

Customer Segmentation and Purchase Behavior Analysis Using Hierarchical Agglomerative Clustering and Machine Learning Techniques

 Lutta David,  Gichoya David,  Korongo Julia and  Gudu Jael

Department of Information Technology, School of Information Science,
Moi University, P.O. Box 3900-30100, Kesses, Eldoret, Kenya

Abstract

The heterogeneity of consumer behavior in digital marketplaces poses a significant challenge for firms seeking to optimize marketing resource allocation and predictive accuracy. This study proposes and validates a hybrid analytical framework that integrates Hierarchical Agglomerative Clustering (HAC) with supervised machine learning techniques to segment customers and predict purchase behavior. Grounded in Machine Learning Theory (MLT) and Computational Learning Theory (COLT), and operationalized through the Cross-Industry Standard Process for Data Mining (CRISP-DM), the research employs a three-stage pipeline: (1) segmentation via HAC using Ward's linkage and Euclidean distance, (2) segment-specific predictive modeling using Choice Model Trees (CMT), Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Logistic Regression, and (3) comprehensive evaluation using accuracy, precision, recall, F1-score, and ROC-AUC. Using a dataset of 50,000 customer records with demographic, behavioral, and transactional features, the HAC algorithm identified four distinct segments: High-Frequency Power Users, Price-Sensitive Deal Hunters, Low-Engagement Browsers, and High-Value Efficiency Shoppers. The hybrid segment-specific approach achieved a predictive accuracy of 84.33%, substantially outperforming global modeling benchmarks. The Choice Model Trees further provided interpretable decision rules such as the predictive thresholds of five product views and 200 seconds of browsing time enabling actionable managerial insights. The findings demonstrate that addressing customer heterogeneity through hierarchical segmentation prior to predictive modeling yields significant performance uplifts and enhances the interpretability of machine learning outputs for strategic marketing deployment.

Keywords: Customer segmentation, hierarchical agglomerative clustering, purchase behavior prediction, machine learning, e-commerce analytics

Correspondence: davelutta@gmail.com

Copyright © 2026 Lutta et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Funding: The author received no financial support for the research, authorship and/or publication of this article

Data Availability Statement: The authors confirm that the data supporting the findings of this study are available within the article [and/or] its supplementary materials or upon reasonable request.

Competing interests: The authors declare no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Introduction

The rapid growth of digital technologies has transformed how organizations manage customer relationships, generating vast volumes of transactional, demographic, and behavioral data through e-commerce platforms, loyalty programmes, and digital payment systems (Gomes & Meisen, 2023; Salminen et al., 2023). Yet data abundance alone does not guarantee effective decision-making; firms require analytical techniques capable of converting complex datasets into actionable insights for targeted marketing and customer retention. Customer segmentation is the division of heterogeneous populations into homogeneous groups based on shared characteristics or behaviors has consequently emerged as a critical analytical approach (Gomes & Meisen, 2023; Salminen et al., 2023). While traditional methods have relied heavily on demographic variables, contemporary analytics increasingly emphasizes behavioral and transactional indicators because customers with similar demographic profiles often exhibit divergent purchasing patterns, and vice versa (Luo et al., 2023). The Recency, Frequency, and Monetary (RFM) framework has long served as a

foundational tool for evaluating purchase behavior, but conventional RFM models are inherently limited by their narrow focus on three transactional dimensions and may overlook broader customer characteristics that shape segment profiles (Ho et al., 2023).

Machine learning techniques, particularly unsupervised clustering algorithms, offer a pathway beyond manually defined categories by identifying naturally occurring customer groups within multidimensional data (Gomes & Meisen, 2023; Salminen et al., 2023). Among these, Hierarchical Agglomerative Clustering (HAC) provides distinct advantages for exploratory segmentation: it initially treats each observation as an individual cluster and progressively merges similar cases, producing a dendrogram that reveals nested customer relationships without requiring a predetermined number of segments (Rungruang et al., 2023; Luo et al., 2023). Recent applications in retail and e-commerce contexts have demonstrated the efficacy of hierarchical approaches for distinguishing valuable customers, occasional buyers, and disengaged users (John et al., 2023; Rungruang et al., 2023). Nevertheless, methodological complexity persists; the

diversity of available algorithms, distance measures, and evaluation criteria can yield substantially different segment solutions, raising questions about which approaches produce the most meaningful and interpretable groups (Salminen et al., 2023). Moreover, many advanced predictive models function as black boxes, limiting managerial usability, while insufficient integration of segmentation with downstream prediction means most models treat heterogeneous customer datasets as uniform populations (Joung & Kim, 2023). These gaps including customer heterogeneity, limited interpretability, and the lack of hybrid frameworks that combine behavioral segmentation with choice-based prediction prevent the development of precise, actionable purchase prediction systems. Accordingly, this study applies Hierarchical Agglomerative Clustering integrated with machine learning techniques to identify distinct customer segments and analyse the purchase behaviours that differentiate them, thereby contributing to more targeted, evidence-based customer management strategies.

Theoretical Review

The present study is anchored in five complementary theoretical perspectives: Market Segmentation Theory, Data Mining and Knowledge Discovery Theory, Discrete Choice Theory, Machine Learning Theory, and Computational Learning Theory. Together, these frameworks address the interconnected dimensions of customer segmentation and purchase behaviour analysis explaining why heterogeneous populations should be divided into homogeneous groups, how meaningful patterns can be extracted from complex datasets, how customers make purchasing decisions among competing alternatives, and how predictive models

can be optimized to manage the trade-offs between exploration of new data and exploitation of existing patterns.

