Retail Income Targeting and Segmentation Technical Report

Data Science Team

February 20, 2026

1 Executive Summary

This project supports two related campaign workflows: (i) ranking households by likelihood of higher income ($> 50K$) for targeting, and (ii) producing broad, interpretable segments for planning and messaging.

On a single stratified hold-out split (`random_state=2026`), XGBoost achieved ROC-AUC 0.9540 (weighted 0.9540), F1 0.5820, and Precision@Top20% 0.2844 (weighted 0.2921). Logistic regression remained competitive (ROC-AUC 0.9435, Precision@Top20% 0.2768), so the incremental gain from XGBoost is real but not a step-change.

For segmentation, features were standardized/encoded, reduced with PCA to 92.17% retained variance, and clustered with KMeans. Under the operational constraint $K \in \{3, 4, 5\}$, $K = 3$ was selected based on elbow and interpretability. Silhouette values are modest (e.g., 0.227 at $K = 3$, 0.233 at $K = 5$), so the clusters should be used as planning buckets rather than as sharply separated behavioral archetypes.

2 Business Objective

The business question is how to concentrate outreach where economic return is likely higher. In this dataset, income above $50K$ is treated as a proxy for purchasing power, so the classifier is used as a *ranking* tool: you choose a targeting depth (e.g., top 10%, top 20%) based on budget and channel constraints.

Because the source is a weighted survey sample, `weight` is part of the modeling design rather than optional metadata. Training with weights steers optimization toward population-level structure instead of sample composition. In evaluation, we report both unweighted and weighted metrics so the results are interpretable under both views.

3 Data Understanding and Exploration

The dataset contains 199,523 rows and 42 columns: 40 predictors, one sample-weight column, and one binary income label. The positive class is rare: unweighted prevalence is 6.21% and weighted prevalence is 6.41%. This imbalance is why Precision@Top20% is emphasized; for campaign operations, concentration within the top-ranked slice is usually more actionable than global accuracy.

The strongest numeric associations with the label are work intensity and investment-related fields (for example, weeks worked in year and capital gains). These are correlation signals, not causal effects. They are still useful for ranking, but interpretation should stay operational.

4 Classification Methodology

All three required models were evaluated on the same 80/20 stratified split (`random_state=2026`): L2-regularized logistic regression, random forest, and XGBoost. Each model was implemented as a leakage-safe `Pipeline` with a `ColumnTransformer`:

- Numeric features: median imputation; standardization for logistic regression only.
- Categorical features: most-frequent imputation; one-hot encoding with unknown handling.

Training incorporates survey weights through weighted empirical risk minimization:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} w_i \, \ell(y_i, f_\theta(x_i)) \,,$$

where w_i is the survey weight for record i.

Precision@Top$k$ is computed by sorting validation records by predicted probability and taking the top fraction ($k = 0.20$):

$$\text{Precision@Top}k = \frac{1}{|S_k|} \sum_{i \in S_k} y_i,$$

where S_k is the top-ranked subset.

5 Classification Results

Table 1: Validation metrics on shared hold-out data (`metrics.csv`).

Model	ROC-AUC	ROC-AUC (w)	P@20%	P@20% (w)	Brier	Brier (w)
XGBoost	0.9540	0.9540	0.2844	0.2921	0.0320	0.0328
Random Forest	0.9449	0.9451	0.2783	0.2858	0.0348	0.0357
Logistic Regression	0.9435	0.9432	0.2768	0.2847	0.0362	0.0372

5.1 Weighted vs. Unweighted Interpretation

We report both unweighted and weighted metrics. Unweighted metrics reflect the hold-out sample behavior directly. Weighted metrics (e.g., weighted ROC-AUC, weighted Precision@Top20%) reweight the validation rows toward population representation. For campaign decisions, the key quantity is precision/lift at a chosen targeting depth. Budgeting uses weighted Precision@Top$k$ (population-facing), while model selection uses unweighted metrics for hold-out comparability.

The validation base positive rates are 6.20% unweighted and 6.47% weighted. At 20% depth, XGBoost reaches 28.44% precision unweighted and 29.21% weighted, equivalent to lift of 4.58× and 4.52×. This is the main operational takeaway: the top quintile is several times denser in positives than the baseline population.

Table 2: XGBoost gains profile by targeting depth (`gains_table.csv`).

