Customer Segmentation using Machine Learning

Razia Sulthana A
Department of Computing and Mathematical Sciences
University of Greenwich, Old Royal Naval College
London, United Kingdom
razia.sulthana@greenwich.ac.uk

Anukriti Jaiswal
Electrical and Computer Engineering, College of Engineering
Carnegie Mellon University
Silicon Valley, California, United States
anukritj@andrew.cmu.edu

Supraja P
Department of Networking and Communication
SRM Institute of Science and Technology
Kattankulathur, India
suprajap@srmist.edu.in

Sairamesh L
Department of Information Science and Technology
Anna University
TamilNadu, India
sairamesh.ist@gmail.com

***Abstract*—Customer segmentation has seen major growth in all sectors in the last decade. Several techniques have been devised to analyze customer behavior through loyalty, purchases, recency, frequency and monetary to develop efficient marketing strategies that cater to each client individually. As the availability of products and services increases, so does the competition. With the spiraling of automation accompanied by its cost-effectiveness and ease of availability, all businesses equip themselves with the required workforce and machinery to conduct experiments such as customer segmentation on an industrial scale. In the proposed work, the datasets are manipulated by extracting features from existing attributes. A widespread approach is RFM that calculates the Recency, Frequency and Monetary values for each customer tuple. This paper aims at laying out a new approach at every step of customer segmentation from pre-processing, clustering, validation and suggesting marketing strategies for customer retention. Two datasets- Online Retail II Set and Mall Customer Segmentation are modelled and the results from analysis of both the datasets are presented and compared to reach a generalized opinion.**

***Index Terms*—Naive Bayes, Decision Trees, Random Forest, K-Nearest Neighbours, Backpropagation, DBScan, Customer Relationship Management**

I. Introduction

Company profits form the foundation of successful businesses. In order to secure ample profits, it is necessary for firms to understand the demands of their clients and provide tailored products and services. This can be achieved by customer segmentation using machine learning. Applying the right marketing tactics to the right set of customers increases the probability of profit maximization. Finding the perfect target group of clients will also prove to increase cost-efficiency since marketing campaigns and resources will not be wasted on unlikely customer bases. Each company must aim to have a distinct business intelligence division that provides comprehensive and coherent visualizations to other branches of the firm. Some of the primary sectors of industry that have benefited from the intelligence of customer segmentation models are retail, marketing, e-commerce, social media and banking. A change in shopping patterns was observed during the pandemic and the need for analysis gained further foreground and so did the concept of Customer Relationship Management (CRM). Segmentation itself can be categorized into three types. These are behavioral segmentation, attitudinal segmentation and demographic segmentation. The first is based on collected behavioral patterns displayed by customers. Observed features involved are browsing patterns, purchasing patterns, past transactions and user interactions. Attitudinal segmentation is based on the intentions and possible actions of customers. Lastly, demographic segmentation focuses on age, gender, religion, and income of customers. The proposed method of segmentation in this paper is of the demographic kind [1].

Data mining helps derive meaningful attributes and even arrive at strong relationships among the data available. The RFM analysis is a data mining technique that computes the values of Recency, Frequency and Monetary. Recency values are based on how recently the customer purchased. Frequency is based on how often the customer purchases and monetary value refers to the amount of revenue generated by the purchases. These values are calculated for each customer sample [2]. An overall score can be computed by assigning weights to each of the RFM values. Equal weights are assigned to the Key Performance Indicators (KPIs) in [2] and [3]. Random weights could be assigned to these values in the case of neural networks, but no research has been able to differentiate between the attributes of recency, frequency and monetary value to determine which of those is more important than the other. Hence, varying weights would not be justified. Correlation can be used to establish the relationship among all attributes. A heatmap is used for this purpose- darker colours indicate a stronger relationship.

