RFM Analysis Applied to Customer Segmentation Using Machine Learning Models

Imhotep, Sayed
Sohag University, Sohag Governorate, Egypt

Abstract: Prediction, classification, and anomaly detection are made easier by a wide range of supervised and unsupervised methods that are included in machine learning (ML). Customer churn prediction is one of the most well-known applications for these approaches. Data scientists use a range of demographic, social, transactional, and behavioral information and traits to predict customer switching. Regretfully, a large number of UK businesses still lack the thorough and flexible customer data needed to conduct precise assessments. Because of this, they frequently rely significantly on data generated by enterprise resource planning systems, which are essentially transactional. As a result, companies are frequently only able to model and forecast using transactional data and are reluctant to make large investments in marketing research or other sources pertaining to customers. Companies are frequently restricted to using transactional data for modeling and forecasting, which are typically not based on sophisticated methodologies like ML and recency, frequency, and monetary (RFM). Therefore, the primary goal of the current effort is to offer a combination of RFM and ML analysis methods for churn prediction with a focus on transactional data. The dataset was extracted from a service that searches for internet retail datasets. Based on the information at hand, each customer's RFM scores are calculated. a churn metric that shows if a consumer has completed a transaction within a certain amount of time. This report presents a comparison of various strategies. Density-Based Spatial Clustering of Applications with Noise clustering and K-means were employed. By the end of this research, it can be concluded that the division of customers into six different clusters is a more easy and practical method.
Keywords: RFM analysis, statistical approaches, data analysis, machine learning models, artificial intelligence.

1. Introduction

Recency, frequency, and monetary (RFM) analysis is a relatively simple yet very powerful method for understanding and evaluating purchase-based customer behavior. In order to identify and target the most valuable customers for the purpose of carrying out targeted and focused marketing campaigns, the RFM technique uses quantitative categorization and classification of customers based on the RFM total of their most recent transactions (Shihab et al., 2019; Smaili & Hachimi, 2023). Based on these criteria, each customer is given a numerical score, making the study data-driven and objective. RFM analysis is anchored in the well-known marketing cliché that “80% of your business comes from 20% of your customers” (Alsayat, 2023; Bratina & Faganel, 2023; Chakraborty et al., 2023).

RFM is a strategic method used to analyze and estimate a customer's value based on the assessment of three important data points: frequency, monetary value, and recency. The frequency statistic asks how often the consumer makes purchases, but the recency metric shows the customer's most recent transaction. Finally, monetary value explores how much the consumer has spent. One helpful tool that might provide insightful information about clients and their behavior is RFM analysis. It should be highlighted, nonetheless, that this strategy ignores a number of other crucial elements that play a significant role in determining the client experience. To improve results, for example, detailed targeted marketing tactics may make use of a variety of factors, including the kind of product bought or consumer campaign reactions (Bahari & Elayidom, 2015). Furthermore, it is critical to recognize that RFM research does not account for client demographics: age, sex, and ethnicity, among others. For a more accurate and useful understanding of customers, marketers must thus incorporate a more thorough and nuanced strategy that takes into consideration a wide range of characteristics. The literature mentions a variety of data integration techniques that could be useful in this situation. Statistical data integration is discussed in Lewaa et al. (2021) and Lewaa et al. (2023).

An effective method for learning important things about consumer behavior is RFM analysis. It does have some restrictions, though. RFM analysis's disregard for important elements such client demographics and the type of goods purchased is one of its drawbacks. A more thorough strategy that considers a wider range of parameters is urgently needed in light of these constraints. A more precise understanding of clients can be attained in this way. Additionally, it should be mentioned that RFM's use is entirely dependent on the customers' history data, which suggests that it may not be able to predict their future actions with any degree of accuracy (Rahim et al., 2021;

Seymen et al., 2021). On the other hand, predictive methods can reveal possible customer behavior patterns that the RFM study might miss (Maryani et al., 2018; Mohammad & Kashem, 2022). This implies that although if RFM is useful for assessing customer data, its shortcomings in forecasting future customer behavior call for the use of more sophisticated predictive techniques.

RFM values are simple to compute and comprehend for the current case study, but they only account for a single facet of consumer behavior. Data scientists require flexible information about client demands, opinions, socioeconomic factors, relationship data, etc. in order to create high-quality prediction models. Since small and mid-sized businesses frequently lack a systematic strategy to data collection, it can be challenging to gather this type of information.

