Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method

Yoseph, Fahed; Heikkilä, Markku

Published in:
2018 International Conference on Machine Learning and Data Engineering (iCMLDE)

DOI:
10.1109/iCMLDE.2018.00029

Published: 01/01/2018

Document Version
Accepted author manuscript

Document License
Publisher rights policy

Segmenting Retail Customers with Enhanced RFM Data using a Hybrid Regression/Clustering Method

Fahed Yoseph, Professor Markku Heikkilä Åbo Akademi University, Finland, Turku

***Abstract*—Targeted marketing strategies attract interest from both industry and academia. A viable approach for gaining insight into the heterogeneity of customer purchase lifecycle is market segmentation. Conventional market segmentation models often ignore the evolution of customers' behavior over time. Therefore, retailers often end up spending their limited resources attempting to serve unprofitable customers. This study looks into the integration of Recency, Frequency, Monetary scores and Customer Lifetime Value model, and applies the resulting data to segment customers of a medium-sized clothing and fashion accessory retailer in Kuwait. A modified regression algorithm is implemented for finding the slope for customer purchase curve. Then K-means and Expectation Maximization clustering algorithms are used to find the sign of the curve. The purpose is to gain knowledge from point-of-sales data and help the retailer to make informed decisions. Cluster quality assessment concludes that the EM algorithm outperformed k-means algorithm in finding relevant segments. Finally, appropriate marketing strategies are suggested in accordance with the results generated by EM clustering algorithm.**

Keywords—Segmentation; Clustering; RFM model; Retailing

I. INTRODUCTION

Mass marketing strategies are mainly based on marketing experts' and sales managers' opinions about customer information [42]. For example, Hossein [25] states that the retail industry is highly competitive and the number of products is many times overwhelming. Consumers are faced with a variety of products, causing the demand to be higher and more complex. To follow this trend, modern marketing is moving from mass-marketing (products-focus) to target-marketing (customer-focus). This is leading the small and medium-sized retailers (SMR) to segment their markets to be able to design and implement successful marketing strategies and retention policies. Segmentation techniques have been shown to have a great impact in empowering retailers to precisely reach a consumer with specific needs and wants by dividing the market into similar and identifiable segments and to help to focus on individuals with similar preferences, choices, needs and interests [28], [31]. The objective of segmentation is to evaluate the value of customers as a segment indirectly instead of evaluating them directly. By segmentation, retailers can then decide how to make full use of their limited resources to serve customers effectively as customer sub-groups [12].

A Kuwaiti retail chain has customers apart from the state of Kuwait also from neighboring countries from the Gulf of Persia, and Arabic-speaking countries, in general, have faced a considerable change and growth of customer base. The demographic features in the customer-base are manifold, and due to the demographic changes in Kuwait and other Arabic–speaking countries it is no longer possible to define customer profiles based on known, previous consumption patterns.

To take on this challenge, this study will focus on incorporating Recency, Frequency, Monetary (RFM) scores with Customer Lifetime Value (CLV) model. RFM and CLV models are two of the most important techniques frequently used in market segmentation. According to Birant [4], RFM distinguishes important customers and identifies customers' purchase behavior by three dimensions, customer's most recent purchase (R), the frequency of purchases (F), and spent money (M). With these variables and with appropriate feature weights, an RFM-score is calculated and used as a key figure in segmentation. On the other hand, CLV is a quantitative measurement of the amount of sales the customer is expected to spend with a retailer over their lifetime [16]. The CLV model is here used as an addition to the RFM model to predict the future cash flows attributed to the customer during his or her entire lifetime with the retailer [42].

In the fast-changing retail industry, there is a clear need for advanced methods for finding market segments with data such as RFM and CLV, and this has also been identified in the case company. In addition, to convert ever-increasing transaction data to knowledge requires proper mechanisms for treating point-of-sales (POS) events [41]. In this study, we will use POS data of a medium-sized retailer from Kuwait. The benefit of POS data is that it is usually generated and stored in a structured way, and is relatively easy to aggregate to customer-level. This characteristic of POS data makes it useful for analyses that require (more or less) complete data, such as K-means clustering.

Our objective is to design an advanced segmentation method, construct a system utilizing the method and test this system (the artefact) by analyzing the results produced by the system. We want to show how the market segmentation process can be improved with our hybrid approach that utilizes both regression and clustering as steps of the analysis of a POS data warehouse and show that with an appropriate design more usable results can be produced. More specifically, we want to answer the following research questions:

1. Can a method with regression and clustering steps help in breaking down structured RDM/CLV data into meaningful segments for marketing in SMR companies?
2. Can the segments be used by the decision-makers in making better-informed decisions?
3. What is the methodological validity of the obtained results?

In the following sections, we will elaborate this by reviewing appropriate literature in section 2, constructing the system in section 3 and analyzing the results in section 4. Finally, our study is concluded in section 5 with answers to our research questions.

II. MARKET SEGMENTATION

The US Small Business Administration has traditionally defined Small to Medium Size Retailers as businesses employing fewer than 500 employees [40]. According to the European Commission, the SMR industry forms the backbone of the economy and are the key players in the creation of new jobs and economic growth [22]. Historically, small size retailers have had the privilege of developing close and mutually beneficial relationships with their customers, thus keeping existing customers and reaching new markets is a major challenge for the retailer [18], [23]. These relationships were possible because consumer's buying behavior did not change much, and the price was less of an issue due to less competition [6]. However, the recent economic and social changes have transformed the retail industry, particularly the relationship between the retailer and customers has changed significantly. As a result, retailers have been forced to seek new marketing strategies to identify the profitable segment of customers, to develop marketing mixes that appeal to those potential segments of customers and to focus on providing value to the key segments of customers [18], [23].

