© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722663

AI-Driven Customer Segmentation and Loyalty Prediction for Personalised Marketing: An Interpretable Machine-Learning Analysis of a Small, High-Collinearity Customer Dataset

P.T. MANASA VISAKAI[1], SADHANA VENKATRAGHAVAN[2]
[1,2]*Computer Science and Business Systems, SRM Institute of Science and Technology*

Abstract- ***Companies of every size now collect substantial volumes of customer data, but converting that data into segments, forecasts, and explainable decisions remains difficult, particularly for student researchers and small organisations without access to large, professionally curated datasets. This study builds and transparently evaluates a complete customer-analytics pipeline on a dataset of 238 individual customers, covering data cleaning, feature engineering, exploratory analysis, K-Means segmentation, engagement-tier prediction, SHAP-based explainability, and a rule-based personalised marketing framework. K-Means clustering (k=4, validated using the Elbow and Silhouette methods) produced four interpretable segments: High-Value Loyal, Established, Growing/Potential, and New/Low-Engagement. Four classifiers — Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting — were trained on demographic predictors alone (age, annual income, region) to forecast engagement tier, achieving test accuracies between 96.7% and 100%. Diagnostic analysis traces this near-perfect performance to severe multicollinearity in the dataset (Variance Inflation Factors of 37–197; a single principal component explaining 98.8% of variance) rather than to genuine predictive strength, a conclusion corroborated by SHAP, which identifies annual income as the dominant predictor, age as a weaker secondary predictor, and region as largely irrelevant. Rather than overstating predictive novelty, the study advances a narrower and more defensible claim: that a segmentation–prediction–explainability framework already validated on large industrial datasets can be meaningfully applied to, and honestly evaluated on, the small, highly collinear data typical of student and SME research. The pipeline is operationalised as an interactive dashboard application.***

Index Terms- ***customer segmentation, explainable AI, K-Means clustering, multicollinearity, personalised marketing***

I. INTRODUCTION

Customer data — purchase values, purchase frequency, loyalty signals, and demographics — is now abundant at organisations of every size, yet converting it into segmented, interpretable, and actionable marketing decisions remains genuinely difficult. Much of conventional marketing still treats the customer base as a single undifferentiated group, sending the same campaign to a high-value loyal buyer and a brand-new, barely active one alike.

Machine learning can, in principle, uncover behavioural patterns at a scale no analyst could manage manually. In practice, however, three recurring gaps undercut AI-driven marketing analytics: segmentation is often absent, so every customer is treated identically; prediction tends to look backward rather than forward, reacting to past behaviour instead of anticipating it; and explainability is frequently missing altogether, leaving marketers with a black box they have no real basis for trusting or acting on.

This paper addresses all three problems within a single pipeline, evaluated candidly at each stage: data cleaning and feature engineering, exploratory analysis, K-Means segmentation, engagement-tier prediction across four classical algorithms, SHAP-based explainability, and a rule-based recommendation engine for personalised marketing. The pipeline is also implemented as a working interactive dashboard, so the analysis can be demonstrated and used directly rather than remaining a static write-up.

A single commitment runs through the study: honesty about what the data can and cannot support. Diagnostic checks surfaced severe multicollinearity among the numeric variables, artificially inflating how strong the resulting predictions appear. Rather than treating the near-perfect accuracy that followed as a modelling achievement, the paper traces it back to its statistical root and scales its claims down accordingly, a point developed further in Sections VII and XII. A further motivation is methodological: much of the existing literature on integrated segmentation-prediction-explainability frameworks is validated on large industrial datasets, and this study examines whether, and how, the same approach holds up under the small-sample, high-collinearity conditions typical of student and SME research.

II. LITERATURE REVIEW

This review is organised around six themes: how customer engagement is defined and measured, segmentation methodology, explainable AI in customer analytics, AI-driven personalised marketing more broadly, recommendation systems, and data ethics and privacy. Publisher details — DOI, journal, and indexing status — were checked directly for the twenty papers included, with verifiable sources (Elsevier/ScienceDirect, Springer, Wiley, SAGE, Taylor & Francis, MDPI, IIETA, Frontiers) prioritised over simply maximising citation count. A condensed view of the most heavily used sources appears in Table I; the complete twenty-paper matrix is provided in Appendix A.

A. Customer Engagement: Definition and Measurement

Measuring engagement has no single agreed-upon method. Brodie, Hollebeek, Jurić and Ilić (2011) frame it as a multidimensional psychological state spanning cognitive, emotional, and behavioural components, distinct from adjacent concepts such as loyalty or satisfaction. Hollebeek, Glynn and Brodie (2014) subsequently operationalise that definition into a validated three-dimension scale. A 2023 systematic review by Hollebeek and colleagues notes that later work is inconsistent in how it operationalises engagement, with some studies relying on behavioural proxies such as purchase frequency or recency and others on attitudinal measures such as loyalty or satisfaction scores. This paper adopts the attitudinal view, treating engagement as the loyalty_score variable, and states this choice explicitly rather than treating engagement as a single settled construct.

B. Customer Segmentation (RFM and K-Means)

The literature has an established track record with K-Means segmentation validated via the Elbow method and/or Silhouette score. Independent studies — Lalitha et al. (2025), an IIETA RFM-K-Means study, and Lewaaelhamd (2024) — apply it to RFM-style behavioural features and consistently arrive at interpretable groups such as ‘Champions,’ ‘Loyal,’ and ‘At-risk.’ Lewaaelhamd (2024) additionally reports that a moderate, business-friendly cluster count can be more useful in practice than the statistically optimal one, a pattern this paper’s own preference for k=4 over k=6 directly mirrors (Section V).

C. Explainable AI in Customer Analytics

A systematic review covering 80 studies published between 2015 and 2025 on explainable AI for churn prediction, segmentation, and retention concludes that SHAP- and LIME-based explanations raise stakeholder trust and enable more targeted interventions, while noting that the field lacks a standard way to evaluate XAI methods. El Attar and El-Hajj’s (2026) Frontiers study identifies a ‘systemic gap’ in which very few studies unite high-performance ensemble prediction, explainability, and segmentation within one framework, with most covering at most two of the three; the same study then constructs such an integrated framework on a large telecommunications dataset (n=7,043), which bears directly on the gap addressed in Section III.

