New RFM-D classification model for improving customer analysis and response prediction

Article *in* Ain Shams Engineering Journal · April 2023

DOI: 10.1016/j.asej.2023.102254

CITATIONS
3

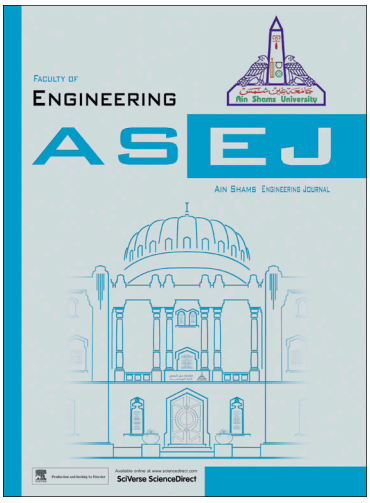
READS
96

2 authors:

Moulay Youssef Smaili
Université Ibn Tofail
6 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Hanaa Hachimi
University Sultan Moulay Slimane
146 PUBLICATIONS **425** CITATIONS
SEE PROFILE

Journal Pre-proofs



New RFM-D Classification Model for Improving Customer Analysis and Response Prediction

Moulay Youssef SMAILI, Hanaa HACHIMI

PII: S2090-4479(23)00143-0
DOI: https://doi.org/10.1016/j.asej.2023.102254
Reference: ASEJ 102254

To appear in: *Ain Shams Engineering Journal*

Received Date: 26 April 2022
Revised Date: 22 February 2023
Accepted Date: 30 March 2023

Please cite this article as: Youssef SMAILI, M., HACHIMI, H., New RFM-D Classification Model for Improving Customer Analysis and Response Prediction, *Ain Shams Engineering Journal* (2023), doi: https://doi.org/10.1016/j.asej.2023.102254

New RFM-D Classification Model for Improving Customer Analysis and Response Prediction

Moulay Youssef SMAILI

Systems Engineering Laboratory LGS, Sultan Moulay Slimane University, Av Med 5, Po Box 591, Beni Mellal, Morocco

Ismaili.youssef@usms.ac.ma

Hanaa HACHIMI

Systems Engineering Laboratory LGS, BOSS Team, Sultan Moulay Slimane University, Av Med 5, Po Box 591, Beni Mellal, Morocco

Hanaa.hachimi@usms.ma

Abstract

Customer segmentation is seen as one of the pillars of a successful advertising campaign. Marketers give great importance to this flagship phase in the process of marketing new products. Successful segmentation will involve successful "Customer Targeting" and therefore a profitable customer marketing campaign. Many works have dealt with customer segmentation using unsupervised Machine Learning algorithms such as K-Means by applying the famous Recency, Frequency and Monetary model. That model suffers from insufficiency by ignoring other important parameters according to the field of application. In this paper, we have modified the model by adding diversity "D" as a fourth parameter, referring to the diversification of products purchased by a given customer. The segmentation based on RFM-D is applied in a retail market in order to detect behavior patterns for a customer. The proposed model increases the quality of prediction of customer behavior; Companies could predict, customers who will respond positively.

Keywords: RFM & RFM-D models, Diversity, K-Means algorithm, customer segmentation, Customer response prediction and targeting optimization.

1. Introduction

The main objective and common to all companies, in a Retail Marketing environment, is to make the Return on investment optimum; either through an acquisition, by influencing and attracting new customers, or by retention by prospecting for new revenues by offering new offers to existing customers [1]. According to Pareto [2, 3], among all customers, only one fifth contributes more to the revenue of the company compared to the rest of the customers. Thus, in this article, we will focus on customer retention by performing efficient segmentation. That segmentation will group customers and offer them new products with the goal of maximizing return on investment. The goal is to achieve a prediction that identifies the customers who will almost certainly accept the new products offered by the company. Customers can have various and different preferences, for that, customer segmentation is considered one of the best ways to manage a company's customer. It is the process that produces similar groups, in terms of characteristics and attributes, from heterogeneous groups of customers [4, 5]. the marketing planning phase becomes clearer thanks to customer segmentation. This segmentation allows companies to take good care of their relationships that bind them to customers [5, 6]. Among many customers relationship management (CRM) analysis models, the RFM model is popularly adopted for segmenting customer. To measure customer value and customer profitability, chi-square automatic interaction detection (CHAID) and RFM model are considered as an important tool (method). CHAID has several strengths. the most important are: it's is a non-parametric statistic, both interval and nominal variables can be considered as predictors and in addition to discrete variables, Continuous ones can be defined as criterion [7]. It is found that CHAID tends to be superior to RFM when the response rate to a mailing is low and the mailing would be to a relatively small portion of the database, however, RFM is efficient procedure in other circumstances [8] and deserves to be improved to better compete with the different segmentation methods. In a dynamic layer, RFM describes in detail the customer's outline. This allows to produce a basis for personalized service and communication. Previous studies suggest that the results of predictions, based on the classic RFM model, could be improved by adding additional variables to predict customer behaviors [9]. In light to that, this work extends RFM model to Recency, Frequency, Monetary and Diversity Model (RFM-D Model) by adding a fourth variable diversity (D) which calculate the diversity in a customer's purchasing behavior; this will allow to identify customers characterized by diversified baskets in terms of products. This same type of customer will have a high probability of buying new products offered by a company as part of its product diversification. Diversification of the offer will allow the company to reach a maximum level of customer retention. Segmentation based on this parameter will identify a customer segment to target by offering new products and ensure better profitability in a marketing campaign. The use of machine learning techniques for machine learning are increasingly used in various fields. With the appearance of big-data and large flows of marketing data from companies, these techniques, and in particular, clustering techniques, are applied to achieve customer segmentation according to their similarities [10, 11]. These techniques of segmentation (or clustering) aim to identify groups with minimal intra-group variation, and maximum inter-group variation using a distance function (dissimilarity) [10, 12]. Although statistical clustering methods are the traditional methods for creating segmentation models, the large flow of data decreases the robustness and fit of these models [13-15] . This work focuses mainly on adding a fourth parameter to the well-known RFM model, namely diversity. The objective is to zoom in on this variable when analyzing the data-set and make a prediction of the customers who will respond positively to new offers. To achieve this objective, customer segmentation will be addressed by the KMEANS algorithm based on the extended RFM-D model. Once the segmentation is performed, both silhouette and Elbow methods will be used to accurately determine the optimal number of clusters to adopt in this analysis. Then, the effectiveness of the segmentation will be measured by calculating the accuracy of customers who have or not accepted a new product during the period following the segmentation. Diversity is a behavioral parameter related to the buying habits of customers who like to diversify their baskets without necessarily linking the items purchased. Contrary to the basket analysis, which mainly seeks to find the relationships between the

products that go together and that are involved in the transactions of each customer [16]. It is an analysis of all the groupings of items in a shopping cart in order to identify affinities that can be exploited to better predict what products to offer to customers [17].

