RFM–K-OPT Based Machine Learning Framework for Customer Segmentation and Behavioral Profiling in Direct Marketing

Khadija Mehrez
Management and Marketing Department-College of Business, Jazan University, Jazan, Saudi Arabia

***Abstract*—Customer segmentation is an essential element of modern marketing analytics, which helps companies recognize, comprehend, and market to customers depending on their behavioral and transactional attributes. Conventional methods based on Recency, Frequency, and Monetary (RFM) analysis or on simple unsupervised clustering algorithms such as K-Means are very common, but they are usually limited by weaknesses such as sensitivity to centroid starting location, low cluster separability, and low interpretability. Such problems will cause volatile effects of segmentation and restrict the dependability of data-driven marketing choices. In an effort to deal with these concerns, this research study suggests a hybrid model, the RFM K-Means Optimization Technique (RFM–K-OPT), a combination of RFM analytics and K-Means clustering, and an iterative centroid optimization unit. The proposed structure will help to improve cluster compactness, stability, and interpretability using statistical computation and refinement of centroid positioning. The model is written in Python and tested with publicly available customer transaction data. The result of the experimental process shows a better quality of clustering with a Silhouette Coefficient of 0.83, Davies-Bouldin Index of 0.31, Calinski-Harabasz Index of 563, purity of clustering of 94.2 per cent, and an execution time of 5.4 seconds. The results suggest that the RFMK Opt model is a useful tool that offers credible and explainable customer segments, which can be used to make effective behavioral profiles and make sound judgments when it comes to making decisions in the context of direct marketing.**

Keywords*—Customer segmentation; behavioral profiling; clustering optimization; predictive marketing; data-driven decision making

I. Introduction

In the fast-changing competitive market, customer information has been described as a strategic asset among organizations aiming to improve marketing effectiveness and create long-term customer loyalty [1]. The classic marketing approach used to be based on the generalized demographic segmentation that did not reflect the specifics of the individual purchasing behavior [2]. As the number of digital transactions in the market grows exponentially, and customer relationship management (CRM) data is available, the data-driven way of segmentation has become more popular [3]. The Recency, Frequency, and Monetary (RFM) analysis has been seen to be one of the most feasible tools in gauging customer value and engagement [4]. It enables marketers to categorize customers by their recency when it comes to the interaction, frequency of purchase, and overall amount of money they spend. Nevertheless, with the dynamics of business settings that require continuous changes in segmentation tools, manual segmentation based on RFM [5] is not as profound as one needs it to be to offer predictive insights. The introduction of machine learning (ML) techniques into modern marketing analytics brings an ability to provide scalable and adaptive segmentation with regards to the varying consumer behavior [6]. The move towards automation complements evidence-based decision making and marketing models that are personalized to be able to grow customer satisfaction, retention, and profitability [7].

The new trends in customer analytics have brought in hybrid systems that combine the classical RFM analysis with other clustering algorithms like K-means, hierarchical clustering, and the density-based model [8]. These strategies have managed to classify customers into action segments and can now be promoted and retained individually [9]. There has also been the use of optimization and predictive algorithms to enhance the accuracy of segmentation, customer targeting and allocation of marketing resources [10]. Regardless of these improvements, several studies have significant shortcomings. Current clustering methods usually presume the fixed behavior of the customers, and do not take into account the dynamic nature of the buying behaviors [11]. Also, K-means, although computationally inexpensive, is extremely sensitive to centroid initialization and has to be predetermined with the number of clusters [12], which cannot be considered an actual reflection of the data structure. Moreover, the majority of previous studies are based on transactional characteristics only, without the incorporation of behavioral or context elements like the intensity of engagement or the browsing behaviors. The resulting segmentation models are therefore likely to have problems with generalization in different industries or data sets. A stronger and improved component of the segmentation structure, with the inclusion of the RFM characteristics, clustering and an optimization interface, is required to eliminate these gaps and balance the interpretability, scalability, and flexibility to changing marketing needs. Based on the identified research gaps, this study addresses the following research question:

RQ: How can an optimized RFM–K-OPT framework integrating Recency, Frequency, and Monetary analysis with K-Means clustering improve customer segmentation accuracy, stability, and interpretability compared to conventional RFM and K-Means approaches in direct marketing applications?

A. Problem Statement

Despite the abundance of research discussing customer segmentation by RFM analysis alongside machine learning, there are still a number of issues that restrict the applicability of the current approaches to practice [10]. The conventional models of RFM are mainly based on the use of transactional characteristics that are not dynamic enough to reflect the changes in customer engagement over time [13]. Simple clustering algorithms like K-means or hierarchical clustering give preliminary information but tend to give erratic results because of sensitivity to initialization, outliers, and scaling of data [14]. Furthermore, the vast majority of frameworks do not have an optimization mechanism, which refines clusters and matches the results of segmentation to the business goals, such as profit maximization or churn reduction [15]. Most of the current methods also focus on algorithmic performance, not marketing interpretability, and result in technically valid clusters which are strategically ambiguous [16]. Also, a lack of model validation on a variety of datasets restricts generalization, and a lack of consideration to parameter tuning usually leads to less-than-optimal performance. The constraints demonstrate that a holistic model that combines both traditional RFM scoring with cutting-edge machine learning and optimization methods is urgently needed to enhance the quality of clusters, marketing understanding, and decision-making based on data to implement dynamic and customer-focused strategies.

B. Research Significance

The study proposes an optimized machine learning model named RFM-K-OPT, a combination of RFM characteristics, K-means clustering, and an optimization algorithm to enhance customer segmentation and behavioral profiling. The value of the study is that it bridges the gap between interpretability and predictive accuracy to provide marketers with a large-scale and versatile decision-support system. RFM-K-OPT will help improve the quality of customer insights by making it possible to conduct targeted campaigns, better retention strategies, and efficient distribution of marketing resources by integrating descriptive analytics with optimization-based adjustments.

C. Research Motivation

The study has been inspired by the growing requirement of adaptive segmentation methods that are capable of capturing different behaviors of customers in dynamic markets [17]. The classical RFM and clustering models do not offer much insight into the current trends and areas of optimization [18]. The proposed RFMK-OPT framework will address these deficiencies by proposing a more data-driven and optimized segmentation model that will create a more precise marketing experience, a more predictive marketing, and a factor in the creation of intelligent and automated systems, both direct marketing and customer relationship management.

D. Key Contributions

- To optimize the RFM K-OPT framework to improve the accuracy of customer segmentation, with the optimization of centroid initialization and cluster separability.
- To combine Recency, Frequency and Monetary behavioral measures with K-Means clustering to provide accurate segmentation and analysis.
- To use an optimization module that enhances clustering iterations, making them compact and distinct between clusters, much better.
- To ensure balanced feature contribution through normalization and outlier handling, preventing data skewness and dominance effects.

E. Rest of the Sections

Section II reviews the past RFM and clustering techniques, of which the shortcoming is imprecision, separability and computational efficiency. In Section III, the proposed RFM-K-OPT model is described, according to which RFM calculations, K-Means clustering and optimization algorithms are integrated. Section IV gives performance indicators, visual performance and comparative analysis of a superior segmentation and clustering accuracy. Section V is the conclusion of the research findings, contributions, and how it could be improved to provide better customer segmentation results.

