Census Income Modeling (1994–1995 CPS): Income Classification & Customer Segmentation

Ishraq Khan

12/26/25

Happy holidays to whoever is taking the time to review this take home project submission of mine. I truly appreciate your time and consideration and hope that the time and consideration I put into the project is well-received!

1 Executive Summary

We build two models from U.S. Census Current Population Survey data (1994–1995):

1. **Income propensity model (binary classifier)** that estimates the likelihood an individual earns $> \$50K$.
2. **Customer segmentation model (clustering)** that groups individuals into a small set of interpretable segments to tailor marketing efforts.

Key Takeaways

- **Business value comes from more than just prediction accuracy.** We explicitly treat the classifier as a ranking tool: marketing can pick a probability threshold (or target the top $X\%$) to match budget and ROI constraints.
- **Interpretability matters.** We pair the best-performing predictive model with a regularized logistic regression "explainability" model to highlight predictive signals and check for biases.
- **Segments enable action beyond a single score.** Segments provide human-friendly personas that can drive differentiated creative, offer structures, and channel strategies.

Practical Considerations

- Use **propensity scores** to prioritize outreach (premium products, loyalty upgrades, cross-sell).
- Use **segments** to tailor messaging and promotions (e.g., premium convenience vs. budget practicality), then measure lift via A/B tests.

2 Understanding the Data

Data source and structure

The dataset is derived from the 1994–1995 U.S. Census Current Population Survey, provided in a processed form commonly used for income modeling. Each row corresponds to an individual with ~ 40 demographic and employment-related attributes and a binary income label.

Target definition

The target is whether an individual earns $> \$50K$ vs. $\leq \$50K$. In the raw file, labels are string-coded and mapped to:

$$y = \begin{cases} 1, & \text{income } > \$50K \\ 0, & \text{income } \leq \$50K \end{cases}$$

Role of sample weights

The data includes a "sampling weight" column representing how many people in the population a record stands for. We treat weights as important for exploratory data analysis, but optional for predictive modeling depending whether we want to optimize for the sample or the population. In our notebook pipeline, weights are primarily used to interpret prevalence and to build weighted segments; predictive training focuses on robust out-of-sample ranking under class imbalance.

Feature types and data quality notes

- **Continuous / count features** (e.g., age, wage/hour, weeks worked/year, capital gains/losses/dividends) are heavily skewed with spike-at-zero patterns.
- **Categorical features** include education, class of worker, marital status, citizenship, and location proxies.
- **Special missing-like categories** appear (e.g., "Not in universe", "?"), which are not necessarily missing-at-random; as such, we treat them as meaningful categories unless clearly erroneous.
- **Class imbalance** is present: the majority class is $\leq \$50K$ (around 94%).

Distributional insights that affect modeling

Key modeling implications from EDA:

- Monetary variables are long-tailed: linear models benefit from log transforms; tree/boosting methods handle skew but still benefit from stabilizing extreme values.
- Some high-cardinality coded variables (e.g., detailed industry/occupation recodes) are sparse and/or redundant relative to broader groupings (major industry/occupation code); using them can add dimensionality and hurt interpretability.
- The dataset includes minors; for an income model aimed at marketing to adults, we filter to working-age individuals.

3 Data Preprocessing & Feature Engineering

Filtering and basic cleaning

- **Population filter:** we restrict to individuals with `age` ≥ 18 to align the task with an adult income propensity use case.
- **Label encoding:** string labels are mapped to $\{0, 1\}$.
- **Weights:** retained for interpretation and segmentation; excluded from the list of features for training classifier.

Handling missing values and "Not in universe"

Rather than imputation, we rely on the dataset's explicit "missing-like" categories. For categorical features, these values remain as categories and are handled by one-hot encoding. For numeric features, our pipeline yields no true NaNs after filtering. The only true NaNs are found in the 'hispanic origin' column and due to the small number (800), we drop these rows. The data is captured through other features such as 'citizenship' and 'country of birth self'.

Feature engineering decisions (and why they matter)

Our transformations are meant to improve signal-to-noise and interpretability (important for marketing contexts):

- **Monetary features:** create both an indicator and a stabilized log transform:

  $$\texttt{has\_x} = 1[x > 0], \quad \widetilde{x} = \log(1 + x)$$

  This separates "presence of any gains/losses" from "how large," which often maps to distinct real-world behaviors.

