Project Report

Census-Based Income Prediction
and Customer Segmentation

Prepared for the Retail Business Client

Anukriti Singh

February 2026

Classification and segmentation models for marketing
using U.S. Census data .

Contents

1 Executive Summary

This report summarizes the development, validation, and business use of two models built on your census dataset:

1. **Income classification model** — Predicts whether an individual is likely to earn more than $50,000 per year, so you can prioritize higher-value prospects for premium or high-ticket campaigns.
2. **Customer segmentation model** — Groups the population into distinct segments based on employment and financial behavior, so you can tailor messaging, products, and channels to each group.

Both models use the same 40 demographic and employment variables you provided, plus the census weight so that results reflect the true distribution of the population. The classification model is ready for use in scoring new prospects; the segmentation model provides four interpretable segments with clear marketing implications. Key recommendations are in Section 8.

2 Business Context and Objectives

Your business needs to:

- **Identify high-income individuals** (income > $50K) for targeted marketing, using the 40 variables you can collect for any prospect.
- **Segment the population** for marketing so that campaigns can be tailored by segment (e.g., mass-market vs. premium vs. long-term nurture).

The dataset you supplied contains weighted census records from the 1994–1995 Current Population Surveys: 40 variables per person, a population weight, and a label indicating whether income is above or below $50K. This report describes how we explored the data, built and evaluated the models, and how you can use them operationally.

3 Data Overview and Exploration

3.1 Dataset size

Approximately 199,500 records after removing rows with missing labels.

3.2 Target (income)

About 94% of records are labeled as income $\leq$ $50K and 6% as > $50K. This **class imbalance** was explicitly handled in modeling and evaluation so that the minority (high-income) group is not ignored. Figure 1 shows the weighted distribution of the target in the population.

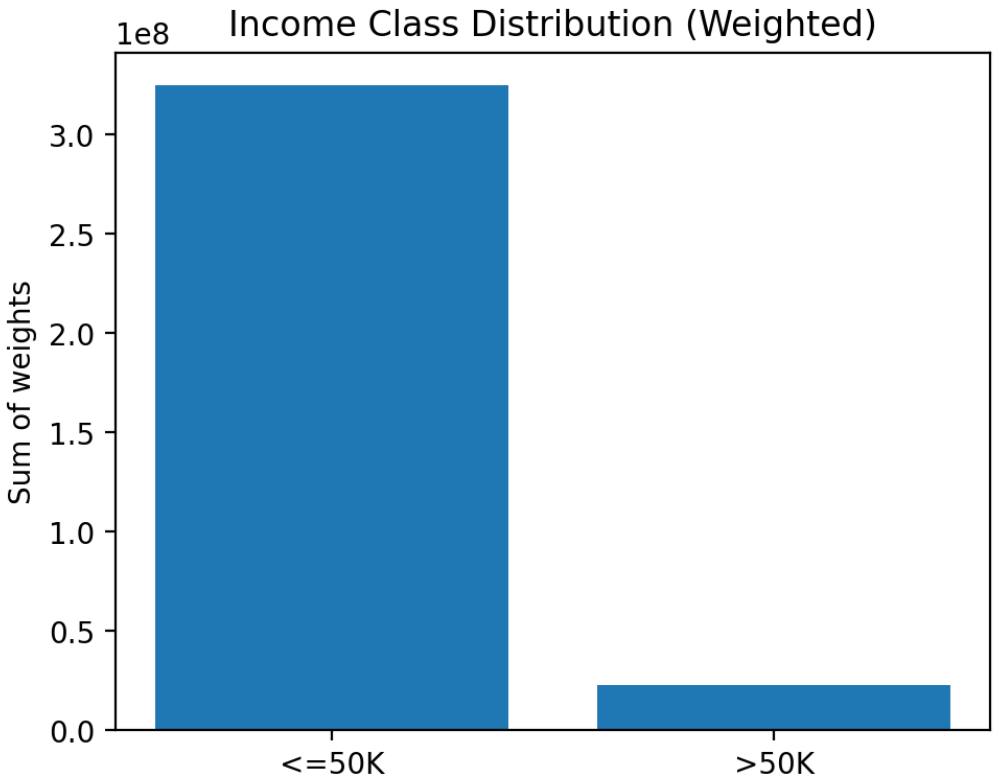

Figure 1: Weighted distribution of income class (% of population). The minority high-income group (>50K) is about 6%.

3.3 Variables

The 40 variables cover demographics (e.g., age, education, marital status, race, sex), employment (e.g., weeks worked, wage per hour, industry, occupation), and financial indicators (e.g., capital gains, capital losses, dividends). Categorical variables were one-hot encoded for classification; a subset of numeric variables was used for segmentation to keep segments interpretable.

3.4 Weight

Each record has a census weight indicating how many people in the population it represents (stratified sampling). We used these weights in training and evaluation so that model performance and segment sizes reflect the real population.