It would be more prudent and advantageous to classify and divide the customers according to their unique attributes, such as age or location, and then separate them into a customer group rather than performing a comprehensive and exhaustive analysis of the entire customer database. It is feasible to create a customized and pertinent offer that would be very alluring to clients who are extremely valuable to the company by putting into practice a methodical, well-organized marketing campaign that is especially suited to each subdivided group. Calculating the RFM ratings for real-world applications requires significant mathematical skills or specific analytical knowledge. Furthermore, RFM models can range in complexity from simple to complex, just like any other model. Sorting clients into the three categories of recency score, frequency score, and monetary score is the first step in the RFM segmentation process. This is often done on a scale of 1 to 4. A score of 1 indicates the top 25% in each category (i.e., the most recent, most frequent, and most purchased individuals), a score of 3 indicates the next 25%, and so on. Through the use of an RFM scoring system Similar to this, client RFM segments can be used to create an effective marketing strategy. We correct the data using the Box–Cox transformation, which was not employed in the prior work, to ensure that the data are normally distributed among the many methods for assessing customer classification that have been provided in the literature. Additionally, this study is the first in the UK to use this kind of client segmentation. This report presents a comparison of various strategies.

We hope to find out through this study which machine learning (ML) method is optimal for customer turnover, which is up for debate. To select the best candidate, data scientists must investigate and evaluate as many as possible. This study suggests and assesses a churn prediction technique using machine learning algorithms using RFM data. Numerous candidate algorithms have been evaluated using various input variables. Using machine learning algorithms to identify which consumers are most likely to end their commercial relationship with a company is known as customer churn prediction. This is commonly performed by training a supervised ML model on historical data such as consumer purchase history, account activity, and customer support interactions. The model is then used to produce predictions based on fresh information, such as whether a new client is likely to leave within a given period of time. The ML model can be used to categorize whether or not a customer will leave, making customer churn a classification problem.

2. Literature Review

According to Jiang and Tuzhilin (2009), buyer targeting and customer segmentation must be used in order to improve marketing performance. Despite the difficulty of unified optimization, these two interrelated tasks are combined into a methodical methodology. As a result, the authors have suggested using the K-Classifiers Segmentation algorithm to solve this problem. This specific strategy places a strong emphasis on allocating more resources to clients who generate more profits for the company. Numerous writers have added to the body of knowledge regarding various approaches to customer segmentation. In their research, Jiang and Tuzhilin (2009) suggest a novel approach to customer clustering that departs from the standard practice of depending only on calculated statistics. To get a more direct clustering result, their method instead uses the transactional data of several clients. The authors also disclose that the work of determining the best segmentation strategy is actually NP-hard, which means that different sub-optimal clustering strategies must be developed. Tuzhilin carefully developed these techniques. The authors then carefully examined the customer segments that were produced using the direct grouping method and discovered that, in contrast to the conventional statistical methodology, this strategy produced far better outcomes.

A unique clustering technique that is comparable to the K-means and K-medoids algorithms was presented by Shah and Singh (2012). It is acknowledged that these algorithms are partitional strategies. However, the recently suggested algorithm does not always provide the best answer. It lowers the cluster error criteria, however. As the number of clusters increases, Saurabh's findings show that the time needed to implement the new strategy reduces, which is a major improvement over conventional techniques.

A recommendation system that is tailored to the consumers' requirements is proposed by Cho and Moon (2013). In order to find the patterns that appear most frequently, the suggested approach uses the weighted frequent pattern mining technique. The authors employed the RFM model, which is frequently used for this purpose, to conduct consumer profiling in order to identify the possible customers. The association rules that can be derived from the mining process are generated by the suggested method using different weights for every transaction. The RFM model can be used to improve the accuracy of the recommendations given to clients, which will raise the company's earnings. An extensive analysis focused on customer attrition prediction was carried out by Lu et al. (2014). To create a completely original prediction model, the authors deftly used logistic regression and successfully separated the transactional data. The clever researchers found that by implementing tailored marketing methods in their experiments, it is possible to identify and retain consumers with the highest churn value. Zhang, on the other hand, believes that determining the underlying reason for customer attrition and meeting individual demands is a necessary condition for every business to survive. In order to improve customer satisfaction, He and Li (2016) have put up a complex and multidimensional three-dimensional methodology.