Depth	Precision	Lift	Precision (w)	Lift (w)
5%	0.6844	11.03	0.6946	10.74
10%	0.4681	7.54	0.4796	7.42
20%	0.2844	4.58	0.2921	4.52
30%	0.2000	3.22	0.2060	3.19

5.2 Gains at Campaign Depths

Campaign decisions are driven by precision/lift at fixed targeting depth, not by a global threshold. Table 2 shows the XGBoost gains profile from `gains_table.csv`.

The curve decays in the expected way: performance is strongest at very shallow depth and tapers as the campaign expands. Operationally, this gives a simple knob for balancing volume and contact efficiency.

5.3 Calibration Check

Because the score may be used in economic reasoning, we include a calibration diagnostic. For XGBoost, the Brier score is 0.0320 (weighted 0.0328). The reliability plot in `calibration_curve.png` does not suggest a severe failure in the score ranges used for targeting, although it is not perfectly calibrated.

5.4 Uncertainty Across Deterministic Splits

To assess stability, we ran five deterministic splits (seeds 2026–2030), summarized in `metrics_summary.csv`. For XGBoost, ROC-AUC is 0.9533 ± 0.0008, Precision@Top20% is 0.2836 ± 0.0010, and lift@20 is 4.571 ± 0.016. Variance is low, so ranking quality is fairly stable under modest split perturbations.

6 Segmentation Methodology

Segmentation intentionally excluded both label and weight from the clustering feature matrix. Numeric variables were standardized after imputation; categoricals were one-hot encoded. PCA was then applied, retaining 92.17% of total variance.

KMeans was evaluated for $K = 2$ through 8 using inertia and silhouette, and the final selection was restricted to $K = 3$ to 5. Within that range, silhouette is 0.2270 at $K = 3$, 0.2239 at $K = 4$, and 0.2325 at $K = 5$. The $K = 5$ silhouette advantage over $K = 3$ is only 0.0055, so $K = 3$ was retained for interpretability and cleaner operational use.

These silhouette levels are low. They imply weak geometric separation and suggest the covariates may not naturally organize into compact, well-separated groups. KMeans also imposes roughly spherical cluster structure in transformed space, which may not be the best match for mixed socioeconomic tabular data.

6.1 Reproducible Cluster Assignment

The segmentation transformation path is serializable. `segmentation_preprocessor.pkl`, `segmentation_pca.pkl`, and `segmentation_kmeans.pkl` are saved after fitting, with run settings in `segmentation_metadata.json`. This enables deterministic assignment of new records to the same cluster space without refitting.

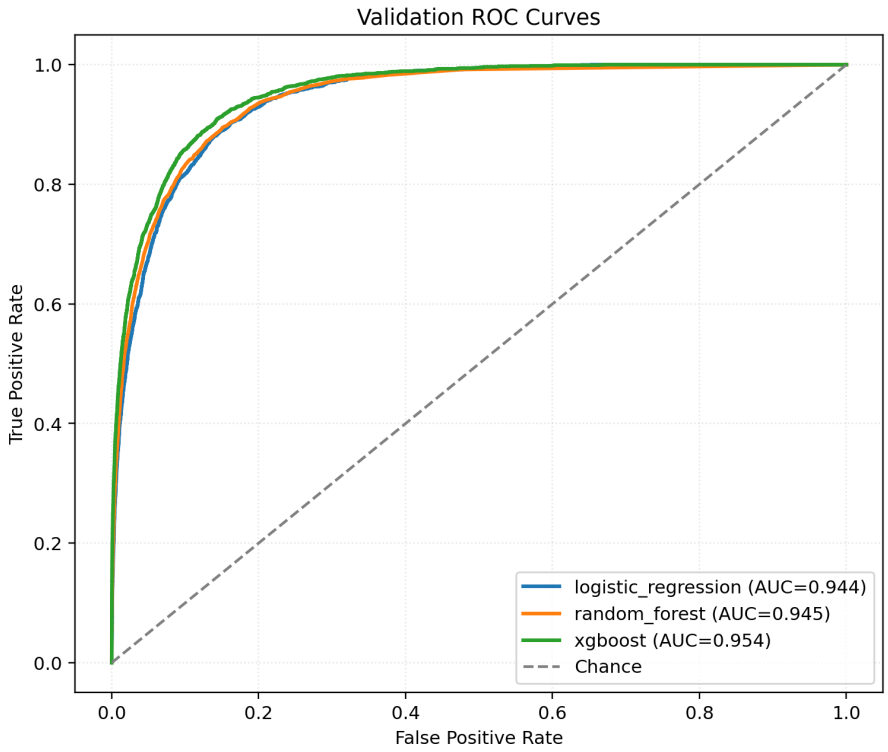

Figure 1: Validation ROC curves for the three classifiers (`classification_roc_curves.png`).

