IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

AI BASED BENEFICIARY SEGMENTATION AND PERSONALIZED RECOMMENDATION

Hershly Bose J[1], Deena Darshan M[2], Lijo Y[3], S. Poornima[4]

Dept. of CSE (Emerging Technologies), SRM Institute of Science and Technology, Vadapalani Campus Chennai, India[1-4]

Abstract: Small retailers depend on customers who come back, yet most treat every customer the same way. This paper presents CustomerSegment ML, a full-stack system that segments the customers of a small shop automatically. The case study is a Tamil Nadu gift shop selling key chains, bouquets and miniature toys. A FastAPI service validates an uploaded customer file, standardises two behavioural features, age and number of repeat purchases, sweeps the number of clusters K from 2 to 8 and selects the K with the highest silhouette score. K-Means with k-means++ seeding and ten restarts is fitted, and each cluster is described by a persona, its top product, its top district and a marketing strategy; new customers are classified in real time through a Next.js dashboard. Three synthetic datasets of 70, 85 and 100 customers from 20 districts were generated with hidden ground-truth groups. The system recovers the correct number of groups on all three, with 100%, 98.8% and 99% accuracy. On the 100-customer file it reaches a silhouette of 0.6654 and ranks first among fourteen methods on accuracy, silhouette and the Calinski-Harabasz index, ahead of Gaussian mixtures, hierarchical clustering, BIRCH and DBSCAN, and far ahead of grouping by product type (33%) and random assignment (30%).

Keywords: customer segmentation, K-Means clustering, repeat purchases, unsupervised learning, silhouette analysis, retail analytics, FastAPI, Next.js

I. INTRODUCTION

Small shops in Tamil Nadu, from gift stores in Chennai to stalls near temples in Madurai, live on returning customers. A customer who comes back fifteen times is worth far more than one who buys once, but shop owners usually rely on instinct or on a single rule such as “students” and “elders”. Market segmentation, formalised by Smith [1], divides customers into homogeneous groups that can each receive a suitable strategy, and clustering discovers such groups directly from data [2].

K-Means [3], [4] is the most widely used clustering algorithm because it is fast and its centroids read as the profile of a typical member. Three problems limit its use by non-experts: K must be chosen in advance, the result depends on the starting centroids, and the output is a set of numbers rather than an action. Repeat purchases, the frequency part of RFM analysis [5], measure loyalty directly, and combined with age they separate loyal fans from occasional visitors in each age group.

This paper presents CustomerSegment ML, a system that addresses these problems end to end. Its contributions are:

- a tolerant upload pipeline that detects columns, cleans invalid rows and standardises age and repeat purchases;
- automatic selection of K by silhouette analysis, supported by the elbow, Davies-Bouldin and Calinski-Harabasz criteria;
- a stable K-Means configuration with k-means++ seeding and ten restarts, validated over 30 random seeds;
- personas with top product, top district and strategy, real-time prediction, and an optional accuracy card when true groups are known;
- three Tamil Nadu gift-shop datasets and a comparison against thirteen alternative and baseline methods.

Section II reviews related work, Section III describes the data, Section IV the method, Section V the implementation, Section VI the results and Section VII concludes.

II. RELATED WORK

A. Clustering Algorithms

Lloyd's algorithm [4] alternates between assigning points to their nearest centroid and moving each centroid to the mean of its points. k-means++ seeding [6] chooses spread-out initial centroids and is O(log K)-competitive with the optimum. Agglomerative clustering with Ward linkage [7] merges the pair of clusters with the smallest variance increase, while single, average and complete linkage use other inter-cluster distances. DBSCAN [8] finds dense regions

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

of arbitrary shape, Gaussian mixture models fitted by EM [9] give soft memberships, and BIRCH [10] summarises very large data in a clustering-feature tree.

B. Choosing K and Validating Clusters

Without labels, partitions are judged by internal indices. The elbow method [11] looks for the K after which WCSS stops falling sharply. The silhouette coefficient [12] compares each point's distance to its own cluster with its distance to the nearest other cluster, and the Davies-Bouldin [13] and Calinski-Harabasz [14] indices measure the ratio of within-cluster scatter to between-cluster separation. When true groups are known, accuracy after Hungarian matching [15] and the adjusted Rand index [16] measure external agreement.