2. Related works

Yosef and Samuels [18] performed customer segmentation based on customer Internet Protocol (IP) transit data in order to perform customer profiling, they mainly used two factors; customer lifetime value factor and loyalty factor. Mussadiq Abdul Rahim (Rahim et al., 2021) prefers to focus on RFM functionality as well as quantitative and statistical analysis of transactions. The objective is to extract the re-purchase behavior of customers, plus the data needed for this study is available and easily accessible to retail store owners. They used the superseded learning methods to process retail data and validate their scheme. In order to solve this problem with high accuracy, a new kernel is used for the SVM method. They claim that this work is a beginning of research on merging retail and machine learning applications. Radio-frequency identification (RFID) technology has been adopted by Han et al [19] in order to achieve customer identification as well as to analyze and predict behavior of a client regarding the product offerings of a company. Rodrigo propose an RFM model per product; a model that estimates the value of customers on an individual basis (by product) and then it aggregates these values to calculate the overall customer value. They confirmed that the RFM prediction by product is better than the classic RFM prediction. That is ensured by grouping together two important marketing perspectives that are usually dealt with separately in the prediction: the customer and product outlook. The results obtained by this study showed that the data related to the products are very useful in estimating CLV. Seyed Mohammad [20] have proposed a new approach to segment customers; They enriched the RFM model with an additional parameter and subsequently combined their model with the K-means algorithm according to the Davies-Bouldin index. This allowed them to classify customers according to product loyalty in a B2B universe. The additional parameter represents the period of activity of the product with determining the weight (W) of each LRFM measurement based on the Eigenvector Technique. The work has shown that the combination of the extended WRFM model in the K-means algorithm clearly improves the precision of the classification. Reinartz et al. assert [15] that companies cannot efficiently segment and clearly differentiate between short-lived and long-lived customers if they are based on the classic RFM model. They therefore extended the model by a fourth variable, which is the length (L). They applied their approach to the clients of children's dental clinic in Taiwan to properly target clients in their marketing operations for dental services. (L) calculates the safety between the first and the last purchase made by a customer. The approach of Qeethara Kadhim et al. [21] is based on the hybridization of the model-trust region concept of confidence with the concept of conjugate gradient. The experimentation of the prediction of customer behavior (buyer / non-buyer) by applying the methodology of artificial neural networks and adopting the Recency, Frequency, Monetary and Time Model (RFMT), gave good results in terms of the 88% classification of cases in the training set. Zuccaro et al. designed a transaction-based model to segment the internet users of a large bank in Canada [22]. This hybridized model identified four clusters using Two-step cluster analysis. Cross-tabulated afterwards with the Mosaic financial segments, gives better results than classical way of segmentation; Highly stable and interpretable segments are created. Thanks to these hybrid segments, banks will have detailed measures to develop models for assigning profitability/loyalty scores to each customer who connects to the bank's internet-banking site. A previous work compared three algorithms to calculate the probabilities for the next month purchase of customers registered in a dataset of more than 10 0 0 0 customers and a total number of 20 0 0 0 0 purchases. Compared to the two first algorithms, logistic Lasso and the extreme learning machine, the third one, which is the gradient tree boosting method, gave best performances[23]. Kaveh Khalili-Damghani et al proposed a hybrid method to deal with the problem of customer segmentation. the model is based on clustering, rule extraction and the decision tree. the proposed approach was applied on a database of insurance and telecommunications customers to determine, on the one hand, the most influential features available and can be offered, and on the other hand, to determine the potentially profitable prospects [24]. Kaveh Khalili-Damghani et al concludes that it is essential to make an extension to this segmentation method to deal with problems whose data are uncertain. Wafa Qadadeh et al compared the performance of two analysis and segmentation methods namely K-MEANS and Self-Organized Maps (SOM) in the analysis of customer interest. The work concluded that SOM is better in terms of data processing speed and clustering quality [25]. Makoto Mizuno et al evoke the concepts of recall and precision to be able to evaluate segmentation methods where precision presented the only indicator adopted to measure the performance of a method. Using these two metrics, a financial metric is modeled to assess lost opportunity and non-optimal use of the company's budget for the customer. The problem of segmentation is therefore treated as a classic problem of optimization of the threshold of the financial measure [26].

3. Methodology

a. RFM and RFM-D Model

RFM model is a very popular model in the analysis of customer values and their segmentation. It is a model That is mainly based, in its analysis, on the behavior of customers in terms of their transaction and purchase, then make a prediction on the database [10]. The Three measures that make up this model are: recency, frequency and monetary summarized in the word RFM; where the parameter recency (R) measures the time elapsed since the last purchase. The frequency parameter (F) reflects the number of purchases made during a period. The parameter Monetary (M) is the total budget spent by a customer in a period [10]. Diversity, in the proposed RFM-D model, measures the number of different products that were purchased by each customer during the period studied. This variable is used to focus on customers open to new product offers and who agree to test new items without reluctance. A customer who buys more than one type of item, regardless of the total budget, should be distinguished from customers who only buy products that they are used to ordering. These loyal customers are also more likely to try new products and consume. A global view of the proposed customer segmentation system is elucidated by Figure 1. Below the description of the different steps: First step: Data pre-processing. The objective of this phase is to eliminate incorrect data that could bias the analysis and the final segmentation. Data with null stock codes, transactions with negative values, lines with missing customer numbers are affected by this phase. During this same phase, the values of the RFM-D components are calculated to build the table of RFM-D measures (Recency, Frequency, Monetary and Diversity). This step also allows an initial exploratory analysis of the data. It allows us to identify the distribution of orders and customers who have recently purchased items. This is an important phase as it determines the correlation between the different parameters Recency, Frequency, Money and Diversity. Second step:

Analysis of the data according to the RFM and RFM-D model. At the end of the data pre-processing, the recency is calculated by comparing the date of the last purchase of a customer with the reference date of the studied period. The reference date is nothing but the date of the last purchase of the period studied.