3.5 Data quality

Missing values were handled (e.g., "?" treated as missing, then imputed or excluded as appropriate). Numeric fields were checked for valid ranges and scaled where needed.

4 Preprocessing and Methodology

4.1 Classification pipeline

- Numeric features: standardized (zero mean, unit variance).
- Categorical features: one-hot encoded with unknown categories ignored at prediction time.
- Data split: 80% training, 20% validation, with stratification by income.

- Sample weights from the census weight column were passed to the models and to all reported metrics (ROC-AUC, PR-AUC) so that performance is representative of the population.

4.2 Segmentation pipeline

- A subset of numeric features was selected (e.g., age, weeks worked, wage per hour, capital gains/losses, dividends, number of employers, self-employment indicator).
- Skewed monetary variables were log-transformed and winsorized (clipping extreme values).
- Features were standardized before clustering.
- Clustering was performed without using the census weight inside the algorithm; the weight was used only after clustering to compute segment sizes and income mix per segment.

5 Classification Model: Approach, Results and Recommendation

5.1 Approach

We trained and compared three model types: Logistic Regression, Random Forest, and XGBoost. For each we used class balancing (e.g., class weights, `scale_pos_weight`) and sample weights. We evaluated using weighted ROC-AUC, weighted Precision-Recall AUC (PR-AUC), precision, recall, and confusion matrices.

5.2 Results

- **XGBoost** gave the best ranking performance (highest ROC-AUC and PR-AUC) and the best balance between finding high-income individuals and limiting false positives.
- After tuning the classification threshold (e.g., to 0.55 instead of 0.5), XGBoost achieved a better tradeoff: strong detection of high-income prospects with fewer false positives.
- This tuned XGBoost model was selected as the **final classification model** for deployment.

Figure 2 shows the top 20 features by importance in the XGBoost model; these are the main drivers of the predicted probability of income > \$50K.

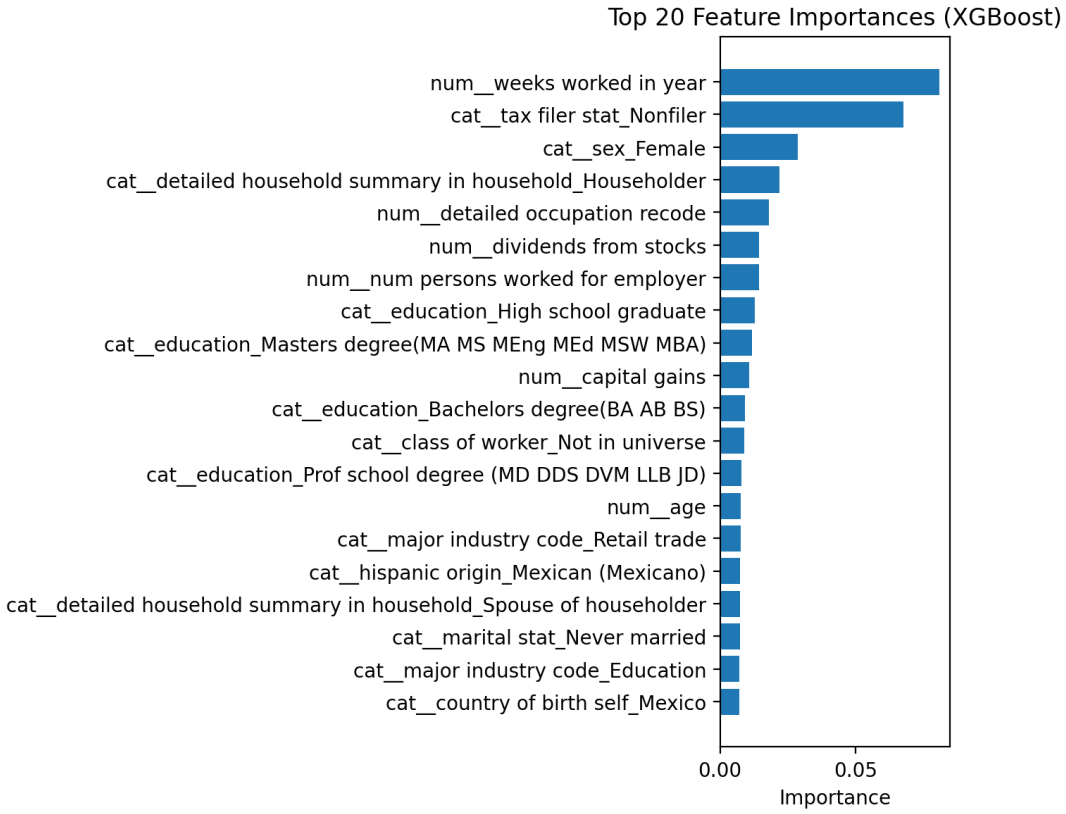

Figure 2: Top 20 feature importances from the XGBoost classification model.

