Automated Market Analysis by RFMx Encoding Based Customer Segmentation using Initial Centroid Selection Optimized K-means Clustering Algorithm

Ahmed Maghawry
Department of Computer Science, College of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport (AASTMT), Alexandria, Egypt.
ahmed_maghawry@efinance.com.eg

Ahmed Alqassed,
Business Solutions Department, E-Finance Cairo, Egypt.
ahmed_alqassed@efinance.com.eg

Mohamed Awad,
Business Solutions Department, E-Finance Cairo, Egypt.
mohamed_awad@efinance.com.eg

Mohamed Kholief
Department of Computer Science, College of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport (AASTMT), Alexandria, Egypt. Email: kholief@gmail.com

***Abstract*— Market analysis including customer segmentation is one of the most important approaches utilized by business owners to analyze customer behavior. Such analysis can provide significant insights and decision support for businesses. Multiple research effort was conducted for market analysis including the Recency, Frequency and Monetary analysis (RFM) in addition to many variations including RFD, RFE, RFM-I and RFMTC. In this research a methodology is proposed to utilize the intermediate vector representation of the introduced RFMx for machine learning toward high precision automatic customer segmentation. In this methodology there's no need to calculate the actual final RFMx score. The RFMx technique introduces a multi-monetary model where each monetary value is assigned different weight to suite the business targets of business owners. The proposed model allowed for finely tuned market analyses on product type or service type level. The results showed significant clustering results that lead to automatic customer segmentation without the need to calculate the final RFMx score.**

Keywords: Customer Segmentation, Artificial Intelligence.

I. INTRODUCTION

The concept of digital transformation is crawling toward all aspects of our lives [7]. Almost all fields are affected with a variety of artificial intelligence techniques being used to maximize the benefit of digital transformation [9][14]. The field of marketing is significantly concerned as it is significantly important for marketers to understand their customer. For a marketer to gain significant business benefit from their business model, they should focus on retention, loyalty and building customer relationship instead of simply generating more clicks by customers visiting their web pages. Analyzing the whole customer base will not be as efficient as segmenting them into homogeneous groups, then understand the characteristics of each group and finally assign the proper action toward each group according to their behavior as well as business needs. There exist various customer segmentation methods to support decision making for business owners, however, RFM is considered one of the most popular and effective segmentation method that enables marketers to analyze customer behavior. In this research, a modified version of the RFM model labeled as the RFMx model is introduced.

Furthermore, a vector representation encoding based on the RFMx model is utilized for machine learning to achieve automatic customer segmentation. A custom data set was used to test the proposed methodology where it showed significant results in terms of the recognized patterns as will be discussed in detail in the results section. This paper is organized as follows, section I presents an introduction with a brief to the proposed methodology and an abstract to the obtained results. Section II lays a background to the concepts that will be utilized this research. Section III discusses the challenges facing this research. Section IV introduces the proposed methodology. Section V reviews the obtained results. And finally, section VI discusses the conclusions of this research.

II. BACKGROUND

A. Recency, Frequency and Monetary Analysis RFM

The RFM is a technique used to manifest customer value based on customer's transaction data. Such technique is used widely in both database marketing as well as direct marketing. The RFM [1] technique is short for the three dimensions: Recency, Frequency and Monetary. Recency that reflects how recently a customer purchased either a good or a service. Frequency refers to how frequent do a customer performs a transaction to purchase a good or a service. Monetary refers to how much money do they spend to get such goods or services.

The targeted customer data may be laid out in a tabular structure that consists of columns for the customer name, their date of purchase as well as purchase amount. One way to apply RFM [8] on such data is by assigning a score on a scale from minimum value to a maximum value to each of the dimensions mentioned above for each customer. In such case, the maximum limit score will refer to the preferred customer behavior for this dimension, and the minimum limit score will refer to customers that requires action from business owners. Such actions can range from corrective measures up to full-scale marketing campaigns to assert that customers are well engaged to the business.