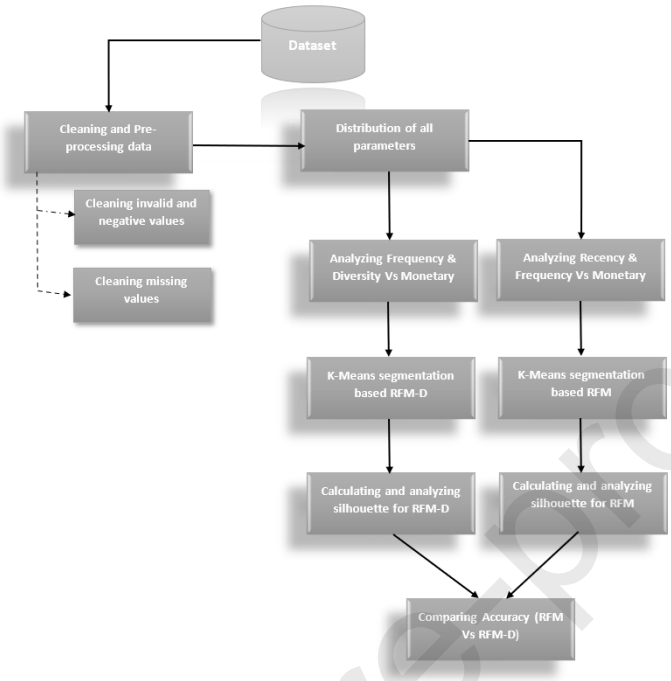


Figure 1. Customer segmentation steps (RFM Vs RFM-D)

Diversification parameter is calculated by determining the number of different products purchased without counting the duplicates. The calculation is made, of course, for the period studied. That parameter will allow marketers to predict the percentage that a new product offered in a future marketing campaign will be accepted by this customer. At the end of this step, each customer will have his own score based on the four RFM-D measures:

- Recency: the date of the customer's last purchase and the reference date of the last purchase of the period studied. A good customer, who visit repeatedly the company, is characterized by a small value of Recency.
- Frequency: During the period studied, the frequency represents the number of purchases made by the customer. He is considered loyal as long as the value of frequency is great.
- Monetary: This measure represents the total amount that a customer spent during the study period. The higher the parameter, the more money the company makes and has an interest in retaining the customer.
- Diversity: This measure represents the number of products that interested a given customer during the period studied. The greater the diversity parameter, the more the customer is open to new products and presents a potential customer to new offers of the company.

For each client, there is a score for the four parameters Recency, Frequency, Monetary and Diversity. The quantiles take scores from five, for the largest, to one for the lowest. It is assumed that these scores represent unique characteristics as described in Table 1.

Score Value	Description
1	Lost
2	Customer at Risk
3	Cannot be lost
4	Promising Customer
5	Potential Customer

Table 1. RFM values description

The Aggregate RFM-D (Recency, Frequency, Monetary and Diversity) node allows using historical customer transaction data, removing unnecessary data, and combining all of their remaining transaction data into a single row, using their unique customer ID as key, which indicates their last purchases (Recency), the number of transactions they made (Frequency), the total value of transactions (Monetary) and the distinct count of products purchased. For each customer, there is a score for the four parameters. The quantiles take scores ranging from 5, for the highest, to 1 for the lowest quantile. By grouping the RFM-D metrics into quintiles from 1 to 5, individual scores are created and allow the recency, frequency, monetary and diversity values of each customer to be compared to those of others.

Customers who have made purchases recently get a 5, while those who have not been seen for a long time get a 1. Customers who have placed many orders get a 5 for frequency, while those who have not 1. Customers who spent the most get a 5 for Monetary, while those who spent the least get a 1. Finally, customers who bought a maximum of different products get a 5, while those who have only purchased one type of product (even multiple times) get a 1. Thus, the table of scores is calculated, based on these aggregates, and on which the program will base itself to calculate the quantiles.
The third step segments the data, having the scores of each parameter (R, F, M and D), by applying the K-MEANS method and then specify the clusters based on the two models RFM and RFM-D. A comparison is then made as a final step to measure and see how well the addition of a parameter, in this case Diversity D, to the RFM model.

b. Clustering by K-MEANS:

in the field of segmentation, K-means is iterative. It is a standard algorithm in that it takes, as input, the data and the number of clusters "k". The output is data divided into "k" clusters so that the resulting intra-cluster similarity is high (minimizing the sum of squares within clusters in equation (1), while the inter-cluster similarity is low. K-Means relies mainly on the Euclidean distance formula to identify the similarity of the data in an iterative way.

$$d = \sum_{k=1}^{K} \sum_{i=1}^{n} (x_i - \mu_k)^2 \quad (1)$$

where k represents K cluster centers, μk is the kth center, and xi represents the ith point in the data set. Taking into account the skewed aspect of the data, the values of the RFM-D model are normalized and the clustering, by K-Means, is performed on these scaled data. The number of clusters will be limited to 10. For each segment, the amount earned will be calculated to determine the best segment in terms of revenue. As elucidated in Algorithm 1, the first step of the K-Means algorithm is initialization; This involves initializing the centroids, randomly, in order to run the iterations of the algorithm. The step of segmentation comes right after assigning every data to the nearest center group. For each point i = [13] :

$$z_{ik}^{(t)} = \begin{cases} 1 \ if \ k = \arg\min \left\| x_i - \mu_z \right\|, \ \forall z \in \{1, \ldots K\} \\ 0 \ \text{otherwise} \end{cases} \quad (2)$$

Then, the algorithm recalculates the centroid of each cluster, resulting from the previous step, by calculating the arithmetic mean of the data belonging to the same cluster:

$$\mu_k^{(t+1)} = \sum_{i=1}^{n} z_{ik}^{(t+1)} x_i \Big/ \sum_{i=1}^{n} z_{ik}^{(t)} \quad (3)$$

The current iteration in the algorithm is represented by the letter "t". The algorithm ends if the maximum number of iterations, defined before has been reached. It ends, also, if the result of the clustering, in an iteration t, is the same as the one in the previous iteration t-1

Algorithm 1 (K-Means Algorithm).