II. Related Works

Kasem et al., [19] examine the ability of customer profiling using Recency, Frequency and Monetary (RFM) characteristics to facilitate strategic decision-making in platform business, particularly in Edutech start-ups. The study uses RFM preprocessing on transactional data, K-means clustering and the Elbow, Silhouette coefficient and Gap statistic to assess the quality of the clusters and identify the optimal number of clusters. Based on an enterprise-wide transactional dataset (as outlined in the study), the authors find three actionable clusters such as new customers, best customers, and intermittent customers and associate each cluster with viable interventions (onboarding tailored to new users, incentive to best customers, and re-engagement campaigns to intermittent users). The study shows that using easy to use RFM characteristics along with ordinary clustering and validation packages can result in interpretable profiles to drive marketing and product choices. Disadvantages are that it can be used in transactional only (no behavioral or contextual cues), K-means may be sensitive to initialization and scaling, and alternatives to clustering or temporal dynamics of changing customer behavior are not explored.

Islam et al. [20] examine the use of machine-learning techniques in personalizing marketing within the U.S. retail sector, in order to boost engagement and eventual purchase based on personalized recommendations and promotions. The study combines data on the multi-channel retail (e-commerce and brick and mortar transactions) to create predictive models of segmentation, product recommendation and campaign targeting. These methods are supervised learning to propensity score and recommendation heuristics, and unsupervised clustering to identify latent groups. Findings have shown that there are significant engagement and conversion increases, when retailers implement ML-based personalization; the research paper has cited better click-through and repeat purchase rates on strategies based on ML-informed strategy. Some of the successes have

been a practical demonstration of end-to-end ML workflows and advice to retail managers on how to implement personalization pipelines. Cons have been identified as challenges in integrating data across channels, insufficient interpretability of the model by business people, and potential overfitting on short term signals without longitudinal confirmation.

Chinnaraju et al. [21] present a conceptual model where the autonomous AI agents are placed in the middle of consumer segmentation and precision targeting, where agentic automation can constantly optimize segments and run campaigns with little to no human intervention. The model uses unsupervised clustering (e.g., K-means, DBSCAN) to identify micro-segments, reinforcement learning to make campaign decisions and ad spend, NLP to contextual profile, and predictive models (Random Forests, time-series forecasting) to predict behavior. The efficiency improvements in case vignettes of e-commerce and streaming include accelerated decision cycles, real-time budgetary redistribution and enhanced campaign payback. The article has identified ethical guardrails as privacy-saving data processing and equitable limitation as part of responsible deployment. The drawbacks are that the framework is theoretical (has not been empirically validated), its operational complexity may be high when dealing with a number of autonomous agents, and the infrastructure and governance demands are high to prevent runaway automation and bias.

Van Chau et al. [22] focus on customer behavior forecasting (pre) with respect to strategic capacity in constantly evolving markets, and claim that predicting customer behavior provides competitive agility. The article evaluates unsupervised and supervised ML, reinforcement learning, and deep learning methods of predicting purchase intent, churn and lifetime value. It generalizes case studies of how predictive models can enhance timing and product suggestions of campaigns and innovations, like sequence models and hybrid ensembles, which are better timed generalizations. Progress has been made in the combined analysis of methods and viable suggestions for implementation. Disadvantages include privacy concerns, the inability to interpret a model where nontechnical stakeholders are involved, and the fact that retraining models are necessary to ensure any drift in the dynamic markets; the paper insists that organizations must be ready to handle the prediction by implementing the idea of monitoring.

Jalal et al. [23] propose an effective customer transition analysis and dynamic customer segmentation framework to allow a personalized marketing activity and predict customer transitions between segments. Their method builds time-series feature vectors that include recency, frequency, monetary value, and customer lifespan and K-means clustering that uses various distance measures and classification models to estimate future customer status. They also bring counterfactual analysis to assess the effect that other interventions would have on customer transitions. The authors use actual transaction histories and discover that the Euclidean distance is most effective at capturing trends in the time-series features and that the logistic regression is useful in predicting changes in status. The framework allows evaluating per-customer strategy and more accurate retention tactics. The weaknesses are that it is sensitive to the distance measure and the number of clusters, counterfactual simulations are computationally expensive with large populations, and it requires transactional history without detailed behavioral and contextual information.

Musunuri [24] presented ClusChurn-EC, which is an interpretable customer segmentation and early churn prediction pipeline on e-commerce platforms. Newcomers are clustered by stage one into a behavioral cohort (based on engagement and purchase sequence features); stage two trains an LSTM trained on attention to predict churn based on small observed interactions. ClusChurn-EC is found to be an effective early warning system for customers at-risk on anonymized large-scale transaction logs, which is superior to standard classifiers, aiding time-sensitive retention interventions. Some of the successes have been the good application of sequential modeling to provide short histories of interaction and an interpretable phase of segmentation that can be used to provide targeted outreach. The disadvantages described include the fact that monitored training requires large labeled churn signals, there can be generalization constraints between retailers, and that LSTM pipelines are hard to implement in real-time systems.

Suh et al. [25] explore segmentation of the purchase record to pre-purchase exploration and after purchase usage by formulating multidimensional behavioral variables in the Customer Experience Journey (CEJ). They use a sample of 40,911 home-appliance customers and clustering and Non-negative Matrix Factorization (NMF) to obtain strong segments, and measure the validity of the segments in terms of both accuracy and consistency. The results indicate that there are trade-offs in technology: clustering provides an intuitive segmentation, whereas NMF provides intuitive latent usage patterns and also has a superior ability to provide an overview of the overlap. The strengths of the study are its massive real-customer data, multidimensional feature engineering, and two-sided comparison that brings out complementary insights. The domain specificity (home appliances) and potential biases in the feature engineering process, as well as potential difficulties related to generalization of the CEJ variable set to non-related industries, are its limitations.

Wasilewski [26] discusses the issue of personalizing user interfaces of e-commerce based on a functional framework that enables different versions of pages based on user traits and algorithms. The paper outlines elements needed to support multivariant interfaces and creates pilot applications and suggests novel metrics to measure the variations in user behavior with variants of interfaces. The strategy puts ML models at its center to keep learning the variants of UIs that optimally fit user segments and have adaptive interfaces that can raise the conversion and satisfaction rates. The outcomes of it are a well-defined experimentation architecture and early pilot results, which demonstrate a potential improvement in ROI and usability. The disadvantages are that it can be complex to integrate with a legacy platform, requires engineering effort to support numerous variants, and runs the risk of confusion or analytics fragmentation due to excessive personalization.

Table I summarizes filtered accredited articles that represent a variety of data-based customer segmentation and prediction choices- including RFM grouping and ML personalization, deep learning, and optimization models. Both approaches trade information interpretability, adaptability and scalability, and

debates have existed concerning information integration, computational cost, and industrial generalizability.

