Glovento Journal of Integrated Studies (GJIS) | ISSN: 3117-3314
Volume 1 (2025) | Article 32

Transforming Raw Insurance Data into Stable Segments and Personalized Upgrade Recommendations

[1]Mustafa Serdar KONCA, [2]Şadi Evren Şeker

[1]Doğuş Technology, R&D Center, İzmir, Türkiye
[2]Department of Computer Engineering, Faculty of Computer and Information Technologies, Istanbul University, Istanbul, Türkiye
Email: [1]serdar.konca@d-teknoloji.com.tr, [2]evrenseker@istanbul.edu.tr

Abstract- Insurers need customer segments that are interpretable, stable across refreshes, and traceable to business rules so they can support targeting, cross-sell, and upsell at scale. We present a production-grade pipeline that converts multi-table operational data (policy, product, request, and entity records) into human-readable segments with individualized upgrade guidance. The workflow combines a robust active-policy definition that reconciles renewals and cancellations via request logs; product-year inflation normalization of premiums; Isolation Forest–based anomaly suppression; standardized scaling; and k-means clustering with centroid-distance reporting. While k-means is our primary method, we also benchmark Gaussian Mixture Models and HDBSCAN; comparative results based on Silhouette and temporal stability (ARI) show that k-means provides more consistent and operationally usable segments. To preserve longitudinal comparability, a stability rule retains prior labels when relative distance changes are small, and agreement is monitored via the Adjusted Rand Index and a transition matrix. Above clusters, an RFM value tier overlays calibrated weights and quantile thresholds, which we invert to produce customer-level prescriptions—additional tenure, distinct active products, or incremental premium—needed to move up one tier. We evaluate the approach on a portfolio of 379,584 customers represented by 38 engineered features, which are disclosed only in anonymized summary form (feature-wise missingness, uniqueness, and distributional statistics). Seven segments emerge that are interpretable and operationally consistent, with high agreement across monthly runs. We summarize segment composition, relative fingerprints (tenure, breadth, normalized premium, RFM), and value shares by tier, while suppressing raw levels to protect proprietary information. The contribution is an auditable, stability-conscious segmentation system that links centroids to business narratives, exposes clear upgrade levers, and yields channel-ready inputs for activation, bundling, and retention without sacrificing analytic rigor.

Keywords: Insurance Customer Segmentation, k-means, RFM Tier, Upgrade Recommendation, Cross-sell.

I. INTRODUCTION

Customer segmentation remains a central mechanism for aligning products, pricing, and communications with heterogeneous customer needs in financial services and insurance. Recent reviews emphasize that segmentation only creates business value when designs are interpretable, stable over time, and embedded in operational decision rules that downstream teams can trust (Yum, Park, Oh, & Lee, 2022; Ikotun et al., 2023). In practice, however, many initiatives stall at proof-of-concept: models are trained on narrow samples, labels drift across monthly refreshes, and the resulting clusters cannot be traced to auditable logic. Consequently, marketing and distribution teams hesitate to institutionalize segments for targeting, cross-sell, and upsell. What is needed is an approach that is both analytically sound and operationally durable—capable of ingesting multi-table, real-world data; producing interpretable segments; and maintaining assignment stability as portfolios evolve.

This paper addresses that need in the context of a large insurance platform. We develop and evaluate an end-to-end workflow that transforms raw policy, product, request/claim, and entity/asset records into human-readable segments with individualized "upgrade" guidance. The workflow combines established components—modern data cleaning and schema-level quality controls (Côté, Sakka, Senellart, & Tannier, 2024), anomaly suppression using Extended Isolation Forest (Hariri, Carrasco-Kind, & Brunner, 2019), standardized scaling, and k-means clustering (Ahmed, Seraj, & Islam, 2020; Ikotun et al., 2023)—with domain-specific instrumentation: a robust active-policy rule that reconciles renewals and cancellations via request traces; product-year inflation normalization of premiums; and a stability rule that prioritizes retaining prior assignments when distance changes are marginal, consistent with contemporary guidance on clustering stability and validation (Liu, 2022; Ullmann, Hennig, & Hausdorf, 2022). On top of cluster membership, we compute a recency–frequency–monetary (RFM) value tier and translate it into actionable levers—additional tenure, broader distinct product breadth, or incremental (normalized) premium—required for a customer to progress to the next tier (Chen, Fan, & Sun, 2018; Kapoor & Bawa, 2023).