Recency can be obtained by getting the number of months or days passed since customers last purchase and then order the data descending by the recency value. And finally assigning the scores for equal sets of records as per scoring definition motivated by business needs. Frequency is obtained by calculating the number of transactions performed by the customer in a specific period of time, then order the data by frequency value and assign the

scores according to the scoring definition the same way as recency. Monetary is obtained the same way as recency and frequency but focusing on the amounts spend by the customers for goods or services.
By the end of this phase, we obtain the form of customer data that will be used to calculate the final RFM score of each customer. The workflow of the basic RFM scoring system is reviewed in Figure.1.

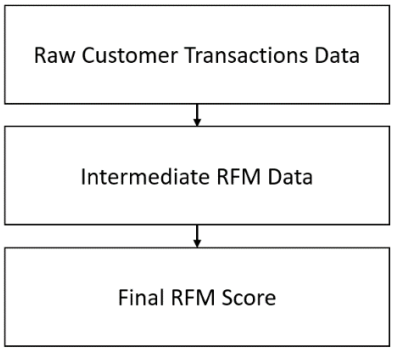

Fig.1. Basic RFM workflow.

The raw customer transaction (Trx) data against the intermediate RFM data are viewed for example in tables 1, 2, 3 and 4.

TABLE 1 CUSTOMER RAW DATA.

Record ID	Customer ID	Trx Date	Trx Amount
1	1001	31-01-2020	500$
2	1002	29-02-2020	300$
3	1003	31-03-2020	100$

TABLE 2 RECENCY SCORING.

Customer ID	Recency	Frequency	Monetary
1003	9 → 3	…	…
1002	10 → 2	…	…
1001	11 → 1	…	…

TABLE 3 ALL DIMENSIONS SCORING.

Customer ID	Recency	Frequency	Monetary
1003	3	10 → 3	500$ → 3
1002	2	5 → 2	300$ →2
1001	1	2 → 1	100$ → 1

TABLE 4 INTERMEDIATE RFM RESULT.

Customer ID	Recency	Frequency	Monetary
1003	3	3	3
1002	2	2	2
1001	1	1	1

Considering the weights of RFM as follows, (R, F, M) : (10,2,10) hypothetically for example purpose, the final RFM score will be as shown in table 5.

TABLE 5 FINAL RFM SCORE.

Customer ID	RFM
1003	66
1002	44
1001	22

Assuming that the scoring weights defined earlier are from pure business needs perspective, the final RFM score in table 5 reflects that customer "1003" is considered the champion satisfying the business needs. Several variations also exist for RFM including RFD [2] – Recency, Frequency, Duration which is a modified variant of the RFM that manifests the customers behavior of viewing business products, for example the time spent viewing a web page on the internet. On the other hand, RFE – Recency, Frequency, Engagement is a generalized version of RFD where engagement reflects visit durations and count of web pages viewed per customer visit. RFM-I [3] where the I stand for interactions is a modified version of RFM that reflect the frequency of interactions with the customer. RFMTC [4][5] – As Time and Churn rate -, is an augmented RFM model [3] that utilizes probability techniques to calculate the probability of a customer buying at the next marketing campaign.

B. The K-Means Clustering Algorithm.

K-means is a non-hierarchical partitioning clustering algorithm [11]. It is widely used and utilized in a variety of science and technology fields [10][12]. It is usually used because it is common on data with different types. It is initialized by several targeted numeric objects N and a specified integer number k. The algorithm then attempts an effort to partition all objects members of N into K clusters while minimizing the sum of squared errors [13].

The algorithm randomly picks K cluster centers from N and attempts to assign each member of N to its closest centroid according to the square of the Euclidean distance [13]. Each centroid is then updated to be the mean vector of each cluster. The algorithm loops to continuously update cluster centers if members allocation is changed, until no more membership changes occur. The following equation is used to calculate how near a data vector is to a cluster's center:

$$d(z_p, a_j) = \sqrt{\sum_{k=1}^{d}(z_{pk} - a_{jk})^2} \quad \text{- (1)}$$

As the final clustering result depends on the quality of the randomly selected initial centroids. Many techniques were introduced to neutralize the algorithms sensitivity to the initially selected centroids including the initial centroid selection optimization technique (ICSO) [15].

