Received: 29 June 2023 | Revised: 21 August 2023 | Accepted: 6 September 2023 | Published online: 8 September 2023

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2024, Vol. 2(1) 165–172
DOI: 10.47852/bonviewJDSIS32021293

BON VIEW PUBLISHING

Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis

Israa Lewaaelhamd[1],*

[1]*Department of Business Administration, The British University in Egypt, Egypt*

Abstract: Machine learning (ML) encompasses a diverse array of both supervised and unsupervised techniques that facilitate prediction, classification, and anomaly detection. Among the many fields of application for such techniques, customer churn prediction is a prominent one. In order to forecast customer switching, data scientists employ a variety of demographic, social, transactional, and behavioral variables and attributes. Unfortunately, many businesses in the United Kingdom still lack the comprehensive and adaptable consumer data required to perform accurate analyses. As a result, they often rely heavily on data produced by enterprise resource planning systems, which is primarily transactional in nature. Consequently, businesses are often limited to modeling and forecasting on transactional data alone and are unlikely to invest significantly in marketing research or other customer-related sources. Businesses are often limited to performing modeling and forecasting on transactional data that are most often not based on advanced techniques like recency, frequency and monetary (RFM) and ML. So, the major objective of the current work is to provide a mix of ML and RFM analysis techniques for churn prediction using mostly transactional data. The dataset was taken from the dataset search website containing online retail datasets. Every customer's RFM scores are computed based on the available data. A churn metric that indicates whether or not the customer has made a transaction in a limited time. Through this paper, different techniques are compared. We used K-means and Density-Based Spatial Clustering of Applications with Noise clustering. By the end of this paper, it may be inferred that the act of dividing customers into six distinct clusters is a more practical and straightforward approach.

Keywords: RFM analysis, statistical approaches, data analysis, machine learning models, artificial intelligence

1. Introduction

The recency, frequency, and monetary (RFM) analysis tool is a relatively straightforward but highly effective means of comprehending and assessing consumer behavior predicated on purchase. The RFM technique functions by quantitatively categorizing and classifying patrons based on the RFM total of their most recent transactions, with the end goal of identifying and targeting the most valuable customers for the purposes of performing focused and precision-targeted marketing campaigns (Shihab et al., 2019; Smaili & Hachimi, 2023). Each consumer is assigned numerical scores based on these parameters, thereby rendering the analysis objective and data-driven. RFM analysis is rooted in the well-known marketing axiom that "80% of your business comes from 20% of your customers" (Alsayat, 2023; Bratina & Faganel, 2023; Chakraborty et al., 2023).

RFM is a strategic approach employed in the analysis and estimation of a customer's worth, predicated upon the evaluation of three crucial data points, namely recency, frequency, and monetary value. The recency metric is indicative of the customer's most recent purchase, while frequency posits the question of how frequently the customer makes purchases. Lastly, monetary value delves into the amount expended by the customer.

RFM analysis is a useful tool that can furnish valuable insights about customers and their behavior. However, it must be noted that this approach does not account for various other important factors that are instrumental in shaping the customer experience. For instance, in-depth targeted marketing strategies may leverage diverse variables such as the type of item purchased or customer campaign responses in order to achieve better outcomes (Bahari & Elayidom, 2015). Moreover, it is important to acknowledge that customer demographics, including but not limited to age, sex and ethnicity, are not taken into consideration by RFM analysis. Therefore, it is imperative for marketers to integrate a more comprehensive and nuanced approach that accounts for a broad range of factors for a more accurate and actionable understanding of customers. There are many ways mentioned in the literature for data integration that may help in this case. For more information about statistical data integration, see Lewaa et al. (2021) and Lewaa et al. (2023).

RFM analysis is a useful tool for gaining valuable insights into customer behavior. However, it is subject to certain limitations. One of the limitations of RFM analysis is its lack of consideration for key factors such as customer demographics or the nature of the purchased items. In light of these limitations, there is a pressing need for a more comprehensive approach that takes into account a broader range of

*Corresponding author: Israa Lewaaelhamd, Department of Business Administration, The British University in Egypt, Egypt. Email: israalewaa@feps.edu.eg

factors. By doing so, a more accurate understanding of customers can be achieved.

Furthermore, it should be noted that the usage of RFM solely depends on the historical data of the customers, thereby implying that it might not be able to accurately forecast the future activities of the customers (Rahim et al., 2021; Seymen et al., 2021). In contrast, predictive techniques possess the capability to unveil the potential customer behavioral patterns which may remain undetected by the RFM analysis (Maryani et al., 2018; Mohammad & Kashem, 2022). This suggests that despite the utility of RFM in analyzing customer data, its limitations in predicting future customer behavior necessitate the incorporation of more advanced predictive methods.

For the current case study, RFM values are easy to calculate and understand, but they cover only one aspect of customer behavior. In order to accomplish high-quality prediction models, data scientists need versatile data about customer needs, opinions, socio and economic characteristics, relationship data, etc. In many cases such data are hard to harvest as small and mid-sized companies do not implement a systematized approach for collecting it.