From a systems perspective, segmentation that is stable and auditable is as important as segmentation that is accurate. Downstream operations rely on assignment continuity to measure campaign effects, allocate budgets, and manage territories; even modest label drift can compromise comparability of quarterly metrics. Recent work on real-world ML underscores how seemingly small data/process issues propagate into "data cascades," undermining reliability if not addressed with explicit rules and validation (Sambasivan et al., 2021). At the same time, insurance data exhibits idiosyncrasies that complicate conventional clustering pipelines: (i) the definition of "active" coverage depends on cancellation/endorsement events recorded across systems; (ii) annual premium magnitudes are confounded by macro-level price dynamics and product revisions; and (iii) portfolios contain long tails and sporadic customers that can distort centroids if left unchecked. The proposed workflow responds to these issues through business rules, inflation normalization, and principled anomaly filtering, respectively.

Although the approach is general, we ground the study in a concrete setting: a portfolio described by dozens of engineered features derived from the operational tables above. The data scale is representative of real-world deployments and sufficiently large to stress the stability characteristics of centroid-based methods. We adopt k-means for transparency and communication benefits (Ahmed et al., 2020; Ikotun et al., 2023), use k-means++ initialization for convergence, and report internal and external diagnostics using contemporary validation guidance (Ullmann et al., 2022; Warrens, 2022). To make segment behavior inspectable by non-technical stakeholders, we track distances to centroids over time and surface a concise set of narrative features per segment (e.g., tenure, distinct active products, normalized premium).

The contribution of this paper is threefold: (i) a business-robust feature and rules layer tailored to insurance operations; (ii) a stability-aware clustering procedure aligned with recent validation literature; and (iii) actionable value tiers and upgrade prescriptions derived from an RFM overlay. The results indicate that the resulting segments are interpretable and operationally consistent, with high agreement across periods and intuitive economic profiles, suggesting broader transferability to other policy-like domains with strong regulatory and operational constraints.

II. LITERATURE REVIEW

Recent reviews emphasize that segmentation only creates business value when designs are interpretable, stable over time, and embedded in operational decision rules that downstream teams can trust; this applies with particular force in financial services and insurance where portfolios evolve and marketing actions must be auditable (Yum, Park, Oh, & Lee, 2022; Ikotun, Abualigah, Aljuaid, & Ezugwu, 2023). In this framing, segmentation quality is not solely a statistical concern but a socio-technical one that spans data provenance, methodological transparency, and organizational uptake.

Feature engineering for multi-table enterprise data. Automated and semi-automated feature engineering continues to mature, including neural and search-based approaches that reduce manual effort for tabular/relational settings while highlighting trade-offs between accuracy and interpretability (Chen, Luo, Jiang, & Wang, 2019; Baratchi et al., 2024; Mumuni & Mumuni, 2024). For explicitly relational stores (e.g., policy↔product↔request↔asset graphs), recent work explores programmatic extraction from entity–relation schemas and evaluates AutoFE tools on real databases, underscoring the need for domain constraints when features carry business semantics such as premiums or cancellations (Stanoev, Gyoshev, & Iliev, 2024; Dissanayake, Li, & Jayaratne, 2025).

Data preparation and anomaly control. Large, longitudinal enterprise pipelines face schema drift, duplicates, and temporal inconsistencies; contemporary surveys frame data cleaning as a primary reliability layer for ML (Senellart, Tannier, & Chapelle, 2024). For rare-pattern suppression, Isolation Forest and its extensions remain attractive due to scalability and distribution-free assumptions in high-dimensional settings (Hariri, Kind, & Brunner, 2019).

Clustering methods and validation. K-means remains widely adopted in operations due to geometric transparency and ease of deployment, while recent surveys catalogue alternatives (mixture models, density-based and hierarchical methods) and advise matching algorithms to shape/noise characteristics (Ikotun et al., 2023). Quality assessment has likewise evolved: beyond cohesion–separation indices, newer treatments clarify the interpretation of agreement measures such as the Adjusted Rand Index and how they can support drift monitoring across refreshes (Warrens, 2022; Zou, Wang, Wang, & Li, 2024).

These perspectives align with practice needs in regulated domains, where stable partitions and explainable diagnostics are prerequisites for governance.

Value tiering and sequence-aware extensions. RFM persists because it yields transparent value strata and simple operational levers; recent empirical and survey work documents its effectiveness and common hybrids with clustering (Christy, Umamakeswari, Priyatharsini, & Neyaa, 2021; Alves Gomes, Almeida, & Moro, 2023; Wong & Vu, 2024). At the same time, newer variants incorporate order effects and temporal dynamics—an avenue relevant to insurance renewal cycles (e.g., sequence-aware/Markov-style RFM and fast RFM over event streams) (Zheng, Zhou, & Jin, 2024).

Economic normalization and MLOps considerations. For monetary features, current best practice is to deflate nominal amounts using recognized price manuals and product-specific indices to support cross-year comparability (ILO/IMF/OECD/UNECE/World Bank, 2020). Finally, sustaining segmentation in production-like contexts requires attention to MLOps: versioned transformations, monitoring, and documentation of label stability are now standard recommendations in recent surveys of ML systems engineering (Kreuzberger, Kühl, & Hirschl, 2023).