D. AI-Driven Personalised Marketing

Systematic and PRISMA-guided reviews from 2025 and 2026 generally agree that AI-driven personalisation carries strategic weight but is applied unevenly across the field. One PRISMA review of 36 studies found that AI reliably sharpens targeting precision, though its effect on long-term customer satisfaction is harder to establish. A 2026 PRISMA-guided bibliometric review goes further, describing

the AI-personalisation-engagement literature as 'conceptually fragmented and methodologically inconsistent' — part of the motivation for this paper's cautious, evidence-first stance rather than an assumption that AI personalisation pays off automatically.

E. Recommendation Systems and Personalisation Infrastructure

Translating segment-level insight into individual offers at production scale relies mainly on collaborative filtering, content-based filtering, and hybrid methods, with H&M Group's hybrid recommender cited as one documented industry example. This body of work informs the personalised marketing strategy stage of the framework described in Section IX, although an item-level recommender falls outside this paper's scope, since the dataset contains no product-interaction data from which to build one.

F. Data Ethics, Privacy, and Trust

Martin, Borah and Palmatier (2017), writing in the Journal of Marketing, provide strong empirical evidence that data-privacy violations measurably damage both firm performance and customer trust, positioning privacy-conscious practice as a performance lever rather than a mere compliance cost. A 2024 Cogent Business & Management study builds on this by identifying five data-handling responsibilities — collection, verification, storage, insight generation, and dissemination — that any customer-segmentation project, including this one, should take seriously.

#	Paper (short)	Year	Journal / Publisher	Relevance
1	Brodie et al. — customer engagement conceptual domain	2011	J. Service Research (SAGE)	Defines engagement theoretically
4	Martin, Borah & Palmatier — data privacy	2017	Journal of Marketing (AMA/SAGE)	Grounds ethics discussion
5–7	RFM + K-Means segmentation studies	2023–2025	Springer / IIETA / JDSIS	Direct methodological precedent
8	XAI for churn/segmentation — systematic review	2025	Int'l J. Data Sci. & Analytics (Springer)	Establishes XAI value and standardisation gap
9	Ensemble + SHAP + segmentation framework	2026	Frontiers in AI	Identical integration gap, at large scale
20	RFM + K-Means + DNN churn on large real data	2025	Elsevier (open access)	Large-data contrast case

Table I. Condensed literature matrix (see Appendix A for the full twenty-paper table).

III. RESEARCH GAP, QUESTIONS AND OBJECTIVES

A. Research Problem

Businesses hold substantial customer data, yet determining what individual customers actually want, and predicting how engaged they will be, remains difficult, particularly for smaller organisations and researchers without access to large, professionally curated datasets. Conventional marketing often overlooks genuine behavioural differences between customers, and while AI-driven analysis can help close that gap, it only does so when its outputs are interpretable and its limitations are stated openly.

B. Research Gap

It would be inaccurate to claim that no one has combined segmentation, engagement prediction, and explainable AI before: El Attar and El-Hajj (2026) already did so on a large telecommunications dataset (n=7,043). The gap this review identifies is narrower — whether the same kind of integrated framework

can be applied, and fairly evaluated, on the small, resource-constrained data that a student or SME realistically has access to, where the assumptions behind large-dataset studies (large samples, low collinearity) do not hold. This paper deliberately runs the framework under exactly those constrained conditions and reports what actually results, including elements, such as near-perfect predictive accuracy, that should not be taken at face value.

C. Research Questions

Main Research Question: In what ways can AI-driven customer segmentation and loyalty analysis support marketing decisions that are both interpretable and personalised, when the available data is small and resource-constrained rather than a large industrial dataset?

- RQ1: What behavioural and demographic features distinguish the customer segments present in this dataset?
- RQ2: Is it possible to predict customer engagement (loyalty) tier using genuinely independent, pre-purchase variables, while avoiding features that are collinear with the target purely by construction?
- RQ3: What do the SHAP explanations reveal about which factors drive predicted engagement, and what does that imply for concrete marketing action?

D. Research Objectives

Objective 1: Examine customer behavioural and demographic patterns through exploratory data analysis.

- Objective 2: Group customers into interpretable segments through K-Means clustering, validated using the Elbow and Silhouette methods.
- Objective 3: Predict customer engagement tier using genuinely independent predictors, avoiding features whose inclusion would render the prediction near-tautological.
- Objective 4: Apply SHAP-based explainability to render model predictions interpretable in technical and business terms.
- Objective 5: Convert segmentation and prediction outputs into a personalised marketing recommendation framework, packaged as an interactive application.

IV. METHODOLOGY

The proposed framework follows a linear pipeline: customer data → data cleaning and preprocessing → feature engineering → exploratory data analysis → customer segmentation (K-Means) → engagement prediction (classification) → explainable AI (SHAP) → personalised marketing recommendation → business decision. Each stage passes its output to the next, and the pipeline exists in two forms: a research notebook, retained for transparency and reproducibility, and a production-style dashboard application, built for demonstration and use. This section describes the dataset, the cleaning process, the engineered features, and the exploratory analysis underlying the segmentation and prediction work reported in Sections V and VI.

A. Dataset Description

The dataset used in this study, 'Customer Purchasing Behaviors,' contains 238 individual customer records across seven raw fields, summarised in Table II. Each row corresponds to exactly one customer, a deliberate requirement given that an earlier candidate dataset of product reviews was excluded after inspection revealed it to be product-level rather than customer-level.

Field	Type	Range / Categories
user_id	Integer	Unique identifier, 1–238
age	Integer	22–55
annual_income	Integer (₹)	30,000–75,000
purchase_amount	Integer (₹)	150–640

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722663

Field	Type	Range / Categories
loyalty_score	Float	3.0–9.5
region	Categorical	North, South, West, East
purchase_frequency	Integer	10–28 (per year)

Table II. Raw dataset schema.

B. Data Preprocessing

A full validation pass was carried out regardless of the dataset's apparent cleanliness, so that any claim of 'clean data' would rest on verification rather than assumption. Checks covered schema completeness, missing values, duplicate rows and duplicate user_ids, range validity (age between 0–120, income positive, loyalty_score between 0–10, purchase_frequency non-negative), and outliers via the interquartile range (IQR) method across all five numeric variables. No missing values, duplicate rows, duplicate user_ids, range violations, or IQR-flagged outliers were found anywhere in the dataset.