Instead of conducting an all-encompassing and exhaustive analysis of the entire customer database, it would be more advisable and beneficial to categorize and divide the customers based on their distinct characteristics such as their age or geographical location and subsequently segregate them into a customer group. Through the implementation of a meticulous and well-structured marketing campaign that is specifically tailored to each segmented group, it is possible to fashion a personalized and relevant offer that would be highly appealing to customers that possess a high value to the business.

The process of computing the RFM scores for practical purposes necessitates specialized analytical expertise or advanced mathematical proficiency. Additionally, like any model, the complexity of RFM models can range from rudimentary to sophisticated. The process of RFM segmentation commences by arranging clients in each of the three categories: recency score, frequency score, and monetary score. Conventionally, this is executed on a scale of 1 to 4. A score of 1 designates the uppermost 25% in each category (i.e., the most recent to transact, the most frequent to transact, and those who made the most purchases), with a 3 denoting the following 25%, and so on. By utilizing an RFM scoring system akin to this, one can fabricate an efficacious marketing strategy by creating customer RFM segments.

Among the different approaches presented in the literature for considering customer classification, we fix the data using Box–Cox transformation, which was not used before in the previous work, to ensure the data are normally distributed. Besides, this study is the first study to apply such customer segmentation in United Kingdom. Through this paper, different techniques are compared.

Through this study, we are interested to discover which is the best machine learning (ML) algorithm for customer churn is debatable. Data scientists must explore and assess as many potential candidates as they can in order to choose the best one. This research proposes and evaluates a method for churn prediction utilizing ML algorithms on RFM data. Different input variables have been used to evaluate a number of candidate algorithms.

Customer churn prediction involves utilizing ML techniques to determine which customers are most likely to discontinue their business relationship with a company. This is typically accomplished by training a supervised ML model on historical data such as customer purchase history, account activity, and customer service interactions. The model is subsequently utilized to generate forecasts based on new data, including whether or not a new customer is expected to churn within a specific time frame. Customer churn is a classification challenge, and the ML model can be used to classify whether a customer will churn or not.

2. Literature Review

Jiang and Tuzhilin (2009) have indicated that the enhancement of marketing performances necessitates the implementation of both customer segmentation and buyer targeting. These two interdependent tasks are merged into a systematic approach, albeit faced with the challenge of unified optimization. Consequently, to address this issue, the authors have proposed the application of the K-Classifiers Segmentation algorithm. This particular approach emphasizes the allocation of additional resources to customers who provide higher returns to the organization. A multitude of authors have contributed to the literature on diverse methodologies for segmenting customers. In their scholarly work, Jiang and Tuzhilin (2009) propose an innovative method for clustering customers that diverges from the conventional practice of relying solely on computed statistics. Instead, their approach taps into the transactional data of multiple customers to achieve a more direct clustering outcome. The authors also reveal that the task of identifying an optimal segmentation strategy is, in fact, NP-hard and, thus, necessitates the development of various sub-optimal clustering techniques, which Tuzhilin thoughtfully devised. Subsequently, the authors meticulously scrutinized the customer segments that were generated via the direct grouping method and found that this approach yielded far superior results compared to the traditional statistical approach.

Shah and Singh (2012) have introduced a novel clustering approach that shares similarities with the K-means algorithm and K-medoids algorithms. These algorithms are recognized as partitional techniques. The newly proposed algorithm, however, does not guarantee an optimal solution in all circumstances. On the other hand, it reduces the cluster error criterion. According to Saurabh's observations, the time required for executing the new approach decreases with an increase in the number of clusters, which is a significant improvement over traditional methods.

Cho and Moon (2013) put forth a proposal for a recommendation system that is customized to the needs of the users. The proposed system employs the technique of weighted frequent pattern mining, which is aimed at identifying the patterns that are most frequently occurring. In order to identify the potential customers, the authors have carried out customer profiling through the RFM model, which is widely used for this purpose. The proposed system utilizes varied weights for each transaction to generate the association rules that can be obtained from the mining process. By using the RFM model, the accuracy of the recommendations provided to the customers can be enhanced, thereby leading to an increase in the profits of the firm.

Lu et al. (2014) conducted an in-depth analysis, centered on customer churn prediction. The authors skillfully employed logistic regression and effectively isolated the transactional data to produce an entirely novel prediction model. Through their experimental implementation, the astute researchers observed that customers with the highest churn value can be identified and retained through the deployment of individualized marketing strategies. On the other hand, Zhang subscribes to the notion that the identification of the root cause for customer churn behavior and the satisfaction of individual needs is an indispensable prerequisite for the sustainable existence of any company.