Dipanjan, Satish, and Goutam [14] defined market segmentation as the process to divide customers into similar or homogeneous groups sharing one or more characteristics such as shopping habits, lifestyle, taste, and food preferences. According to McCarty and Hastak [38] these characteristics, such as demographics, age, location, nationality, gender, interests and spending habits, are relevant to marketing and sales (see Table I for a summary of market segmentation bases). Also, [28] and [29] state that segmentation methods have great importance in empowering retailers to precisely reach a consumer with specific needs. Today, companies look for customer service as a market differentiator, and many companies have started to segment customers for service delivery [39]. There are several both qualitative and quantitative methods for collecting, analyzing and processing data [3]. According to Elby [17], a combination of qualitative and quantitative methods is the best practice in most cases.

TABLE I. MARKET SEGMENTATION BASES

Base: Description	Goals Benefits	Ref.
Demographic; Identifiable population features	The goal is to have a precise customer purchase profile and focuses on measurable criteria of consumers and their households.	[9]
Geographic; Location-related features	The goal is to map consumer wants and needs from nationality, region to another etc.	[23]
Behavioral; Product attitudes, customer relationship-related features	The goal is to identify behavioral variables such as occasions, benefits, user status, usage rate, buyer-readiness stage, loyalty status and attitude.	[9], [5]
Psycho-graphic; Lifestyle-related features	The goal is to target specific groups such as more budget-conscious customers, i.e. smart shoppers who value a good deal.	[5]

A. Segmentation models

The literature has traditionally defined RFM analysis as the standard approach to assess and understand customer lifetime value, and it is quite popular, especially in the retail industry. According to Tsiptsis and Chorianopoulos [46], RFM involves the calculation and the examination of three variables – Recency, Frequency, and Monetary (RFM). Recency refers to the inverse of the most recent interval from the time when the latest consuming behavior happens to the present moment. Frequency is the number of events the consumer purchases in a period. Monetary is simply the amount of money consumed during a period. As the weighted average of its individual components, the RFM score and is calculated as

$$\text{RFM score}=(rs \times rw)+(fs \times fw)+(ms \times mw)\ , \quad (1)$$

where *rs* = recency score and *rw* = recency weight, *fs* = frequency score and *fw* = frequency weight, *ms* = monetary score and *mw* = monetary weight.

One limitation of RFM analysis in market segmentation is that the features are assumed static, and they ignore behavioral changes. The recency parameter, however, indicates a momentary change, but it only shows one static, transient event and cannot properly capture long-term dynamic changes in customer behavior. This is the reason for our proposal to apply a new dynamic variable (C) to show the quantity and sign of change in customer purchase behavior.

Dwyer [16] defines customer lifetime value (CLV) as a quantitative measurement of the amount of sales the customer is expected to spend with a retailer over their lifetime. Furthermore, Safari [41] considers CLV as the present value of all future profits obtained from a customer over his or her lifetime relationship with the retailer. To better utilize CLV in every-day decision making, Marcus [37] introduced CLV matrix as a variant of the RFM analysis for small-business retailers. In CLV matrix, F, the frequency of purchase and M, the average purchase amount are used for the segmenting customers. The easiness to understand quadrant identifiers is considered as its main advantage. In Marcus' approach, the average values for the number of purchases and the average amount spent per customer are calculated. After identifying these, each customer is segmented to one of the four resulting

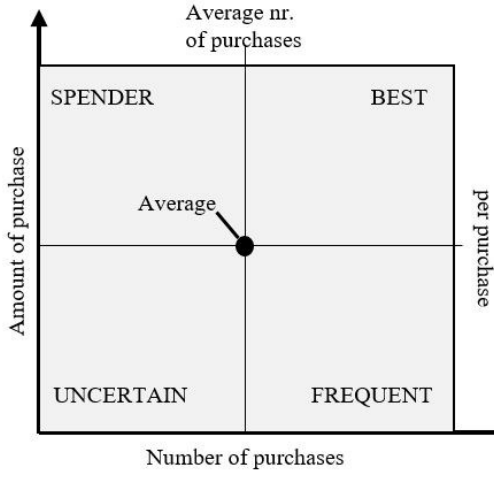

Fig. 1. The customer value matrix

categories (quadrants) based on whether customers are above or below the axis averages (Fig. 1).

B. Data mining

Modern information and communication technology generates massive amounts of data to databases, data warehouses, and other repositories. Transforming the insights about (big) data into knowledge can help retailers to make better business decisions [9]. Tufféry [47] sees data mining as a powerful analytical tool for finding insight into the retail industry. According to Azevedo [2], data mining is used to provide the analyses on product sales, customer buying habits, data and identify naturally occurring clusters of behavior, which then form the basis of segments. Ramageri and Desai [41] say that in the retail sector, data mining offers insightful measures, taking into account all the factors that affect the value of the customer to the retailer over the entire course of customer relationship [40].

Gunaseelan and Uma [24] stated that the main aim in data mining is to discover valuable patterns from a large collection of data for users. It can identify patterns, and apply data analysis and discovery algorithms to produce a data mining model. Models help in generating a model, a hypothesis about the data, that key executives can use to make better-informed decisions [51]. There are two primary data mining process goals, which are verification and discovery. Verification is verifying the user's hypothesis about the data while discovery is automation of finding unknown patterns.