Market Segmentation Theory provides the foundational rationale for dividing heterogeneous customers into relatively homogeneous groups. Originating with Smith (1956), the theory posits that markets consist of consumers who differ in needs, preferences, and behavioural patterns, and that organizations achieve superior targeting by developing differentiated strategies for specific subgroups rather than applying uniform approaches to entire populations. Wedel and Kamakura (2000) emphasized that effective segments must be sufficiently homogeneous internally and sufficiently distinct externally to justify differentiated marketing strategies. In contemporary data-rich environments, this principle has become increasingly relevant as businesses move beyond demographic classification toward behavioural and algorithmic segmentation (Gomes & Meisen, 2023; Salminen et al., 2023). The theory justifies the application of Hierarchical Agglomerative Clustering (HAC) as a computational operationalization of segmentation, wherein customers are progressively grouped according to similarity without requiring predetermined categories. However, because traditional segmentation may overlook hidden patterns within large datasets and treat customer groups as static despite evolving preferences (Luo et al., 2023), the theory must be complemented by computational frameworks capable of discovering segments directly from data.

Data Mining and Knowledge Discovery Theory supplies this computational foundation, explaining how valid, novel, and actionable patterns can be extracted from large and complex datasets. Fayyad et al. (1996)

conceptualized knowledge discovery as a systematic process involving data selection, preprocessing, pattern identification, and interpretation. This perspective is essential for customer analytics because transactional databases contain volumes of information that conventional descriptive methods cannot effectively analyse. Within this framework, machine learning and clustering algorithms serve as critical tools for uncovering naturally occurring customer groups that are not known in advance. Hierarchical Agglomerative Clustering is particularly consistent with this perspective because it produces a hierarchical structure that reveals nested relationships among customers, enabling researchers to explore data-driven segmentations at different levels of aggregation (Salminen et al., 2023; John et al., 2023). The theory also emphasizes that discovered patterns must be interpretable and connected to practical contexts; statistically distinct clusters become meaningful only when their behavioural characteristics can inform marketing or customer relationship strategies.

Discrete Choice Theory provides the behavioural mechanism underlying purchase decisions. Rooted in random utility theory and extensively developed by McFadden (1974), the theory assumes that individuals select the alternative yielding the greatest utility from a finite set of options, where utility comprises an observable systematic component and an unobservable random component. This framework is highly relevant to purchase behaviour analysis because customers facing identical alternatives may make different choices based on varying sensitivities to price, quality, brand, convenience, or promotional incentives. Traditional choice models often estimate average preferences across entire populations, potentially masking

heterogeneity in decision-making processes. By integrating segmentation with choice modelling, the study can examine whether the determinants of purchase decisions differ across customer groups. Choice Model Trees operationalize this integration by combining tree-based segmentation with discrete choice models, allowing utility functions and decision pathways to be estimated separately for each homogeneous subgroup identified through HAC.

Machine Learning Theory (MLT) and Computational Learning Theory (COLT) provide the mathematical and analytical foundations for building and optimizing the predictive components of the study. MLT furnishes the principles for constructing models that can handle the complexity of high-dimensional consumer data, while COLT offers the theoretical tools required to optimize these models specifically managing the bias-variance trade-off and the balance between exploring new customer data patterns and exploiting established behavioural regularities. These theories ensure that the segmentation and prediction layers are not merely algorithmic exercises but are grounded in formal learning principles that guarantee generalizability and predictive validity. Together, the five theoretical perspectives form a coherent intellectual architecture: Market Segmentation Theory explains why differentiation is necessary; Data Mining and Knowledge Discovery Theory explains how segments can be discovered from raw data; Discrete Choice Theory explains the behavioural mechanisms driving purchase decisions within segments; and Machine Learning Theory alongside Computational Learning Theory ensures that the predictive models are mathematically sound, optimally configured, and capable

of generating actionable managerial insights.

Empirical Review

Machine Learning Techniques for Customer Segmentation

The application of machine learning techniques to customer segmentation has received increasing scholarly attention as organizations seek to extract actionable insights from large and complex customer datasets. Machine learning-based segmentation enables organizations to identify patterns within customer data that may not be readily observable through conventional statistical approaches. In particular, unsupervised learning techniques such as clustering are increasingly used to identify groups of customers with similar demographic, transactional, and behavioural characteristics. Gomes and Meisen (2023) conducted a review of customer segmentation methods for personalized customer targeting in e-commerce. The objective was to examine existing segmentation approaches and assess their suitability for personalized customer targeting in e-commerce environments. The study reviewed existing literature on customer segmentation methods and examined how different analytical techniques could be used to identify customer groups. The findings demonstrated the increasing importance of data-driven and machine learning-based approaches in identifying heterogeneous customer groups and supporting personalized targeting. The study concluded that machine learning and algorithmic segmentation offer substantial opportunities for improving customer targeting, although the choice of method should be guided by the characteristics of the data and the intended managerial application.

Similarly, Salminen et al. (2023) systematically reviewed the application of algorithms in customer and user segmentation. The objective was to examine how algorithmic approaches have been applied to segment customers and identify methodological gaps for future research. The researchers conducted a systematic review of 172 relevant studies and examined the algorithms, data types, segmentation approaches, and evaluation metrics used across the literature. The findings revealed considerable diversity in the algorithms employed for customer segmentation, with clustering approaches representing a major component of algorithmic segmentation. The review also identified substantial variation in the evaluation metrics used to assess segmentation quality. The authors concluded that algorithmic customer segmentation provides considerable potential for identifying meaningful customer groups, but emphasized the need for greater methodological consistency and stronger evaluation of the quality and practical usefulness of identified segments.

John et al. (2023) examined the application of clustering algorithms for customer segmentation in the UK retail market. The objective of the study was to explore the effectiveness of clustering algorithms in identifying meaningful customer groups from retail transaction data. The researchers used an online retail dataset and applied clustering techniques to examine customer purchasing patterns and generate customer segments. The findings demonstrated that clustering algorithms could effectively identify distinct customer groups with different purchasing characteristics. The study further showed that the resulting segments could provide useful information for developing differentiated

marketing strategies. The authors concluded that machine learning-based clustering provides an effective approach to retail customer segmentation, particularly where businesses have access to large volumes of transactional data.