Input:
DS: Datasets having n instances of segments (clusters)
K: number of clusters
Algorithm:
1. randomly choose k points in DS to form the cluster centers
2. **repeat:**
3. According to the mean value, reassign each object to the cluster
4. Evaluation of each data point chosen as centroids using Euclidian distance
5. **until:**
6. the maximum number defined for the clusters is reached,
7. the process is aborted also if the resulting segments are the same as the previous step

Output:
K clusters segmenting customer database

The main problem for the K-Means algorithm is to determine the optimal value of the number of clusters "k" to segment the data set. A large K number can lead to a too fragmented partitioning of the data. This will prevent the discovery of interesting patterns in the data. On the other hand, a number of clusters that is too small will lead to having, potentially, too generalized clusters containing a lot of data. In this case, there will be no "fine" patterns to discover. Previous works, such as[27] have proven that the appropriate choice of an initial value of the number of clusters will improve the accuracy of K-MEANS. The Elbow method and the silhouette index are two examples of methods used to calculate the optimal number of clusters; both methods are used in this paper.

c. Silhouette and Elbow Method Algorithm

The Elbow method is considered the most efficient method for accurately calculating the optimal number of clusters during segmentation [28, 29]. The Elbow rule consists in generating a series of possible values for K while using a square of the distance connecting

the sample points of each cluster and its centroid. To measure the performance, the sum of the squared errors (SSE) is adopted as an indicator. Clusters are convergent as long as the SSE values are small. Once the number of clusters approaches the optimal number, a rapid decline is noticed for the SSE. Once the optimal number of clusters is exceeded, the SSE decreases, but in a very slow way [30]. The silhouette method was adopted to confirm the optimal cluster number determined by the Elbow method. Silhouette analysis can be used to determine the separation between the obtained clusters. The silhouette plot provides a way to visually assess parameters such as the number of clusters by considering how close each point in a cluster is to points in neighboring clusters.

4. Experimentation and results discussion

a. Results

The segmentation performed, using K-Means algorithm, is based on the classical RFM model and also on the proposed RFM-D model. The data set used comes from the purchase history of a customer between 12/01/2010 and 12/09/2011. A data set made available to the researchers on the UCI repository [31]. The objective is to compare the segmentation according to the two models and make a comparison that proves that taking into account the parameter "Diversity" in the customer segmentation is very important in predicting the promoter customers to be targeted in the next Marketing campaigns of the company. The data set consists of eight attributes described in detail in Table 2. It contains more than one million transactions that will undergo pre-processing to eliminate rows containing erroneous information. The dataset concerns 4383 customers making purchases on 4481 different products.

The analysis based on RFM and RFM-D showed that there is a significant correlation not only for the couple (Frequency, Monetary), but also for the couples (Diversity, Monetary) and (Diversity, Frequency). The confusion matrix, represented by “Figure 2” and “Figure 3”, shows well the significant correlations when adding the “Diversity” parameter to the model. The parameters R, F, M and D of the model have been normalized by the min-max method in order to avoid the asymmetry of the values of these variables in the data set. K-Means is, therefore, applied on the scaled data.

No	Attribute	Attribute Description	Type Of data
1	Invoice Number	Unique transaction identifier : six characters	String
2	Stock Code	Unique product identifier: five characters	String
3	Description	Name and description of each product	String
4	Quantity	Quantity purchased of a given product per transaction	String
5	Invoice Date	Date and time of the invoice	Date
6	Price (Unit)	Unit price of each product	String
7	Customer Identifier	unique customer identifier: five characters	String
8	Country	Name of the country from which the purchase is made	String

Table 2. Dataset Attributes description

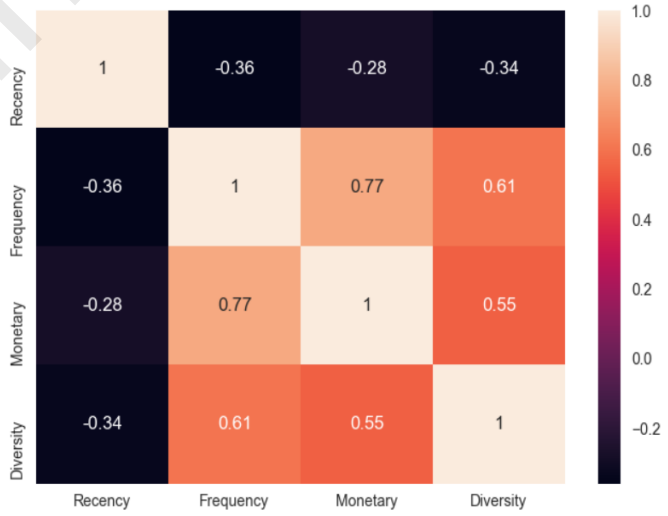


Figure 2. Confusion Matrix

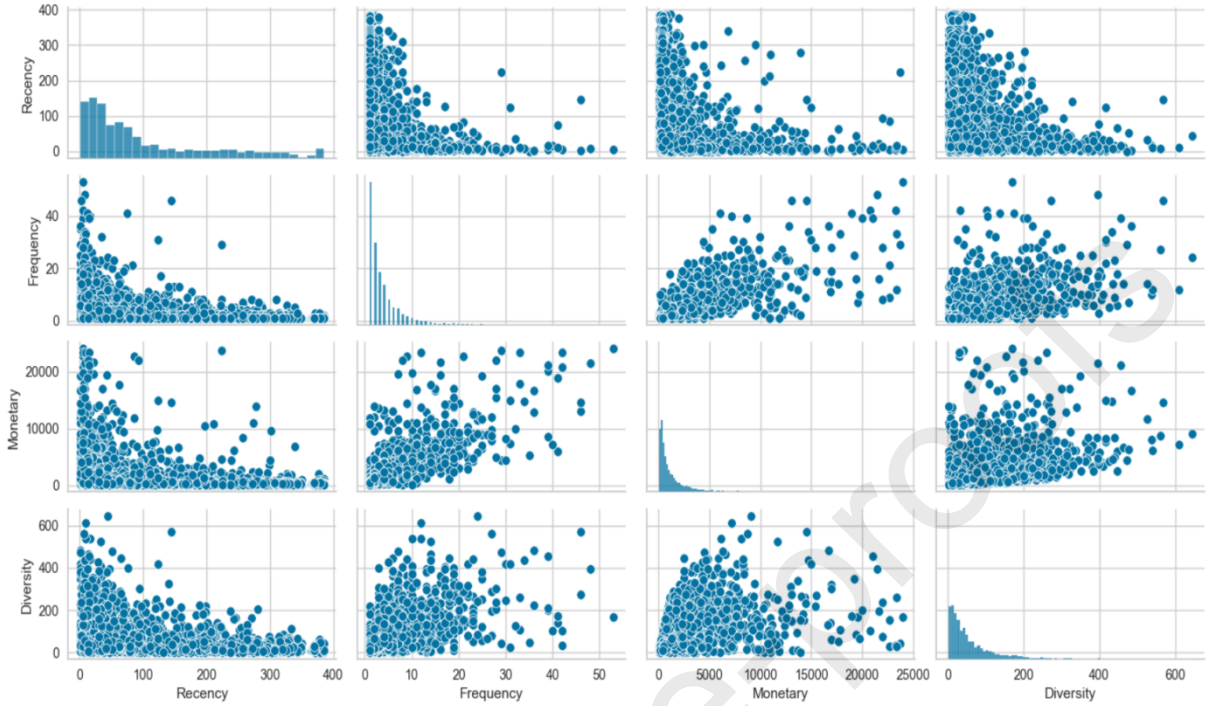


Figure 3. Confusion Matrix Distribution

This correlation is very clear in “Figure 4” and “Figure 5”, and shows that the (Diversity, Frequency) Vs correlation raised in the proposed model RFM-D, is more important compared to the one raised in the classical RFM model, i.e., Recency, Frequency Vs Monetary.