He and Li (2016) have proposed an intricate and multifaceted three-dimensional methodology aimed at enhancing customer

lifetime value, augmenting customer satisfaction, and positively influencing customer behavior. Through their extensive research and analysis, the authors have astutely observed that consumers are indeed a heterogeneous group, each with their unique set of needs, desires, and aspirations. To cater to this diversity and intricacy of customer preferences, segmentation techniques can be effectively employed to identify and understand their demand and expectations, which in turn facilitates the delivery of superior services to these valuable customers. Sheshasaayee and Logeshwari (2017) developed a novel and innovative approach that is integrated, aimed at segmentation through the use of the RFM and lifetime value methods. This approach was carried out in two distinct phases, with the first phase being a statistical approach and the second phase entailing the performance of clustering. The primary objective of this approach was to perform K-means clustering after the two-phase model and subsequently employ a neural network to enhance the overall segmentation process.

Zahrotun (2017) utilized customer data obtained through online channels to conduct identification of the most superior customers by means of customer relationship management (CRM). This application of the CRM paradigm in the context of online shopping allowed the author to effectively pinpoint potential customers through segmentation, which in turn contributes to the maximization of company profits. To facilitate the accurate implementation of customer segmentation and marketing strategies, the fuzzy C-means clustering method was employed. This methodology ultimately affords customers the opportunity to receive specialized amenities across multiple categories, all in accordance with their unique needs and preferences.

3. Methodology

The data for the experiments have been extracted from Online Retail Dataset. For every customer, a set of features and metrics have been calculated, as "recency" (recency from the ending date), "frequency" (number of transactions till the ending date), and "monetary" (total amount of transactions till the ending date) (Chaudhary et al., 2022; Gustriansyah et al., 2020).

RFM scores have been calculated with its RFM analysis feature. RFM scores are calculated using both nested and independent binning with 4 bins. A modification of RFM analysis could be done by using K-means clustering instead of this classic approach. This approach is compared with the classic one (Li et al., 2022; Shirole et al., 2021; Wu et al., 2021). The application of ML algorithms was performed in Python (Joung & Kim, 2023; Khajvand et al., 2011).

Any endeavor aimed at manipulating raw data in order to optimize it for subsequent data processing operations is known as data preprocessing, which is an integral part of data preparation (Shim et al., 2012; Weng, 2017). This process has consistently been recognized as a pivotal initial stage in the data mining process. Recently, data preparation techniques have undergone significant modifications in order to facilitate the training of artificial intelligence (AI) and ML models, as well as to enable conducting inferences against these models.

The process of data preprocessing involves a series of operations aimed at transforming data into a structured format that can be readily processed in various data science tasks, such as ML and data mining, in a more expeditious and efficient manner (Wu et al., 2022; Yoseph & Heikkila, 2018). It is a critical step that is typically implemented at the outset of the ML and AI development pipeline to guarantee dependable findings.

Based on the data description provided in Table 1, the aim of our study is for the United Kingdom. So, we select all observations for online retail related to this country. Then, null values were checked as a step of preprocessing. A null value, which is commonly encountered in the context of relational databases, represents a scenario wherein the value contained within a particular column is either unknown or missing. It is important to note that a null value should not be conflated with an empty string, which is characteristic of character or datetime data types, nor should it be mistaken for a zero value, which is typical for numeric data types. Such distinctions are crucial to ensure the accurate interpretation and manipulation of data in a given database. In our data for the United Kingdom, we have 133,600 null values out of the total sample size of 495,478.

Table 1
Online retail dataset description

No	Attribute name	Description
1	Invoice number	6-digit unique number for each
2	Stock code	5-digit unique number for each product
4	Quantity	Quantity of product per transaction
5	Invoice date	Invoice date and time
6	Unit price	Product price per unit
7	Customer ID	5-digit unique number for each customer
8	Country	Country name

RFM values will be calculated. Recency is calculated by counting how many days exist from the maximum date for each customer to the maximum invoice date. Frequency factor is then calculated by considering the number of customer ID is repeated. Moreover, the total price is calculated for being able to calculate monetary. Total price for each customer is calculated by multiplying quantity by unit price. Then, the monetary for each customer is calculated by summing the total price for each customer till the end date.

So, the fitness of models to study the behavior of the customer based on various sets of input factors has been explored through a number of tests using various input variables:

(I) RFM values;
(II) R. F, M, RFM scores;
(III) R. F, M, RFM scores, Count of Objects per every customer.

The above various inputs are created for the two methods for comparing them together and reach to the best method for segmenting type of customers.

Each customer is allocated with a triumvirate of discrete scores for the temporal proximity, frequency, and monetary variables. The act of scoring is executed on a graduated scale from 1 to 4. The uppermost percentile is bestowed with a score of 4, and the remaining patrons are awarded scores of 3, 2, and 1 correspondingly. The scores are capable of being assumed to possess idiosyncratic attributes as itemized in Table 2.

Table 2
Scores of RFM

Score	Characteristics
1	Potential
2	Can't lose them
3	At risk
4	Lost

Ultimately, all of the customers are furnished with scores ranging from 444, 443, to 111. The customers who obtain a score of 111 may be referred to as the potential customer of the entity as they are anticipated to furnish a greater quantum of proceeds to the entity, and conversely, the customer possesses a score of 444. Based on this RFM score, each customer can be sorted by its own segmentation.

