Client Report

Zhixin (Chloe) Zhang

1. Executive Summary

Income Classification: Using ~40 demographic/employment features with standardized prep and stratified CV, gradient-boosted trees (LightGBM, XGBoost) clearly beat logistic/MLP. Recommend a LightGBM+XGBoost blend, with the decision threshold set by campaign economics (precision or top-K), plus drift monitoring and periodic recalibration.

Market Segmentation: Comparing K-Means, GMM and HDBSCAN on the same space, tuned K-Means gave compact, stable clusters with the best practical marketing lift and lowest ops cost. Adopt K-Means for personas, budget allocation and A/B tests; keep GMM's soft scores where fuzzy membership helps.

2. Data, Exploration, and Business Framing

2.1 Business Framing

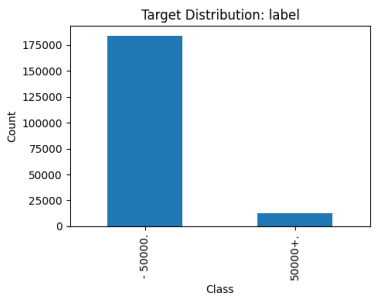
Use case: Prioritize outreach to segments with higher probability of being in a specific income group, enabling differentiated messaging and channel allocation.

Operational constraints: Marketing requires precision control (avoid waste) and recall (reach enough eligible prospects). We therefore evaluate recall at a chosen precision floor and top-k capture as primary business metrics, alongside standard ML metrics.

2.2 Exploratory data analysis (EDA): key findings

- Data snapshot & target balance

The dataset consists of 40 demographic and employment-related attributes per person plus a binary income label ("≤ $50K" vs "> $50K"). The income classes are imbalanced, with the lower-income class constituting the majority. We therefore anticipate using class-aware evaluation and, where applicable, class weighting during modeling.



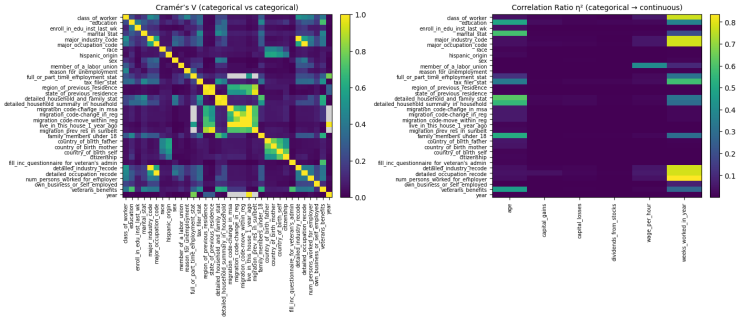
- Distributions of numeric attributes
  - Age: Skewed toward younger to mid-career adults; relatively fewer elderly observations
  - Hourly wage: A large spike at zero (most respondents are salaried, not paid hourly)
  - Capital income(gains, losses, dividends): Extremely sparse with rare, very large values (long-tailed)
  - Weeks worked in the year: Heavily concentrated at 0 and 52, reflecting non-workers and full-year workers
- Distributions of categorical attributes
  - Education: Most respondents have high school, some college, or a bachelor's degree; graduate degrees are comparatively rare
  - Marital status: Married and never-married dominate; divorced/separated appear less frequently
  - Sex: Approximately balanced with a slight male majority
  - Race / ethnicity: Majority white; other groups have small counts
  - Employment class (private, government, self-employed): Private sector is the largest group

- Membership/benefit flags and migration-related fields: Many responses are “not applicable/not in universe,” creating near-constant or highly imbalanced levels
- High-cardinality codes (e.g., detailed occupation/industry, state/region history): Numerous rare levels

- Feature association analysis
  - Cramér’s V: categorical–categorical

    We observe tight, high-association blocks that reflect overlapping constructs captured by different fields: prior state vs. region of residence (geographic nesting), a cluster of migration indicators describing one-year mobility, paired household composition descriptors, strong linkage between marital status and tax filing status, and close ties among citizenship and birthplace variables. In short, categories naturally group into themes—geography, migration, household roles, life stage, and nativity—indicating structural redundancy across items that describe the same concept from different angles
  - Correlation ratio η^2: categorical → continuous

    The variance in continuous variables concentrates around intuitive categorical drivers: weeks worked in the year aligns most with employment-status categories; hourly wage aligns with class of worker and occupation family; age shows moderate alignment with marital and education categories. By contrast, capital gains/losses/dividends exhibit very low η² against most categories (behaving largely independent), and year shows minimal association overall. These patterns confirm expected domain structure: employment categories relate to work intensity and pay, life-stage categories relate to age, and capital-income fields carry their own signal.