C. Clustering Methods

According to Lefait and Kechadi [33], clustering consists of "creating groups of objects based on their features in such a way that the objects belonging to the same groups are similar, and those belonging to different groups are dissimilar. Clustering analysis is one of the most important and prominent market segmentation techniques, and it has long been the dominant and preferred method for market segmentation [49], [28]. For example, [26] combined weighted RFM model into K-Means algorithms to improve customer relationship management. Khavand and Tarokh [29] proposed a data mining tool to prepare a framework for segmenting customers based on their estimated future CLV value in an Iranian private bank, and the method was implemented in a health and beauty company, as well [30]. In retail sales clustering methods have been applied at least in groceries [34], online retail [8] and for identifying strategies for new ventures [7].

MacQueen [34] was the first to name k-means clustering, which is one of the simplest and fastest learning algorithms and is applicable to market segmentation to arrive at appropriate forecasting and planning decisions [48]. K-means clustering is one of the most common methods of data mining and machine learning and useful when the objective is to find patterns and structures from data or to guide researchers to choose appropriate supervised machine learning methods. K-means, on the other hand, as well as other clustering methods learns from data only, i.e. learning is unsupervised. K in K-means represents the number of clusters chosen. Observation is assigned to a particular cluster of which its distance to the cluster mean is the smallest [11], [32].

Expectation-maximization (EM) clustering algorithm [13] is closely related to the K-Means algorithm. In this algorithm, two subsequent steps are iterated until there are no more changes in the current hypothesis [27]. In the Expectation-step (E-step) the probability that each observation is a member of each of the chosen class is calculated. Maximization-step (M-step) alters the parameters of each class with the objective to maximize those probabilities. The iteration is then repeated until converging to a (local) optimum.

III. METHODS APPLIED

Here we present data mining techniques used in this study. The analysis process takes four phases. The first phase focuses on the data preprocessing, i.e., data cleansing, feature selection, and data transformation. Regression and clustering algorithms are applied in the second and third phase, respectively. We enhanced the data to consist of four different variables, (R, F, M, and C). The variable C is the Customer Change Rate (a trend) that shows the quantity and sign of a change of customer purchase behavior. The modified regression algorithm is used to find this trend. The output is then fed into the clustering algorithms with RFM data. K-means and EM clustering algorithms are used for the segmentation. At the final phase, the accuracy of these partitions is measured by the cluster quality assessment introduced by Drăghici [15].

The point-of-sales database consists of all product sales and shows that the client sells diverse products like clothing, shoes, schools supplies, and accessories. Transaction data from three years were retrieved and extracted. Each transaction represents a purchase event, with each line consisting of a transaction cashier, store code, item code, brand code, products (quantity) sold, unit price and the total (quantity multiplied by the unit price), date and time of the purchase and information about the customer. Both active and inactive customers are included in the study. This is vital to enable the marketing team to develop appropriate marketing and retention campaigns.

A. Phase 1: Data Pre-processing

To prepare for this-this stage, several interviews were conducted with marketing experts, sales directors, the IT manager, in-store employees, and POS engineers. Interviews intend to maximize variation in responses to gain a deep understanding on the challenges experienced at the client company in general and, more specifically, to find information and company insights about the retail industry, the market, and the customer base.

a) Data transformation using RFM model

Based on the client information, RFM values were assigned. The latest purchase date of the customer R is found from a set of 1095 days (records from 2013 to 2016), a number of transactions during this 1095-day period F comes from totally 11550 transactions, and total amount purchased M comes from the total sales of 1,300,000 KD. The RFM attributes were weighted with category 10. The data is extracted from POS database and fed into a unified, generic POS data warehouse in a common format with consistent definitions for keys and fields. In this phase, the string variables must be converted to numeric variables. The missing values were checked and deleted or replaced by default or mean values of each parameter.

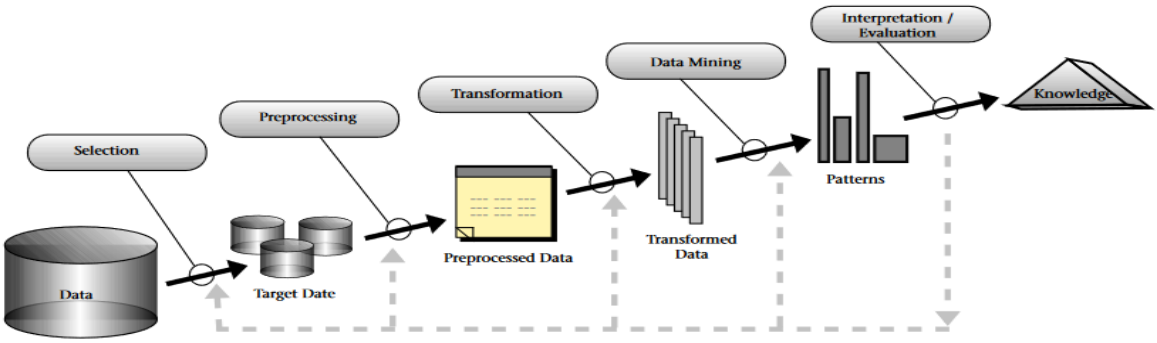

Fig. 2. Data mining process

The rest of the data preprocessing phase handle noisy data, missing values and makes attributes reduction and transformation. In this step, the data must be transformed into an appropriate format, to make the discovery of patterns easier. Continuous customer-related attributes were encoded by decreasing the original values into a small number of value ranges. The age of the customer was encoded into six categories.

Gender attribute was encoded as 1 for Male, 2 for Female and 3 for Companies. Furthermore, demographic concepts like cities and countries were replaced by higher level concept nationality.