4. Results

Through this section, data visualization is presented. Also, the normality issue is solved using the Box–Cox transformation. In addition, the results after applying a number of ML approaches, K-means, and DBSCAN are presented. Also, a comparison between these methods is considered.

4.1. Data visualization

Before conducting the clustering process, an investigation was done for the distribution of RFM values.

It is very clear from Figures 1, 2, and 3 that we have outliers. With the use of the statistical method known as the Box–Cox transformation, our goal variable is changed such that your data closely resemble a normal distribution.

Figure 1
Box plot for recency values

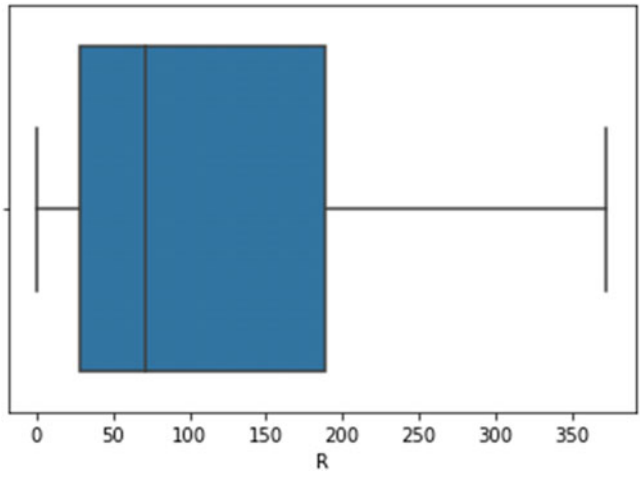


Figure 2
Box plot for frequency values

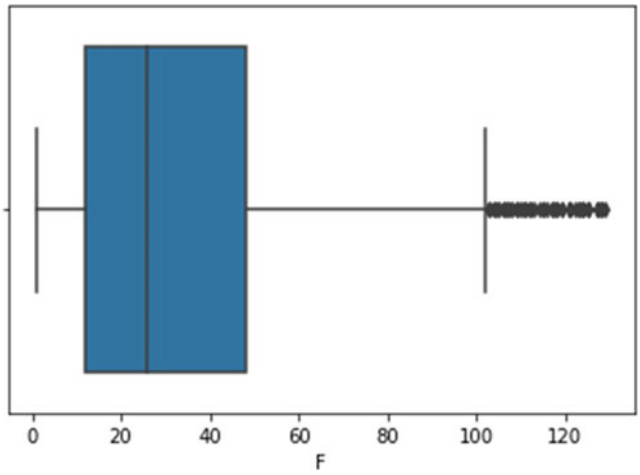
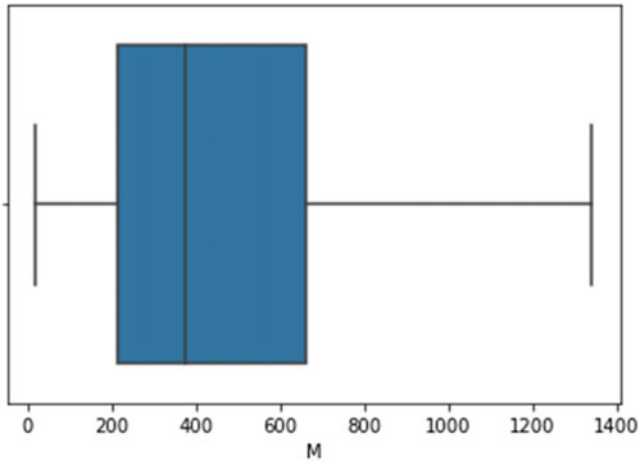
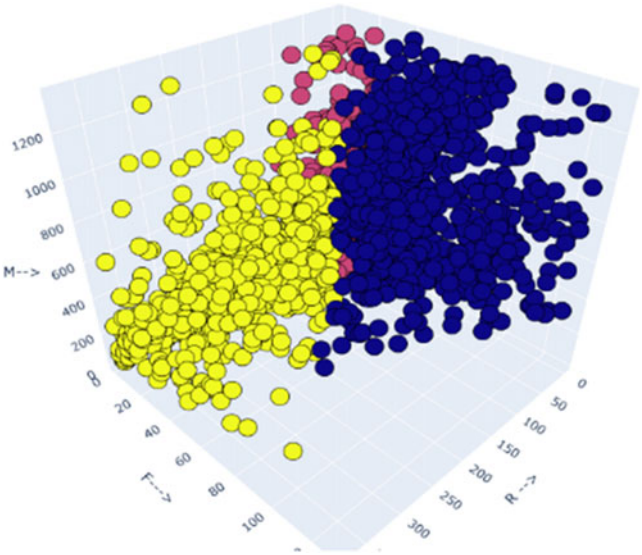


Figure 3
Box plot for monetary values



From Figure 4, we can notice that there are few outliers that exist which may affect the results of clustering. Outliers are observations that deviate significantly from the norm, either in a positive or negative direction. These aberrant data points can exert a disproportionate impact on statistical outcomes, particularly on measures of central tendency such as the mean, thereby potentially leading to erroneous inferences and conclusions. So, we remove these outliers and run our clustering methods without it.