- **Migration features:** collapse multiple migration fields into a smaller set of binary indicators (e.g., non-mover vs. mover; within-state vs. cross-state).

- **Education simplification:** collapse pre-HS categories into broader buckets.

- **Sensitive attributes:** sex and race were converted into binary indicators for analysis and monitoring.

Encoding and scaling

Our modeling pipeline cleanly separates numeric and categorical processing:

- **Numeric:** standardization via `StandardScaler`.

- **Categorical:** one-hot encoding with `handle_unknown=ignore` to ensure robustness to unseen categories at scoring time.

4 Modeling Approach (Prediction Task)

Objective

Our task is binary classification where the practical marketing goal would be ranking individuals by their likelihood of making $> \$50K$.

Models considered

We compare multiple models in our analysis to develop a comprehensive picture of model performance, strengths, and weaknesses (especially considering our goal of addressing class imbalance):

- **Dummy baseline** (most-frequent class) to anchor expectations.

- **Logistic regression** with class weighting (strong baseline; interpretable).

- **Decision tree / random forest** (nonlinearities; partial interpretability).

- **Gradient boosting** and **XGBoost** (typically strong for tabular data; industry standard).

Final model selection logic

Because the data is imbalanced, we selected the final production candidate based on **cross-validated PR-AUC (average precision)**, which better reflects performance on the positive class (high-income) than accuracy.

- **Validation approach:** stratified train/validation split with a stratified k-fold CV loop on the training partition.
- **Metrics tracked:** PR-AUC, ROC-AUC, balanced accuracy, F1, and accuracy.

Class imbalance handling

We address imbalance via:

- **Metric choice:** optimize for PR-AUC.
- **Class weights:** enabled for models that support it (e.g., logistic regression, trees/forests).
- **Threshold tuning:** treat 0.5 as a starting point, then we can tune threshold to marketing goals (precision vs. recall tradeoff).

5 Model Evaluation & Results

Cross-validated model comparison

Table 1 summarizes CV performance.

Table 1: Cross-validated performance summary (training folds). Selection based on PR-AUC.

Model	PR-AUC	ROC-AUC	Balanced Acc.	F1	Acc.
Dummy (most frequent)	0.0863	0.5000	0.5000	0.0000	0.9137
Logistic Regression (balanced)	0.6071	0.9210	0.8448	0.4742	0.8364
Decision Tree (balanced)	0.2517	0.7021	0.7017	0.4475	0.9021
Random Forest (balanced subsample)	0.6002	0.9148	0.6856	0.4977	0.9326
Gradient Boosting	0.6428	0.9241	0.6981	0.5274	0.9366
XGBoost	0.6739	0.9327	0.7208	0.5665	0.9398

Validation results (held-out split)

We fit the best-by-CV model on the training split and evaluated on the validation split:

