Thanh Ho et al.

Article

An extended RFM model for customer behaviour and demographic analysis in retail industry

Business Systems Research (BSR)

Provided in Cooperation with:
IRENET - Society for Advancing Innovation and Research in Economy, Zagreb

Suggested Citation: Thanh Ho et al. (2023) : An extended RFM model for customer behaviour and demographic analysis in retail industry, Business Systems Research (BSR), ISSN 1847-9375, Sciendo, Warsaw, Vol. 14, Iss. 1, pp. 26-53, https://doi.org/10.2478/bsrj-2023-0002

This Version is available at:
https://hdl.handle.net/10419/318810

sciendo

An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry

Thanh Ho, Suong Nguyen, Huong Nguyen, Ngoc Nguyen, Dac-Sang Man, Thao-Giang Le
University of Economics and Law, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

Background: Customer segmentation has become one of the most innovative ways which help businesses adopt appropriate marketing campaigns and reach targeted customers. The RFM model and machine learning combination have been widely applied in various areas. **Motivations:** With the rapid increase of transactional data, the RFM model can accurately segment customers and provide deeper insights into customers' purchasing behaviour. However, the traditional RFM model is limited to 3 variables, Recency, Frequency and Monetary, without revealing segments based on demographic features. Meanwhile, the contribution of demographic characteristics to marketing strategies is extremely important. **Methods/Approach:** The article proposed an extended RFMD model (D-Demographic) with a combination of behavioural and demographic variables. Customer segmentation can be performed effectively using the RFMD model, K-Means, and K-Prototype algorithms. **Results:** The extended model is applied to the retail dataset, and the experimental result shows 5 clusters with different features. The effectiveness of the new model is measured by the Adjusted Rand Index and Adjusted Mutual Information. Furthermore, we use Cohort analysis to analyse customer retention rates and recommend marketing strategies for each segment. **Conclusions:** According to the evaluation, the proposed RMFD model was deployed with stable results created by two clustering algorithms. Businesses can apply this model to deeply understand customer behaviour with their demographics and launch efficient campaigns.

Keywords: Customer segmentation, RFMD model, K-Means, One hot encoding, K-Prototypes, Cohort analysis, machine learning

JEL classification: C61; C63; C67
Paper type: Research article

Received: Dec 25, 2022
Accepted: Jul 23, 2023

Citation: Ho, T., Nguyen, S., Nguyen, H., Nguyen, N., Man, D-S, Le, T-G. (2023). An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry. Business Systems Research, 14(1), 26-53.
DOI: https://doi.org/10.2478/bsrj-2023-0002

Introduction

Customer data is a foundation for successful business strategies. Exploring data to find customer insights and supporting decision-making increases business interest. Instead of applying marketing strategies for all customers who interact with the business in the same way and collectively, clustering customers helps businesses to identify target customers (Dawane et al., 2021), from which they can understand the characteristics of each segment and devise appropriate business strategies. Therefore, applying clustering methods to identify potential customers is today's leading trend. Combining machine learning algorithms with user data is a perfect example of customer segmentation that can help businesses identify customer segments that are difficult to detect through intuition and manual data inspection (Kumar, 2023).

RFM has found extensive utilisation as a method for analysing customers in various contexts (Ha & Park, 1998). The model is a behavioural prediction model used to analyse and predict customer behaviour (Yeh et al., 2009). RFM combined with clustering algorithms is an effective method many businesses apply to search for optimal customer segments. RFM is also applied in many fields based on customer clusters. Many studies have proposed extended models, incorporating new variables or applying various analytical techniques to differentiate customer groups in more detail and from multiple angles. Some extended RFM models include the RFMV model (Variety of products) by (Moghaddam et al., 2017), incorporating variables on products that customers frequently purchase and return to or incorporating variables such as Category to segment according to the characteristics of transactions performed by customers in the RFMC model, the study by (Allegue et al., 2020). Several other studies have delved deeper into the application of techniques in the clustering process to achieve clear analysis: the AHP technique in the RFM scoring process in the study by Liu and Shih (2005a) and Liu and Shih (2005b) or the basic clustering methods in the RFM segmentation, such as K-means, Cheng and Chen (2009).

Although RFM combined with clustering algorithms is easy to understand and apply, according to the study by Wei et al. (2010), the RFM model can only help businesses group customers based on transaction variables without considering other variables. Therefore, businesses have requested a specific customer segmentation model or method to address this issue in detail. Research questions have also been raised to clarify the research direction. Businesses have not yet evaluated or analysed which customer groups they belong to, their demographic characteristics, and how these characteristics affect their buying behaviour. Is there a method or model to apply and solve this issue easily? From there, appropriate strategies can be developed for each customer segment. The objective of this article is also to answer these research questions.

This study proposes the RFMD model, which combines traditional RFM with demographic data to provide businesses with a more specific understanding of customer behaviour and demographic characteristics within each cluster. Additionally, the model is applied with two clustering techniques, K-means, and K-prototypes, to obtain results and compare the two clustering outcomes. The proposed RFMD model combined with machine learning clustering techniques aims to generate customer clusters with similar behaviour and demographic characteristics on mixed data sets. Furthermore, the study applies Cohort analysis techniques to gain deeper insights into the data and combine the clustering results to provide business recommendations. Evaluation of the clustering results using AMI and ARI indices is also applied to ensure objectivity in selecting the clustering method. With the traditional RFM model, it can be offered to understand customer behaviour and identify customer segments.

However, these factors only evaluate behavioural variables based on how customers interact with the business but cannot provide specific information about the characteristics of these customers, as argued earlier. Therefore, integrating demographic factors into customer segmentation provides a more comprehensive understanding of each customer group. The combination of behavioural analysis and demographic characteristics is necessary and helps to enrich customer segmentation, according to the study (Sarvari et al., 2016). However, in this study, RFM analysis is performed first, followed by demographic analysis to understand each cluster better, which may be time-consuming and costly. Therefore, this study proposes the integrated RFMD model to simultaneously segment all four input variables and evaluate and collect results. This model is tested using popular clustering methods with mixed data sets. The RFMD model is a new contribution to the field of science, improving existing models and applying them to many execution methods. Furthermore, the proposed RFM improvement has contributed a model for applying machine learning in science, helping to classify data and search for data groups with similar characteristics.

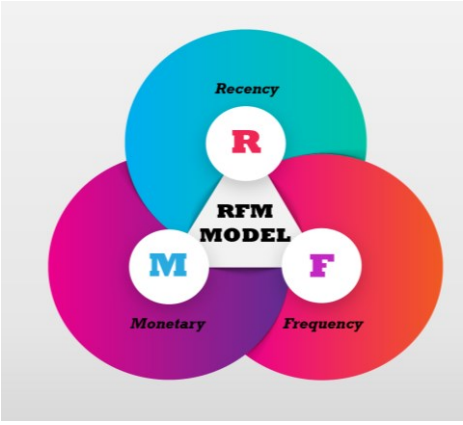
The research consists of six sections, with the introduction as the first part of the study. The next section reviews relevant research and theoretical foundations related to the research topic. Section 3 outlines the method proposed for constructing the process and model. The next section discusses the analysis, results collection, and content. In section 5, the achieved results are discussed in more detail, along with the proposed solutions. Finally, the conclusion summarises the research conducted, with limitations and implications.

Literature review

RFM models

In this section, various literature related to this paper has been reviewed, and research gaps also were discovered in articles about RFM models (Figure 1). Moreover, a series of extended RFM model studies is summarised as a diagram in Figure 2.

Figure 1
RFM Model Illustration



Source: Authors' summary

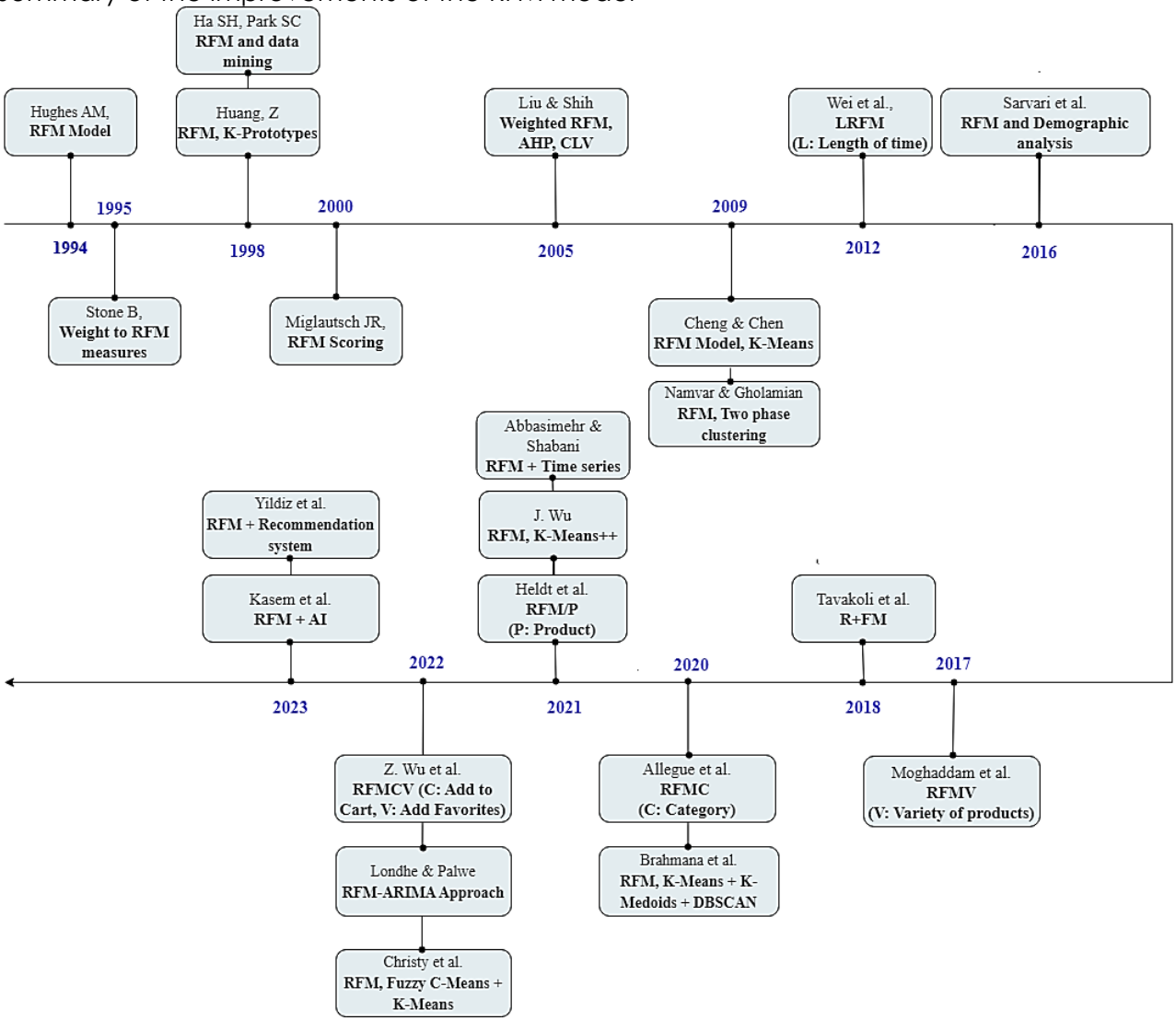
The RFM model is famous for dividing customers into segments based on analysing their past transactional data. This includes factors like the Recency of a customer's purchases, the Frequency of their buying activity, and the monetary value of their spending (Wei et al., 2010).