III. CHALLENGES

A. RFM Challenges

The main purpose of the RFM model is to introduce a score for each customer that refers to the customer's value for the business owner. However, score calculation will highly influence information delivered to business owners via the RFM score [6]. As the primary values of R, F and M are calculated, the data is ordered descending by each primary value and a score within the defined scale is assigned to corresponding groups of customers, such that, if the scale is from 1 to 3 then the customers will be split into three groups where the first 33.33% receives a score of 3 and the next 33.33% of customers will received 2 and finally the

last 33.33% customers will receive a score of 1. Then such score is multiplied by parameters weights and the summation of the three values is introduced as the final RFM score to a customer. In this research the introduced RFMx technique is used to get the intermediate RFM results as will be discussed in section 4. In this research we propose the view that the basic RFM score will be deficient in multi category transaction data. For example, a bank customer that performs two transactions, one transaction withdraw 100$ and another transaction deposit 100$.
Basic RFM will either combine both values to deliver a monetary amount of 200$ or produce two different RFMs for each transaction type. Section 4 will discuss the proposed methodology to achieve precise RFM based on business needs. In conclusion, to achieve a solid RFM model that delivers insights and decision support, the final scores of all customers is usually empirically segmented by the analyst. Such empirical segmentation is a subject of trial and error, since the analyst will have no clue of the optimal number of segments before assuming some. This also makes the final segmentation result subject to error. In this research, the proposed methodology aims to introduce automatic segmentation through pattern recognition and without even having to calculate the final RFM score.

IV. Auto Segmentation using RFMx based encoding with K-Means

The main idea of RFMx is to target multi category transactions, for example, multiple transactions types (deposit, withdrawal), multiple product types (product 1, product 2) etc... Our view toward such data is to grant business owners the ability to define the RFM scoring on product type level, such that, the RFM definer will split the M value into x M values one corresponding to each product or transaction type. Doing this, will also enable the RFM scoring definer to define x weights for each M that will direct the RFM score toward the motivation and strategy defined by business owners to fulfil their business needs. Consider hypothetical data in table 6.

Table 6 Multi product customers transactions.

Trx ID	C ID	C Name	Trx Date	Product Type ID	Trx Amt
1	1	Mark	1-1-2021	1001	1200
2	1	Mark	2-1-2021	1020	110
3	1	Mark	3-1-2021	1003	75
4	1	Mark	3-1-2021	1033	120
5	2	John	30-12-2020	1001	1000
6	2	John	31-12-2020	1020	100
7	2	John	31-12-2020	1033	75
8	2	John	31-12-2020	1022	15
9	3	Sarah	5-11-2020	1001	800
10	3	Sarah	5-11-2020	1022	10
11	3	Sarah	5-11-2020	1003	65

Defining a business efficient RFM model requires a solid scoring and weighing decided by experts from strong business point of view and motivated by business needs. For experiment, example of the proposed methodology is configured with recency weight (100), frequency weight (50) with several monetary weights as in table 7.

Table 7. Product types.

Product Type ID	Product Type Weight
1001	2
1003	4
1020	3
1022	10
1033	5

The proposed methodology is combined of two main models, model 1 and model 2. Model 1 focuses on calculating the final RFMx score of multi-product purchases in one M value to introduce a non-product driven RFMx score value to the business owner.

On the other hand, Model 2 proposes the splitting of the M value itself into x Ms one M for each product type to deliver a detailed and a product type driven vector representation that can be later used for unsupervised machine learning approaches to achieve automatic customer segmentation for decision makers as shown in Figure.2 and Figure.3.

