International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Data-Driven Customer Acquisition using Machine Learning for Business Growth

Customer Intelligence Framework for Customer Targeting and Campaign Optimization

Syeda Zoya[1], Insiya Maryam[2], Zabiullah Khan[3], Raju Kumar Allam[4]
[1] Data Solutions Specialist, Falcon Informatics
[2] Platform Head, Falcon Informatics
[3] Data Scientist, Falcon Informatics
[4] Practice head, Falcon Informatics

Abstract - Customer acquisition and business growth increasingly depend on understanding customer purchasing behaviour. However, traditional customer segmentation approaches based on demographic information or conventional Recency–Frequency–Monetary analysis often fail to capture the complexity of customer behaviour. This research proposes a Behaviour-Driven Customer Segmentation Framework that integrates behavioural feature engineering, EDA, feature selection, Principal Component Analysis and K-Means clustering to identify meaningful customer segments.PCA reduced the feature space from 32 variables to 15 principal components, retaining 95.62% of the total variance. Cluster validation identified an optimal four-cluster solution representing distinct customer purchasing behaviours. The resulting customer segments support personalized marketing, customer retention, promotional planning, inventory management, and strategic decision-making. The solution provides a practical approach for converting retail transaction data into actionable customer intelligence, supporting data-driven customer acquisition and sustainable business growth.

Keywords: Customer Segmentation, Machine Learning, Business Growth, Retail, Data-Driven, Customer Acquisition, Behavioral Analytics

1. INTRODUCTION

Customer acquisition remains one of the most significant investments for retail organizations. However, traditional marketing approaches often rely on broad demographic segmentation or generalized promotional campaigns that overlook differences in customer behaviour. This can reduce marketing effectiveness and lead to inefficient use of organizational resources. Behaviour-driven customer segmentation offers a more targeted approach by grouping customers according to their purchasing patterns.

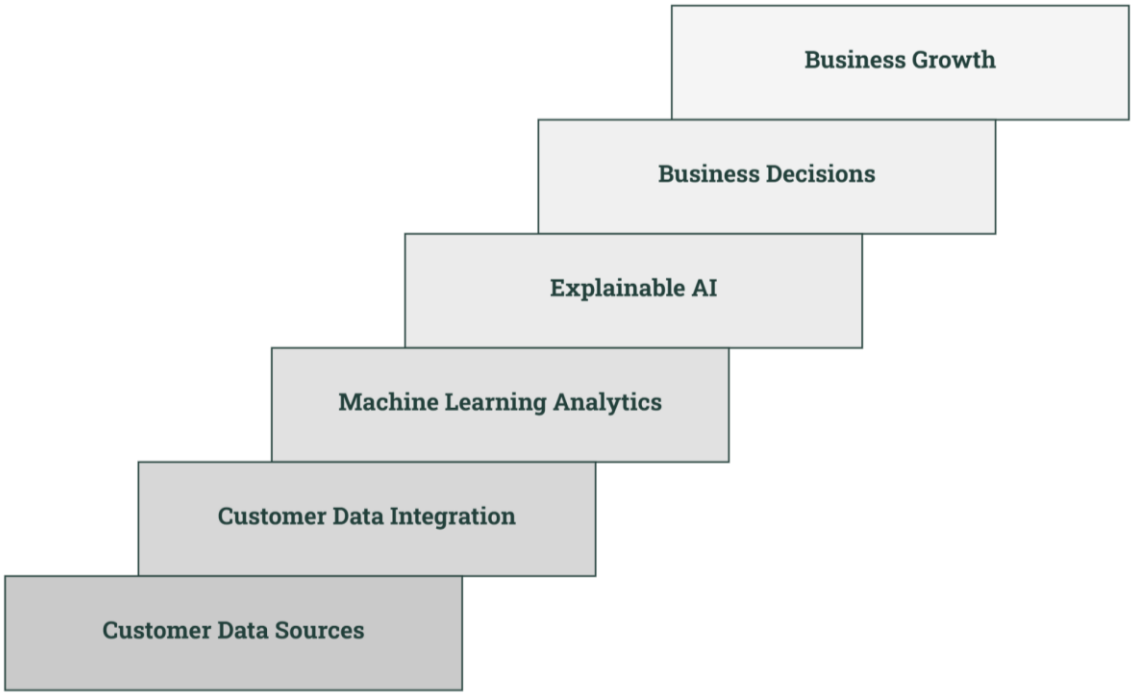
This research proposes a **Behaviour-Driven Customer Segmentation Framework** that combines customer master construction, behavioural feature engineering, exploratory data analysis, feature selection, Principal Component Analysis (PCA), and K-Means clustering to identify meaningful customer segments from retail transaction data. The framework is evaluated using the dunnhumby "The Complete Journey" dataset.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

The identified customer segments provide practical insights for personalized marketing, customer retention, promotional planning, inventory management, and strategic decision-making. By integrating statistical analysis with machine learning, it offers a practical and scalable approach to customer analytics that supports customer acquisition and sustainable business growth.

Figure 1.1. Business Growth through customer data



2. LITERATURE REVIEW

Retail transaction data has spurred research into customer analytics for improved marketing and decision-making. Studies have utilized statistical and machine learning methods for segmentation, purchase behavior analysis, and recommendation systems. Recently, focus has shifted toward behavioral analytics to better understand purchasing patterns and personalize strategies. This chapter reviews literature on segmentation, feature engineering, dimensionality reduction, and clustering techniques to identify current trends and the research gap.

2.1 Customer Analytics and Data-Driven Marketing

Customer analytics has become a fundamental component of modern marketing, enabling organizations to leverage customer data for understanding consumer behavior, optimizing marketing strategies, and improving customer lifetime value. With the rapid growth of digital platforms, businesses now generate large volumes of transactional, behavioral, and demographic data that support evidence-based decision-making.

Theodorakopoulos et al. (2026) investigated the role of big data analytics and artificial intelligence in digital consumer behavior. Their study demonstrated that AI-driven customer analytics significantly improves customer targeting, campaign optimization, and customer engagement by integrating data from online platforms, mobile applications, loyalty programs, and transactional systems. The authors also highlighted emerging concerns regarding privacy, algorithmic fairness, and consumer trust, emphasizing that organizations must balance personalization with ethical data governance. The research establishes customer analytics as a strategic capability that enhances marketing effectiveness while requiring responsible AI implementation.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Similarly, **Sun (2025)** proposed a data-driven personalized marketing framework that combines machine learning with customer behavioral analysis to optimize digital marketing strategies. Using predictive models trained on customer interactions, purchase histories, and browsing behavior, the study reported substantial improvements in customer engagement, conversion rates, and average transaction value. The research further demonstrated that integrating interactive analytics dashboards enables marketers to interpret customer patterns more effectively and supports timely business decision-making. However, the study primarily focuses on digital commerce applications and does not extensively evaluate the generalizability of the content across diverse industries.

Both studies demonstrate that customer analytics has evolved beyond descriptive reporting toward intelligent, data-driven decision support systems capable of delivering personalized marketing interventions. While previous research has established the effectiveness of big data analytics and machine learning in improving customer engagement and campaign performance, limited attention has been given to developing integrated frameworks that combine customer segmentation, predictive campaign response modeling, and explainable artificial intelligence. This gap motivates the present research, which aims to develop a comprehensive machine learning-based customer analytics framework that not only predicts marketing outcomes but also provides interpretable insights to support managerial decision-making.

2.2 Machine Learning in Customer Analytics

Machine learning (ML) has transformed customer analytics by enabling organizations to extract meaningful insights from large-scale customer data and predict future customer behavior with greater accuracy than traditional statistical techniques. Unlike conventional analytical methods that primarily describe historical trends, ML algorithms learn complex relationships within data to support customer segmentation, churn prediction, purchase forecasting, recommendation systems, and personalized marketing. The increasing availability of high-dimensional customer data has accelerated the adoption of supervised and ensemble learning techniques for customer-centric decision-making.

Zhe Yuan. (2025) investigated the application of machine learning for consumer behavior prediction and precision marketing. The study evaluated multiple supervised learning algorithms using customer behavioral data to predict purchasing intentions and optimize personalized marketing strategies. Experimental results demonstrated that machine learning models significantly outperformed conventional analytical approaches in identifying customer preferences and improving marketing precision. The authors further highlighted that feature engineering and data quality substantially influence predictive performance, emphasizing the importance of robust preprocessing before model development. However, the research primarily focused on prediction accuracy and provided limited discussion on model interpretability, which is essential for managerial adoption.

Mittal (2025) proposed an intelligent customer acquisition framework based on machine learning-driven campaign channel attribution. The research addressed the limitations of traditional attribution models, such as last-click and linear attribution, by employing machine learning algorithms capable of learning complex customer journey patterns across multiple marketing channels. The proposed approach demonstrated improved customer acquisition prediction and more efficient marketing budget allocation, enabling organizations to optimize campaign effectiveness. Nevertheless, the study concentrated on customer acquisition and attribution, while offering limited integration with broader customer analytics tasks such as customer segmentation and response prediction.

Langen and Huber (2022) explored the application of causal machine learning for evaluating marketing interventions through coupon campaigns. Rather than merely predicting customer responses, the study estimated heterogeneous treatment effects across different customer groups and identified which customers should receive promotional offers to maximize campaign effectiveness. Their findings demonstrated that causal machine learning provides deeper insights than conventional predictive models by supporting data-driven targeting strategies and improving decision quality. The research also illustrated the growing importance of combining predictive analytics with causal inference for evidence-based marketing decisions.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Collectively, these studies demonstrate that machine learning has evolved from a predictive tool into a comprehensive decision-support technology capable of enhancing customer acquisition, behavioral prediction, and marketing optimization. Despite significant advances, most existing research focuses on isolated applications such as purchase prediction or campaign attribution. Addressing this gap forms a key motivation for the proposed research, which seeks to develop an end-to-end machine learning framework for customer analytics using interpretable predictive models.

2.3 Customer Segmentation Techniques

Customer segmentation, a key part of customer analytics, classifies customers into groups based on demographic, behavioral, or transactional traits to drive personalized marketing and optimize business value. Advances in machine learning have evolved this process from rule-based methods to data-driven clustering, with Recency-Frequency-Monetary (RFM) analysis becoming a standard approach in retail.

John, Shobayo, and Ogunleye (2023) conducted a comprehensive comparative study of state-of-the-art clustering algorithms for customer segmentation using a UK online retail dataset. The research employed the RFM framework alongside K-Means, Gaussian Mixture Models (GMM), DBSCAN, Agglomerative Clustering, and BIRCH to evaluate segmentation quality. Experimental results indicated that the Gaussian Mixture Model achieved the highest clustering performance, outperforming the other techniques based on Silhouette Score and cluster separation metrics. The study demonstrated that selecting an appropriate clustering algorithm substantially influences segmentation quality and the resulting business insights. However, the research primarily focused on clustering performance without integrating predictive analytics or explainable machine learning for downstream marketing applications.

Wong et al. (2024) explored customer segmentation in e-commerce by integrating RFM analysis with Hierarchical Clustering and K-Means algorithms. The study compared clustering performance using multiple internal validation measures and reported that Hierarchical Clustering produced more coherent and interpretable customer groups than K-Means. The resulting customer segments enabled businesses to distinguish between loyal, high-value, and inactive customers, thereby facilitating personalized marketing strategies and improved customer relationship management. Nevertheless, the research concentrated primarily on segmentation accuracy and did not investigate how the identified customer groups could be utilized for campaign response prediction or business decision support.

Lewaaelhamd et al. (2023) proposed a machine learning-based customer segmentation framework that combines RFM analysis with unsupervised clustering techniques to identify customer groups from transactional data. The study demonstrated that integrating behavioral metrics with clustering algorithms provides a more objective and scalable segmentation process than manual classification approaches. The authors emphasized that machine learning-based segmentation enables organizations to identify high-value customers, customers at risk of churn, and potential growth segments more effectively. However, the idea focused mainly on customer classification and did not incorporate explainable artificial intelligence or predictive models capable of forecasting customer responses to future marketing campaigns.

Overall, the existing literature demonstrates that machine learning-based customer segmentation techniques significantly outperform traditional rule-based methods in identifying meaningful customer groups and supporting personalized marketing strategies. Most recent studies emphasize clustering performance using RFM features and unsupervised learning algorithms, yet relatively few integrate customer segmentation with campaign response prediction, explainable artificial intelligence, and managerial decision-support systems within a unified framework. This limitation motivates the present research, which combines customer segmentation with supervised predictive modeling and explainable AI to provide actionable insights for marketing decision-makers.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

2.4 Campaign Response Prediction

Machine learning-driven campaign response prediction is vital for modern marketing, allowing organizations to target likely responders, optimize ROI, and lower costs. Unlike traditional mass marketing, predictive analytics uses customer demographics, transaction history, and behaviors to enable personalized strategies. Research increasingly leverages supervised, ensemble, and causal learning techniques to improve predictive effectiveness.

El-Hajj et al. (2024) developed a decision tree-based predictive framework to identify customers most likely to respond to direct marketing campaigns. The study evaluated multiple customer-related features, including demographic characteristics, purchasing behavior, and previous campaign interactions, to construct an interpretable classification model. Experimental results demonstrated that predictive modeling substantially improved campaign targeting efficiency while reducing the cost associated with contacting non-responsive customers. The authors also emphasized that interpretable models such as decision trees facilitate managerial understanding of the factors driving campaign success. However, the research focused primarily on a single classification algorithm and did not investigate the performance of more advanced ensemble learning techniques or explainable AI methods.

Choi (2023) assessed the predictive performance of several machine learning algorithms for direct marketing response prediction, including Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting models. The comparative analysis revealed that ensemble learning methods consistently outperformed traditional statistical approaches in predicting customer responses while achieving higher precision and recall. The study highlighted the importance of feature engineering and class imbalance handling in improving predictive accuracy. Nevertheless, its primary objective was algorithm comparison rather than developing an integrated decision-support framework capable of combining prediction with customer segmentation and model interpretability.

Langen and Huber (2022) extended campaign response prediction by incorporating causal machine learning to estimate heterogeneous treatment effects of coupon campaigns. Instead of merely predicting whether customers would respond, the study evaluated which customer segments would benefit most from receiving promotional offers. Using causal inference and optimal policy learning, the authors demonstrated that data-driven customer targeting significantly improved campaign effectiveness and marketing resource allocation. The research showed that causal machine learning provides richer decision support than conventional predictive models by distinguishing between customers who are likely to respond naturally and those whose behavior is influenced by marketing interventions. However, the framework primarily focused on treatment-effect estimation and did not integrate explainable artificial intelligence for transparent managerial interpretation.

