Take Home Project

Mohsen Taheri

1 Income Classification

1.1 Executive summary

I trained and validated a production ready classifier to predict whether an individual's income exceeds \$50,000 using 40 demographic and employment variables with survey weights. The model of choice is **XGBoost (gradient–boosted decision trees)** optimized for **weighted Average Precision (AUPRC)** to reflect class imbalance and the marketing objective of high–quality positive identifications. I performed systematic feature screening, removed a small set of high–cardinality/low–signal fields, and tuned both model hyperparameters and the **decision threshold** to meet budget–constrained outreach goals.

Balanced operating point (current setting):

Weighted Accuracy	**0.9520**
Weighted Precision	**0.6255**
Weighted Recall	**0.6211**
Weighted F1	**0.6233**
ROC–AUC	**0.9555**

I also provide a thresholding routine that maps a marketing budget to the corresponding probability cutoff, enabling precision–heavy or recall–heavy strategies.

1.2 Business objective

Identify two groups for marketing: people with income $\leq$\$50k and $>$\$50k. For targeting premium offers, high precision in the $>$\$50k class is desirable when budgets are tight; for broad awareness, higher recall may be preferred.

1.3 Data understanding

1.3.1 Structure

40 demographic and employment variables (mix of numeric and categorical) + weight + label.

1.3.2 Class imbalance

The $>$\$50k class is the minority. All training and evaluation use the provided weights.

1.4 Pre–processing

1.4.1 Encoding and cardinality

Standard categorical variables were one–hot encoded where category counts were manageable. A subset of fields exhibited *very high cardinality* and/or noisy semantics. After quantitative screening (feature gain rankings), I removed them to prevent dimensionality explosion and reduce overfitting.

1.4.2 Feature screening and removals

Using gain-based importance from an initial boosted-tree fit, I flagged fields that consistently ranked at the bottom and showed either very high cardinality or unstable/noisy behavior. To avoid an oversized sparse feature matrix and because these variables added little incremental signal, I removed the following:

- `country of birth father`
- `country of birth mother`
- `country of birth self`
- `migration code - change in MSA`
- `migration code - change in region`
- `migration code - move within region`
- `live in this house 1 year ago`
- `region of previous residence`
- `state of previous residence`
- `detailed household and family status`

I also used a simple earnings/workload signal—hourly wage × weeks worked—and applied $\log(1+x)$
.

1.4.3 Weights

I treat sampling weights as observation weights in both training and validation so that behavior reflects the real–world distribution.

1.5 Model choice and architecture

1.5.1 Algorithm

XGBoost binary classifier (gradient–boosted trees). Handles mixed data types and nonlinear interactions; supports observation weights; strong performance on tabular, imbalanced data; regularization and early stopping mitigate overfitting.

1.5.2 Objective/metric

Trained with `eval_metric = aucpr` (area under Precision–Recall) to align with minority–class marketing goals.

1.6 Training procedure and hyperparameter tuning

- **Split:** Stratified train/test preserving class balance and weight distribution.
- **Cross–validation:** 5–fold stratified CV with weights; score = mean weighted Average Precision.
- **Hyperparameters searched:** learning rate, max depth, min child weight, gamma (split penalty), subsample, colsample_bytree, L1/L2 regularization (`reg_alpha`/`reg_lambda`), and `scale_pos_weight` (from weighted class ratio). Conservative early stopping was used throughout.
- **Chosen configuration (balanced):** Shallow, regularized ensemble with conservative learning rate; histogram trees; early stopping for stability.

1.7 Evaluation

Balanced operating point metrics are summarized in the table in §1.1.

- **Primary metric:** Weighted Average Precision (AUPRC).

- **Secondary metrics:** ROC–AUC, weighted F1, weighted accuracy (at a selected threshold).
- **Weighted confusion matrix:** Computed under sampling weights to reflect population proportions.

1.8 Post–processing: decision threshold

Marketing actions are budget–constrained. Instead of a fixed 0.50 cutoff, I *search over thresholds* and select the one that yields a target **weighted positive prediction rate** (PPR), e.g., "contact the top 8% most likely individuals." This provides a direct, auditable link between spend and precision/recall trade–offs.

Operational use:

1. Choose a budget–driven PPR target (e.g., 5%, 8%, 15%).
2. Find the probability threshold t^* whose weighted PPR is closest to the target.
3. Report weighted precision/recall at t^* to quantify expected yield and leakage.

1.9 Business levers: dialing precision vs. recall

1.9.1 Primary lever (recommended): decision threshold

- **More precision / lower volume:** increase the threshold.
- **More recall / higher volume:** decrease the threshold.