6.2 Stability Check

Cluster stability was checked on a fixed 20k PCA sample using seeds 2026–2030 and pairwise ARI (`segmentation_stability.json`). The observed ARI range is 1.00 to 1.00 with mean 1.00, indicating numerical stability under KMeans initialization changes on that fixed sample. This check does not test stability under resampling or alternative feature specifications.

7 Segmentation Results

Table 3: Weighted cluster profile summary (`cluster_summary.csv`).

Cluster	Records	Weighted Population	Weighted Income Rate	Mean Age
0	99,221	176.22M	0.1175	38.15
1	52,791	88.30M	< 0.001	7.83
2	47,511	82.72M	0.0185	55.09

The profiles are internally coherent. Cluster 0 is working-age with the strongest income rate and higher labor attachment (e.g., mean weeks worked 44.9, mean capital gains \$761). Cluster 1 is a youth-heavy segment with minimal work intensity and essentially zero high-income incidence. Cluster 2 is older with low work intensity and low income incidence. The PCA projection still shows overlap among groups, so these should be used as operational segments rather than as hard population boundaries.

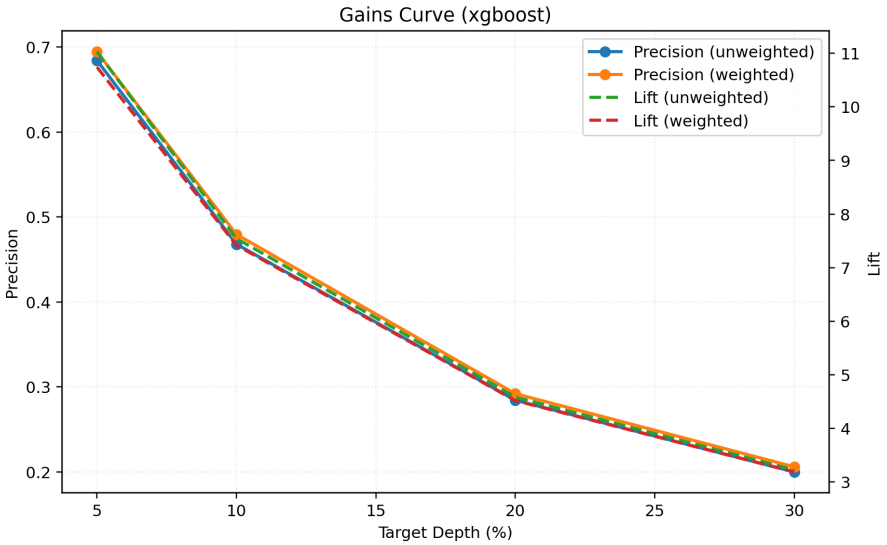

Figure 2: XGBoost gains curve (precision and lift by depth) (`gains_curve.png`).

Table 4: Messaging hypothesis by segment (`segment_messaging.csv`).

Cluster	Short profile	Suggested message	Channels
0	Mean age 38.1, weighted income rate 11.8%, top occupation: administrative support.	Lead with premium bundles and convenience-focused value.	Email, search, web
1	Mean age 7.8, weighted income rate $<$ 0.1%, top occupation: not in universe.	Lead with essentials and family-value framing; avoid premium positioning.	Social, mobile, email
2	Mean age 55.1, weighted income rate 1.9%, top occupation: not in universe.	Emphasize reliability and value; prioritize low-friction retention offers.	Email, direct mail, call center

8 Business Recommendations

Use the classifier score as the primary allocation axis, then apply segment overlays for creative and channel strategy. Cluster 0 is the strongest candidate for premium and upsell campaigns. Cluster 2 is better suited to lower-cost retention or value-oriented messaging. Cluster 1 should typically be excluded from income-driven offers unless the product strategy explicitly targets household-level decision makers.

This combination balances accuracy and operational simplicity: one score for ranking and one segment label for messaging context.