Studies show machine learning improves marketing efficiency through precise targeting. However, few frameworks unify segmentation, response prediction, and explainable AI. This research addresses this gap, providing accurate predictions and transparent insights for strategic decision-making.

2.5 Explainable Artificial Intelligence in Marketing

The increasing adoption of machine learning in marketing has significantly improved predictive accuracy; however, the complexity of advanced algorithms often limits their interpretability. Many high-performing models operate as "black boxes," making it difficult for marketers to understand how predictions are generated or which customer attributes influence decision-making. Explainable Artificial Intelligence (XAI) addresses this challenge by providing transparent and interpretable explanations for machine learning models, thereby improving trust, accountability, and managerial acceptance. In marketing applications, XAI techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) enable organizations to identify the factors influencing customer behavior, campaign responses, and purchasing decisions, facilitating more informed business strategies.

Haag et al. (2022) proposed an explainable artificial intelligence framework for cross-selling in the energy retail sector by integrating machine learning with SHAP-based explanations. Using transactional data from over

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

220,000 customers, the study developed predictive models to identify customers likely to purchase additional products. The authors demonstrated that SHAP explanations successfully revealed the contribution of individual customer attributes to each prediction, enabling marketers to understand why particular customers were recommended for cross-selling. The research concluded that explainability not only improves managerial trust in AI systems but also supports more effective customer relationship management. However, the study focused specifically on cross-selling applications and did not examine broader customer analytics tasks such as customer segmentation or campaign response prediction.

Karimzadeh et al. (2024) developed an explainable machine learning framework for analyzing online customer reviews to identify the product attributes that most strongly influence customer satisfaction. The proposed approach combined machine learning with SHAP to quantify the importance of individual product features and provide interpretable insights for product design and marketing decisions. Experimental results showed that explainable models enabled businesses to better understand customer preferences while maintaining high predictive performance. The authors emphasized that integrating explainability into machine learning enhances decision quality by translating complex model outputs into actionable business knowledge. Nevertheless, the research primarily focused on product attribute analysis rather than developing an integrated framework for customer segmentation and marketing campaign optimization.

The reviewed studies demonstrate that Explainable Artificial Intelligence has become an important advancement in marketing analytics by improving the transparency and interpretability of machine learning models. Although recent research has successfully applied SHAP-based explanations to cross-selling and customer preference analysis, limited work has integrated explainable AI with customer segmentation and campaign response prediction within a comprehensive customer analytics framework. The proposed research addresses this limitation by incorporating SHAP-based model explanations into predictive customer analytics, enabling both accurate marketing predictions and transparent managerial decision-making.

2.6 Business Decision Support Systems

Business Decision Support Systems (BDSS) have evolved from traditional reporting tools into intelligent platforms that assist managers in making data-driven strategic and operational decisions. The integration of artificial intelligence, machine learning, and business analytics has enhanced the capability of these systems to process large volumes of structured and unstructured data, identify hidden patterns, and generate actionable recommendations. In the marketing domain, decision support systems enable organizations to optimize customer acquisition strategies, evaluate campaign performance, forecast customer behavior, and improve resource allocation. Modern decision support systems increasingly incorporate predictive analytics and visualization tools to facilitate timely and informed managerial decision-making.

Abedin et al. (2023) proposed an intelligent decision support framework that integrates machine learning with business analytics to improve organizational decision-making. The study demonstrated that predictive models embedded within decision support systems significantly enhanced the accuracy of forecasting and strategic planning across various business functions. The authors emphasized that combining data-driven analytics with interactive visualization enables managers to interpret complex analytical results more effectively and make evidence-based decisions. However, the framework was designed as a general-purpose business solution and did not specifically address customer analytics or marketing campaign optimization.

Kumar and Sharma (2024) developed an AI-enabled decision support system for marketing analytics that combined predictive modeling with business intelligence dashboards. The proposed system utilized customer transaction data to generate recommendations for customer targeting, campaign planning, and performance monitoring. Experimental findings indicated that integrating machine learning predictions with visual decision support improved managerial efficiency and facilitated quicker responses to changing market conditions. Nevertheless, the study primarily emphasized predictive performance and dashboard visualization while providing limited model interpretability, making it difficult for managers to understand the reasoning behind the generated recommendations.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

The reviewed studies indicate that modern Business Decision Support Systems have become essential tools for transforming analytical insights into strategic business actions. Although recent research demonstrates the effectiveness of integrating predictive analytics and visualization into decision support platforms, relatively few studies combine customer segmentation, campaign response prediction, and Explainable Artificial Intelligence within a unified system tailored for marketing decision-making. The present research addresses this gap by developing an integrated customer analytics framework that not only predicts campaign responses but also provides interpretable insights through explainable AI, thereby supporting transparent and data-driven managerial decision-making.

3. RESEARCH GAP

Digital commerce and data-driven marketing has significantly increased the adoption of machine learning techniques for customer analytics. Existing research has extensively explored customer segmentation, purchasing behaviour analysis, customer lifetime value (CLV) estimation, churn prediction, recommendation systems, and personalized marketing using various statistical and machine learning approaches. Clustering techniques such as K-Means, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models have been widely applied to identify customer groups based on transactional and demographic information. However, despite these advancements, several important research gaps remain.

First, most existing studies focus on individual analytical tasks, such as customer segmentation, RFM analysis, purchase prediction, or recommendation systems, without integrating these components into a unified behaviour-driven customer analytics framework. As a result, customer insights are often generated independently, limiting their ability to support comprehensive business decision-making and long-term marketing strategy.

Second, many previous studies rely primarily on demographic information or conventional Recency–Frequency–Monetary (RFM) variables for customer segmentation. While these approaches provide useful customer classifications, they often overlook richer behavioural characteristics such as purchasing consistency, customer engagement, product diversity, shopping behaviour, purchase velocity, and promotion sensitivity. The limited use of multidimensional behavioural features restricts the ability of segmentation models to capture the complexity of real-world customer purchasing behaviour.

Third, existing customer segmentation research frequently emphasizes clustering performance while providing limited discussion on business interpretation and practical applicability. Although clustering algorithms successfully identify statistically distinct customer groups, relatively few studies translate these analytical results into actionable business insights that support personalized marketing, customer retention, promotional planning, inventory management, and strategic decision-making.

Furthermore, many studies employ dimensionality reduction or clustering techniques independently, with limited attention given to developing an integrated analytical pipeline that combines behavioural feature engineering, exploratory data analysis, feature selection, dimensionality reduction, cluster optimization, and customer segmentation within a single framework. Such an integrated approach is essential for producing meaningful customer intelligence that is both analytically reliable and practically useful.

To address these research gaps, this study proposes a **Behaviour-Driven Customer Segmentation Framework** that integrates customer master construction, behavioural feature engineering, exploratory data analysis, feature selection, Principal Component Analysis (PCA), and K-Means clustering into a unified analytical workflow. Unlike many existing studies that rely primarily on conventional customer metrics, the concept captures multiple dimensions of purchasing behaviour and converts retail transaction data into meaningful customer segments that support personalized marketing, customer relationship management, promotional planning, and strategic business decision-making.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

4. RESEARCH OBJECTIVES

Customer transaction data has created significant opportunities for organizations to apply machine learning for improving customer understanding and business decision-making. However, converting large volumes of retail transaction data into meaningful customer insights remains a challenging task due to the complexity of customer purchasing behaviour and the multidimensional nature of retail data. To address these challenges, this research proposes a **Behaviour-Driven Customer Segmentation Framework** that integrates behavioural feature engineering, exploratory data analysis, dimensionality reduction, and unsupervised machine learning to support data-driven customer analytics and business growth.

The primary objective of this study is to develop a behaviour-driven customer segmentation framework capable of transforming retail transaction data into meaningful customer segments that support data-driven marketing strategies and business decision-making.

The study pursues the following specific objectives:

RO1: To analyze customer transactional and behavioural data in order to identify meaningful purchasing patterns, shopping behaviour, and customer engagement characteristics.

RO2: To develop comprehensive behavioural features, including Recency, Frequency, Monetary Value (RFM), purchase velocity, shopping consistency, basket characteristics, product diversity, customer engagement, behavioural loyalty, and promotion sensitivity for enhanced customer profiling.

RO3: To perform exploratory data analysis and feature selection to understand behavioural relationships and prepare the customer dataset for machine learning analysis.

RO4: To apply Principal Component Analysis (PCA) for reducing feature dimensionality while preserving the majority of behavioural information required for customer segmentation.

RO5: To segment customers into meaningful behavioural groups using the K-Means clustering algorithm and determine the optimal number of customer segments through multiple cluster validation techniques.

RO6: To analyze and interpret the generated customer segments in order to identify distinct purchasing behaviours and derive actionable customer insights for business applications.

RO7: To develop an integrated customer analytics framework that converts behavioural customer data into meaningful business intelligence supporting personalized marketing, customer relationship management, inventory planning, and strategic business decision-making.

Table 4.1 Research Gap Analysis

Research Domain	Identified Research Gaps	Framework Contribution
Clustering-Based Segmentation	Predominantly utilizes conventional RFM metrics or limited behavioral variables.	Employs an extensive suite of engineered behavioral features to facilitate granular segmentation.
Traditional RFM Analysis	Provides a restricted view of multifaceted consumer purchasing behaviors.	Develops multidimensional features encompassing engagement, loyalty, and promotion sensitivity.
Behavioral Analytics Studies	Behavioral variables are frequently analyzed in isolation, lacking a holistic framework.	Provides a unified analytical pipeline integrating feature engineering and dimensionality reduction.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Segmentation Applications	Insufficient focus on systematic cluster validation and optimal selection methods.	Implements rigorous validation using Silhouette Scores and the Davies–Bouldin Index.
Business Intelligence Integration	Analytical results are seldom translated into strategic or actionable business insights.	Generates distinct segments to optimize personalized marketing and inventory planning.
Retail Analytics Frameworks	Absence of end-to-end workflows connecting data preparation to managerial interpretation.	Proposes a comprehensive workflow bridging behavioral engineering and strategic intelligence.

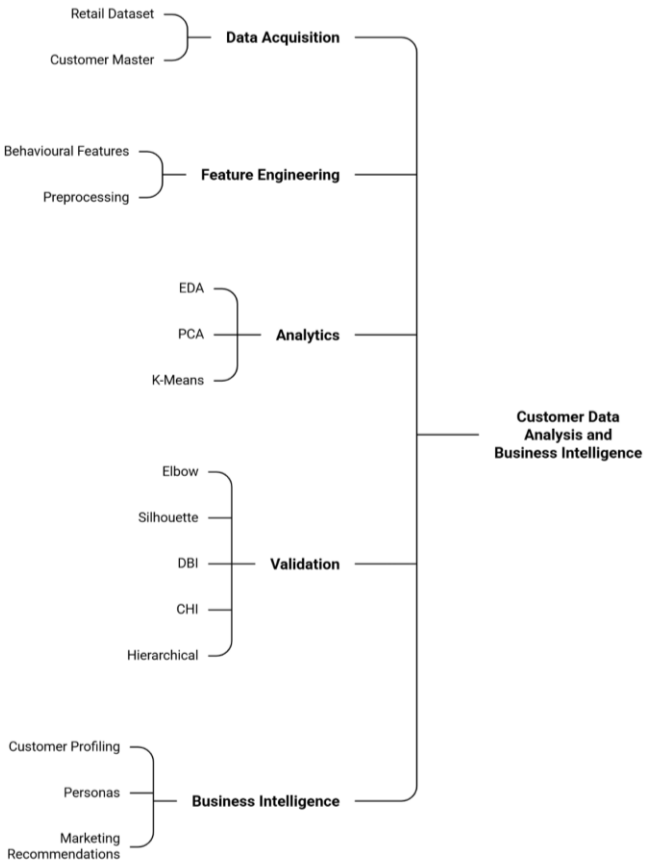
5. PROPOSED METHODOLOGY

5.1 Overview

The intent is to develop a behavior-driven customer intelligence framework capable of identifying meaningful customer segments from retail transaction data using unsupervised machine learning. Unlike traditional segmentation approaches that primarily rely on demographic variables or conventional Recency Frequency Monetary (RFM) analysis, it integrates comprehensive behavioral feature engineering, dimensionality reduction, clustering, validation, and business-oriented customer profiling to generate actionable marketing insights.

The approach follows a systematic analytical pipeline beginning with customer-level data construction and concluding with customer persona development. Each stage progressively transforms raw transaction records into interpretable customer intelligence suitable for business decision-making.

Figure 5.1. Proposed Customer Segmentation Framework



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

The workflow begins with customer transaction aggregation and feature engineering, followed by data preprocessing, exploratory data analysis, dimensionality reduction using Principal Component Analysis (PCA), customer segmentation using K-Means clustering, hierarchical validation, and customer profiling for business interpretation.

5.2 Customer Master Construction

It begins by transforming transaction-level retail records into a customer-centric analytical dataset. Rather than treating individual transactions as independent observations, all purchase activities belonging to each customer are aggregated to create a unified customer master. This transformation enables the analysis of long-term purchasing behavior instead of isolated shopping events.

The customer master serves as the foundation for the subsequent analytical pipeline, where each record represents a unique customer summarized through multiple behavioral characteristics.

5.3 Behavioral Feature Engineering

After constructing the customer master, behavioral features are engineered to capture different dimensions of customer purchasing behavior. Instead of relying solely on conventional RFM metrics, it derives a richer representation of customer behavior through multiple categories of engineered variables.

These behavioral dimensions include:

- Customer activity metrics
- Spending behavior
- Basket characteristics
- Product diversity
- Shopping consistency
- Purchase velocity
- Promotion sensitivity
- Customer engagement
- Behavioral loyalty
- Household purchasing power

The engineered features collectively provide a comprehensive description of customer purchasing patterns suitable for behavioral segmentation.

5.4 Data Preprocessing

Before model development, the customer dataset undergoes a structured preprocessing pipeline to improve data quality and ensure compatibility with machine learning algorithms.

The preprocessing stage includes:

- duplicate verification,
- missing value handling,
- removal of identifier variables,
- feature encoding where required,
- feature selection,
- feature scaling,
- correlation analysis.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

These preprocessing operations reduce redundancy, eliminate non-informative attributes, and standardize the feature space before dimensionality reduction and clustering.

5.5 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is performed to understand the statistical characteristics of the engineered customer dataset before clustering.