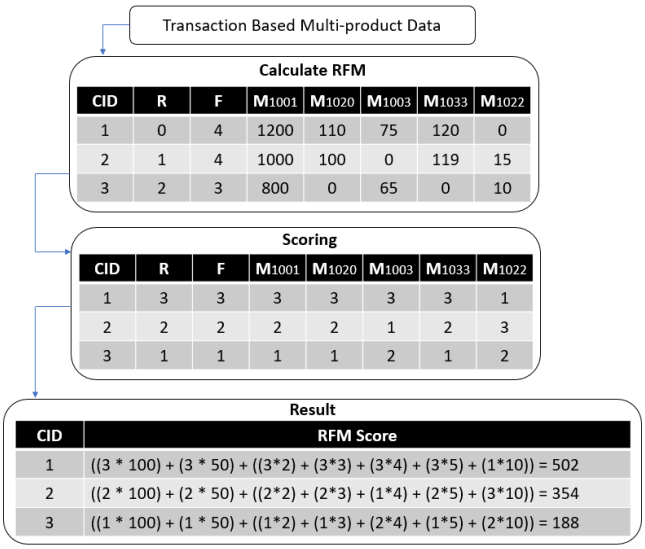

Fig.2 Model full cycle from data to score.

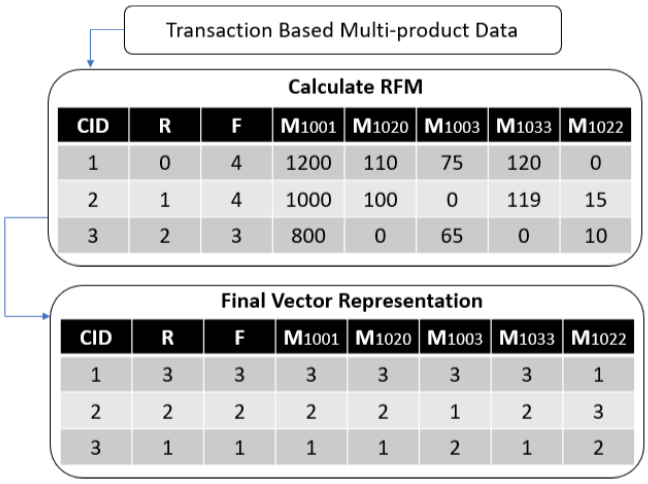

Fig.3 Model encoding cycle from data to vector representation.

The proposed methodology will also utilize a modified version of the K-means clustering algorithm to further target the RMFx encoded data vectors. The clustering aims to explore patterns of transactions of similar behavior. And finally, the patterns are aimed to be introduced with analytics to marketers for further insights and decision

support. Figure.4 introduces an outline of the proposed methodology.

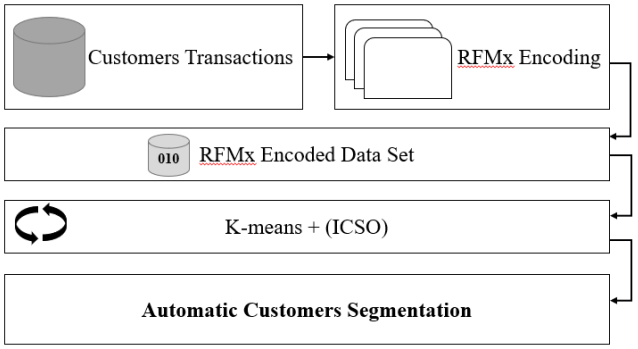

Fig.4 Proposed methodology outline.

V. RESULTS

The proposed methodology has been applied against a data set that consists of 7682 customers with a total of 9,828,428 transactions performed by all customers on 6 product types. For the targeted data set, the RFM and RFMx were calculated on a 10 key customer segmentation bases and the results of both cases are as shown in figures 5 to 8.

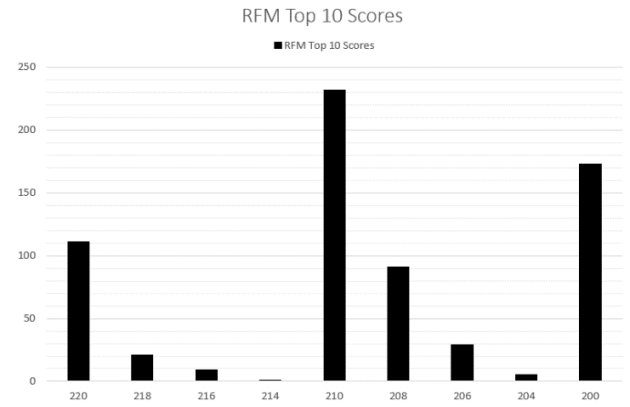

Fig.5 RFM Top 9 Scores Set Counts.