Figure 4
3D for the three values after correcting the outlier problem



4.2. Approach Q-CUT

R, F, M and RFM scores are imputed according to Q-CUT method and after obtaining RFM values. Each score for recency took scores from 1 to 4. The first quarter of the values of recency takes 1 and the fourth quarter takes 4. Therefore, All scores with

value one represented the customers with the highest category (i.e., 111) and has made a transactional acquisition shall continue to retain the said product in their cognitive faculties, thereby increasing the probability of their propensity to either procure or utilize the product at a future instance. In the same manner, frequency and monetary took scores from 1 to 4 where the first quarter of the values of frequency and monetary takes 4 and the fourth quarter takes 1. Frequency distribution for RFM scoring after considering Q-CUT method is summarized in Table 3.

Table 3
Frequency distribution for RFM scoring

RFM scoring	Frequency
444	186
111	169
344	105
433	100
211	96
...	
441	8
241	7
431	7
314	6
414	2

4.3. Clustering approach using K-means

The primary objective of the K-means clustering, which is a widely used vector quantization approach that emerged in the signal processing domain, is to divide n observations into k clusters, wherein each observation is assigned to the cluster that has the closest mean value (also known as cluster centroid or cluster center). The outcome of this process is the creation of Voronoi cells, which partition the data space. It is worth noting that the geometric median is the only measure that minimizes Euclidean distances. Nevertheless, K-means clustering is effective in minimizing within-cluster variances (squared Euclidean distances) rather than regular Euclidean distances, which is a more complex Weber problem. Consequently, k-medians and k-medoids can be employed to obtain superior Euclidean solutions (Ahmed et al., 2020).

In the realm of cluster analysis, which is a widely used technique for grouping objects into similar subsets in an unsupervised manner, the elbow method has emerged as a popular heuristic for determining the optimal number of clusters in a given dataset. The essence of this approach involves generating a plot of the explained variation, which is a measure of the total variance accounted for by the clustering algorithm, as a function of the number of clusters considered. Subsequently, the point on the curve where a noticeable change in the slope occurs, known as the elbow of the curve, is selected as the most appropriate number of clusters to employ. It is worth mentioning that this method can be extended to other data-driven models, such as principal component analysis, where one aims to capture the maximum amount of variance in a dataset using a smaller set of variables, by utilizing the same underlying principle of identifying the elbow point on the relevant curve.

The initial stage of the clustering algorithm is to select a set of k random points carefully as the initial centroids, based on a predetermined value of k. Then, each and every data point in the dataset is subjected to a thorough evaluation process, whereby the Euclidian distance between the data point and the previously chosen centroids is meticulously computed. As a third step, upon completion of the distance evaluation process, the computed values are carefully compared to determine which centroid has the shortest Euclidian distance value, after which the data point in question is assigned to the corresponding centroid, utilizing an efficient and effective assignment mechanism. Finally, the preceding steps are then meticulously repeated, with the ultimate goal being to obtain an optimal clustering solution. The entire process is carefully monitored and is only halted once it is determined that the clusters obtained are identical to those obtained in the previous iteration (Sinaga & Yang, 2020).

The input for this analysis is the customer dataset containing "*n*" instances *k*, which is the number of clusters. The output is the customer data partitioned to *k* clusters.

Based on Figure 5, the best number of clustering is three. So, we are going to make customer segmentation to Figures 6, 7, and 8.