The analysis investigates:

- customer spending behavior,
- purchase frequency,
- basket characteristics,
- promotion responsiveness,
- customer engagement,
- feature correlations,
- distribution patterns,
- potential outliers.

The insights obtained during EDA provide an understanding of customer purchasing behavior while validating the suitability of the engineered features for clustering.

5.6 Dimensionality Reduction

Since the engineered customer dataset contains numerous correlated behavioral variables, dimensionality reduction is performed using Principal Component Analysis (PCA).

PCA transforms the standardized behavioral variables into a lower-dimensional orthogonal feature space while preserving the majority of information contained within the original dataset. This process reduces redundancy, minimizes multicollinearity, improves computational efficiency, and facilitates visualization of customer distributions prior to clustering.

5.7 Customer Segmentation

The transformed feature space is segmented using the K-Means clustering algorithm to identify groups of customers exhibiting similar purchasing behavior.

To determine the optimal number of customer segments, multiple internal cluster validation metrics are employed, including:

- Elbow Method
- Silhouette Coefficient
- Davies–Bouldin Index
- Calinski–Harabasz Index

The combined use of these evaluation measures ensures statistically reliable and business-interpretable customer segmentation.

5.8 Cluster Validation

To further assess the stability and structural consistency of the identified customer groups, Hierarchical Agglomerative Clustering is employed as an additional validation technique.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

The hierarchical clustering process generates a dendrogram that illustrates relationships among customer clusters and provides complementary evidence regarding the robustness of the segmentation results.

5.9 Feature Importance Analysis

Although customer segmentation is fundamentally an unsupervised learning problem, an additional feature importance analysis is performed to improve the interpretability of the generated customer groups.

A Random Forest classifier is trained using the cluster assignments as target labels solely for explanatory purposes. Feature importance scores derived from the trained model identify the behavioral variables that contribute most significantly to cluster differentiation.

This analysis provides valuable insight into the behavioral drivers underlying customer segmentation without altering the unsupervised nature.

5.10 Customer Persona Development

The final stage of the proposed methodology converts statistically generated customer clusters into business-oriented customer personas.

Each cluster is interpreted according to its dominant behavioral characteristics and assigned descriptive labels representing distinct purchasing behaviors. These customer personas provide actionable insights for customer relationship management, personalized marketing, promotional strategy development, and customer retention planning.

Rather than relying on numerical cluster identifiers, customer personas facilitate intuitive interpretation of the segmentation results by business stakeholders.

5.11 Proposed Methodology Workflow

The complete methodology adopted in this research can be summarized as follows:

1. Construction of the customer master dataset from transaction-level records.
2. Engineering of behavioral customer features.
3. Data preprocessing and feature standardization.
4. Exploratory data analysis.
5. Dimensionality reduction using Principal Component Analysis.
6. Customer segmentation using K-Means clustering.
7. Validation using Hierarchical Agglomerative Clustering.
8. Cluster profiling and customer persona generation.
9. Feature importance analysis for interpretation.
10. Business insight generation for marketing decision support.

The suggested paradigm unifies machine learning, statistical analysis, and business interpretation into a single consumer analytics pipeline. By focusing on engineered behavioral characteristics instead of demographic attributes or unavailable campaign-response labels, the methodology provides a practical and scalable approach for behavior-driven customer segmentation within retail environments.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

CHAPTER 6: EXPERIMENTAL DESIGN

6.1 Experimental Environment

The proposed customer segmentation framework was implemented using Python 3.x within a virtual environment to ensure reproducibility and dependency management. Data preprocessing and feature engineering were performed using Pandas and NumPy, while machine learning algorithms were implemented using the Scikit-learn library. Hierarchical clustering analysis utilized SciPy, and data visualization was performed using Plotly and Matplotlib.

The experiments were conducted on a standard desktop computing environment without GPU acceleration, as the adopted unsupervised learning algorithms are computationally efficient for medium-scale datasets.

Table 6.1. Experimental Environment

Component	Specification
Programming Language	Python 3.x
Development Environment	Virtual Environment
Data Processing	Pandas, NumPy
Machine Learning	Scikit-learn
Hierarchical Clustering	SciPy
Visualization	Plotly, Matplotlib
Dataset Size	Approximately 50,000 customer records
Learning Approach	Unsupervised Machine Learning
Dimensionality Reduction	Principal Component Analysis (PCA)
Primary Segmentation Algorithm	K-Means Clustering
Validation Techniques	Elbow Method, Silhouette Score, Davies–Bouldin Index, Calinski–Harabasz Index

6.2 Dataset Description

The experiments were conducted using an engineered customer master dataset derived from retail transaction records. Individual transaction-level data were aggregated to construct customer-level behavioural profiles, where each record represents a unique household.

The final analytical dataset consists of **50,000 customer records** and **49 engineered features** representing multiple dimensions of customer purchasing behaviour. These include customer activity, spending behaviour, basket characteristics, product diversity, shopping patterns, promotional behaviour, and composite behavioural scores. Identifier variables and non-informative attributes were excluded prior to modelling.

The behavioural features provide a comprehensive representation of customer purchasing patterns and form the basis for the subsequent clustering analysis.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Table 6.2. Categories of Engineered Customer Features

Feature Category	Representative Variables	Purpose
Customer Activity	Recency, Frequency, Purchase Velocity, Active Shopping Days	Measure shopping frequency and customer activity
Customer Value	Monetary Value, Basket Value, Total Spending	Quantify customer economic contribution
Basket Behaviour	Basket Size, Basket Units, Maximum Basket Value	Characterize purchasing missions
Product Diversity	Unique Products, Category Diversity, Brand Diversity	Measure purchasing breadth
Shopping Behaviour	Weekday Share, Weekend Share, Shopping Consistency	Capture shopping patterns
Promotional Behaviour	Promotion Sensitivity, Coupon Metrics	Assess responsiveness to promotions
Behavioural Scores	Engagement Score, Loyalty Score, Purchasing Power	Summarize overall customer behaviour

6.3 Experimental Configuration

The experimental pipeline consisted of sequential preprocessing, dimensionality reduction, clustering, validation, and customer profiling stages. Before clustering, behavioural features were standardized using **StandardScaler** to eliminate differences in measurement scales. Principal Component Analysis (PCA) was then applied to reduce feature dimensionality while preserving approximately **95% of the total variance**.

Customer segmentation was performed using the K-Means clustering algorithm. The optimal number of clusters was determined using multiple internal validation measures, including the Elbow Method, Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. Hierarchical Agglomerative Clustering was additionally employed to validate the stability of the generated customer groups.

Table 6.3 Experimental Parameters

Parameter	**Value**
Scaling Method	StandardScaler
Dimensionality Reduction	Principal Component Analysis
Variance Retained	95%
Clustering Algorithm	K-Means
Cluster Validation	Hierarchical Agglomerative Clustering
Random State	42
Distance Metric	Euclidean

6.4 Evaluation Metrics

The quality of the generated customer segments was assessed using multiple internal clustering evaluation metrics.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

- **Elbow Method** was used to determine the optimal number of clusters by analysing the Within-Cluster Sum of Squares (WCSS).

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

- **Silhouette Score** measured the compactness and separation of customer clusters.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- **Davies–Bouldin Index** evaluated cluster similarity, where lower values indicate better-defined clusters.

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

- **Calinski–Harabasz Index** assessed the ratio between inter-cluster and intra-cluster variance, with higher values indicating improved clustering quality.

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right)$$

The combination of these metrics ensured that the selected clustering solution was both statistically robust and business interpretable.

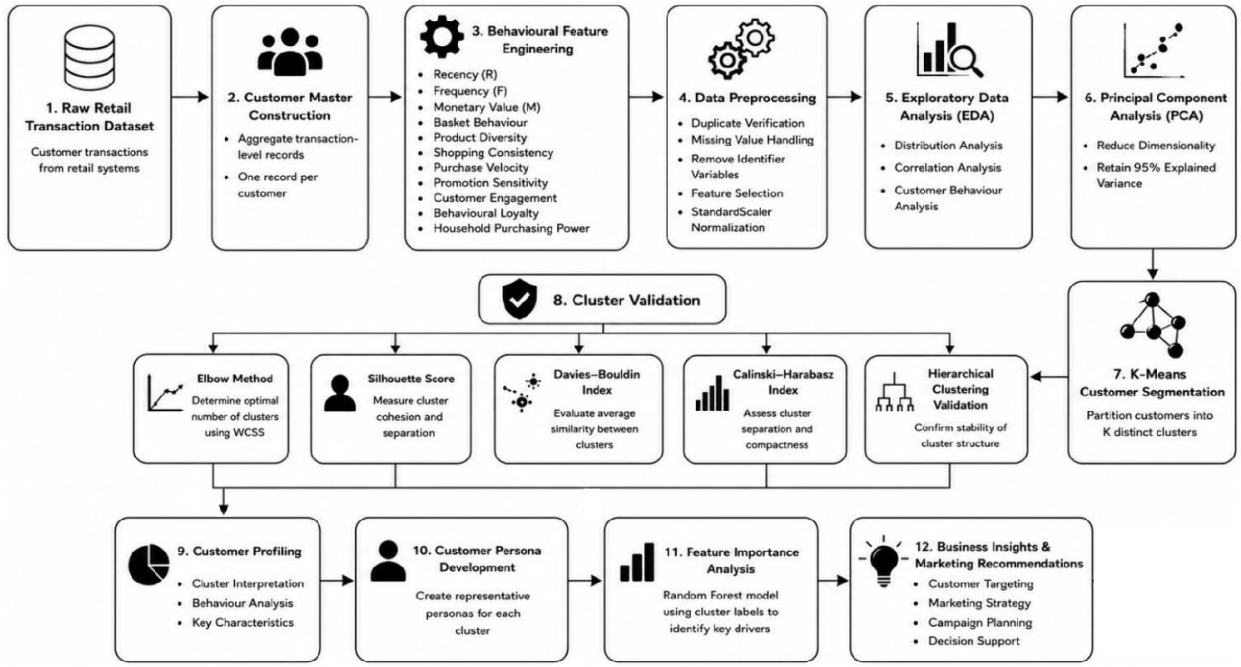
6.5 Experimental Workflow

The complete experimental workflow adopted in this research is illustrated in **Figure 6.1**.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 6.1 Overall Experimental Workflow



The workflow begins with the construction of the customer master dataset, followed by feature engineering, preprocessing, feature scaling, and dimensionality reduction using Principal Component Analysis. The reduced feature space is subsequently analysed using K-Means clustering, validated through hierarchical clustering, and finally interpreted through customer profiling, persona development, and business insight generation.

7. EXPERIMENTAL RESULTS

7.1 Dataset Validation

Before performing exploratory data analysis and applying machine learning techniques, the engineered customer master dataset was validated to ensure its completeness, consistency, and suitability for customer segmentation. Dataset validation is a critical step in unsupervised learning because the quality of the input data directly influences the reliability and interpretability of the resulting customer segments.

The validation process focused on assessing the structural integrity of the dataset rather than performing additional preprocessing operations. Since the customer master dataset had already undergone feature engineering and preprocessing during the experimental design stage, this phase aimed to verify that the final analytical dataset was complete, free from inconsistencies, and ready for exploratory analysis and clustering.

The validation included an assessment of the number of customer records, feature composition, data completeness, duplicate observations, memory utilization, and the distribution of feature types. These evaluations confirm that the dataset satisfies the requirements for large-scale behaviour-driven customer segmentation.

7.1.1 Dataset Summary

The final customer master dataset consists of 50,000 unique customer records, each representing a customer through engineered behavioural attributes. It includes numerical and categorical features capturing purchasing behaviour, shopping patterns, product diversity, promotional responsiveness, and composite behavioural scores. The dataset occupies approximately 24.65 MB, contains no missing values or duplicate records, and is well suited for subsequent machine learning analysis.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Table 7.1 Dataset Summary

Parameter	Value
Total Customer Records	50,000
Total Features	45
Numerical Features	42
Categorical Features	3
Missing Values	0
Duplicate Records	0
Memory Usage	24.65 MB

Table 7.1 summarizes the principal characteristics of the engineered customer master dataset used throughout this research.

7.1.2 Feature Categories

Instead of relying solely on traditional Recency–Frequency–Monetary (RFM) variables, the engineered customer master captures multiple dimensions of customer purchasing behaviour. The behavioural variables were organized into functional categories to provide a comprehensive representation of customer shopping patterns and purchasing habits.

Table 7.2 Categories of Engineered Behavioural Features

Feature Category	**Representative Variables**	**Purpose**
Customer Activity	Recency, Frequency, Active Shopping Days, Purchase Velocity	Measure customer activity and shopping frequency
Customer Value	Monetary Value, Total Spend, Average Basket Value	Quantify customer economic contribution
Basket Behaviour	Basket Size, Basket Units, Maximum Basket Value	Characterize purchasing missions
Product Diversity	Unique Products, Category Diversity, Brand Diversity, Store Diversity	Measure purchasing breadth
Shopping Behaviour	Shopping Consistency, Weekday Share, Weekend Share	Capture purchasing patterns

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Promotional Behaviour	Promotion Sensitivity, Premium Price Orientation	Evaluate pricing and promotional responsiveness
Behavioural Scores	Customer Engagement Score, Behavioural Loyalty Score, Household Purchasing Power	Summarize multidimensional customer behaviour

The engineered feature categories collectively describe multiple aspects of customer purchasing behaviour and provide richer customer representations than conventional segmentation approaches based solely on RFM analysis.

7.1.3 Dataset Quality Assessment

A comprehensive quality assessment was performed to verify the integrity of the customer master dataset prior to exploratory analysis.

The validation confirmed that the dataset contains **50,000 unique customer records**, with no duplicate observations detected. Furthermore, all behavioural variables were complete, and no missing values were present across the analytical dataset. This level of completeness eliminates the need for additional imputation or record removal before machine learning.

The majority of variables are numerical, making the dataset well suited for statistical analysis, dimensionality reduction, and distance-based clustering algorithms. The limited number of categorical variables consists primarily of descriptive identifiers and supporting information rather than modelling features.

The relatively compact memory footprint of the dataset also enables efficient execution of preprocessing, dimensionality reduction, and clustering algorithms on conventional desktop computing environments.

7.1.5 Validation Summary

The dataset validation confirms that the engineered customer master is complete, internally consistent, and suitable for subsequent analytical stages. The absence of missing values and duplicate records ensures that the dataset accurately represents the underlying customer population without introducing unnecessary bias into the clustering process.