However, this traditional RFM model does not take advantage of additional consumer attributes and only considers transactional variables like Recency, Frequency, and Monetary (Hoegele et al., 2016). As a result, numerous research studies have been conducted to add additional variables to the traditional RFM model and apply machine learning to improve customer segmentation performance.

A summary of developments in consumer segmentation using enhanced RFM models between 1994 and 2021 was prepared. Figure 2 shows the evolution of the RFM model during the periods.

Figure 2

Summary of the improvements of the RFM model



Source: Authors' work

In the earliest period of the above timeline (from 1994 to 2000), the first concepts of the RFM model were introduced. Hughes first proposed the definition of the RFM model. In this initial RFM model (Hughes, 1994), three variables, R, F, and M, held equal significance in computing the overall score. For instance, if R, F, and M were assigned scores of 4, 5, and 3, respectively, the resulting combined score would be 12. However, Stone (1995) implied that product and industry features should be considered, and distinct weights would be assigned to each measure of RFM. In 1998, the RFM model

was first integrated with data mining technologies applied to datasets from enterprise data warehouses to boost convenience store sales in Ha and Park (1998). This time, two methods were also proposed in the research (Huang, 1998): extended K-Means for categorical variables and mixed data processing algorithms that integrate the K-Modes using the measurement method. The K-Prototypes algorithm, combining features from both the K-Means and K-Modes algorithms, not only handles categorical data but also removes restrictions when handling data from huge datasets. K-Means is only used with numerical data. RFM was also referred to as the customer quantile-based method in Miglautsch (2000), which arranges customers in decreasing order. The numbers of customers in each segment were in the same proportion.

From 2005 to 2012, many important improvements were proposed to the RFM model. In Liu and Shih (2005a) and Liu and Shih (2005b), the authors determined the weights of the three factors within the RFM model identified by the AHP technique instead of predetermining them. This approach enabled firms to evaluate customer lifetime value with the RFM model more precisely when applied in different industries. RFM model was also explored with another dimension with additional variables, for instance, variable L in the study of Wei et al. (2012). Variable L represents the duration from a patient's initial to final hospital admission. Moreover, the analysis of RFM models, which improved classification accuracy with the first use of K-Means in the study of Cheng and Chen (2009), enabled firms to successfully implement CRM tools when gaining knowledge of client segments' consumption patterns.

The third stage observed for summary, from 2015 to 2023, is when the RFM model developed with both clustering algorithms application and integration of other factors. The article of Moghaddam et al. (2017) considered the characteristic of the product category with additional variable V for the RFMV model, while the research by Heldt et al. (2021) took the product factor into account and suggested a product-specific RFM (RFM/P) model to predict the value of customer per product. Moreover, the article by Tavakoli et al. (2018) suggested using the R+FM model to help firms consider the relationship between customers and changes in their behaviour. The article by Allegue et al. (2020) also proposed an extended RFM model, which was RFMC with variable C as the feature for the Category of transactions made by the customer. Since variable C was categorical data, one hot encoding transforms the data into numeric values.

Additionally, the study by Sarvari et al. (2016) compared two clustering methods, Kohonen and K-Means, and three algorithms for generating association rules: FP-growth algorithm, Éclat algorithm, and A-priori under 42 scenarios with different RFM weights, segmentation factors, and input for Association Rule Mining. Analysis results demonstrate the importance of the demographics factor when merged with the RFM model. In 2022, for the dataset of the e-commerce platform, the extended RFM model with two variables, C and V, was proposed (Wu et al., 2022). Variable C represents the Frequency of customers adding items to their shopping carts, and variable V refers to the Frequency of customers adding items to their favourites list. Also, this year, the RFM model was combined with the ARIMA model in Londhe and Palwe (2022) to segment customers and predict sales revenue, referred to as a hybrid multi-step model. The results of these methods showed that RFM-ARIMA achieved better accuracy and was presented as a better approach to analysing customer behaviour and making precise sales revenue predictions. RFM model analysis also improved clustering efficiency with many clustering algorithms. K-Means, K-Medoids, and DBSCAN clustering algorithms—were applied in the article of Brahmana et al. (2020) using the customer transactional dataset. In addition, K-Means and Fuzzy C-Means were used, and the results were evaluated based on the comparison between their iteration number, cluster cohesion, and execution time of the clustering process (Christy et al., 2021). Besides, Wu et al.

(2021) precisely calculate customer value with an enhanced RFM model and employ K-Means++ for user segmentation on E-commerce platforms. Moreover, various techniques were combined with RFM; for example, Abbasimehr and Shabani (2021) proposed customer behaviour forecasting at the segment level with RFM and represented as a time series; Yıldız et al. (2023) presented the integration of the recommendation system with a rule-based heuristic algorithm and customer segmentation whose parameters for the clustering phase were RFM value and demographic data; RFM analysis and boosting tree were also combined in (Kasem et al., 2023) to create the customer profiling system and predict product sales.

Research gap

The reviewed customer segmentation using RFM analysis studies in general use factors related to product or transaction characteristics but have not taken advantage of many customer demographics such as age, gender, and Region. Although demographics have been considered in some studies, its data is not a main variable in customer clustering. Therefore, the demographic variable cannot demonstrate the relationship or impact on the R, F, and M variables of the RFM model in the process of clustering or customer segmentation. For example, Namvar et al. (2010) built a new customer segmentation model with two distinct clustering phases: first, customers were clustered with K-Means based on R, F, and M variables, and then with demographics data, these clusters would be partitioned into new clusters. In addition, in the study of Sarvari et al. (2016), with the transaction dataset from the global pizza chain, customer segmentation was analysed utilising the RFM model and demographic information by applying K-Means clustering and association rule mining. The study has designed 42 scenario types with various inputs for clustering processes, including the value of each RFM variable, scores for each RFM variable, and the average RFM score to segment customers more precisely. Three RFM variables can be assigned weights or combined with demographic data. Categorical data such as Age, Region, and Gender were also converted into numeric data for the K-Means clustering process by being assigned ordinal numbers. The study has carefully evaluated and compared the efficiency of scenarios by applying Neutral Networks. The proposed research method was considered to have achieved the best results. Moreover, with the limited amount of demographic data, the study shows that demographic characteristics play an undeniably important role when combined with RFM data. However, in the processing process, because of applying many clustering methods and rule extraction techniques, the study also needs to use more machine learning methods to improve the speed, time, and ability to extract rules more comprehensively of the data mining tools.

To help businesses gain a deeper understanding of both customer behaviour and demographic characteristics by customer segments, the article proposed the RFMD model with a combination of RFM and demographic factors. The new point, as well as the way to fill in the research gap in this study, is to include all the variables of the RFMD model for the input data of the clustering process with the application of the K-Means algorithm combined with One hot encoding method and K-Prototypes algorithm. Therefore, the relationship or impact of four variables, R, F, M, and D, on customer segmentation is demonstrated. Besides, two clustering algorithms also help the clustering process on a mixed dataset of numeric and categorical data to become more easily and efficiently. In addition, from the works of Covoes et al. (2013) and Romano et al. (2014), this study implemented the ARI and AMI to evaluate the similarity between the two clustering results. AMI and ARI, in general, are quite stable and often used to evaluate clustering results. Cohort analysis is also applied in the study to help

marketers provide valuable insights into how different segments of customers behave over time.

Background

After finding the research gap, the theoretical definitions relevant to building background knowledge and concepts for the research are presented.

Customer segmentation

Customer segmentation was first introduced by Smith (1956), which is an act of dividing into groups. Customers in a group have similar behaviour, characteristics, and needs (McDonald, 2012). Customer segmentation builds customer profiles, the foundation for a customer-centric information system (Bose & Chen, 2015).

The general customers of the market and the business have many different characteristics. In customer segmentation, general factors and characteristics related to each product make up the two main categories of customer values (Wedel & Kamakura, 2000). Customer demographics such as age, gender, and geography are included in general variables. Contrarily, product-specific variables are related to transactional information and consumer behavior, including shopping habits, customer lifetime value (LTV), and spending. On the other hand, product-specific variables pertain to transactional details and customer behavior, such as purchase patterns, customer LTV, and expenditure. Product-specific variables are very effective as the purchase behaviour can be summarised using a very small number of variables, like the RFM model.

Consumer behaviour and experience

Brands must understand how consumers behave and think when making decisions, particularly now, when customers are more aware and able to access information more rapidly online (Jacoby, 1975; Kicova et al., 2018). Many factors influence behaviour, such as demographics, religion, and geographical location (Gajjar, 2013). Research on consumer behaviour is important because it helps marketers understand the impact on customers' purchasing decisions.