Unsupervised learning is the preferred method to conduct customer segmentation. It makes the chances of biased results minimal. Unsupervised learning models interpret the actual structure of data. Classification algorithms such as Naive Bayes, K Nearest Neighbours and Improved Decision trees serve no other purpose than to verify the results from clustering analysis. If a majority of customers are classified and clustered similarly, then the data collected is appropriate and there is no discrepancy between final results. The clustering approaches applied in related works are K-means, mini-batch k-means, hierarchical, Density-based spatial clustering of applications with noise (DBSCAN), (Gaussian Mixture Models) GMM, and MeanShift clustering. K-means clustering is the most simplified and commonly used clustering algorithm. This algorithm makes use of the sum of squared errors, Silhouette and Davies-Bouldin scores for the purpose of determining the ideal number of clusters. Positions of the k-centroids are optimized iteratively. Mini-batch k-means on the other hand is slightly faster than the usual k-means since it takes a small number of random samples to form clusters in every iteration. The results may differ from the k-means algorithm. Hierarchical clustering also comes in very handy in finding the number of clusters through dendrograms. DBSCAN selects two parameters epsilon and minPoints to cluster data points with similar densities. MeanShift algorithm follows unsupervised learning and does not require input parameters to be defined explicitly. The GMM algorithm is a distinct method in which each Gaussian distribution represents a separate cluster. This algorithm determines the number of clusters by minimizing the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) [4]. DBSCAN ignores sparse data and only models dense clusters while MeanShift clustering does not take into consideration all data points available. Hence, both of these clustering methods are infeasible. Hybrid clustering in [5] is so named because of its combination of hierarchical and non-hierarchical methods. The hierarchical clustering method here is the average linkage method that considers the distance between two clusters as the average distance between points in one cluster and another cluster. The non-hierarchical technique is the k-means clustering algorithm.

Rank-based stepwise in [1] is comparatively complex and a rare approach. It comprises five steps- unique classes of each attribute are identified, all attributes are ranked based on the number of unique classes, compute attribute combinations, identify meaningful clusters, and finally validate the clusters. Stream clustering and incremental clustering [6] solve the issue of dynamic clustering for ever-changing data. Any random tuple can act like a mean and the distance of other tuples are calculated from this selected random tuple. A similarity threshold is set. Items with similarity less than the threshold are put in one group and similarities greater than the threshold are placed into another cluster [7]. If any new tuples are added, clustering need not be performed from the beginning as only similarities have to be computed. Besides this, computing the Hopkins statistic is of utmost importance as checking for cluster tendency in data helps see if meaningful clusters can be formed [8].

In the proposed model, we will carry out cluster validation through Silhouette and Davies-Bouldin indicator methods. Essentially, models with the highest Silhouette score and lowest Davies-Bouldin score are selected. All previous existing models have failed to incorporate classification, clustering, dynamic clustering, validation and majority voting all in one model, making simple algorithms very accurate [9]–[12].

II. Methodology

The proposed method will incorporate a 3-step approach for both datasets.

A. Data cleaning

Find and fill missing data points. Standardize data to provide a uniform data format. Further, normalization is done to rescale all numeric values. Feature extraction is also performed using correlation for classification and RFM analysis for clustering [13].

B. Clustering analysis

The data is unlabeled. Hence, unsupervised algorithms will be used. First, Hopkin's statistic is run thrice to check the cluster tendency of data. If the value is high enough, clustering can be performed. The optimal number of clusters is determined using the Silhouette method and the Elbow method by majority voting. Hierarchical, K-means, mini-batch k-means, Birch clustering and Spectral clustering are applied, and the best approach is chosen [14], [15].

C. Validation of performance metrics

Accuracy is the performance metric used for classification models [16]. The Davies-Bouldin index is used to measure the accuracy of clustering algorithms. Further, visualizations are created for the best results obtained. The mall customer segmentation data is used to establish a relationship between customer behavior based on demographics such as age, gender, income and spending amount. On the other hand, the online retail dataset focuses on product consumption and price. Although the data are not related, in linking these two models, we would be able to suggest some insights relating to both models (Fig. 1).

Libraries used: Pandas, Numpy, Seaborn, Sklearn, Matplotlib, and Scipy.