Furthermore, the engineered behavioural variables capture diverse aspects of customer purchasing behaviour, including spending patterns, shopping frequency, basket composition, purchasing consistency, and customer engagement. These characteristics provide a comprehensive behavioural representation that supports robust exploratory analysis, dimensionality reduction, and unsupervised customer segmentation.

Overall, the validated customer master dataset provides a reliable foundation for the exploratory analyses presented in the following section, where the statistical characteristics and behavioural patterns of the customer population are examined in greater detail.

7.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to examine the statistical characteristics of the engineered customer master dataset prior to dimensionality reduction and clustering. The objective of this phase was to understand customer purchasing behaviour, identify underlying distribution patterns, detect potential outliers, and evaluate relationships among behavioural variables.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

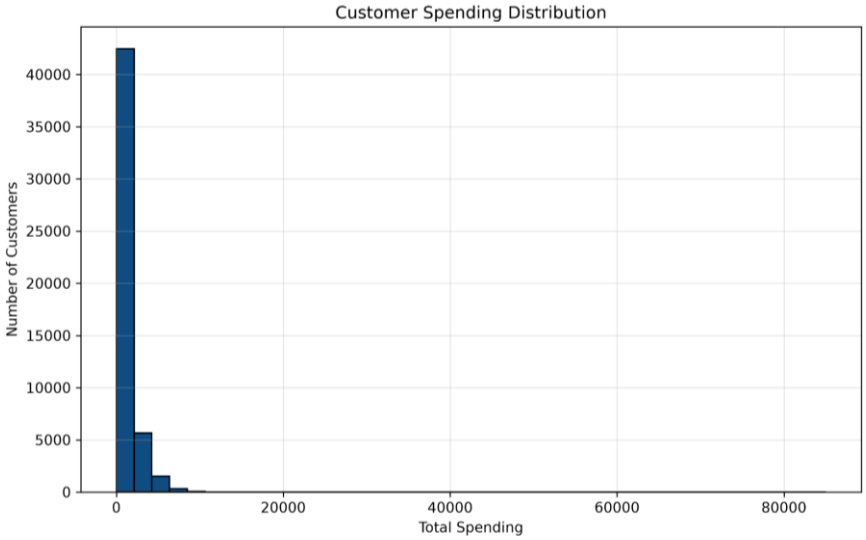
Unlike predictive modelling, customer segmentation relies heavily on understanding the inherent structure of the data before applying unsupervised learning algorithms. Therefore, EDA provides both statistical validation of the engineered features and valuable business insights into customer purchasing behaviour.

The analysis includes spending behaviour, purchasing frequency, customer value, behavioural diversity, feature relationships, and correlation analysis. These observations establish the foundation for the feature selection and clustering stages discussed in the subsequent sections.

7.2.1 Customer Spending Distribution

Customer spending distribution was analysed to understand how revenue is distributed across the customer base. Examining the spending pattern provides an initial indication of customer value concentration and identifies whether a small group of customers contributes disproportionately to total revenue.

Figure 7.1 Customer Spending Distribution



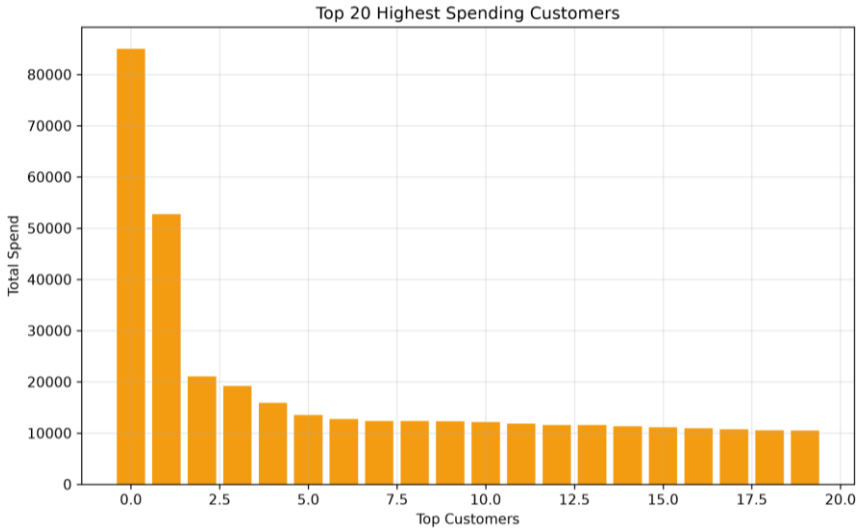
7.2.2 Top Revenue Customers

To identify the concentration of revenue among individual customers, the highest revenue-generating customers were examined.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

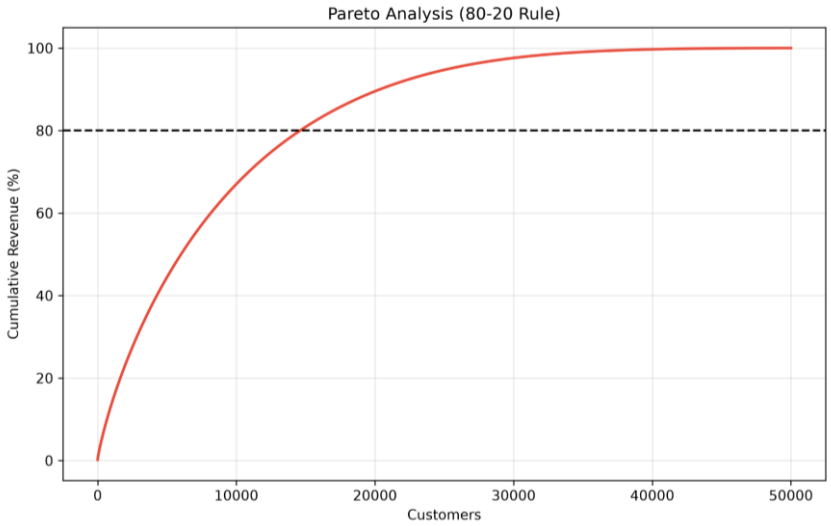
Figure 7.2 Top Revenue-Contributing Customers



7.2.3 Pareto Analysis

The Pareto principle was evaluated to determine the cumulative contribution of customers toward overall business revenue.

Figure 7.3 Pareto Analysis of Customer Revenue



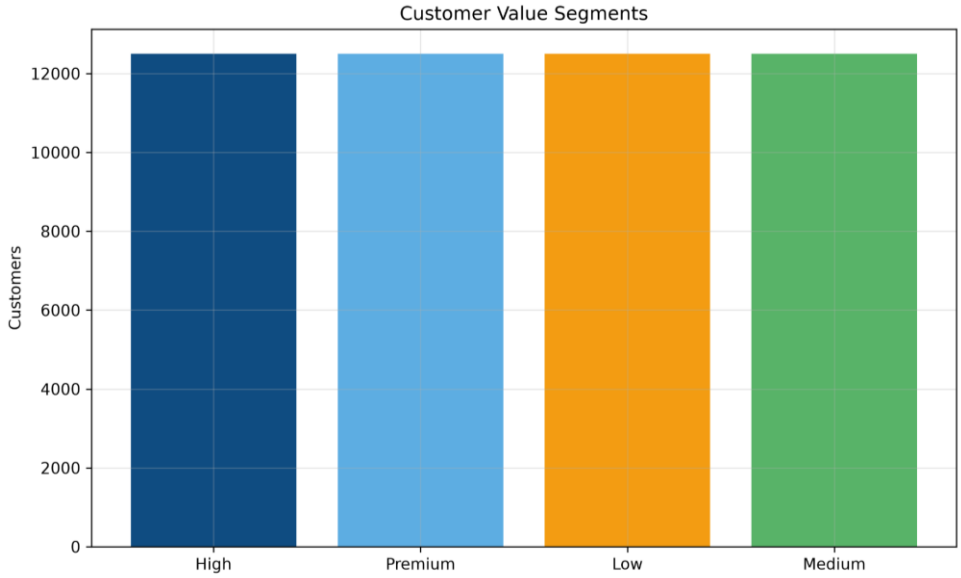
7.2.4 Customer Value Segmentation

Customers were grouped into preliminary value categories based on their purchasing behaviour to provide an initial business-oriented understanding of customer value before clustering.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

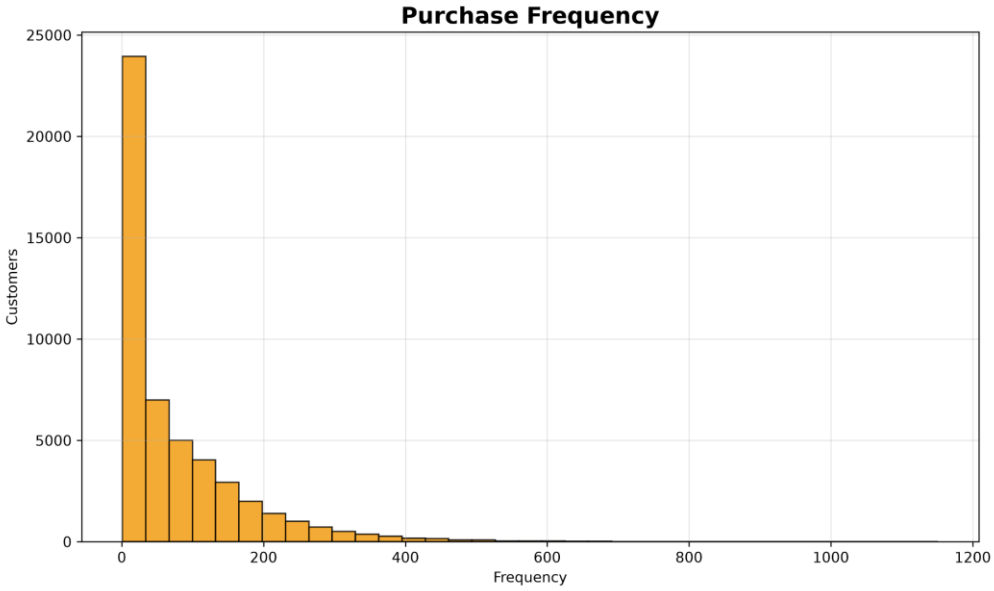
Figure 7.4 Preliminary Customer Value Segments



7.2.5 Behavioural Distribution Analysis

To understand the characteristics of the engineered behavioural variables, distribution analysis was performed for the principal customer metrics.

Figure 7.5 Purchase Frequency Distribution



7.2.6 Product Diversity Analysis

Customer purchasing diversity was examined through product, category, brand, and store diversity metrics.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 7.18: Diversity Distribution

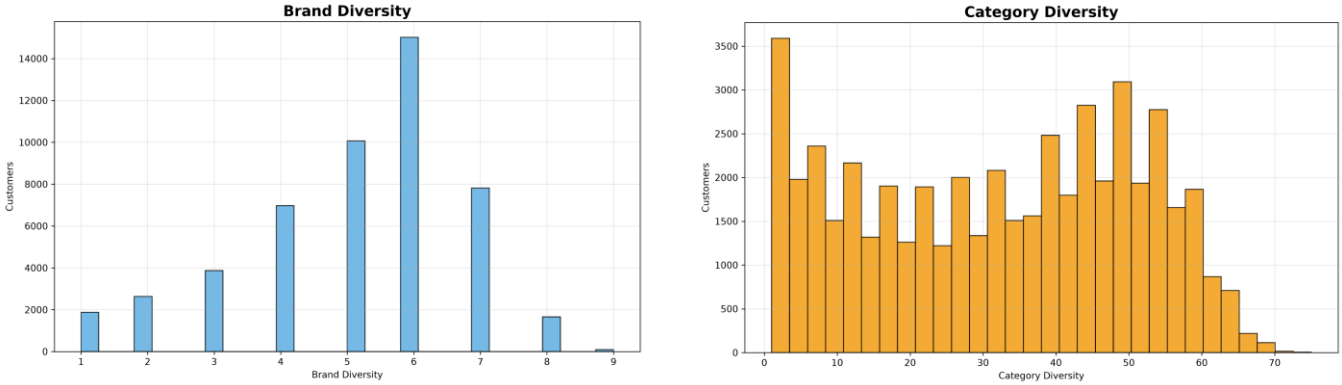
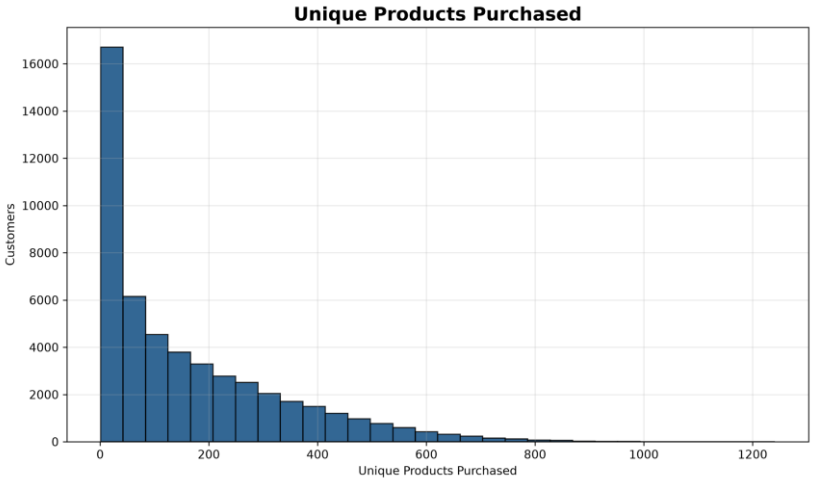