Customer experience is the consumer's cognitive and emotional evaluation of their direct and indirect interactions with the business about their purchase behavior.

Clustering

Clustering is a crucial and popular tool for client segmentation (Chiu & Tavella, 2008). Grouping objects into sets of related objects is known as clustering (Al-Augby et al., 2015). Clustering is divided into two popular categories: Hierarchical and Partitional clustering. Hierarchical clustering generates a cluster tree (or dendrogram) using heuristic splitting or merging techniques, which does not need to specify a specific number of clusters. Agglomerative algorithms are well-known ones that create the cluster tree through merging. Each pattern is initially given to a single cluster using a dividing hierarchical technique. Then, until all data points are in the same cluster, splitting is done to a cluster in each stage. Partitional clustering is a method of dividing data into a specific number of clusters. This method is more efficient than hierarchical algorithms (Omran et al., 2007).

K-Means

The K-Means algorithm is one of the machine learning methods of Partitional clustering. K-Means has been popular and in the top 10 most used in data mining and knowledge discovery since 2000 (Wu et al., 2008). Several academics from other fields have

discovered this algorithm, most notably Lloyd (1982), Forgey (1965) and Friedman and Rubin (1967). Indeed, the K-Means algorithm minimises the distance from the values in the cluster to the centroids based on the theory of vector distances in Euclidean space.

$$d(x, C_i) = \sqrt{\sum_{j=1}^{m} (x_j - C_{ij})} \quad (1)$$

K-Means and One hot encoding

For the K-Means algorithm to work with categorical data, we have to encode that data into a numerical format. The method we use is One hot encoding.

"One hot encoding is the most common approach to converting categorical features to a suitable format for input to a machine-learning model" (Seger, 2018). It is an encoding method that expands an initial vector into a multidimensional matrix. Each dimension of the matrix is the number of states in this feature, and each dimension represents a specific state. All other state dimensions are zero due to this processing, and just one feature matrix dimension is typically asserted for each state (Yu et al., 2022). Table 1 presents part of the encoding data.

Table 1

Encoding data

Regular representation	One-hot encoding
0	1
1	10
2	100
3	1000
4	10000
5	100000

Source: Authors' work

K-Prototypes

The K-Prototype algorithm combines K-means and K-Modes (Lakshmi et al., 2018). K-Prototypes is a powerful approach for clustering datasets of different types. The results of Cheng and Chen (2009) also show that the K-Prototypes algorithm is effective when clustering on large and complex datasets in the number of data lines and clusters.

$$d(x_{il}(t), c_j) = \sum_{r=1}^{m_r} \sqrt{(x_{ir}(t) - c_{jr})^2} + \gamma_j \sum_{t=1}^{m_t} \delta(x_{is}(t), c_{js}) \quad (2)$$

Extended RFM model (RFMD)

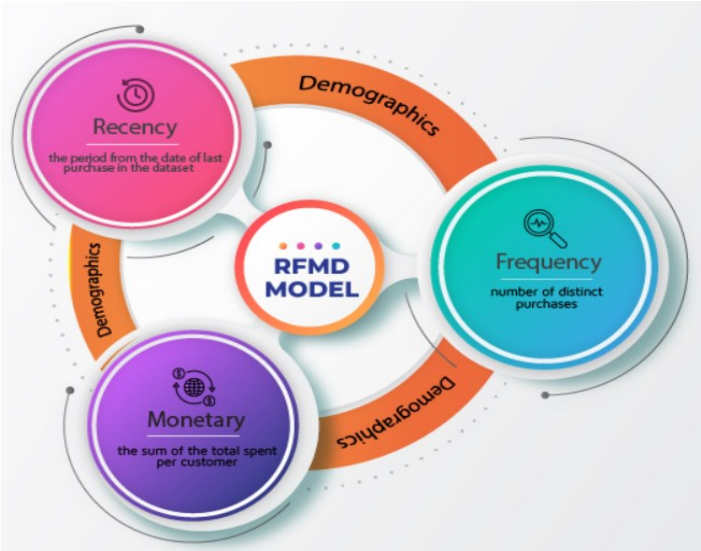
A new model based on the research gap and conceptual background is proposed. The mathematical formula is also demonstrated below.

The RFM model has limitations in accommodating only a limited number of selection factors (Wei et al., 2010). This model does not exploit the influence of customer demographic features. However, customers' demographic values impact the formulation of business and marketing strategies. "Demographics play a specific role in marketing; clearer concept/measure relationships, better techniques, and expanded applications can improve the quality of work seen" (Pol, 1991).

With the affirmation of the importance of Demographic in the customer segment from previous reputable studies, as stated above, this study was conducted to incorporate demographic variables into the RFM model, thereby effectively optimising customer segmentation. Our RFMD model is built from transaction data of business, with variables R(Recency), F(Frequency), M(Monetary), and D(demographic). Figure 3 depicts our RFMD model. Variable D includes categorical variables corresponding to

the customer's demographic attributes. The selected attributes differ depending on the customer's business purpose and strategy. In this model, we assume that the numeric variables in the model R, F, and M have the same impact on the model, and the results are the same.

Figure 3
RFMD model



Source: Authors' work

To evaluate which segment a customer is placed in, we built formula (3) with the same rating assigned to a segment.

RFMDscore = Rscore + Fscore + Mscore + Dscore (3)

The scores of each variable R, F, M, and D are from 1 to 5 depending on the criteria of each different enterprise. In this study, we used 2 machine learning methods, K-Means and K-Prototypes, to cluster the RFMD data model to find customer groups with the most similarity. K-Means and K-Prototype use Euclidean or squared Euclidean distance on the numeric attributes to measure the distortion between data objects and centroid K. (Prabha & Visalakshi, 2014). Assuming the data set is U = {x_1,x_2,x_3,...x_N}, let C be the centre vector of 1 cluster, and N is the number of rows of data. Formulas (4) and (5) demonstrate the Euclidean distance calculation between data points. Example in case of RFM model: there is 1 data centre point c(1,2,3) and one data point x(2,3,4) the distance between two points is equal to $d_{(c,x)} = \sqrt{(1-2)^2 + (2-3)^2 + (3-4)^2} \approx 1.73$. The same calculation method is applied for the RFMD model as formula (4). These 2 formulas show the difference between using RFM and RFMD data models. Adding a customer Demographic attribute variable directly affects the clustering results of machine learning, thereby changing the results in which customers are classified. The distance from a point to cluster entered in the case of the RFM model:

$$d(c;x) = \sqrt{(C_R - Xi_R)^2 + (C_F - Xi_F)^2 + (C_M - Xi_M)^2} \quad (4)$$

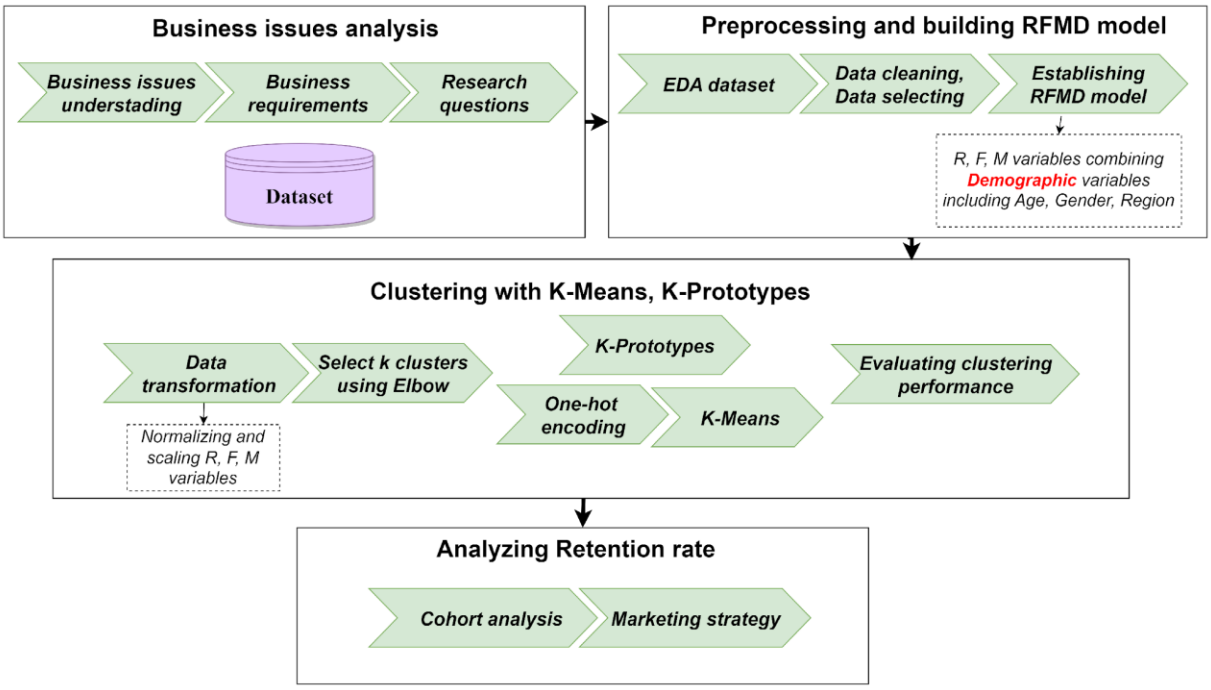
The distance from a vector x to the centre of the cluster of the RFMD model:

$$d(c;x) = \sqrt{(C_R - Xi_R)^2 + (C_F - Xi_F)^2 + (C_M - Xi_M)^2 + (C_D - Xi_D)^2} \quad (5)$$

Methodology

This section elaborates on the experimental framework and explains all the steps. As shown in Figure 4, the research process involves four main stages.

Figure 4
Proposed Methodology



Source: Authors' work

Proposed stages:

(1) At this stage, it is necessary to understand the business problems of the limitations when using the traditional RFM model, thereby setting out the business requirements related to customer segmentation that need to be solved. Next are proposing the questions that need to be answered.

(2) Pre-processing and building RFMD. At this step, the researcher performs data mining and selects suitable features to build the proposed data model. Data pre-processing is then conducted to make the input data compatible with the proposed machine learning methods. This is a very important stage for machine learning to be highly effective. Finally, the proposed RFMD data model is finalised.