Joung and Kim (2023) investigated an interpretable machine learning approach to customer segmentation for new product development using online product reviews. The objective was to identify customer segments based on customer preferences for product attributes and to develop an interpretable approach that could provide actionable information for product development. The study analysed online product reviews and applied machine learning techniques to identify distinct customer groups according to their product-feature preferences. The results demonstrated that machine learning could identify meaningful customer segments based on differences in customer preferences and provide insights into the product characteristics valued by different groups. The study concluded that interpretable machine learning can improve the practical usefulness of customer segmentation by enabling managers to understand why customers are classified into particular segments.

Hierarchical Agglomerative Clustering and Customer Segmentation

Hierarchical Agglomerative Clustering (HAC) has emerged as an important approach for customer segmentation because of its ability to reveal hierarchical relationships among observations. Unlike clustering approaches that require researchers to specify the final number of clusters in advance, hierarchical agglomerative clustering progressively combines observations or clusters according to their degree of similarity. The resulting

hierarchical structure can be represented through a dendrogram, allowing researchers to examine alternative numbers and levels of customer segments. This characteristic makes HAC particularly relevant for exploratory customer analysis where the underlying structure of customer groups is initially unknown.

Rungruang et al. (2023) examined RFM-based customer segmentation using a hierarchical approach combined with Formal Concept Analysis (FCA). The objective of the study was to develop an approach for segmenting customers according to their recency, frequency, and monetary characteristics while improving the identification of relationships among customer groups. The researchers used transactional customer data and applied RFM analysis before employing a hierarchical segmentation approach and FCA. The findings demonstrated that the hierarchical approach could identify distinct customer groups based on differences in purchasing behaviour and customer value. The results further indicated that customer segments could be interpreted according to their purchasing characteristics and subsequently linked to differentiated marketing strategies. The study concluded that hierarchical customer segmentation provides a useful approach for identifying and interpreting customer groups, particularly when combined with behavioural measures such as RFM.

Luo et al. (2023) investigated dynamic customer segmentation using hierarchical fragmentation-coagulation processes. The objective was to develop a framework capable of capturing changes in customer segments over time rather than treating customer groups as static entities. The researchers developed a hierarchical modelling approach that allowed customer segments to fragment

and combine as customer behaviours changed. The findings demonstrated that customer segmentation could be understood as a dynamic process in which customers may move between groups over time. The study concluded that hierarchical approaches can provide valuable insights into the evolving nature of customer behaviour and may be more informative than static segmentation models in environments characterized by changing customer preferences and purchasing patterns.

John et al. (2023), although examining several clustering algorithms in a retail context, provided additional evidence of the relevance of hierarchical approaches for customer segmentation. The study used retail transaction data to identify customer groups based on purchasing patterns and compared clustering approaches for determining meaningful customer segments. The findings indicated that clustering methods could distinguish customers according to their purchasing characteristics and that the choice of clustering technique influenced the structure and interpretation of the resulting segments. The study concluded that organizations should carefully select clustering methods based on the characteristics of their customer data and the intended purpose of segmentation.

RFM Analysis and Customer Purchase Behaviour

Customer purchase behaviour analysis is central to understanding differences in customer value and engagement. One of the most widely used approaches for analysing transactional purchase behaviour is the Recency, Frequency, and Monetary (RFM) framework. RFM analysis evaluates customers according to the recency of their most recent purchase, the frequency of their transactions, and the

monetary value of their purchases. These dimensions provide an accessible means of differentiating customers according to their transactional behaviour and are frequently used as inputs for clustering-based segmentation.

Ho et al. (2023) examined an extended RFM model for analysing customer behaviour and demographic characteristics in the retail industry. The objective was to enhance conventional RFM analysis by incorporating additional customer demographic information to obtain a more comprehensive understanding of customer characteristics. The researchers used retail customer data and analysed customers according to transactional and demographic characteristics. The findings demonstrated that combining RFM measures with demographic information could provide more detailed customer profiles than relying exclusively on the traditional three-dimensional RFM model. The study concluded that an extended RFM framework can improve customer segmentation by integrating behavioural and demographic characteristics, thereby providing organizations with more comprehensive information for customer targeting.

Rungruang et al. (2023) also applied RFM measures in a hierarchical customer segmentation framework. The objective was to classify customers according to purchasing behaviour and identify groups that could be targeted using differentiated strategies. Using transactional data, the researchers calculated RFM measures and subsequently applied hierarchical segmentation techniques. The findings revealed meaningful differences among customer groups in terms of purchase recency, purchase frequency, and monetary value. High-value customers could be differentiated from less active or lower-value customers, providing a basis

for differentiated marketing strategies. The study concluded that RFM-based hierarchical segmentation is useful for identifying customer groups according to purchasing behaviour and supporting targeted customer relationship management.

John et al. (2023) examined retail transaction data to explore customer purchasing patterns through clustering techniques. Although the study was not exclusively based on the RFM framework, its findings reinforce the importance of transactional variables in distinguishing customer groups. The researchers demonstrated that customer purchase data could be transformed into meaningful behavioural profiles using clustering algorithms. The resulting segments reflected differences in customer engagement and purchasing activity, indicating that transaction-based characteristics provide valuable inputs for data-driven customer segmentation.

Customer Segmentation and Personalized Marketing and Customer Relationship Management

Customer segmentation is ultimately intended to support managerial decision-making by enabling organizations to develop differentiated strategies for customers with different needs and behaviours. Effective segmentation can facilitate personalized marketing, targeted promotions, customer retention, product recommendations, and improved allocation of marketing resources. Rather than treating all customers uniformly, organizations can design strategies that correspond to the characteristics and value of specific customer groups.

Gomes and Meisen (2023) examined customer segmentation methods in the context of personalized customer targeting in e-commerce. The objective was to assess how

segmentation approaches could support the identification of customer groups and facilitate personalized marketing interventions. The researchers reviewed segmentation methods used in e-commerce and examined their implications for targeted customer management. The findings indicated that data-driven segmentation can help businesses identify heterogeneous customer groups and develop personalized targeting strategies. The study concluded that effective customer segmentation is an important foundation for personalized marketing, although organizations must select appropriate segmentation approaches according to their data and strategic objectives.

