Income Classification & Customer Segmentation using U.S. Census Data

Report by: Gauravi Patankar | **Email:** gsp2137@columbia.edu

Executive Summary

This project utilizes data science methods to analyze U.S. Census Bureau data for **income classification** and **customer segmentation**, thereby supporting a retail marketing strategy. The dataset, comprising nearly 200,000 individuals, includes demographic, educational, occupational, and socioeconomic attributes, enabling both predictive and exploratory insights.

In the **classification task**, models like Logistic Regression, Random Forest, XGBoost, and LightGBM were trained to predict whether an individual earns more than $50K per year. Since the high-income group represents only ~6% of the data, the focus was on optimizing the **F1-score**, balancing precision (cost-effective targeting) and recall (maximizing reach). Key predictors included education, work duration, and capital income.

The **segmentation task** applied unsupervised clustering (K-Means, GMM) to discover interpretable population groups based on socioeconomic patterns. Five clear personas emerged, from *young working-class earners* to *high-wealth retirees*, each with distinct demographic and behavioral profiles.

Problem Definition & Business Context

The goal of this project is to use data from the U.S. Census to implement:
(1) a **Classification Model** to predict income category (>$50K vs ≤$50K), and
(2) a **Segmentation Model** to identify distinct population personas.

Both approaches serve the broader business goal of understanding and targeting customer groups more effectively, enabling data-driven decision-making.

A. Income Classification

The classification task predicts whether an individual earns above or below $50K annually, a key indicator of **purchasing power**.

- **High-income individuals** → ideal for luxury, credit, or subscription-based offerings.

- **Low-income individuals** → better suited for discounts, mass-market products, or budget plans.

Because the data is **heavily imbalanced** (only ~6% high-income), standard accuracy would be misleading. Instead, the model optimizes the **F1-score**, which balances:

- **Precision** → reduces false positives; critical when campaigns have **budget constraints** (e.g., sending targeted offers to verified high-income customers).
- **Recall** → increases true positives; valuable for **reach-driven campaigns** (e.g., identifying all potential premium customers).

B. Customer Segmentation

The segmentation analysis explores **natural clusters** within the population using unsupervised learning (MiniBatch K-Means). The aim is to identify **interpretable socio-demographic segments**. Each segment captures unique behavioral and economic traits, such as:

- *Young working-class earners* → price-sensitive, respond to promotions or financing offers.
- *Mid-career professionals* → stable earners, ideal for brand loyalty programs.
- *High-wealth retirees* → respond better to investment or lifestyle campaigns.

Data Overview

Source: 1994–1995 U.S. Census Current Population Survey (CPS) containing **~199K records and 42 features**

Features:

- **Demographics:** age, sex, race, citizenship, country of birth.
- **Education & Household:** education level, marital status, family composition.
- **Employment:** class of worker, occupation, industry, weeks worked, veteran status.
- **Socio-economic Factors:** wages, capital gains/losses, dividends, and household income proxies.

Target Variable (for classification): label→ **Income > $50K (6.2%)** vs **≤ $50K (93.8%)**, indicating a significant class imbalance.

Sampling Weights: Each observation includes a **weight variable** representing how many individuals in the population that record stands for, ensuring analyses reflect true population-level proportions.

Data Preprocessing

The following steps were undertaken for data preprocessing:

- **Duplicates:** Kept (~1.6 %) since they represent real individuals.
- **Data Types:** Converted coded variables (occupation, industry, veteran status) to categorical.
- **Category Simplification:** Merged education levels (7 tiers), grouped countries into 6 regions, and condensed household roles for interpretability.
- **Missing Values:** "Hispanic origin" NaN + "Do not know" → "Unknown."
- **Outliers:** Z-score for near-normal (age, weight), Log + IQR for skewed financial features; Winsorized post-split to avoid leakage.
- **Encoding & Scaling:** One-Hot + Ordinal encoders for categoricals; RobustScaler for numerics within pipelines.

Feature Engineering

Feature transformations were created for interpretable, model-ready segmentation.

- Net Capital Income = Capital Gains − Losses + Dividends → proxy for asset wealth.
- Annual Income Proxy = Wage per Hour × Weeks Worked → approximate earnings.
- Age Group and Education Tier were derived for stratified insight.
- Categorical regroupings reduced noise and highlighted socio-economic patterns for both classification and segmentation.

Part A – Income Classification

Pipeline Design

- To ensure robust evaluation and prevent overfitting, the data was divided into **Train, Validation, and Test** subsets using a **stratified split**, preserving class proportions.
- The training set was resampled using **SMOTE** (Synthetic Minority Oversampling Technique) to balance classes by synthetically generating minority-class samples.
- All preprocessing steps were integrated into Scikit-learn Pipelines for reproducibility and leakage prevention.