Figure 5
The elbow method for the second method

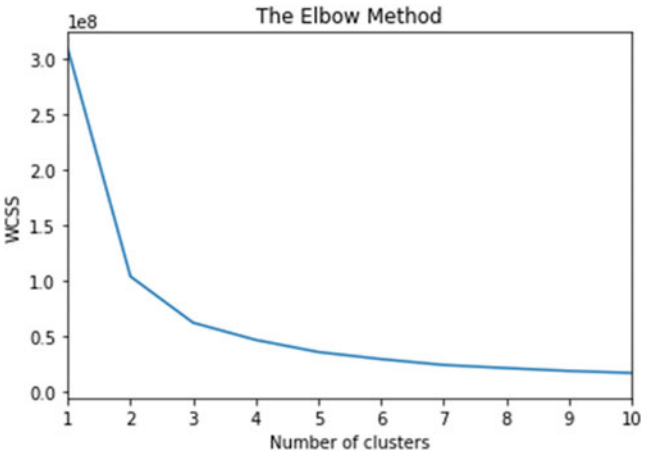


Figure 6
Boxplot within clusters of customers among recency values

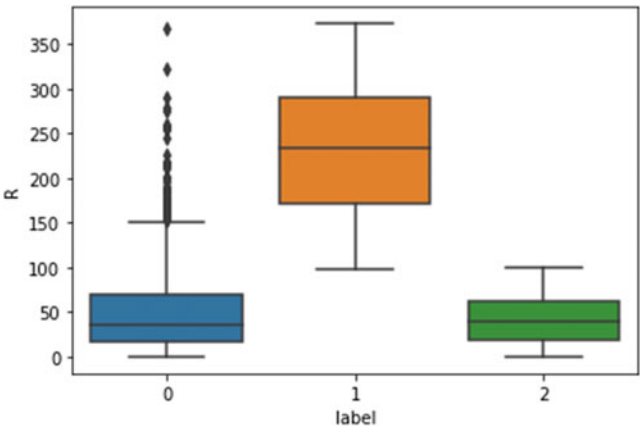


Figure 7
Boxplot within clusters of customers among frequency values

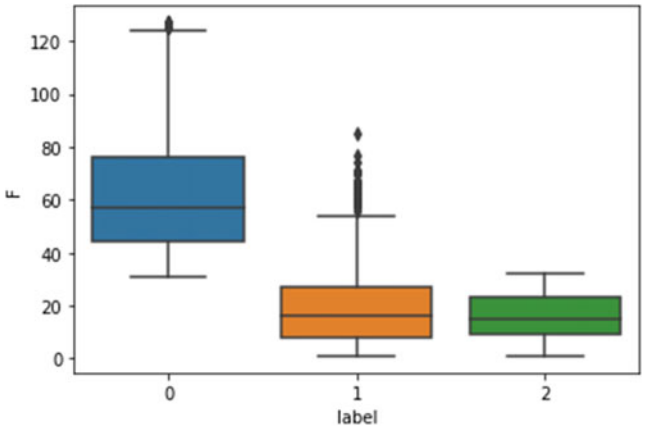
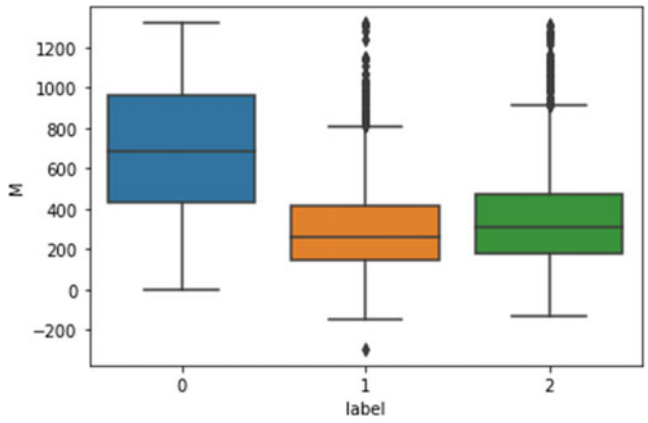


Figure 8
Boxplot within clusters of customers among monetary values



4.4. Clustering approach using DBSCAN

The present algorithmic methodology for DBSCAN clustering initiates by randomly selecting a data point within the dataset and iteratively repeating this process until all points have been visited. If a minimum number of "minPoint" data points exist within a specified radius of "ϵ" to the chosen point, then all such data points are deemed to belong to the same cluster. Finally, the process involves the iterative expansion of the clusters through the repeated computation of the surrounding area for each adjacent point. Figure 11 shows the full process of clustering approach using DBSCAN.

Through Figures 9, 10 and 12, it may be inferred that the act of dividing customers into six distinct clusters is a more practical and straightforward approach.