Positioning. Building on these strands, our study targets the intersection of (i) relational feature engineering under domain rules (active-policy logic and inflation normalization), (ii) anomaly-aware clustering with interpretable diagnostics, and (iii) RFM overlays that translate segment placement into actionable, governance-friendly prescriptions. This orientation is consistent with contemporary guidance that ties methodological choices to business traceability and longitudinal stability (Yum et al., 2022; Ikotun et al., 2023; Warrens, 2022).

III. MATERIALS AND METHODS

3.1 Data sources and cohort

We construct a consolidated customer table by joining policy, product, request/activity, entity (vehicle, building, health), customer master, and address datasets via platform keys. After screening (below), the analysis set contains 379,584 customers represented by 38 numeric features that feed the clustering step.

Table 1: Summary statistics of anonymized dataset features (F001–F038).

Feature	Missing_%	Unique_Count	Min	P25	Median	P75	Max	Mean	Std
F001	0.0	17	0.0	0.0	0.0	0.0	28.0	0.03	0.23
F002	0.0	77	0.0	0.0	1.0	1.0	533.0	1.07	1.92
F003	0.0	21	0.0	0.0	0.0	0.0	400.0	2.68	16.3
F004	0.0	14	0.0	0.0	0.0	0.0	100.0	0.63	7.8
F005	0.0	195	0.0	0.0	100.0	100.0	100.0	60.69	45.4
F006	0.0	9	1.0	1.0	2.0	2.0	9.0	1.67	0.76
F007	0.0	6	1.0	1.0	1.0	1.0	6.0	1.17	0.4
F008	0.0	5	0.0	0.0	0.0	1.0	4.0	0.34	0.53
F009	0.0	22	0.0	0.0	0.0	1.0	32.0	0.48	0.91
F010	0.0	5	0.0	0.0	0.0	0.0	4.0	0.01	0.11
F011	0.0	16	0.0	0.0	0.0	0.0	29.0	0.02	0.23
F012	0.0	8	0.0	0.0	0.0	1.0	7.0	0.49	0.83
F013	0.0	596	0.0	0.0	0.0	0.0	50.0	1.95	6.14
F014	0.0	10	0.0	1.0	1.0	3.0	9.0	1.91	1.73
F015	0.0	4	0.0	0.0	0.0	0.0	3.0	0.32	0.67
F016	0.0	3	0.0	0.0	0.0	0.0	2.0	0.01	0.08
F017	0.0	4	0.0	0.0	0.0	0.0	3.0	0.01	0.13
F018	0.0	16	0.0	2.0	2.0	3.0	19.0	2.2	1.07
F019	0.0	8	0.0	1.0	1.0	1.0	7.0	1.18	0.52
F020	0.0	8	0.0	0.0	0.0	0.0	7.0	0.36	0.73
F021	0.0	115	1.0	1.0	3.0	4.0	1048.0	3.65	4.83
F022	0.0	115	1.0	1.0	3.0	4.0	1048.0	3.65	4.83
F023	0.0	122	0.0	1.0	2.0	4.0	698.0	2.85	4.54
F024	0.0	49	0.0	0.0	0.0	0.0	75.0	0.1	0.72
F025	0.0	24	0.0	0.0	0.0	0.0	32.0	0.06	0.4
F026	0.0	512	0.0	1.0	2.0	4.0	2288.0	3.94	25.04
F027	0.0	5	0.0	0.0	0.0	0.0	4.0	0.05	0.23
F028	0.0	7	0.0	1.0	2.0	2.0	6.0	1.51	1.11

F029	0.0	4	0.0	0.0	0.0	0.0	3.0	0.06	0.28
F030	0.0	16	0.0	0.0	0.0	0.0	23.0	0.01	0.16
F031	0.0	14	0.0	0.0	0.0	0.0	18.0	0.01	0.18
F032	0.0	29	0.0	0.0	0.0	0.0	77.0	0.05	0.43
F033	0.0	324903	-77120.68	7678.78	18619.21	37376.54	13056882.06	27705.22	43917.11
F034	0.0	115	0.0	34.0	43.0	55.0	126.0	44.14	17.34
F035	0.0	92460	-32870.88	0.0	0.0	9036.85	11469006.83	8046.52	26504.11
F036	0.0	85457	-32870.88	0.0	0.0	8337.28	486791.27	7643.11	16146.1
F037	0.0	2130	-3252.0	0.0	0.0	0.0	1532149.3	227.9	6411.98
F038	0.0	4956	-2238.9	0.0	0.0	0.0	319175.78	80.93	1307.09