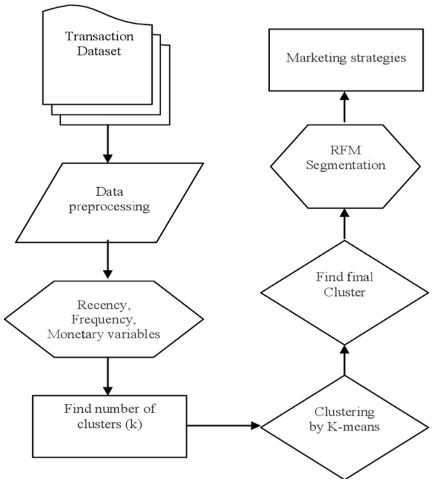


Fig. 1. Machine Learning approach for RFM analysis and clustering

III. Results and Discussion

This research makes use of two datasets- The Online Retail Dataset for 2010-2011 and Mall Customers Dataset [11]. The first data was obtained from a UK based e-commerce company that deals in gift items. Let us discuss the results of the two implementations separately

A. Online Retail

This dataset initially has eight attributes including Invoice number, Stock code, Country, Description, Quantity, Price and Customer ID. First, consumer behavior must be judged through periods of time. So, data is transformed to get cohort month. This determines the user behavior depending on when they began using the product. The number of users retained can be calculated by finding out the difference between purchase month and cohort month. The retention levels of these consumers can be seen in Fig. 2 with 27% in Month 1.

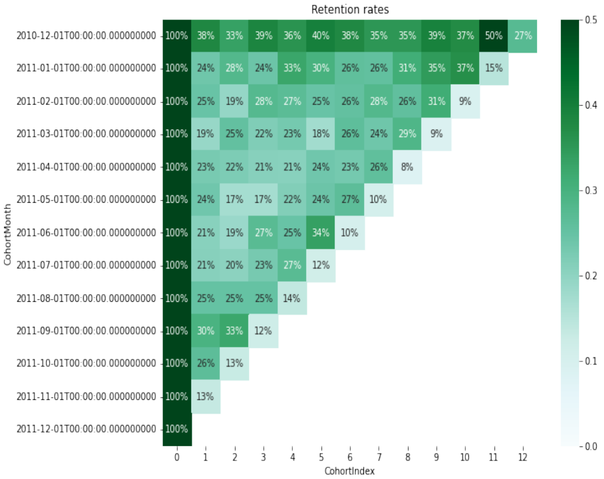
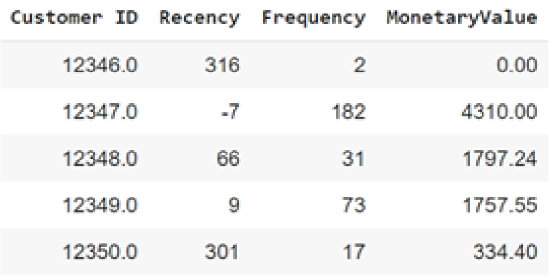


Fig. 2. Retention levels of consumers by cohort month

Next, RFM analysis is conducted on the data and a new data frame is created to observe new values (Fig. 3).



Customer ID	Recency	Frequency	MonetaryValue
12346.0	316	2	0.00
12347.0	-7	182	4310.00
12348.0	66	31	1797.24
12349.0	9	73	1757.55
12350.0	301	17	334.40

Fig. 3. Dataframe with RFM values

The data must be processed further to remove outliers and is standardized. The Silhouette score, Elbow Method and Davie-Bouldin Index are used to determine the optimal number of clusters k. The Silhouette score was highest- 0.4234 for k=5 and lowest- 0.3491 for k=2. In general, the higher the silhouette score, the better the clustering method is. The elbow method determined optimal k to be 4 (Fig. 4).