The above RFM results show the score set counts with extreme differences indicating that too many subjects have the same score, our vision is that it will be better if these large blockes can be divided into smaller groups so that marketers can have better insights.

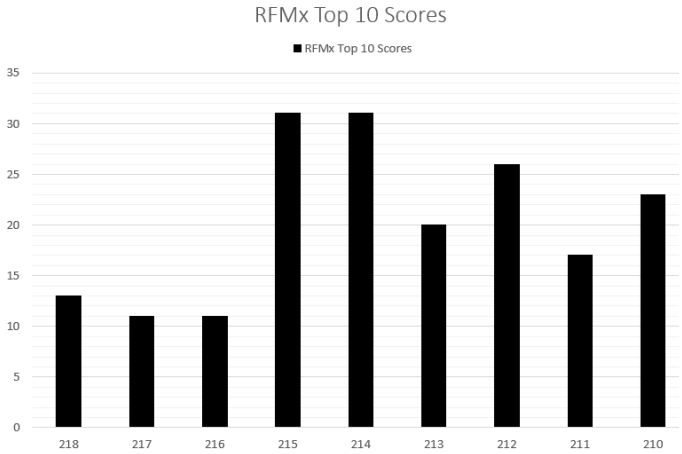

Fig.6 Model 1 RFMx Top 9 Scores Set Counts.

The RFMx results of the modified RFM shoes a more homogeneous distribution of the score set counts, indicating that many customers were removed from large score groups into other groups as the modified RFM model succeeded to distinguish them with different score. Hence providing better customer score distribution.

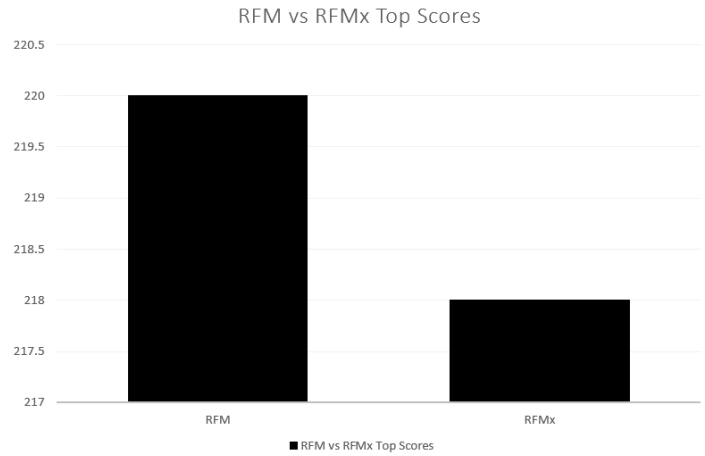

Fig.7 RFM vs RFMx top scorers.

Figure.7 shows a comparison between the top score achieved by the best customer in the RFM model which was 220, and the top score achieved by the best customer in the RFMx model which was 218. Obviously the top score decreased indicating that the customer that was able to achieve a score of 220 in the RFM model, only achieved 218 in the RFMx model. This change will have impact in decision making and insights.

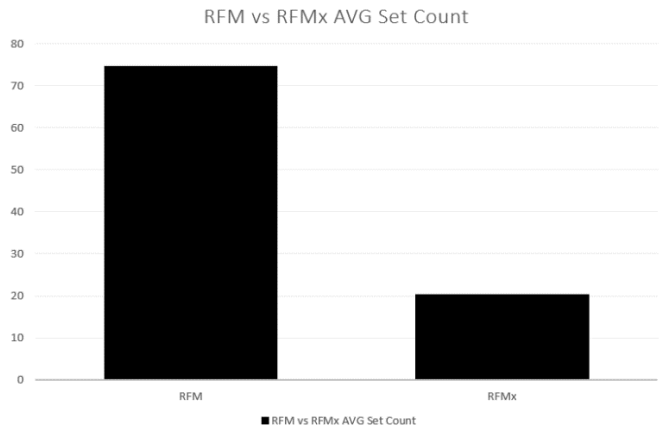

Fig. 8 RFM vs RFMx Average Set Count.