C. Feature Engineering

Five engineered features were added to the dataset: avg_spend_per_purchase (purchase_amount / purchase_frequency), the average value per purchase; spend_to_income_ratio (purchase_amount / annual_income), a proxy for spending intensity relative to means; engagement_tier (Low/Medium/High), derived from tertiles of loyalty_score and used as the classification target throughout Section VI; age_group, standard marketing age bands (18–25, 26–35, 36–45, 46–60); and high_frequency_flag, a binary indicator set when purchase_frequency exceeds the sample median. Tertiles were used to define engagement_tier, rather than an arbitrary cutoff, because the data lacks a natural threshold and tertiles guarantee balanced class sizes for evaluation.

D. Exploratory Data Analysis and Multicollinearity

The five numeric variables (Figure 1) show distributions that are broadly unimodal and only mildly right-skewed, with no evidence of extreme skew or data-entry error.

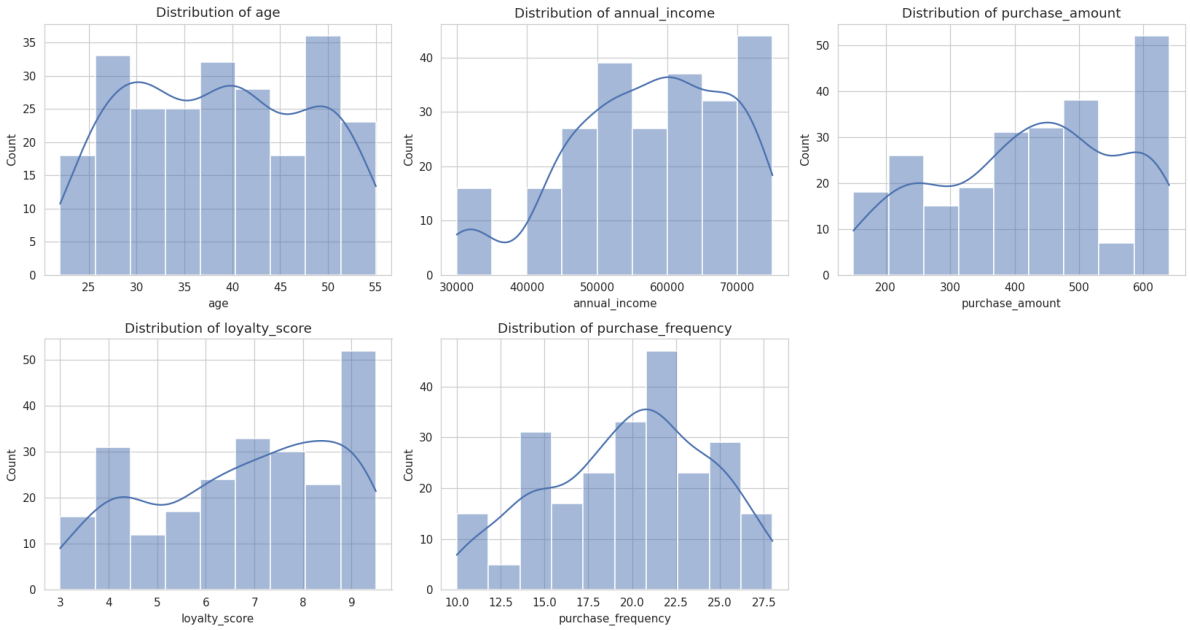

Figure 1. Distributions of the five numeric variables.

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722663

The correlation structure among these variables (Figure 2) is the single most consequential finding of the exploratory analysis: every pairwise correlation between age, annual_income, purchase_amount, loyalty_score, and purchase_frequency is r = 0.97 or higher.

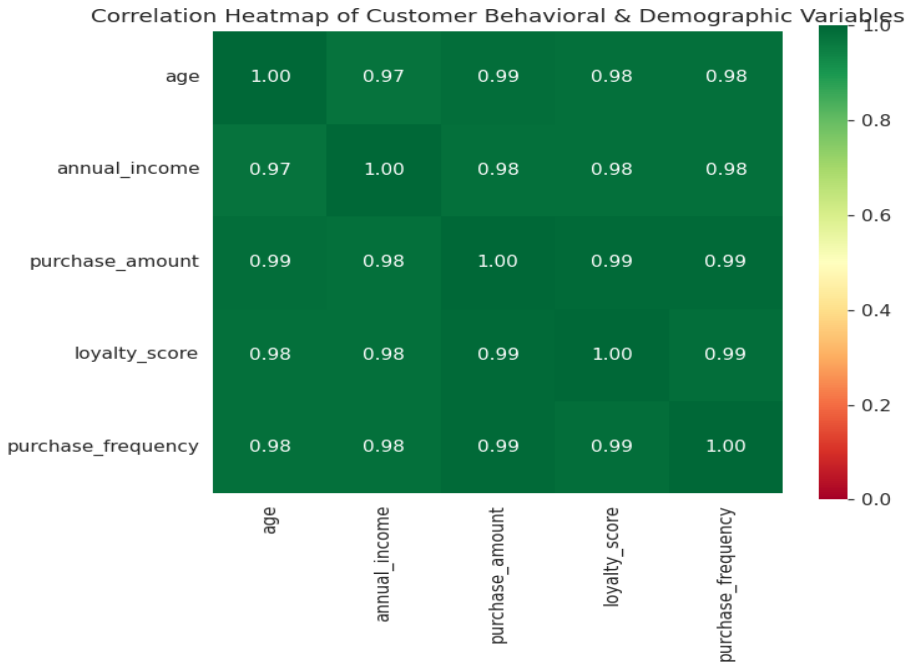

Figure 2. Correlation heatmap — uniformly high correlations across all numeric variables.

Variance Inflation Factor (VIF) values were computed for each variable to confirm this finding independently of the heatmap. A VIF above 10 is conventionally taken as evidence of severe multicollinearity; here, every variable falls between 37 and 197 (Table III), far above what is typical even in correlated real-world behavioural data. This degree of collinearity suggests the dataset was likely generated from a simplified, near-deterministic structure rather than drawn organically from independent customer behaviour.