C. Segmentation of Retail Customers

Demographic and behavioural clustering is a standard segmentation method [2], and RFM scoring [5] is a common rule-based alternative. Most studies apply clustering in a notebook to large supermarket or online datasets; our work packages it as a validated tool for a small shop with 70 to 100 customers, with automatic model selection, explanations and real-time prediction.

III. DATASET AND PROBLEM FORMULATION

Real customer records of small shops are private, so we generated synthetic but realistic datasets. Each customer is drawn from one of five behavioural groups with a typical age, number of repeat purchases and product mix; names are common Tamil names, districts come from 20 Tamil Nadu districts and products are Key Chains, Bouquet and Miniature Toys. The hidden group is not written to the files and is used only for evaluation. Table I summarises the main 100-customer file; the other files have 70 customers (3 groups) and 85 customers (4 groups).

TABLE I: Summary of the 100-Customer Dataset

Attribute	Type	Range / values	Statistics	Role in model
Name	Text	Tamil first name and initial	93 unique names	Display only
Age	Integer	16 to 62 years	Mean 38.2, SD 15.1, median 36.5	Clustering feature
Gender	Categorical	Female, Male	F 51, M 49	Female share per cluster
District	Categorical	20 Tamil Nadu districts	Top: Tiruchirappalli 10, Chennai 9	Top district per cluster
Product Type	Categorical	Key Chains, Bouquet, Miniature Toys	Bouquet 42, Miniature Toys 34, Key Chains 24	Top product per cluster
Repeat Purchases	Integer	1 to 19 purchases	Mean 9.0, SD 5.6, median 9	Clustering feature
All records	-	100 customers, 6 columns	0 missing values	5 hidden groups (evaluation only)

Age and repeat purchases are almost uncorrelated ($r = -0.042$), so they carry independent information about who the customer is and how loyal they are. Product, district and gender are kept as descriptive attributes. Given n customers $x_i = [age_i, repeats_i]$, the task is to find a partition into K clusters with high cohesion and separation, choose K automatically and assign new customers to the correct cluster.

IV. PROPOSED METHOD

A. Preprocessing

Headers are lower-cased and stripped of non-alphanumeric characters so that variants such as “Repeat_Purchases” and “Repeats” match. Values are coerced to numbers, and rows with missing, negative or non-numeric features are dropped and reported. Each feature is standardised with its mean μ and standard deviation σ:

$$z = (x - \mu) / \sigma \qquad (1)$$

B. K-Means With k-means++ Seeding

Each customer is assigned to its nearest centroid and each centroid moves to the mean of its members:

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

$$c_i = \arg\min_j \| x_i - \mu_j \|^2 \qquad (2)$$

$$\mu_j = (1 / |C_j|) \cdot \Sigma\, x_i\,, \;\; x_i \in C_j \qquad (3)$$

Both steps decrease the within-cluster sum of squares, which guarantees convergence:

$$J = \Sigma_{j=1}^{K} \Sigma_{i \in C_j} \| x_i - \mu_j \|^2 \qquad (4)$$

Initial centroids are drawn with k-means++, choosing a point with probability proportional to its squared distance $D(x)^2$ from the nearest seed:

$$P(x) = D(x)^2 / \Sigma_{x'} D(x')^2 \qquad (5)$$

The procedure is repeated ten times and the run with the lowest J is kept (random state 42).

C. Automatic Selection of K and Evaluation

For K = 2 to 8 the system computes the mean silhouette S(K), where a(i) and b(i) are the mean distances of customer i to its own and to the nearest other cluster, and selects the maximum:

$$s(i) = (b(i) - a(i)) / \max\{ a(i), b(i) \} \qquad (6)$$

$$K^* = \arg\max S(K), \;\; K \in \{2, \ldots, 8\} \qquad (7)$$

Davies-Bouldin and Calinski-Harabasz indices are reported as supporting evidence. When true groups are available, accuracy is computed after the best one-to-one matching π of clusters to groups on the confusion matrix M [15]:

$$\text{Accuracy} = \max_{\pi} (1 / n) \Sigma_j M[j, \pi(j)] \qquad (8)$$

D. Personas and Prediction

Each cluster's mean age is banded into young, adult or senior, and its mean repeat purchases into occasional, regular or loyal, using half a standard deviation around the overall mean. The nine combinations give names such as Loyal Young Fans or Occasional Senior Buyers with matching strategies. The top product, top district and female share are reported per cluster. A new customer is scaled and assigned to the nearest centroid.