Figure.8 shows the average score set count of the RFM compared to the RFMx model. The comparison shows that the average score set count decreased from approximately 75 members per score set to approximately 20 members. This comparison is an indicator to the redistribution of score set memberships, which concludes that the RFMx model changed the previous insights obtained from the RFM model.

The proposed methodology was applied to cluster the produced RFMx based intermediate vector representation before calculating the actual RFMx score. This was applied in order to achieve automatic customer segmentation using unsupervised machine learning without having to calculate the actual RFMx score and without human intervention. Tables 8 and 9 shows the clustering results for 10 consecutive runs.

TABLE 8 ITERATIONS AND CLUSTERS COUNT FOR CLUSTERS 0 TO 4.

ITR	0	1	2	3	4
28	736	725	332	2191	582
31	1096	2349	330	471	369
35	557	2293	695	439	508
34	590	757	695	2141	437
63	2169	512	254	338	598
25	608	1439	413	325	92
37	2229	413	1548	790	506
42	2432	347	753	915	458
44	1176	330	723	98	1012
32	342	772	1299	611	260

TABLE 9 CLUSTERS COUNT FOR CLUSTERS 5 TO 9.

5	6	7	8	9
1145	410	447	772	342
1500	472	427	327	341
255	1107	335	952	541
972	242	347	394	1107
682	1107	1037	553	432
1034	2502	648	249	372
95	297	722	750	332
327	405	427	1248	370
484	327	384	2563	585
331	2700	342	475	550

Average members count per cluster shown for applying normal K-means vs K-means + ICSO shown in Figure.9.

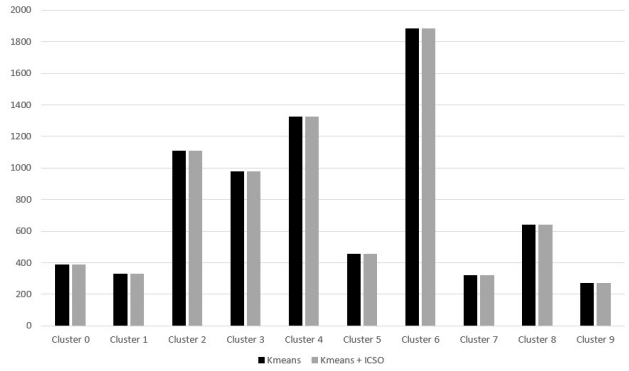

Fig.9 KMeans VS KMeans+ ICSO.

Average iterations of the two execution configurations are shown in Figure.10.

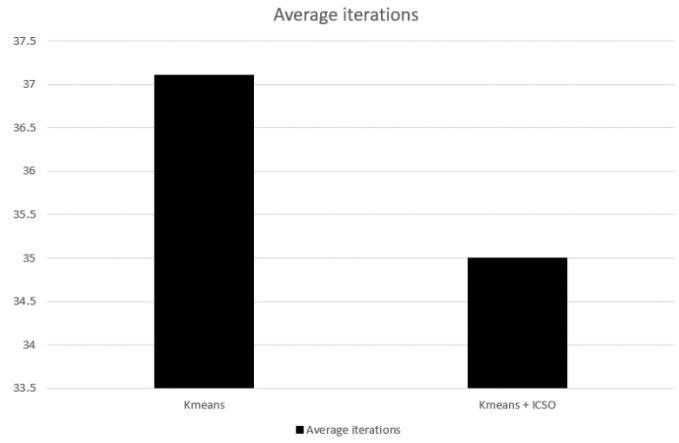

Fig.10 KMeans VS KMeans+ ICSO AVG iterations.

Figure.9 shows a comparison between the average iteration consumed to cluster the RFMx encoded data. The first test case utilized only the K-means clustering algorithm that resulted in approximately 37 iterations in average of 10 consecutive runs. While the same data was clustered using the K-means + ICSO and the average iterations decreased due to the efficiency of selection of the initial random centroids.