lifetime value, increasing client contentment, and favorably impacting consumer conduct. The writers have wisely noted that consumers are a diverse group with a wide range of needs, wants, and goals based on their thorough investigation and analysis. The delivery of superior services to these valued clients is made easier by the successful use of segmentation techniques to identify and comprehend their demands and expectations, which may be used to accommodate the diversity and complexity of customer preferences. In order to segment using the RFM and lifetime value methodologies, Sheshasaayee and Logeshwari (2017) created a new and creative integrated strategy. This strategy was implemented in two stages, the first of which involved a statistical technique and the second of which involved clustering performance. This method's main goal was to use a neural network to improve the segmentation process overall after performing K-means clustering following the two-phase model.

Through the use of customer relationship management (CRM), Zahrotun (2017) identified the most exceptional clients using customer data collected online. The author was able to efficiently identify new clients through segmentation by using the CRM paradigm in the context of online buying, which ultimately helps to maximize business earnings. The fuzzy C-means clustering method was used to help with the precise implementation of marketing plans and client segmentation. In the end, this approach gives clients the chance to obtain customized amenities in a variety of categories, all in line with their own requirements and tastes.

3. Methodology

The Online Retail Dataset is where the experiment data was taken from. "Recency" (recentness from the ending date), "frequency" (number of transactions until the ending date), and "monetary" (total amount of transactions until the ending date) are among the features and metrics that have been computed for each client (Chaudhary et al., 2022; Gustriansyah et al., 2020).
Its RFM analysis feature has been used to determine RFM scores. Both independent and nested binning with four bins are used to compute RFM scores. K-means clustering could be used in place of this traditional method to modify RFM analysis. The classic method and this methodology are contrasted (Li et al., 2022; Shirole et al., 2021; Wu et al., 2021). Python was used to apply machine learning techniques (Joung & Kim, 2023; Khajvand et al., 2011).

Data preprocessing, a crucial component of data preparation, is any effort to modify raw data in order to optimize it for further data processing activities (Shim et al., 2012; Weng, 2017). This procedure has long been acknowledged as a crucial first step in the data mining process. In recent years, there have been substantial changes made to data preparation methods to make it easier to train AI and ML models and to make conclusions against them.
Data preprocessing is a set of procedures designed to convert data into a structured format that can be easily handled in a variety of data science tasks, including machine learning and data mining, more quickly and effectively (Wu et al., 2022; Yoseph & Heikkila, 2018). To provide reliable results, it is a crucial step that is usually applied at the beginning of the ML and AI development pipeline.

Our study's focus is on the UK, as indicated by the data description in Table 1. Therefore, we choose every observation pertaining to this nation's online retail. Next, as a preprocessing step, null values were examined. A null value, which is frequently seen in relational databases, indicates that the value in a specific column is either absent or unknown. Note that a null value should not be confused with an empty string, which is a feature of datetime or character data types, or with a zero value, which is a feature of numeric data types. To guarantee accurate data manipulation and interpretation in a particular database, these distinctions are essential. Out of the 495,478 total sample size, we have 133,600 null values in our data for the UK.

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

We'll compute RFM values. The number of days between each customer's maximum date and the maximum invoice date is used to compute recency. The number of client IDs that are repeated is therefore taken into account when calculating the frequency factor. Additionally, the entire cost is computed in order to perform financial calculations. Each customer's total price is determined by multiplying the quantity by the unit price. The total price for each customer up until the end date is then added up to determine the monetary amount for each consumer.
Therefore, a variety of tests utilizing different input variables have been conducted to examine the models' suitability for studying customer behavior based on different sets of input factors:

(I) RFM values; (II) RFM scores; (III) RFM scores, RFM counts, and the number of objects for each customer.

The aforementioned diverse inputs are developed for the two approaches in order to compare them collectively and determine the most effective technique for consumer segmentation.
A triad of discrete scores for the variables of frequency, monetary value, and temporal proximity are assigned to each consumer. On a graded scale ranging from 1 to 4, the scoring process is carried out. Patrons in the top percentile receive a score of 4, while those in the lower percentile receive ratings of 3, 2, and 1, respectively. It is possible to presume that the scores have the peculiar characteristics listed in Table 2.