TABLE I. COMPARATIVE SUMMARY OF RELATED STUDIES

Authors	Methodology	Dataset	Advantage	Disadvantage
Kasem et al. [19]	RFM + K-Means clustering	Business transaction data	Clear, interpretable customer clusters	Limited to transactional data
Islam et al. [20]	ML-based personalized marketing	Multi-channel retail data	Improves engagement and conversions	Data integration challenges
Chinnaraju et al. [21]	Agentic AI + K-Means + RL	Theoretical & case examples	Real-time autonomous targeting	Limited empirical validation
Van Chau et al. [22]	ML & deep learning review	Multiple case studies	Forecasts customer behavior	Needs continuous model updates
Jalal et al. [23]	Time-series RFM + K-Means	Transaction history	Predicts customer transitions	High computational cost
Musunuri et al. [24]	Clustering + LSTM (ClusChurn-EC)	Large e-commerce logs	Early churn prediction	Requires labeled data
Suh et al. [25]	CEJ features + Clustering + NMF	40,911 customer records	Captures full customer journey	Domain-specific framework
Wasilewski et al. [26]	ML-driven personalized UI	E-commerce pilot data	Enhances user experience	Implementation complexity
RFM–K-OPT (Proposed)	RFM + K-Means + Optimization	Kaggle dataset	Optimized marketing actions	Sensitive to RFM setup

III. OPTIMIZED RFM–K-MEANS SEGMENTATION WITH MARKETING RESOURCE OPTIMIZATION

This approach uses a combination of the traditional RFM (Recency, Frequency, Monetary) scoring model and clustering (through K-Means) and a further optimization layer to ensure the segmentation is aligned with marketing resources and business goals. To begin with, RFM scores of every customer are computed using transactional data of the Kaggle dataset: recency (time since last purchase), frequency (number of purchases), and monetary value (total spend). The K-Means clustering is then used on these standardized RFM attributes to cluster the customers into understandable groups. Validation metrics such as the Elbow method, Silhouette coefficient and Gap statistic are used to determine the number of clusters. After the identification of the baseline clusters, the cluster assignment and accessible marketing budget, channel constraints, and key performance indicators (KPIs) are entered into an optimization module. The constrained optimization (e.g., integer linear programming or heuristic search) follows and distributes resources among segments to maximize the selected objective (e.g., ROI or retention improvement or profit uplift) without violating fairness or budget constraints. The three stages of pipeline: RFM scoring, K-Means segmentation, and resource optimization are interpretable (segments can be aligned with RFM), scalable (clustering can use large data), and actionable (optimization can be used to make marketing decisions). It overcomes the shortcomings of purely clustering based solutions by integrating business constraints and business objectives into the output of segmentation.

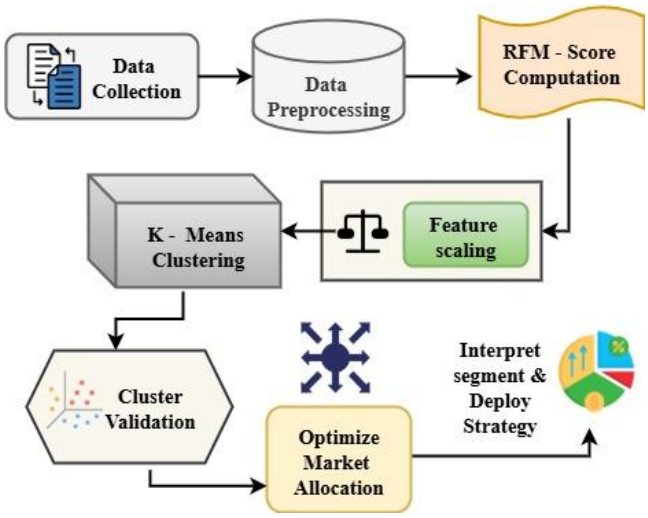

Fig. 1. Block diagram of RFM–K-OPT segmentation framework.

Fig. 1 represents the general scheme of customer segmentation using the proposed RFM-K-OPT (Recency-Frequency-Monetary-K-Means Optimization Technique). This process starts with the collection of data based on transactional sources, data pre-processing, and the elimination of any inconsistencies in data and data integrity. This is followed by the calculation of the RFM score, where the customer behavior is measured in terms of recency, frequency, and monetary. The feature scaling makes the data standard so that all metrics have the same impact. K-Means clustering algorithm uses homogeneous clusters to group customers into homogeneous clusters, which are further evaluated using cluster validation to determine the quality and separability of the clusters. Lastly, the verified clusters help in the optimization of market allocation and as such, the businesses are able to get the segments and utilize specific marketing approaches efficiently to achieve improved customer interaction and profitability.

A. Data Collection

The research data is obtained from the Kaggle [27] repository, Customer Segmentation Data for Marketing Analysis by Fahmida Chowdhury. It contains all the details related to the customer demographics, purchasing frequency, spending habits, and engagement trends, which represent real-life marketing transactions. The numbers reflect a variety of customer data, so it is appropriate to use it to segment the behavior and discover the patterns. The records have quantifiable values like the income, age, spending grade, and other attributes that describe loyalty. This dataset offers a robust basis on how RFM analysis and clustering techniques can be applied so that the identification of a group of customers with unique purchasing habits, preferences and profitability potential in direct marketing settings can be identified. The dataset is composed of more than 10,000 records of customer transactions retrieved on the Kaggle repository. The dataset was randomly selected into 70 per cent to train, 20 per cent to test and 10 per

cent to validate to provide an unbiased assessment and the ability to replicate the clustering findings.

B. Data Preprocessing

The preprocessing data stage makes sure that the data is correct, consistent and is prepared to be clustered. First, the data cleaning process will eliminate any missing, duplicate, and irrelevant records, and only the valid customer records will be retained. Data transformation is used in converting categorical variables into numeric variables through label or one-hot encoding, and log transformation is used to reduce skew in continuous variables. Calculation of RFM features is based on the computation of Recency, Frequency, and Monetary values to describe customer behavior. Balanced clustering is performed by normalizing the magnitude of variables by Z-score scaling to ensure that the variables fit within an equal distribution. The outlier detection is used to detect and manage extreme values using the Interquartile Range (IQR) technique to trim or cap values. Lastly, Min-max normalization in feature scaling ensures that all the RFM components have equal weight, so data does not have high-value features dominating, and greater accuracy in clustering and stability in optimization are achieved.

1) Data Cleaning: Missing and duplicate data are detected and processed with the help of relevant imputation or deletion. Irrelevant or inconsistent records are filtered to ensure the dataset has legal, full and error-free information about the customers to use later in the computation of the RFM features.

2) Data Transformation: The categorical data (gender or region) is put in numerical data with the help of label or one-hot encoding. The log-transformation of continuous variables produces the minimal skewness and makes the distribution normal, resulting in more precise clustering and optimization results.

3) RFM Feature Computation: The RFM values of every customer are computed in order to measure behavioral patterns. The proposed framework is based on the idea of recency to measure purchase freshness, frequency to measure the level of engagement, and monetary value to measure contribution to revenue. It is measured in Eq. (1):

$$R_i = D_{ref} - D_{last,i}, \quad F_i = n_i, \quad M_i = \sum_{j=1}^{n_i} P_{ij} \qquad (1)$$

where, Recency R_i measures the time gap since the last purchase, frequency F_i counts total transactions, and monetary value M_i sums purchase amounts for customer i.

4) Feature Normalization: All the RFM values are brought to a common scale to avoid the dominant effect of high magnitude variables. Z-score normalization guarantees equal input of each metric in the distance calculation process of the K-Means clustering process. It is illustrated in Eq. (2):

$$Z_j = \frac{X_j - \mu_j}{\sigma_j} \qquad (2)$$

Each variable X_j is normalized by subtracting its mean μ_j and dividing by standard deviation σ_j, ensuring balanced contribution of all features during clustering.