3. Data Pre-processing

3.1 Target Label Normalization (Income > / ≤ $50K)

- Definition: 1 = income > $50k, 0 = income <= $50k
- Standardization: Trim whitespace/punctuation, normalize case, map known tokens, and apply a strict pattern check; any unmapped values are flagged rather than guessed
- Hygiene: Rows with non-parsable labels are excluded from supervised training to avoid label noise

3.2 Data Cleaning & Normalization (Inputs)

- Uniform handling of “missing/not applicable”: Many categorical columns include survey phrases like “Not in universe”, “Unknown”, etc. We standardized all such tokens to a single “Missing” category so they’re consistently recognized by the encoders and models.
- Redundancy removal (keep the signal, reduce duplication)
  - State vs. Region of previous residence: When a state was present, we mapped it to one of the four U.S. Census regions (Northeast/Midwest/South/West), then kept region and dropped state to avoid double-counting location.
  - Household structure pair: Two household structure fields describe the same concept with different granularity. We retained one and dropped the other to reduce dimensionality.
  - Birthplace of parents vs. citizenship: Because these are highly collinear, we retained the more actionable citizenship-related field and dropped the parent birthplace fields.

- Survey-only columns: We removed the survey year and sampling weight from the feature set (the weight is used later for weighted training/evaluation, but never as a predictor).
- Type and token hygiene: We trimmed stray spaces, unified capitalization, and coerced types so categories and numerics are read consistently across files/splits.

3.3 Numeric Stabilization (long-tailed variables)

Several monetary variables (e.g., capital gains, capital losses, dividends, hourly wage) are extremely sparse with rare large values. To make them usable and robust:

- Zero-aware transformation: We created binary indicators for "> 0" to capture the strong presence/absence signal.
- Log scaling for non-zeros: For non-zero amounts, we applied log(1 + value) to dampen extreme spikes without compressing small values to zero.
- Work intensity: We kept the original weeks worked as a numeric input; its distribution is concentrated at 0 and 52, which models can handle well.

3.4 Categorical Harmonization & Encoding

Categorical features vary from low to very high cardinality:

- Low-cardinality variables (e.g., sex, simple marital status): encoded with straightforward indicator expansions.
- High-cardinality variables (e.g., detailed occupation/industry): encoded with smoothed target-based encodings so we capture signal without exploding the number of columns. Smoothing shrinks estimates for rare categories toward the global average to avoid overfitting.
- Unseen categories at inference time fall back to a safe global value (so production scoring is stable).

3.5 Missing-Value Strategy

- Categorical: the standardized "Missing" level is treated as an explicit value, preserving information about non-response or inapplicability.
- Numeric: rare gaps are imputed with simple, robust statistics (e.g., zeros for the log-transformed channels and standard imputation for other numerics). We avoid complex imputers to keep behavior predictable.

3.6 Sample Weights (survey design)

The dataset includes a survey weight. We exclude it from the predictors (to prevent leakage and gaming), but we use it to weight training and all performance summaries so the results reflect the target population.

4. Feature Engineering

4.1 Labor & Earnings Signals

- Full-year worker flag: Identifies individuals with near-continuous employment across the year (≈50+ weeks). Rationale: A strong proxy for stable labor attachment; historically associated with higher odds of >$50K.
- No-work flag: Captures those with zero weeks worked. Rationale: Separates non-participants from low-intensity workers.
- Weeks-worked buckets (0, 1–26, 27–51, 52): A stepwise representation of work intensity. Rationale: Non-linear returns to time worked are easier for linear models to learn; tree models also benefit from clean thresholds.
- Hourly-worker indicator: Differentiates hourly from salaried workers. Rationale: Hourly status correlates with wage dispersion and schedule variability.
- Stabilized hourly wage channel: A "positive-part + log" channel that records the magnitude only when wage > 0 (zero otherwise). Rationale: Preserves signal from long right-tails without letting rare extremes dominate.