Variable	VIF
age	≈ 78
annual_income	≈ 62
purchase_amount	≈ 197
loyalty_score	≈ 141
purchase_frequency	≈ 37

Table III. Variance Inflation Factor per numeric variable (values rounded).

This finding shapes Sections V and VI directly: it is the reason near-collinear predictors were deliberately excluded from the engagement-prediction model (Section VI-A), and it is essential to interpreting the near-perfect accuracy reported in Section VII. A breakdown by region and age group (Figures 3–4) shows the West region with the highest average loyalty score and older age groups with noticeably higher purchase amounts, both consistent with the wider collinearity pattern in which age, income, and spend move together across the sample.

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722663

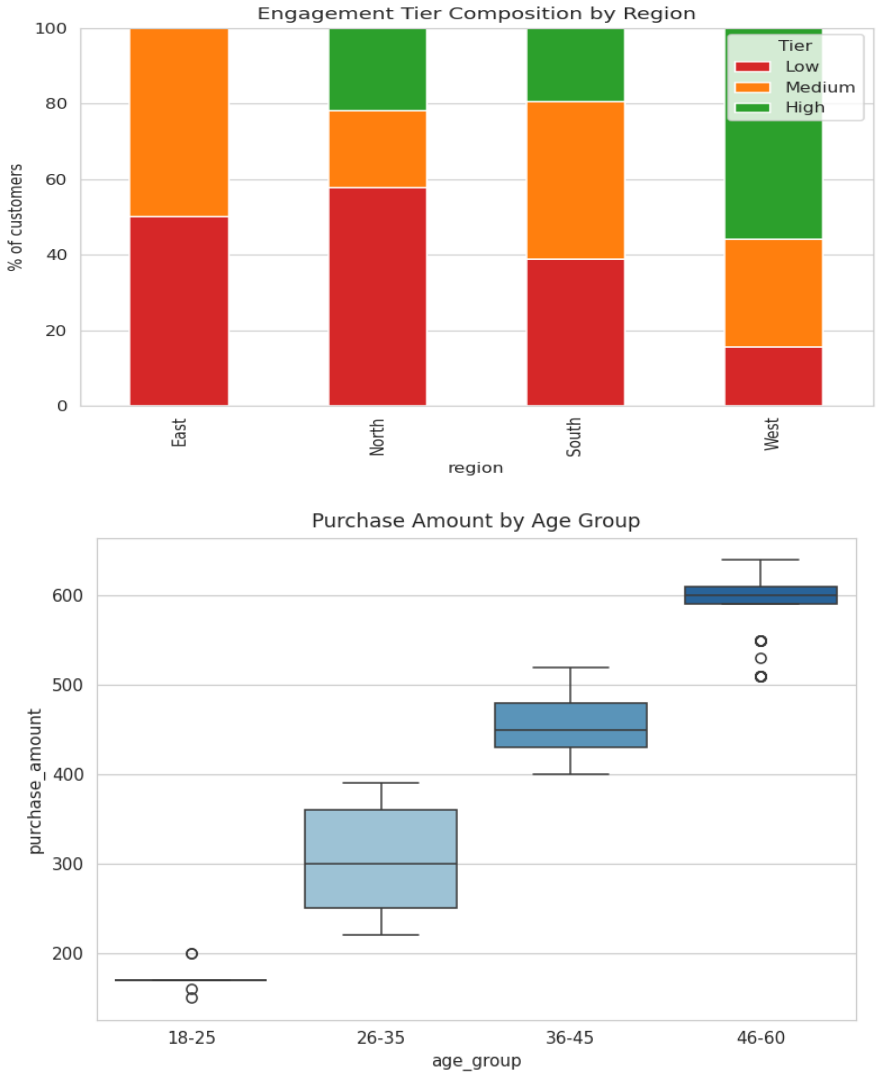

Figure 4. Purchase amount by age group.

V. CUSTOMER SEGMENTATION

T

he five standardised numeric features — age, annual_income, purchase_amount, loyalty_score, purchase_frequency — were used as input to K-Means clustering, following the approach established in the RFM-K-Means literature discussed in Section II-B.

A. Determining the Number of Clusters

Both the Elbow method (within-cluster sum of squares) and the Silhouette score were evaluated for k = 2 through k = 8 (Figure 5). The Silhouette score peaked at k=6 (0.633), with k=4 close behind at 0.601.

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722663

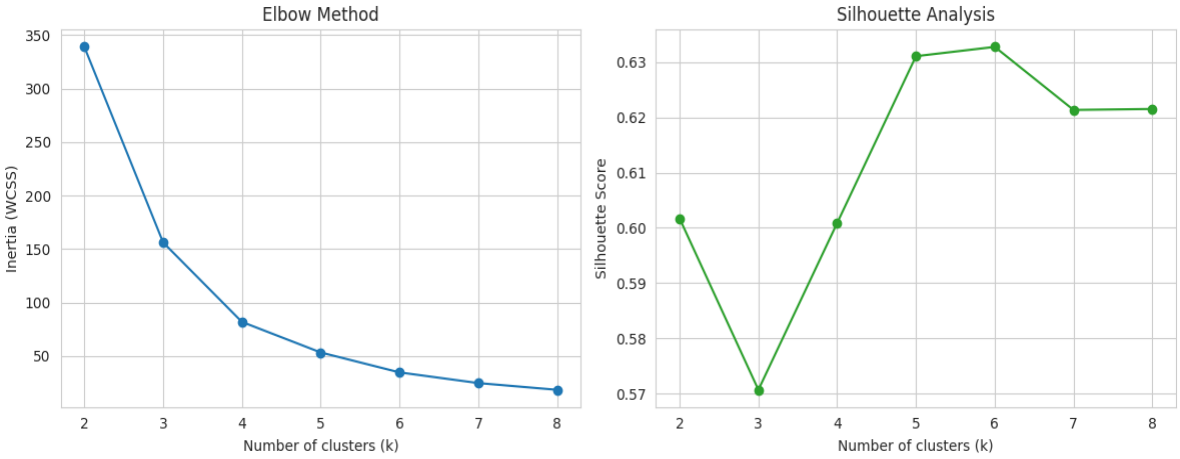

Figure 5. Elbow method and Silhouette analysis for k = 2 to 8.