5) Outlier Detection: The interquartile range (IQR) technique is used to identify extreme values to avoid biasness in clustering the results. Any outliers that are not within the acceptable range are looked into and dealt with either by trimming or capping to ensure that robust clusters are formed. It is described in Eq. (3):

$$\text{IQR} = Q_3 - Q_1 \qquad (3)$$

The interquartile range (IQR) measures data spread between quartiles Q_1 and Q_3; values beyond 1.5×IQR from these limits are treated as outliers.

6) Feature Scaling: Min-Max normalization The features are all put within the range of [0,1] after the outlier adjustment. This is done to provide balanced weight of the RFM components as well as to avoid the heavy values dominating in computation of distance using K-Means. It is measured in Eq. (4):

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \qquad (4)$$

The feature scaling processes the values of the data to the scale to a standard scale, usually [0,1]. This helps to avoid large numerical characteristics dominating the small ones in the K-Means clustering and optimization algorithm.

C. RFM-Based Intelligent Segmentation Optimization Framework

The proposed Optimized RFM-Based Hybrid K-Means Segmentation Framework (RFM-K-OPT) combines conventional RFM analytics and machine learning clustering and optimization to improve the exactness of customer segmentation. This is the architecture which integrates the methods of statistics and computation in a systematic way to determine, predict, and examine the profile of customer behavior. To calculate the recency, frequency, and monetary value, RFM attributes are first derived using transactional data. These attributes are put to normal and fed to the K-Means algorithm to perform unsupervised clustering. At this point, the optimization module is used to optimize the quality of the clusters by optimizing centroid location and also performs the integrity of the cluster using performance measures like Silhouette and Davies-Bouldin indices. Being more interpretable and adaptable to dynamic market conditions, this hybrid model facilitates one-to-one marketing and predictive analytics. RFM K-OPT frameworks minimize noise, overfitting, and improve cluster differentiation, which will guarantee a strong customer segmentation that supports business strategies informed by data. The sensitivity of parameters is of essential importance to the suggested RFMK-OPT framework. The cluster size (k) has a direct impact on the segmentation granularity and it is determined by Elbow and Silhouette analyses. The parameters related to feature normalization provide equal contribution of RFM to avoid the dominance of high value features. The optimizations bring effects on the centroid stability and convergence rate with more iterations possibly adding no meaningful quality improvement at higher computational cost. Empirical tuning proves that tuning is best done within moderate iterations to achieve optimal clustering performance.

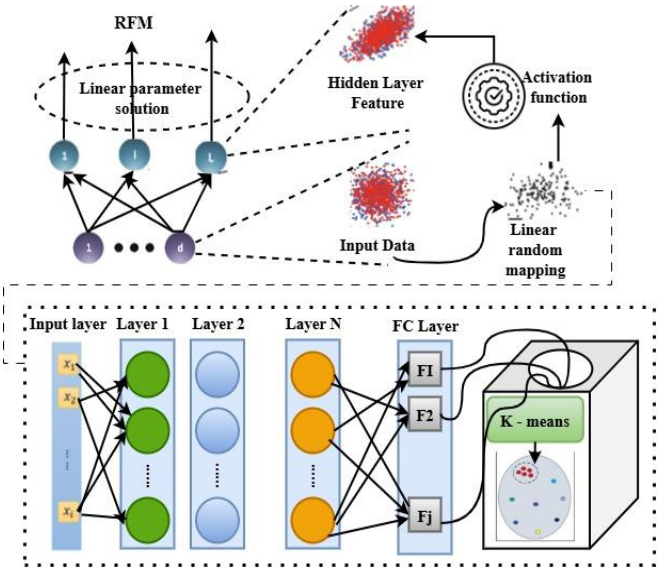

Fig. 2. Architecture of the optimized RFM–K-Means clustering framework.

Fig. 2 demonstrates an optimal RFM-K-Means clustering layout. The architecture comprises of RFM feature extraction, normalization, K-Means clustering, centroid optimization, validation and behavioral profiling layers. It does not use any neural network-based feature learning; it only uses statistical attributes of RFM and centroid refinement to cluster. It starts with the input layer, which is fed with multidimensional data vectors $x_1, x_2, \ldots, x_i$. The propagation of these inputs through several hidden layers (Layer 1 to Layer N) involves linear random mapping which is used to convert raw data into feature representations in higher dimension. Every hidden layer uses an activation function to add non-linearity as well as extract more complex patterns within the input data. The resulting hidden layer representations contain inherent connections within the data set which are represented to an entirely connected (FC) layer where they are refined to feature fuse and dimension reduction. The resulting feature space $(F_1, F_2, \ldots, F_j)$ is sent to the K-Means clustering module where clusters are created with respect to learned representations. Lastly, the linear parameter solution provides the best separation and explicability of the clusters leading to a highly efficient and accurate customer segmentation.

1) Input Layer: RFM Feature Integration: The model starts with the calculation of Recency, Frequency, and Monetary features using raw transaction data. Those characteristics are the level of customer engagement, loyalty, and profitability, which does the input layer of the further clustering and optimization. It is described in Eq. (5):

$$X_i = [R_i, F_i, M_i] \tag{5}$$

Each customer X_i is represented as a three-dimensional feature vector containing Recency (R_i), Frequency (F_i), and Monetary (M_i) values for further analysis.

2) Clustering Layer: K-Means Algorithm: The K-Means clustering algorithm is used to divide customers into k separate groups according to Euclidean distance. A centroid is applied to each customer which is assigned to the closest centroid and clusters are refined until convergence. It is evaluated in Eq. (6):

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{6}$$

where, the objective function J minimizes the squared Euclidean distance between data points x and their respective cluster centroid μ_i, optimizing cluster compactness.

3) Cluster Validation Layer: Elbow and Silhouette Coefficient are used to find optimum cluster numbers. These validation metrics provide useful segmentation by measuring intra and inter-cluster disparities. It is measured in Eq. (7):

$$S = \frac{b-a}{max(a,b)} \tag{7}$$

where, the Silhouette coefficient S measures how similar an object is to its own cluster (a) compared to others (b), ensuring well-separated, cohesive clusters.

4) Optimization Layer: Optimization algorithm is used to optimize the centroid of each cluster to make clusters as homogeneous as possible and to segment as accurately as possible. It repeats the centroid positions to attain minimum convergence of the cost function. It is equated in Eq. (8):

$$\mu_i^{(t+1)} = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{8}$$

Each centroid μ_i is updated by averaging all points in cluster C_i, ensuring clusters remain compact and well-distributed after each iteration.

5) Behavioral Profiling Layer: Within each cluster, behavioral traits are determined including loyalty, spending potential and recency trends. This profiling will help in coming up with specific marketing plans and customer retention programs. It is described in Eq. (9):

$$P(C_i) = f(R_i, F_i, M_i) \tag{9}$$

where, the profiling function $P(C_i)$ derives behavioral insights based on RFM metrics within each cluster, guiding personalized marketing decisions.