Figure 9
2-D clustering using DBSCAN

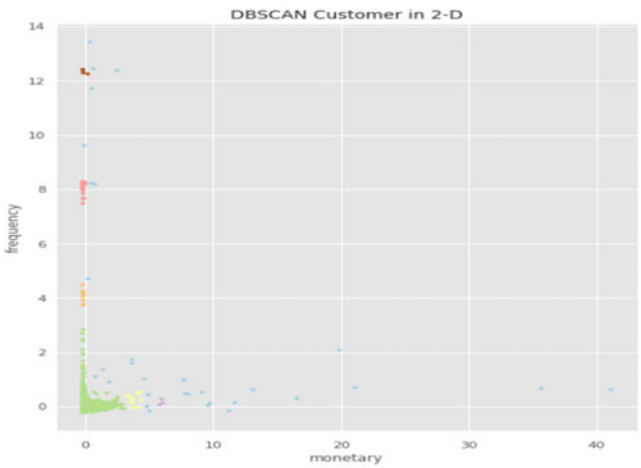


Figure 10
3-D clustering using DBSCAN

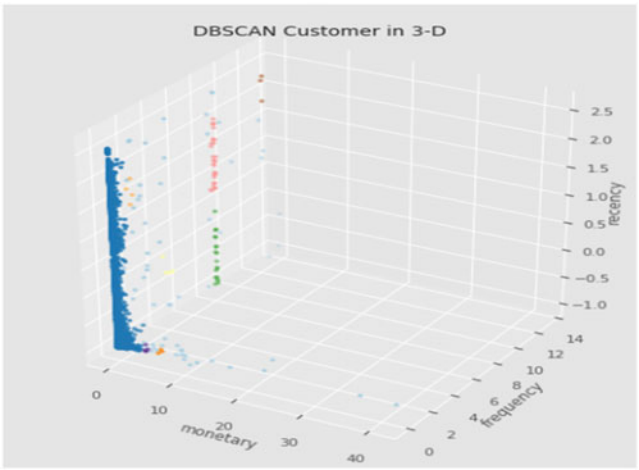


Figure 11
DBSCAN clustering

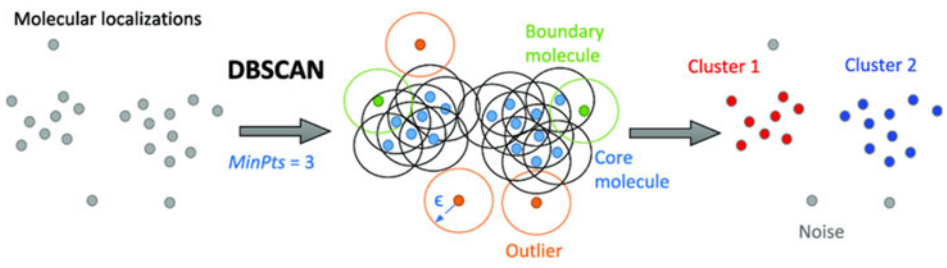


Figure 12
Clustering using DBSCAN among frequency

5. Conclusion

The major objective of the current work is to provide a mix of ML and RFM analysis techniques for churn prediction using mostly transactional data. The dataset was taken from the online retail dataset. Every customer's RFM scores are computed based on the available data. A churn metric that indicates whether or not the customer has made a transaction in a limited time. This is the first study among different approaches presented in the literature to consider customer segmentation to deal with the outlier using Box–Cox transformation. Besides, this study is the first study to apply such customer segmentation in the United Kingdom. Moreover, we make a transformation to resemble a normal distribution and this step improves the results of the study.

Through this paper, different techniques are compared. We used K-means and DBSCAN clustering. By the end of this paper, it may be inferred that the act of dividing customers into six distinct clusters is a more practical and straightforward approach. This is further substantiated by the evident separation of the groups as depicted in the various clustering method plots. Consequently, it is now incumbent upon marketing managers and customer insight teams to deliberate on the most effective mode of communication or promotional strategy to be utilized, with the intention of converting individuals from one segment to another or potentially directing more customers toward a new segment, ideally positioned at the top right corner of the plots.

In future work, data scientists can improve the prediction process and produce a more accurate estimate of customer turnover by including more pertinent input variables based on the subject area. On the other hand, this would provide customer relationship strategists a competitive advantage to keep their profitable clients and reduce unwelcome turnover. Also, future studies should consider other classification methods like decision trees, Support Vector Machine, neural networks, and logistic regression.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The author declares that she has no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in [Google Drive] at https://drive.google.com/file/d/1qme8WeYkmXWfWkLa87jG37owWPCsNY65/view?usp=drive_web

References

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295. https://doi.org/10.3390/electronics9081295

Alsayat, A. (2023). Customer decision-making analysis based on big social data using machine learning: A case study of hotels in Mecca. *Neural Computing and Applications*, *35*(6), 4701–4722. https://doi.org/10.1007/s00521-022-07992-x

Bahari, T. F., & Elayidom, M. S. (2015). An efficient CRM-data mining framework for the prediction of customer behaviour. *Procedia Computer Science*, *46*, 725–731. https://doi.org/10.1016/j.procs.2015.02.136

Bratina, D., & Faganel, A. (2023). Using supervised machine learning methods for RFM segmentation: A casino direct marketing communication case. *Market-Tržište*, *35*(1), 7–22. https://doi.org/10.22598/mt/2023.35.1.7

Chakraborty, A., Mitra, S., Bhattacharjee, M., De, D., & Pal, A. J. (2023). Determining human-coronavirus protein-protein interaction using machine intelligence. *Medicine in Novel Technology and Devices*, *18*, 100228. https://doi.org/10.1016/j.medntd.2023.100228

Chaudhary, P., Kalra, V., & Sharma, S. (2022). A hybrid machine learning approach for customer segmentation using RFM analysis. In *International Conference on Artificial Intelligence and Sustainable Engineering: Select Proceedings of AISE 2020, 1*, 87–100. https://doi.org/10.1007/978-981-16-8542-2_7

Cho, Y. S., & Moon, S. C. (2013). Weighted mining frequent pattern based customer's RFM score for personalized u-commerce recommendation system. *Journal of Convergence*, *4*(4), 36–40.

Gustriansyah, R., Suhandi, N., & Antony, F. (2020). Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 470–477. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477

He, X., & Li, C. (2016). The research and application of customer segmentation on e-commerce websites. In *2016 IEEE 6th*

International Conference on Digital Home, 203–208. https://doi.org/10.1109/ICDH.2016.050

Jiang, T., & Tuzhilin, A. (2009). Improving personalization solutions through optimal segmentation of customer bases. *IEEE Transactions on Knowledge and Data Engineering*, *21*(3), 305–320. https://doi.org/10.1109/TKDE.2008.163

Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management*, *70*. https://doi.org/10.1016/j.ijinfomgt.2023.102641

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, *3*, 57–63. https://doi.org/10.1016/j.procs.2010.12.011