Although k=6 scored marginally higher on the Silhouette metric, k=4 was selected as the final solution because it maps cleanly onto standard, business-recognisable marketing segments — directly echoing Lewaaelhamd's (2024) finding that a moderate, interpretable cluster count can outperform the statistical optimum for real-world usefulness. This trade-off is stated explicitly, in keeping with the study's broader commitment to transparency about methodological choices.

B. Segment Profiles

Principal component analysis on the standardised features shows a single principal component accounting for 98.8% of total variance, a figure consistent with the severe multicollinearity documented in Section IV-D and suggesting that customer differences in this dataset are driven predominantly by one underlying 'customer value' axis rather than several independent dimensions.

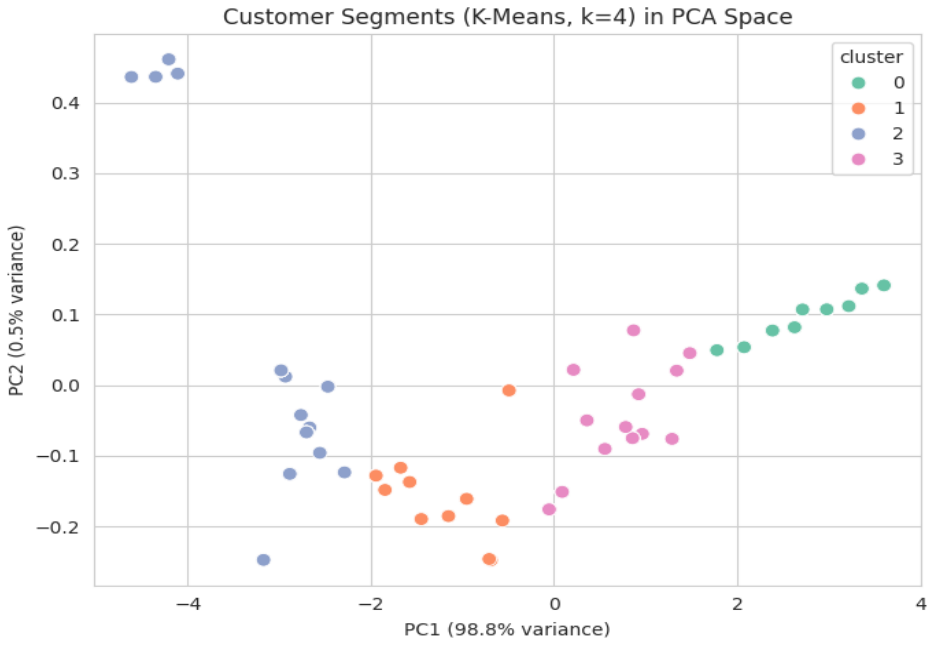

Figure 6. Customer segments visualised in PCA space (k=4).

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722663

Segment	n	Age	Income (₹)	Spend (₹)	Loyalty	Frequency
High-Value Loyal	60	51.1	71,067	601	9.05	25.5
Established	77	41.2	61,208	469	7.53	21.2
Growing / Potential	50	32.6	52,100	358	5.85	17.7
New / Low-Engagement	51	26.3	40,804	219	3.95	13.1

Table IV. Segment profiles (mean values per segment).

All four segments follow a clear, monotonic order across every variable, a direct consequence of the single-dominant-axis structure PCA revealed: High-Value Loyal customers rank highest on age, income, spend, and loyalty, while New/Low-Engagement customers rank lowest on all four. This ordering is genuinely useful for marketing purposes, but should be interpreted alongside the multicollinearity finding — these segments largely split customers along one axis rather than capturing independent behavioural dimensions.

VI. ENGAGEMENT PREDICTION

A. Predictor Selection: Avoiding Near-Tautological Prediction

Careful consideration was given to which features would serve as predictors of engagement_tier. Because purchase_amount and purchase_frequency correlate at $r \geq 0.98$ with loyalty_score — the variable from which engagement_tier is derived — including them as predictors would render the prediction task nearly circular, with the model effectively reading the target back off an almost-identical variable rather than genuinely forecasting an unknown outcome from independent inputs. The predictor set was therefore narrowed to age, annual_income, and region (one-hot encoded), variables conceptually independent of how engagement_tier was constructed, though, as Section VII shows, they remain empirically correlated with it because of the dataset's overall collinearity.

B. Models Compared

Four classifiers were compared: Logistic Regression, a simple and interpretable linear baseline; Decision Tree, interpretable and capable of non-linear splits; Random Forest, an ensemble of trees generally resistant to overfitting; and Gradient Boosting, a sequential ensemble typically the strongest tabular-data performer of the four. This selection follows the standard approach used throughout the segmentation-prediction literature reviewed in Section II.

C. Training and Validation Protocol

Stratified sampling split the data 80/20 into training and test sets, preserving engagement-tier proportions across both. Five-fold stratified cross-validation was applied to every model. Numeric predictors were standardised to zero mean and unit variance prior to training, and macro-averaged F1-score — which weighs all three engagement tiers equally regardless of class size — served as the primary basis for model selection.

VII. MODEL EVALUATION

Table V presents the full comparison of all four models on the held-out test set, together with 5-fold cross-validation statistics.

Model	Test Acc.	Precision	Recall	F1 (macro)	ROC-AUC	CV Mean ± SD
Logistic Regression	96.7%	0.967	0.965	0.964	0.994	95.4% ± 2.5%
Decision Tree	100%	1.000	1.000	1.000	1.000	100% ± 0.0%
Random Forest	98.3%	0.983	0.983	0.982	1.000	99.2% ± 1.0%

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722663

Model	Test Acc.	Precision	Recall	F1 (macro)	ROC-AUC	CV Mean ± SD
Gradient Boosting	100%	1.000	1.000	1.000	1.000	100% ± 0.0%

Table V. Model comparison on held-out test set and 5-fold stratified cross-validation.