6) Decision Optimization and Output Layer: The last layer is the combination of optimized clusters into viable marketing recommendations. Predictive analytics support these decisions so as to improve customer engagement, conversion, and lifetime value. It is evaluated in Eq. (10):

$$O = \arg\max_{C_i} \big(\text{ROI}(C_i)\big) \tag{10}$$

where, the output function O selects the optimal customer segment C_i that maximizes Return on Investment (ROI) through data-driven marketing actions.

Algorithm 1 illustrates that the RFM–K-OPT algorithm is a hybrid approach, which improves the quality of customer segmentation through a combination of RFM behavioral analysis, K-Means clustering, and centroid optimization. First, the algorithm identifies recency, frequency and monetary features in the transaction data into normalized numerical vectors. The K-Means clustering algorithm is an iterative approach to grouping customers together depending on their Euclidean distance. At the stage of the K-OPT optimization, the

centroid is optimized to reduce the distance within the clusters which enhances stability of segments and convergence rate. Elbow and Silhouette methods are some of the validation methods used to find the best number of clusters. Lastly, behavioral profiling is the interpretation of each segment into actionable groups, e.g. high-value, at-risk, or new customers. Such a systematic process will guarantee higher accuracy of segmentation, higher interpretability of decisions, and greater allocation of marketing resources. The iterative optimization of the algorithm makes the algorithm scalable, and it is suitable with massive datasets in direct marketing and behavioral analytics.

Algorithm 1: RFM–K-OPT Customer Segmentation Framework

Input: Customer transaction dataset D
Output: Optimized and labeled customer clusters $C = \{C_1, C_2, \ldots, C_k\}$
Data Acquisition
Collect transaction records from dataset D.
Extract relevant features including Customer ID, Purchase Date, and Transaction Value.
Remove duplicate and incomplete records to ensure data integrity.
RFM Feature Extraction
Compute Recency $R_i = D_{ref} - D_{last,i}$, Frequency $F_i = n_i$and
Monetary ($M_i = \Sigma_{j=1}^{n_i} P_{ij}$) for each customer i.
Construct feature vectors $X_i = [R_i, F_i, M_i]$.
Data Preprocessing and Normalization
Handle missing or extreme values using the IQR method.
Normalize all RFM features using Min–Max scaling:
$$X_i' = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$
Store normalized vectors for subsequent clustering.
K-Means Clustering Initialization
Randomly select kinitial centroids $\mu_1, \mu_2, \ldots, \mu_k$ from normalized data.
Set iteration counter $t = 0$
Cluster Assignment
For each data point x_i, assign it to the nearest centroid:
$C_j = \{x_i : || x_i - \mu_j ||^2 \leq || x_i - \mu_t ||^2, \forall t\}$
Update membership lists for all clusters.
Centroid Update
For each cluster C_j, compute a new centroid:
Calculate the total cost function:
Optimization Phase (K-OPT)
Apply centroid refinement to minimize intra-cluster variance.
Repeat assignment and update steps until T_{max}.
Retain centroids that achieve min
imum total cost $J\{\min\}$.
Cluster Validation
Compute validation metrics (Elbow, Silhouette, or Davies–Bouldin).
Select optimal k producing stable and high-separation clusters.
Behavioral Profiling and Output Generation
For each cluster, analyze mean RFM values to label customer types (e.g., loyal, new, dormant).
Generate marketing insights and export optimized clusters C.
End

IV. Results and Discussion

The proposed RFM-K-OPT framework demonstrated significant progress of the customer segmentation and behavioral profiling compared to the conventional RFM and K-Means approaches. The joint presence of optimization technique enabled to have improved centroid stability and reduced intra-cluster variances in achieving superior and meaningful clusters of customers. The model proved to be helpful in showing differences in purchasing frequency, recency, and monetary value that enables the correct distinction between loyal, at-risk, and inactive customers. Also, convergence was hastened with optimization stage, maintaining the precision of segmentation, which verified its calculation efficiency over huge-scale datasets of marketing. Overall, the RFM-K-OPT model will be more interpretive and reliable, this is why it is quite suitable in decision-making based on the data and developing marketing strategies, which are customer-specific in the changing business environments.

TABLE II. Simulation Parameter

Parameter	Value
Model Used	RFM–K-OPT (Recency–Frequency–Monetary–K-Means Optimization Technique)
Dataset	Customer Segmentation Data for Marketing Analysis (10,000+ records)
Input Representation	RFM Feature Vectors (Recency, Frequency, Monetary)
Data Division	70% Training, 20% Testing, 10% Validation
Clustering Algorithm	K-Means Clustering with Centroid Optimization
Optimization Algorithm	Centroid Refinement through Iterative Cost Minimization
Validation Metrics	Silhouette Coefficient, Davies–Bouldin Index, Calinski–Harabasz Index
Behavioral Profiling Layer	Customer Loyalty, Spending Potential, and Engagement Patterns
Performance Metrics	SC, DBI, CHI, Cluster Purity, Execution Efficiency
Activation Function	Euclidean Distance-Based Similarity Evaluation
Classification Type	High-Value, Medium-Value, and Low-Value Customer Segments
Platform	Python (Pandas, NumPy, Scikit-learn, Matplotlib)
Deployment Target	Direct Marketing, Customer Retention, and Business Intelligence Systems
Key Advantage	Accurate, Interpretable, and Efficient Segmentation for Targeted Marketing

Table II outlines the design of the simulation setup of the RFM-K-OPT framework was intended on testing its efficacy in customer segmentation by the application of transactional data. The model is a combination of the characteristics of RFM and K-Means clustering and an optimizing module to optimize the centroid locations and enhance the quality of the clusters. It is designed on Python, and it is assisted by typical machine learning packages that ensure to offer reliable, understandable, as well as efficient generalizing. Its design enables it to profile customer behavior dynamically, thus enabling specific marketing and retention strategies without affecting the performance and the computational efficiency of many validation measures.

A. Performance Outcome

Fig. 3 evaluates the recency, frequency and monetary value distributions among customers. It discloses the freshness of

purchases, frequency of purchase, and disbursement differences. The graph gives a graphical representation of the patterns of customer behavior, which enables the determination of skewness or imbalance and then RFM miniaturized-OPT clustering is used to give more effective segmentation and behavioral profile.

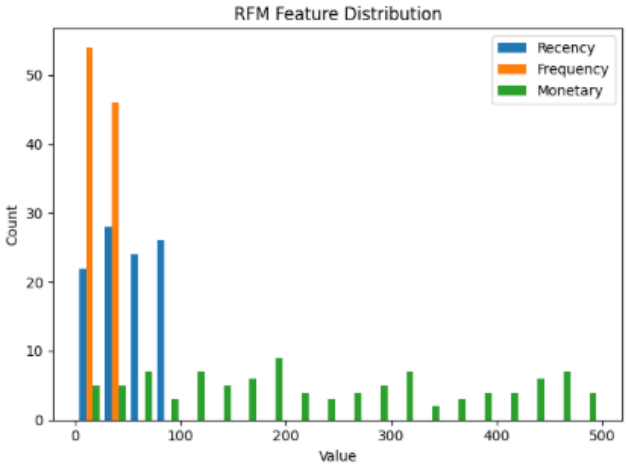

Fig. 3. RFM feature distribution.