Joung and Kim (2023) examined customer segmentation for new product development using interpretable machine learning and online customer reviews. The objective was to identify customer groups with different product-feature preferences and demonstrate how segmentation could inform product development. The study analysed customer-generated review data and used machine learning methods to identify preference-based customer segments. The findings revealed that different customer groups valued different product attributes, suggesting that organizations could use customer segmentation to develop more targeted product strategies. The study concluded that integrating customer segmentation with interpretable machine learning can provide actionable insights for product development and marketing decisions.

Rungruang et al. (2023) investigated customer segmentation using hierarchical RFM analysis and demonstrated how differentiated customer groups could inform marketing strategies. The study found that customers with different levels of recency, frequency, and monetary value

required different approaches to customer engagement. High-value and active customers, for instance, could be managed differently from customers with low purchase frequency or declining engagement. The study concluded that customer segmentation can improve marketing effectiveness by enabling organizations to align strategies with the characteristics and value of specific customer groups.

Luo et al. (2023) extended the discussion by demonstrating that customer segments may change dynamically over time. Their study showed that customers can move between different segments as their behaviours and relationships with businesses change. This finding is particularly important for customer relationship management because it suggests that customer segmentation should not necessarily be viewed as a one-time classification exercise. Instead, organizations may need to continuously monitor customer behaviour and update customer segments to reflect changing purchasing patterns.

Research Gap

The empirical literature reviewed demonstrates the increasing application of machine learning and clustering techniques in customer segmentation and purchase behaviour analysis. Four major themes emerge from the reviewed studies. First, machine learning techniques have demonstrated considerable potential for identifying heterogeneous customer groups from large datasets (Gomes & Meisen, 2023; John et al., 2023; Salminen et al., 2023). Second, hierarchical clustering approaches can reveal relationships and similarities among customers and provide flexible structures for identifying customer segments (Luo et al., 2023; Rungruang et al., 2023). Third,

transactional measures, particularly RFM variables, provide important indicators for distinguishing customers according to their purchase behaviour and value (Ho et al., 2023; Rungruang et al., 2023). Fourth, customer segmentation can support personalized marketing, product development, customer retention, and customer relationship management by enabling organizations to differentiate strategies across customer groups (Gomes & Meisen, 2023; Joung & Kim, 2023).

Despite these contributions, several gaps remain. Existing studies have frequently focused on specific techniques or application areas, including RFM-based segmentation, e-commerce personalization, retail clustering, and product preference analysis. In addition, considerable methodological diversity exists in the algorithms and evaluation approaches used in customer segmentation research (Salminen et al., 2023). Some studies have also concentrated on identifying customer clusters without sufficiently examining the detailed purchase behaviours that distinguish the resulting groups. Furthermore, although hierarchical approaches have demonstrated potential for customer segmentation, there remains a need for further empirical application of Hierarchical Agglomerative Clustering integrated with machine learning techniques to simultaneously identify customer segments and provide deeper insights into their purchase behaviour.

The present study seeks to address this gap by applying Hierarchical Agglomerative Clustering and machine learning techniques to customer data. Specifically, the study aims to identify distinct customer segments, examine the purchase behaviours associated with each segment, and demonstrate how the resulting insights can contribute to data-

driven customer management and marketing decisions. This approach is expected to extend existing empirical research by combining hierarchical segmentation with machine learning-based analysis and by placing greater emphasis on the behavioural interpretation and practical relevance of identified customer segments.

Conceptual Framework: The CRISP-DM Methodology

To ensure a systematic and industry-standard approach, this study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) as the overarching methodological framework. CRISP-DM, originally developed by Chapman et al. (2000), is widely regarded as the de facto standard for data mining projects due to its reliability, structured nature, and flexibility across research domains (Sivakami, 2018). The methodology partitions the research process into six interconnected phases, each contributing to the development and refinement of the hybrid predictive model:

1. **Business Understanding:** This initial phase involves defining the concrete objectives of the research. In the context of e-commerce, the primary goal is a binary classification task identifying whether a browsing session will culminate in a "Purchase" or "No Purchase." This phase ensures that the technical development of the model aligns with the strategic goal of improving marketing efficiency.
2. **Data Understanding:** Descriptive statistics and visualization techniques are employed to explore the raw dataset. This phase is crucial for identifying variables such as session duration, product views, and user

demographics. By visualizing data distributions, the researcher can identify patterns or anomalies before moving into technical preparation.

3. **Data Preparation:** Recognized as one of the most resource-intensive stages, data preparation involves selecting relevant variables, cleaning the data (handling missing values and outliers), and transforming it (normalization and encoding). These tasks are essential to ensure the HAC algorithm can calculate Euclidean distances accurately without being skewed by raw feature scales.
4. **Modeling:** The core of the hybrid architecture is constructed during this phase. The process is split into two steps: (a) segmentation, applying HAC to group heterogeneous data into distinct clusters; and (b) prediction, training Choice Model Trees and comparative models on each segment.
5. **Evaluation:** Models are not judged on accuracy alone. This study employs a comprehensive evaluation layer including precision, recall, F1-score, and ROC-AUC. A critical component of this phase is confusion matrix analysis, which allows the researcher to quantify and minimize false positives and false negatives, ensuring the model is cost-effective for business deployment.
6. **Deployment:** The final phase involves the integration of the validated model into a decision support system. Effective deployment relies on the model's ability to provide interpretable rules—such as the specific

engagement thresholds identified by the CMT allowing marketing managers to automate personalized customer interventions. Through systematically progressing through these six phases, CRISP-DM ensures that the theoretical objectives of customer segmentation and choice modelling are translated into a reproducible, end-to-end analytical pipeline capable of generating both statistical validity and managerial actionability.

Methodology

Research Design

A quantitative research design was employed, involving data preparation, clustering, predictive modeling, and model evaluation. The study utilized a hybrid three-layer architecture comprising a segmentation layer (HAC), a prediction layer (CMT and

comparative models), and an evaluation layer (performance metrics and cross-validation). All computational procedures were implemented following the CRISP-DM protocol to ensure replicability and alignment with business objectives.