V. SYSTEM IMPLEMENTATION

Fig. 1 shows the architecture. A Next.js 16 dashboard in TypeScript with Tailwind CSS and Recharts calls a FastAPI service through a typed client. Pydantic schemas validate every request, and every model endpoint returns HTTP 409 until a dataset has been uploaded. The engine, built on scikit-learn [17], fits the scaler and model into local objects and publishes them together, so a concurrent prediction never sees a half-trained model.

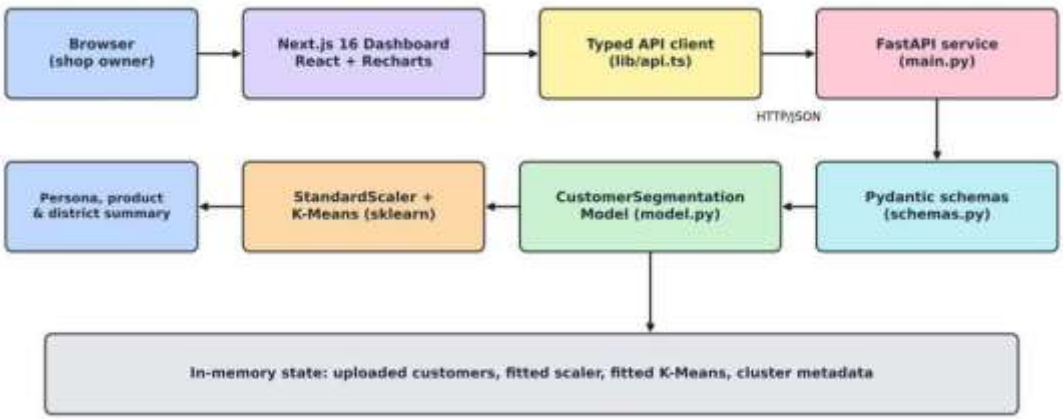


Fig. 1. Architecture of the CustomerSegment ML system.

Six endpoints cover health, upload, dataset, elbow, cluster and predict. The dashboard shows processing stages, a searchable customer table, the elbow chart, a K slider, an age against repeat-purchases scatter plot with centroids,

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

persona cards with top product and district, an accuracy card when true groups are present, a prediction form and a CSV export. Automated tests cover every endpoint and all three datasets.

VI. RESULTS AND DISCUSSION

A. Selection of K

Table II reports the indices for K = 2 to 8 on the 100-customer file. WCSS drops from 117.3 to 13.4 up to K = 5 and then flattens (Fig. 2); silhouette peaks at 0.6654, Davies-Bouldin reaches its minimum of 0.442 and Calinski-Harabasz its maximum of 331.4, all at K = 5, the true number of groups.

TABLE II: Validity Indices and Accuracy for K = 2 to 8

K	WCSS	Silh.	DB	CH	Acc.
2	117.3	0.407	1.146	69.0	39%
3	68.4	0.502	0.649	93.2	58%
4	30.6	0.612	0.548	177.2	78%
5	13.4	0.665	0.442	331.4	99%
6	11.5	0.589	0.611	307.2	92%
7	9.9	0.569	0.656	297.6	81%
8	8.6	0.503	0.764	293.1	72%