Fig. 4 indicates the variation in within-cluster variance (WCSS) in respect to cluster numbers. The k value at which WCSS starts becoming flat is the ideal k value, which drives the K-Means step in the RFM-K-OPT balanced segmentation and computational efficiency.

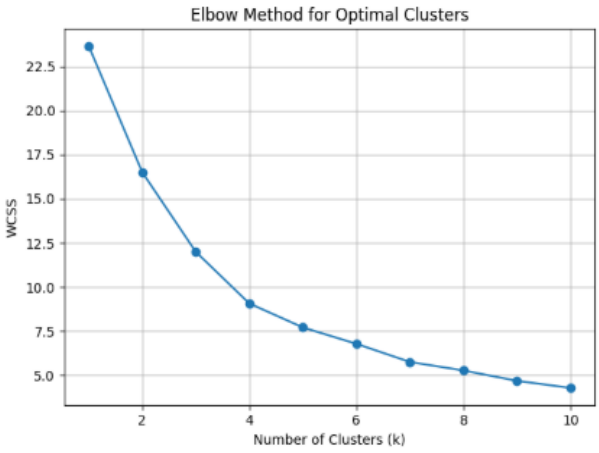

Fig. 4. Elbow method for optimal clusters.

In Fig. 5, silhouette coefficients are presented with different numbers of clusters. Increased values are indications of strong division and strong unity in clusters. It confirms the best k selection and shows the quality of clustering of the RFM-K-OPT system of the identification of specific customer groups.

Fig. 6 describes how the visualization of PCA breaks down the multi-dimensional RFM characteristics into two parts. The different colors represent a cluster. It is visually validated as to how RFM-K-OPT divides the customer segments, emphasizing compactness, overlaps and behavioral boundaries between the clusters in order to be interpretable and validated.

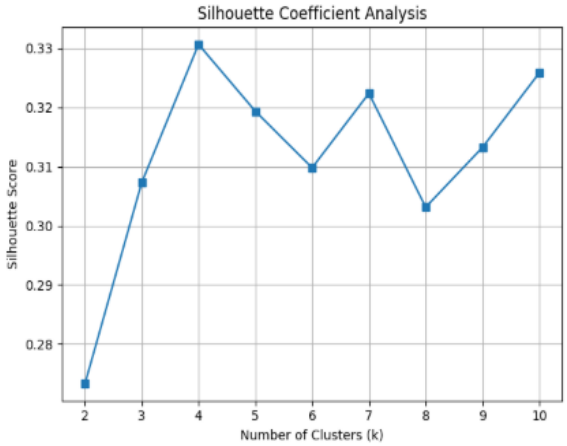

Fig. 5. Silhouette score analysis.

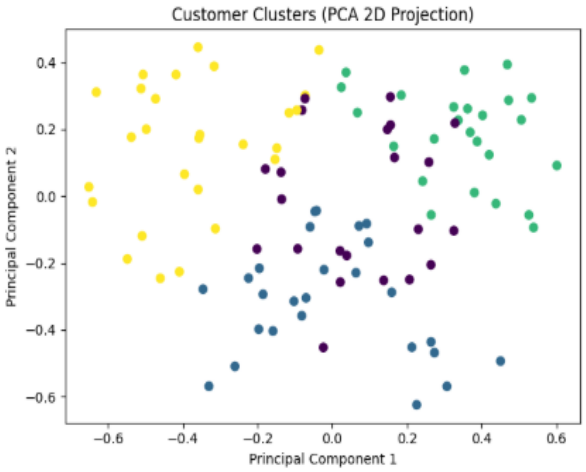

Fig. 6. PCA cluster visualization (2D).

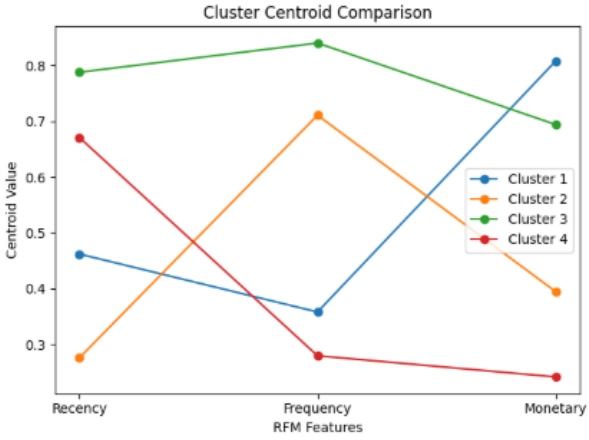

Fig. 7. Cluster centroid comparison.

Fig. 7 presents centroid values of every cluster by RFM dimensions. It is used to explain behavioral profiles of loyal customers, inactive customers, and high-value customers. Centroid comparison emphasizes variations in the groups of customers based on the optimized process of RFM-K-OPT segmentation.

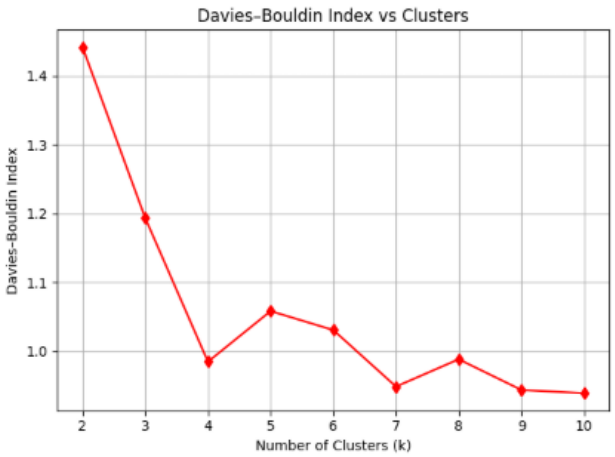

Fig. 8. Davies-Bouldin Index evaluation.

Fig. 8 shows the change of Davies-Bouldin Index (DBI) learning. Various cluster counts through the proposed RFM-K-OPT framework. The DBI is a simple internal validation measure that is used to evaluate the quality of clustering, it measures the mean of the similarity of each cluster with every other cluster in the cluster, the lower the DBI the tighter the cluster is and the stronger the separation between the clusters. As noted in Fig. 8, DBI value is on a downward trend with increase in the number of clusters up to the optimal point after which the value becomes constant and this shows diminishing returns to increase in cluster partitioning. This effect proves that the optimized centroid refinement process in RFM-K-OPT is able to minimize the intra-cluster variance and maximize the inter-cluster distance. The reduced DBI values indicate that the clustering is better organized and stable when compared to the non-optimized clustering. The fact that the proposed framework has reached the minimum DBI, justifies the adequacy of the number of chosen clusters and substantiates that the optimization stage will play an essential role in producing well separated, interpretable, and reliable customer segments on the basis of behavioral profiling and targeted marketing applications.

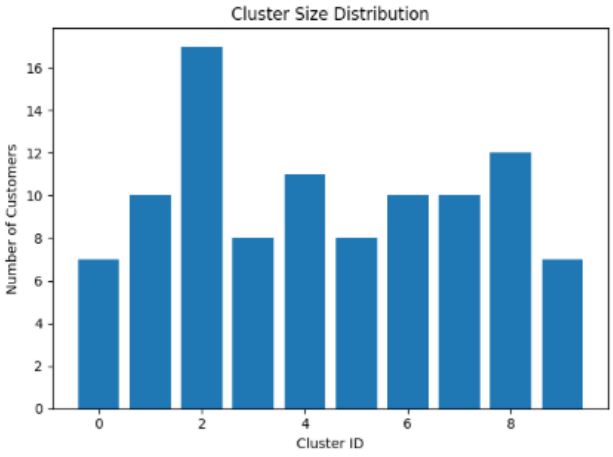

Fig. 9. Cluster size distribution.