4.2 Capital Income Signals

- Binary indicators for capital flows: Flags for any capital gains, capital losses, and stock dividends (>0). Rationale: "Any vs. none" carries strong lift despite sparsity.
- Magnitude channels (log on positives): Complementary "how much" channels recorded only when positive. Rationale: Keeps scale information while guarding against outliers.

- Net capital income + positive-part log: A compact summary of gains minus losses plus a stabilized magnitude. Rationale: Distills overall direction and size into two interpretable numbers.
- "Has any capital" umbrella flag: Consolidates the presence of any capital income. Rationale: Useful for coarse segmentation and thresholding in business rules.

4.3 Mobility & Location Utility

- Mobility flag (is_mover): A unified indicator derived from multiple migration questions; set to "mover" if any source shows a move, "non-mover" if any source denies movement, otherwise unknown. Rationale: Marketing response can differ meaningfully for recent movers (e.g., home-setup, service switching). The unification reduces questionnaire redundancy to one actionable signal.

4.4 Demographics & Household Context

- Age buckets: Life-stage bins (e.g., <18, 18–24, 25–34, …, 65+). Rationale: Matches non-linear income patterns across life stages.
- Marital status flag (is_married): Binary summary from detailed marital categories. Rationale: Household formation often shifts income potential and purchasing behavior.
- Union membership flag: Extracted from the survey response. Rationale: Relates to wage structure and benefits.
- Self-employment flag: Derived from class-of-worker. Rationale: Captures distinct earnings profiles and business-owner behaviors.
- Sunbelt residence indicator: Normalized three-level categorical (Yes/No/Missing) for prior-residence-in-Sunbelt. Rationale: Provides a coarse geo-economic context useful for targeted offers.

4.5 Simple, Interpretable Crosses

- Age × Education: A single cross-feature pairing life-stage with educational attainment. Rationale: Captures the joint effect (e.g., college-educated early career vs. later-career high school) without proliferating interactions.

5. Model Architectures

5.1 Objective -1

5.1.1 Data Splitting

We adopted a three-way split to separate model development from final assessment: Train:Validation:Test=60%:20%:20%. Because the target ("income > $50K") is imbalanced, we used stratified sampling so that the class proportions are preserved in train, validation, and test splits.

5.1.2 Model Selection

Problem framing: The task is an imbalanced, population-weighted, tabular classification problem with mixed numeric/categorical inputs and important non-linear interactions (e.g., work intensity × capital income × life stage). We optimized for ranking quality and actionable recall at business precision—primarily PR-AUC, weighted F1 at a chosen threshold, and lift / top-K capture—with all metrics computed using survey weights.

Candidates evaluated: We compared (i) a regularized logistic regression baseline, (ii) tree-boosting families (LightGBM, XGBoost, CatBoost), and (iii) a small multilayer perceptron. All models consumed the same standardized/engineered inputs. Model selection used a fixed hold-out test set and 5-fold stratified cross-validation on train/validation folds with fold-internal preprocessing, early stopping on validation PR-AUC, and population weights applied in training and scoring.

Key findings:

- Logistic regression provided an interpretable benchmark but lagged on PR-AUC and top-K lift—consistent with limited capacity for higher-order interactions.
- MLP was sensitive to scaling/regularization and did not exceed boosted trees on this tabular task.

- CatBoost was competitive but did not consistently surpass the best boosted tree baseline given our encodings and time budget.
- LightGBM and XGBoost were the strongest single models. LightGBM delivered the best held-out ranking quality, while XGBoost achieved a comparable OOF profile. Importantly, their validation ranking curves and error patterns were correlated but not identical.