(3) Clustering with K-Means, K-Prototypes. At this stage, K-means and K-prototype clustering techniques are used to identify customer groups whose demographic characteristics affect different purchasing behaviour. Firstly, a method is tested and chosen for normalising and scaling the transactional variables to ensure the model is efficient and stable. The Elbow method selects the appropriate number of clusters for the data model. After clustering using two machine learning methods, the indexes ARI and AMI compare the similarity and agreement of the two clustering results.

(4) Analysing Retention Rate. At this final stage, Cohort analysis is performed to dig deeper into the built data model. Besides, the researcher recommends several marketing campaigns based on the results obtained.

Business issues understanding

After summarising domestic and foreign research on customer segmentation and specifically the application of the RFM model to the customer segment, some limitations

of the RFM model have been mentioned. Customer data is increasing, not only focusing on common transactional data such as purchase date, purchase quantity, and total amount but also expanding in terms of customer demographic data such as age, and gender. Therefore, businesses must use this data source to make appropriate decisions.

In addition, the traditional RFM model cannot segment customers and evaluate them comprehensively because it only focuses on customers' shopping behaviour while ignoring other customer characteristics such as demographics. There is a requirement for a new data model that helps businesses acknowledge both the behavioural and demographic characteristics of customers. There are also indispensable clustering machine learning methods for this new data model. From there, businesses can use demographic data well, have a more comprehensive view of customers, and promote accurate marketing and customer service strategies. The current requirements are not only to answer the question of which customer group they belong to but also the question of who they are and their demographic characteristics.

Data preparation and customer segmentation

A dataset of Online Sales in the USA is about the sales of different products, merchandise, and electronics in different states. It contains 286,392 transactions of 64,248 customers during the period between October 2020 and September 2021 with four basic statuses, including Cancelled orders (112,166 transactions), Returned orders (25,713 transactions), Received orders (51,775 transactions), Completed orders (88,968 transactions). There are no missing or duplicate values. Part of the data is presented in Table 2.

Table 2
A piece of original input data

Code	Order_date	Status	Qty_ordered	Price	Cust_id	Gender	Age	Region	Zip	...	Item_id
100354678	10/1/2020	Received	21	89.9	60124	F	43	South	73571	...	574772
100354678	10/1/2020	received	11	19.0	60124	F	43	South	73571	...	574774
100354680	10/1/2020	complete	9	149.9	60124	F	43	South	73571	...	574777
100354680	10/1/2020	complete	9	79.9	60124	F	43	South	73571	...	574779
100367357	11/13/2020	received	2	99.9	60124	F	43	South	73571	...	595185
100367357	11/13/2020	received	2	39.9	60124	F	43	South	73571	...	595186

Source: Authors' work

Firstly, the transactions are removed with Cancelled and Returned status as they do not create value for the company. Following that, the order_id is grouped to create a new dataset where each row represents a distinct order_id. Therefore, the final dataset records 93,873 orders from 42,589 customers. Attributes are selected based on RFMD model features such as order_id (a transaction has only one order_id), order_date (time to execute that transaction), qty_ordered (the number of products in that order), price (price per product), cust_id (Each customer is provided with a unique cust_id), Gender, Age (customer age), Region (The Region where the customer lives with 4 unique values is South, Northeast, West, Midwest). The new dataset is created by grouping each cust_id:

- **Recency:** calculated by the number of days since the newest order_date to the reference day;
- **Frequency:** counts total order_id of each customer;
- **Monetary:** is calculated by adding a column name 'payment' using formula (qty_ordered times to price) as a cell of that column and sum all payments of each customer;

- **Age:** Integer;
- **Region:** Coded;
- **Gender:** Male or Female.

After obtaining the RFMD data model, the customers with a Monetary transaction value of 0 are removed from the data set because these customers do not bring value to the company. The final data set includes 42,492 transactions for each specific customer. Table 3 presents a part of the data of the RFMD model.

Table 3
A part of the data of the RFMD model

Customer ID	Recency	Frequency	Monetary	Gender	Region	Age
115322	1	1	209.6	F	Northeast	56
115323	1	1	8,839.8	M	Northeast	51
115324	1	1	79.8	M	South	52
115325	1	2	179.8	F	South	38
115326	1	1	7,119.8	M	South	28

Source: Authors' work

Clustering with K-Means and K-Prototypes

After completion of the RFMD model, clustering using two machine learning methods, K-Means and K-Prototypes, is initiated. However, before that, normalisation and scaling of the transactional data are selected and performed.

Data transformation

The data normalisation procedure is commonly applied in the data transformation phase during the clustering process of machine learning algorithms. The value of transactional data, specifically Recency, Frequency, and Monetary, is studied and analysed first.

As demonstrated in Table 4, the quartiles of the three indices R, F, and M are considered, revealing Recency's minimum and maximum values as 1 and 365.

Table 4
Descriptive statistics of the Recency, Frequency and Monetary variables

Statistics	Recency	Frequency	Monetary
Count	42,492.00	42,492.00	42492.00
Mean	196.90	2.21	4332.18
Standard Deviation	89.32	5.29	12743.78
Min	1.00	1.00	0.20
25%	136.00	1.00	239.60
50%	192.00	1.00	643.95
75%	279.00	2.00	3200.00
Max	365.00	770.00	582665.40

Source: Authors' work

Some customers have purchased recently, and some have not in 1 year. The mean is larger than the median. The Frequency value ranges from 1 to 770. The Frequency is much distributed at the value of 1 purchase. Especially the mean is higher than the 3rd percentile (75%) due to the large outlier's influence. The data distribution tends to be much left skewed. The Monetary value is distributed from 0.2 to 582,665. Like Frequency, outliers heavily influence monetary value, and the data distribution tends to be much left-skewed. As can be seen, the value domain of R, F, and M has a clear difference,

especially the domain of Monetary is much larger than the two values of Recency and Frequency.

Before scaling the data, the data must be asymptotic to a normal distribution. At this step, the distribution of three transactional data variables is tested: Recency, Frequency, and Monetary. Then, Box-Cox transformation is used as Box-Cox can constitute a best practice in data transformation (Osborne, 2010). The Box-cox transformation method is commonly used to stabilise variance (remove changed variances) and transform non-normal dependent variables into normal shapes.

$$y_i^{(\lambda)} = \{\frac{y_i^{\lambda}-1}{\lambda}, if\lambda \neq 0 \; ln(y_i), if\lambda \neq 0 \quad (6)$$

Table 5
Result of Skewness before and after using Box-Cox transformation

Indicator	Recency	Frequency	Monetary
Original distribution	-0.35	76.10	11.9
Box-Cox transformation	-0.33	0.71	0.04

Source: Authors' work

Based on the result in Table 5, normalising by the box-cox method for all 3 variables gives a good result close to normal distribution. The Frequency and Monetary variables with the original data are very right-skewed. Besides, the R, F, and M scales differ, as they are measured in days, occurrences, and monetary units, respectively. Therefore, the study uses the standard method (StandardScaler) after the data is normally distributed. The Recency, Frequency, and Monetary variables are scaled to a domain of values suitable for clustering algorithms (Table 6).

Standardisation:

$$z = \frac{x-\mu}{\sigma} \quad (7)$$

with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \quad (x_i) \quad (8)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \quad (x_i - \mu)^2} \quad (9)$$

Table 6
Part of RFMD data with normalised and scaled transaction variables

CustomerID	Gender	Region	Age	Recency	Frequency	Monetary
115322	F	Northeast	56	-2.16	-0.73	-0.86
115323	M	Northeast	51	-2.16	-0.73	1.29
115324	M	South	52	-2.16	-0.73	-1.59
115325	F	South	38	-2.16	0.99	-0.97
115326	M	South	28	-2.16	-0.73	1.19

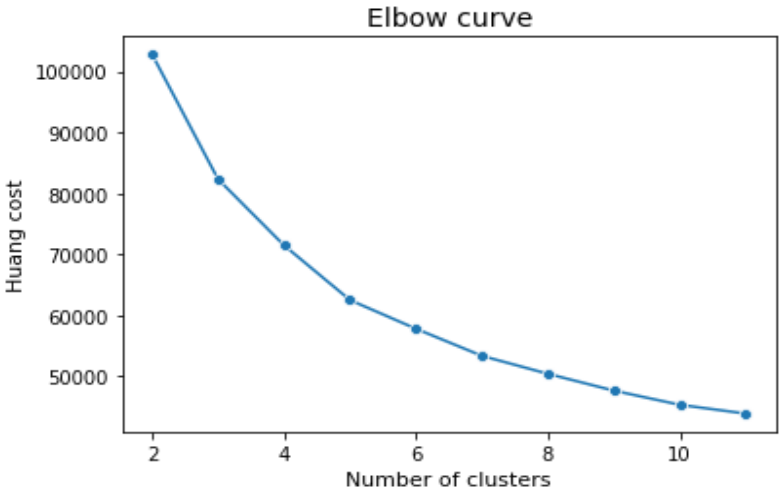
Source: Authors' work

Selecting the number of clusters using the Elbow method

The Elbow method is one of the popular methods to select k clusters for both. The Elbow method is illustrated as a curve graph with the horizontal axis being the number of K clusters (the count of customer segments using the values extracted from the RFMD data model) and the vertical axis being the Huang Cost Function. An index that measures the difference between points in the cluster. "Assume sr is the dissimilarity measure on numeric attributes defined by the squared Euclidean distance and sc is the

dissimilarity measure on categorical attributes defined as the number of mismatches of categories between two objects. The Huang Cost Function defines the dissimilarity measure between sr + γsc, where γ is a weight to balance the two parts to avoid favouring either type of attribute. Because these algorithms use the same clustering process as k-means, they preserve the efficiency of the k-means algorithm, which is highly desirable for data mining" (Cheng & Chen, 2009).

Figure 5
Result of Elbow curve



Source: Authors' work

As shown in Figure 5, the elbow with K = 5 is chosen as the appropriate number of clusters. It is observed that, from cluster number 6 onwards, the value of the cost function is almost uniformly decreasing, or in other words, the Huang Cost is almost linear, indicating the homogeneity between the points in the cluster.