For normalizing the RFM score, we use the following rescaling method (2)

$$\text{RFM score}_{\text{scaled}} = \frac{\text{RFM score- min(RFM score)}}{\max \quad \text{(RFM score)-min} \quad \text{(RFM score)}}. \qquad (2)$$

B. Phase 2: Steps of Data Transformation to RFMC

As usual, the client company has limited resources, and they are not able to serve all customers with the same intensity. To manage the available resources more efficiently, the client wants to distinguish between relatively low and high shoppers, choose only profitable market segments and concentrate all efforts on the strategic value of these segments to increase profitability and customer's retention. We decided to incorporate RFM scores with CLV matrix before running the modified regression algorithm to generate the C data set. The CLV matrix will divide the RFM data set into four categories.

The CLV matrix calculates the number of purchases by taking the total number of purchases for the customer and then dividing by the total number of customers in the customer database. The average purchase amount is derived by taking the total revenue and dividing it by the total number of purchases. Comparing the average number of purchases, F between customers and the average purchase amount, M with total average values is the next step. M and F are used to classify each customer into one of the four fields Best, Spender, Frequent and Uncertain.

TABLE 1 ENCODING OF THE AGE

Age	**1-17**	**18-24**	**25-34**	**35-44**	**45-54**	**55 +**
Category	1	2	3	4	5	6

a) Modified regression algorithm to generate C data set

In this study, we propose a combination of two analytical data mining steps. For finding C, the change in customer purchase behavior, we use supervised linear regression method. Then C dataset is then put into the unsupervised clustering algorithm, to split the customers into different groups based on pattern dissimilarities. The variable C should answer a very important question if the customer is at high risk of shifting his or her service to another retailer. One of the most common indicators of high-risk customers is a drop off in purchases and decrease of visits.

A major limitation of RFM and market segmentation models is that they ignore behavioral changes of customers during the period of analysis. Although the recency parameter is one of the indicators of such behavior, it suffers from the transient behavior of the customer. Therefore, introducing a new analysis parameter is essential for retailers to narrow down a group of customers in high risk. From a retailer perspective, each customer has different average values of purchases. If these average purchase amounts decrease continuously, then the customer is on the verge of shifting his or her services to another retailer or falling from the beneficial segment for the non-beneficial segment. Similarly, a customer with an increasing average purchase value during the period of analysis shows the potential that could be harnessed with appropriate marketing actions.

To capture this change of behavior, we first calculate purchase amounts of all customers in each selected period of analysis. The parameter C for time k is defined as

$$C(k) = \begin{Bmatrix} \frac{pa_k - pa_{k-1}}{pa_{k-1}}, pa_{k-1} \neq 1 \\ 1, \text{else} \end{Bmatrix}, \qquad (3)$$

where pa_k indicates the total amount purchased by a customer at time k.

If the data is divided into $n+1$ analysis periods, n changes are calculated. If the change rate values of the latest analysis periods have the same sign (negative or positive), then the average of these values will be used as the final change rate parameter C for each the customer. Customers changing signs

are assigned neutral $C = 0$. The final set consists of positive, negative and neutral change rate values.

Furthermore, since the C dataset consists of numbers with six or more digits, we decided to use the logarithmic of C to get normalized C data the analysis database.

$$C_{scaled} = \log_b C, b>0, b\neq 1 \quad (4)$$

The next step is to apply the modified regression algorithm on each segment separately by using the demographic variables Age, Gender and Nationality. Then the purchase behavior change rate C values are prepared for market segmentation by applying clustering algorithms for every segment.

The solution starts with the calculation of the purchase amount slope of the regression algorithm in the time series of the known k and known M_k, purchases at k. The result is m_1, the slope of linear regression curve

$$M_k = m_0 + m_1 k + \varepsilon_k, \quad (5)$$

where m_0 is the intercept and ε_k is the random parameter.

The m_1 rate determines the expected value and sign of the change rate C, based on past purchases.

To obtain the regression curve we use the purchase slope discount rate w_k to modify the effect of the purchase M_k for each period k. More recent spending gets a higher rate. For instance, for a customer with 4-time periods in the analysis, with purchase slope discount rate 0.7, the latest time M_{k-1} is obtained by multiplying the actual purchase amount by $w_{k-1} = 0.7^1$, M_{k-2} by $w_{k-2} = 0.7^2 = 0.49$, and so on. This method to compute the total purchase amount slope leads to decreasing effect from older purchase is used to capture the alternative cost of customer capital.

Therefore, the technique is as follow, transform the Mk and k shape using the attenuation factor att explained below and then multiplying it to get the right calculated purchase slope.

The final formula is shown in Equation (6).

$$(\text{cps})\ P = \frac{n\sum(\Delta\bar{k}\Delta\overline{M_k}) - (\sum\Delta\bar{k})(\sum\Delta\overline{M_k})}{n\sum\Delta\bar{k}^2 - (\sum\Delta\bar{k})^2} \quad (6)$$

where $\Delta\bar{k} = \Delta k \times att$ and $\Delta\overline{M_k} = \Delta M_k \times att$.

Where k is the selected period time for customer purchase in the time series of purchases, and M_k represents the total amount of a single transaction per customer multiplied with w_k. CPS stands for the customer purchase amount trend, and n is the number of time segments.