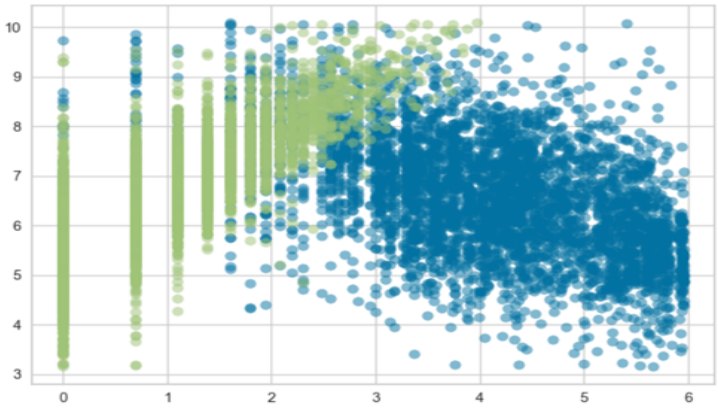


Figure 4. Data distribution: Recency & Frequency Vs Monetary (RFM case)

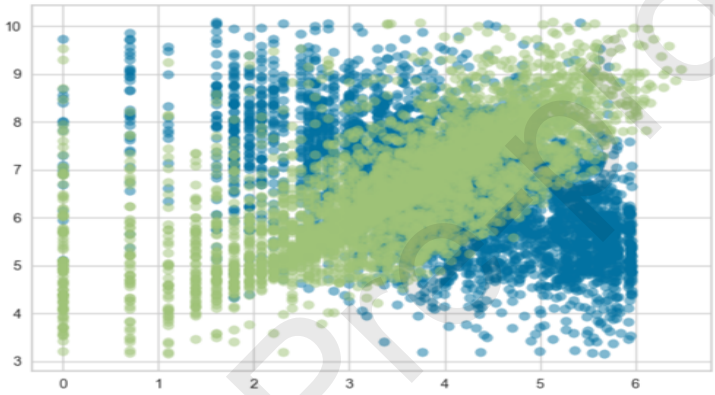


Figure 5. Data distribution: Frequency & Diversity Vs Monetary (RFM-D case)

The K-Means algorithm is applied using both the RFM and RFM-D models to perform the segmentation and compare between the segments containing the best customers for both models, in terms of revenue. To determine the optimal number of clusters, the Elbow method is applied on the segmentation performed according to the two models; as shown in "Figure 6" and "Figure 7", the segmentation using the RFM model gives four clusters as results, while the segmentation using the RFM-D model gave a result of six different clusters ("Figure 8" and "Figure 9").