Clustering with K-Means and K-Prototypes algorithms

The clustering results are obtained for K-Prototypes machine learning using the scaled RFMD dataset and the number of previously selected clusters as input (Table 7).

Table 7
Clustering result of K-Prototypes machine learning

Cluster	Number of customers	Mean R	Mean F	Mean M	F	M	Mid-west	North-east	South	West	Age 18-24	Age 25-64
0	7,902	109.0	5.4	12,289.1	3,555	4,352	2,169	1,364	2,888	1,481	1,093	5,474
1	10,463	118.7	1.0	464.9	5,691	4,772	2,890	1,801	3,888	1,884	1,461	7,176
2	7,163	231.7	1.0	3,589.7	3,814	3,349	1,894	1,278	2,686	1,305	996	4,937
3	10,053	277.7	1.0	258.5	4,294	5,759	2,703	1,844	3,707	1,799	1,345	7,037
4	6,911	262.3	3.4	7,784.4	3,853	3,058	1,885	1,232	2,586	1,208	956	4,773

Source: Authors' work

For the special K-Means machine learning that only works with numerical data, One hot encoding method is chosen for all three variables: Age, Gender, and Region. As the K-Means algorithm cannot interpret categorical variables, demographic variables must be encoded with numbers for the algorithm to understand and interpret. One hot encoding technique is chosen because all three variables, Age, Gender, and Region, are nominal categorical variables, and this method creates new (n-1) variables for n values of a categorical value. It uses only 0 or 1 while encoding a variable. The label

coding approach, inappropriate for these three variables, cannot be used for ordinal categorical variables.

Table 8
Descriptive statistics of the Age variable

Statistics	Age
Count	42,492.00
Mean	46.40
Standard Deviation	16.73
Min	18.00
25%	32.00
50%	46.00
75%	61.00
Max	75.00

Source: Authors' work

Table 8 displays the descriptive statistics for the Age variable, encompassing an age range from 18 to 75 within the datasets. Customer ages are categorised into three groups: Youth (18-24), Adults (25-64), and Elderly (over 65) (Ritchie & Roser, 2019), corresponding to the three columns that have been designated: 'Age 18-24', 'Age 25-64', 'Age 65+'. Table 9 displays a dataset segment following the utilisation of One-hot encoding.

Table 9
A part of the dataset after using One-hot encoding

Customer ID	Recency	Frequency	Monetary	Gender F	Gender M	Age 18-24	Age 25-64	Age 65+	Region Midwest	Region Northeast	Region South	Region West
4	-2.155221	1.935947	2.066071	0	1	0	1	0	1	0	0	0
15	0.389741	1.344794	-0.391672	1	0	0	0	1	1	0	0	0
20	-2.155221	1.436696	2.026171	0	1	0	1	0	0	0	1	0
21	0.479979	-0.977193	-1.234476	0	1	0	1	0	0	0	1	0

Source: Authors' work

After using One hot encoding, the Gender variable is divided into two columns, including 'Gender F' as a Female column and 'Gender M' as a Male; the variable Region is divided into 4 columns corresponding to 4 geographic regions of the United States: 'Region Midwest', 'Region Northeast, 'Region South, 'Region West; especially with the Age variable. The clustering results of the K-Means algorithm are presented in Table 10.

Table 10
Clustering result of the K-Means algorithm

Cluster	Number of customers	Mean R	Mean F	Mean M	F	M	Midwest	Northeast	South	West	Age 18-24	Age 25-64
0	9,629	282.58	1.0	262.5	4,722	4,907	2,593	1,779	3,527	1,730	1,098	6,811
1	6,591	266.86	3.3	7,242.8	3,308	3,283	1,812	1,161	2,479	1,139	795	4,532
2	8,248	112.19	5.3	12,481.3	4,112	4,136	2,250	1,438	3,006	1,554	999	5,740
3	10,928	122.03	1.0	445.2	5,604	5,324	3,024	1,863	4,066	1,975	1,320	7,492
4	7,096	229.43	1.0	3,665.1	3,456	3,640	1,862	1,278	2,677	1,279	893	4,862

Source: Authors' work

Evaluation of clustering results

After obtaining two clustering results with the same number of clusters, we use the Adjusted Rand Index (ARI) and Adjusted Mutual Information Index (MI) to compare the similarity and agreement of the two results. Given the knowledge of the method's clustering algorithm assignments between the two-clustering methods K-Means and K-

Prototypes of the same sample, the ARI is a function that measures the similarity. The AMI is a function that quantifies the concordance between two assignments, while disregarding permutations and incorporating chance normalisation (Vinh et al., 2009).

Adjusted Mutual Information:

$$AMI(U,V) = \frac{MI(U,V) - E\{MI(U,V)\}}{\{H(U),H(V)\} - E\{MI(U,V)\}} \quad (10)$$

Adjusted Rand Index:

$$ARI = \frac{\Sigma_{ij} \left(\frac{n_{ij}}{2}\right) - \frac{[\Sigma_i \left(\frac{a_i}{2}\right) \Sigma_j \left(\frac{b_j}{2}\right)]}{\left(\frac{n_{ij}}{2}\right)}}{\frac{1}{2}\left[\Sigma_i \left(\frac{a_i}{2}\right) + \Sigma_j \left(\frac{b_j}{2}\right)\right] - \frac{[\Sigma_i \left(\frac{a_i}{2}\right) \Sigma_j \left(\frac{b_j}{2}\right)]}{\left(\frac{n_{ij}}{2}\right)}} \quad (11)$$

where n_{ij}, a_i, b_i are values from the contingency table.

Table 11
Result of ARI and AMI Index

Adjusted Rand Index (ARI)	Adjusted Mutual Information (AMI)
0.880	0.864

Source: Authors' work

The results presented in Table 11 show that both AMI and ARI > 86%, i.e., the results of the two clusters have high similarity. In other words, our RFMD data model can be tested on both K-Means and K-Prototypes clustering methods.

Results

Cluster description and evaluation

Customers with the same characteristics were categorised into 5 clusters.

Cluster 0: Loyal Customers. This cluster includes 7,902 customers with the most recent purchase (109), the highest Frequency (5), and the first group of customers who spend the most money in the 5 clusters to shop (12,289). According to the Pareto principle, this cluster plays the most important role in contributing revenue to the business. Therefore, it is labelled as a *Loyal Customer* based on purchasing behaviour that shows the great engagement of customers in a business. Customers in this cluster are evenly distributed among age groups. However, the most valuable customer group belongs to women aged 18-24 in the Northeast, age 25-65 in the Southern Region, age 18-24 in the West region, and males aged 25-65 in the Midwest, age 65+ in the Northeast, age 25-65 in the Southern Region, and age 65+ in the West. To increase customer loyalty, businesses must regularly receive feedback and provide better customer experiences. Moreover, businesses can choose up-selling campaigns with higher-value products and services or cross-selling with other accompanying products to increase Monetary value. Table 12 presents the descriptive statistics relevant to cluster 0.

Table 12
Descriptive statistics of variables Recency, Frequency and Monetary and Demographics variables of cluster 0

Variable	Count	Mean	Standard Deviation	Min	25%	50%	75%	Max
Recency	7,902	109.05	55.22	1	68	112	155	203
Frequency	7,902	5.38	11.37	1	2	3	5	770
Monetary	7,902	12,289.12	25,005.52	135.8	1,089.58	4,101.75	12,119.4	582,665.4
Age 18-24	1,093	.	.	.	.	.	.	.
Age 25-65	5,474	.	.	.	.	.	.	.
Age 65+	1,335	.	.	.	.	.	.	.
F	3,550	.	.	.	.	.	.	.
M	4,352	.	.	.	.	.	.	.
Midwest	2,169	.	.	.	.	.	.	.
Northeast	1,364	.	.	.	.	.	.	.
South	2,888	.	.	.	.	.	.	.
West	1,481	.	.	.	.	.	.	.

Source: Authors' work

Cluster 1: New Customers. This cluster includes 10,463 customers who recently purchased with Recency (119) but did not make regular purchases with Frequency (1) and spent only a small amount of money, Monetary equal to (465). This group of customers revealed the features of buying products recently and showed business interest. However, they only purchased with small amounts of money because they were unfamiliar with the business. This cluster is labeled as *New Customers*. The potential customers in this cluster belong to the female customer group, especially those aged 25-65 in the Northeast region. In addition, male customers of all ages in the South region contribute significantly to revenue, and male customers aged 65+ in the Northeast are also prominent customers. Businesses must stimulate new customers to shop with preferential policies and try to convert them into loyal customers with personalised care services. Besides, businesses should help customers experience and understand the value of products and services to use them without problems later and build friendly relationships with customers. Table 13 presents the descriptive statistics relevant to cluster 1.

Table 13
Descriptive statistics of variables Recency, Frequency and Monetary and Demographics variables of cluster 1

Variable	Count (number of customers)	Mean	Standard Deviation	Min	25%	50%	75%	Max
Recency	10,463	118.67	55.1	1	77	136	156	213
Frequency	10,463	1.01	0.11	1	1	1	1	2
Monetary	10,463	464.9	716.95	0.2	158.8	260	479.1	11,263
Age 18-24	1,461	.	.	.	.	.	.	.
Age 25-65	7,176	.	.	.	.	.	.	.
Age 65+	1,826	.	.	.	.	.	.	.
F	5,691	.	.	.	.	.	.	.
M	4,772	.	.	.	.	.	.	.
Midwest	2,890	.	.	.	.	.	.	.
Northeast	1,801	.	.	.	.	.	.	.
South	3,888	.	.	.	.	.	.	.
West	1,884	.	.	.	.	.	.	.

Source: Authors' work

Cluster 2: Needing Attention Customers. This cluster includes 7,163 customers who have not returned to the shopping business for a long time, with Recency (232), spending with small Frequency (1) but with a relatively high amount of (3,590). These customers used to contribute a lot to business but somehow have not come back to business. Therefore, they are labelled as *Needing Attention Customers* because businesses must focus on them to win them back. Male customers aged 18-24 and 25-65 in the Northeast with high Monetary value tend to leave the business in this cluster.