Models Tested

- Logistic Regression
- Random Forest
- XGBoost
- LightGBM
- CatBoost

Evaluation Metrics

Given the imbalance, the **F1-score** was used as the primary metric, balancing precision (avoiding false positives) and recall (capturing true high-income cases).

Hyperparameter Tuning & Threshold Optimization

Both **RandomizedSearchCV** (broad exploration) and **GridSearchCV** (fine-tuning) were used for hyperparameter tuning, ensuring optimal tree depth, learning rate, and regularization parameters.

After model selection, decision thresholds were adjusted on validation data to maximize the F1-score, as default thresholds (0.5) often underperform with imbalanced targets.

Results & Insights

Across all five models, Logistic Regression, Random Forest, XGBoost, LightGBM, and CatBoost, overall performance was **comparable**, with F1-scores ranging between **0.50 and 0.53** on the test set. While LightGBM and XGBoost maintained slightly better precision–recall balance, differences were small, and no single model consistently dominated across all metrics.

Overall Performance Summary

Each model's results were analyzed at three levels: baseline threshold (0.5), tuned threshold (optimized for F1 on validation), and final test performance. The table below summarizes key results for the minority (high-income) class:

Random Forest	Precision	Recall	F1-score	True Positives (Minority Class)	False Negative (Minority Class)
Baseline	0.55	0.45	0.49	558	680
Threshold Tuning (0.37)	0.37	0.61	0.53	759	479
TestSet	0.44	0.57	0.5	708	530
XGB	**Precision**	**Recall**	**F1-score**	**True Positives (Minority Class)**	**False Negative (Minority Class)**
Baseline	0.47	0.58	0.52	729	509
Threshold Tuning (0.68)	0.48	0.61	0.54	763	475
TestSet	0.47	0.58	0.52	722	516
LightGBM	**Precision**	**Recall**	**F1-score**	**True Positives (Minority Class)**	**False Negative (Minority Class)**
Baseline	0.55	0.48	0.51	597	641
Threshold Tuning (0.28)	0.45	0.62	0.52	768	470
TestSet	0.44	0.61	0.51	759	479
CatBoost	**Precision**	**Recall**	**F1-score**	**True Positives (Minority Class)**	**False Negative (Minority Class)**
Baseline	0.54	0.48	0.51	600	638
Threshold Tuning (0.34)	0.49	0.54	0.52	678	560
TestSet	0.48	0.54	0.51	675	563
LR	**Precision**	**Recall**	**F1-score**	**True Positives (Minority Class)**	**False Negative (Minority Class)**
Baseline	0.26	0.88	0.4	1097	141
Threshold Tuning (0.87)	0.52	0.54	0.53	680	558
TestSet	0.5	0.52	0.51	647	591

Key Observations

- F1-scores clustered tightly **(0.50–0.53)**, showing all models captured the same core income patterns.
- *LightGBM* and *XGBoost* offered the best recall, identifying more true high-income individuals → useful for *reach-maximizing* campaigns.
- *CatBoost* and *Logistic Regression* achieved slightly higher precision, reducing false positives → better for *budget-constrained* targeting.
- *Random Forest* sat in the middle ground, showing balanced but slightly noisier predictions.

Feature Insights

All models converged on similar high-impact predictors:

- Education level (↑) – Strongest indicator of higher income.
- Weeks worked per year (↑) – Longer employment duration correlated directly with higher income.
- Capital gains (↑) – Reflects asset ownership and investment income.
- Tax filer status (↑) – Strong proxy for stable earnings and financial participation.
  Household structure (↓) – Dependents and "group quarters" residents were less likely to earn above $50K.

Part B – Customer Segmentation

The customer segmentation task was designed to uncover natural socio-demographic clusters within the U.S. population, essentially identifying **personas** that reflect distinct profiles.

Preprocessing & Feature Preparation

The dataset used the same cleaned and encoded version from the classification task, excluding the target label and sampling weights.