9 Production Deployment Considerations

Thresholding should be cost-sensitive rather than fixed at 0.5. A practical rule is to contact an account when expected value is positive:

$$\hat{p}(x) \cdot V_{\text{response}} - C_{\text{contact}} > 0.$$

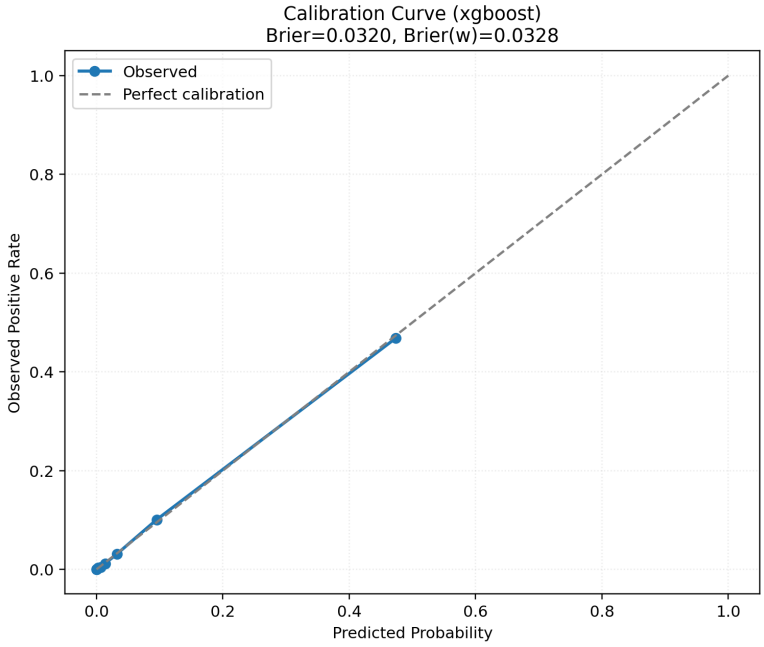

Figure 3: XGBoost calibration curve (`calibration_curve.png`).

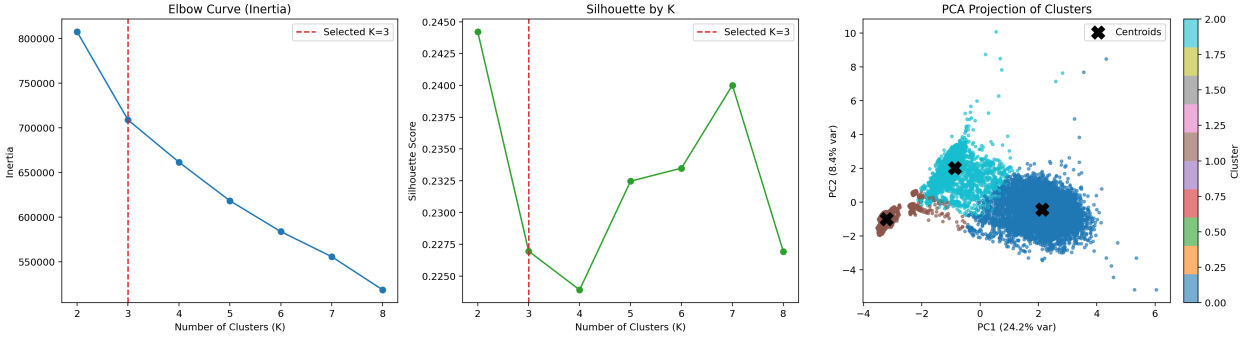

Figure 4: KMeans diagnostics and 2D PCA projection of clusters (`segmentation_plots.png`).

Here $\hat{p}(x)$ is the predicted proxy (income-above-threshold probability), V_{response} is expected contribution value, and C_{contact} is channel cost.

Operational deployment also depends on feature availability at scoring time. Some high-signal variables in this dataset (for example, annual work and income components) may be delayed, estimated, or absent in real-time systems, which can reduce realized performance unless fallback features are engineered. Monitoring should include calibration drift, segment share drift, and subgroup-level performance checks. Drift risk is elevated because the training data reflects 1994–95 population structure.

10 Limitations

External validity is limited by the 1994–95 data origin; labor-market structure and demographic distributions have changed materially since then. The model is predictive, not causal, so it should

not be used to infer policy effects. Bias risk is non-trivial because protected-class proxies can appear in socioeconomic attributes; fairness checks are necessary before operational rollout.

Clustering adds another limitation: KMeans assumes roughly spherical structure in transformed space. Given the observed overlap, segment boundaries are partly algorithmic convenience. A stronger follow-up would include stability checks under resampling and sensitivity to alternative feature subsets.

11 Future Improvements

Immediate next steps are: (i) cost-optimized depth selection with campaign economics, (ii) calibration monitoring and decision-curve analysis, and (iii) uplift modeling to prioritize incremental impact rather than likelihood alone. On segmentation, periodic retraining and stability tests under resampling should be added before relying on segments for long-term budget planning.