Table 2: Scores of RFM

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

Finally, scores between 444, 443, and 111 are provided to each and every consumer. Because they are expected to provide a larger amount of cash to the company, consumers with a score of 111 may be considered potential customers of the entity; on the other hand, clients with a score of 444 are considered potential customers. Every consumer can be categorized according to their unique segmentation based on this RFM score.

4. Results

According Data visualization is shown in this section. Additionally, the Box–Cox transformation is used to resolve the normalcy problem. Furthermore, the outcomes following the application of several machine learning techniques, K-means, and DBSCAN are displayed. A comparison of various approaches is also taken into account.

4.1. Visualization of data

Prior to starting the clustering process, the distribution of RFM values was examined.

Figures 1, 2, and 3 clearly show that we have outliers. The Box-Cox transformation is a statistical technique that alters our aim variable so that your data closely resembles a normal distribution.

Figure 4 shows that there aren't many outliers that could have an impact on the grouping results. Observations that substantially depart from the norm, either positively or negatively, are known as outliers. Erroneous inferences and conclusions may result from these aberrant data points' excessive influence on statistical results, especially on measures of central tendency like the mean. Therefore, we eliminate these anomalies and use our clustering techniques without them.

Figure 1: Box plot for recency values

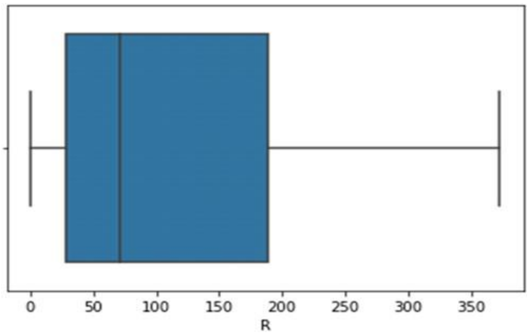


Figure 2: Box plot for frequency values

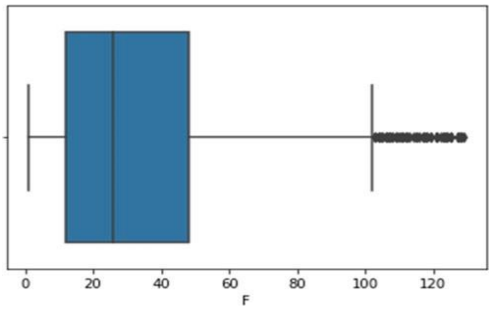
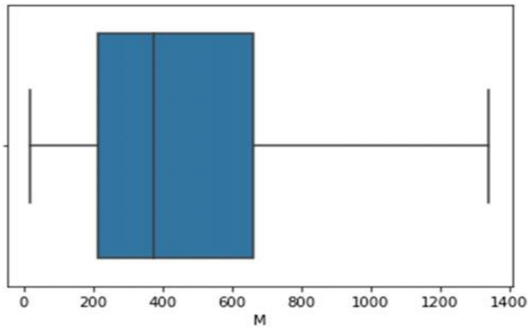


Figure 3: Box plot for monetary values



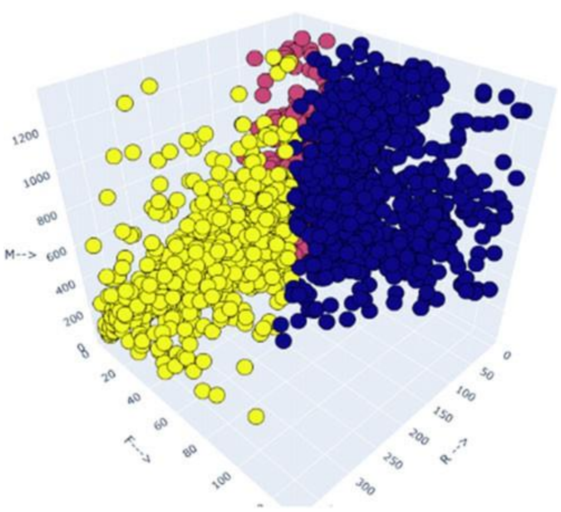


Figure 4: 3D for the three values after correcting the outlier problem

4.2. Use Q-CUT

Following the acquisition of RFM values, the Q-CUT method is used to impute the R, F, M, and RFM scores. Each recency score ranged from 1 to 4. Recency values take one in the first quarter and four in the fourth. As a result, all scores with a value of one indicate that the customers who fall into the highest category (i.e., 111) and have completed a transaction will continue to remember the product in their minds, which will increase the likelihood that they will purchase or use it in the future. Similarly, frequency and money received scores ranging from 1 to 4, with the first quarter receiving a score of 4 and the fourth quarter receiving a score of 1. Table 3 summarizes the frequency distribution for RFM score after taking the Q-CUT approach into account.