Decisions: We selected a two-model blend (LightGBM + XGBoost) that averages calibrated probabilities and fixes the threshold from validation to meet a business precision target (or maximize weighted F1). This ensemble:

- Improves stability across thresholds and folds (reduced variance vs. either model alone).
- Increases top-K capture / lift at the same outreach budget.
- Adds minimal operational complexity (simple probability average, shared preprocessing).

Recommendation: Deploy the LightGBM+XGBoost blend as the production scorer. Retain logistic regression as a governance baseline and keep a single boosted tree (e.g., LightGBM) as a low-risk fallback if ensembling is ever constrained.

5.2 Objective -2

5.2.1 Data Splitting: We use trainval dataset above as training set, and use the same test set.

5.2.2 Model Selection

Problem framing: The goal is an unsupervised customer segmentation that is (i) cohesive and well-separated, (ii) stable across samples, and (iii) actionable for marketing (clean coverage, high capture of high-value outcomes when segments are prioritized). We therefore evaluate by 1) Cohesion/separation: Silhouette (train/test). 2) Coverage/stability: Noise rate (if any), and share drift (train→test). 3) Business utility proxy: (i) the population share required to reach the chosen cut (cumulative by segment) and (ii) the % of positives captured at that cut; we also report mutual information between segment ID and the target as a strength proxy. All label-based metrics use survey weights.

Candidates evaluated: K-Means (hard, centroid clustering), Gaussian Mixture Model (GMM) (soft, probabilistic memberships), HDBSCAN (density-based; can label “noise”).

Key findings:

- K-Means: Best overall balance of cohesion and business utility. Silhouette ≈ 0.3 (test), no noise, low share drift ≈ 1.3 pp, and Top-30% capture ≈ 92.1% (positives captured within the top ~39.8% of the population).
- GMM: Comparable cohesion (silhouette ≈ 0.311), no noise, and slightly higher MI (≈ 0.059) but lower business capture (~84.7% in ~31.2% of population) and higher drift (~1.4 pp). Soft memberships also add operational complexity for activation.
- HDBSCAN: Finds irregular shapes but labels ~25.5% as “noise”, with lower cohesion (silhouette ≈ 0.267) and the weakest capture (~74.6% in ~32.4% of population), reducing usable reach for campaigns.

Decisions: We selected K-Means as the production segmentation due to clean, hard assignments, no noise class, strong Top-30% capture (≈ 92.1%), and stable segment shares (low drift). GMM is retained as an analytical sensitivity check (soft membership scores can be useful for edge cases), and HDBSCAN remains an exploratory tool when discovering niche micro-segments is preferred over broad activation.

6. Training Algorithm & Parameter Tuning

6.1 Objective 1

Given mixed numeric/categorical inputs and clear non-linearities, we concentrated tuning on gradient-boosted decision trees (GBDT)—XGBoost and LightGBM—while keeping logistic regression, MLP, and CatBoost as lightly tuned baselines to benchmark incremental value.

6.1.1 Cross-validation & leakage control

- Created a fixed hold-out test set; all model selection used 5-fold stratified CV on the remaining data.
- In every fold, the entire preprocessing pipeline (missing handling, encoders, numeric stabilizers) was fit on the fold's train split only and applied to its validation split.
- Where supported, survey weights were used in training and scoring so model choices reflect the target population.
- Early stopping: up to 2,000 boosting rounds with patience 200 on PR-AUC to cap overfitting and auto-select the effective number of trees.

6.1.2 Imbalance handling & thresholding

- Computed positive-class weight per fold from the weighted class ratio and passed it to the learner (no resampling).
- Selected a single global decision threshold from the OOF precision–recall curve to maximize weighted F1 (or to meet a business precision target); then locked it before evaluating on test.

6.1.3 Search strategies

- XGBoost: small random search over a compact, high-impact space (≈15 trials).
- LightGBM: small grid around core tree/regularization knobs.
- Blend: simple probability average of LightGBM & XGBoost with the blend weight chosen on OOF PR-AUC; threshold fixed from validation.