- **Validation ROC-AUC: 0.9352**
- **Validation PR-AUC: 0.6861**

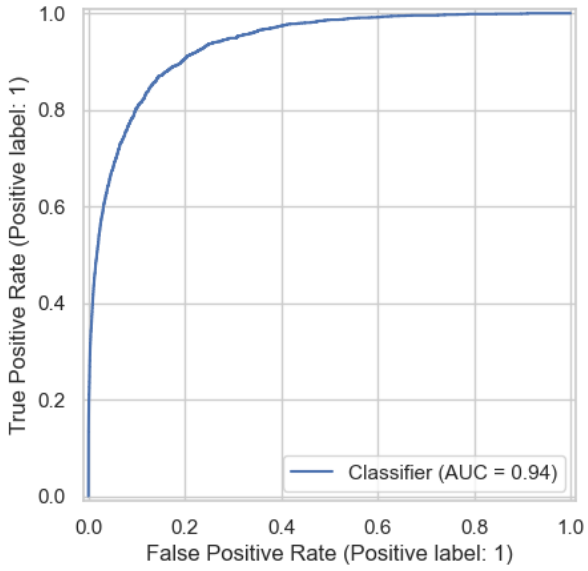

Figure 1: ROC Curve

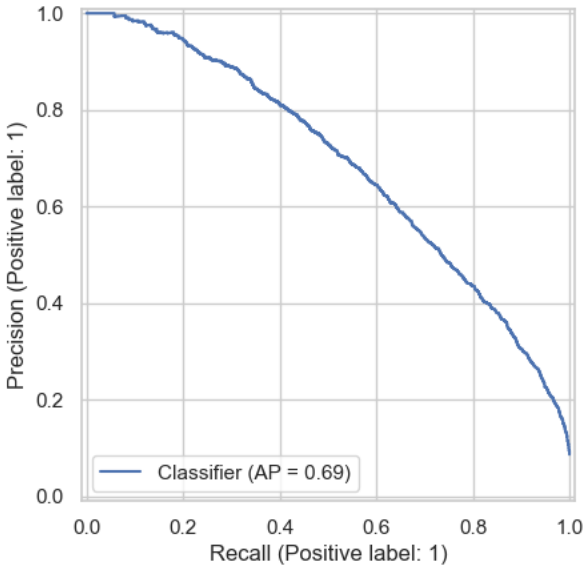

Figure 2: Precision–Recall Curve

Figure 3: Model performance on the validation set. The ROC curve summarizes overall ranking performance, while the precision–recall curve emphasizes performance on the minority (high-income) class.

Business interpretation of metrics

For a marketing use case, PR-AUC and precision/recall at operating thresholds matter most:

- **High precision** reduces wasted spend (fewer low-income individuals incorrectly targeted for premium offers).
- **High recall** captures more high-income customers (useful when the campaign goal is reach/coverage).

5.1 Weaknesses / cautions

- The model is predictive, not causal: it captures correlations in 1994–1995 data that may not hold today.
- There are potential sensitivity/fairness issues if demographic features are used directly for targeting.

6 Explainability & Insights

Approach

In addition to the best-performing predictive model, we fit a regularized logistic regression (elastic net) as a proof-of-concept. This provides directionality (is there a +/- association with high income), sparsity and stability, and a way to check against bias.

What features matter most (non-causal)

We rank features by absolute coefficient magnitude in the explainability model.

Table 2: Top drivers from elastic-net logistic regression (validation-fit coefficients). Positive coefficients increase the likelihood of income $> \$50K$, while negative coefficients decrease it.

Feature	Coefficient
Log capital gains	5.0919
Has capital gains	-4.8792
Log capital losses	1.6867
Has capital losses	-1.5007
Graduate degree	1.4282
Tax filer status: nonfiler	-1.3577
Occupation: private household services	-1.2540
Employment status: children or Armed Forces	-1.1702
Class of worker: without pay	-1.1643
Occupation: Armed Forces	1.1209

Interpretation

- Variables related to capital income dominate the model, reflecting the strong association between investment-related income and higher total earnings.
- Education also plays a meaningful role with graduate degrees positively associated with income above \$50K.
- Negative coefficients largely correspond to non-working or marginal labor-force attachment categories (non-filers, unpaid workers) which makes sense.

7 Segmentation Analysis

A single propensity score is useful for prioritization, but segments can provide distinct personas to tailor marketing messages and offers:

Method and preprocessing

We use K-means clustering on the same preprocessed feature space from our propensity model:

- Continuous features are standardized.
- Binary indicators pass through.
- Categorical variables are one-hot encoded into a sparse representation.
- We fit both unweighted and weighted K-means to ensure segment sizes reflect population prevalence.

Choosing the number of segments

We evaluated K over a range (2 to 30) using inertia (elbow method) for both unweighted and weighted fits, then selected **K=7** for how it gives us enough segments to act on that are meaningfully different.