4.3. K-means clustering technique

Splitting n observations into k clusters and assigning each observation to the cluster with the closest mean value (also called the cluster centroid or cluster center) is the main goal of the K-means clustering, a popular vector quantization technique that first appeared in the signal processing field. Voronoi cells, which divide the data space, are the result of this procedure. It should be noted that the only metric that minimizes Euclidean distances is the geometric median. However, for a more complicated Weber problem, K-means clustering works better at minimizing within-cluster variances (squared Euclidean distances) than normal Euclidean distances. Thus, improved Euclidean solutions can be obtained by using k-medians and k-medoids (Ahmed et al., 2020).
The elbow approach has become a well-liked heuristic for figuring out the ideal number of clusters in a given dataset in the field of cluster analysis, which is a common technique for unsupervisedly classifying items into related subsets. Plotting the explained variation—a measure of the total variance that the clustering algorithm accounts for—as a function of the number of clusters taken into consideration is the fundamental task of this method. The elbow of the curve, which is the point on the curve where the slope noticeably changes, is then chosen as the ideal number of clusters to use. It should be noted that by applying the same fundamental idea of locating the elbow point on the pertinent curve, this technique can be applied to other data-driven models, such as principal component analysis, which seeks to capture the greatest amount of variance in a dataset using a smaller set of variables.

A predefined value of k is used to carefully choose a set of k random points as the initial centroids in the first stage of the clustering method. The Euclidian distance between each data point and the previously selected centroids is then painstakingly calculated as part of a comprehensive evaluation procedure that is applied to every single data point in the dataset. Third, after the distance evaluation process is finished, the calculated values are carefully compared to see which centroid has the shortest Euclidian distance value. Then, using an effective and efficient assignment mechanism, the data point in question is assigned to the appropriate centroid. The last objective is to obtain an ideal clustering solution by carefully repeating the previous processes. The entire procedure is closely watched, and it is only stopped when it is established that the clusters produced are the same as those from the prior iteration (Sinaga & Yang, 2020).

The customer dataset with "n" instances k—the number of clusters—is the input for this study. The customer data divided into k clusters is the output.

Three is the ideal number of clusters, according to Figure 5. Therefore, we will partition our consumer base into Figures 6, 7, and 8.

4.4. Using DBSCAN for clustering

Starting with a randomly chosen data point from the dataset, the current algorithmic methodology for DBSCAN clustering iteratively repeats this process until all points have been visited. All "minPoint" data points are considered to be part of the same cluster if there are a minimal number of them within a given "ϵ" radius of the selected point. At the end of the process, the clusters are iteratively expanded by repeatedly calculating the surrounding area for each neighboring point. The entire clustering procedure using DBSCAN is depicted in Figure 11.

It can be concluded from Figures 9, 10, and 12 that grouping clients into six different groups is a more sensible and direct strategy.