Dataset Description

The study utilized a comprehensive dataset including demographic, behavioral, and transactional variables. The dataset contains 50,000 customer records with the following variables: Age, Gender (0: Female, 1: Male), Annual Income, Number of Purchases, Product Category (0: Electronics, 1: Clothing, 2: Home Goods, 3: Beauty, 4: Sports), Time Spent on Website (minutes), Loyalty Program membership (1: Yes, 2: No), Discounts Available (0–5), and Purchase Status as the binary target variable (0: No, 1: Yes). The target variable distribution is 48% No Purchase and 52% Purchase, indicating a relatively balanced dataset.

Table 1. Descriptive Statistics of the Dataset

Variable	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
Age	50,000	44.30	15.54	18	31	45	57	70
Gender	50,000	0.50	0.50	0	0	1	1	1
Annual Income	50,000	84,249.16	37,629.49	20,001.51	53,028.98	83,699.58	117,167.77	149,785.18
Number of Purchases	50,000	10.42	5.89	0	5	11	15	20
Product Category	50,000	2.01	1.43	0	1	2	3	4

Data Preprocessing

Data preprocessing followed established protocols to ensure analytical validity. Missing values were addressed using mean or median imputation

depending on the distributional properties of each variable. Numerical data were standardized using min-max and z-score normalization methods to ensure uniformity across features,

thereby preventing variables with larger magnitudes from dominating distance calculations in the clustering phase. Categorical variables were transformed via one-hot encoding to render them compatible with machine learning algorithms. The processed data were divided into an 80% training set and a 20% testing set to facilitate unbiased model evaluation.

Segmentation Layer: Hierarchical Agglomerative Clustering

The first layer of the hybrid model focuses on grouping consumers into homogeneous segments. HAC was implemented as a bottom-up technique that iteratively merges individual observations into clusters based on similarity. The study utilized the Euclidean distance metric and Ward's linkage criterion, the latter of which minimizes the variance increase following cluster

fusion and is particularly suited to generating balanced, compact segments for marketing applications.

To determine the optimal number of clusters and ensure validity, the researchers employed multiple validation techniques: dendrogram visualization to identify natural cut-off thresholds, the Silhouette score to assess cluster cohesion and separation, the Dunn index to quantify cluster compactness relative to inter-cluster distance, and inconsistency coefficients to evaluate the hierarchical structure. The resulting clusters corresponded to customer segments with distinct behavioral and demographic characteristics.

The resulting clusters corresponded to customer segments with distinct behavioral and demographic characteristics. The hierarchical merging process is illustrated in Figure 1.

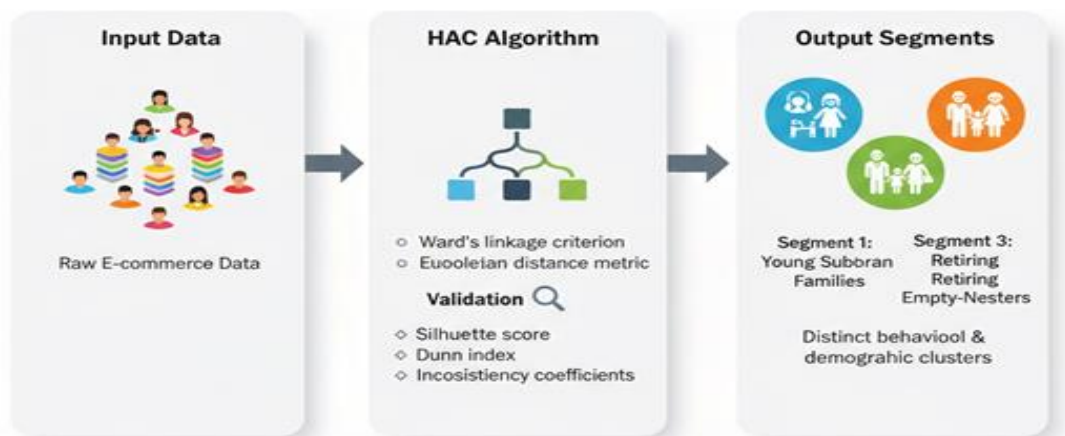


Figure 1: HAC Clustering Process.

Adapted from Kim and Lee (2018).

Following cluster identification, predictive models were applied to each segment to forecast purchase behavior.

Prediction Layer: Choice Model Trees and Comparative Models

Tree growth was managed using Bayesian Information Criterion (BIC)

pruning to prevent overfitting and ensure generalizability. The decision pathway structure of the CMT is illustrated in Figure 2.

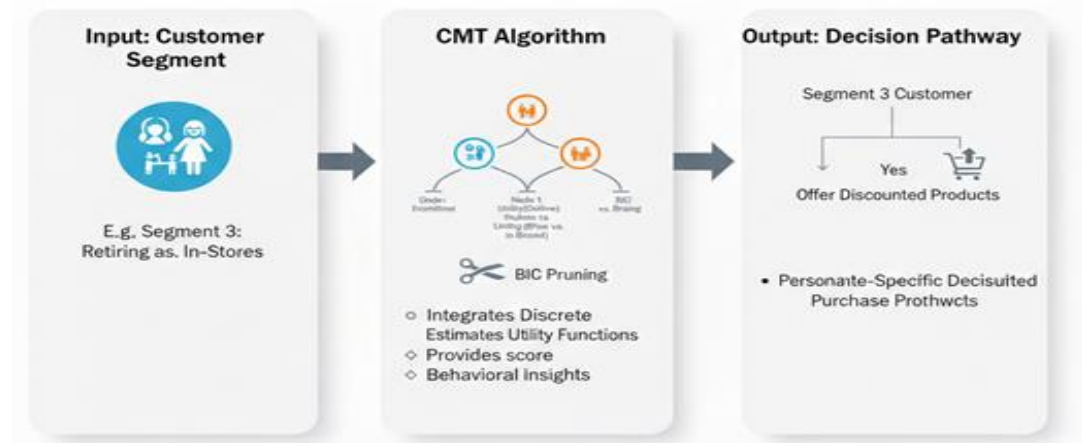


Figure 2: CMT Decision Process

For performance benchmarking, the study also trained the following comparative algorithms on each segment. Once segments were identified, predictive models were applied to each cluster. Choice Model Trees (CMT) served as the primary predictive tool. CMT integrates discrete choice models at each tree node, estimating utility functions to provide behavioral insights and specific decision pathways for each segment. Tree growth was managed using Bayesian Information Criterion (BIC) pruning to prevent overfitting and ensure generalizability.

