Income Classification and Market Segmentation using Weighted Census Data

Dhruvil Gorasiya

2025-10-29

1 Executive Summary

This project was conducted for a retail client seeking to understand income-based customer segments and optimize marketing strategies. Using weighted census data from the 1994–1995 U.S. Census Bureau surveys, two objectives were addressed: (1) to predict whether an individual earns more or less than $50 000 per year, and (2) to segment the population into distinct, actionable groups for targeted outreach.

A calibrated Logistic Regression model delivered strong and reliable performance, achieving a ROC-AUC of 0.94 and PR-AUC of 0.60, with consistent 5-fold cross-validation results (ROC-AUC 0.9379 $\pm$ 0.0023, PR-AUC 0.5771 $\pm$ 0.0119). Despite a class imbalance—only 6.39 % of individuals earn above $50 000—the model produced well-calibrated probability estimates ranging from 0 to 0.98, confirming its robustness.

Weighted K-Means segmentation uncovered eight distinct clusters representing different socioeconomic strata. Clusters 0 and 7 stood out as high-value micro-segments, with positive income rates of 33 % and 30 % and lifts of 5.17× and 4.71× over baseline, while clusters 2 and 4 contained no high-income individuals. Collectively, these two top clusters represent just 5.7 % of the population but more than 30 % of potential high-income targets—offering substantial marketing efficiency gains.

Overall, this project demonstrates that interpretable, well-calibrated statistical modeling and data-driven segmentation can yield actionable insights for customer targeting, revenue optimization, and strategic resource allocation.

2 Data Understanding

The dataset originates from the 1994–1995 U.S. Census Bureau Current Population Survey (CPS) and contains 40 demographic and employment variables, plus a sample weight and an income label indicating whether an individual earns more or less than $50 000 per year.

Each record represents a weighted observation reflecting a portion of the U.S. population. Variables span demographics (age, gender, marital status), education and employment (degree, occupation, industry, hours worked), and financial attributes (wage per hour, capital gains/losses, dividends).

The target variable (label_bin) is binary:

- $1 \rightarrow$ income $> \$50\,000$
- $0 \rightarrow income < \$50\ 000$

Only 6.39% of records fall into the high-income category, making this a class-imbalanced problem. Roughly two-thirds of the fields are categorical, with some containing "Not in universe" or rare responses typical of survey data. The weight column was used throughout to correct for sampling bias.

Preliminary inspection revealed that education level, occupation, marital status, and hours worked per week correlate strongly with income. These early insights informed later feature transformations and the model's focus on socio-economic factors.

3 Data Preparation

The preprocessing and feature-engineering pipeline was implemented through modular scripts (data_processor.py) and executed in the demo.ipynb notebook. Each step was designed to ensure data integrity, prevent target leakage, and maintain the population weighting scheme.

3.1 Data Loading and Normalization

Column names were read from census-bureau.columns and aligned to each record in census-bureau.data. Whitespace, capitalization, and inconsistent delimiters were normalized using regex cleaning functions.

The target label was mapped to a binary variable label_bin, where

1 = income > $50 000

0 = income < $50 000.

The sampling weight column was retained as weight; missing or zero weights were imputed to 1.0 to preserve relative proportions.

3.2 Target Integrity and Filtering

Records with ambiguous or missing income labels were removed. Observations with non-positive weights were excluded. Potential "leakage" fields such as income recode, target, or literal ">50 K" text entries were dropped automatically via regex pattern matching.

3.3 Handling Missing and Rare Categories

Missing categorical values were replaced with the mode ("most-frequent" strategy). To reduce sparsity, categories representing less than 1 % of total weighted frequency were grouped under a unified token _RARE_. This step improved model stability and reduced the dimensionality of One-Hot Encoding from over 2400 to roughly 1000 binary features.

3.4 Feature Transformation

Numeric variables (e.g., age, hours worked per week, capital gains/losses) were standardized using the Yeo-Johnson power transformation, mitigating skewness without requiring strictly positive values. Categorical variables were encoded through One-Hot Encoding with handle_unknown='ignore', preserving unseen test categories. The combined transformations were managed by a unified ColumnTransformer within a Scikit-learn Pipeline, ensuring consistency across training, validation, and segmentation stages.

3.5 Train–Test Split and Weighting

The dataset was split into training (80%) and testing (20%) partitions using stratified sampling on the binary label to preserve the 6.39% positive-class ratio. All training and evaluation metrics—ROC-AUC, PR-AUC, F1, Precision, Recall—were computed using sample weights, aligning model performance with true population proportions.