Additionally, male and female customers in the Northeast have not shopped for a long time. Businesses should review their marketing campaigns in the Northeast region and try to re-engage their interest with coupons and personalised emails. Table 14 presents the descriptive statistics relevant to the cluster.

Table 14
Descriptive statistics of variables Recency, Frequency and Monetary and Demographics variables of cluster 2

Variable	Count	Mean	Standard Deviation	Min	25%	50%	75%	Max
Recency	7,163	231.65	64,34	33	174	260	281	365
Frequency	7,163	1	0	1	1	1	1	1
Monetary	7,163	3,589.7	3,686.72	399.8	1,258.65	2,400.0	4,385.5	49,996.5
Age 18-24	996	.	.	.	.	.	.	.
Age 25-65	4,937	.	.	.	.	.	.	.
Age 65+	1,230	.	.	.	.	.	.	.
F	3,814	.	.	.	.	.	.	.
M	3,349	.	.	.	.	.	.	.
Midwest	1,894	.	.	.	.	.	.	.
Northeast	1,278	.	.	.	.	.	.	.
South	2,686	.	.	.	.	.	.	.
West	1,318	.	.	.	.	.	.	.

Source: Authors' work

Cluster 3: At-risk Customers. These 10,053 customers made their last purchase a long time ago, with Recency (277), low Frequency (1), and minimal transaction amount of (259). This is a customer cluster that shows signs of leaving the business. However, the number of customers in this cluster is very large, so if the business does not retain them, it will lose significant revenue and receive negative feedback. With the importance of this cluster and its churn behaviour, this cluster is labelled as *At Risk Customers* because if the business does not respond to this cluster immediately, they can all be churned out soon. These customers are concentrated in the Midwest and Southern regions, aged 65+.

Additionally, male customers aged 18-24 in the Southern Region also need attention. Businesses need to focus more on marketing campaigns for customers aged 65+. Besides, businesses must provide many useful information sources for this customer cluster to retain them. Table 15 presents the descriptive statistics relevant to cluster 3.

Table 15
Descriptive statistics of variables Recency, Frequency and Monetary and Demographics variables of cluster 3

Variable	Count	Mean	Standard Deviation	Min	25%	50%	75%	Max
Recency	10,053	277,66	38,23	174	266	280	288	365
Frequency	10,053	1,01	0,12	1	1	1	1	4
Monetary	10,053	258,54	182,27	0,4	139,6	211,8	330	1.658,3
Age 18-24	1,345	.	.	.	.	.	.	.

Age 25-65	7,037	.	.	.	.	.	.	.
Age 65+	1,671	.	.	.	.	.	.	.
F	4,294	.	.	.	.	.	.	.
M	5,759	.	.	.	.	.	.	.
Midwest	2,703	.	.	.	.	.	.	.
Northeast	1,844	.	.	.	.	.	.	.
South	3,707	.	.	.	.	.	.	.
West	1,799	.	.	.	.	.	.	.

Source: Authors' work

Cluster 4: **Cannot lose them.** This cluster includes 6,911 valuable customers with the second-highest total purchase value (7,784) and an impressive frequency of (3). They contribute a sizable revenue stream to the business if they return to make more purchases. However, they have not purchased in a long time, with an alarmingly high Recency value of (262). These customers are labelled as *Cannot lose them* since they had a relatively important impact on the business's revenue, so the business should not let them go. However, these customers still showed they were leaving the business with that Recency value. In this cluster, women are more likely to leave the business, even though they contribute much higher revenue than men.

Additionally, women over 65 spend a lot of money shopping, but their R-value is the highest. Furthermore, men in the Western Region of the United States aged 18-24 and 25-65 also need attention. Businesses should focus on retaining female customers in all regions and try to entice them with new products to prevent them from switching to a rival company. Businesses must implement the most personalised customer service and strategies to retain this valuable cluster if necessary. Table 16 presents the descriptive statistics relevant to cluster 4.

Table 16
Descriptive statistics of variables Recency, Frequency and Monetary and Demographics variables of cluster 4

Variable	Count	Mean	Standard Deviation	Min	25%	50%	75%	Max
Recency	6,911	262,28	38,23	165	246	277	282	365
Frequency	6,911	3,38	2,4	2	2	3	4	56
Monetary	6,911	7.784,39	11.709,35	99,6	843,3	3.347,2	10.750,5	166.460,9
Age 18-24	956	.	.	.	.	.	.	.
Age 25-65	4,773	.	.	.	.	.	.	.
Age 65+	1,182	.	.	.	.	.	.	.
F	3,853	.	.	.	.	.	.	.
M	3,058	.	.	.	.	.	.	.
Midwest	1,885	.	.	.	.	.	.	.
Northeast	1,232	.	.	.	.	.	.	.
South	2,586	.	.	.	.	.	.	.
West	1,208	.	.	.	.	.	.	.

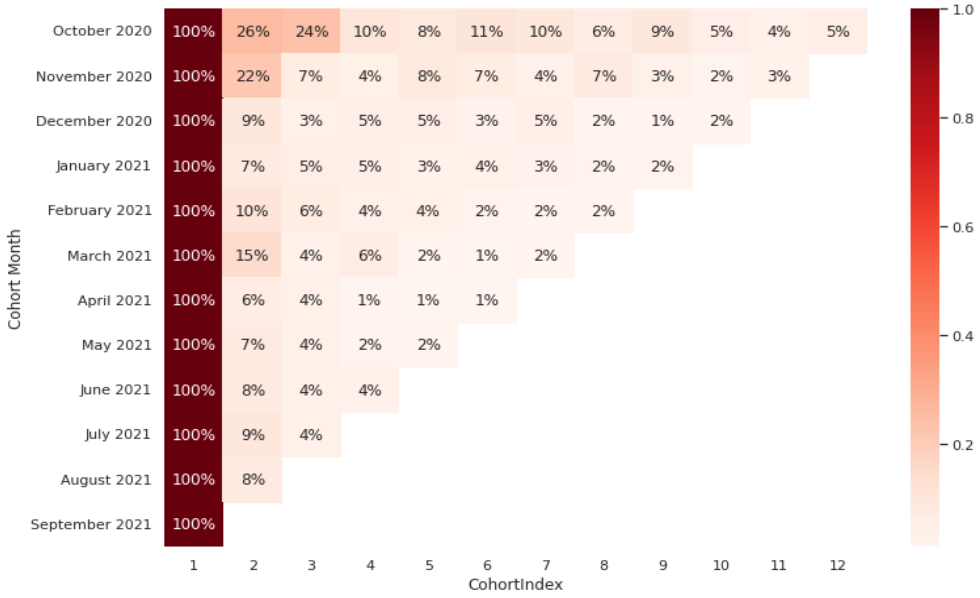
Source: Authors' work

By applying the RFMD model, businesses can have more information about the age, gender, and Region of each cluster. Therefore, businesses can customise marketing strategies based on these demographic features.

Cohort analysis

After analysing all business segments, more than 50% of customers have extremely high R-value>230 days (~ approximately 7 months). The higher the R-value, the higher the churn rate the customer can have. This status needs to be considered, and other aspects of business status, we look at the Cohort analysis below (Figure 6).

Figure 6
Result of Cohort analysis



Source: Authors' work

Cohort analysis is used to measure customer engagement over time:

- Looking at the chart horizontally, the customer retention rate was measured since November 2020 (26%), then it only decreased by 2% in the following month (24%). However, there is a sharp decrease in the retention rate to only 10%, which was reduced by approximately half compared to the previous month. Although it slightly increases by 2-3% in the next months, it continues to decrease to 5% in the last month.
- Looking at the chart vertically, the business obtains the average retention rate after each cycle (one month) with the average value. After a month, the business maintained a ratio of 22%, only declining 4% compared to the previous month. By December 2020, the customer retention rate had declined by 11%. Then, it only increased slightly in February 2021 and March 2021.

In general, this retention rate is very low. Between October 2020 and November 2020, the retention rate remained stable, but in December 2020, there was a severe decline of more than 50%. Enterprise needs to find out the causes and solutions for the dramatic decrease in retention rate in the following months (starting from December 2020). Therefore, a business can improve the churn of customers and reduce R-value.

Discussion

To have a comprehensive overview of the differences, this part divides the previous studies and models into four main types: studies using the RFM model combined with demographics, studies segmenting customers using the RFM model, studies using the RFM model combined with new technologies, and studies segmenting customers using the improved RFM model. This shows the scientific and practical significance of the RFMD model, which is summarised in Figure 2 and the literature review section.

Applying K-Means combined One-hot encoding and K-Prototypes on the RFMD model, five customer segments are determined with different and detailed behaviour and demographic characteristics. However, customer clusters have not been distinctly differentiated; there are similarities in many factors.

Namvar et al. (2010) conducted the first clustering of customers based on the RFM model and then on demographic variables, including Education level, Occupation level and Age, using K-Means algorithms. After two consecutive clustering phases, customers were divided into nine groups compared to three, resulting from the first clustering process. Profiles of nine customer segments were created with the rank of each attribute: RFM, demographic attributes such as Education level, Occupation level, Age, and other characteristics like LTV and Largeness. Sarvari et al. (2016) compared the performance evaluation of 42 scenario types, which had three phases: RFM analysis phase with Weighted RFM analysis, Segmentation Phase with different RFM outputs and Demographic factors, and Association Rules Mining Phase with Demographics if they were not included in Segmentation Phase. Scenario 25 was discovered to be the best with proper numbers of cluster and elapsed time and clusters customers with both R, F, and M scores and demographic data in the segmentation phase. The result produced from this scenario was 5 clusters of customers characterised by R, F, and M scores and demographic factors labelled with ordinal numbers as categorisation.

These studies both include demographic factors in customer segmentation and discover that demographics considerably impact the cluster's effectiveness when integrating with the RFM model to group customers.