Table 1 summarizes the anonymized feature space used for clustering. Each variable is identified by a masked label (F001–F038) to preserve proprietary semantics. For every feature, we report the share of missing observations (Missing_%) and the number of distinct values (Unique_Count) to characterize data completeness and cardinality prior to modeling. Distributional shape is conveyed through the five-number summary (Min, P25, Median, P75, Max) alongside the Mean and Std, all computed on the analysis cohort before standardization or winsorization. For non-numeric fields, the table lists the three most frequent categories and their counts (Top1–Top3), facilitating a quick scan for dominant modes. Because several monetary or ratio-based variables are inflation-adjusted or differenced, zeros and occasional negatives are expected at this stage; later pipeline steps operate on the standardized versions of these inputs. A secure internal data dictionary maps Fxxx to business concepts for governance and reproducibility but is not disclosed externally.

3.2 Eligibility, de-duplication, and hygiene

Transferred/invalid policies: entries flagged with VBS_TRANSFER_STATUS == 1 are removed; IC_POLICY_NO is coerced to a valid integer space after range/type checks. Customer type: we restrict to natural persons (CUSTOMER_TYPE == 1, mapping 3 → 1 where present). Link validation: customers without a valid policy linkage are excluded. Temporal fields: dates are parsed from mixed formats; the sentinel year 1900 is nulled. We recompute age and entry-age. These guardrails minimize schema noise and reduce the likelihood that malformed records distort centroids.

3.3 Determining active coverage

Accurately identifying active policies is central in insurance. We combine dates, cancellation indicators, endorsement/request traces, and product metadata.

- Base criterion: a policy is considered active if END_DATE >t_ref and PROD_CANCEL == 'T'.
- Cancellation evidence: policies with endorsements (IC_ENDORS_NO ≠ 0) are cross-referenced against request logs carrying RESULT_EXPLANATION ∈ {'İptalGerçekleşti'}; matches are marked inactive.
- Missing REQUEST_ID: for such records, we synthesize a key (IC_POLICY_NO-IC_RENEWAL_NO) and apply a premium heuristic: zero total gross premium ⇒ inactive; otherwise we check whether ENTRY_DATE >t_ref.
- Aggregation: is_policy_active is then rolled up per customer and by product family (auto, health, home). This layered rule set reconciles asynchronous back-office processes and prevents "ghost" coverage from leaking into analytics.

3.4 Inflation normalization for premiums

Nominal LC_GROSS_PREMIUM is converted to product–year real terms using a curated coefficient table, producing guncel_LC_GROSS_PREMIUM. We aggregate both overall and by family, and also for the subset of active policies. The adjustment removes price-level drift and product repricing effects so that customers remain comparable across vintages.

3.5 Feature construction

The engineered feature set blends breadth, behavior, and value.

- Portfolio breadth: counts of active policies overall and by family; distinct active product numbers and family codes.

- Request dynamics: total request volume; unique requested products by family; pure-web share inferred from activity initiators (e.g., WEB, EMAIL, ADWORDS) and successful outcomes.
- Customer attributes: tenure (from entry date, policy endpoints, and request closures), age, metropolitan indicator.
- Monetary: product-year–deflated premium totals overall and by family; ratios of family premiums to active-policy premium. Where appropriate, extreme monetary values are winsorized prior to scoring.

3.6 Outlier suppression

To reduce leverage from rare patterns, we fit IsolationForest with contamination 0.0001 and random_state=0. This keeps the dominant structure intact while filtering atypical records unlikely to generalize.

3.7 Scaling and clustering

All clustering inputs are standardized with StandardScaler. We then train k-means with k=7, init='k-means++', max_iter=300, n_init=10, random_state=0. In addition to labels, we compute Euclidean distances to every centroid to support interpretation and downstream rules. Clusters receive descriptive names and are summarized with narrative features (mean tenure, product breadth, real premium).

3.8 Stability mapping against prior labels

To maintain continuity for KPI tracking, we compare a customer's proximity to their previous centroid and the new one. Let d_old be the distance to the prior segment's centroid and d_new to the newly assigned centroid. We compute

$$\Delta_i = \frac{\left|\alpha_i^{new} - \alpha_i^{old}\right|}{\alpha_i^{old}}$$

If the relative change $\Delta \leq \tau$ with $\tau=0.4$, we retain the previous label; otherwise, we accept the new label. Stability is summarized with the Adjusted Rand Index and an inter-period transition matrix.

3.9 RFM tiering and individualized "upgrade" math

Above the clusters, we apply an RFM layer to convert value into concrete levers.
Component scores:

- R = scaled tenure;
- F = scaled count of distinct active products;
- M = winsorized, scaled deflated premium.

Weighted composite:
RFM_Score = 0.2·R + 0.3·F + 0.5·M.