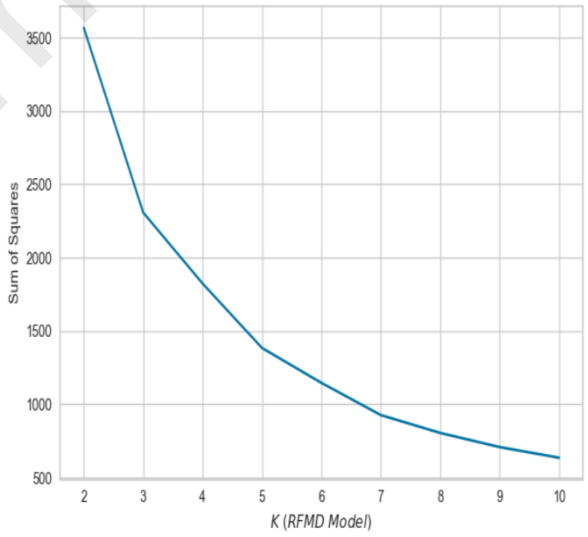


Figure 6. Elbow Graph for RFM-D model

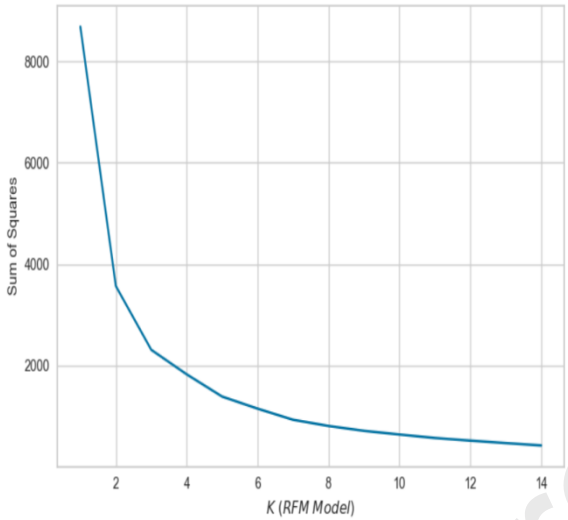


Figure 7. Elbow Graph for RFM model

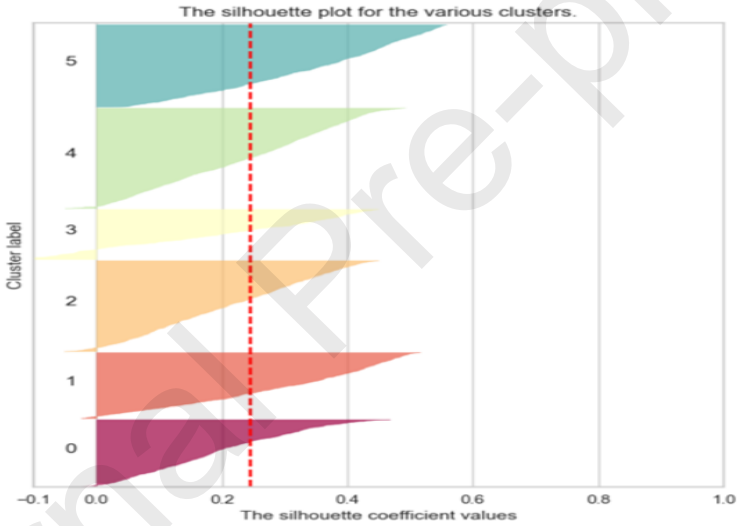


Figure 8.a RFM-D Silhouette coefficient values, Cluster = 6

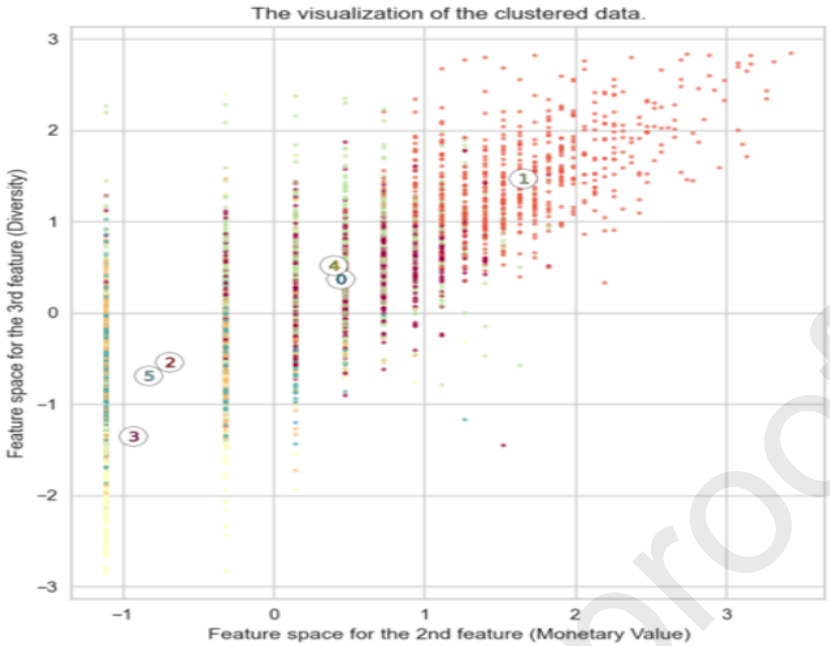


Figure 8.b RFM visualization of the clustered data, Cluster = 6

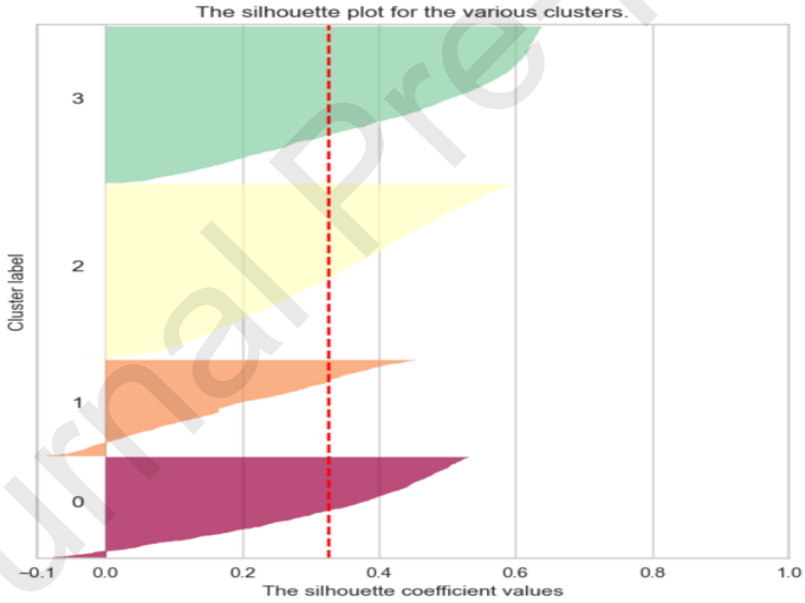


Figure 9.a RFM Silhouette coefficient values, Cluster = 4

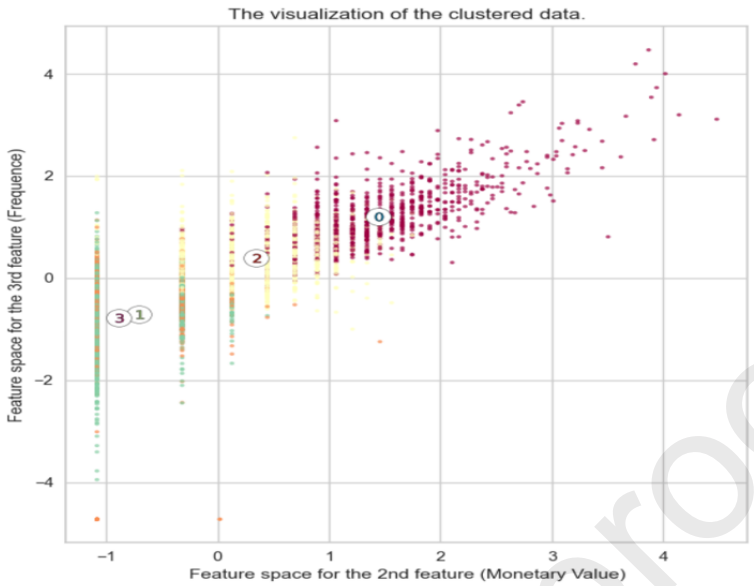


Figure 9.b RFM visualization of the clustered data, Cluster = 4

Table 3 displays the result of the average expenditure per cluster for RFM and for RFM-D ordered from the largest average expenditure to the smallest average expenditure. This average is calculated using equation (4):

$$\sum_{i=1}^{N} \frac{b_i}{k} \tag{4}$$

Where N represents number of customers in the cluster, and b_i is the total purchases made by customer i in the cluster. The second comparison will focus on the accuracy of the clusters. Accuracy will be calculated on the basis of the customers belonging to the best identified segments, for each model, and who bought again during the second part of the period studied between 01/04/2012 and 12/09/2012 using the equation (5).

$$Accuracy = \frac{\sum_{i=1}^{k} c_i}{K} \tag{5}$$

where c_i represents the customers belonging to cluster K and having made at least one purchase during the second period. "K" represents the total number of customers belonging to cluster K.

The cluster 1, champions of the RFM-D model, has a probability of responding positively to product offers in the second period, not only better than the second cluster of the same model, but also better than the two best clusters of the classical RFM model as shown in the table 2.