Figure 7.20 Unique Products Purchased Distribution



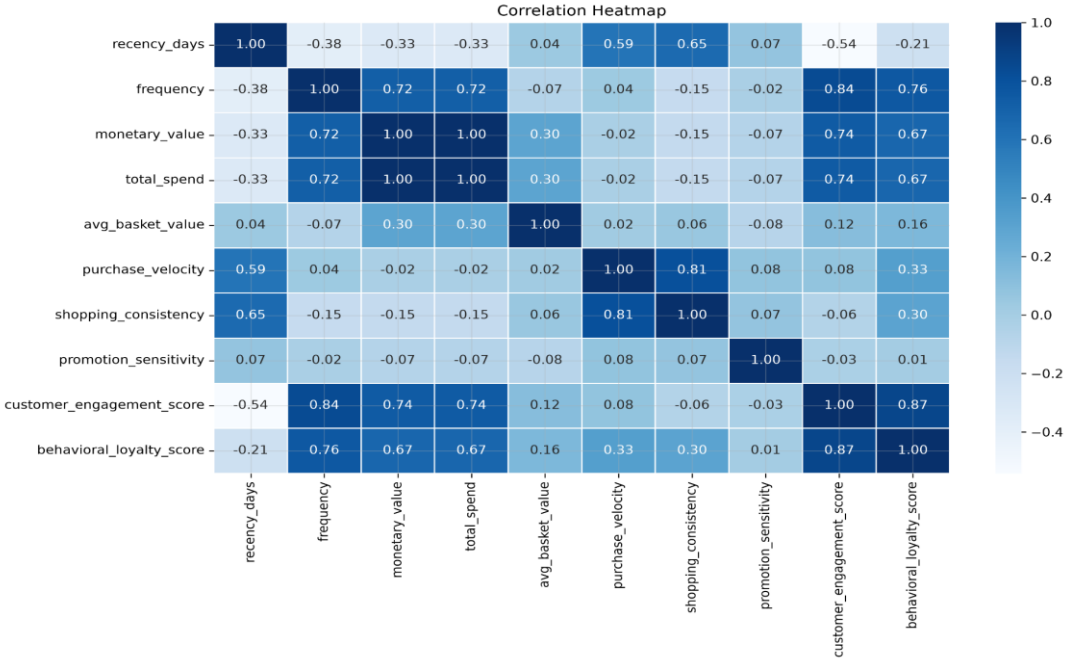
7.2.7 Correlation Analysis

Pearson correlation analysis identified significant behavioral relationships, including strong positive links between spending, engagement, and loyalty. Conversely, recency was negatively associated with activity, as recent shoppers exhibited higher spending and engagement levels. These complementary findings validate the feature set and justify the use of dimensionality reduction in subsequent stages.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

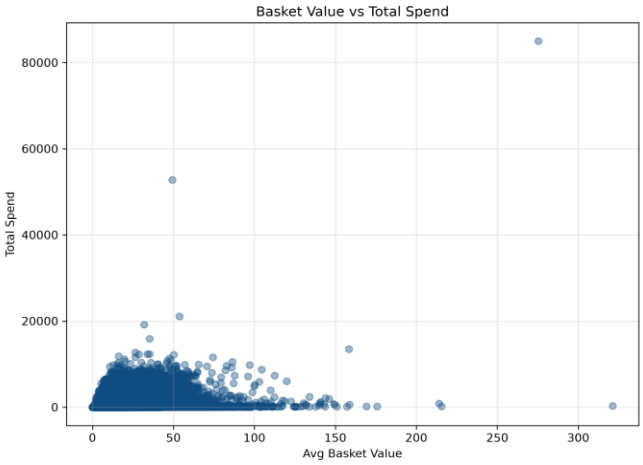
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

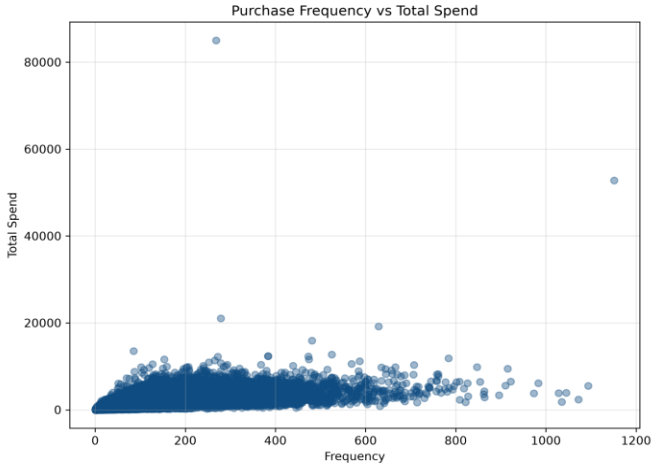
Figure 7.21 Correlation Heatmap of Behavioural Features

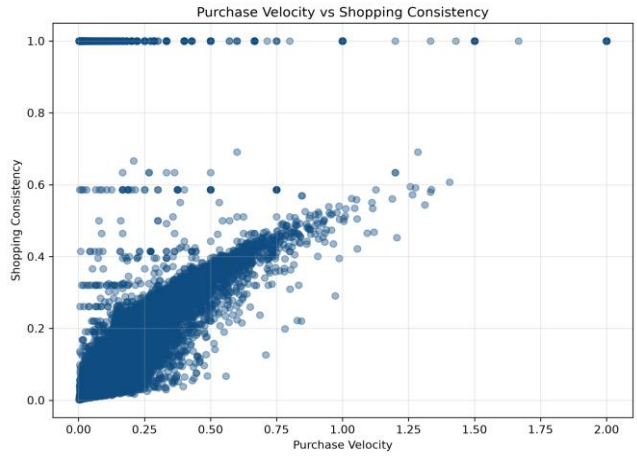


7.2.8 Feature Relationship Analysis

Scatter plot analysis was conducted to investigate interactions between key behavioural variables.





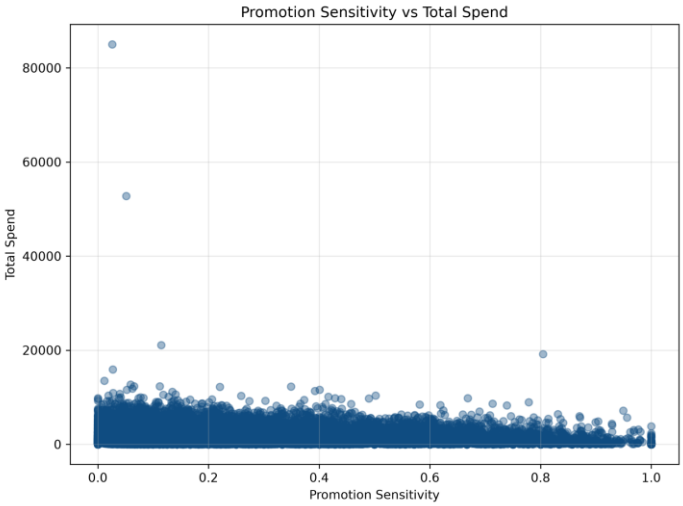


(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

7.2.9 Summary of Exploratory Data Analysis

The exploratory analysis confirms that the engineered customer master dataset captures multiple dimensions of customer purchasing behaviour. Customer spending exhibits a highly skewed distribution, indicating the presence of high-value customer groups, while purchasing frequency,



engagement, and behavioural loyalty emerge as important behavioural characteristics. Product diversity and shopping consistency further reveal meaningful variations in customer purchasing habits, supporting the suitability of the engineered features for customer segmentation.

Overall, the EDA demonstrates that the customer dataset possesses sufficient behavioural diversity and statistical quality to support feature selection, dimensionality reduction, and unsupervised clustering in the subsequent stages.

7.3 Feature Selection

Before applying dimensionality reduction and clustering, feature selection was performed to improve the quality of the customer segmentation model by eliminating redundant, non-informative, and highly correlated variables. Since clustering algorithms rely on similarity measures, the inclusion of irrelevant or duplicate features can adversely affect cluster quality and reduce the interpretability of the resulting customer segments.

A multi-stage feature selection strategy was adopted to preserve meaningful behavioural information while reducing redundancy within the customer master dataset. The selection process involved the removal of identifier variables, low-variance features, and highly correlated attributes that provided overlapping information.

7.3.1 Removal of Non-Informative Features

The first stage involved removing variables that did not contribute meaningful behavioural information for clustering. Variables representing customer identifiers and operational transaction statistics were excluded because they primarily served administrative purposes and did not describe purchasing behaviour.

The following variables were removed:

- Total Transactions
- Transaction Line Count
- Average Days Between Transactions

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

- Standard Deviation of Days Between Transactions
- Transactions per Active Week

These variables were considered redundant because their information was already represented through higher-level behavioural metrics such as purchase frequency, shopping consistency, and purchase velocity.

7.3.2 Removal of Low-Variance Features

The second stage focused on identifying variables exhibiting little or no variation across customers. Features with extremely low variance contribute minimally to cluster formation because they provide limited discriminatory information.

The following promotional variables exhibited negligible variance and were therefore excluded from the modelling dataset:

- Campaign Exposure Count
- Coupons Received
- Coupons Redeemed
- Coupon Redemption Rate

The limited variability of these attributes is consistent with the characteristics of the available dataset, where campaign-related information contained insufficient variation to meaningfully influence customer segmentation.

7.3.3 Correlation-Based Feature Selection

To further reduce redundancy, Pearson correlation analysis was performed on the remaining numerical variables. Highly correlated feature pairs were examined to identify attributes representing similar behavioural information.

The correlation analysis revealed several strong positive relationships among spending-related variables. In particular, **Monetary Value** and **Total Spend** exhibited an almost perfect positive correlation, indicating that both variables represented nearly identical purchasing behaviour.

To avoid multicollinearity while preserving behavioural meaning, **Total Spend** was removed and **Monetary Value** was retained as the representative indicator of customer spending.

Other moderately correlated variables were retained because they captured complementary aspects of customer behaviour rather than duplicating information.

7.3.4 Feature Selection Summary

The multi-stage selection process substantially reduced the dimensionality of the original customer master while preserving the behavioural information required for segmentation.

Table 7.3 Feature Selection Summary

Item	Count
Original Features	45

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Identifier Features Removed	5
Low-Variance Features Removed	4
Highly Correlated Features Removed	1
Final Selected Features	35

The final feature set provides a balanced representation of customer purchasing behaviour across multiple behavioural dimensions, including:

- Customer Activity
- Spending Behaviour
- Basket Characteristics
- Product Diversity
- Shopping Behaviour
- Promotion Sensitivity
- Customer Engagement
- Behavioural Loyalty
- Household Purchasing Power

These variables collectively retain the essential behavioural characteristics required for effective customer segmentation while reducing unnecessary redundancy.

7.3.5 Discussion

The feature selection process reduced the original feature space from **45 engineered variables to 35 representative behavioural features**, resulting in a more compact and informative analytical dataset. Removing redundant and low-information attributes improves clustering efficiency, reduces computational complexity, and minimizes the influence of multicollinearity during subsequent dimensionality reduction.

The retained variables capture diverse aspects of customer purchasing behaviour without introducing excessive overlap between features. Consequently, the selected feature set provides a robust foundation for feature scaling and Principal Component Analysis, discussed in the following section.

7.4 Principal Component Analysis

Principal Component Analysis (PCA) was performed on the standardized behavioural feature set to reduce dimensionality while preserving the maximum amount of information contained within the original dataset. Since several behavioural variables exhibited varying degrees of correlation, PCA was employed to transform the feature space into a smaller set of orthogonal principal components, thereby minimizing redundancy and improving the efficiency of the subsequent clustering process.

The analysis was conducted on the **32 standardized behavioural features** obtained after feature selection. A cumulative explained variance threshold of **95%** was adopted to determine the optimal number of principal components. The PCA results indicated that **15 principal components** were sufficient to retain **95.62%** of the

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

total variance, reducing the dimensionality of the dataset by **53.1%** while preserving nearly all relevant behavioural information.

The transformed principal component dataset was subsequently used as the input for customer segmentation using the K-Means clustering algorithm.

7.4.1 PCA Summary

Table 7.4 Summary of Principal Component Analysis

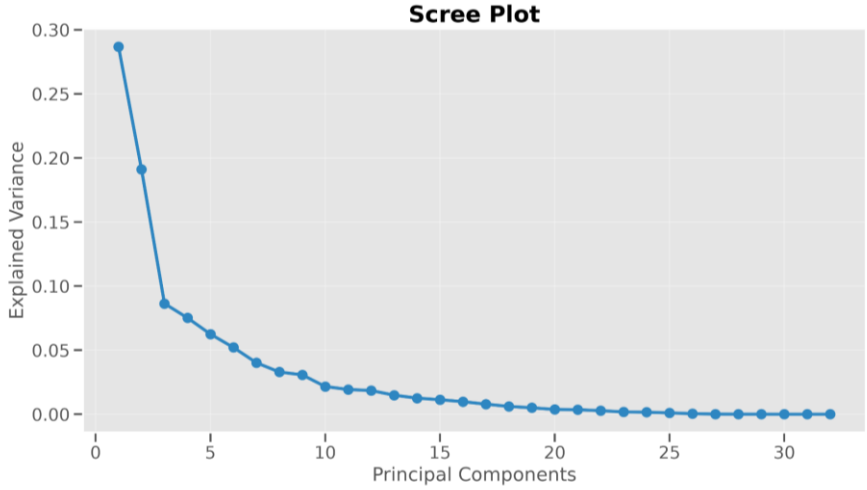
Metric	**Value**
Original Behavioural Features	32
Selected Principal Components	15
Variance Retained	95.62%
Dimensionality Reduction	53.1%

Table 7.4 summarizes the dimensionality reduction achieved through PCA. By reducing the original feature space from 32 behavioural variables to 15 principal components, the model retained the vast majority of customer behavioural information while significantly decreasing feature redundancy.

7.4.2 Scree Plot Analysis

The Scree Plot illustrates the proportion of variance explained by each principal component and assists in identifying the point beyond which additional components contribute only marginal improvements.

Figure 7.28 Scree Plot of Principal Components



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

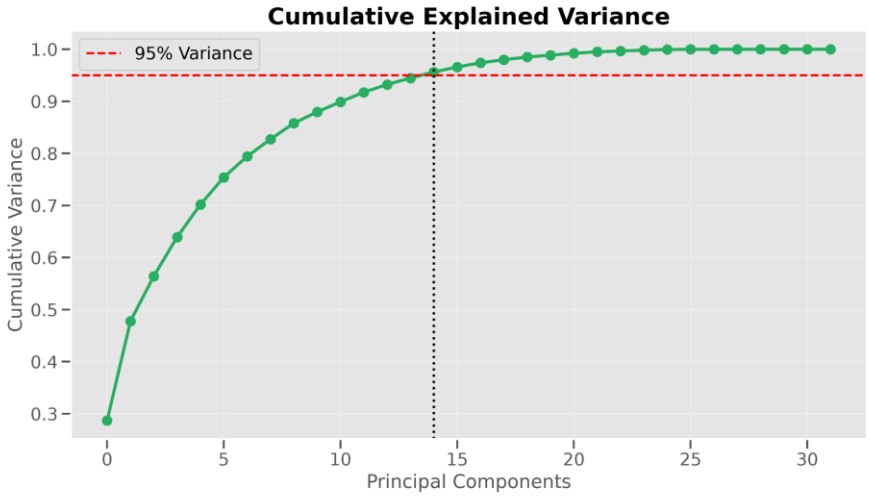
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Initial principal components capture the most variance, with subsequent contributions declining. An elbow appearing between the **10th–15th principal component** signals diminishing returns. This suggests that the first fifteen components encompass the bulk of relevant customer behavior data.