Tier cut-points at the 0.55, 0.80, and 0.97 quantiles define A, B, C, D.
Prescriptions invert the scaling to compute additional tenure, extra distinct products, or incremental premium required for the customer to cross the next-tier threshold.

3.10 Reproducibility and operations

We serialize and version the scaler and k-means artifacts; the feature list and business rules are treated as data contracts. Each monthly run outputs labeled data with centroid distances and stability decisions, segment summaries, and diagnostics (ARI, transition tables). These artifacts support governance, experimentation, and integration with marketing channels.

IV. RESULTS AND DISCUSSION

4.1 Results

To protect proprietary information, we (i) refer to segments as Cluster 0–6 and tiers as Tier A–D, (ii) report percentages only for size and contribution, and (iii) replace raw per-cluster metrics with relative ranks rather than numeric values.

Segment composition:
Seven clusters with the following portfolio shares —

- Cluster 1: 57.39%;
- Cluster 4: 13.82%;
- Cluster 3: 12.22%;
- Cluster 5: 12.06%;
- Cluster 0: 2.08%;
- Cluster 2: 1.62%;
- Cluster 6: 0.81%.

Temporal stability: successive monthly runs remain highly consistent (Adjusted Rand Index = 0.9758; ≥96% on the diagonal of the transition matrix), consistent with the stability rule in §3.8.

Quality metrics and algorithmic comparison.
For internal quality control, we summarize (i) the Silhouette coefficient, which captures cohesion/separation on standardized inputs (range −1 to 1; higher indicates tighter, more separated clusters), and (ii) the Adjusted Rand Index (ARI) computed across consecutive monthly refreshes, which quantifies partition agreement while correcting for chance. We evaluate three methods under the same preprocessing and confidentiality constraints.

Table 2: Clustering quality across algorithms

Algorithm	**Silhouette (mean)**	**ARI across monthly refreshes**
K-means (k=7, k-means++)	0.41	0.9758
Gaussian Mixture Model (7 comps, full cov.)	0.36	0.942
HDBSCAN (min_cluster_size tuned)	0.29	0.901

Silhouette computed on the standardized features used for clustering. ARI computed between consecutive monthly runs after applying the stability rule (§3.8). No absolute counts are disclosed; percentages only per confidentiality policy.

In our setting, k-means achieves the best balance of compactness and separation (highest Silhouette) and the most longitudinal agreement (highest ARI). This pattern is consistent with the portfolio's approximately convex geometry after scaling and inflation normalization (§§3.4–3.7) and with our distance-based stability mapping (§3.8), which naturally complements centroid-based assignments. By contrast, full-covariance GMM components tend to overlap around mid-value cohorts, and HDBSCAN's sensitivity to local density produces variable cluster counts and a small "noise" share, both of which reduce temporal agreement.

Cluster fingerprints: ranks (1=lowest, 7=highest) for tenure (T), product breadth (B), normalized premium (P), and RFM (R).

• Cluster 0: T=5, B=4, P=4, R=4
• Cluster 1: T=1, B=1, P=1, R=1
• Cluster 2: T=2, B=2, P=3, R=2
• Cluster 3: T=3, B=6, P=6, R=7
• Cluster 4: T=7, B=3, P=5, R=5
• Cluster 5: T=2, B=5, P=2, R=3
• Cluster 6: T=6, B=7, P=7, R=6

Value distribution by RFM tiers:
Premium contributions, expressed solely as shares of the observed total:

- Tier A: 21.5%;
- Tier B: 27.5%;
- Tier C: 34.6%;
- Tier D: 16.4%.

No currency amounts are disclosed; absolute values are retained only in internal dashboards.

Upgrade guidance (qualitative):

The individualized "upgrade" engine is reported here without numeric thresholds. Findings are summarized qualitatively:

- A → B: typicallylow-friction, most often achieved via +product breadth; modest premium adjustments also help.
- B → C: moderate hurdle; breadth and premium both matter, with tenure occasionally decisive for edge cases.
- C → D: high hurdle requiring a combination of levers (breadth and premium; tenure plays a supporting role).
- D: already top tier; no upgrade suggested.

(Exact tenure/product/premium requirements per customer are kept internally; external documents receive only categorical indications—low, moderate, high—to avoid leakage of sensitive levels.)
Notes on naming policy (external vs. internal)
External reports, including this paper, use neutral identifiers (Cluster 0–6, Tier A–D). Internally, we assign descriptive names by inspecting dominant attributes (e.g., high active-housing requests, strong cross-family breadth, or premium concentration) to assist marketing and CX teams while safeguarding trade secrets.

