **Choosing optimal hyperparameters:** The fifth step is to select the best hyperparameters. "Tuning" or "hyperparameter optimization" is the process of selecting the optimal set of hyperparameters for an algorithm. Based on our past work, this is the next stage since it assists us in identifying the most accurate and rewarding client groups.

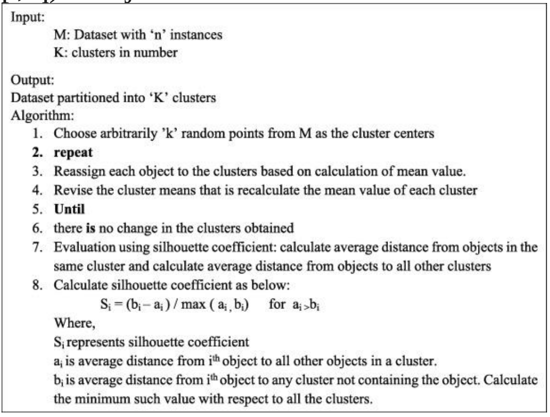
vi. **Visualization and interpretation:** Visualization and interpretation are the final steps in the process. Now it's time to visualize and interpret your findings. Businesses can improve marketing campaigns by targeting features, launches, andproduct roadmaps when they have profitable customer profiles at their fingertips. This gives the company a much clearer picture of which customers are most likely to stick around.

### 4.1 Mathematical model

Clustering using K-means algorithm is a method of unsupervised learning used for data analysis. This algorithm identifies ‘K’ centroids from the dataset ‘D’ and assigns the non- overlapping data points to each of the nearest clusters. The intra-cluster distance is maximum compared to inter-cluster distance in K-means algorithm. Since it is an iterative approach, data points are moved to different clusters, based on the centroid’s calculation.

As per the pseudo algorithm shown in fig , the mathematical model for the manual calculation of silhouette for an object is given below. Consider K clusters of which each cluster contains variable objects. Since K-Means is applied twice in the present experiment, objects are clustered based on a customer transaction data for recency vs monetary and frequency vs monetary values. K=p1,q1,p2,q2…px,qx,p1,q1,p2,q2…py,qy,…p1,q1,p2,q2…pz,qz where,

- K = number of clusters, (p, q) = object in a cluster.



```
Input:
        M: Dataset with ‘n’ instances
        K: clusters in number

Output:
Dataset partitioned into ‘K’ clusters
Algorithm:
    1. Choose arbitrarily ’k’ random points from M as the cluster centers
    2. repeat
    3. Reassign each object to the clusters based on calculation of mean value.
    4. Revise the cluster means that is recalculate the mean value of each cluster
    5. Until
    6. there is no change in the clusters obtained
    7. Evaluation using silhouette coefficient: calculate average distance from objects in the
       same cluster and calculate average distance from objects to all other clusters
    8. Calculate silhouette coefficient as below:
               $S_i = (b_i – a_i) / \max(a_i, b_i)$    for $a_i > b_i$
       Where,
       $S_i$ represents silhouette coefficient
       $a_i$ is average distance from $i^{th}$ object to all other objects in a cluster.
       $b_i$ is average distance from $i^{th}$ object to any cluster not containing the object. Calculate
       the minimum such value with respect to all the clusters.
```

Fig-2. Algorithm -K means.

Identify any point, for example $\{p_1, q_1\}$ in cluster 1. Objects in a cluster represents RFM values. Calculate the average distance from $\{p_1,q_1\}$to all objects of the same cluster (intra distance value $a_i$). Calculate average distance from $\{p_1,q_1\}$to the objects of other clusters as given in the Eq. (1).(1)√∑i=1n(p1-pi)+(q1-qi)2

the minimum average distance from object $\{p_1, q_1\}$to all other objects in the same cluster.

- $b_i$ is the minimum average distance from $\{p_1, q_1\}$to all other clusters, which does not contain $\{p_1, q_1\}$.

comparably calculate silhouette values for cluster 2,3…n by repeating the above steps. The cluster with highest silhouette value is the best as per the evaluation method. Compute the mean silhouette value of all objects to evaluate for whole cluster.