7.4.3 Cumulative Explained Variance

To determine the number of components required for effective dimensionality reduction, the cumulative explained variance was examined.

Figure 7.29 Cumulative Explained Variance



The cumulative explained variance increases rapidly during the initial principal components before gradually approaching saturation. The predefined threshold of **95% variance** is achieved at the **15th principal component**, where the cumulative explained variance reaches **95.62%**. This confirms that the reduced feature space effectively preserves the underlying behavioural structure of the original dataset with minimal information loss.

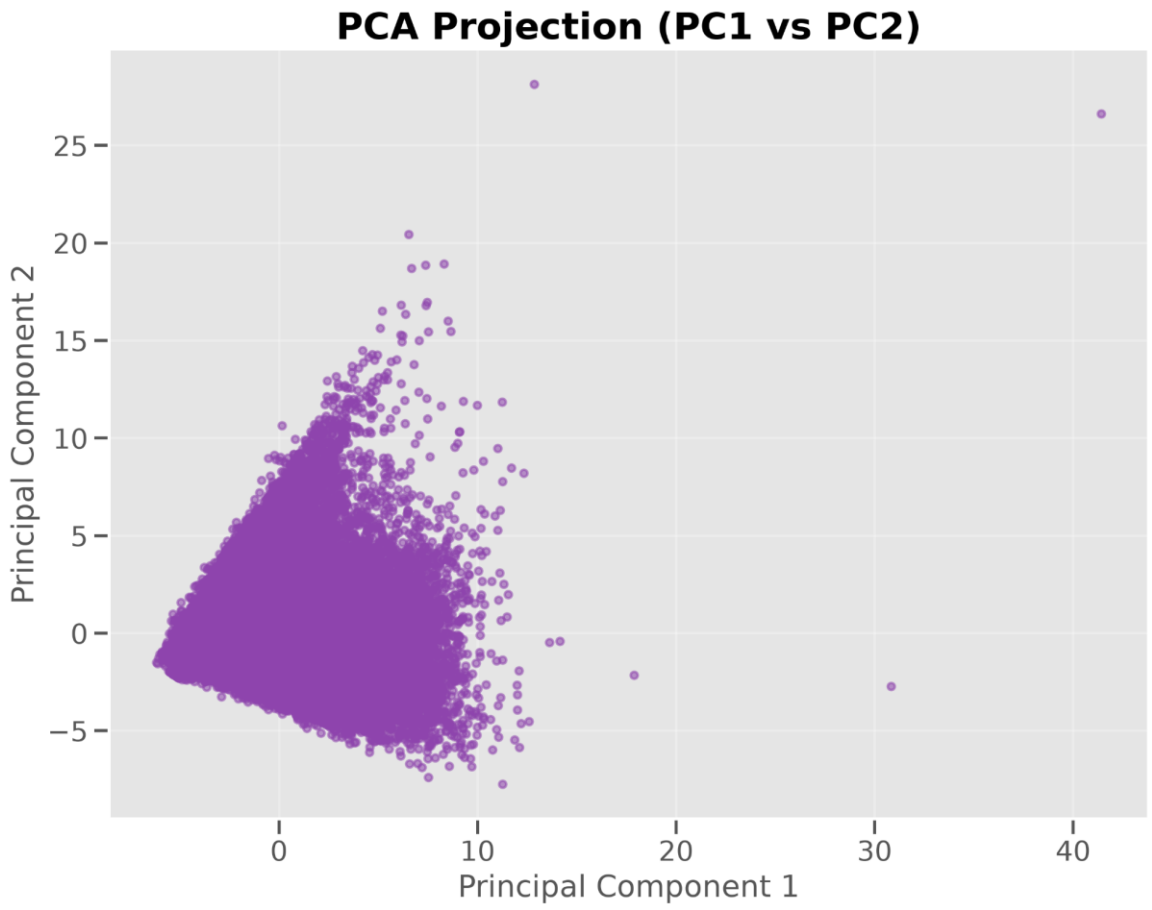
7.4.4 PCA Projection

A two-dimensional projection using the first two principal components was generated to visualize the overall distribution of customer behaviour.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 7.30 PCA Projection of Customer Behaviour



The PCA projection reveals a dense central distribution containing the majority of customers, together with several isolated observations representing customers exhibiting distinct purchasing behaviours. Although the first two principal components provide an intuitive visualization of the behavioural space, considerable overlap remains between observations. Consequently, clustering was performed using all **15 retained principal components** rather than relying solely on the first two dimensions.

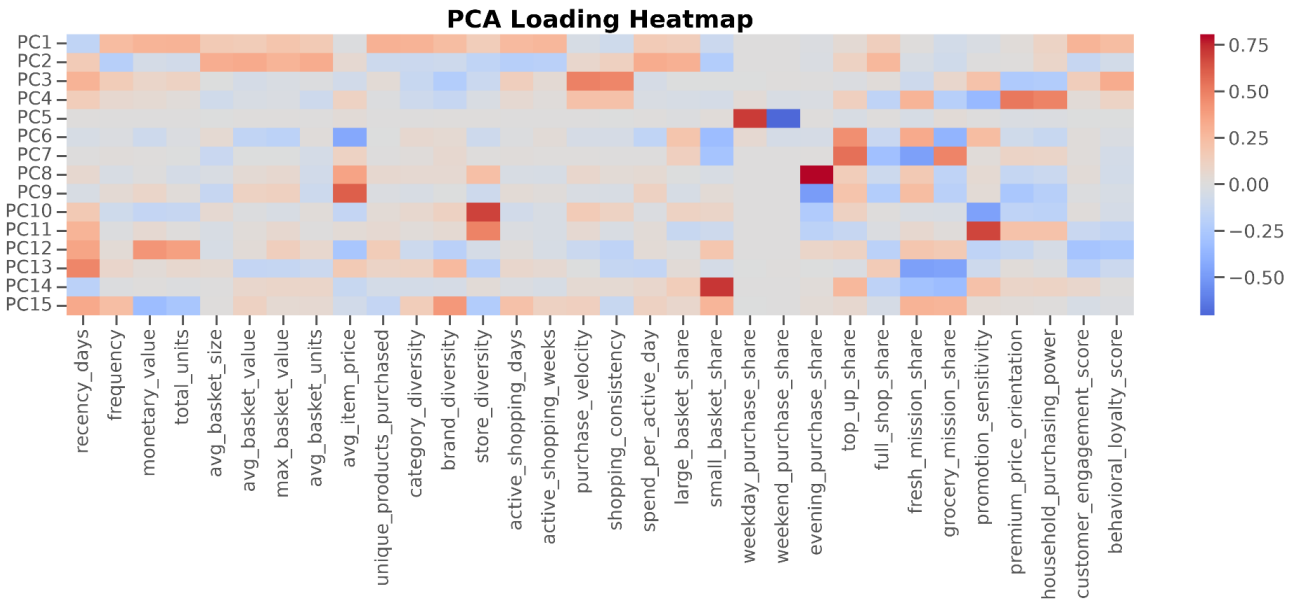
7.4.5 PCA Loading Analysis

The contribution of the original behavioural variables to the retained principal components was examined using the loading matrix.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 7.31 PCA Loading Heatmap



The loading heatmap indicates that variables related to customer spending, purchasing frequency, shopping consistency, customer engagement, behavioural loyalty, and product diversity contribute strongly across multiple principal components. Conversely, temporal purchasing variables such as weekday and weekend purchasing shares exhibit comparatively smaller loading magnitudes. These observations demonstrate that each principal component captures different behavioural characteristics, providing a balanced representation of customer purchasing patterns.

7.4.6 Discussion

The PCA results demonstrate that substantial redundancy existed among the original behavioural variables. By reducing the feature space from **32 variables to 15 principal components**, the dimensionality of the dataset was reduced by more than half while retaining **95.62%** of the original variance. This transformation improves computational efficiency, reduces multicollinearity, and provides a compact representation of customer behaviour without sacrificing essential information.

The reduced principal component space therefore provides an effective input for the clustering algorithms employed in the subsequent stage of customer segmentation.

7.5 Optimal Cluster Determination

Before performing customer segmentation, the optimal number of customer clusters was determined to ensure that the resulting segments were both statistically meaningful and practically interpretable. The transformed dataset containing **15 principal components** obtained through Principal Component Analysis (PCA) was used as the input for cluster optimization.

The optimal value of **K** was evaluated over a range of **2 to 10 clusters** using four complementary internal validation metrics: **Elbow Method (Inertia), Silhouette Coefficient, Calinski–Harabasz Index,** and **Davies–Bouldin Index**. Employing multiple evaluation measures provides a more reliable assessment of clustering quality than relying on a single metric.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

7.5.1 Cluster Validation Results

Table 7.5 Cluster Validation Metrics

K	Inertia	Silhouette Score	Calinski–Harabasz Index	Davies–Bouldin Index
2	1,218,264.14	0.1994	12,788.06	1.6859
3	1,046,901.27	0.2170	11,532.37	1.6761
4	**938,215.05**	**0.2214**	10,509.29	1.7265
5	877,365.19	0.1425	9,295.32	1.7766
6	820,507.58	0.1543	8,644.28	1.6530
7	774,846.92	0.1554	8,118.91	1.6240
8	743,242.42	0.1502	7,558.51	1.6158
9	712,464.12	0.1579	7,169.21	1.6519
10	691,175.66	0.1625	6,739.86	**1.5979**

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

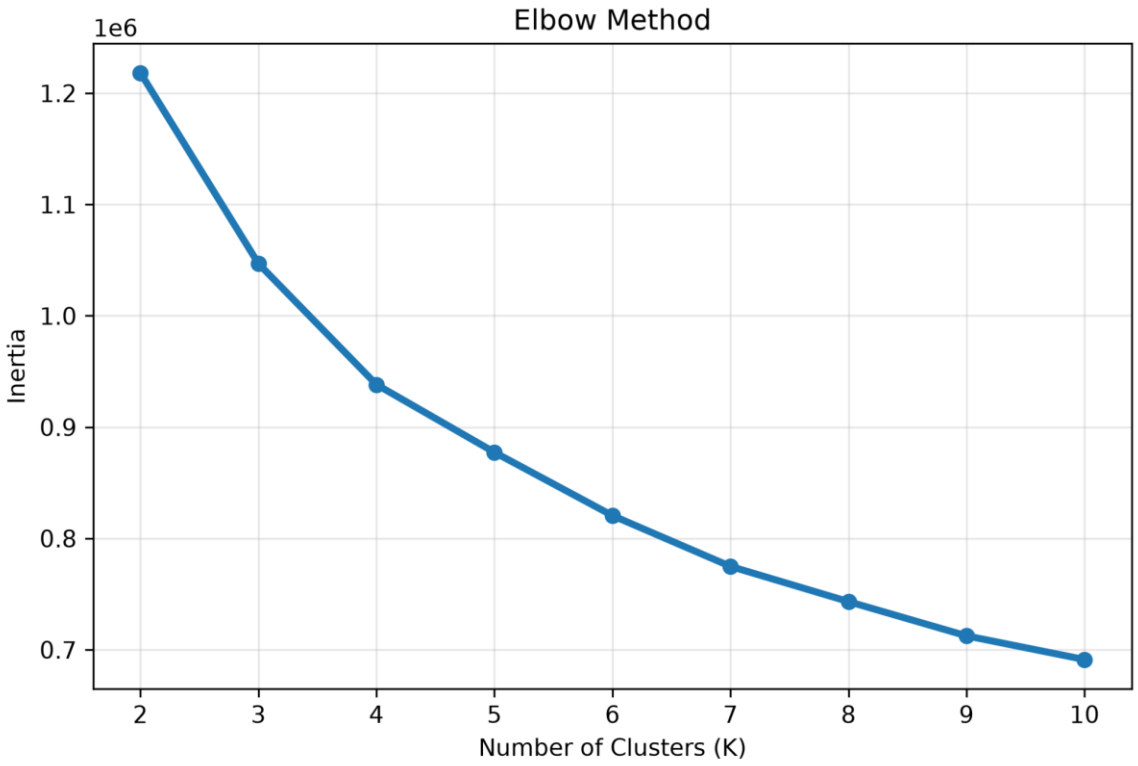
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Table 7.5 summarizes the clustering performance obtained for different values of **K** using the selected validation metrics.

7.5.2 Elbow Method Analysis

The Elbow Method was used to evaluate the reduction in within-cluster variation as the number of clusters increased.

Figure 7.32 Elbow Method for Optimal Cluster Selection



The Elbow Curve demonstrates a substantial reduction in inertia between **K = 2** and **K = 4**, after which the rate of improvement decreases considerably. Although inertia continues to decline with larger values of **K**, the reduction becomes progressively smaller, indicating diminishing returns. This behaviour suggests that four clusters adequately capture the underlying behavioural structure of the customer dataset without introducing unnecessary segmentation complexity.