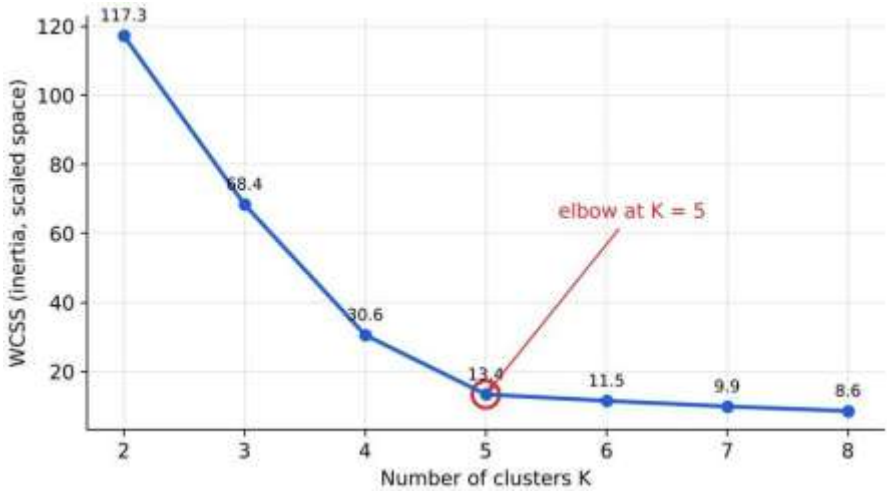


Fig. 2. Elbow curve of WCSS against K.

B. Results on All Datasets

TABLE III: Results on the Three Datasets

Dataset	n	True K	Rec. K	Silh.	Acc.
70 cust.	70	3	3	0.698	100.0%
85 cust.	85	4	4	0.705	98.8%
100 cust.	100	5	5	0.665	99.0%

The recommended K equals the true number of groups on every file and accuracy is 98.8% to 100%. Fig. 3 shows that accuracy peaks exactly at the recommended K and falls on both sides.

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

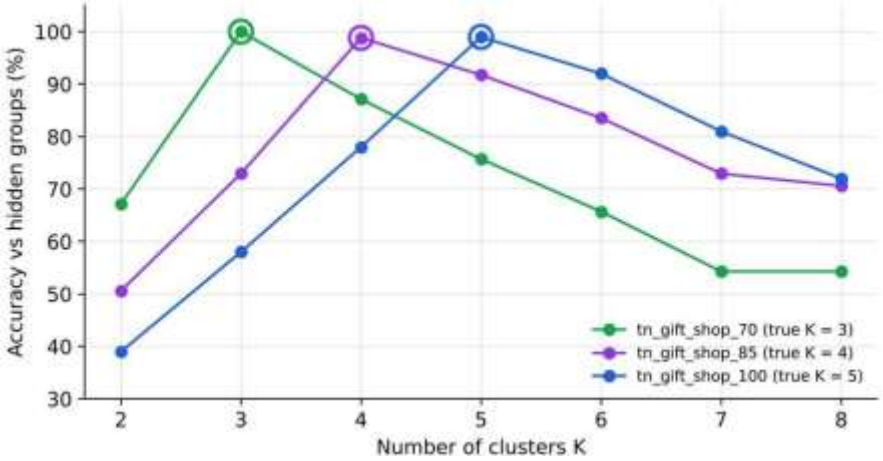


Fig. 3. Accuracy against the hidden groups for each K on the three datasets.

C. Discovered Segments

Fig. 4 and Table IV show the five segments. K-Means converged in 4 iterations and trained in about 17 ms, so the dashboard retrains interactively.

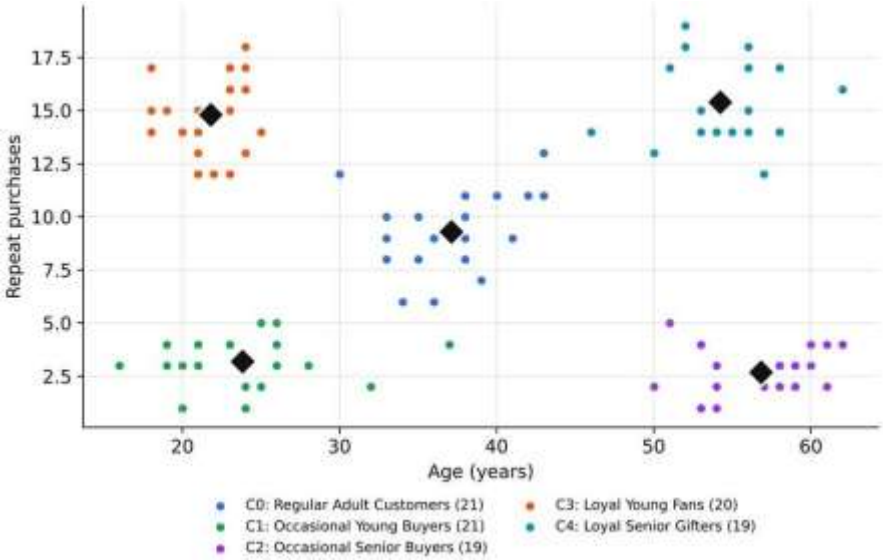


Fig. 4. Final segmentation with K = 5; diamonds mark centroids.