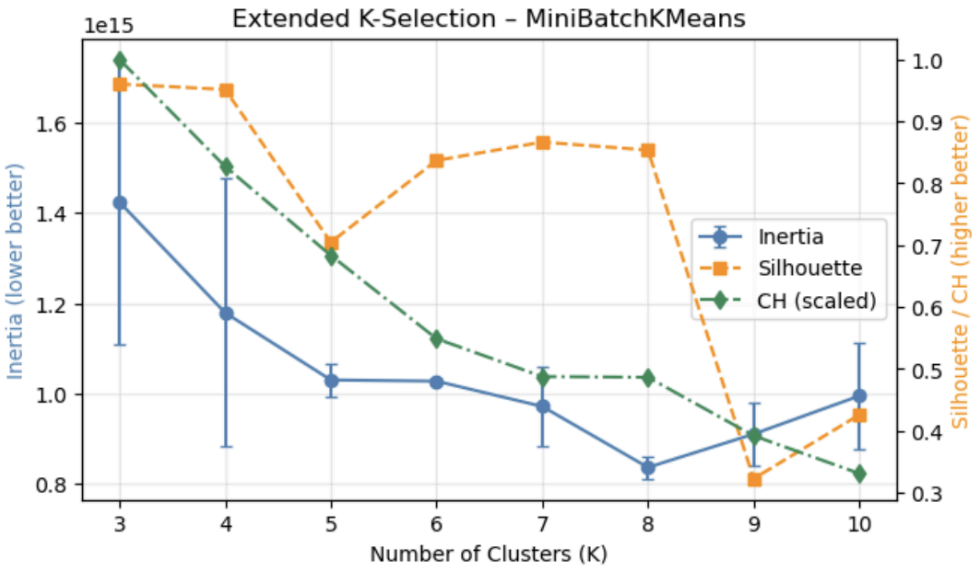
Key features included *age, education, class of worker, marital status, weeks worked, capital income proxies, and household role*.

All variables were encoded (One-Hot/Ordinal) and scaled using RobustScaler to handle wide value ranges and outliers. Feature-engineered columns like *annual income proxy* and *net capital income* were also included to enrich the socioeconomic dimension.

Clustering Method

MiniBatch K-Means was employed, chosen for its scalability and stability with large datasets (~200K rows). It approximates standard K-Means but processes data in mini-batches, enabling faster convergence without compromising interpretability.

Cluster Evaluation



To determine the optimal number of clusters, several internal metrics were analyzed:

- **Inertia (Elbow Method):** To observe diminishing returns in variance reduction.
- **Silhouette Score:** To measure how well individuals fit within their cluster.
- **Calinski–Harabasz (CH) Index:** To balance cohesion and separation.

All three metrics consistently pointed to an optimal K = 5, balancing interpretability and granularity.

Results

The final five clusters revealed clear, interpretable personas with distinct economic and lifestyle traits, forming the foundation for actionable marketing insights presented in the next section.

Cluster Profiles

- **Cluster 0 – Mid-Career Professionals:** Well-educated, full-time workers with strong earning potential; ideal for loyalty programs or premium financial products.
- **Cluster 1 – Dependents / Non-Workers:** Students, homemakers, or retirees not in the labor force; suited for education, essential goods, or awareness campaigns.
- **Cluster 2 – Steady Working Households:** Consistently employed service or clerical workers with moderate income; respond well to family-value or discount offers.
- **Cluster 3 – Educated Asset-Rich Retirees:** Older, financially independent individuals with investment income; good targets for health, finance, or retirement planning.
- **Cluster 4 – Young Emerging Workers:** Early-career, tech-savvy earners with rising income; strong potential for brand loyalty and subscription-based products.

Cluster	Economic Tier	Employment Type	Profile Summary
0 – Young Working-Class Employees	Lower income	Private wage earners	Young adults (18–25), mainly female, employed in retail/service sectors. Low education (Below HS/HS Grad), never married, low income (~400K), minimal assets.
1 – Married Skilled Professionals	Upper-middle income	Private / Government professionals	Mid-career men (36–50), married, high wages (~3.6M), modest capital income. Educated, steady jobs in manufacturing and hospital sectors.
2 – Dependents / Non-workers	Low / inactive	Not in labor force	Children, students, or unemployed adults; minimal or zero income and assets. Largely dependents supported by others in household.
3 – Steady Working Households	Middle income	Private / Clerical / Service roles	Early-to-mid-career adults (26–50), mix of genders, high-school educated, stable full-time employment (~1.5M income), modest investments.
4 – Educated Asset-Rich Retirees	High wealth	Self-employed / Retired investors	Older men (50+), highly educated (Bachelor's+), married, no wage income but high capital income (~21K). Financially independent households.

Each cluster represents a **socioeconomic persona**, defined by combinations of education, occupation, and income proxies.

Future Scope

Future improvements that I believe can further strengthen both the classification and segmentation models:

- **Test additional classification algorithms** (Neural Networks, Gradient Boosted Ensembles with stacking, and Support Vector Machines)
- **Handle class imbalance more effectively** using cost-sensitive learning, adaptive resampling, or focal loss functions.
- **Use advanced optimization techniques** such as Bayesian Optimization or Optuna to automate hyperparameter tuning.
- **Incorporate dimensionality reduction methods** like PCA or UMAP before clustering to improve separation, visualization, and reduce redundancy.
- **Experiment with more clustering algorithms** (e.g., DBSCAN, Hierarchical, or Gaussian Mixture Models) to capture non-spherical and overlapping population structures.