3.6 Output Artifacts

The preprocessing pipeline produced:

- **X_train, X_test** — encoded sparse matrices for model input
- **y_train, y_test** — binary labels
- **w_train, w_test** — corresponding sample weights
- **preprocessor.pkl** — serialized transformer for reuse in scoring and clustering

4 Exploratory Data Analysis (EDA)

Exploratory analysis in census_eda.ipynb examined demographic and occupational patterns influencing income. Weighted summaries and visualizations revealed clear socioeconomic trends aligned with prior census research.

- **Income Imbalance**: Only 6.39 % of individuals earned above $50 000, confirming the strong class imbalance.
- **Education**: Income rose sharply with education level — high school graduates rarely exceeded $50 K, while college graduates exceeded 30% positive rate.
- **Occupation**: Executive, managerial, and technical professions dominated the high-income group; service and retail sectors were primarily low-income.
- **Marital Status**: Married individuals had the highest income rates, followed by widowed or divorced, while single respondents were least likely to earn $>$ $50 K
- **Work Intensity**: Individuals working longer hours or full-time schedules were far more likely to belong to the high-income group
- **Capital Gains**: Highly skewed but strongly correlated with high income, justifying later power transformation

Overall, education, occupation, marital status, and work hours emerged as the most influential predictors, providing an early signal of the model's key decision drivers.

5 Modeling Approach and Evaluation

5.1 Methodology

To predict whether individuals earn above $50 000, several supervised models were evaluated — Logistic Regression, Random Forest, and XGBoost — using a consistent preprocessing pipeline. Each model operated on features processed through a unified ColumnTransformer (median imputation, Yeo–Johnson scaling, and One-Hot Encoding). Weighted training and validation were applied to account for population stratification

After baseline comparisons, XGBoost was selected as the final model due to: Strong performance with low variance across folds, Excellent calibration of probability outputs, and Interpretability of coefficients and SHAP attributions.

5.2 Model Calibration and Validation

The XGBoost model was fitted with sample weights and calibrated using a sigmoid (Platt) scaling procedure to improve probabilistic predictions. A 5-fold stratified cross-validation was then performed to measure generalization.

Table 1: Classification Model Performance Summary (Calibrated Logistic Regression)

Metric	**Mean**	**Std. Dev.**
ROC-AUC	**0.9451**	± 0.0021
PR-AUC	**0.6200**	± 0.0114
Precision	0.590	—
Recall	0.601	—
F1-Score	0.595	—
Optimal Threshold	0.19	—

5.3 Results

On the held-out test set, the calibrated Logistic Regression achieved: ROC-AUC = 0.9429, PR-AUC = 0.6004, confirming strong separability despite the 6.39% positive rate. Predicted probabilities were well-distributed (0–0.98), showing good calibration. The narrow cross-validation deviations indicate high model stability.

5.4 Feature Importance and Explainability

SHAP analysis confirmed that:

- Education level, occupation type, hours worked per week, and marital status were the strongest positive predictors of higher income.
- Age, industry, and capital gains provided additional differentiating power.

These findings align closely with economic intuition and EDA observations, reinforcing the model's interpretability and business credibility.

6 Segmentation Analysis

6.1 Methodology

To identify distinct population segments for targeted marketing, Weighted K-Means clustering was applied on the transformed feature space using census weights as sample importance. The optimal number of clusters was selected through inertia minimization and interpretability analysis, resulting in eight well-separated clusters. Each cluster was profiled across demographic, occupational, and income characteristics using weighted averages and top-category shares.

6.2 Results Overview

The overall positive rate for income $>$ \$50 000 was 6.39%, but significant variation appeared across clusters:

Table 2: Summary of Weighted K-Means Clusters for Income Segmentation

Cluster	Pop. (%)	Pos. Rate (%)	Lift	Segment Type
0	3.7	33.1	**5.17**×	High-value micro-segment
7	2.0	30.1	**4.71**×	High-value professionals
3	20.2	9.0	1.44×	Mid-income group
6	19.1	10.4	1.62×	Mid-income group
1	21.7	4.0	0.63×	Transitional segment
5	5.5	1.4	0.22×	Low-income segment
2	13.3	0.0	0.00×	Non-target
4	14.5	0.0	0.00×	Non-target

6.3 Key Insights

High-value segments (Clusters 0 & 7): Together, they represent only 5.7 % of the population but contain over 30 % of high-income individuals. These clusters include well-educated professionals, often married, with full-time employment and managerial or technical occupations.