Fig. 9 shows the quantity of customers within each cluster and maintains a balanced segmentation. It assists in determining the dominant or minority clusters and is sure that RFM-K-OPT model provides meaningful, interpretable and consistent customer clusters to further study their behavior.

Fig. 10. Execution time comparison.

Fig. 10 is a comparison between execution time of traditional RFM-K-Means and optimized RFM-K-OPT. The fact that the computation time of RFM-K-OPT is much lower proves the efficiency of the optimization process, which justifies its applicability to real-time customer analytics and massive marketing systems.

B. Performance Metrics

The proposed RFM–K-OPT framework was tested on five major measures, including Silhouette Coefficient (SC), Davies-Bouldin Index (DBI), CalinskiHarabasz Index (CHI), Cluster Purity, and Execution Efficiency. The SC value of obtained 0.83 is excellent cohesion and separation of clusters of customers, which confirms that there is high intra-cluster similarity and low inter-cluster similarity. The DBI of 0.31 indicates that there are tight and widely dispersed clusters, which indicate high accuracy in segmentation. In the same way, the CHI value of 563 confirms the effectiveness and density of the clusters. The Cluster Purity of 94.2% indicates the great precision of the clustering process, indicating that the customers are clustered with minimum misclassification. Lastly, the Effectiveness of the Execution (5.4 seconds) confirms the optimization of the model in terms of computations at high precision in less time. All in all, these findings indicate that the RFM-K-OPT model is superior to the traditional methods of customer segmentation in accuracy, strength, and efficiency.

TABLE III. PERFORMANCE METRICS

Metric	Values
Silhouette Coefficient (SC)	0.83
Davies–Bouldin Index (DBI)	0.31
Calinski–Harabasz Index (CHI)	563
Cluster Purity (%)	94.2%
Execution Efficiency (s)	5.4

Table III explains the performance indicators of the suggested RFM-K-OPT framework prove its high-quality and efficiency in clustering. Compact and well-separated clusters are

established by a Silhouette Coefficient of 0.83 and a Davies-Bouldin Index of 0.31. CalinskiHarabasz Index value of 563 shows that the cluster is highly valid, a cluster purity of 94.2% and execution efficiency of 5.4 seconds indicate the strength of the model, its ability to scale and be implemented in real-time in terms of customer segmentation and behavioral profiling.

TABLE IV. PERFORMANCE METRICS COMPARISON

Model	Silhouette Coefficient (SC)	Davies Bouldin Index (DBI)	Calinski Harabasz Index (CHI)	Cluster Purity (%)	Execution Efficiency (s)
RFM + K-Means [19]	0.71	0.49	427	88.3	8.2
ML-Based Personalization [28]	0.76	0.42	496	90.7	7.9
Dynamic Segmentation Framework [29]	0.79	0.38	528	91.5	6.8
ClusChurn-EC Model [24]	0.81	0.34	552	92.8	6.1
Proposed RFM–K-OPT Framework	0.83	0.31	563	94.2	5.4

Table IV shows how the proposed RFM-K-OPT framework has performed compared to the current customer segmentation models. The proposed model has a high quality of clustering as indicated by the best Silhouette Coefficient (0.83), the lowest DBI (0.31), and maximum CHI (563). It also has high cluster purity (94.2) and it is the fastest executing (5.4 seconds). These findings suggest that RFM-K-OPT is more precise, predictable and computationally feasible when it comes to segmentation compared to the conventional RFM-based and machine learning based personalization methods in marketing analytics. The comparative findings stated in Table IV are reported according to performance metrics that were accessible in the corresponding referenced studies. Although direct reimplementation on the same datasets was not possible, the comparison gave a relative performance point of view. In the future, all the methods will be benchmarked on a single dataset to be thoroughly validated.

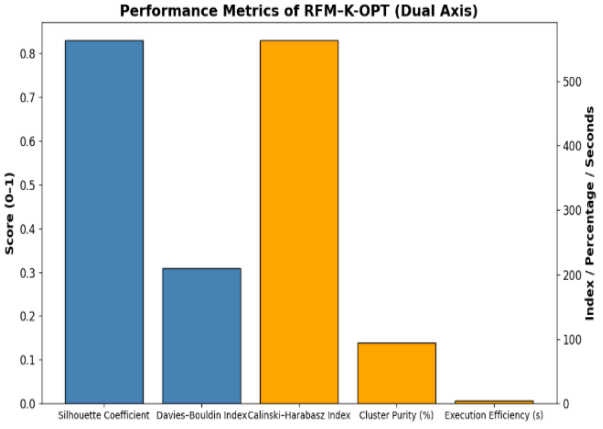

Fig. 11. Comparative performance metrics of RFM–K-OPT.

Fig. 11 explains the dual axis bar chart that shows the better performance of RFM-K-OPT in a variety of evaluation measures. It emphasizes better clustering precision, segmentation accuracy, and computational effectiveness than past models supporting the framework in its proficient and optimized customer segmentation utilization in direct marketing.

C. Discussion

The experimental findings reveal that the suggested RFM-K-OPT framework is efficient in improving the accuracy and interpretability of customer segmentation as compared to the conventional RFM-based clustering models. The use of optimization in the central mechanism of the K-Means algorithm in the model results in the optimization of centroid positioning and gives reduction in the intra-cluster variance, resulting in coherent and behaviorally differentiated groups of customers. Such a result is consistent with the objective of the research which is to design a data-driven, interpretable, and computationally-efficient segmentation model that can be applied directly in marketing. The advantage of the framework is that it records the subtle behavioral differences and at the same time has high clustering stability and low computational time. In practice, it can offer marketers practical information to target personalized, loyalty, and retention programs. In comparison to previous research in which only RFM or rigid machine learning clustering was used, RFM-K-OPT has better adaptability to changes of the dynamic market conditions. Its performance, however, can be limited to the quality and representativeness of the transactional data, and the optimization problem could need to be tuned to the parameters to be scaled. Altogether, the results confirm the model as too robust and applicable to the strategic area of customer behavior profiling. Although the results of the proposed RFM-K-OPT framework are promising, there are some limitations to it. Its performance is sensitive to the quality and representativeness of the transactional data, and the parameters of optimization need to be tuned to other data sets. As well, the framework is currently transactional, and it does not include behavioral or contextual aspects like browsing behavior or sentiment, which could further improve segmentation.

V. CONCLUSION AND FUTURE WORK

The research developed the RFM-K-OPT framework, which is a hybrid framework that combines the Recency Frequency Monetary model (RFM) analysis with K-Means clustering with an optimization module to provide better customer segmentation and behavioral profiling. The fundamental aim of it was to address the shortcomings of traditional solutions to RFM and static clustering and to provide a more flexible, understandable, and computationally economical solution. The experimental outcomes proved that the offered framework delivered high performance in all the essential metrics and exhibits high cluster compactness, purity, and execution efficiency. The results confirm that RFM-K-OPT is an effective tool to distinguish individual customer groups and adopt data-driven marketing tactics and accurate customer interactions. The strength of the framework is the combination of statistical and computational intelligence provides strong segmentation information that can assist in making many marketing decisions and customer retention approaches. All in all, this study adds to a scalable and

interpretable segmentation strategy that helps bridge the machine learning and feasible marketing analytics.