Next, we use the new technique by applying attenuations factor. The attenuation factor is a natural arithmetical value of signal transmission over long distances. The attenuation function is used for reduction in the strength of a signal or digits, and it has been used along with other numerical methods to generate and validate test data. In this study, we use to reduce the effect of the old customer's transactions and strengthen the effect of the new transactions. Meaning, old transactions have less effect on the final customer's purchase slop, where newer transactions have more effect on the customer's final purchase slope. att represents the Attenuations Factor in the newly modified equation.

C. Phase 4: Market Segmentation using Clustering

a) k-means algorithm

According to [45], the K-means algorithm is partitional, non-hierarchical clustering method and is often considered suitable for large datasets commonly used in marketing. It has been widely used in market segmentation and pattern recognition because of its good performance, simplicity in implementation and fast execution [48]. According to [44], The K-means algorithm steps are:

Step 1: Initializing the position of the clusters using random cluster centers.

Step 2: Assignment step, where each observation is assigned to the cluster with shortest within-cluster sum of square (WCSS). Because the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean. K-The Euclidean distance of observation, x in three-dimensional coordinates, such as RFM, to a centroid, c is calculated as

$$d(x,c) = \sqrt{(x_r - c_r)^2 + (x_f - c_f)^2 + (x_m - c_m)^2}, \quad (7)$$

Where x_r, x_f, and x_m are normalized recency, frequency and monetary values of observation x, respectively, and c_r, c_f, and c_m are corresponding coordinates of the cluster center c. For additional features such as C additional dimensions are needed.

Step 3: Updating centers, calculates new means to the clusters. Equations (8) calculate the N-dimensional centroid point amid k n-dimensional points,

$$CP(x_1, x_2, \ldots, x_k) = \left(\frac{\sum_{i=1}^{k} x1st}{k}, \frac{\sum_{i=1}^{k} x2nd}{k}, \ldots, \frac{\sum_{i=1}^{k} xnth_i}{k}\right). \quad (8)$$

Step 4: Repeats Step 3 until the centers have converged, i.e. do not change. The centroids are now these centers. Convergence criteria are found by minimizing sum of squared error (SSE) measure. For each observation, x the error is the distance to the nearest cluster. To find the SSE, square the errors and then sum them as shown below as shown in equation (9)

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2 (m_i, x), \quad (9)$$

Where x is a data point in cluster c_i and m_i is the representative point for cluster c_i. The m_i corresponds to the center (mean) of the cluster c_i.

b) Expectation- Maximization (EM) Clustering

The Expectation-Maximization (EM) is an iterative estimation algorithm developed by Dempster, Laird, and Rubin [13]. EM is historically very important algorithm in market segmentation and data mining. EM algorithm has also proven its efficiency in a good performance, decreasing sensitivity to noise and estimation problem involving unlabeled data.

Step 1. Expectation- Maximization (EM) Clustering Initialization, the E-step

Every class j, of M classes (or clusters), is formed by a vector parameter (θ), composed by the mean (μ_j) and by the

covariance matrix (P_j), which defines the Gaussian probability distribution (Normal) features used to characterize the observed and unobserved entities of the data set as shown in Equation (11)

$$\theta(t)=\left(\mu_j(t),P_j(t)\right),j=1,\ldots\ldots,M. \quad (11)$$

On the initial instant ($t=0$) the implementation can generate randomly the initial values of mean (μ_j) and of covariance matrix (P_j). The EM algorithm aims to approximate the parameter vector (θ) of the real distribution.

Fraley and Raftery [20] suggested another alternative to initialize (EM) with the clusters obtained by a hierarchical clustering technique. The relevance degree of the points of each cluster is given by the likelihood of each element attribute in comparison with the attributes of the other elements of cluster C_j as shown in Equation (12) The E-step

$$\left(C_j|x\right)=\frac{|\sum_j(t)^{-\frac{1}{2}}|e^{n_j}P_j(t)}{\sum_{k=1}^{M}|\sum_j(t)|^{\frac{1}{2}}e^{n_j}P_k(t)}, \quad (12)$$

Step 2. M-Step

First is computed the mean (μ) of class j obtained through the mean of all points in function of the relevance degree of each point, as shown in Equation (13)

$$\mu_j(t+1)=\frac{\sum_{k=1}^{N}P(C_j|x_k)x_k}{\sum_{k=1}^{N}P(C_j|x_k)}, \quad (13)$$

To compute the covariance matrix for the next iteration the with the *Bayes Theorem, P(A|B) = P(B|A) * P(A)P (B)* conditional probabilities of the class occurrence are calculated, as shown in Equation (14)

$$\sum_j(t+1)=\frac{\sum_{k=1}^{N}P(C_j|x_k)(x_k-\mu_j(t))(x_k-\mu_j(t))}{\sum_{k=1}^{N}P(C_j|x_k)}. \quad (14)$$

The probability of occurrence of each class is computed through the mean of probabilities (C_j)

$$P_j(t+1)=\frac{1}{N}\sum_{k=1}^{N}P(C_j|x_k). \quad (15)$$

Step 3. Cluster Convergence

After performing each iteration, a convergence inspection which verifies if the difference of the attributes vector of an iteration to the previous iteration is smaller than an acceptable error tolerance, given by parameter.

IV. Analysis Results and Discussion

A sample data of the generated segment Best is shown in tables Table 4.1 and include information such as the number of samples for each cluster, the average purchase change rate of the cluster as well as the average age, gender and nationality of the cluster.

The k-means analysis shows cluster 4 is the best cluster with positive purchase slope (37.12), while cluster 3 (-44.41) has the worst negative purchase slope.