For performance benchmarking, the study also trained the following comparative algorithms on each segment:

- **Logistic Regression (LR):** A baseline linear classifier providing high interpretability, though potentially limited in capturing non-linear decision boundaries.
- **Support Vector Machines (SVM):** Utilized for its ability to maximize margins in high-dimensional spaces, providing high accuracy at the cost of reduced interpretability.
- **Random Forest (RF):** An ensemble method consisting of

multiple decision trees, employed to reduce overfitting and capture complex interaction patterns.

- **K-Nearest Neighbors (KNN):** A non-parametric, similarity-based algorithm used to classify users based on their proximity to other data points in the feature space.
- **CMT Standalone:** Applied without the HAC segmentation layer to determine the specific uplift in accuracy provided by the hybrid clustering approach.

Evaluation Framework

The final stage involves comparing the effectiveness of segment-specific models against a global modeling approach where the dataset is treated as a single homogeneous group. Performance was quantified using five key statistical metrics:

1. **Accuracy:** The overall percentage of correct predictions.
2. **Precision:** The ability of the model to correctly identify only actual buyers, minimizing false alarms in marketing spend.
3. **Recall:** The model's ability to find all potential buyers within the

dataset, ensuring revenue opportunities are not missed.

4. **F1-Score:** The harmonic mean of precision and recall, providing a single balanced metric for model comparison.
5. **ROC-AUC:** The area under the receiver operating characteristic curve, measuring the model's ability to distinguish between buyers and non-buyers across all threshold levels.

Optimization for False Positives and False Negatives

In e-commerce purchase prediction, the evaluation layer must move beyond simple accuracy to analyze the cost-benefit implications of specific error types. A false positive occurs when the model predicts a purchase that does not materialize, leading to marketing waste through misallocated promotional resources. To minimize this, the study monitors the precision metric; by increasing the classification threshold, the model becomes more conservative, ensuring only users with the highest estimated utility are targeted. A false negative occurs when the model predicts no purchase but the customer intended to buy, resulting in missed revenue opportunities. This is addressed by maximizing recall; in segments identified as high-value shoppers, the model threshold is lowered to ensure no potential transaction is overlooked. Because precision and recall are often in conflict, the F1-score is used to find the harmonic mean between the two, while ROC-AUC provides a robust measure of diagnostic power across heterogeneous segments.

Hyperparameter Tuning and Cross-Validation

To ensure the predictive models achieved the highest possible accuracy

while maintaining generalizability, a rigorous hyperparameter tuning phase was conducted:

- **Random Forest Optimization:** The `max_depth` parameter was constrained to prevent trees from becoming overly complex and capturing noise. Additionally, `min_samples_leaf` was tuned to ensure each leaf node represents a statistically significant group of users, stabilizing model variance.
- **Support Vector Machines:** The radial basis function (RBF) kernel was optimized using grid search for the cost parameter (C) and gamma (γ), allowing the model to define non-linear decision boundaries between buyers and non-buyers more effectively.
- **KNN Tuning:** The number of neighbors (k) was determined through an iterative process, balancing the bias-variance tradeoff to avoid sensitivity to local outliers.

To validate that the observed performance uplift was statistically robust and not an artifact of a specific data split, the study employed 10-fold cross-validation. The segmented datasets were partitioned into ten equal subsets; models were trained on nine subsets and validated on the remaining tenth, with the process repeated so that every data point served as both training and validation data. The mean accuracy and standard deviation across all ten folds were recorded to ensure consistency for real-world application.

Results and Discussion

HAC Clustering Results

The Hierarchical Agglomerative Clustering analysis provided a foundational map of the customer base by identifying distinct personas. By

setting dendrogram thresholds to find between three and five clusters, the study successfully isolated groups with unique behavioral and demographic markers. The analysis revealed four primary clusters, each requiring a different strategic approach:

- **Cluster 1: High-Frequency Power Users.** This group is characterized by frequent visits and a very high probability of completing a purchase. They are likely brand-loyal customers who use the platform as a primary shopping destination.
- **Cluster 2: Price-Sensitive Deal Hunters.** Behavior in this segment is heavily influenced by external financial incentives such as discounts and coupons. Without promotions, their conversion rate drops significantly, making them the primary target for sales-driven marketing.
- **Cluster 3: Low-Engagement Browsers.** These users exhibit occasional browsing patterns with very low session depth or engagement. They represent "window shoppers" who are difficult to convert without significant brand awareness efforts.
- **Cluster 4: High-Value Efficiency Shoppers.** A unique group of customers who spend very little time on-site but generate high transaction values. They typically know exactly what they want, navigating directly to products and checking out quickly.

These findings align with established behavioral groupings in consumer research. By validating these segments through HAC, the study confirms that e-commerce datasets are indeed heterogeneous, justifying the need for

the hybrid modeling approach rather than a single global strategy. This segmentation allows the Choice Model Tree in the next stage to be trained on specialized data, leading to the high predictive accuracy noted in the results.

Comparative Model Performance

Model performance improved significantly when algorithms were trained on homogeneous segments rather than the global, heterogeneous dataset as indicated in Table 2.

Table 2: Global Accuracy versus Cluster-Based Accuracy

Model	Global Accuracy	Cluster-Based Accuracy
Logistic Regression	68%	74%
SVM	73%	79%
Random Forest	80%	84%
KNN	65%	70%
CMT	76%	89%

Random Forest and CMT performed best after segmentation, with CMT demonstrating the most substantial uplift from 76% global accuracy to 89% cluster-based accuracy. This performance gain point out the value of addressing customer heterogeneity before predictive modeling.