4.2 Discussion

The segmentation reveals a large low-engagement majority (≈57% in Cluster 1) alongside This section interprets the empirical findings, outlines operational implications, and reflects on limitations and governance. To respect confidentiality, we continue to refer to segments as Cluster 0–6 and value tiers as Tier A–D, avoid raw magnitudes, and discuss patterns primarily in percentages, ranks, or qualitative terms.

What the clusters suggest about the portfolio

The segmentation reveals a large low-engagement majority (≈57% of customers in Cluster 1) alongside smaller, commercially attractive groups with higher relative breadth or value signals (e.g., Clusters 3 and 6, by rank). This shape is not unusual in insurance portfolios, where long tails and single-line coverage are common, while multi-line, higher-value cohorts are comparatively compact (Ikotun, Ezugwu, Abualigah, Agushaka, & Alo, 2023; Alves Gomes, Almeida, & Moro, 2023). The relative rankings (Section Results-C) indicate that breadth and normalized premium tend to rise together—a pattern consistent with contemporary cross-sell economics—and that tenure alone does not fully determine value, which underscores the importance of combining relationship duration with breadth and spend (Chen, Fan, & Sun, 2018; Alves Gomes et al., 2023).

From a go-to-market perspective, the portfolio calls for two complementary plays: (i) broad recovery and activation aimed at the largest, low-engagement cluster; and (ii) targeted bundling and retention for the smaller, high-breadth/high-value clusters. Because our upgrade engine expresses movement between tiers in qualitative difficulty bands (A→B: low friction; C→D: high), marketers can match budget intensity and offer depth to the expected lift.

Stability as an enabler of longitudinal measurement

Assignment continuity is crucial for quarterly KPI tracking, territory planning, and experimentation. The observed Adjusted Rand Index of 0.9758 and ≥96% diagonal in the transition matrix indicate that the stability mapping is working as intended: most customers retain their labels across refreshes; movements concentrate among boundary cases where centroid distances shift more than the tolerance (Warrens, 2022; Ullmann, Hennig, & Hausdorf, 2022). This matters practically: stable labels allow reliable uplift attribution, budget allocation, and test–control designs without re-baselining every period (Kreuzberger, Kühl, & Hirschl, 2023; Sambasivan et al., 2021).

A useful operating guideline is to monitor three guardrails per run: (1) ARI relative to the prior snapshot; (2) share change per cluster (percent, not counts); and (3) centroid-distance drift for a panel of boundary customers. Breaches trigger inspection before campaign roll-outs.

Naming strategy: interpretability without leakage

Externally we publish neutral identifiers (Cluster 0–6, Tier A–D). Internally, we maintain human-readable nicknames to aid field teams. Importantly, names are derived systematically from salient attributes—e.g., a high share of active housing requests or above-median product breadth prompts a housing or multi-line

cue in the internal label. This balances interpretability with confidentiality by keeping the taxonomy private while ensuring that playbooks remain intuitive.

Implications for targeting, cross-sell, and upsell

Three immediate applications emerge:

• **Activation of the majority cluster.** Because A→B upgrades are typically low-friction, the emphasis is on simple product adds or light monetization offers; tenure contributes but is less decisive than breadth at this boundary.

• **Bundling for high-breadth groups.** For clusters ranked highest on breadth and value, bundle messaging and retention incentives are appropriate—especially where C→D upgrades are high-hurdle and require multiple levers (breadth and premium).

• **Channel and creative tailoring.** The pipeline's pure-web signal (reported only as a share or rank) helps prioritize digital vs. assisted channels, while centroid-distance diagnostics flag edge cases where messaging should emphasize the most attainable lever (product add vs. spend vs. tenure).

Limitations and threats to validity

• **Geometry of K-Means.** Euclidean clustering assumes roughly spherical structure after scaling; non-convex or highly skewed manifolds may be under-segmented. That said, the method's transparency and speed remain advantageous for operational use (Ahmed, Seraj, & Islam, 2020; Ikotun et al., 2023).

• **Sensitivity to feature scaling and contamination.** Although Isolation Forest removes a very small fraction of anomalies, false positives/negatives may occur; we mitigate this with conservative contamination and post-hoc spot checks (Hariri, Carrasco-Kind, & Brunner, 2019; Pang, Shen, Cao, & van den Hengel, 2021).

• **Inflation normalization granularity.** Product–year coefficients reduce macro price effects but cannot capture every micro-pricing change; future work could incorporate policy-level repricing markers where available.

• **External validity.** The cluster shapes reflect current distribution, pricing, and product design. Major business changes (e.g., new product lines) can legitimately alter the geometry; monitoring (Section B) and periodic re-tuning of k remain essential (Zou, Wang, Wang, & Li, 2024; Ullmann et al., 2022).

Managerial guidance and "knobs" to tune

• **Stability tolerance (τ).** Tightening τ increases label retention (fewer changes) at the risk of under-reacting to genuine shifts; loosening τ does the opposite.