Table 4.2 shows the four clusters generated by EM. The analysis shows cluster 2 is the least cluster with negative purchase slope (+3.82) and it is the most beneficial segment because it is superior to the other clusters 2. The analysis indicates that customers between the age of (18-24) to (25-34), 3. Average Gender (1.74) indicates that Female customers and very small percentage of Male customers and 4. Average Nationality (9.11) citizens of Kuwait and Saudi Arabia are best customers in in this cluster. The analysis also shows cluster 1 (-150.98) has the worst negative purchase slope.

A. Accuracy and Effectiveness Determination (Inter-cluster Distance)

Customer dataset can be clustered in many ways, but how can we know the clusters are accurate and meaningful. There is one unique way to find out the meaningful of the cluster. Clustering method has become a key technique in analyzing quality assessment in variety of the recent study. There are several studies suggestions for a measuring the similarity between clustering algorithms. Those measures are used to compare how accurate different clustering algorithms perform on a particular dataset. These measures are usually tied to the type of benchmark being considered in assessing the cluster quality method. The approach of Draghici [15] is to compare the size of the clusters vs. the distance to the nearest cluster (the inter-cluster distance vs. the size (diameter) of the cluster). In another word, the distance between the members of a cluster and the cluster's center, and the diameter of the smallest sphere containing the cluster. If the inter-cluster distance is much larger than the size of the clusters, then the clustering method is considered to be trustworthy. Figure 9 shows the quality can be assessed simply by looking at the cluster diameter. Therefore, the cluster can be formed by heuristic even when there is no similarity between clustered patterns. This is occurring because the algorithm forces K clusters to be created.

The above figure shows the quality can be assessed simply by looking at the cluster diameter. Therefore, the cluster can be formed by heuristic even when there is no similarity between clustered patterns.

TABLE 2 FOUR CLUSTERS (CL) BASED ON THE BEST SEGMENT GENERATED BY K-MEANS ALGORITHM

Cl	Customers	Purchase Change Rate	Av. Age	Av. Gender	Av. Nationality
1	102	-27.39	3.83	2.04	2.59
2	55	-5.39	3.73	2.00	2.71
3	61	-44.41	3.36	2.08	11.79
4	43	37.12	3.07	2.02	16.28
Tot.	**261**	**-16.11**	**3.57**	**2.04**	**7.02**

TABLE 3 FOUR CLUSTERS (CL) BASED ON THE BEST SEGMENT GENERATED BY EM ALGORITHM.

Cl	Customers	Purchase Change Rate	Av. Age	Av. Gender	Av. Nationality
1	54	-150.98	3.80	1.09	2.37
2	19	+3.82	2.53	1.74	1.7
3	232	-199.13	3.81	1.94	17.85
4	560	-128.54	3.41	2.00	2.95
Tot.	**865**	**-147.82**	**3.52**	**1.92**	**7.05**

Comparing both algorithms using the cluster quality assessment results for the Best segment shows the EM clustering algorithm quality assessment results with the size of the cluster (2.45967329) is far more accurate than the K-means algorithm with the size of the cluster (0.001661999).

B. RFMC Dataset Summary

According to the EM analysis results of the Best Segment, all four clusters in the Best segment have a negative purchase slope, with a total average purchase slope of (-147.82). This ranks the Best segment as the highest negative purchase slope. In the Spender segment, the EM analysis results show the effectiveness of the new adapted target marketing strategy. Cluster 3 is showing positive purchase slope, but the total average purchase slope is (-16.10), this is the segment with the biggest churn rate. The Frequent segment EM analysis results show cluster 2 has a positive purchase slope with (7.17), but the total average purchase slope is (-6.01). This ranks the Frequent Segment slightly above the Spender Segment with respect to the purchase slope.

To the K-means analysis results, the Uncertain segment has the Best positive purchase slope, with a total average purchase slope of (1.78). Also, the analysis shows the positive purchase slope is based on smaller purchases. The overall analysis shows young females from the age 20 – 40 and citizens of the GCC region are the biggest customers with negative purchase slope. Customers who are at high risk of canceling their services or falling to less beneficial segment are easily identified based on the analysis generated from the C dataset. The retailer can develop more targeted retail service model to retain those profitable customers who are the foundation of the retailer interest. These valuable insights derived only from the capability of new proposed RFMC method in identifying customer purchase trend.

The research and data analysis, though often rushed, is the most important stage in the implementation process of market segmentation solution. Our Analysis was conducted under the supervision of internal marketing and sales manager to identify which variable and segment make the most sense to pursue. This framework identifies the model's Strengths and Weaknesses, with special attention paid to the all implications stemming from each. The reason to quire the help of the internal marketing and sales managers is because they have the understanding of the client's capabilities and resources. This will also help the client to focus on the customer and develop marketing mixes for very specific market segment.

Detail description and specification of all segments found in our case study were presented. Based on these specifications, some useful strategies were proposed by the marketing experts.

C. Summary

In this study, we used two methods

1. Draghici approach is to compare the size of the clusters vs. the distance to the nearest cluster.

2. Industry human experts to validating the accuracy and intelligence of the results.

The analysis results, as well as the marketing experts, have agreed that classifying customer purchase behavior using CLV matrix against RFM Dataset revealed the most accurate and important information about customer purchase behavior. The new proposed C proved to have the advantage of taking into consideration the changes in the customer average purchase power over the time with respect to the simple RFM model. The age and gender variables showed strong Indication of accuracy, with an estimated accuracy of 75%. However, the other the nationality variable gave low accuracy, possibly because of the missing data related to these variables.