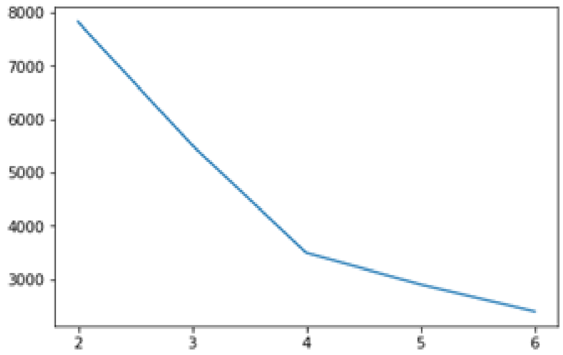


Fig. 4. Elbow curve for dataset 1

Then, several clustering algorithms such as k-means, agglomerative clustering, mini-batch k-means, birch clustering and Spectral clustering are modelled for the best results. The Davies-Bouldin score for each k and clustering model can be seen in Table 1. According to these results and majority voting, the optimal value of k is 5 when used for the k-means clustering algorithm. Finally, a box plot is drawn to visualize the clusters by amount and by recency. The number of data points are more clusters 3 and 5 (highest) by amount while the distribution is fairer in the boxplot by recency (Fig. 5) with the highest number of data points in cluster 2.

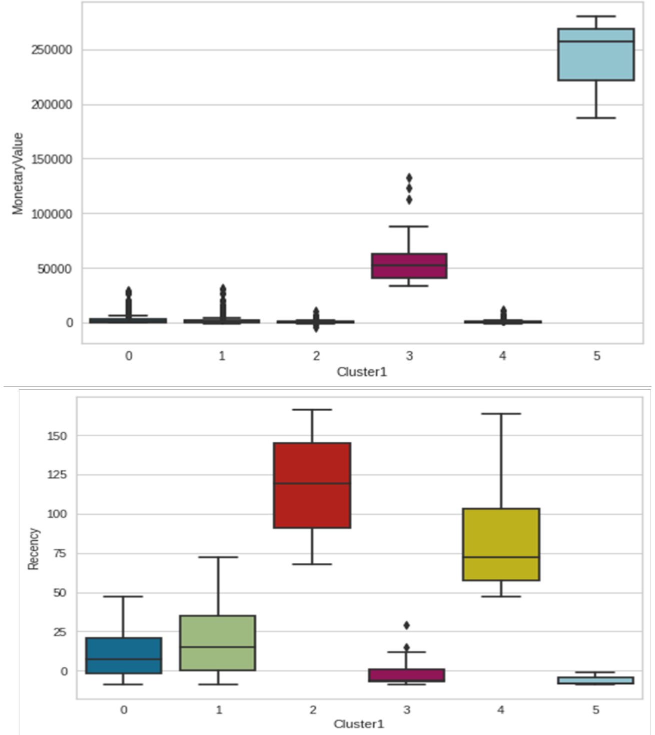


Fig. 5. Box plot by Monetary value and recency

Fig. 6 displays the number of data points in each cluster when modelled through k-means with cluster 2 (shown as 1) at the top.

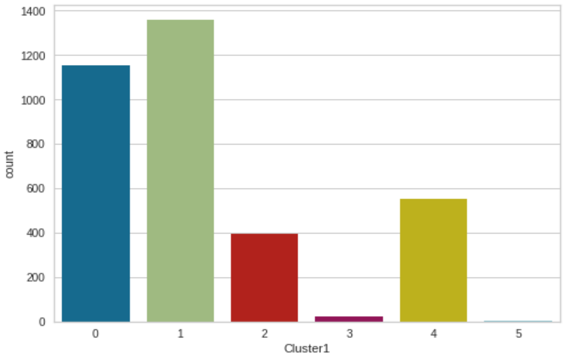


Fig. 6. Bar graph to show data point count for each cluster

B. Mall Customers

This dataset has only five attributes namely- Customer ID, Age, Gender, Annual Income and Spending Score. There are a total of 200 data items only. After preprocessing of data, the distribution of points is visualized using bar graphs by spending score and age (Fig. 7 and Fig. 8).