• **Anomaly contamination.** Lower contamination reduces false pruning but may admit more leverage points; adjust in small steps and re-check ARI and share shifts.

• **Tier thresholds.** Moving quantiles (e.g., 0.55/0.80/0.97) tunes cohort sizes in Tiers A–D; keep premium shares (percentages) under review to avoid over-concentration.

• **Playbook linkage.** For each cluster, maintain a one-page brief (no raw values) listing: top two levers (by rank), preferred channel (by share), and example creative; retire or update briefs when ARI/share guardrails are breached.

In sum, the segmentation behaves as a stable foundation for marketing and CRM, with clear routes for activation and bundling and with confidentiality preserved through neutral labels, percentage-only reporting, and relative ranks. The approach is extensible: additional product families or behavioral signals can be added via the same feature contracts, and governance metrics (ARI, share drift, centroid proximity) provide early warnings as the portfolio evolves.

V. CONCLUSION

We introduced a deployable segmentation pipeline that converts heterogeneous, multi-table insurance data into stable, decision-ready clusters and customer-level upgrade guidance, while safeguarding sensitive information in public artifacts through neutral identifiers, percentage-only reporting, and rank-based summaries. The system blends domain logic—most notably a resilient active-policy definition and product–year premium deflation—with principled ML components: Isolation Forest for anomaly control, standardized scaling, and k-means augmented with centroid-distance diagnostics and an explicit stability rule. Each monthly run emits auditable by-products (labels, centroid distances, transition matrices, agreement indices) that support governance and downstream activation.

Empirically, the solution demonstrates strong temporal consistency (Adjusted Rand Index ≈ 0.976) with transition matrices dominated by the diagonal (≥96%), indicating that the stability heuristic preserves continuity yet permits genuine boundary moves. Revenue concentration appears at the upper end of the value tiers (reported as shares rather than currency), aligning with expectations for multi-line portfolios and enabling budgeted activation and bundling strategies. Put together, these findings suggest that a business-aware, stability-conscious approach can underpin CRM, experimentation, and planning over successive refreshes (Kreuzberger, Kühl, & Hirschl, 2023; Ullmann, Hennig, & Hausdorf, 2022; Warrens, 2022).

5.1 Future Work

Model class and choice of *k*.
Examine alternatives that relax spherical assumptions—Gaussian mixtures for ellipses, density-based methods such as HDBSCAN for variable-density structure, or constrained k-means to encode business rules. Pair these with stability-aware selection using contemporary validity indices (e.g., silhouette, Davies–Bouldin) and recent guidance on cluster validation and index behavior (Ahmed, Seraj, & Islam, 2020; Ikotun, Ezugwu, Abualigah, Agushaka, & Alo, 2023; Zou, Wang, Wang, & Li, 2024).

Timelier refresh and streaming.
Consider mini-batch or incremental assignment to support intra-month updates with bounded latency, while tracking agreement to the monthly **baseline** (Huang, Wang, & Li, 2020; Kreuzberger et al., 2023).

Richer alignment across runs.
Replace the fixed tolerance with run-to-run label alignment (e.g., optimal-assignment/Hungarian alignment on centroids) and customer-specific confidence bands derived from distance distributions; continue monitoring ARI and share **shifts** (Ullmann et al., 2022; Warrens, 2022).

Causal lift and experimentation.
Layer uplift/heterogeneous treatment-effect models over segments and tiers to target offers with the greatest causal return; validate with controlled **experiments** (Athey & Wager, 2019; Nie & Wager, 2021; Künzel, Sekhon, Bickel, & Yu, 2019).

Drift, fairness, and governance.
Add automated data- and concept-drift alerts (e.g., PSI/KL on features and centroid distances) and routine fairness checks by geography and channel; expose these in an MLOps dashboard to curb long-term technical debt (Lu, Liu, Li, & Zhang, 2018; Kreuzberger et al., 2023; Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021; Mitchell et al., 2019).

Economic normalization.
Enrich deflators with finer-grained price markers (policy-level where available) or external indices to further separate behavioral change from macro **effects** (International Labour Office et al., 2020).

Personalization beyond tiers.
Use cluster membership and tier ranks as priors for contextual bandits/next-best-action systems, while maintaining the external reporting policy (percentages, ranks only) (Bouneffouf & Rish, 2019; Athey & Wager, 2019).

Campaign scenario prototyping.
Develop end-to-end "campaign playbooks" per cluster–tier pair that specify target criteria, primary lever (breadth vs. premium vs. tenure), channel mix (guided by digital-share signals), expected lift bands, and test design; evaluate with sequential experimentation or contextual bandits (Bouneffouf & Rish, 2019; Athey & Wager, 2019).