TABLE IV: Profiles of the Five Segments

Persona	n	Age	Rep.	Top product
Regular Adult Customers	21	37.1	9.3	Miniature Toys
Occasional Young Buyers	21	23.8	3.2	Key Chains
Occasional Senior Buyers	19	56.8	2.7	Bouquet
Loyal Young Fans	20	21.8	14.8	Key Chains
Loyal Senior Gifters	19	54.2	15.4	Bouquet

Young groups buy mostly Key Chains and Miniature Toys while senior groups buy mostly Bouquets, even though product type was never used for clustering (Fig. 5). Loyal groups should be rewarded with loyalty cards and early access to new designs, regular adults converted with festival combos, and occasional buyers won back with occasion reminders: social-media offers for the young and phone, SMS and festival offers for seniors. Five hand-made test customers placed in each region were all assigned to the expected segment.

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

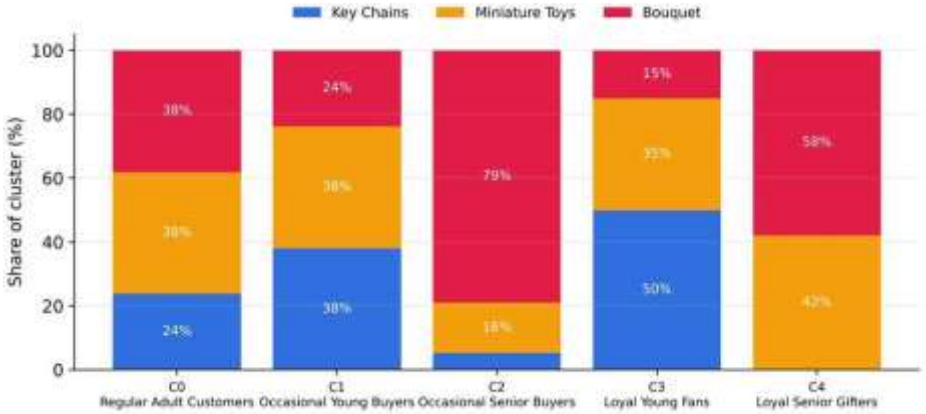


Fig. 5. Product mix inside each segment.

D. Comparison With Alternative and Baseline Methods

All methods were run on the same scaled 100-customer data with K = 5 (Table V, Figs. 6 and 7).

TABLE V: Comparison With Alternative and Baseline Methods

Method	Silh. ↑	CH ↑	Acc. ↑
CustomerSegment ML	0.6654	331.4	99%
Gaussian mixture	0.6647	330.2	98%
K-Means, random, 1 run	0.6580	320.8	97%
Agglom. (Ward)	0.6498	295.4	97%
Agglom. (average)	0.6479	310.9	96%
Agglom. (complete)	0.6456	309.1	95%
BIRCH	0.6191	246.9	93%
DBSCAN	0.5212	78.6	81%
Agglom. (single)	0.3877	99.1	80%
Age bands (rule)	0.2510	44.1	60%
Age-only K-Means	0.1552	23.8	64%
Repeats-only K-Means	0.1379	24.2	62%
Group by product	-0.0195	7.7	33%
Random assignment	-0.0970	0.9	30%