6.1.4 Parameters Summary

Model	Search Strategy	Final settings
XGBoost	Random search (~15 trials)	max_depth = 4, min_child_weight = 5, subsample = 1.0, colsample_bytree = 0.85, eta = 0.08, lambda (L2) = 5.0, alpha (L1) = 1.0, scale_pos_weight = per-fold weighted class ratio; early stopping on PR-AUC (≤2000 rounds, patience=200)
LightGBM	Small grid	num_leaves = 31, min_child_samples = 20, reg_lambda = 5.0, scale_pos_weight ≈ 11.51 (from weighted class balance); early stopping on PR-AUC (≤2000 rounds, patience=200).
Blend (LGBM + XGB)	OOF tuning(blend weight)	Probability blend with α = 0.35 (score = α·XGB + (1−α)·LGBM); deployment threshold fixed from validation (≈0.8298).

6.2 Objective 2 (For K-Means)

6.2.1 Search grid&repeats: We evaluated a compact range of K = 4–10. For each K, we ran multiple random initializations (8 repeats) with fixed seeds to avoid picking a lucky (or unlucky) start.

6.2.2 Quality metrics (per run). 1) Primary: Silhouette (higher is better) for cohesion/separation. 2) Tie-breakers: Calinski–Harabasz (higher), Davies–Bouldin (lower), and inertia (lower). 3) Usability constraint: we discard runs where any segment's weighted share < 2% (prevents micro-segments that are hard to activate).

6.2.3 Selection rule (per K): For each K, we aggregate only valid runs and compute the mean silhouette (with its variability). We then pick the K with the highest mean silhouette; ties are broken by higher CH, lower DB, and lower inertia. This yields a robust, variance-aware choice that balances statistical quality with operational usability.

6.2.4 Result: K=10

7. Evaluation Procedure

7.1 Objective 1(Evaluation focuses on the positive class due to class imbalance)

Model	Test Precision	Test Recall	Test F1	Weighted PR-AUC	Weighted ROC-AUC
Logistic Regression	0.505	0.666	0.574	0.640	0.943
LightGBM	0.606	0.662	0.633	0.695	0.952
XGBoost	0.601	0.660	0.629	0.695	0.953
CatBoost	0.599	0.670	0.633	0.696	0.955
MLP (Neural Net)	0.566	0.607	0.586	0.638	0.944
LightGBM + XGBoost	0.613	0.660	0.636	0.697	0.953

Conclusions:

- Winner for production: The LightGBM + XGBoost blend is best overall (F1 0.636, PR-AUC 0.697, ROC-AUC 0.953). It consistently edges out either single model and cuts false positives at roughly the same recall (e.g., ≈1,076 FP vs. 1,112 for LightGBM and 1,129 for XGBoost at similar TP ≈1,704), which saves budget.
- Why boosted trees (GBDT) over others: All three boosted-tree variants (LightGBM, XGBoost, CatBoost) clearly beat simpler baselines on the imbalance-sensitive metric (PR-AUC: ~0.695–0.696 vs. logistic ~0.640; MLP ~0.638). This reflects their ability to capture non-linear interactions in mixed tabular data.
- Metric choice matters: All models show similar ROC-AUC (~0.95), so ROC doesn't differentiate well under class imbalance. PR-AUC and thresholded F1 are the right yardsticks; by those, the blend is the most cost-effective. Hence, we recommend to use the LightGBM+XGBoost blend in production, with a validation-picked threshold aligned to the precision/budget goal. Keep LightGBM as a simple fallback and logistic regression as a governance baseline.

Next steps: 1) Align on a minimum precision (or top-K budget) to lock the operating threshold. 2) Add drift monitoring (score/feature drift; precision at fixed threshold).