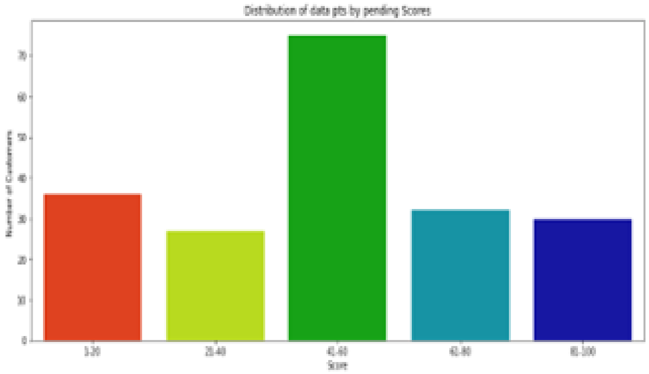


Fig. 7. Distribution of data by spending score

It was observed that most of the customers have a spending score between 41 and 60. Also, most customers are aged between 21 and 40, and on a whole, there are more female shoppers than male. Only one algorithm is used to segment the data. The k-means clustering algorithm and the optimal value of k by the elbow method is found to be 4 (Fig. 9).

Finally, the cluster data is calculated for the dataset, grouped by each cluster, out of which Cluster 2 (shown as 1) had the highest aggregate spending score of 81.487. The number of data points is the maximum in the first cluster with a count of 65 while cluster 3 has only 37 items. Cluster analysis by age is the most uniform followed by annual income in thousand dollars. Finally, a dendrogram representing single linkage hierarchical clustering for the first 20 records of data is shown in Fig. 10 to give an overview of agglomerative or hierarchical clustering.