Lewaa, I., Hafez, M. S., & Ismail, M. A. (2021). Data integration using statistical matching techniques: A review. *Statistical Journal of the IAOS*, *37*(4), 1391–1410. https://doi.org/10.3233/sji-210835

Lewaa, I., Hafez, M. S., & Ismail, M. A. (2023). Mixed statistical matching approaches using a latent class model: Simulation studies. *Journal of Statistics Applications & Probability*, *12*(1), 247–265. https://dx.doi.org/10.18576/jsap/120123

Li, X. Q., Song, L. K., & Bai, G. C. (2022). Deep learning regression-based stratified probabilistic combined cycle fatigue damage evaluation for turbine bladed disks. *International Journal of Fatigue*, *159*. https://doi.org/10.1016/j.ijfatigue.2022.106812

Lu, N., Lin, H., Lu, J., & Zhang, G. (2014). A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, *10*(2), 1659–1665. https://doi.org/10.1109/TII.2012.2224355

Maryani, I., Riana, D., Astuti, R. D., Ishaq, A., Sutrisno, & Pratama, E. A. (2018). Customer segmentation based on RFM model and clustering techniques with K-means algorithm. In *2018 IEEE Third International Conference on Informatics and Computing*, 1–6. https://doi.org/10.1109/IAC.2018.8780570

Mohammad, J., & Kashem, M. A. (2022). Air pollution comparison RFM model using machine learning approach. In *2022 IEEE 7th International Conference for Convergence in Technology,* 1–5. https://doi.org/10.1109/I2CT54291.2022.9824248

Rahim, M. A., Mushafiq, M., Khan, S., & Arain, Z. A. (2021). RFM-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services*, *61*. https://doi.org/10.1016/j.jretconser.2021.102566

Seymen, O. F., Dogan, O., & Hiziroglu, A. (2021). Customer churn prediction using deep learning. In *Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition*, 520–529. http://doi.org/10.1007/978-3-030-73689-7_50

Shah, S., & Singh, M. (2012). Comparison of a time efficient modified K-mean algorithm with K-mean and K-medoid algorithm. In *2012 IEEE International Conference on Communication Systems and Network Technologies*, 435–437. https://doi.org/10.1109/CSNT.2012.100

Sheshasaayee, A., & Logeshwari, L. (2017). An efficiency analysis on the TPA clustering methods for intelligent customer segmentation. In *2017 IEEE International Conference on Innovative Mechanisms for Industry Applications*, 784–788. https://doi.org/10.1109/ICIMIA.2017.7975573

Shihab, S. H., Afroge, S., & Mishu, S. Z. (2019). RFM based market segmentation approach using advanced k-means and agglomerative clustering: A comparative study. In *2019 IEEE International Conference on Electrical, Computer and Communication Engineering*, 1–4. https://doi.org/10.1109/ECACE.2019.8679376

Shim, B., Choi, K., & Suh, Y. (2012). CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns. *Expert Systems with Applications*, *39*(9), 7736–7742. https://doi.org/10.1016/j.eswa.2012.01.080

Shirole, R., Salokhe, L., & Jadhav, S. (2021). Customer segmentation using RFM model and k-means clustering. *International Journal of Scientific Research in Science and Technology*, *8*(3), 591–597. https://doi.org/10.32628/IJSRST2183118

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Smaili, M. Y., & Hachimi, H. (2023). New RFM-D classification model for improving customer analysis and response prediction. *Ain Shams Engineering Journal, 14*(12). https://doi.org/10.1016/j.asej.2023.102254

Weng, C. H. (2017). Revenue prediction by mining frequent itemsets with customer analysis. *Engineering Applications of Artificial Intelligence*, *63*, 85–97. https://doi.org/10.1016/j.engappai.2017.04.020

Wu, J., Shi, L., Yang, L., Niu, X., Li, Y., Cui, X., . . . , & Zhang, Y. (2021). User value identification based on improved RFM model and k-means++ algorithm for complex data analysis. *Wireless Communications and Mobile Computing*, *2021*, 1–8. https://doi.org/10.1155/2021/9982484

Wu, Z., Zang, C., Wu, C. H., Deng, Z., Shao, X., & Liu, W. (2022). Improving customer value index and consumption forecasts using a weighted RFM model and machine learning algorithms. *Journal of Global Information Management*, *30*(3), 1–23. https://doi.org/10.4018/JGIM.20220701.oa1

Yoseph, F., & Heikkila, M. (2018). Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method. In *2018 IEEE International Conference on Machine Learning and Data Engineering*, 108–116. https://doi.org/10.1109/iCMLDE.2018.00029

Zahrotun, L. (2017). Implementation of data mining technique for customer relationship management (CRM) on online shop tokodiapers.com with fuzzy c-means clustering. In *2017 IEEE 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering*, 299–303. https://doi.org/10.1109/ICITISEE.2017.8285515

How to Cite: Lewaaelhamd, I. (2024). Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis. *Journal of Data Science and Intelligent Systems* 2(1), 165–172, https://doi.org/10.47852/bonviewJDSIS32021293