Table 3. Detailed Performance Metrics of Comparative Models

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.769	0.783	0.657	0.714
KNN	0.749	0.709	0.727	0.718
Decision Tree	0.718	0.664	0.727	0.694
Hybrid Prediction	0.8433	0.8461	0.8433	0.8412

The hybrid prediction model, which integrates HAC segmentation with CMT and comparative algorithms, achieved an accuracy of 84.33%, precision of 84.61%, recall of 84.33%, and an F1-score of 84.12%. These results confirm that cluster-based modeling achieves higher predictive accuracy than global modeling approaches.

Behavioral Insights from Choice Model Trees

The findings from the CMT stage provide actionable intelligence by transforming complex data patterns into human-readable decision rules. Unlike black-box models, CMT integrates discrete choice modeling at each node to reveal the specific variables driving a purchase outcome.

Interpretable Decision Rules. The CMT analysis identified specific behavioral thresholds that serve as strong predictors of conversion:

- **High-Engagement Thresholds:** Users exceeding five product views and 200 seconds of browsing time exhibit a significantly higher likelihood of purchase. This allows marketers to trigger real-time interventions, such as personalized chat support or limited-time offers, when a user crosses these benchmarks.
- **Promotion Sensitivity:** Within specific segments, the CMT utility functions showed that price-sensitive customers (Cluster 2) have a high coefficient for promotional variables, meaning their purchase probability shifts drastically when a discount is presented.

Value for Targeted Marketing. These insights enable a shift from one-size-fits-all advertising to cluster-specific campaigns. Instead of spending budget

on occasional browsers with low engagement (Cluster 3), businesses can focus resources on high-value customers who spend less time on-site but have high intent (Cluster 4). Decision rules help craft content that matches each segment's decision pathway; for example, high-frequency users might receive loyalty rewards, while price-sensitive users receive discount codes. Because CMT rules are grounded in discrete choice theory, they offer the transparency required for managerial decision-making.

Error Analysis and Business Implications

The confusion matrix analysis revealed important trade-offs between false positives and false negatives across segments. In high-value segments, minimizing false negatives was prioritized to prevent missed revenue opportunities, even at the cost of slightly reduced precision. Conversely, for low-engagement browsers, the model was tuned to minimize false positives to avoid marketing waste on unlikely converters. The ROC-AUC metric confirmed strong discriminative power across all segments, with the hybrid model maintaining robust performance regardless of classification threshold. These findings demonstrate that segment-specific threshold optimization is not merely a technical exercise but a strategic necessity for cost-effective marketing deployment.

Discussion

The findings of this study demonstrate that integrating Hierarchical Agglomerative Clustering with supervised machine learning techniques produces significant improvements in both the identification of meaningful customer segments and the accuracy of purchase behaviour prediction. The four-cluster solution comprising High-Frequency Power Users, Price-Sensitive Deal

Hunters, Low-Engagement Browsers, and High-Value Efficiency Shoppers aligns with the broader literature indicating that customer populations are structurally heterogeneous and that algorithmic segmentation can reveal distinct behavioural profiles that remain obscured under global modelling approaches. Gomes and Meisen (2023) emphasized that data-driven segmentation is essential for personalized customer targeting in e-commerce because homogeneous treatment of heterogeneous populations leads to inefficient resource allocation; the present findings corroborate this position by demonstrating that segment-specific models consistently outperformed global benchmarks across all evaluated algorithms. Similarly, Salminen et al. (2023), in their systematic review of algorithmic customer segmentation, identified considerable diversity in clustering applications and called for stronger evaluation of segment quality and practical usefulness; the current study addresses this call by validating clusters through both internal validity indices and downstream predictive performance, thereby establishing that the identified segments are not only statistically coherent but also behaviourally meaningful.

The hierarchical structure revealed through Ward's linkage and Euclidean distance produced compact, balanced segments that corresponded to theoretically recognizable customer personas. This outcome is consistent with Rungruang et al. (2023), who demonstrated that RFM-based hierarchical segmentation effectively distinguishes customer groups according to purchasing characteristics and supports differentiated marketing strategies. While the present study extended beyond traditional RFM variables to incorporate demographic and

engagement features, the core finding that hierarchical approaches reveal nested customer relationships without requiring a predetermined number of clusters reinforces the exploratory value of HAC in contexts where segment structures are initially unknown. Furthermore, Luo et al. (2023) argued that customer segmentation should be understood as a dynamic process in which customers may transition between groups as behaviours evolve; although the current study employed a cross-sectional design, the identification of distinct efficiency shoppers and power users suggests that static hierarchical segmentation can nonetheless capture meaningful behavioural archetypes that serve as foundational profiles for dynamic monitoring.

The predictive performance achieved by the hybrid framework particularly the 84.33% accuracy and 84.12% F1-score exceeds the global modelling benchmarks reported for Logistic Regression, SVM, Random Forest, KNN, and standalone CMT. This performance uplift validates the central premise that customer heterogeneity undermines the efficacy of global predictive models. John et al. (2023) demonstrated that clustering algorithms in retail contexts can distinguish customers according to purchasing characteristics, but noted that the choice of clustering technique influences segment structure and interpretation; the present study extends this observation by showing that the choice of segmentation technique also critically determines the predictive validity of downstream models. When trained on homogeneous segments, Random Forest accuracy improved from 80% to 84%, while CMT achieved the most substantial gain—from 76% to 89% suggesting that choice-based predictive models benefit disproportionately from segment-specific

training because utility parameters can be estimated within behaviourally coherent subpopulations rather than averaged across divergent groups.