## 5. RESULTS AND DISCUSSION SCREEN SHOTS

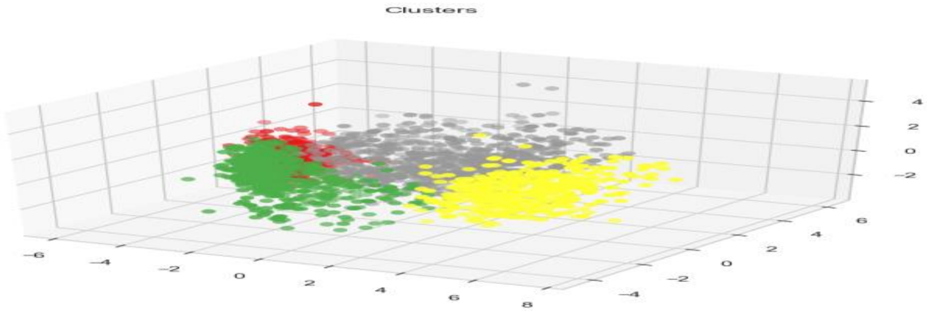


**Fig 2:- forming the data clusters of customer**

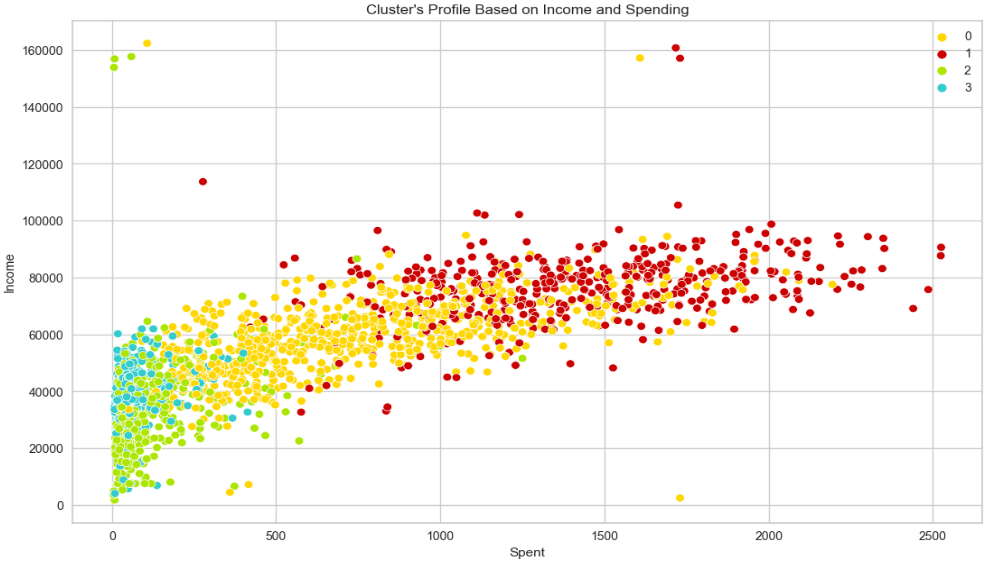


**Fig 3:-** Clusters Profile Based on Income and Spending

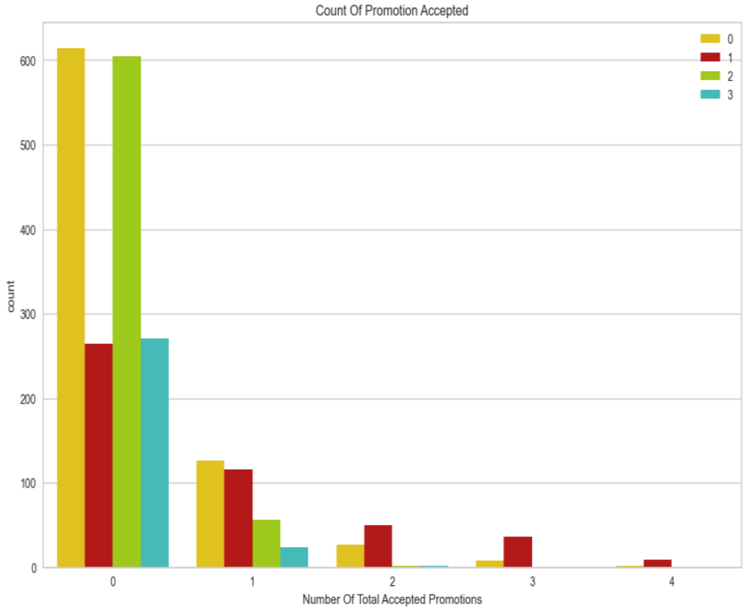


**Fig 4:-** number of total accepted promotions

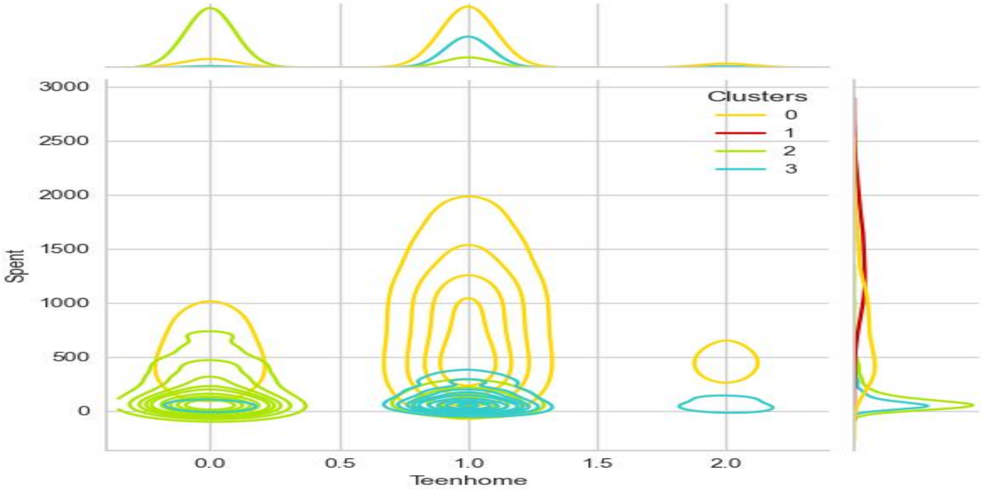


**Fig 5:-** clusters were formed based on promotions

## 6. CONCLUSION & Future work

### 6.1 Conclusion

Our contribution to this research paper is to highlight and solve the challenges of product recommendation based on the customer segmentation engine. Data science can use Customer segmentation to build a stronger relationship with their customers. It enables them to make educated retention choices, develop new features, and position their product strategically in the market.

### 6.2 Future Work

As part of their customer segmentation strategy, retailers with extensive, multi-category offers must show their things in such a way that target customers may search and pick from those offerings. In the first piece of this dissertation, a product segmentation approach is described. The proposed approach offers merchants a methodology for identifying customer-centric, cross-category product segments from large numbers of goods across multiple categories,where products within a segment are bought by the same type of customers. The research also looks at the relationship between the recommended product segmentation approach and a customer segmentation method. Because the approaches are so closely linked, the product and customer groups inferred by each are likely to be identical.

.

## 7 REFERENCES

[1] Wayne Xin Zhao;Sui Li;Yulan He;Edward Y. Chang;Ji-Rong Wen;Xiaoming Li"Connecting Social Media to E-Commerce: Cold-Start Product Recommendation Using Microblogging Information" IEEE Transactions on Knowledge and Data Engineering,volume: 28, pages: 1147-1159, 17 December 2015