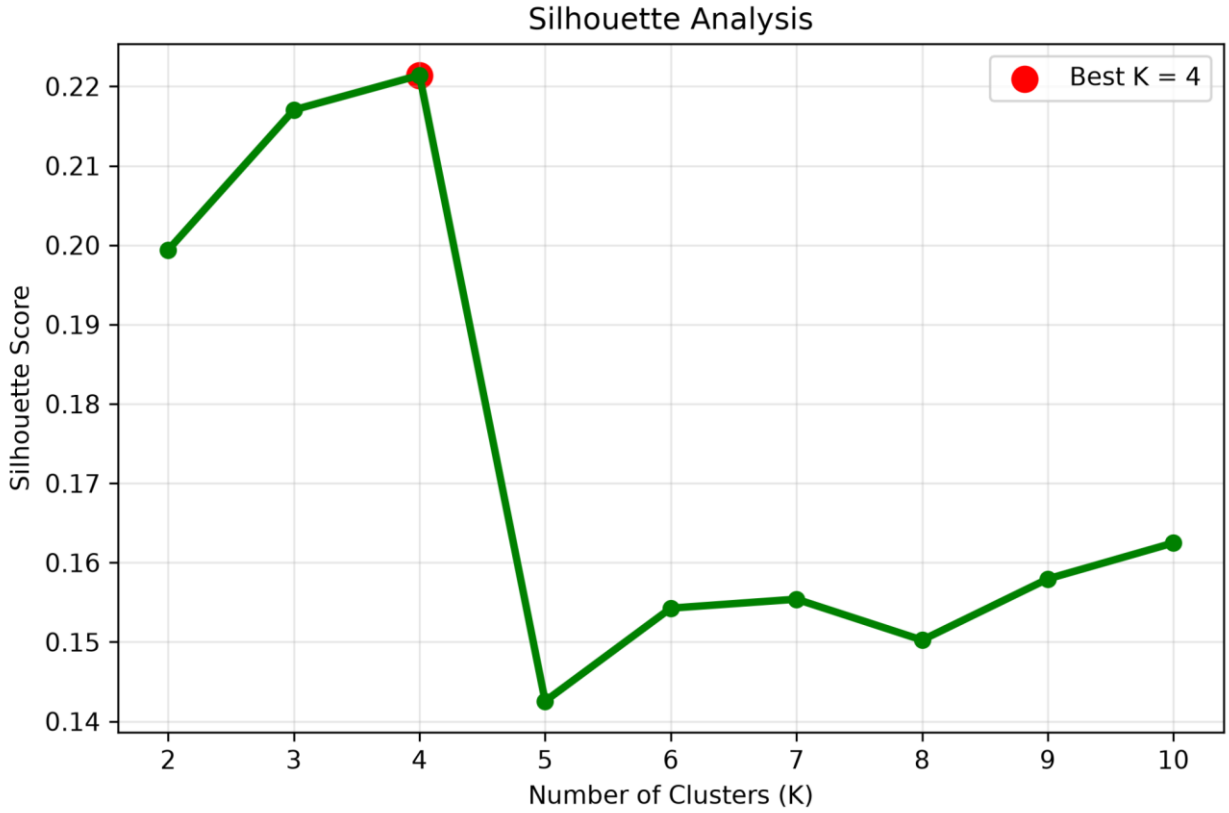
7.5.3 Silhouette Analysis

The Silhouette Coefficient was used to evaluate the balance between intra-cluster cohesion and inter-cluster separation.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 7.33 Silhouette Score Analysis



Among all evaluated clustering configurations, the highest Silhouette Score (**0.2214**) was obtained at **K = 4**. Although the absolute value is moderate, such values are commonly observed in large-scale retail datasets where customer behaviour changes gradually rather than forming perfectly separated groups. The result indicates that four clusters provide the most balanced segmentation of the behavioural feature space.

7.5.4 Calinski–Harabasz and Davies–Bouldin Analysis

The Calinski–Harabasz Index and Davies–Bouldin Index were examined to further assess cluster quality.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 7.34 Calinski–Harabasz Index

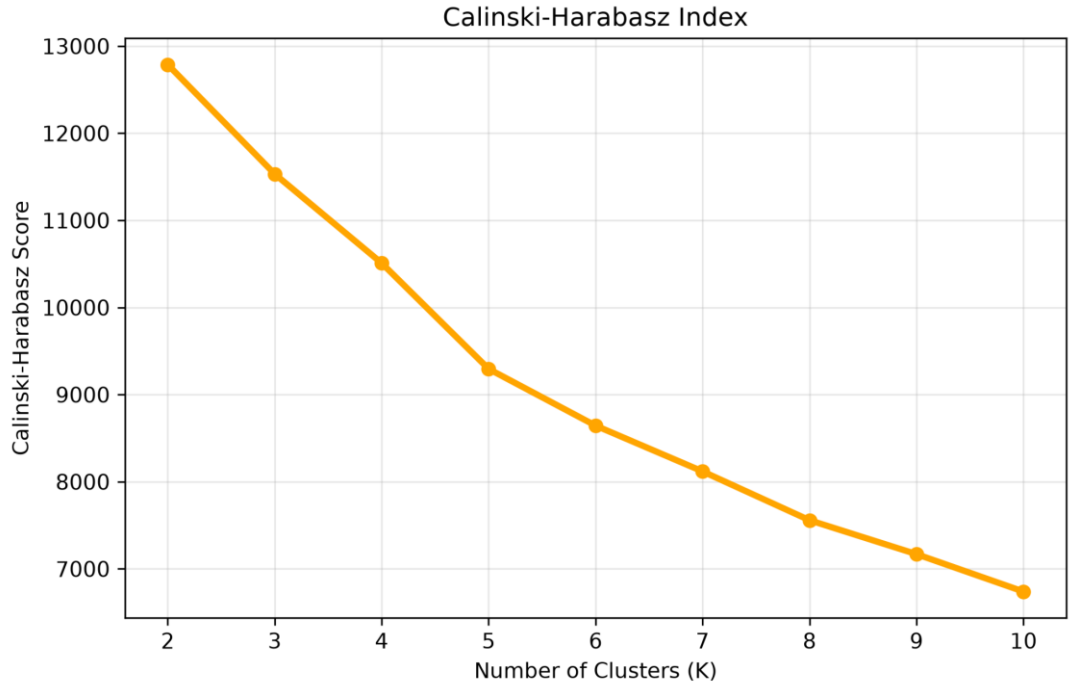
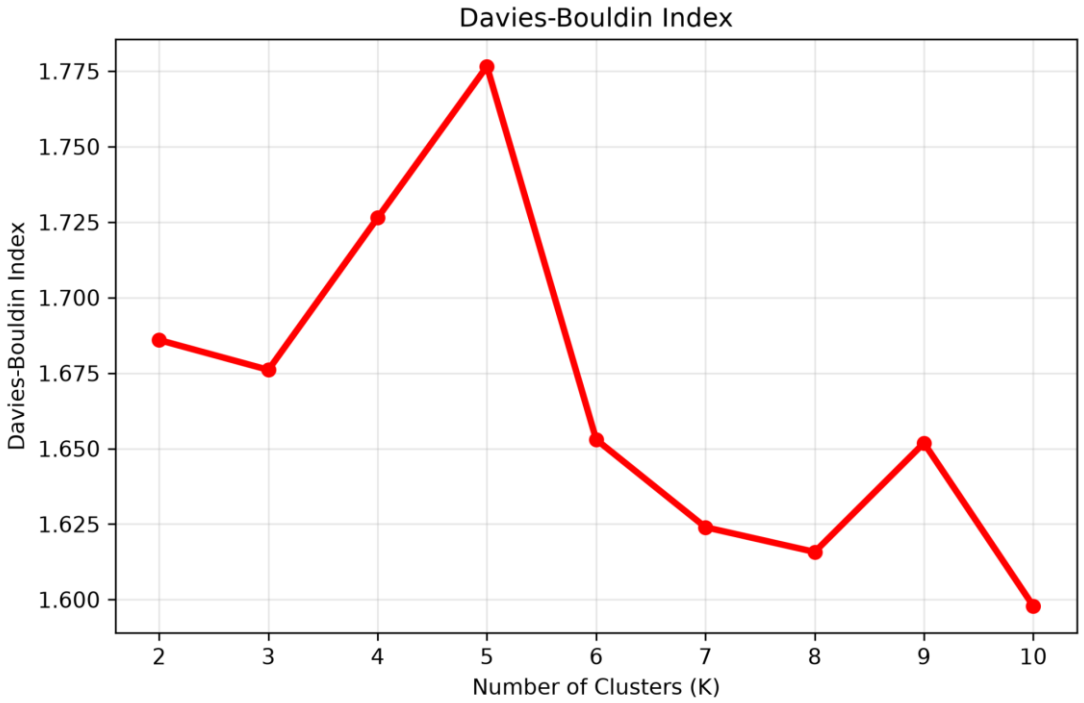


Figure 7.35 Davies–Bouldin Index



The Calinski–Harabasz Index achieved its highest value at **K = 2**, indicating strong statistical separation between two broad customer groups. However, such coarse segmentation provides limited business value by failing to capture the diversity of customer purchasing behaviour.

Conversely, the Davies–Bouldin Index continued to improve as the number of clusters increased, reaching its minimum value at **K = 10**. While this suggests increasingly compact clusters, selecting ten customer groups would substantially reduce interpretability and complicate marketing decision-making.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

7.5.5 Selection of Optimal Number of Clusters

The final number of customer segments was determined through a balanced interpretation of all validation metrics rather than relying on a single evaluation criterion.

The Elbow Method indicated that improvements beyond four clusters were relatively small, while the Silhouette Coefficient achieved its maximum value at **K = 4**. Although the Calinski–Harabasz and Davies–Bouldin indices favoured alternative values of **K**, these metrics alone were not considered sufficient because they prioritise statistical optimisation over business interpretability.

Considering both clustering quality and practical applicability, **four customer clusters** were selected as the optimal segmentation solution for the subsequent analysis.

7.5.6 Discussion

The cluster validation results demonstrate that the selected **four-cluster solution** provides an effective balance between statistical performance and business usability. This level of segmentation is sufficiently detailed to capture meaningful behavioural differences while remaining manageable for customer relationship management, targeted marketing, and strategic decision-making.

The selected clustering configuration therefore forms the basis for the final customer segmentation presented in the following section.

7.6 Customer Segmentation using K-Means Clustering

Following feature selection, feature scaling, dimensionality reduction, and cluster optimization, the K-Means clustering algorithm was applied to the transformed behavioural dataset to identify groups of customers exhibiting similar purchasing behaviour. Based on the cluster validation results presented in the previous section, the optimal number of clusters was selected as **K = 4**.

The clustering model was trained using the **15 principal components** retained after Principal Component Analysis, representing **95.62%** of the total variance in the original behavioural feature space. This reduced feature representation minimizes redundancy while preserving the essential behavioural characteristics required for effective customer segmentation.The objective of this stage was to partition customers into homogeneous behavioural groups that could subsequently support customer-centric business analysis and decision-making.

7.6.1 Cluster Distribution

The final K-Means model successfully segmented the dataset containing **50,000 customers** into four distinct behavioural clusters.

Table 7.6 Distribution of Customer Clusters

Cluster	Number of Customers	Percentage
Cluster 0	7,309	14.62%
Cluster 1	13,642	27.28%
Cluster 2	23,231	46.46%

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Cluster	Number of Customers	Percentage
Cluster 3	5,818	11.64%

The distribution demonstrates that **Cluster 2** represents the largest customer segment, accounting for approximately **46.46%** of the customer population. **Cluster 1** forms the second-largest segment with **27.28%**, while **Clusters 0** and **3** represent comparatively smaller but behaviourally distinct customer groups.The variation in cluster sizes indicates that customer purchasing behaviour is heterogeneous and cannot be adequately represented by a single customer profile.

7.6.2 Cluster Distribution Analysis

Figure 7.36 Customer Distribution Across Clusters

The customer distribution illustrates that nearly half of the customers belong to Cluster 2, whereas the remaining customers are distributed among three additional behavioural groups. The presence of multiple customer segments highlights the diversity of purchasing patterns within the retail dataset and demonstrates the effectiveness of the clustering algorithm in identifying naturally occurring customer groups.

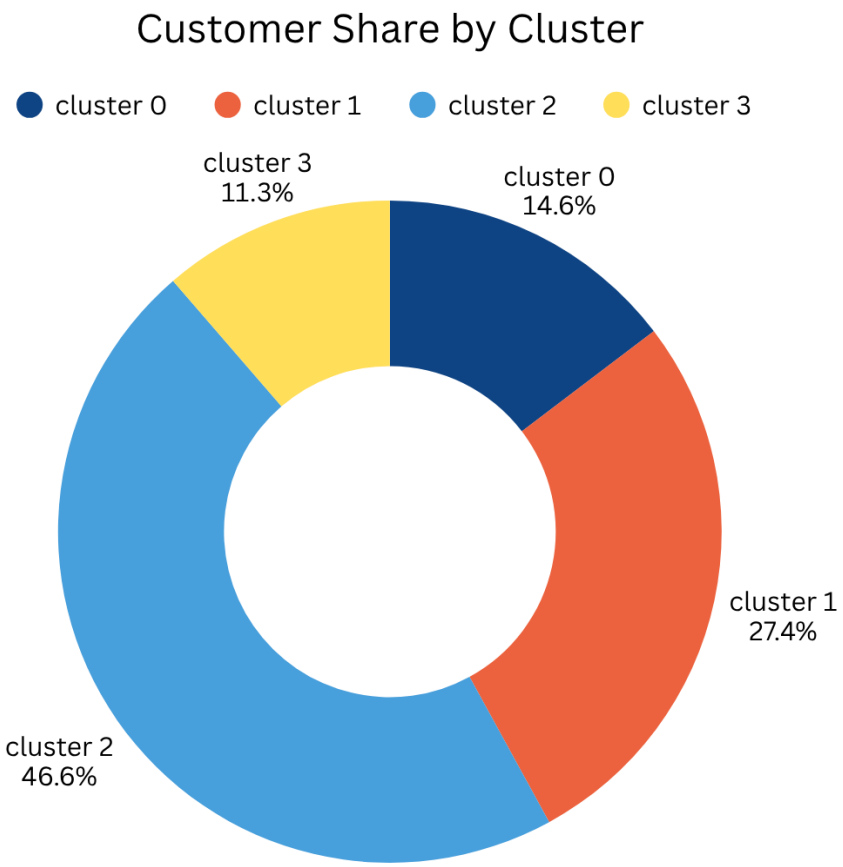
7.6.3 Customer Share Analysis

The proportional distribution further confirms that while one dominant behavioural segment exists, more than half of the customer population belongs to alternative purchasing profiles. This observation supports the need for customer-specific marketing strategies instead of applying uniform promotional campaigns across the entire customer base.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 7.37 Customer Share by Cluster



7.6.4 PCA Visualization of Customer Segments

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 7.38 PCA Visualization of Customer Clusters

The two-dimensional PCA projection provides a visual representation of the four customer segments. Although the clustering algorithm utilised all fifteen retained principal components, the projection onto the first two components illustrates that customers with similar behavioural characteristics are grouped together while maintaining reasonable separation from neighbouring clusters.