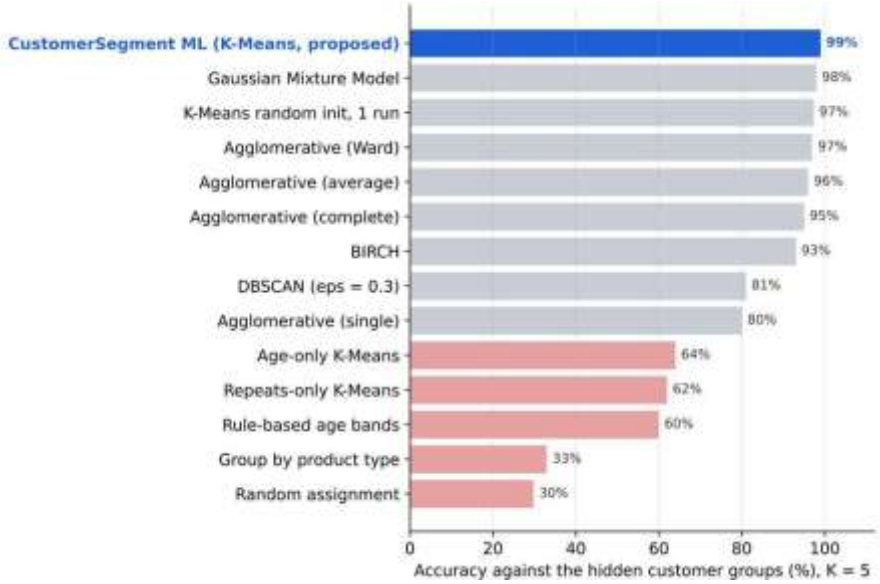


Fig. 6. Accuracy of CustomerSegment ML versus alternative and baseline models.

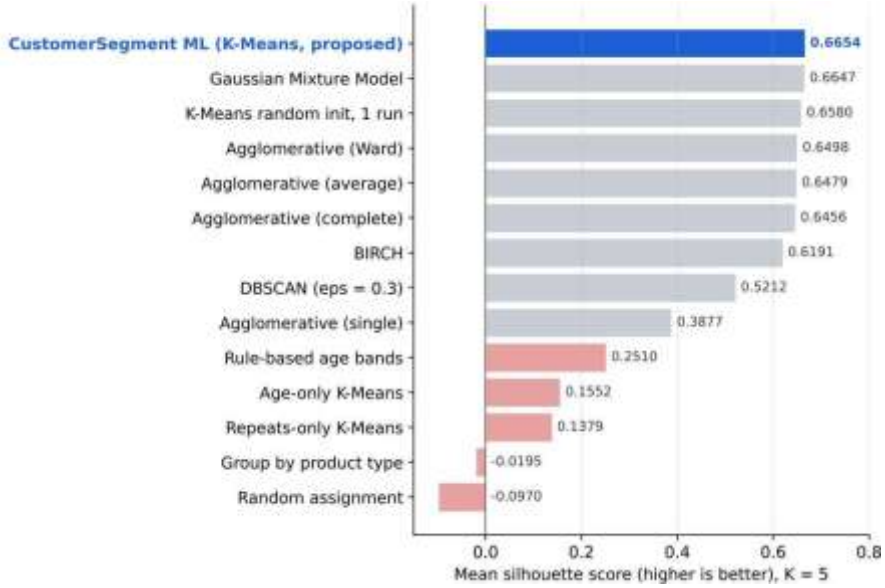


Fig. 7. Silhouette of CustomerSegment ML versus alternative and baseline models.

CustomerSegment ML ranks first on accuracy (99%), silhouette (0.6654) and Calinski-Harabasz (331.4). The Gaussian mixture comes closest (98%, 0.6647) but has more parameters and no simple centroid explanation; hierarchical methods reach 95% to 97%, BIRCH 93%, and DBSCAN and single linkage about 80%. Using only age (64%) or only repeats (62%) misses half of the structure, fixed age bands reach 60%, grouping by product only 33% and random assignment 30%. On Davies-Bouldin, the Gaussian mixture (0.435) and complete linkage (0.425) are marginally lower than ours (0.442) but less accurate.

E. Stability

Over 30 random seeds the proposed configuration returned the same partition every time (adjusted Rand index 1.000 between runs), while K-Means with a single random start averaged 95.0% accuracy and dropped to 0.390 silhouette and 71% accuracy on an unlucky seed (Fig. 8).