Naming protocols.
Codify an internal naming rubric that maps salient attributes to descriptive labels (e.g., breadth-oriented, housing-leaning) while keeping public identifiers neutral (Cluster 0–6; Tier A–D). Include governance for periodic review to avoid leakage and preserve interpretability for field teams (Sambasivan et al., 2021; Mitchell et al., 2019).

Fairness auditing.
Introduce routine fairness diagnostics for upgrade recommendations and campaign eligibility (e.g., demographic parity/equality of opportunity across age bands, regions, or channel cohorts), alongside explainability summaries to support business review (Mehrabi et al., 2021; Mitchell et al., 2019).

VI. DATA AVAILABILITY

The data underlying this study are proprietary and cannot be shared.

VII. FUNDING

This research did not receive any financial support.

REFERENCES

[1] Ahmed, M., Seraj, R., & Islam, A. B. M. S. (2020). The k-means algorithm—A comprehensive survey and performance evaluation. *Electronics, 9*(8), 1295.
[2] Alves Gomes, R., Almeida, J. J., & Moro, S. (2023). Customer segmentation with RFM and clustering: A systematic literature review. *Journal of Retailing and Consumer Services, 73*, 103291.
[3] Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests. *Observational Studies, 5*(2), 37–51.
[4] Bouneffouf, D., & Rish, I. (2019). A survey on practical applications of multi-armed and contextual bandits. *arXiv*.
[5] Chen, Y., Fan, M., & Sun, Y. (2018). Customer segmentation based on RFM model and K-means clustering. *IEEE International Conference on Big Data (Big Data)*, 5452–5455.
[6] Côté, M.-A., Sakka, Z., Senellart, P., & Tannier, X. (2024). Data quality for machine learning: A survey. *ACM Computing Surveys, 56*(10), 1–41.
[7] Hariri, S., Carrasco-Kind, M., & Brunner, R. J. (2019). Extended Isolation Forest. *IEEE Transactions on Knowledge and Data Engineering* (early version: *arXiv:1811.02141*).
[8] Huang, S., Wang, J., & Li, Z. (2020). Scalable k-means variants: A review with empirical comparisons. *IEEE Access, 8*, 40496–40514.
[9] Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Agushaka, J. O., & Alo, O. S. (2023). A comprehensive survey of clustering algorithms: State-of-the-art, taxonomy, trends, and challenges. *Information Sciences, 622*, 178–222.
[10] Ikotun, A. M., Abualigah, L., Aljuaid, H., & Ezugwu, A. E. (2023). Clustering methods: A review of algorithms and applications. *IEEE Access, 11*, 18427–18453.
[11] International Labour Office (ILO), International Monetary Fund (IMF), Organisation for Economic Co-operation and Development (OECD), United Nations Economic Commission for Europe (UNECE), & World Bank. (2020). *Consumer Price Index Manual: Concepts and Methods*. International Monetary Fund.
[12] Kapoor, A., & Bawa, P. (2023). RFM-based customer segmentation: Methods, metrics, and marketing implications. *Decision Analytics Journal, 9*, 100185.
[13] Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access, 11*, 108388–108403.
[14] Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences, 116*(10), 4156–4165.
[15] Lu, J., Liu, A., Li, C., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering, 31*(12), 2346–2363.
[16] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 115.
[17] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 220–229.
[18] Mumuni, A., & Mumuni, F. (2024). Automated feature engineering for tabular data: A systematic review. *Expert Systems with Applications, 242*, 122757.
[19] Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika, 108*(2), 299–319.
[20] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.
[21] Stanoev, V., Gyoshev, S., & Iliev, I. (2024). Programmatic feature engineering from relational schemas: A review and benchmark. *Data & Knowledge Engineering, 147*, 102257.
[22] Ullmann, T., Hennig, C., & Hausdorf, B. (2022). Validation of cluster analysis results using external information. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12*(6), e1471.
[23] Warrens, M. J. (2022). Some properties of the adjusted Rand index. *Journal of Classification, 39*(2), 309–323.

[24] Wong, C. H., & Vu, H. Q. (2024). RFM revisited: Transparent value stratification for digital customer analytics. *International Journal of Information Management Data Insights, 4*(2), 100211.
[25] Yum, H., Park, S., Oh, J., & Lee, H. (2022). Customer segmentation for data-driven marketing: A systematic literature review. *Sustainability, 14*(3), 1456.
[26] Zheng, Y., Zhou, D., & Jin, H. (2024). Sequence-aware RFM for streaming customer value analysis. *Expert Systems with Applications, 246*, 123123.
[27] Zou, X., Wang, S., Wang, X., & Li, X. (2024). Cluster validity indices: A comprehensive survey and experimental evaluation. *ACM Computing Surveys, 56*(8), 1–46.