However, there are differences between the three studies. First, Namvar et al. (2010) analysis was conducted based on the result with the rank of each attribute. So, it was difficult for marketers to understand more about customer groups. For example, cluster 1 presents the quick general view with the rank of RFM, education level, occupation level and age level, respectively 6, 2, 1, instead of giving detailed information about customers, whether their Education level, Occupation and Age range was. Furthermore, the number of clusters as well as the number of ranks was quite large, which could lead to the fact that each customer segment was not specific or well-defined and the difficulty in targeting customers (Larivière & Van den Poel, 2005; Verhoef et al., 2009). As previously discussed, it was challenging to effectively target its marketing efforts and tailor its products and services to meet the specific needs of each group. Besides (Sarvari et al., 2016), the new finding was a 42-scenario design using K-Means and extracting rules. Using Neutral Network, the clustering with the 25th scenario has been chosen and evaluated to be better than other approaches in comparison with similar works. Despite this, the methodology of this study, with a combination of various clustering and rule extraction approaches, needs to improve the performance of extracting rules with others. Compared to the two studies above, RFMD model research applies K-Means combined One-hot encoding and K-Prototypes to deal with the mixed dataset, including numeric data (R, F, M values) and categorical data (demographic variable), which are popular and proven to be effective clustering algorithms (Hamerly & Elkan, 2002; Huang, 1998). In contrast to previous studies, demographic data are converted into numeric data with One hot encoding when combined with K-Means algorithms in the clustering process instead of being assigned with ordinal numbers. Four variables of the RFMD model can, therefore, be the main inputs for the clustering process and demonstrate the impact of each other in customer segmentation. As a result, customers in each segment are identified with a comprehensive understanding of behaviour and demographic characteristics.

Currently, many improved RFM models have been studied. How does the proposed RFMD model in this study differ in approach or clustering results? To provide a comprehensive overview of the differences, we categorised previous studies into three main types: customer segmentation using the RFM model, customer segmentation using the RFM model combined with new technologies, and customer segmentation using

improved models developed from the RFM model. All are summarised in Figure 2 and the review section.

Many studies have applied new machine learning clustering methods to segment customers using the RFM model. Notable examples include the K-Means++ machine learning method proposed, which was used to segment customers based on the RFM model (Wu et al., 2021). In addition, the DBSCAN and K-Medoids methods have also been applied in clustering. Brahmana et al. (2020) experimented with customer segmentation using the RFM model with three machine learning methods on the same dataset and evaluated them using the Davies Bouldin and Silhouette indices. The results showed that K-Means had the best performance.

The methods are completely different from the techniques used in this study. While newer methods may have better clustering effectiveness, the main objective of this study is to develop an improved model. Therefore, it would be best to use the two most popular clustering methods, K-Means and K-Prototypes, which have been proven effective and serve as a basis for future development methods.

Below is a chronological order comparison of the RFMD model and the existing improved RFM models.

Moghaddam et al. (2017) proposed the RFMV model by adding various products (V), calculated as the number of products a customer has purchased in a given period. Furthermore, the CRISP-DM and K-Means algorithm was used for clustering. Furthermore, Allegue et al. (2020) proposed the RFMC model, using data mining tools to perform customer segmentation based on it. When using the RFMC model, each customer is identified for each Category (C) of the purchased products, and then clustering is performed with the transaction data of each Category. The above studies are completely different from the RFMD model's approach. Rather than proposing a new model for a deeper analysis of purchasing behaviour, we fully utilise customer attribute data, which provides many different perspectives on customers beyond transaction history.

Another study also proposed an improved model to exploit buying behaviour. Wu et al. (2021) proposed an improved model, which includes two additional attributes, S (customer contribution time, referring to the time interval between a user's first and last transactions) and P (repeat purchase attributes), referring to the Frequency of purchases of a specific category of goods made by a particular user within a specified period. Then, they experimented with clustering using the improved machine learning method K-Means++. The results of this study show that some customers are classified as "General" instead of "Loyal", as analysed by the RFM model. That finding is based on a very low S value, indicating that they have never used the app, and a high P indicates that customers only focus on certain products. In the future, the model combined with Demographics values as an input variable for customer segmentation is an interesting topic (Wu et al., 2021).

Conclusion

Summary of research

This study proposes the RFMD model by integrating demographic variables into clustering. With the results of 5 customer segments, detailed information related to behavioural and demographic variables of each segment was extracted to provide useful suggestions for businesses. By successfully testing two popular clustering algorithms, K-Means and K-prototypes, the model demonstrated both the dataset's feasibility and the algorithm selection's suitability. Furthermore, the study also used indicators to evaluate the results and obtained results >0.86 on both clustering

techniques. Cohort analysis is also applied to support in-depth data analysis and propose business marketing strategies.

In business, customer segmentation always plays a crucial role for companies. Finding customer segments allows businesses to identify target customer groups, understand the detailed characteristics of each segment, and develop appropriate business and marketing strategies. Sarvari et al. (2016) developed a study closely related to this research; it was found that analysing demographic factors in customer segmentation is necessary. However, this study proposes a new method by directly integrating demographic variables as input variables into the clustering process along with R, F, and M instead of performing customer segmentation with RFM first and then analysing demographic factors within each cluster. This would support the speed and details of the segmentation process.

This article proposed an extended model RFMD which causes the demographic variables to have a direct impact on the variable's Recency, Frequency and Monetary, making the new model totally different from the old one and completely distinguished from the clustering method using the traditional RFM model and then performing the demographic analysis. Next, two clustering methods are proposed to perform clustering on a mixed dataset with both variables (quantitative variables) and categorical variables (qualitative variables) that are suitable for the RFMD model. Finally, Cohort analysis was conducted based on the clustering results to understand customers better. ARM and AMI are used to compare the clustering results of two proposed clustering methods, which show that the results of both methods are relatively similar. As a result, the RFMD model is a stable data model that can be utilised in customer segmentation to get insight into purchase behaviour and demographic features and evaluate customers more comprehensively.

Implications

The traditional RFM model helps businesses recognise their customer segmentations and purchasing behaviour but cannot reveal the demographic characteristics to launch better-customised marketing campaigns or customer services. The RFMD proposed model can solve this problem, helping businesses reduce marketing and service costs while building effective strategies. For example, instead of sending serial marketing messages or promotions to all customers, which is costly and inefficient, we can customise our marketing strategy based on purchasing behaviour and demographics to target customers. Besides, businesses analyse the relationship between customers' shopping behaviour and demographic characteristics to make the right decisions.

Limitations

Although the research has successfully addressed the research question and yielded good clustering results, the dataset used for testing in this study was from a US retail business that only recorded data for one year. While the mixed data used is suitable for the research objective, the study has only tested on a limited dataset with few demographic attributes, which may limit the generalizability of the method to other attributes or fields. The techniques used in the research are also not diverse; clustering is performed on two common clustering algorithms, and the evaluation index for clustering results is also relatively simple. The research has also not determined the trend of changes in customer shopping behaviour over time; therefore, there is a need for further proposals to improve the method and model.

Future research recommendations

From the above limitations, some recommendations are proposed for future studies. Further studies are suggested to address issues related to the dataset. It is necessary to search and experiment with longer and more diverse demographic variables (such as income and occupation) to make the research results more diverse; each cluster obtained is detailed, and easy to find differences between clusters. One, experiments should also be conducted on multiple datasets in different fields and with many demographic factors to compare accuracy and results and draw conclusions about the impact of each factor. For example, in this study, the variable Region was used to refer to customer addresses, and future studies may further break down this variable into cities, districts or experiment with one demographic variable at a time to obtain more detailed comparison results. In addition, other clustering algorithms should be researched and applied to increase the method's applicability, such as classifying data encoding or another clustering algorithm performed on the entire mixed data. In addition to the above recommendations, the study proposes to apply Time series analysis to analyse the detailed behaviour of each customer cluster in different periods and continuously analyse to help businesses deploy quickly and effectively. Predicting user behaviour is also a suggestion for future studies to support businesses in predicting after analysing detailed results. Although purchasing behaviour and machine learning models can be easily tested and applied, demographic factors, which are diverse, difficult to calculate, and complex, still play a crucial role in customer segmentation.

References

1. Abbasimehr, H., & Shabani, M. (2021). A new framework for predicting customer behaviour in terms of RFM by considering the temporal aspect based on time series techniques. *Journal of ambient intelligence and humanised computing*, *12*(1), 515-531. https://doi.org/10.1007/s12652-020-02015-w
2. Al-Augby, S., Majewski, S., Majewska, A., & Nermend, K. (2015). A comparison of k-means and fuzzy c-means clustering methods for a sample of gulf cooperation council stock markets. *Folia Oeconomica Stetinensia*, *14*(2), 19-36. https://doi.org/10.1515/foli-2015-0001
3. Allegue, S., Abdellatif, T., & Bannour, K. (2020, September). RFMC: a spending-category segmentation. In 2020 *IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* (pp. 165-170). IEEE.
4. Bose, I., & Chen, X. (2015). Detecting the migration of mobile service customers using fuzzy clustering. *Information & Management, 52*(2), 227-238. https://doi.org/10.1016/j.im.2014.11.001
5. Brahmana, R. S., Mohammed, F. A., & Chairuang, K. (2020). Customer segmentation based on RFM model using K-means, K-medoids, and DBSCAN methods. *Lontar Komput. J. Ilm. Teknol. Inf, 11*(1), 32. https://doi.org/10.24843/LKJITI.2020.v11.i01.p04.
6. Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert systems with applications*, *36*(3), 4176-4184. https://doi.org/10.1016/j.eswa.2008.04.003
7. Chiu, S., & Tavella, D. (2008). *Data mining and market intelligence for optimal marketing returns*. Routledge.
8. Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking–An effective approach to customer segmentation. *Journal of King Saud University-Computer and Information* Sciences, *33*(10), 1251-1257. https://doi.org/10.1016/j.jksuci.2018.09.004
9. Covoes, T. F., Hruschka, E. R., & Ghosh, J. (2013). A study of k-means-based algorithms for constrained clustering. *Intelligent Data Analysis*, *17*(3), 485-505. https://doi.org/ 10.3233/IDA-130590
10. Dawane, V., Waghodekar, P., & Pagare, J. (2021). RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention. In *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*.