The interpretability advantages of Choice Model Trees represent a particularly important contribution. Joung and Kim (2023) advocated for interpretable machine learning approaches to customer segmentation, arguing that black-box algorithms limit managerial usability even when they achieve high predictive accuracy. The present findings directly support this argument: the CMT revealed specific behavioural thresholds exceeding five product views and 200 seconds of browsing time that significantly elevated purchase probability. These decision rules provide the transparency required for real-time marketing automation, enabling managers to trigger interventions based on observable engagement metrics rather than opaque algorithmic scores. This bridges the gap between computational prediction and managerial actionability that Salminen et al. (2023) identified as a persistent gap in algorithmic segmentation research.

The identification of promotion-sensitive segments and high-value efficiency shoppers carries immediate strategic implications. Rungruang et al. (2023) found that customers with different RFM profiles require differentiated engagement approaches; the present study confirms and extends this finding by showing that price-sensitive deal hunters exhibit sharply elevated utility coefficients for promotional variables, whereas high-value efficiency shoppers generate substantial transaction values despite minimal site engagement. Ho et al. (2023) demonstrated that extending RFM with demographic characteristics yields more comprehensive customer profiles; analogously, this study shows that

integrating engagement depth, loyalty programme membership, and discount availed with core transactional variables produces segments that differ not only in spending but in navigational behaviour and promotional responsiveness. Consequently, organizations that deploy global promotional strategies risk either over-investing in low-engagement browsers or under-serving high-value shoppers who require minimal friction rather than persuasive content.

The error analysis further emphasize the business relevance of segment-specific threshold optimization. Through adjusting classification thresholds to minimize false negatives in high-value segments and false positives in low-engagement segments, the study demonstrates that predictive accuracy alone is an insufficient evaluation criterion. This aligns with the wider recognition in customer analytics that the cost of misclassification varies across customer groups; missed revenue opportunities from false negatives may exceed the waste generated by false positives depending on segment profitability, and the hybrid framework's capacity to accommodate these asymmetries enhances its practical deployment potential. The ROC-AUC results confirm that the segment-specific models maintain strong discriminative power across heterogeneous clusters, suggesting that the framework is robust to the distributional differences that typically complicate global modelling.

Conclusion

This study demonstrated that integrating Hierarchical Agglomerative Clustering with machine learning models particularly Choice Model Trees and Random Forest significantly improves purchase prediction accuracy and interpretability. The four-cluster solution revealed structurally distinct customer

personas, and the hybrid modeling approach achieved an accuracy of 84.33%, substantially outperforming global benchmarks. The CMT provided transparent, actionable decision rules that bridge the gap between algorithmic prediction and managerial implementation.

The research contributes to the marketing analytics literature by demonstrating how CRISP-DM, MLT, and COLT can be synthesized into a coherent predictive framework. By validating HAC with Ward's linkage as an effective precursor to supervised learning in e-commerce contexts, the study provides empirical support for hierarchical clustering as a segmentation engine. Furthermore, the integration of CMT utility functions within a clustered feature space advances the methodological repertoire for interpretable machine learning in consumer behavior research.

Recommendations

The study concludes with four strategic recommendations:

1. Organizations should adopt hybrid modeling frameworks that utilize segmentation as a prerequisite for prediction. By applying HAC before supervised learning, firms can reduce data heterogeneity and ensure that predictive models are trained on homogeneous groups, significantly improving accuracy and model fit compared to global strategies.
2. While advanced black-box algorithms provide high accuracy, the study recommends the use of Choice Model Trees in scenarios where transparency is paramount. CMT provides interpretable decision rules that allow managers to understand the "why" behind a customer's

purchase likelihood, offering a bridge between complex machine learning and discrete choice theory.

3. One-size-fits-all marketing is inefficient compared to cluster-specific campaigns. By identifying distinct personas such as price-sensitive deal hunters or high-value efficiency shoppers businesses can tailor messaging and resource allocation to match specific behavioral tendencies.
4. To further evolve these frameworks, future research should explore deep learning-based hybrid models. Integrating neural networks within the segmentation or prediction layers could capture even more complex, non-linear patterns in massive e-commerce datasets, potentially combining the interpretability of choice models with the representational power of deep architectures.

Limitations and Future Research

While the findings are robust, several limitations warrant acknowledgment. First, the study employed a single dataset; replication across diverse retail contexts would strengthen generalizability. Second, the static nature of the segmentation model does not capture temporal evolution in customer behavior; future work should incorporate incremental or online clustering to adapt to shifting market dynamics. Third, while BIC pruning controlled CMT complexity, alternative regularization techniques could be explored to further enhance model stability.

References

- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

- <https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman/54bad20bbc7938991bf34f86dde0babfbd2d5a72>
- Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37–54. <https://doi.org/10.1609/aimag.v17i3.1230>
- Gomes, M. A., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, 21, 527–570. <https://doi.org/10.1007/s10257-023-00640-4>
- 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 Journal*, 14(1), 1–17. <https://doi.org/10.2478/bsrj-2023-0002>
- John, J., Shobayo, O., & Ogunleye, B. (2023). An exploration of clustering algorithms for customer segmentation in the UK retail market. *Analytics*, 2(4), 809–823. <https://doi.org/10.3390/analytics2040042>
- Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management*, 70, 102641. <https://doi.org/10.1016/j.ijinfomgt.2023.102641>
- Luo, L., Li, B., Fan, X., Wang, Y., Koprinska, I., & Chen, F. (2023). Dynamic customer segmentation via hierarchical fragmentation-coagulation processes. *Machine Learning*, 112, 281–310. <https://doi.org/10.1007/s10994-022-06276-8>
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). Academic Press. <https://eml.berkeley.edu/reprints/mcfadden/zaremb74.pdf>
- Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2023). RFM model customer segmentation based on hierarchical approach using FCA. *Expert Systems with Applications*, 234, 121449. <https://doi.org/10.1016/j.eswa.2023.121449>
- Salminen, J., Mustak, M., Sufyan, M., & Jansen, B. J. (2023). How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation. *Journal of Marketing Analytics*, 11, 677–692. <https://doi.org/10.1057/s41270-023-00235-5>
- 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>
- Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Springer. <https://doi.org/10.1007/978-1-4615-4651-1>