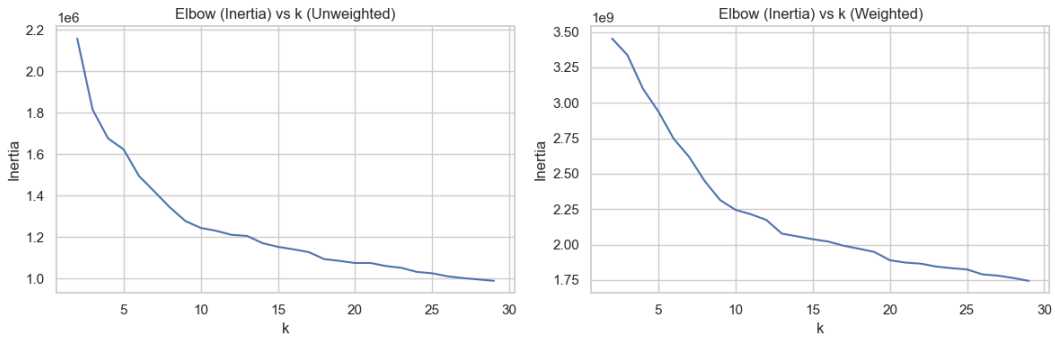

Figure 4: Elbow plot for K-means (weighted vs. unweighted inertia).

Segment profiling

For each segment, we summarize:

- weighted segment size (population share),
- average age and key employment/education markers,
- observed rate of $> \$50K$ (to align segments with propensity),
- dominant categories (education, occupation/industry groupings, marital status).

Table 3: Segment summary (weighted). Segments are derived via K-means clustering on standardized demographic and employment features. Population shares reflect survey weights.

Segment	Pop. Share	Avg Age	$> \$50K$ Rate
Segment A	23.9%	38.9	8.9%
Segment B	23.7%	38.7	7.7%
Segment C	17.2%	70.1	0.9%
Segment D	10.8%	31.6	0.2%
Segment E	10.7%	33.1	6.2%
Segment F	9.7%	51.4	26.4%
Segment G	5.2%	48.5	33.5%

How marketing teams can use segments

Segments can drive differentiated playbooks:

- **Premium/affluent segments (high $> \$50K$ rate):** prioritize premium assortments, convenience, loyalty tiers, and upsell bundles.
- **Value-driven segments (lower $> \$50K$ rate):** emphasize promotions, essentials, and high-utility bundles; use cost-efficient channels.
- **Younger segments:** mobile-first and social channels; installment/payment-plan framing where appropriate.
- **Older segments:** reliability, service, and trust messaging; email/direct mail and retention-focused cadence.

8 Business Recommendations

Using the classifier

- **Adopt a "ranking + threshold" workflow:** score the customer list, then choose a cutoff aligned to budget and acceptable waste.
- **Run controlled experiments:** A/B test targeting rules (e.g., top 5% vs. top 10%) and optimize for incremental lift and margin, not raw response rate.
- **Calibrate if needed:** if decisions depend on well-calibrated probabilities (not just ranking), add Platt scaling or isotonic calibration.

Using segmentation

- **Operationalize segment playbooks:** map each segment to a recommended channel mix, creative angle, and offer structure.
- **Measure segment lift:** evaluate performance by segment to refine the persona definitions and reduce spend in low-fit groups.

Risks, ethics, and governance

- **Fairness:** demographic attributes can introduce disparate impact. We recommend policy review and, at minimum, slice-based monitoring (performance by group).
- **Use limitation:** this model is appropriate for *marketing prioritization*. It should not be repurposed for credit eligibility or other regulated decisions without a substantially stronger compliance and validation framework.

9 Limitations & Future Work

Limitations

- **Dataset Time** 1994–1995 Census patterns won't generalize to today's economy, labor market (quite tough right now if I do say so myself), or demographics.
- **Label simplicity:** a binary cutoff at $50K hides scale ($51K vs. $250K). We should explore a multi-class or regression approach if specific income levels are available.
- **Causality:** observed relationships are correlational; avoid policy conclusions about "drivers" of income.
- **Bias and sensitivity:** using demographic features for targeting may be unethical even if predictive.

Future work

- Apply cost-sensitive thresholding tied to campaign costs.
- Compare additional strong learners (CatBoost/LightGBM) and evaluate stability across time splits if newer data becomes available.

- Expand segmentation validation: silhouette scores and business validation via campaign lift by segment.
- Incorporate additional data sources (purchase history, engagement signals, etc) to build segments tied more directly to retail behaviors.

References

[1] Kaggle, *https://www.kaggle.com/code/impratiksingh/unsupervised-learning.*

[2] *https://www.kaggle.com/code/jacopoferretti/how-to-target-money-donors-in-electoral-campaigns*

[3] COMS4776: Applied Machine Learning notes *(Columbia University course)*