Decision Tree and Gradient Boosting both achieved perfect scores across every metric on the held-out test set, with Random Forest and Logistic Regression trailing only slightly. Decision Tree was selected as the representative best model on the basis of macro F1-score, a tie with Gradient Boosting broken in Decision Tree's favour for being easier to interpret. Its confusion matrix (Figure 7) records zero misclassifications across all three engagement tiers on the test set.

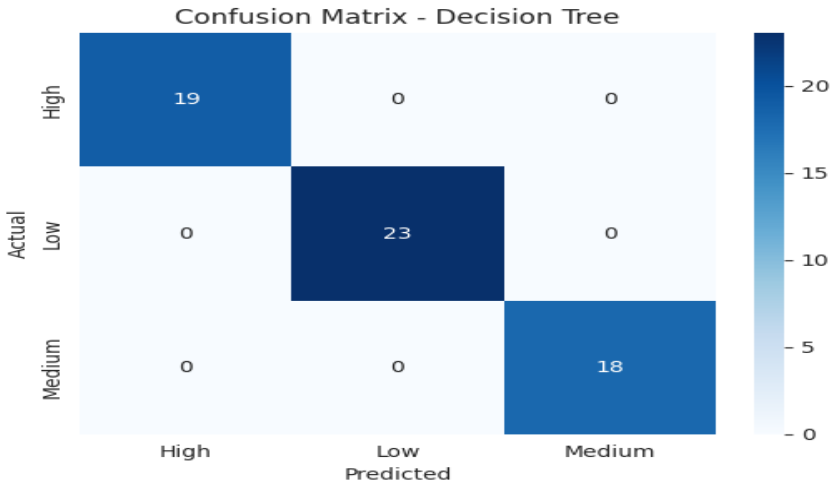

Figure 7. Confusion matrix, Decision Tree model, held-out test set (n=60).

A. Why Is Accuracy Near 100%? An Honest Interpretation

These figures should not be mistaken for evidence of an unusually powerful model. Section IV-D already established that age and annual_income, the two continuous predictors used here, are themselves strongly collinear with loyalty_score and, by extension, engagement_tier, with VIF values of 78 and 62 respectively. What appears to be near-perfect classification accuracy in fact reflects the dataset's structure rather than a genuinely difficult prediction problem being solved. Accuracy remained at or near 100% even with only demographic variables in play, after the most obviously circular predictors (purchase_amount and purchase_frequency; Section VI-A) were deliberately excluded — a result that reinforces this interpretation rather than contradicting it.

This contrasts with studies operating on genuinely large, real-world transactional data: a 2025 study combining RFM features, K-Means, and deep neural networks to predict churn on the Olist and Instacart datasets achieved above 99% accuracy on datasets several orders of magnitude larger, where such performance cannot be explained away by a single dominant latent variable. Section XII (Limitations) revisits this contrast.

VIII. EXPLAINABLE AI

The selected Decision Tree model was analysed using SHAP (SHapley Additive exPlanations), tracing individual predictions back to specific feature contributions and directly addressing Research Objective 4 and Research Question 3.

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722663

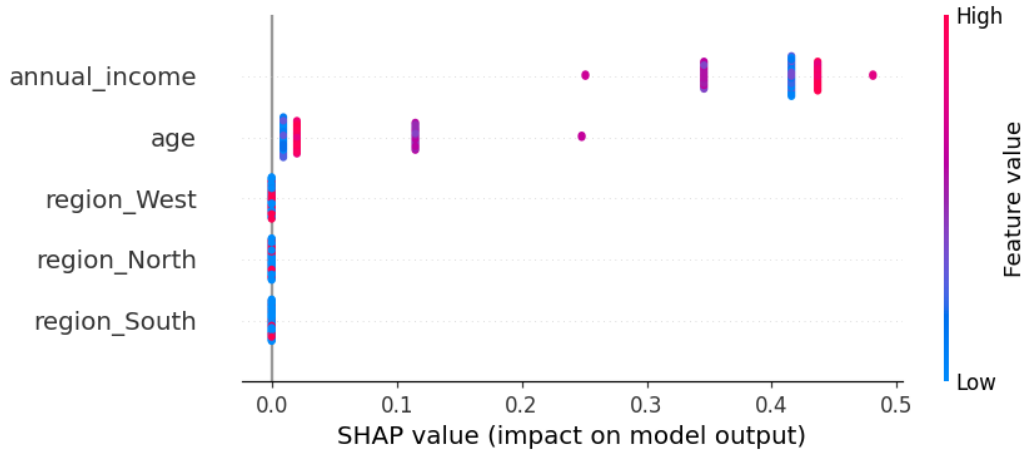

Figure 8. SHAP summary plot for the engagement-tier prediction model.

A. Technical Interpretation

Annual_income's mean absolute SHAP value is roughly an order of magnitude above that of age, while the one-hot region features (region_North, region_South, region_West) barely register at all. This matches the tree-based feature importance shown in Figure 9, and the VIF-confirmed collinearity between annual_income and the loyalty-derived target noted in Section IV-D.

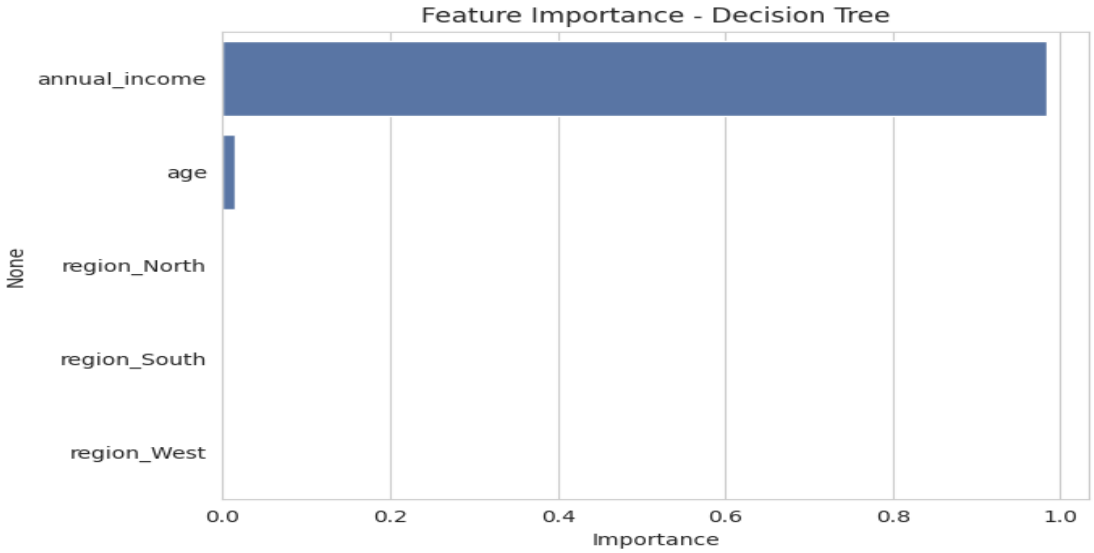

Figure 9. Tree-based feature importance, Decision Tree model.