V. Conclusion, Contribution, and Answers to the Study Questions

Giving the nature of our target industry, simplicity was our key success for designing and developing our market segmentation data mining model. It has been proven that complex statistical modeling methods can provide useful information for experts, but at the same time they are very expensive and very hard to implement and are likely to present a challenge to the implementation of marketing strategies. In this study, RFM method has been used for market segmentation using demographic and behavioral variables. We proposed a novel steps approach which uses CLV and RFM analysis in two data mining tasks, regression, and clustering methods applied separately for every RFM variation. Treating customers with different reflected purchase behavior was one of our main focus in this study. In order to convert this idea into a computable parameter, we proposed a new modified regression algorithm and newly proposed RFM variable (C).

To break down the SMR market into meaningful segments, customer purchase behavior profile exists of demographic variables like age, gender and nationality are segmented and presented accordingly. According to experimental analysis results that market segmentation data mining model using newly proposed methods have shown that the CLV and RFMC methods provided a solid base for measuring and understanding not only for market segmentation with different values of Frequency, Monetary, Average Purchase Power and Purchase Change Rate. The model can identify VIP customers who are at the high risk of shifting their business to another retailer or falling from a

Table 4 Cluster Quality Assessment for the Best Segment Using the EM Algorithm

Best SEGMENT	**EM**		
Inter-Cluster Distance		**Instance 1**	**Instance 2**
D12	7920	2675	2672
D13	165	2675	2669
D14		2675	2671
D23		2672	2669
D24		2672	2671
D34		2669	2671
	Size of Cluster (Diameter)		
Cluster	**Diameter**	**Instance ID**	
D1	724	2675	
D2	1593	2672	
D3	2567	2669	
D4	3220	2671	
Cluster Quality	**2.4596**		

higher segment to lower segments, and also highly profitable products.

The model provided simplicity to measure customer purchase behavioral power and trend, which is the main contribution of this study. The second contribution of this study is proposing new market segmentation method using classification and regression based on the modified RFM followed by clustering on demographic data. The study provided key attributes describing customer's purchase behavior in different demographic eras like gender, age, and nationality. The analytical information gained from demographic data was found exceptionally useful for decision making and strategic planning. By our domain experts, we can conclude that the analysis results of market segmentation study in the SMR industry using demographic attributes can contribute to the body of knowledge in consumer purchase behavior and assist retailers in meeting the needs of consumers in specific product classification.

TABLE 5 SUMMARY ANALYSIS OF CLV SEGMENTS BASED ON THE C DATASET

Segment	Nr. of customers	Best Cluster	Av. (cps)P	Av. Gender	Av. Age	Av. Nationality	Best Algorithm	Rank
Best	3221	2	-147.82	Female	35-44	UAE	EM	4
Spender	10066	4	-16.11	Female	25-34	Egypt	EM	3
Frequent	4771	2	-6.02	Female	25-34 - 35-44	Egypt - Lebanon	EM	2
Uncertain	24114	2	1.79	Female	35-44	Qatar	K-means	1

TABLE 6 STRATEGY RECOMMENDATIONS BASED ON ANALYSIS RESULTS

Segment		R	F	M	C	Recommended Strategy
Best		High	High	High	High	Recognizing the importance of these customers VIP communication VIP and Special Services Preferential discounts Inform timely about new products services
Spender		Low	Low	High	Positive	Communication. Discounts and price matching Inform timely about new products services
Frequent		Low	High	Low	Fluctuate Positive/Negative	Communication. Discounts and price matching Inform timely about new products services.
Uncertain	**Segment 1**	High	Low	Low	Negative	Get rid of them
	Segment 2	High	Low	Low	Positive	Frequently, promotion plans Cross-selling Special discount Providing online shopping POS.

Our extensive review of the literature shows that there are no demographic characteristics that clearly determine the SMR industry profitable consumers, and also traditional statistical models like (RFM) failed to predict actual consumer's behavior. Thus, instead of using traditional statistical methods, as most previous studies do, data-driven exploratory methods ed for extracting consumers hidden knowledge from real POS data warehouse and verified by industry experts

Finally, enhancing the Model to explore more about customer's lifestyle and understanding the interests and hobbies of particular customer segment, more into customer's Price sensitivity and Brand loyalty is a recommended future research.

REFERENCES

[1] APS Meeting Abstracts (Vol. 1, p. 11002).

[2] Azevedo, A. (Ed.). (2014). Integration of Data Mining in Business Intelligence Systems. IGI Global.

[3] Bernard, H. R. (2011). Study methods in anthropology: Qualitative and quantitative approaches. Rowman Altamira.

[4] Birant, D. (2011). Data Mining Using RFM Analysis, INTECH Open Access Publisher.

[5] Broderick, A., & Pickton, D. (2005). Integrated marketing communications. Pearson Education UK.

[6] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). Discovering data mining: from concept to implementation. Prentice-Hall, Inc.

[7] Carter, N. M., Stearns, T. M., Reynolds, P. D., & Miller, B. A. (1994). New venture strategies: Theory development with an empirical base. Strategic Management Journal, 15(1), 21-41.

[8] Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19(3), 197-208.

[9] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to a big impact. MIS Quarterly, 1165-1188.

[10] Cleveland, M., Papadopoulos, N., & Laroche, M. (2011). Identity, demographics, and consumer behaviors: International market segmentation across product categories. International Marketing Review, 28(3), 244-266.

[11] Dasgupta, S., & Freund, Y. (2009). Random projection trees for vector quantization. IEEE Transactions on Information Theory, 55(7), 3229-3242.