5.3 Recommendation for the client

Use the trained XGBoost classifier to **score new prospects** on the probability of income > \$50K. You can:

- Target the top decile (or another percentile) for premium campaigns.
- Use the score as a continuous priority signal (e.g., for lead routing or budget allocation).
- Re-tune the threshold if your cost of false positives vs. missed high-income individuals changes.

We recommend monitoring model performance over time and retraining when you have new labeled data or when the distribution of prospects shifts.

6 Segmentation Model: Approach, Results and Marketing Use

6.1 Approach

We used K-Means clustering on the preprocessed numeric feature set. The number of clusters was chosen using the elbow method and silhouette scores; hierarchical clustering (dendrogram) on a sample was used to validate the structure. The chosen solution uses **four segments** for a balance between interpretability and separation. Figure 3 shows the elbow curve used to support the choice of four clusters.

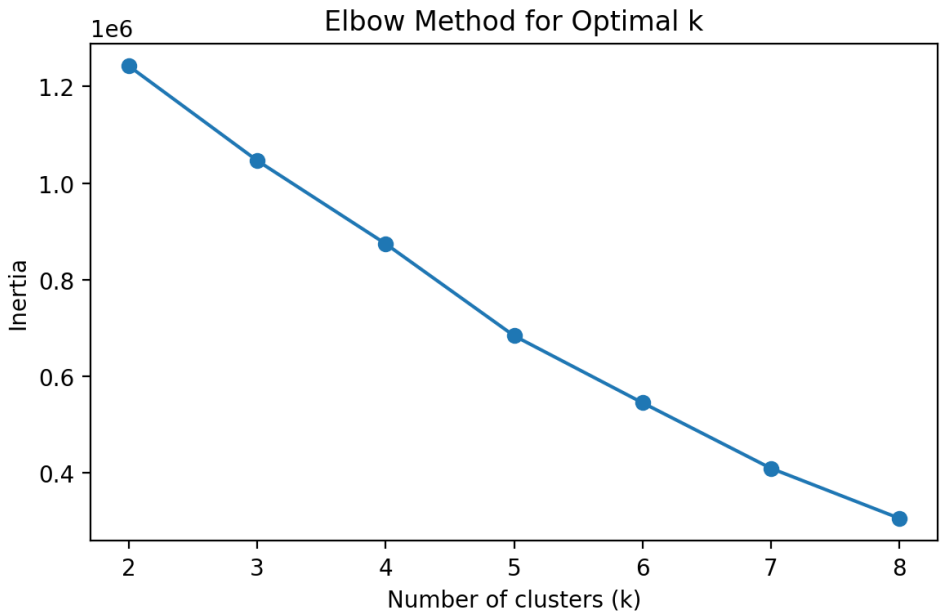

Figure 3: Elbow plot for K-Means: within-cluster sum of squares vs. number of clusters. The bend at $k = 4$ supports the choice of four segments.

6.2 Segment profiles (population-weighted)

Table 1: Segment profiles and marketing implications

Segment	Share	Description	Marketing implication
0. Core working	∼44.5%	Late 30s on average, steady work, moderate wages, little investment income. ∼10% high-income.	Large, stable mass-market segment; suited for broadly targeted, value-oriented products.
1. Wealth / capital-income	∼3.7%	Older, high capital gains and dividends, strong high-income share (∼33%).	Small but high-value; ideal for premium, wealth, and high-ticket offerings.
2. Financially active mid-career	∼2%	Mid-career, moderate wages, notable capital losses (investment activity). ∼30% high-income.	Financially engaged; suitable for investment and financial products and upselling.
3. Low workforce participation	∼49.8%	Younger, low employment and wages, minimal capital income. Very low high-income rate (∼0.5%).	Students, unemployed, or out of labor force; lower immediate value but potential for long-term nurture.

Figure 4 shows the four segments in the space of the first two principal components; the clusters are well separated and support the segment descriptions above.