B. Business Interpretation

In practical terms, predicted engagement rises consistently with income, rises more modestly with age, and barely moves with region. For a marketing team, this means income-tier information alone accounts for most of what this model can convey about likely engagement — a conclusion tied to this dataset's particular structure rather than a general marketing rule, as discussed further in Section XII.

IX. PERSONALISED MARKETING FRAMEWORK

A rule-based recommendation engine draws on both segment membership (Section V) and predicted engagement tier (Sections VI–VII), mapping each of the four empirically derived segments onto a concrete marketing strategy grounded in that segment's actual

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722663

behavioural profile (Table IV) rather than a one-size-fits-all template.

Segment	Strategy	Example Actions
High-Value Loyal	Loyalty and VIP treatment	First access to new product launches; tiered loyalty programme; priority customer support
Established	Retention and cross-sell	Recommend complementary products; reward account milestones; periodic check-ins
Growing / Potential	Personalised onboarding push	Targeted discount on next order; onboarding content; referral incentives
New / Low-Engagement	Win-back campaign	Win-back discount; low-friction product suggestions; closer monitoring going forward

Table VI. Segment-to-strategy mapping.

Where a customer's behavioural segment and independently predicted engagement tier disagree — for instance, a customer in a lower-value segment whose demographics alone point to high engagement — the engine flags the mismatch for a human marketer to review rather than resolving it automatically. This logic is implemented on the dashboard application's Personalised Marketing page, operating against live data rather than remaining a static example.

X. RESULTS AND DISCUSSION

A. Summary of Results

- Across all 238 records, preprocessing found 0 missing values, 0 duplicates, 0 range violations, and 0 IQR-flagged outliers.
- Multicollinearity was severe: pairwise correlation across all five numeric variables reached $r \geq 0.97$, and VIF ranged from 37 to 197.
- K-Means segmentation at k=4 (Silhouette = 0.601, against a statistical optimum of k=6 at 0.633) produced four segments of between 50 and 77 customers each, with PC1 alone accounting for 98.8% of variance.
- Trained on age, annual_income, and region, the four classifiers reached test accuracy from 96.7% (Logistic Regression) to 100% (Decision Tree and Gradient Boosting).
- SHAP places annual_income well ahead as the dominant predictor — roughly ten times the impact of age — with region contributing almost nothing.
- Statistical evidence (VIF, PCA) attributes the near-100% predictive accuracy to dataset-level multicollinearity rather than to genuine predictive achievement.

B. Discussion

The segmentation results (Section V) align well with the broader RFM-K-Means literature (Section II-B): Elbow/Silhouette-validated K-Means consistently produces interpretable and usable segments, and the finding here — that a moderate cluster count (k=4) can outperform the statistical optimum (k=6) on interpretability — directly echoes Lewaaelhamd (2024).

The prediction results tell a different story from the large-dataset XAI literature (Section II-C). El Attar and El-Hajj (2026), together with the RFM+K-Means+DNN churn study (Section VII-A), achieve high accuracy on datasets running into the thousands or millions of records; this paper's near-identical accuracy figures arise from an entirely different mechanism — extreme feature collinearity within a small (n=238) dataset, rather than a genuinely difficult prediction problem solved at scale. Rather than contradicting the earlier work, this represents a boundary condition on it: identical techniques can produce similar-looking performance numbers for very different underlying reasons depending on a dataset's scale and structure, and only diagnostics

such as VIF and PCA reveal which explanation actually applies.

The SHAP findings (Section VIII) are consistent with feature-importance patterns reported elsewhere in customer analytics, where income-related variables tend to dominate engagement and value predictions across the literature, though the near-total irrelevance of region observed here reflects this specific dataset more than customer behaviour generally.

C. Business Implications

Given data structured like this — small sample, limited feature diversity — the primary business value of this framework lies in the segmentation output (Section V) and the resulting strategy mapping (Section IX), neither of which rests on the multicollinearity-inflated predictive claims. The engagement-prediction component's business value should be treated cautiously until validated on a larger, less collinear dataset (Section XII).

XI. RESEARCH CONTRIBUTION

Academic contribution: Using VIF and PCA diagnostics, this paper identifies the specific statistical conditions under which apparently near-perfect model accuracy is a data artefact rather than genuine predictive skill — a caution that large, industrial-scale studies rarely need to make explicit, but one that matters substantially for student and small-organisation research.

Practical contribution: A four-segment customer typology (Table IV), supported by SHAP-based driver analysis (Section VIII), mapped onto actionable marketing strategies (Table VI) and implemented as a working dashboard application.

Methodological contribution: A transparent decision process addressing two choices that recur throughout applied customer analytics: selecting a business-interpretable cluster count over the statistically optimal one, and excluding collinear predictors to avoid near-tautological prediction — both justified explicitly rather than resolved without comment.

What is not claimed: This paper makes no claim to state-of-the-art predictive accuracy — the near-100% figures reflect dataset structure rather than model superiority. Nor does it claim to be first to combine segmentation, prediction, and explainable AI; El Attar and El-Hajj (2026) already did so at larger scale. The novelty claimed here remains deliberately narrow, confined to the small-data, transparency-first contribution described above.