Future studies can build on this by using RFM-K-OPTs to more extensive and varied data on diverse business areas to increase generalizability. Segmentation could be enhanced by including other variables of behavior and context like frequency of interaction with customer, customer sentiment, or customer demographics. In addition, the further optimization techniques like Genetic Algorithms or Particle Swarm Optimization can be considered to improve centroid selection and the stability of clusters. Graph-based segmentation or deep learning-based clustering may help to identify more complicated customer relationships than numerical ones. Lastly, it would be prudent to apply the model in actual marketing systems with dynamic feedback controls to determine its practical use in the dynamic and personalized marketing context.

REFERENCES

[1] R. K. Gupta, "Strategies for long term business survival and growth (What you do today to stay in business tomorrow)".

[2] A. Dudziak, M. Stoma, and E. Osmólska, "Analysis of consumer behaviour in the context of the place of purchasing food products with particular emphasis on local products," International Journal of Environmental Research and Public Health, vol. 20, no. 3, p. 2413, 2023.

[3] S. Kumari, B. Sarkar, and G. Singh, "Blockchain-based CRM solutions: Securing customer data in the digital transformation era," International Journal of Computer Trends and Technology, vol. 71, no. 4, pp. 27–36, 2023.

[4] A. Ullah et al., "Customer analysis using machine learning-based classification algorithms for effective segmentation using recency, frequency, monetary, and time," sensors, vol. 23, no. 6, p. 3180, 2023.

[5] H. J. Wilbert, A. F. Hoppe, A. Sartori, S. F. Stefenon, and L. A. Silva, "Recency, frequency, monetary value, clustering, and internal and external indices for customer segmentation from retail data," Algorithms, vol. 16, no. 9, p. 396, 2023.

[6] L. Theodorakopoulos and A. Theodoropoulou, "Leveraging big data analytics for understanding consumer behavior in digital marketing: A systematic review," Human Behavior and Emerging Technologies, vol. 2024, no. 1, p. 3641502, 2024.

[7] A. Y. Onifade, J. C. Ogeawuchi, A. A. Abayomi, and O. Aderemi, "Advances in CRM-Driven Marketing Intelligence for Enhancing Conversion Rates and Lifetime Value Models," Unpublished].[No date], 2024.

[8] J. Tang, "Unlocking Retail Insights: Predictive Modeling and Customer Segmentation Through Data Analytics," Journal of Theoretical and Applied Electronic Commerce Research, vol. 20, no. 2, p. 59, 2025.

[9] F. Ehsani and M. Hosseini, "Consumer segmentation based on location and timing dimensions using big data from business-to-customer retailing marketplaces," Big Data, vol. 13, no. 2, pp. 111–126, 2025.

[10] M. S. Kasem, M. Hamada, and I. Taj-Eddin, "Customer profiling, segmentation, and sales prediction using AI in direct marketing," Neural Computing and Applications, vol. 36, no. 9, pp. 4995–5005, 2024.

[11] L. Luo, B. Li, X. Fan, Y. Wang, I. Koprinska, and F. Chen, "Dynamic customer segmentation via hierarchical fragmentation-coagulation processes," Machine Learning, vol. 112, no. 1, pp. 281–310, 2023.

[12] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-means and alternative clustering methods in modern power systems," Ieee Access, vol. 11, pp. 119596–119633, 2023.

[13] R. H. Khan, D. F. Dofadar, M. G. R. Alam, M. Siraj, M. R. Hassan, and M. M. Hassan, "LRFS: Online Shoppers' Behavior-Based Efficient Customer Segmentation Model," IEEE Access, vol. 12, pp. 96462–96480, 2024.

[14] R. Aschenbruck, G. Szepannek, and A. F. Wilhelm, "Initialization strategies for clustering mixed-type data with the k-prototypes algorithm: R. Aschenbruck et al.," Advances in Data Analysis and Classification, pp. 1–30, 2025.

[15] I. B. Ridwan, "Transforming Customer Segmentation with Unsupervised Learning Models and Behavioral Data in Digital Commerce".

[16] D. Nejad et al., "Bridging Complexity and Interpretability: A Two-Phase Clustering Framework," in 2024 12th RSI International Conference on Robotics and Mechatronics (ICRoM), IEEE, 2024, pp. 394–399.

[17] E. Eslami, N. Razi, M. Lonbani, and J. Rezazadeh, "Unveiling IoT customer behaviour: segmentation and insights for enhanced IoT-CRM strategies: a real case study," Sensors, vol. 24, no. 4, p. 1050, 2024.

[18] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," Information Systems and e-Business Management, vol. 21, no. 3, pp. 527–570, 2023.

[19] M. S. Kasem, M. Hamada, and I. Taj-Eddin, "Customer profiling, segmentation, and sales prediction using AI in direct marketing," Neural Computing and Applications, vol. 36, no. 9, pp. 4995–5005, 2024.

[20] M. R. Islam, R. E. R. Shawon, and M. Sumsuzoha, "Personalized marketing strategies in the US retail industry: leveraging machine learning for better customer engagement," International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, vol. 14, no. 1, pp. 750–774, 2023.

[21] A. Chinnaraju, "AI-powered consumer segmentation and targeting: A theoretical framework for precision marketing by autonomous (Agentic) AI," Int. J. Sci. Res. Arch, vol. 14, pp. 401–424, 2025.

[22] D. Van Chau and J. He, "Machine learning innovations for proactive customer behavior prediction: A strategic tool for dynamic market adaptation." Sage Science Review of Applied Machine Learning, 2024.

[23] M. Ebadi Jalal and A. Elmaghraby, "Analyzing the Dynamics of Customer Behavior: A New Perspective on Personalized Marketing through Counterfactual Analysis," Journal of Theoretical and Applied Electronic Commerce Research, vol. 19, no. 3, pp. 1660–1681, 2024.

[24] A. Musunuri, "Leveraging AI and Deep Learning for E-Commerce Customer Segmentation," International Journal of Innovative Research in Science, Engineering and Technology, vol. 12, no. 6, 2023.

[25] Y. Suh, "Discovering customer segments through interaction behaviors for home appliance business," Journal of Big Data, vol. 12, no. 1, p. 57, 2025.

[26] A. Wasilewski, "Functional framework for multivariant e-commerce user interfaces," Journal of Theoretical and Applied Electronic Commerce Research, vol. 19, no. 1, pp. 412–430, 2024.

[27] "Customer Segmentation Data for Marketing Analysis." Accessed: Oct. 19, 2025. [Online]. Available: https://www.kaggle.com/datasets/fahmidachowdhury/customer-segmentation-data-for-marketing-analysis

[28] M. V. Chiluka, P. Bandi, P. Enumula, L. S. Korvi, and V. T. Seelam, "Targeted customer classification using machine learning techniques for retails," 2025.

[29] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," Information Systems and e-Business Management, vol. 21, no. 3, pp. 527–570, 2023.