1.9.2 Secondary levers (model hyperparameters)

- **Favor precision (budget–tight, premium offers):** increase `reg_alpha`/`reg_lambda`; lower `max_depth`; increase `min_child_weight`; increase `gamma`; moderately lower `subsample`/`colsample_bytree`; *decrease* `scale_pos_weight` slightly; keep a lower learning rate with early stopping.
- **Favor recall (broad awareness):** ease regularization (lower `reg_alpha`/`reg_lambda`); raise `max_depth`; lower `min_child_weight`; lower `gamma`; raise `subsample`/`colsample_bytree`; *increase* `scale_pos_weight`; pair with a lower threshold.

2 Clustering

2.1 Preprocessing

I selected marketing-relevant numeric and categorical fields, scaled numerics, and one-hot encoded categoricals so everything is on a comparable scale while staying sparse and robust to unseen categories.

2.2 Embedding

I compressed the high-dimensional sparse matrix into 40 latent components with Truncated SVD so distances are meaningful and clustering is fast and stable.

2.3 Algorithm

I clustered the SVD embedding with K-Means and used survey weights during fitting so segments reflect the population mix.

2.4 Choosing K (Number of Clusters)

The silhouette curve rises from K=4 and forms a broad plateau between K=7 and K=10 with only a small gain at K=10. I chose K=7 as the simplest point on the plateau, avoiding over-fragmentation while capturing nearly all the separation, which yields more interpretable, larger, and actionable segments.

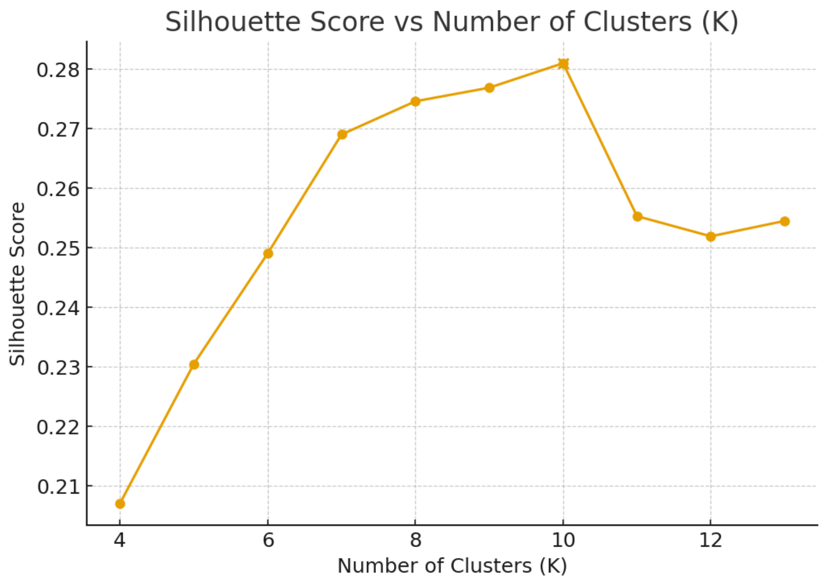

Figure 1: Choosing number of clusters.

2.5 Segments & Actions

I profiled each cluster with survey weights to reflect the population: computing weighted means for numeric features and weighted proportions for categorical levels. I then compared clusters to overall baselines, turning differences into numeric z-scores and categorical lifts to highlight what's most distinctive per cluster. Next, I extracted each cluster's top signals (largest |z| and highest lifts), added income-rate enrichment (> $50k) for business context, and summarized them into short, human-readable labels.

2.5.1 Working Households, Steady Schedules — 16.7%

Signals: Working, many weeks worked, little dividend income; many single/heads of household.
Use: Value packs, fuel rewards.

2.5.2 Children — 44.8%

Signals: Not working; "Children/Nonfiler/Child in household" patterns.
Use: Treat as **exclusion** for adult targeting, or seasonal messaging only.

2.5.3 Married Full Timers, Some Self Employment — 21.3%

Signals: Full year workers; joint filers; more self employed.
Use: Family bundles, bulk club sizes, pickup.

2.5.4 Working Investors with Losses (Highly Educated) — 2.0%

Signals: Capital losses; professional/doctoral education.
Use: Premium electronics/home office, financial services, travel/luggage.

2.5.5 Skilled Trades — 5.6%

Signals: High wage$\times$weeks; strong union.
Use: Pro grade consumables, tools, hot meals, extended hours.

2.5.6 Older Dividend Households — 9.5%

Signals: Older; dividend income; advanced degrees.
Use: Pharmacy & wellness, premium pantry, delivery subscriptions.

2.5.7 Affluent Investors — 0.2%

Signals: Capital gains + dividends; self employed incorporated; professional/doctoral.
Use: VIP microsegment: premium labels, delivery subscriptions, gift cards.

3 References

- T. Chen and C. Guestrin (2016). "XGBoost: A Scalable Tree Boosting System."
- Scikit–learn documentation: metrics for imbalanced classification (AUPRC).
- XGBoost documentation: sample weights, regularization, class imbalance handling.