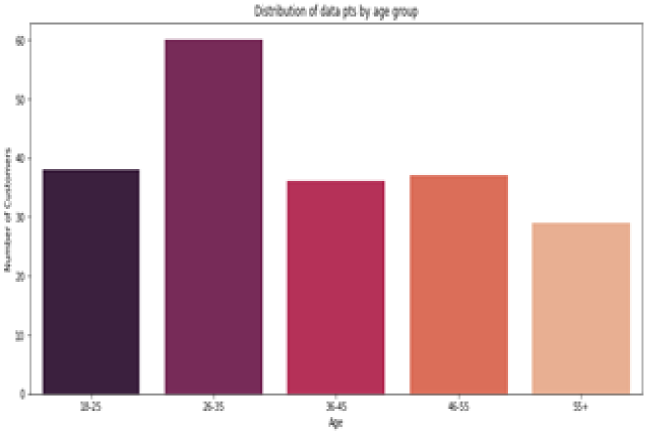


Fig. 8. Distribution of data by age

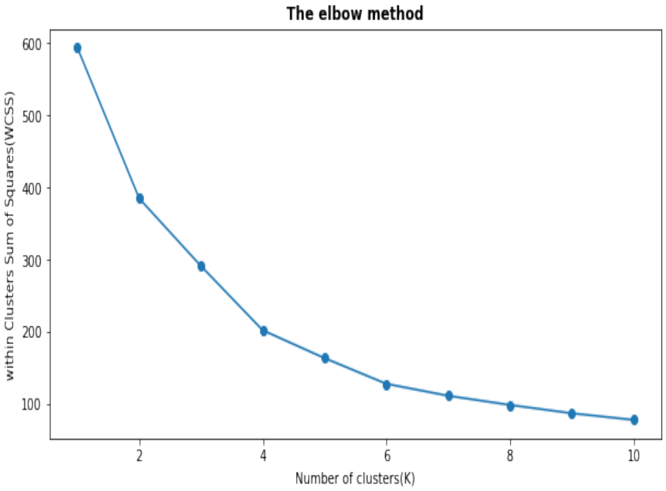


Fig. 9. Elbow curve: k=4

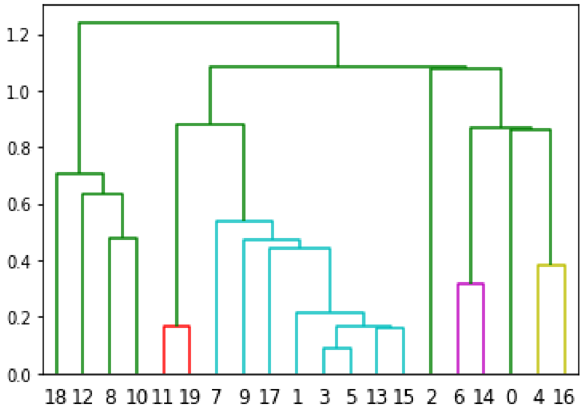


Fig. 10. Dendrogram for single linkage clustering

IV. Conclusion

Customer segmentation can monitor as well as improve retention levels and help develop intelligent strategies to target specific groups of customers. A single dataset with combined features of both datasets used in this research, would prove extremely helpful and accurate in upcoming research. From dataset 1 we realized that clustering is more efficient when done through Recency analysis. From dataset 2, we can conclude that both Age and Spending Score play a vital role is segmentation. Also, different clustering algorithms must be used to determine which works best for which data along with using multiple metrics such as Silhouette method, Davies-Bouldin index and Elbow method to determine optimal number of clusters. It is important to try out different combinations of values of k and clustering algorithms to find the one with the lowest Bouldin index.

References

[1] T. K. Sheng and P. Subramanian, "Proposition of rank-based stepwise interactive visualization for customer segmentation in e-commerce," in *Proceedings of the 2nd International Conference on Software Engineering and Information Management*, 2019, pp. 244–248.

[2] L. Abidar, D. Zaidouni, and A. Ennouaary, "Customer segmentation with machine learning: New strategy for targeted actions," in *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*, 2020, pp. 1–6.

[3] B. E. Cahyana, U. Nimran, H. N. Utami, and M. Iqbal, "Hybrid cluster analysis of customer segmentation of sea transportation users," *Journal of Economics, Finance and Administrative Science*, 2020.

[4] V. Dawane, P. Waghodekar, and J. Pagare, "Rfm analysis using k-means clustering to improve revenue and customer retention," in *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*, 2021.

[5] A. R. Sulthana and R. Subburaj, "An improvised ontology based k-means clustering approach for classification of customer reviews," *Indian Journal of Science and Technology*, vol. 9, no. 15, pp. 1–6, 2016.

[6] M. Phridviraj and C. G. Rao, "A novel approach for unsupervised learning of transaction data," in *Proceedings of the 5th International Conference on Engineering and MIS*, 2019, pp. 1–5.

[7] I. Maryani, D. Riana, R. D. Astuti, A. Ishaq, E. A. Pratama *et al.*, "Customer segmentation based on rfm model and clustering techniques with k-means algorithm," in *2018 Third International Conference on Informatics and Computing (ICIC)*. IEEE, 2018, pp. 1–6.

[8] A. R. Sulthana, M. Gupta, S. Subramanian, and S. Mirza, "Improvising the performance of image-based recommendation system using convolution neural networks and deep learning," *Soft Computing*, vol. 24, no. 19, pp. 14 531–14 544, 2020.

[9] P. Monil, P. Darshan, R. Jecky, C. Vimarsh, and B. Bhatt, "Customer segmentation using machine learning," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 8, no. 6, pp. 2104–2108, 2020.

[10] J. S. Devi and A. R. Sulthana, "Video object segmentation guided refinement on foreground-background objects," *Multimedia Tools and Applications*, pp. 1–17, 2022.

[11] E. Yadegaridehkordi, M. Nilashi, M. H. N. B. M. Nasir, S. Momtazi, S. Samad, E. Supriyanto, and F. Ghabban, "Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques," *Technology in Society*, vol. 65, p. 101528, 2021.

[12] S. Allegue, T. Abdellatif, and K. Bannour, "Rfmc: a spending-category segmentation," in *2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*. IEEE, 2020, pp. 165–170.

[13] B. Kaur and P. K. Sharma, "Implementation of customer segmentation using integrated approach," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 6S, pp. 770–772, 2019.

[14] M. Zhang, Z. Zhang, and S. Qiu, "A customer segmentation model based on affinity propagation algorithm and improved genetic k-means algorithm," in *International Conference on Intelligent Information Processing*. Springer, 2018, pp. 321–327.

[15] R. Sulthana and S. Ramasamy, "Ontology based grouping of products using clustering and classification approaches," *J Adv Res Dyn Control Syst*, pp. 1032–1048, 2017.

[16] A. R. Sulthana and A. Jaithunbi, "Varying combination of feature extraction and modified support vector machines based prediction of myocardial infarction," *Evolving Systems*, pp. 1–18, 2022.