[2] Xiaojun Chen; Yixiang Fang; Min Yang; Feiping Nie; Zhou Zhao; Joshua Zhexue Huang "PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data" IEEE Transactions on Knowledge and Data Engineering, volume: 30, page 559-572,26 October 2017

[3] Qinglu Gao;De Xia;Yangyan Shi;Ji Quan"Policies Adoption for Supply Disruption Mitigation Based on Customer Segmentation" IEEE Access, volume: 7, page 47329 - 47338, 29 March 2019

[4] Fuxiang Liu"3D Block Matching Algorithm in Concealed Image Recognition and E- Commerce Customer Segmentation" IEEE Sensors Journal, volume: 20, pp. 11761 - 11769, 19 August 2019

[5] Caroline Gobbo Sá Cavalcante;Diego Castro Fettermann"Recommendations for Product Development of Intelligent Products" IEEE Latin America Transactions, volume: 17, pages: 1645-1652, October 2019

[6] Itay Lieder; Meirav Segal;Eran Avidan;Asaf Cohen;Tom Hope "Learning a Faceted Customer Segmentation for Discovering new Business Opportunities at Intel", In Proceedings of 2019 IEEE International Conference on Big Data (Big Data),24 Dec. 2019

[7] Yong Huang;Mingzhen Zhang;Yue He "Research on improved RFM customer segmentation model based on K-Means algorithm", In Proceedings of 2020 5th International Conference on Computational Intelligence and Applications (ICCIA), 27 August 2020

[8] Yuxuan Yuan;Kaveh Dehghanpour;Fankun Bu;Zhaoyu Wang "A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand" IEEE Transactions onPower Systems, volume: 35, pp. 4026 - 4035, 10 March 2022

[9] Zhenyong Wu;Lu Li;Haotian Liu"Process Knowledge Recommendation System for Mechanical Product Design" IEEE Access, volume: 8, pages: 112795-112804, 16 June 2020

[10] Sahraoui Dhelim;Huansheng Ning;Nyothiri Aung;Runhe Huang;Jianhua Ma"Personality-Aware Product Recommendation System Based on User Interests Mining and Metapath Discovery" IEEE Transactions on Computational Social Systems, volume: 8, pages: 86-98, 24 November 2020