Figure 5: The elbow method for the second method

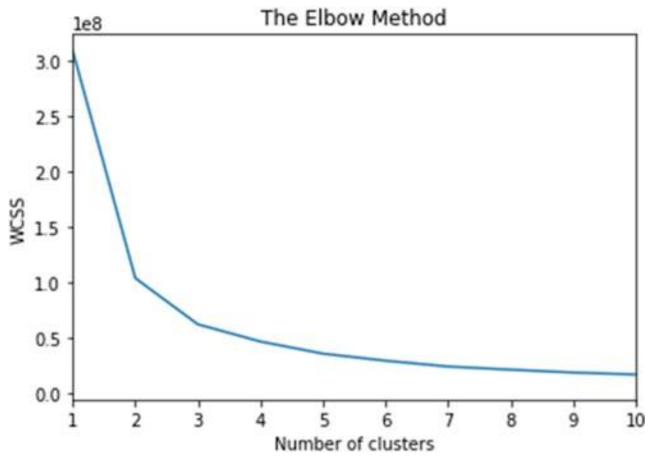


Figure 6: Boxplot within clusters of customers among recency values

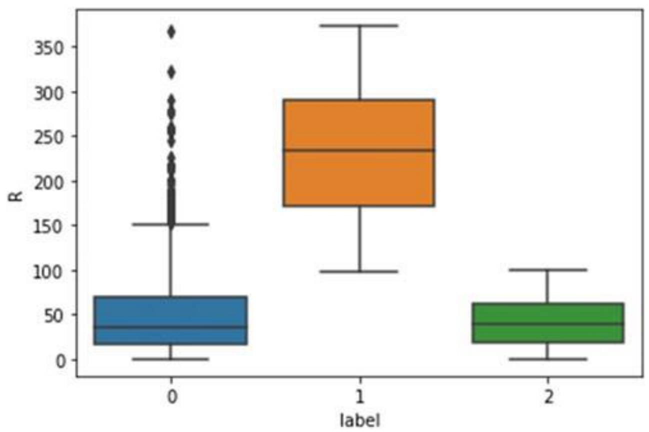


Figure 7: Boxplot within clusters of customers among frequency values

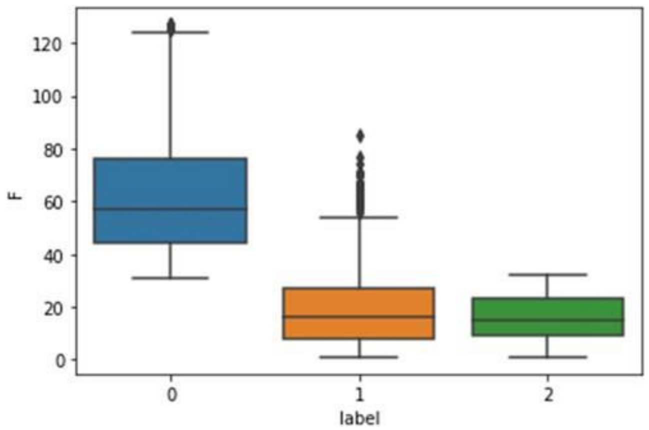


Figure 8: Boxplot within clusters of customers among monetary values

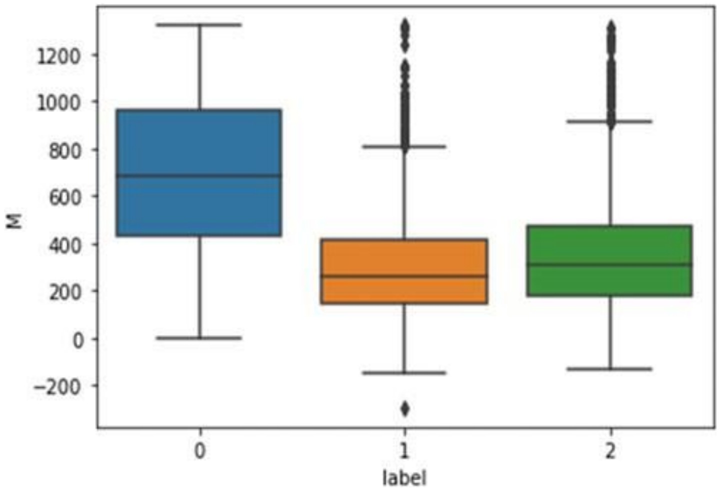


Figure 9: 2-D clustering using DBSCAN

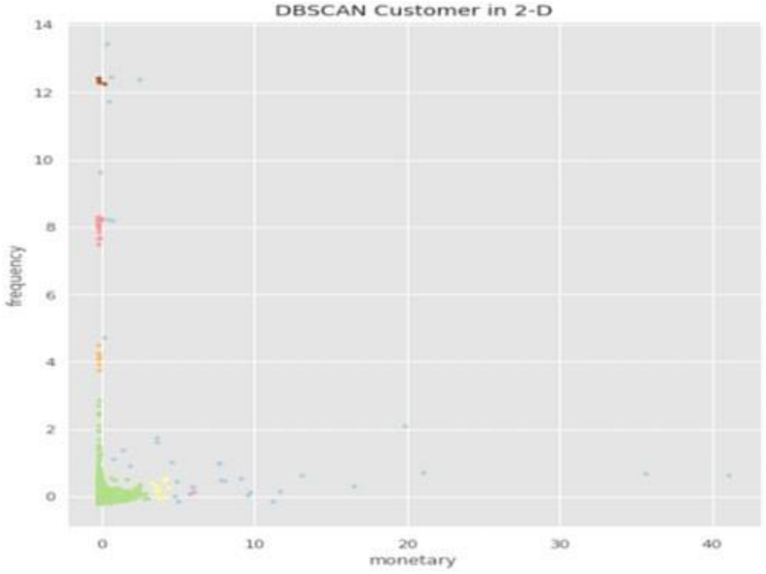


Figure 10: 3-D clustering using DBSCAN

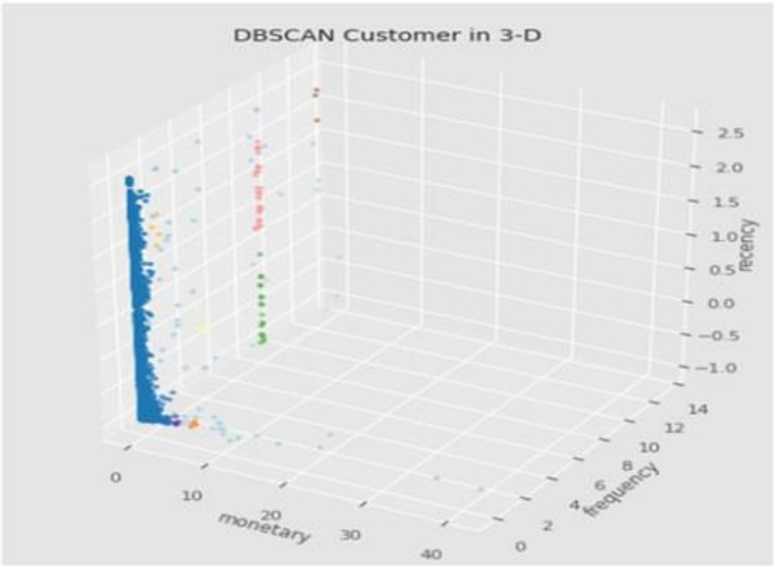


Figure 11: DBSCAN clustering

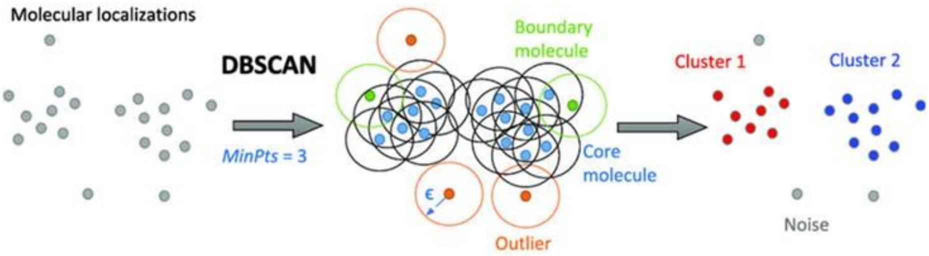


Figure 12: Clustering using DBSCAN among frequency

5. Conclusion

Using mostly transactional data, the current work's main goal is to offer a combination of ML and RFM analysis tools for churn prediction. The internet retail dataset served as the source of the dataset. Based on the information at hand, each customer's RFM scores are calculated. a churn metric that shows if a consumer has completed a transaction within a certain amount of time. Among the several methods discussed in the literature, this is the first study to take into account customer segmentation in order to use Box-Cox transformation to deal with the outlier. Additionally, this study is the first in the UK to use this kind of client segmentation. Additionally, we alter it to look more like a normal distribution, which enhances the study's findings.

This report presents a comparison of various strategies. K-means and DBSCAN clustering were employed. By the end of this research, it can be concluded that the division of customers into six different clusters is a more easy and practical method. The clear division of the groups as shown in the several clustering method plots supports this even more. Therefore, it is now the responsibility of marketing managers and customer insight teams to determine the best communication or promotional strategy to use in order to move people from one segment to another or possibly attract more customers to a new segment, which should be placed in the upper right corner of the plots.

By adding more relevant input factors based on the topic area, data scientists can enhance the prediction process and generate a more accurate estimate of customer turnover in subsequent work. However, this would provide customer relationship strategists a competitive edge in retaining their lucrative clientele and lowering unwanted turnover. Other classification techniques including logistic regression, decision trees, support vector machines, and neural networks should also be taken into account in future research.

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