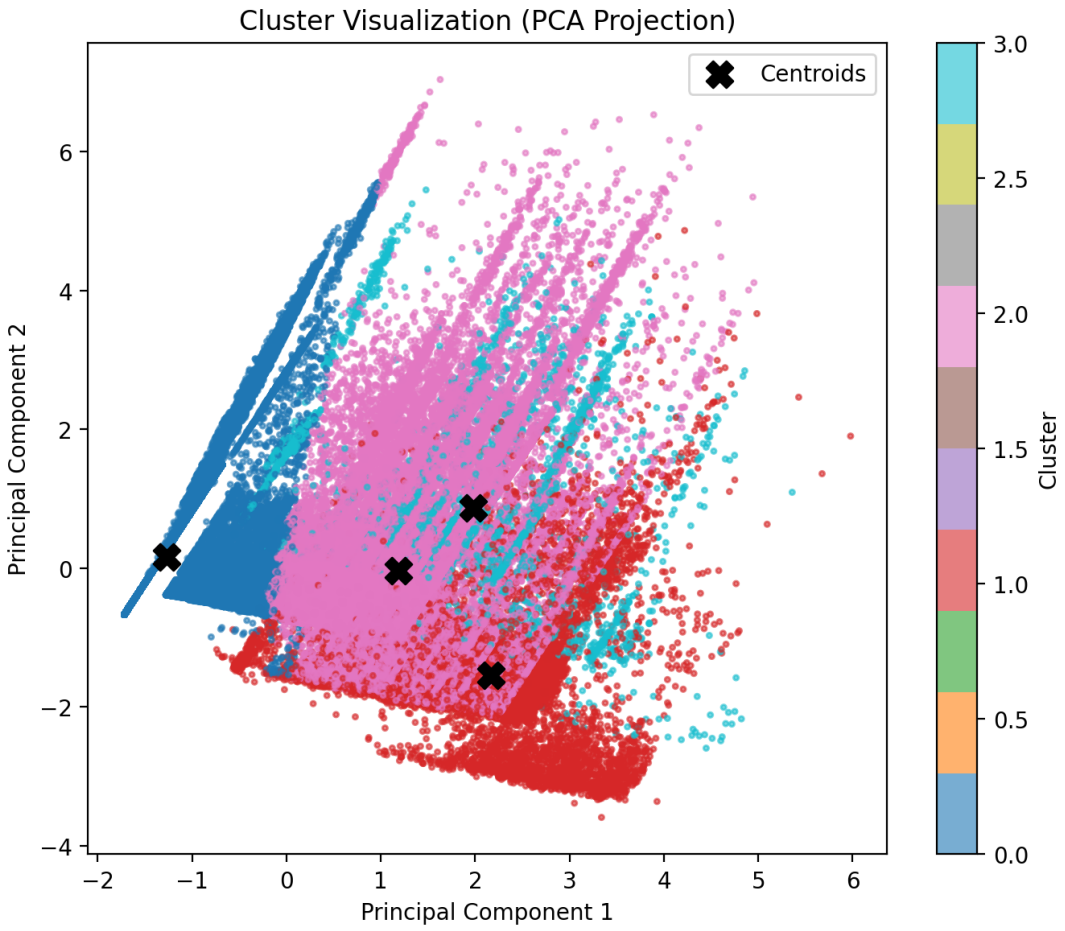

Figure 4: Segments in the space of the first two principal components (sample of points). Colors correspond to the four segments in Table 1.

6.3 Using the segments for marketing

- Use **Segment 1** for high-touch, premium campaigns and high-value product focus.
- Use **Segment 0** for mass-market campaigns and standard product messaging.
- Use **Segment 2** for financial and investment-related cross-sell and upsell.
- Use **Segment 3** for nurture, entry-level offers, or lower-cost channels.

New prospects can be assigned to a segment by applying the same preprocessing and using the fitted cluster centers (nearest centroid). Segment membership can be combined with the classification score for finer targeting.

7 Limitations and Assumptions

- **Limited behavioral information.** Segmentation is based on demographic and financial attributes, not direct consumer behavior or purchasing patterns.
- **K-means modeling assumptions.** K-means assumes spherical, non-overlapping clusters, which may oversimplify real population structure.
- **Static population snapshot.** The segmentation reflects a single time point and does not capture changes in individuals over time.

- **Feature and data scope limitations.** Results depend on available variables and scaling choices, and exclude behavioral or psychographic factors.

8 Recommendations for the Client

1. **Deploy the classification model** to score prospects on probability of income > \$50K and use this score (and/or a threshold) for prioritization and targeting.
2. **Use the four segments** to tailor messaging and channel strategy (premium for Segment 1, mass for Segment 0, financial focus for Segment 2, nurture for Segment 3).
3. **Combine both models** where possible: e.g., segment-specific campaigns with an additional filter or ranking by income score within segment.
4. **Plan for maintenance:** Periodically re-check model performance and segment stability; retrain or re-segment when you have new data or significant distribution shift.
5. **Document usage:** Keep a short record of how thresholds and segments are used in campaigns for future iterations and audits.

9 References

1. Mitchell, T. (1997). *Machine Learning.* McGraw-Hill.
`https://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf`
2. Scikit-learn Machine Learning Library Documentation
`https://scikit-learn.org/stable/`
3. Hastie, T., Tibshirani, R., & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
https://hastie.su.domains/ElemStatLearn/
4. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD.*
XGBoost Documentation

This report is intended for the client as a summary of the work performed and the business use of the classification and segmentation models. For reproduction and extension, see the code and README in the project repository.