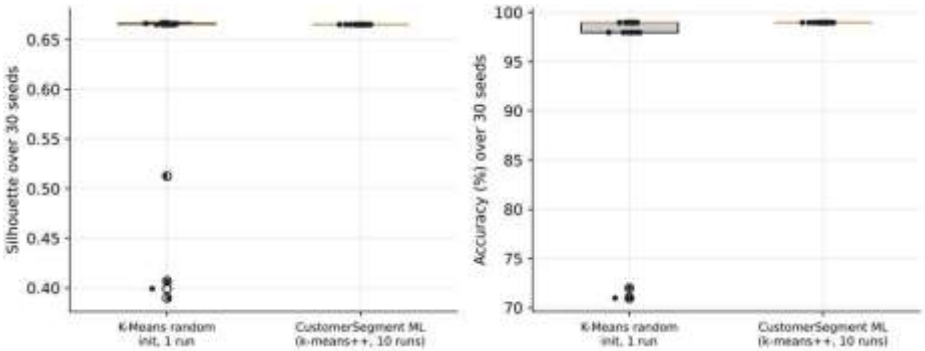


Fig. 8. Silhouette and accuracy over 30 random seeds.

F. Limitations

The datasets are synthetic and generated from well-separated groups, so real shop data will be noisier. The current version clusters two behavioural features, assumes compact clusters and keeps data in server memory. Internal indices and accuracy measure grouping quality, not campaign uplift, which requires A/B testing with real customers.

VII. CONCLUSION AND FUTURE WORK

CustomerSegment ML turns a small shop's customer file into clear, actionable segments with no manual tuning. Scaling, silhouette-based selection of K and k-means++ with restarts recover the hidden groups of three Tamil Nadu gift-shop datasets with 98.8% to 100% accuracy, rank first among fourteen clustering approaches and give perfectly reproducible results, while personas, product and district insights, real-time prediction and an interactive dashboard make the result usable by shop owners. Future work will validate the system on real anonymised shop data, add

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

spending and recency (full RFM) features, let users compare algorithms on their own data, persist per-shop models and connect segments to WhatsApp or SMS campaigns to measure how many customers return.

ACKNOWLEDGMENT

The authors thank Dr. P. Chitra, Professor and Head, Department of Computer Science and Engineering (Emerging Technologies), SRM Institute of Science and Technology, Vadapalani Campus, for her guidance and support.

REFERENCES

[1]. W. R. Smith, "Product differentiation and market segmentation as alternative marketing strategies," J. Marketing, vol. 21, no. 1, pp. 3–8, 1956.
[2]. M. Wedel and W. A. Kamakura, Market Segmentation: Conceptual and Methodological Foundations, 2nd ed. Boston, MA, USA: Kluwer, 2000.
[3]. J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Math. Statist. Probab., vol. 1, 1967, pp. 281–297.
[4]. S. P. Lloyd, "Least squares quantization in PCM," IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129–137, 1982.
[5]. A. M. Hughes, Strategic Database Marketing. Chicago, IL, USA: Probus, 1994.
[6]. D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in Proc. 18th ACM-SIAM Symp. Discrete Algorithms (SODA), 2007, pp. 1027–1035.
[7]. J. H. Ward Jr., "Hierarchical grouping to optimize an objective function," J. Amer. Statist. Assoc., vol. 58, no. 301, pp. 236–244, 1963.
[8]. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. KDD, 1996, pp. 226–231.
[9]. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. R. Statist. Soc. B, vol. 39, no. 1, pp. 1–38, 1977.
[10]. T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD, 1996, pp. 103–114.
[11]. R. L. Thorndike, "Who belongs in the family?," Psychometrika, vol. 18, no. 4, pp. 267–276, 1953.
[12]. P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, pp. 53–65, 1987.
[13]. D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224–227, 1979.
[14]. T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," Commun. Statist., vol. 3, no. 1, pp. 1–27, 1974.
[15]. H. W. Kuhn, "The Hungarian method for the assignment problem," Naval Res. Logist. Quart., vol. 2, no. 1–2, pp. 83–97, 1955.
[16]. L. Hubert and P. Arabie, "Comparing partitions," J. Classification, vol. 2, pp. 193–218, 1985.
[17]. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

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