RFM Model		RFM-D Model	
Cluster 1	Cluster 2	Cluster1	Cluster2
78,63	73,67	**96,86**	72,31

Table 2. Accuracy RFM Vs RFM-D

	RFM Model		RFM-D Model	
	Customer count	Average amount spent	Customer count	Average amount spent
Cluster 1	787	492,16734	574	**533,8251218**

Cluster 2	1303	476,346454	950	512,3856443
Cluster 3	1014	336,3788708	655	360,3745574
Cluster 4	1210	267,6642871	804	317,8655453
Cluster 5	-	-	878	333,8960917
Cluster 6	-	-	453	209,7684424

Table 3. Average amount spent per cluster RFM Vs RFM-D

b. Managerial earning

Companies operating in retail markets have, in their data warehouses, several terabytes of transactions data concerning their customers or customers of their partners. Thanks to this data, marketing departments can predict the probability of purchases of their products, determine the loyalty index of each customer, the Customer Lifetime Value and of course the probability of customer churn[22]. However, a customer segmentation based on this data, without analyzing the different correlations between different parameters, allows for non-optimal segments when measuring customer profitability and thus allows for poor Return-On-Investment when targeting customers during advertising campaigns. Previous works presented the improvement of RFM segmentation by adding more parameters depending on the context. This work supports this hypothesis by adding, the first time, the diversity parameter during customer segmentation in a retail market. Marketers can use RFM-D, which is an extension of the classic RFM model, to:

- Better understand customers: A thorough understanding of customer trading behaviors can be the key to effective calculation of the probabilities of acceptance of product offers by customers. Diversity will be able to identify this type of customer addicted to purchases and who likes to test new products, especially when they are not too expensive.
- Improve customer targeting: The RFM-D model is a purely behavioral model and the high correlation of diversity with other model variables will enable effective analysis of customer behavior and positioning in the most relevant segment. This will ensure better prediction for customer who will respond positively to new product offerings, which will lead to optimization of customer targeting and therefore maximization of Return-On-Investment of the company.

5. Conclusion

In previous works, it was confirmed that segmentation via the RFM model is less efficient compared to other methods such as CHAID segmentation [8], segmentation based CLV [32] or segmentation based on machine Learning methods [8] . In this paper, it's first shown that the addition of the diversity parameter significantly improves the RFM model by analyzing the strong correlation between diversity and the old model parameters. Then, the segmentation with the RFM-D model generated different clusters than the RFM, and whose quality is better especially in total amount spent by clusters (figure.10) and by the prediction of customers who will respond positively to new offers. With this segmentation that improves the prediction of customer behavior, marketers will perform more effective and profitable customer targeting

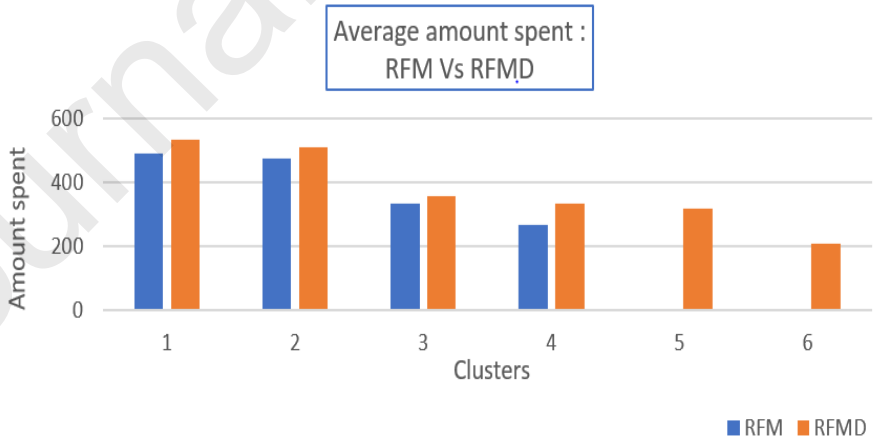


Figure 10. Average Amount spent by cluster

The RFM and RFMD models are then compared to classical and efficient Machine Learning methods and show the clear improvement of the model by adding the divisiveness (Table 4).

Segmentation method	Accuracy	Execution time (s)

SVM	89,7	204
Decision Tree	96,2	60
RFM (KMEANS)	78,63	58
RFMS (KMEANS)	96,86	62

Table 4. RFMD Vs Machine Learning segmentation's methods

In terms of response time, the decision tree method remains the best performing, followed by the RFMD model (Fig). While the best performance in terms of accuracy goes to the RFMD model. Finally, Other parameters can be added to RFM, other than divisiveness, namely "L" which represents the customer's membership period. Taking these different parameters into account will produce correlations that can allow for even more precise and finer segmentation. In a future work, the parameters L, R, F, M, and D will all be combined. The analysis will be more reliable when reasoning by product in order to have a more selective segmentation of customers based on an aggregation by product [33]. A customer may be loyal to one type of product but not to other types of products. This paper has shown that it is important to analyze all the variables and their correlations to choose the right parameters to take into account in the marketing model in addition to the classical variables R, F and M. In the case of a retail market, for example, the parameter "Diversity" is important to consider in the analysis and segmentation of the customer data set.

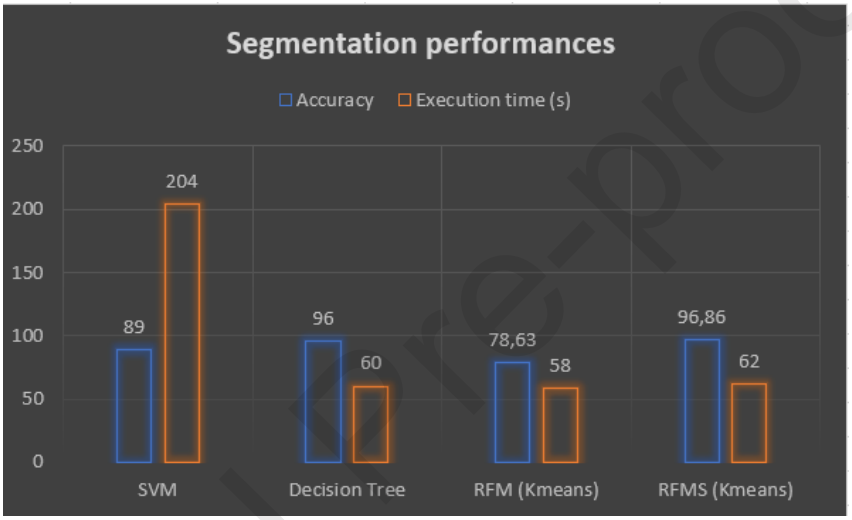


Figure 11. RFMD performances: Accuracy and time execution

Finally, it is important to study in the future a segmentation that search for clusters with customers that can belong to one or more clusters; In this way, the prediction of customer behavior will be more accurate and will avoid excluding customers who are promoters for accepting a given product during a marketing campaign.

REFERENCES

[1] M. Y. Smaili and H. Hachimi, "Hybridization of Improved Binary Bat Algorithm for Optimizing Targeted Offers Problem in Direct Marketing Campaigns," *Advances in Science, Technology and Engineering Systems Journal,* vol. 5, no. 6, pp. 239-246, 2020.