VI. CONCLUSION

In this research, the RFMx as an efficient modified RFM model was utilized to calculate the RFMx score targeting customer transaction data where a customer can perform more than one transaction type manifested in product\service type. The utilized methodology succeeded to introduce a score that's influenced by the weights of different products\services. Such scoring system provides a significantly flexible score definition mechanism to help business owners optimize and configure the scoring definition on transaction type level to achieve their business goals. Furthermore, the automatic customer segmentation without calculating the RFMx score was also achieved. The automatic customer segmentation was achieved by utilizing the intermediate vector representation of the RFMx model to be targeted by unsupervised machine learning technique for clustering. The algorithm succeeded on both configurations the basic and +ICSO to cluster the customer data and provide 10 segments as defined in this research's test case.

By achieving automatic customer segmentation depending on unsupervised machine learning techniques, business owners can utilize their data from collection to segmentation fully automated and without human intervention. The output of the clustering algorithm and further be investigated by data analysis to visualize the different patterns recognized by the clustering algorithm, understand them and finally assign proper countermeasures toward each segment according to business goals.

References

[1] Wei, Jo-Ting & Lin, Shih-Yen & Wu, Hsin-Hung. (2010). A review of the application of RFM model. African Journal of Business Management December Special Review. 4. 4199-4206.
[2] Tkachenko, Yegor. Autonomous CRM Control via CLV Approximation with Deep Reinforcement Learning in Discrete and Continuous Action Space. (April 8, 2015). arXiv.org: https://arxiv.org/abs/1504.01840
[3] Yeh, I-Cheng, Yang, King-Jang, and Ting, Tao-Ming, "Knowledge discovery on RFM model using Bernoulli sequence," Expert Systems with Applications, 2009.
[4] "GitHub - it21208/RFMTC-Implementation-Using-the-CDNOW-dataset". 2018-12-17.
[5] "RFMTC (New Marketing Predictive Model / Bernoulli Sequence) Using the Blood Transfusion Dataset: It21208/RFMTC-Using-the-Blood-Transfusion-Dataset". 2018-12-17.
[6] Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. Journal of Marketing Research, 42(4), 415-430.
[7] Vacca, A., Simpson, C., & Smith, E. (n.d.). Worldwide Digital Transformation Spending Guide. Retrieved January 12, 2021, from https://www.idc.com/getdoc.jsp?containerId=IDC_P32575.
[8] Yeh, I-Cheng, Yang, King-Jang, and Ting, Tao-Ming, "Knowledge discovery on RFM model using Bernoulli sequence," Expert Systems with Applications, 2009.
[9] Maghawry, A., Hodhod, R., Omar, Y. et al. An approach for optimizing multi-objective problems using hybrid genetic algorithms. Soft Comput (2020). https://doi.org/10.1007/s00500-020-05149-3.
[10] Xu, R., Wunschii, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. 16, 645–678 (2005). doi:10.1109/tnn.2005.845141.
[11] An Introduction to Classification and Clustering. Cluster Analysis Wiley Series in Probability and Statistics, pp. 1–13 (2011). doi:10.1002/9780470977811.ch1.
[12] Hamerly, G., Drake, J.: Accelerating Lloyd's algorithm for k-means clustering. Partitional Clust. Algorithms (2014). doi:10.1007/978-3-319-09259-1_2.
[13] Shrivastava, P., Kavita, P., Singh, S., Shukla, M.: Comparative analysis in between the k-means algorithm, k-means using with Gaussian mixture model and fuzzy c means algorithm. Commun. Comput. Syst. (2016). doi:10.1201/9781315364094-186.
[14] Boneva, Miroslava. (2018). Challenges Related to the Digital Transformation of Business Companies.
[15] Maghawry, Ahmed & Omar, Yasser & Badr, Amr. (2018). Initial Centroid Selection Optimization for K-Means with Genetic Algorithm to Enhance Clustering of Transcribed Arabic Broadcast News Documents. Advances in Intelligent Systems and Computing. 662. 86-101. 10.1007/978-3-319-67621-0_8