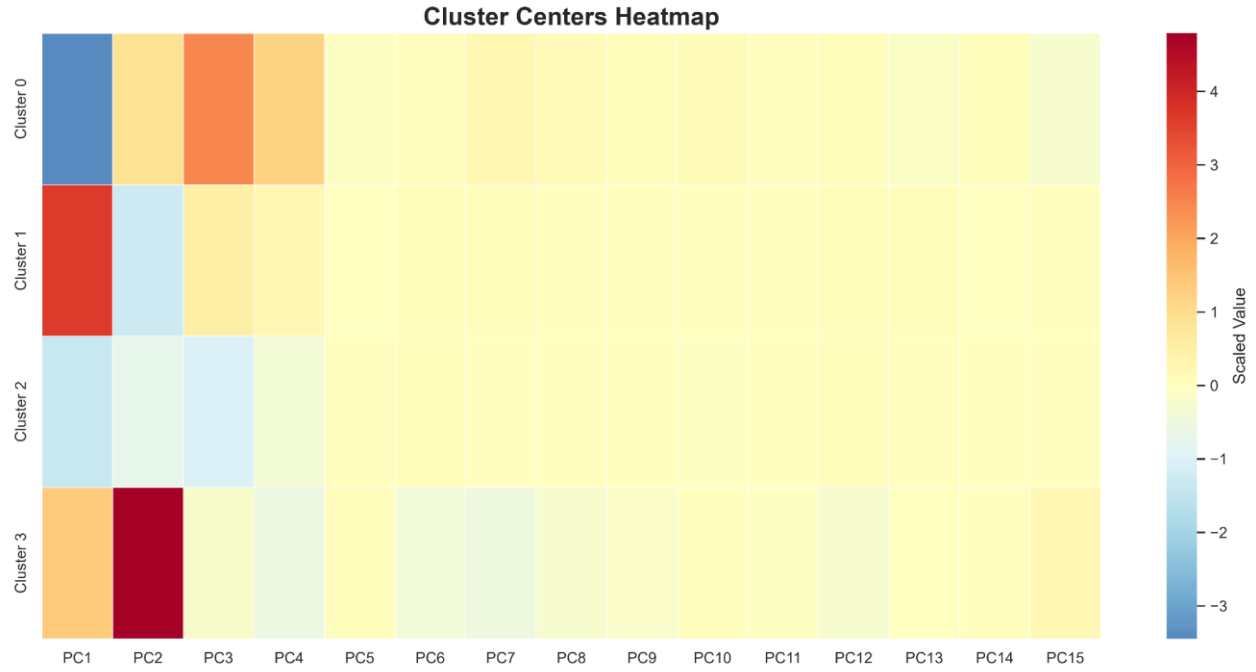
Minor overlap between clusters is expected due to the complex nature of customer purchasing behaviour; however, the visualization confirms that the identified segments exhibit meaningful behavioural differentiation.

7.6.5 Cluster Centre Analysis

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 7.39 Cluster Centre Heatmap



The cluster centre heatmap illustrates the centroid values corresponding to each cluster across the retained principal components. Distinct centroid patterns are observed for all four clusters, indicating that each customer segment possesses a unique multidimensional behavioural signature.

Greater variation is observed within the initial principal components, suggesting that these components contribute most significantly to cluster differentiation, while later components capture more subtle behavioural characteristics.

7.6.6 Discussion

The K-Means clustering algorithm successfully identified four statistically meaningful customer segments exhibiting distinct purchasing behaviours. The resulting clusters demonstrate adequate separation while maintaining sufficient behavioural diversity to support customer segmentation.

The identified customer groups provide a structured foundation for understanding customer purchasing patterns and establish the basis for business-oriented decision-making, including customer targeting, personalized marketing, and relationship management.

7.7 Summary of Experimental Results

This chapter presented the complete implementation and experimental evaluation of the proposed behaviour-driven customer segmentation framework. The engineered customer master dataset was first validated to confirm its completeness and suitability for machine learning analysis. Exploratory Data Analysis revealed significant variation in customer spending behaviour, purchasing frequency, engagement, and shopping consistency, demonstrating the richness of the engineered behavioural features.

A structured feature selection process reduced the original feature space by eliminating redundant and low-information variables while preserving meaningful behavioural characteristics. Principal Component Analysis further reduced the dimensionality of the dataset from **32 behavioural features to 15 principal components**, retaining **95.62%** of the total variance and improving computational efficiency.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Subsequent cluster optimization using multiple validation metrics identified **four** as the optimal number of customer segments. The final K-Means clustering model successfully partitioned **50,000 customers** into four behaviourally distinct groups, demonstrating the effectiveness of the proposed segmentation framework.

Overall, the experimental results confirm that the combination of behavioural feature engineering, dimensionality reduction, and unsupervised clustering provides a robust and scalable approach for customer segmentation. The generated customer groups establish a strong analytical foundation for deriving business insights and supporting data-driven marketing strategies, which are discussed in the subsequent chapter.

8 Business Implications and Practical Applications

8.1 Overview

The customer segmentation framework proposed in this research demonstrates the practical application of machine learning for transforming retail transaction data into actionable customer intelligence. By integrating behavioural feature engineering, dimensionality reduction, and unsupervised clustering, the framework successfully identifies distinct customer segments that reflect diverse purchasing behaviours.

Unlike conventional segmentation approaches that primarily rely on demographic variables or Recency–Frequency–Monetary (RFM) analysis, the proposed methodology captures multiple behavioural dimensions, enabling retailers to develop more precise and data-driven business strategies. The identified customer segments provide valuable support for customer relationship management, marketing optimization, inventory planning, and strategic decision-making.

8.2 Personalized Customer Engagement

The behavioural customer segments generated through the proposed framework enable retailers to replace generalized marketing campaigns with customer-centric engagement strategies. Since each cluster represents customers exhibiting similar purchasing characteristics, organizations can tailor promotional communication, loyalty initiatives, and customer interactions according to the behavioural profile of each segment.

For example, customers demonstrating higher purchasing frequency and stronger behavioural loyalty may be engaged through premium membership programs, exclusive offers, and personalized rewards. Conversely, customer groups exhibiting lower engagement or infrequent purchasing behaviour may benefit from targeted reactivation campaigns designed to increase participation and encourage repeat purchases.

Behaviour-driven customer engagement improves marketing relevance, enhances customer satisfaction, and supports stronger long-term customer relationships.

8.3 Customer Retention and Loyalty Enhancement

The proposed segmentation framework provides organizations with a systematic approach for identifying customers with varying levels of purchasing consistency, engagement, and loyalty.

Rather than implementing identical retention strategies across the entire customer base, businesses can allocate resources according to customer behaviour. High-value customer groups may receive loyalty incentives and personalized services, while emerging or low-engagement segments can be supported through tailored retention initiatives and customer experience improvements.

Such targeted retention strategies contribute to improved customer lifetime value, stronger customer relationships, and more efficient allocation of marketing resources.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

8.4 Promotion and Pricing Optimization

Behavioural attributes such as promotion sensitivity, purchasing frequency, and spending behaviour provide valuable insights for designing more effective promotional strategies.

Instead of applying uniform discount policies, organizations can develop customer-specific promotional campaigns based on behavioural responsiveness. Customers demonstrating higher promotional sensitivity may benefit from targeted discount campaigns, whereas customers exhibiting lower price sensitivity can be engaged through personalized recommendations, exclusive product offerings, or value-added services.

This targeted approach improves promotional efficiency while supporting sustainable revenue growth and profitability.

8.5 Inventory Planning and Product Strategy

The identified customer segments also support inventory management by providing insights into purchasing diversity, basket composition, and shopping behaviour.

Customers exhibiting broader purchasing diversity may require wider product assortments, while specialized customer groups may demonstrate consistent preferences for particular product categories or brands. These behavioural insights enable retailers to optimize assortment planning, improve inventory allocation, and identify opportunities for cross-selling and product bundling.

Consequently, the solution contributes not only to customer analytics but also to operational planning and inventory optimization.

8.6 Strategic Decision Support

The proposed system works as a decision-support system, turning massive amounts of consumer transaction data into interpretable behavioral segments that aid in evidence-based business decisions.

The generated customer groups can assist retail organizations in customer acquisition planning, loyalty management, promotional strategy, pricing decisions, resource allocation, and customer relationship management. By integrating behavioural analytics into business processes, organizations can make informed decisions based on observed customer purchasing patterns rather than intuition alone.

Furthermore, the analytical workflow developed in this research provides a scalable foundation for integrating customer intelligence into enterprise decision-making systems.

8.7 Scalability and Practical Applicability

One of the major strengths of the proposed framework is its adaptability across different retail environments. Since the methodology primarily relies on transactional purchasing behaviour, it can be implemented across supermarkets, e-commerce platforms, departmental stores, pharmacy chains, fashion retailers, and other customer-oriented businesses with minimal modification.

The modular architecture of the framework also enables future integration of additional behavioural variables, advanced analytical techniques, and automated decision-support systems as richer customer data become available. Consequently, the proposed methodology provides a flexible and scalable solution for behaviour-driven customer analytics across a wide range of retail applications.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

9 FUTURE SCOPE

9.1 Framework Enhancement

The proposed customer segmentation framework establishes a robust foundation for behaviour-driven customer analytics and offers several opportunities for future enhancement. As organizations continue to generate increasingly diverse customer data, the framework can be expanded to incorporate additional behavioural indicators, enabling more comprehensive customer understanding and improved segmentation accuracy.Future implementations may integrate demographic attributes, geographic information, online browsing behaviour, customer feedback, and digital interaction data to create richer customer profiles capable of supporting more personalized business strategies.

9.2 Advanced Machine Learning Integration

Although the proposed framework demonstrates the effectiveness of K-Means clustering for behavioural customer segmentation, future studies may explore additional machine learning techniques to investigate alternative customer structures.Methods such as Gaussian Mixture Models, DBSCAN, Spectral Clustering, Self-Organizing Maps, and deep clustering approaches may provide complementary perspectives on customer behaviour, particularly in highly complex retail environments.

9.3 Predictive Customer Analytics

The current framework focuses primarily on descriptive customer segmentation. Future extensions may integrate predictive analytics to further enhance customer relationship management.Potential research directions include customer lifetime value prediction, customer churn prediction, campaign response modelling, personalized recommendation systems, and next-best-action prediction. Combining behavioural segmentation with predictive machine learning models would enable organizations to transition from descriptive analytics toward proactive customer intelligence.

9.4 Real-Time Customer Intelligence

As modern retail increasingly adopts digital platforms, future implementations may incorporate real-time customer transaction streams to support continuous customer segmentation.

The integration of streaming analytics, cloud computing, and interactive business intelligence dashboards would enable organizations to monitor evolving customer behaviour dynamically and respond rapidly to changing purchasing patterns.

Such real-time analytical capabilities would further improve marketing responsiveness and operational decision-making.

9.5 Enterprise Integration

The proposed framework can also be extended through integration with enterprise information systems, including Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), and Marketing Automation platforms.

Embedding behaviour-driven customer segmentation into organizational workflows would support automated customer profiling, intelligent campaign management, personalized customer engagement, and strategic business planning.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

These extensions would further enhance the practical applicability of the framework within modern data-driven retail organizations.

10 CONCLUSION

This research presented a comprehensive behaviour-driven customer segmentation framework designed to support data-driven customer acquisition and business growth within retail environments. Beginning with transaction-level retail data, the proposed methodology systematically integrated customer master construction, behavioural feature engineering, exploratory data analysis, feature selection, dimensionality reduction, and unsupervised machine learning to identify meaningful customer segments.

The experimental evaluation demonstrated that the engineered behavioural features effectively captured multiple dimensions of customer purchasing behaviour. Principal Component Analysis successfully reduced the original behavioural feature space while preserving the majority of the underlying information, and the subsequent K-Means clustering process identified four statistically meaningful customer segments representing distinct purchasing patterns.

Beyond the technical implementation, the study demonstrated how machine learning can generate actionable business intelligence that supports personalized marketing, customer relationship management, promotional optimization, inventory planning, and strategic decision-making. By focusing on behavioural characteristics rather than demographic attributes alone, the proposed framework provides a scalable and adaptable approach to customer analytics that can be applied across diverse retail environments.

Overall, the research successfully achieved its objective of developing an integrated customer segmentation framework that combines statistical analysis, machine learning, and business interpretation into a unified analytical pipeline. The proposed methodology provides a practical foundation for organizations seeking to leverage customer behavioural data to improve decision-making, strengthen customer relationships, and support sustainable business growth.

REFERENCES

[1] Haag, M., Fischer, M., & Gimpel, H. (2022). *Augmented cross-selling through explainable AI -- a case from energy retailing*. arXiv. https://doi.org/10.48550/arXiv.2208.11404

[2] John, J. M., 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

[3] Karimzadeh, S., et al. (2024).An explainable machine learning-based approach for analyzing customers' online data to identify the importance of product attributes. arXiv. https://doi.org/10.48550/arXiv.2402.05949

[4] Langen, H., & Huber, M. (2022). *How causal machine learning can leverage marketing strategies: Assessing and improving the performance of a coupon campaign*. arXiv. https://doi.org/10.48550/arXiv.2204.10820

[5] Zhe Yuan., Consumer Behavior Prediction and Enterprise Precision Marketing Strategy Based on Deep Learning. (2024). Informatica, 48(15). https://doi.org/10.31449/inf.v48i15.6260

[6] Lewaaelhamd, A., et al. (2023). Machine learning-based customer segmentation using RFM analysis and clustering techniques. *Journal of Data Science and Intelligent Systems*, 2(4).https://doi.org/10.47852/bonviewJDSIS32021293

[7] Saurabh Mittal. (2025). Intelligent Customer Acquisition Modeling via Campaign Channel Attribution: A Machine Learning Approach. International Journal of Computational and Experimental Science and Engineering, 11(4). https://doi.org/10.22399/ijcesen.4219

[8] Sun B. (2025). Data-driven personalized marketing strategy optimization based on user behavior modeling and predictive analytics: Sustainable market segmentation and targeting. PloS one, 20(7), e0328151. https://doi.org/10.1371/journal.pone.0328151

[9] Theodorakopoulos, L., Theodoropoulou, A., & Klavdianos, C. (2026). Big Data Analytics and AI for Consumer Behavior in Digital Marketing: Applications, Synthetic and Dark Data, and Future Directions. Big Data and Cognitive Computing, 10(2), 46. https://doi.org/10.3390/bdcc10020046

[10] Wong, C.-G., Tong, G.-K., & Haw, S.-C. (2024). Exploring Customer Segmentation in E-Commerce using RFM Analysis with Clustering Techniques. *Journal of Telecommunications and the Digital Economy*, 12(3).https://doi.org/10.18080/jtde.v12n3.978

[11] Choi, J. (2023). Assessing the predictive performance of machine learning in direct marketing response. In *International Journal of Applied Metaheuristic Computing*. IGI Global. https://doi.org/10.4018/IJEBR.321458

[12] El-Hajj, M., et al. (2024). Predictive Modeling of Customer Response to Marketing Campaigns. *Electronics*, 13(19), 3953. https://doi.org/10.3390/electronics13193953

(This work is licensed under a Creative Commons Attribution 4.0 International License.)