11. Forgey, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics*, *21*(3), 768-769.
12. Friedman, H. P., & Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, *62*(320), 1159-1178.
13. Gajjar, N. B. (2013). Factors affecting consumer behaviour. *International Journal of Research in Humanities and Social Sciences*, *1*(2), 10-15.
14. Ha, S. H., & Park, S. C. (1998). Application of data mining tools to hotel data mart on the Intranet for database marketing. *Expert Systems with Applications*, *15*(1), 1-31. https://doi.org/10.1016/S0957-4174(98)00008-6
15. Hamerly, G., & Elkan, C. (2002, November). Alternatives to the k-means algorithm that find better clusterings. In *Proceedings of the eleventh international conference on Information and knowledge management* (pp. 600-607).
16. Heldt, R., Silveira, C. S., & Luce, F. B. (2021). Predicting customer value per product: From RFM to RFM/P. *Journal of Business Research*, *127*, 444-453. https://doi.org/10.1016/j.jbusres.2019.05.001
17. Hoegele, D., Schmidt, S. L., & Torgler, B. (2016). The importance of key celebrity characteristics for customer segmentation by age and gender: Does beauty matter in professional football?. *Review of Managerial Science*, *10*(3), 601-627. https://doi.org/10.1007/s11846-015-0172-x
18. Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, *2*(3), 283-304. https://doi.org/10.1023/A:1009769707641
19. Hughes, A. M. (1994). Strategic database marketing: the masterplan for starting and managing a profitable. *Customer-based Marketing Program, Irwin Professional*.
20. Jacoby, J. (1975). Consumer psychology as a social psychological sphere of action. *American Psychologist*, *30*(10), 977-987.
21. Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2023). Customer Profiling, Segmentation, and Sales Prediction using AI in Direct Marketing. *arXiv preprint arXiv:2302.01786*.
22. Kicova, E., Kral, P., & Janoskova, K. (2018). Proposal for Brand's Communication Strategy Developed on Customer Segmentation Based on Psychological Factors and Decision-Making Speed in Purchasing: Case of the Automotive Industry. *Economics and Culture*, *15*(1), 5-14. https://doi.org/10.2478/jec-2018-0001
23. Kumar, A. (2023). Customer Segmentation of Shopping Mall Users Using K-Means Clustering. In *Advancing SMEs Toward E-Commerce Policies for Sustainability* (pp. 248-270). IGI Global.
24. Lakshmi, K., Shanthi, S., & Parvathavarthini, S. (2018). Clustering mixed datasets using k-prototype algorithm based on crow-search optimisation. In *Developments and Trends in Intelligent Technologies and Smart Systems* (pp. 191-210). IGI Global.
25. Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert systems with applications*, *29*(2), 472-484. https://doi.org/10.1016/j.eswa.2005.04.043
26. Liu, D. R., & Shih, Y. Y. (2005a). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information & Management*, *42*(3), 387-400. https://doi.org/10.1016/j.im.2004.01.008
27. Liu, D. R., & Shih, Y. Y. (2005b). Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences. *Journal of Systems and Software*, *77*(2), 181-191. https://doi.org/10.1016/j.jss.2004.08.031
28. Lloyd, S. (1982). Least squares quantisation in PCM. *IEEE transactions on information theory*, *28*(2), 129-137. https://doi.org/10.1109/TIT.1982.1056489
29. Londhe, S., & Palwe, S. (2022). Customer-Centric Sales Forecasting Model: RFM-ARIMA Approach. *Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy*, *13*(1), 35-45. https://doi.org/10.2478/bsrj-2022-0003
30. McDonald, M. (2012). *Market segmentation: How to do it and how to profit from it*. John Wiley & Sons.
31. Miglautsch, J. R. (2000). Thoughts on RFM scoring. *Journal of Database Marketing & Customer Strategy Management*, *8*(1), 67-72. https://doi.org/10.1057/palgrave.jdm.3240019

32. Moghaddam, Q.S., Abdolvand, N., & Harandi, R.S. (2017). A RFMV Model and Customer Segmentation Based on Variety of Products. *Journal of Information Systems and Telecommunication (JIST)*, *3*(19), 155.
33. Namvar, M., Gholamian, M. R., & KhakAbi, S. (2010). A Two Phase Clustering Method for Intelligent Customer Segmentation. *2010 International Conference on Intelligent Systems, Modelling and Simulation*. https://doi.org/10.1109/isms.2010.48
34. Omran, M. G., Engelbrecht, A. P., & Salman, A. (2007). An overview of clustering methods, *Intelligent Data Analysis*. 11(6), 583-605. https://doi.org/10.3233/ida-2007-11602
35. Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*, *15*(1), 12. https://doi.org/10.7275/qbpc-gk17
36. Pol, L. G. (1991). Demographic contributions to marketing: An assessment. *Journal of the Academy of Marketing Science*, *19*(1), 53-59. https://doi.org/10.1007/BF02723424
37. Prabha, K. A., & Visalakshi, N. K. K. (2014). Improved Particle Swarm Optimization Based K-Means Clustering. *2014 International Conference on Intelligent Computing Applications.* https://doi.org/10.1109/icica.2014.21
38. Ritchie, H. and Roser, M. (2019, Sept 20) *Age Structure - Our World in Data*. Retrieved July 31, 2023, from https://ourworldindata.org/age-structure.
39. Romano, S., Bailey, J., Nguyen, V., & Verspoor, K. (2014, June). Standardised mutual information for clustering comparisons: one step further in adjustment for chance. In International conference on machine learning (pp. 1143-1151). PMLR..
40. Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. *Kybernetes*, *45*(7), 1129-1157. https://doi.org/10.1108/K-07-2015-0180.
41. Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.
42. Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing*, *21*(1), 3-8. https://doi.org/10.1177/002224295602100102
43. Stone, B. (1995). Successful Direct Marketing Methods, Lincoln-wood. *IL: NTC Business Books*, 29-35..
44. Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., & Rahmani, R. (2018, October). Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: a case study. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)* (pp. 119-126). IEEE.
45. Verhoef, P. C., Lemon, K. N., Parasuraman, A., Roggeveen, A., Tsiros, M., & Schlesinger, L. A. (2009). Customer experience creation: Determinants, dynamics and management strategies. *Journal of retailing*, *85*(1), 31-41. https://doi.org/10.1016/j.jretai.2008.11.001
46. Vinh, N. X., Epps, J., & Bailey, J. (2009, June). Information theoretic measures for clusterings comparison: is a correction for chance necessary?. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1073-1080)..
47. Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations*. Springer Science & Business Media.
48. Wei, J. T., Lin, S. Y., & Wu, H. H. (2010). A review of the application of RFM model. *African Journal of Business Management*, *4*(19), 4199.
49. Wei, J. T., Lin, S. Y., Weng, C. C., & Wu, H. H. (2012). A case study of applying LRFM model in market segmentation of a children's dental clinic. *Expert Systems with Applications*, *39*(5), 5529-5533. https://doi.org/10.1016/j.eswa.2011.11.066
50. Wu, J., Shi, L., Yang, L., XiaxiaNiu, Li, Y., Cui, X., Tsai, S-B. & Zhang, Y. (2021). User value identification based on improved RFM model and k-means++ algorithm for complex data analysis. *Wireless Communications and Mobile Computing*, *2021*, 1-8. https://doi.org/10.1155/2021/9982484
51. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and SteiD. (2008). *Top 10 algorithms in data mining. Knowledge and information systems*, *14*(1), 1-37. https://doi.org/10.1007/s10115-007-0114-2

52. Wu, Z., Jin, L., Zhao, J., Jing, L., & Chen, L. (2022). Research on Segmenting E-Commerce Customer through an Improved K-Medoids Clustering Algorithm. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/9930613
53. Yeh, I. C., Yang, K. J., & Ting, T. M. (2009). Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with applications*, *36*(3), 5866-5871. https://doi.org/10.1016/j.eswa.2008.07.018.
54. Yıldız, E., Güngör Şen, C., & Işık, E. E. (2023). A Hyper-Personalised Product Recommendation System Focused on Customer Segmentation: An Application in the Fashion Retail Industry. *Journal of Theoretical and Applied Electronic Commerce Research*, *18*(1), 571-596. https://doi.org/10.3390/jtaer18010029
55. Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2022). Missing data preprocessing in credit classification: One-hot encoding or imputation?. Emerging Markets Finance and Trade, 58(2), 472-482. https://doi.org/10.1080/1540496X.2020.1825935.

About the authors

Thanh Ho received an M.S. degree in Information Technology from the University of Information Technology, VNU-HCM, Vietnam, in 2009 and a Ph.D. in Information Technology from the University of Information Technology, VNU-HCM, Vietnam, in 2018. He is an Associate Professor and Senior Lecturer in the Faculty of Information Systems, University of Economics and Law, VNU–HCM, Vietnam. His research interests are Management Information Systems, Business and Data Analytics, Business Intelligence, and Artificial Intelligence for Economics. The author can be contacted at **thanhht@uel.edu.vn**

Suong Nguyen is a senior student in Management Information, Systems University of Economics and Law, VNU-HCM, Vietnam. Her research interests are Data Analytics and Customer Analytics. The author can be contacted at **suongntn20406@st.uel.edu.vn**

Van-Huong Nguyen is a senior student in Management Information Systems, University of Economics and Law, VNU-HCM, Vietnam. His research interests are Data mining and Business Intelligence. The author can be contacted at **huongnv20406c@st.uel.edu.vn**

Ngoc Nguyen is a senior in Management Information Systems, University of Economics and Law, VNU-HCM, Vietnam. Her research interests are Data mining and Data Analytics. The author can be contacted at **ngocntb20406c@st.uel.edu.vn**

Dac-Sang Man is a senior in Management Information Systems, University of Economics and Law, VNU-HCM, Vietnam. His research interests are Data Analytics and Business Intelligence. The author can be contacted at **sangmd20406c@st.uel.edu.vn**

Thao-Giang Le is a senior student in Electronic Commerce, University of Economics and Law, VNU-HCM, Vietnam. Her research interests are Digital marketing, Data Analytics, and Customer Analytics. The author can be contacted at **gianglt20411@st.uel.edu.vn**