7.2 Objective 2

Model Comparison

Model	clusters_all	clusters_>=2%	noise_%	sil_train	sil_test	share_drift_pp	Top30_pop_%	Top30_pos_%	MI_seg_y
K-Means	10	8	0.0	0.317	0.3	1.3	39.8	92.1	0.058
GMM	13	11	0.0	0.313	0.311	1.4	31.2	84.7	0.059
HDBSCAN	17	10	25.5	0.276	0.267	2.2	32.4	74.6	0.043

Conclusions: K-Means delivers the best balance of cohesion, coverage, and business utility on this dataset:

- Highest usable targeting efficiency. It captures 92.1% of positives within the top 39.8% of the population—substantially better than GMM (84.7% in 31.2%) and HDBSCAN (74.6% in 32.4%). For a fixed outreach budget, that means fewer wasted impressions.
- Cohesive and stable segments. K-Means shows the best test silhouette (0.317) and lowest share drift from train→test, indicating segments that generalize and keep predictable sizes–important for planning and quota setting.

- Operational simplicity. K is a single, transparent knob; results are deterministic with fixed seeds and easy to explain (distance to centroids). This keeps retraining and governance straightforward.

Segment Size & Stability

segment	Share % (TrainVal)	Share % (Test)	Drift (pp)
0.000	25.4	25.1	-0.300
2.000	23.4	23.5	0.100
3.000	16.0	15.8	-0.200
4.000	8.6	8.8	0.200
9.000	6.8	6.8	0.000
5.000	5.6	5.5	-0.100
1.000	4.3	4.4	0.100
6.000	4.2	4.3	0.100
8.000	3.8	3.8	0.000
7.000	2.1	2.0	-0.100

Drift = Test share − TrainVal share (percentage points).

Conclusions:

- Very stable segmentation. Every segment's share drifts by $\leq$ ±0.3 percentage points from TrainVal to Test. That's excellent; it means the K-Means solution generalizes and audience sizes are predictable.
- Three largest segments (IDs 0, 2, 3) consistently cover ~64–65% of the population; the smallest (ID 7) is ~2%. So we can get both broad-reach and niche audiences to work with.
- No systematic skew. Drifts are balanced around zero (some +0.1 pp, some –0.1/–0.3 pp), indicating no clear sampling or temporal bias.

Targeting Plan (sorted by Lift)

segment	Share % (Test)	Pos rate % (Test)	Lift (vs overall)	Cum. Pop %	Cum. Pos %
8.000	3.8	34.4	5.24	3.8	19.9
7.000	2.0	32.6	4.97	5.8	29.8
4.000	8.8	20.3	3.09	14.6	57.1
6.000	4.3	11.3	1.72	18.9	64.5
2.000	23.5	7.1	1.08	42.4	89.9
1.000	4.4	5.5	0.84	46.8	93.6
5.000	5.5	3.7	0.56	52.3	96.7
9.000	6.8	1.6	0.24	59.1	98.3
3.000	15.8	0.7	0.11	74.9	100.0
0.000	25.1	0.0	0.00	100.0	100.0

Overall positive rate (Test) ≈ 6.56% (weighted). Lift = segment pos% / overall pos%.

Segment Profile Summary (TrainVal)

segment	Share % (TrainVal)	Age (mean)	Weeks worked (mean)	Log wage/ hour (mean)	Female %	Bachelor %	Private worker %	Pos rate % (TrainVal)
0.000	25.4	9.500	0.520	0.000	48.9	0.6	1.9	0.0
2.000	23.4	35.960	45.860	0.000	47.2	15.1	80.1	7.0

3.000	16.0	59.790	1.020	0.000	68.2	6.1	3.0	0.7
4.000	8.6	50.080	32.670	0.000	53.3	27.0	43.9	21.4
9.000	6.8	25.690	21.320	0.000	52.1	8.8	42.5	1.9
5.000	5.6	35.990	44.840	6.720	51.9	9.5	86.8	4.2
1.000	4.3	43.810	44.630	0.000	64.7	9.3	76.9	4.7
6.000	4.2	43.510	45.050	0.000	37.6	16.0	0.0	11.4
8.000	3.8	48.340	39.310	0.450	26.7	22.3	53.2	32.8
7.000	2.1	43.870	41.380	0.430	27.7	21.5	58.1	29.7

Conclusions:

We can combine segment size & stability table with Targeting Plan table and Segment Profile Summary table to make the tiering plan.