[12] Dash, P., & Mishra, S. (2010). Developing RFM model for customer segmentation in retail industry. International Journal of Marketing & Human Resource Management (IJMHRM), 1(1), 58-69.

[13] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38.

[14] Dipanjan, D., Satish, G., & Goutam, C. (2011). Comparison of Probabilistic-D and k-Means Clustering in Segment Profiles for B2B Markets. SAS Global Forum.

[15] Drăghici, S. (2003). Data analysis tools for DNA microarrays. CRC Press.

[16] Dwyer, F. R. (1997). Customer lifetime valuation to support marketing decision making. Journal of interactive marketing, 11(4), 6-13.

[17] Elby, A. (2015, April). The new AP Physics exams: Integrating qualitative and quantitative reasoning. In APS Meeting Abstracts (Vol. 1, p. 11002).

[18] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[19] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

[20] Fraley, C., & Raftery, A. E. (2006). MCLUST version 3: an R package for normal mixture modeling and model-based clustering. WASHINGTON UNIV SEATTLE DEPT OF STATISTICS.

[21] Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. AI magazine, 13(3), 57.

[22] Gal, A. (2010). Competitiveness of small and medium sized enterprises-a possible analytical framework. HEJ: ECO-100115-A.

[23] Goyat, S. (2011). The basis of market segmentation: a critical review of literature. European Journal of Business and Management, 3(9), 45-54.

[24] Gunaseelan, D., & Uma, P. (2012). An improved frequent pattern algorithm for mining association rules. International Journal of Information and Communication Technology Study, 2(5).

[25] Hossein Javaheri, S., (2008), Response Modeling in Direct Marketing: a data mining based approach for target selection, Master's thesis, epubl.luth.se/1653-0187/2008/014/LTU-PB-EX-08014-SE.pdf.

[26] Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. Expert Systems with Applications, 37(7), 5259-5264.

[27] Kak, A. (2014). Expectation-Maximization Algorithm for Clustering Multidimensional Numerical Data. An RVL Tutorial Presentation at Purdue University.

[28] Kashwan, K. R. & C. Velu (2013). Customer Segmentation Using Clustering and Data Mining Techniques. International Journal of Computer Theory & Engineering 5(6): 856-861.

[29] Khajvand, M., & Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. Procedia Computer Science, 3, 1327-1332.

[30] 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.

[31] Kolyshkina, I., Nankani, E., Simoff, S., & Denize, S. (2010). Retail analytics in the context of "Segmentation, Targeting, Optimisation" of the operations of convenience store franchises. Anzmac.

[32] Kumar, N., Verma, V., & Saxena, V. (2013). Cluster Analysis in Data Mining using K-Means Method. International Journal of Computer Applications, 76(12), 11-14.

[33] Lefait, G., & Kechadi, T. (2010, February). Customer segmentation architecture based on clustering techniques. In Digital Society, 2010. ICDS'10. Fourth International Conference on (pp. 243-248). IEEE.

[34] Lewis, P., & Thomas, H. (1990). The linkage between strategy, strategic groups, and performance in the UK retail grocery industry. Strategic Management Journal, 11(5), 385-397.

[35] Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. Expert systems with applications, 39(12), 11303-11311.

[36] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.

[37] Marcus, C. (1998). A practical yet meaningful approach to customer segmentation. Journal of consumer marketing, 15(5), 494-504.

[38] McCarty JA, Hastak M (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. J. Bus. Res., 60: 656-662.

[39] Milgramm, A. (2011). Effective Customer Segmentation Strategies: New CCC Work in Progress.

[40] Nwankwo, S., & Gbadamosi, T. (Eds.). (2010). Entrepreneurship marketing: principles and practice of SME marketing. Routledge.

[41] Ramageri, B. M., & Desai, B. L. (2013). Role of data mining in retail sector. International Journal on Computer Science and Engineering, 5(1), 47.

[42] Safari, M. (2015). Customer Lifetime Value to managing marketing strategies in the financial services. International Letters of Social and Humanistic Sciences, 1(2), 164-173.

[43] Sumathi, S., & Sivanandam, S. N. (2006). Introduction to data mining and its applications (Vol. 29). Springer.

[44] Tan, P. N., Steinbach, M., & Kumar, V. (2006). Cluster analysis: basic concepts and algorithms. Introduction to data mining, 8, 487-568.

[45] Tayal, M. A., & Raghuwanshi, M. M. (2010). Review on various clustering methods for the image data. Journal of Emerging Trends in Computing and Information Sciences, 2, 34-38.

[46] Tsiptsis, K. K., & Chorianopoulos, A. (2011). Data mining techniques in CRM: inside customer segmentation. John Wiley & Sons.

[47] Tufféry, S. (2011). Data mining and statistics for decision making, John Wiley & Sons.

[48] Wang, H., Huo, D., Huang, J., Xu, Y., Yan, L., Sun, W., & Li, X. (2010, July). An approach for improving K-means algorithm on market segmentation. In System Science and Engineering (ICSSE), 2010 International Conference on (pp. 368-372). IEEE.

[49] Wedel, M., & Kamakura, W. A. (2012). Market segmentation: Conceptual and methodological foundations (Vol. 8). Springer Science & Business Media.

[50] Zaiane, O. R. (1999). "Principles of knowledge discovery in databases." CMPUT690, Department of Computing Science, University of Alberta.

[51] Zhao, Y. (2011). R and Data Mining: Examples and Case Studies.

[52] Ziafat, H., & Shakeri, M. (2014). Using Data Mining Techniques in Customer Segmentation. International Journal of engineering Study and Applications, 1(4), 70-79.