[2] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – An effective approach to customer segmentation," *Journal of King Saud University - Computer and Information Sciences,* vol. 33, no. 10, pp. 1251-1257, 2021.

[3] R. J. I. J. o. B. A. Srivastava and Intelligence, "Identification of customer clusters using RFM model: a case of diverse purchaser classification," vol. 4, no. 2, pp. 45-50, 2016.

[4] F. Safari, N. Safari, and G. A. Montazer, "Customer lifetime value determination based on RFM model," *Marketing Intelligence & Planning,* vol. 34, no. 4, pp. 446-461, 2016.

[5] S. J. M. I. Dibb and Planning, "Market segmentation: strategies for success," 1998.

[6] V. L. Miguéis, A. S. Camanho, and J. F. J. E. S. w. A. e Cunha, "Customer data mining for lifestyle segmentation," vol. 39, no. 10, pp. 9359-9366, 2012.

[7] F. M. Díaz-Pérez, M. J. J. o. D. M. Bethencourt-Cejas, and Management, "CHAID algorithm as an appropriate analytical method for tourism market segmentation," vol. 5, no. 3, pp. 275-282, 2016.

[8] J. A. McCarty and M. Hastak, "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression," *Journal of Business Research,* vol. 60, no. 6, pp. 656-662, 2007.

[9] M. Khajvand and M. J. Tarokh, "Estimating customer future value of different customer segments based on adapted RFM model in retail banking context," *Procedia Computer Science,* vol. 3, pp. 1327-1332, 2011.

[10] A. Amine, B. Bouikhalene, R. J. I. J. o. C. Lbibb, and I. Engineering, "Customer segmentation model in e-commerce using clustering techniques and LRFM model: The case of online stores in Morocco," vol. 9, no. 8, pp. 1993-2003, 2015.

[11] S.-C. Huang, E.-C. Chang, and H.-H. J. E. S. w. A. Wu, "A case study of applying data mining techniques in an outfitter's customer value analysis," vol. 36, no. 3, pp. 5909-5915, 2009.

[12] J.-T. Wei, M.-C. Lee, H.-K. Chen, and H.-H. J. E. S. w. A. Wu, "Customer relationship management in the hairdressing industry: An application of data mining techniques," vol. 40, no. 18, pp. 7513-7518, 2013.

[13] R. Florez-Lopez and J. M. Ramon-Jeronimo, "Marketing Segmentation Through Machine Learning Models," *Social Science Computer Review,* vol. 27, no. 1, pp. 96-117, 2008.

[14] P. D. Berger and N. I. J. J. o. i. m. Nasr, "Customer lifetime value: Marketing models and applications," vol. 12, no. 1, pp. 17-30, 1998.

[15] W. J. Reinartz and V. J. J. o. m. Kumar, "On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing," vol. 64, no. 4, pp. 17-35, 2000.

[16] V. Kotu and B. Deshpande, "Chapter 6 - Association Analysis," in *Data Science (Second Edition)*, V. Kotu and B. Deshpande, Eds.: Morgan Kaufmann, 2019, pp. 199-220.

[17] D. Loshin, *Business intelligence: the savvy manager's guide*. Newnes, 2012.

[18] I. Yosef and C. I. Samuels, "Multifactor Customer Classification model for IP Transit product," in *2014 8th International Conference on Telecommunication Systems Services and Applications (TSSA)*, 2014, pp. 1-9: IEEE.

[19] J. Han *et al.*, "Cbid: A customer behavior identification system using passive tags," vol. 24, no. 5, pp. 2885-2898, 2015.

[20] S. M. S. Hosseini, A. Maleki, and M. R. Gholamian, "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty," *Expert Systems with Applications,* vol. 37, no. 7, pp. 5259-5264, 2010.

[21] Q. K. Al-Shayea and T. K. Al-Shayea, "Customer behavior on RFMT model using neural networks," in *Proceedings of the World Congress on Engineering*, 2014, vol. 1, pp. 49-52.

[22] C. Zuccaro and M. J. I. J. o. B. M. Savard, "Hybrid segmentation of internet banking users," vol. 28, no. 6, pp. 448-464, 2010.

[23] A. Martínez, C. Schmuck, S. Pereverzyev Jr, C. Pirker, and M. J. E. J. o. O. R. Haltmeier, "A machine learning framework for customer purchase prediction in the non-contractual setting," vol. 281, no. 3, pp. 588-596, 2020.

[24] K. Khalili-Damghani, F. Abdi, and S. J. A. S. C. Abolmakarem, "Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries," vol. 73, pp. 816-828, 2018.

[25] W. Qadadeh and S. J. P. c. s. Abdallah, "Customers segmentation in the insurance company (TIC) dataset," vol. 144, pp. 277-290, 2018.

[26] M. Mizuno, A. Saji, U. Sumita, and H. J. E. J. o. O. R. Suzuki, "Optimal threshold analysis of segmentation methods for identifying target customers," vol. 186, no. 1, pp. 358-379, 2008.

[27] C. Subbalakshmi, G. R. Krishna, S. K. M. Rao, and P. V. J. P. C. S. Rao, "A method to find optimum number of clusters based on fuzzy silhouette on dynamic data set," vol. 46, pp. 346-353, 2015.
[28] D. Marutho, S. H. Handaka, and E. Wijaya, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," in *2018 international seminar on application for technology of information and communication*, 2018, pp. 533-538: IEEE.
[29] P. Bholowalia and A. J. I. J. o. C. A. Kumar, "EBK-means: A clustering technique based on elbow method and k-means in WSN," vol. 105, no. 9, 2014.
[30] C. Yuan and H. J. J. Yang, "Research on K-value selection method of K-means clustering algorithm," vol. 2, no. 2, pp. 226-235, 2019.
[31] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm," *Journal of King Saud University - Computer and Information Sciences,* 2019.
[32] Y. Sun, D. Cheng, S. Bandyopadhyay, and W. J. M. S. S. J. Xue, "Profitable Retail Customer Identification Based on a Combined Prediction Strategy of Customer Lifetime Value," vol. 24, no. 1, p. 10, 2021.
[33] R. Heldt, C. S. Silveira, and F. B. Luce, "Predicting customer value per product: From RFM to RFM/P," *Journal of Business Research,* vol. 127, pp. 444-453, 2021.

Author declaration

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Author's name (Fist, Last)	Signature	Date
1. Moulay Youssef SMAILI	_____	___21/01/2022____
2. HANAA HACHIMI	_____	___21/01/2022____