Tier A – High return, limited volume

Segments 8 (3.8% @ 34.4% pos; Lift 5.24) and 7 (2.0% @ 32.6%; Lift 4.97)

- Why: Extremely high incidence, small but potent audiences; by 5.8% pop we already capture ~30% of all positives.
- Action: Prioritize premium/CPA-tolerant campaigns; use as lookalike seeds. Apply frequency caps to avoid saturation.

Tier B – Good return, scalable

Segments 4 (8.8% @ 20.3%; Lift 3.09) and 6 (4.3% @ 11.3%; Lift 1.72)

- Why: Solid incidence at meaningful scale; by 18.9% pop (A+B) we cover ~64.5% positives.
- Action: Make this the main performance engine. Tailor creatives/offers to mid-career & steady work patterns; run iterative A/B tests.

Tier C – Broad reach, moderate efficiency

Segments 2 (23.5% @ 7.1%; Lift 1.08) and 1 (4.4% @ 5.5%; Lift 0.84)

- Why: Large reach with near-average incidence; useful when we need volume beyond A/B.
- Action: Use additional filters or the classifier score to tighten efficiency; expand here once A/B saturate.

Tier D – Low priority / learning

Segments 5 (5.5% @ 3.7%; Lift 0.56) and 9 (6.8% @ 1.6%; Lift 0.24)

- Why: Sub-average incidence.
- Action: Keep for learning budgets or when inventory is abundant; otherwise deprioritize.

Exclude / Control

Segments 0 (25.1% @ 0.0%) and 3 (15.8% @ 0.7%)

- Why: Essentially no positives; Seg 0 is predominantly under-18 (exclude for compliance), Seg 3 skews older/retired.
- Action: Exclude from targeting; use Seg 3 as a control group to measure spillover/uplift.

How the tiering aligns with profiles (for sharper messaging)

- Tier A personas:

- Seg 8: Older working age (mean age ≈ 48), steady weeks worked, capital-income flags present → wealth/premium value props.
- Seg 7: Working age (≈ 44), steady weeks, higher bachelor share → "upgrade" messaging (career/earnings acceleration).

- Tier B personas:
  - Seg 4: Mid-career, part-year work patterns, higher education → "career step-up," retirement catch-up, refinancing themes.
  - Seg 6: Steady weeks; self-employed signal is strong → tax-time financing, business banking lite, stability messaging.
- Tier C personas:
  - Seg 2: Large, mainstream working cohort (steady weeks, private workers) → broad, value-oriented offers; use score to prioritize
  - Seg 1: Mid-age steady workers; keep tighter thresholds to maintain efficiency.

Budget, channel & control plan

- Budget split: A (constrained by volume) 25–35%, B 40–50%, C 15–25%, D minimal/learning. Reallocate weekly by CPA/ROAS.
- Channels/creative:
  - A: Paid social + email; premium value props; higher bids; tight frequency caps.
  - B: Search + social for scale; persona-specific landing pages; constant A/B on offers.
  - C: Cheaper inventory (open web/programmatic/partner email); apply score thresholds.
  - D: Owned/earned channels or experimentation only.
- Compliance: Hard-exclude Seg 0 (under-18). Keep documentation of suppression rules for audits.

Governance & monitoring

- Drift alerts: Trigger retraining if any segment's share shifts by > 1.0 pp or > 20% relative for two consecutive runs.
- KPI dashboard (by tier & segment): CPA/ROAS, conversion rate, share of positives captured, frequency-adjusted reach.
- Two-stage activation: (1) Target tiers A/B to set audience, (2) rank within tiers by classifier score to maximize ROI.
- Experimentation: Run uplift tests in B/C; vary offers and friction; use Tier-A responders to train high-quality lookalikes.

The segmentation is size-stable and economically ordered. Start with A for yield, B for scale, C for reach when needed, deprioritize D, and exclude 0/3 to save budget and stay compliant.

8. Reference

- L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems (NeurIPS 2018)
- Anderberg, M.R., Cluster Analysis for Applications (Academic, New York, 1973)
- Hartigan, J.A. and M.A. Wong, A k-means clustering algorithm, Applied Statistics, 28 (1979) 100–108
- HDBSCAN library documentation