Medium-value segments (Clusters 3 & 6): Represent stable middle-class populations with moderate lift; ideal for aspirational or mid-tier marketing strategies.

Low/No-value segments (Clusters 2 & 4): Large portions of the population with zero high-income individuals, suitable for campaign exclusion to reduce wasted reach.

Transitional clusters (1 & 5): Contain mixed demographic profiles with limited purchasing power; may require tailored retention or upskilling initiatives.

6.4 Business Implications

The segmentation demonstrates a clear hierarchy of economic potential. Targeting Clusters 0 and 7 can yield the highest return on marketing investment with minimal population coverage. Moreover, removing Clusters 2 and 4 from broad campaigns can reduce costs by approximately 25% while maintaining the majority of high-value prospects.

7 Discussion and Recommendations

The calibrated XGBoost model demonstrated high predictive reliability and strong generalization across folds. Its gradient-boosted tree architecture effectively captured nonlinear feature interactions, while post-training sigmoid calibration ensured well-behaved probability outputs for downstream decision-making.

7.1 Key Takeaways

Model Strength: The model achieved ROC-AUC = 0.94 and PR-AUC = 0.62, confirming excellent discrimination despite a 6.39 % positive-class rate.

Feature Drivers: SHAP analysis showed that education level, occupation, marital status, hours worked per week, and capital gains were the most influential predictors of high income.

Segmentation Value: Weighted K-Means revealed eight distinct socioeconomic segments. Clusters 0 and 7, together representing only = 6 % of the population, contained > 30 % of all high-income individuals—offering substantial marketing leverage.

7.2 Business Recommendations

Prioritize high-value clusters (0 & 7). Focus premium and personalized marketing here; their 5× lift over baseline yields the highest ROI per impression.

Engage mid-value clusters (3 & 6). These groups form a large, upwardly mobile middle-income base suitable for aspirational or subscription-tier products.

De-prioritize clusters (2 & 4). With zero high-income incidence, excluding them from paid campaigns could cut marketing spend by 25% without hurting reach.

Monitor transitional clusters (1 & 5). Track these segments over time; demographic or economic mobility could shift them toward higher value.

Operational integration. Deploy the calibrated XGBoost model within the CRM pipeline to score new leads. Use SHAP explanations for model transparency and quarterly retraining to maintain calibration.

7.3 Strategic Impact

Combining a powerful calibrated XGBoost classifier with segmentation analytics enables data-driven targeting rather than demographic guessing. By concentrating on only 6 % of the population, the client can capture roughly one-third of the potential high-income market—significantly improving campaign efficiency and reducing acquisition cost.

8 Limitations and Future Work

8.1 Limitations

Historical Data Context: The dataset originates from the 1994–1995 U.S. Census Bureau CPS. Although representative of its time, economic structures and workforce demographics have evolved substantially since then, which may limit generalizability to today's population.

Survey-Design Noise: Some features contain "Not in universe" or self-reported values that introduce non-systematic noise. Even after cleaning and rare-category handling, residual bias may remain.

Class Imbalance: With only 6.39 % of observations labeled > \$50 000, model calibration accuracy for the minority class depends heavily on weighting. Small sampling errors could affect tail-probability estimates.

Model Interpretability: Although SHAP values improve transparency, boosted trees remain complex compared to linear models, and business stakeholders may require simplified rule-based summaries.

Cluster Volatility: Weighted K-Means assumes spherical cluster shapes; slight shifts in scaling or feature composition can alter cluster boundaries. Re-evaluation is needed if new data are added.

8.2 Future Work

Data Modernization: Extend the analysis with recent ACS (American Community Survey) data to reflect current labor and income trends.

Model Enhancement: Explore TabNet, FT-Transformer, or CatBoost for richer nonlinear interactions while maintaining explainability.

Fairness and Bias Audits: Evaluate demographic parity, equal opportunity, and calibration across gender and race groups to ensure ethical deployment.

Dynamic Segmentation: Replace static clustering with probabilistic or streaming segmentation (e.g., Gaussian Mixture Models or DBSCAN) to adapt to evolving customer behavior.

Business Integration: Embed the calibrated XGBoost model into the client's CRM scoring pipeline, linked with automated monitoring dashboards for drift detection and re-training triggers.

9. References

1. **Pedregosa, F. et al. (2011)** – *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12*, 2825–2830.
   (Used for model training, preprocessing pipelines, and evaluation metrics.)

2. **Chen, T., & Guestrin, C. (2016)** – *XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
(Foundation of the gradient boosting algorithm used in this project.)