XII. LIMITATIONS AND FUTURE WORK

A. Limitations

- With n=238, the small sample size limits statistical power and generalisability, particularly for the East region subgroup (n=6).
- Severe multicollinearity (VIF 37–197) inflates apparent predictive performance and makes genuinely independent predictive relationships difficult to isolate.
- The dataset's structure — near-perfect pairwise correlations and PC1 explaining 98.8% of variance — is more consistent with synthetic or heavily simplified data generation than with organically observed customer behaviour, though this could not be verified directly since the source does not document the dataset's original generation process.
- No transaction timestamps exist, which precludes constructing a true Recency feature (the 'R' in RFM analysis).
- The personalised marketing recommendation engine relies on rules rather than a learned recommender system, a direct consequence of the dataset containing no item-level product-interaction data.
- Feature-importance findings reported here, such as income dominance and the irrelevance of region, are tied to this dataset's particular structure and should not be generalised to customer populations broadly without independent validation.

B. Future Work

- Apply the same framework to a larger, real transactional dataset that includes timestamps, enabling genuine RFM features and testing whether the segmentation structure and predictor-

importance rankings hold outside conditions of extreme collinearity.

- If product-interaction data becomes available, extend the personalised marketing framework into an item-level recommender system built on collaborative or content-based filtering.
- Add individual-level SHAP waterfall explanations to the application's Explainable AI page, beyond the current global feature-importance view.
- Explore real-time or streaming customer-behaviour analysis and A/B-testing infrastructure to determine whether the recommended marketing strategies (Section IX) produce measurable engagement improvements in practice.

XIII. CONCLUSION

This paper set out to build, and honestly evaluate, a complete pipeline for AI-driven customer segmentation and engagement prediction, from raw data through to an interpretable, working dashboard application. K-Means segmentation produced four business-interpretable customer segments; engagement-tier prediction achieved high accuracy that was traced, through VIF and PCA diagnostics, back to dataset-level multicollinearity rather than presented at face value as a modelling achievement; and SHAP-based explainability identified annual income as the dominant driver of predicted engagement. Rather than overstating what is novel here, the paper deliberately narrows its contribution to something more defensible: demonstrating how a segmentation-prediction-explainability framework already validated on large industrial datasets behaves, and should be evaluated, when applied to the small, highly collinear data typical of student and small-organisation research. The resulting framework, methodology, and candidly reported limitations serve as a transferable template for similarly resource-constrained customer-analytics projects.

REFERENCES

[1] Brodie, R. J., Hollebeek, L. D., Jurić, B., & Ilić, A. (2011). Customer engagement: Conceptual domain, fundamental propositions, and implications for research. Journal of Service Research, 14(3), 252-271. https://doi.org/10.1177/1094670511411703

[2] El Attar, A., & El-Hajj, M. (2026). Explainable AI-driven customer churn prediction: A multi-model ensemble approach with SHAP-based feature analysis. Frontiers in Artificial Intelligence. https://doi.org/10.3389/frai.2026.1748799

[3] Hollebeek, L. D., Glynn, M. S., & Brodie, R. J. (2014). Consumer brand engagement in social media: Conceptualization, scale development and validation. Journal of Interactive Marketing, 28(2), 149-165. https://doi.org/10.1016/j.intmar.2013.12.002

[4] Hollebeek, L. D., et al. (2023). Hallmarks and potential pitfalls of customer- and consumer engagement scales: A systematic review. Psychology & Marketing. https://doi.org/10.1002/mar.21797

[5] Lalitha, S., Reddy, M. U., Hussain, G. N., & Reddy, N. V. L. (2025). RFM analysis using K-means algorithm for customer segmentation. In Lecture Notes in Electrical Engineering (Springer). https://doi.org/10.1007/978-981-97-4711-5_12

[6] Lewaaelhamd, I. (2024). Customer segmentation using machine learning model: An application of RFM analysis. Journal of Data Science and Intelligent Systems. https://doi.org/10.47852/bonviewJDSIS32021293

[7] Martin, K. D., Borah, A., & Palmatier, R. W. (2017). Data privacy: Effects on customer and firm performance. Journal of Marketing, 81(1), 36-58. https://doi.org/10.1509/jm.15.0497

[8] The Implementation of RFM Analysis to Customer Profiling Using K-Means Clustering. (2023-24). Mathematical Modelling of Engineering Problems (IIETA). https://doi.org/10.18280/mmep.100135

[9] A Systematic Literature Review on Explainable AI for Churn Prediction, Customer Segmentation and Retention. (2025-

26). International Journal of Data Science and Analytics (Springer). https://doi.org/10.1007/s41060-025-01018-0

[10] A Data-Driven Approach with Explainable AI for Customer Churn Prediction in Telecommunications. (2025). ScienceDirect (Elsevier).

[11] A Systematic Review of Artificial Intelligence-Based Customer Analytics in Personalized Digital Marketing (2019-2026). (2026). American Journal of Data Science and Analytics.

[12] The Role of Artificial Intelligence in Customer Engagement and Social Media Marketing — Implications for Tourism and Hospitality. (2025). Journal of Theoretical and Applied Electronic Commerce Research (MDPI).

[13] A Systematic Review on AI in Marketing: Personalization of Advertising Campaigns and Impact on Customer Satisfaction. (2025). Springer conference volume. https://doi.org/10.1007/978-981-96-3077-6_11

[14] AI-Driven Personalization and Customer Engagement: A PRISMA-Guided Systematic Literature and Bibliometric Review. (2026). Frontiers in Artificial Intelligence.

[15] Data Security and Privacy Concerns of AI-Driven Marketing. (2024). Cogent Business & Management (Taylor & Francis). https://doi.org/10.1080/23311975.2024.2393743

[16] Product Collaborative Filtering Based Recommendation Systems for Large-Scale E-commerce. (2025). ScienceDirect (Elsevier, open access).

[17] Artificial Intelligence and Recommender Systems in E-commerce: Trends and Research Agenda. (2024). ScienceDirect (Elsevier, open access).

[18] A Personalized Product Recommendation Model in E-commerce Based on Retrieval Strategy. (2024). ScienceDirect (Elsevier, open access).

[19] Recommendation System for E-Commerce Using Collaborative Filtering. (2024). Journal Européen des Systèmes Automatisés (IIETA). https://doi.org/10.18280/jesa.570421

[20] Enhancing Customer Retention in Online Retail Through Churn Prediction: A Hybrid RFM, K-means, and Deep Neural Network Approach. (2025). ScienceDirect (Elsevier, open access).