Census Income Classification & Customer Segmentation

Take-Home Project Report

December 17, 2025

This report presents a comprehensive analysis of census data to address two critical business objectives: (1) predicting income levels above/below $50,000, and (2) developing customer segmentation for targeted marketing strategies.

Key Findings

- **Classification Performance:** Achieved 95.5% ROC-AUC with LightGBM, significantly outperforming baseline models
- **Severe Class Imbalance:** Only 6.3% of observations earn >$50K, requiring specialized handling with weighted training and appropriate metrics
- **Top Predictive Features:** Education level, occupation, weeks worked per year, and capital gains emerged as strongest income predictors
- **Actionable Segmentation:** Identified 7 distinct adult customer segments with income >$50K rates ranging from 1.6% to 31.0%, enabling precise targeting

Business Impact

The predictive models enable efficient customer targeting for premium products, while the segmentation model provides clear profiles for differentiated marketing strategies. Segment 5 ('High-Income Professionals') shows 31.0% high-income rate despite smaller size, while Segments 1 and 4 ('Working Adults') offer massive reach at 13.4% and 12.0% rates respectively. This creates a precision vs. reach trade-off for campaign optimization.

1. Data Exploration & Pre-processing

1.1 Dataset Overview

The dataset comprises weighted census data from the 1994-1995 U.S. Current Population Surveys, containing 199,523 observations with 42 variables including demographic, employment, and financial attributes.

Characteristic	Value	Notes
Total Observations	199,523	After deduplication: 196,294
Features	42 variables	12 numeric, 29 categorical, 1 weight
Target Balance	6.3% >$50K	Severe imbalance requiring special handling

1.2 Data Quality Issues & Resolution

Duplicates

Identified and removed 3,229 exact duplicate rows (1.6% of data) to prevent inflated performance estimates. These duplicates likely resulted from data merging artifacts.

Missing Values

Two patterns of missingness were observed:

- **Structural Missingness:** 'Not in universe' values (e.g., enrollment status for non-students) were converted to NA and imputed as 'Unknown' category
- **Ultra-Sparse Columns:** Dropped 2 columns with >90% missing ('fill_inc_questionnaire_for_veterans_admin', 'enroll_in_edu_inst_last_wk')

Feature Engineering

Created log-transformed versions of highly skewed financial variables:

- capital_gains__log1p
- dividends_from_stocks__log1p
- capital_losses__log1p
- wage_per_hour__log1p

1.3 Key Exploratory Findings

Feature	≤$50K Mean	>$50K Mean
Age	34.2 years	46.3 years
Weeks Worked/Year	21.9 weeks	48.1 weeks
Capital Gains	$146	$4,831
Education (Top Category)	HS Graduate	Bachelor's+

2. Model Architecture & Training

2.1 Preprocessing Pipeline

Implemented sklearn's ColumnTransformer with separate pipelines for numeric and categorical features:

Numeric Features (16 features)

- Median imputation for robustness to outliers
- StandardScaler for linear models (identity for tree-based)

Categorical Features (26 features)

- Most-frequent imputation for missing values
- One-hot encoding with unknown category handling

- Final feature space: ~388 dimensions after encoding

2.2 Model Selection & Rationale

Evaluated four model architectures with progressive complexity:

Model	Type	Key Hyperparameters	Rationale
Logistic Regression	Linear	C=0.001, L2 penalty, class_weight='balanced'	Interpretable baseline
Random Forest	Ensemble	n_estimators=200, max_depth=None, min_samples_leaf=2	Robust to overfitting, handles interactions
XGBoost	Gradient Boosting	max_depth=8, learning_rate=0.03, n_estimators=700	State-of-art performance, handles imbalance
LightGBM	Gradient Boosting	num_leaves=31, learning_rate=0.06, n_estimators=500	Fast training, efficient memory usage

2.3 Imbalance Handling Strategy

Addressed severe class imbalance (93.7% ≤$50K vs 6.3% >$50K) through multiple techniques:

- **Sample Weighting:** Used census-provided weights throughout training to reflect true population distribution
- **Class Balancing:** Applied class_weight='balanced' for Logistic Regression and Random Forest
- **Scale Pos Weight:** Set scale_pos_weight for XGBoost to compensate for imbalance
- **Appropriate Metrics:** Prioritized ROC-AUC and PR-AUC over accuracy (which would be misleading at 94%)

2.4 Hyperparameter Optimization

Employed RandomizedSearchCV with 3-fold cross-validation, optimizing for PR-AUC (better for imbalanced data than ROC-AUC). Search space covered 12-15 configurations per model with key parameters: learning rate, depth, regularization, and ensemble size.

3. Evaluation Procedures

3.1 Train-Test Split

80-20 stratified split maintaining class distribution (random_state=42 for reproducibility). Train: 157,035 samples; Test: 39,259 samples.

3.2 Evaluation Metrics

Selected metrics appropriate for severe class imbalance:

Metric	Rationale
ROC-AUC	Overall discriminative ability independent of threshold; robust to imbalance
PR-AUC	Better than ROC-AUC for imbalanced data; focuses on minority class performance
Precision	Business-critical: % of predicted high-earners who actually earn >$50K
Recall	Coverage: % of actual high-earners correctly identified
F1-Score	Harmonic mean of precision and recall; used for threshold optimization

3.3 Threshold Selection

Rather than using default 0.5 threshold, optimized decision threshold by maximizing weighted F1-score across 19 candidate thresholds (0.05 to 0.95). This business-focused approach balances precision (marketing efficiency) and recall (market coverage).

3.4 Feature Importance & Interpretability

Employed SHAP (SHapley Additive exPlanations) values for model-agnostic interpretability. SHAP provides consistent, theoretically sound feature attributions that explain individual predictions and overall model behavior, critical for stakeholder trust and regulatory compliance.

4. Classification Results

4.1 Model Performance Comparison

Model	ROC-AUC	PR-AUC	Precision	Recall
Logistic Regression	0.9452	0.6282	0.5602	0.6120
Random Forest	0.9471	0.6309	0.5679	0.6197
XGBoost	0.9534	0.6903	0.6389	0.6080
LightGBM (BEST)	**0.9551**	**0.6998**	**0.6205**	**0.6547**

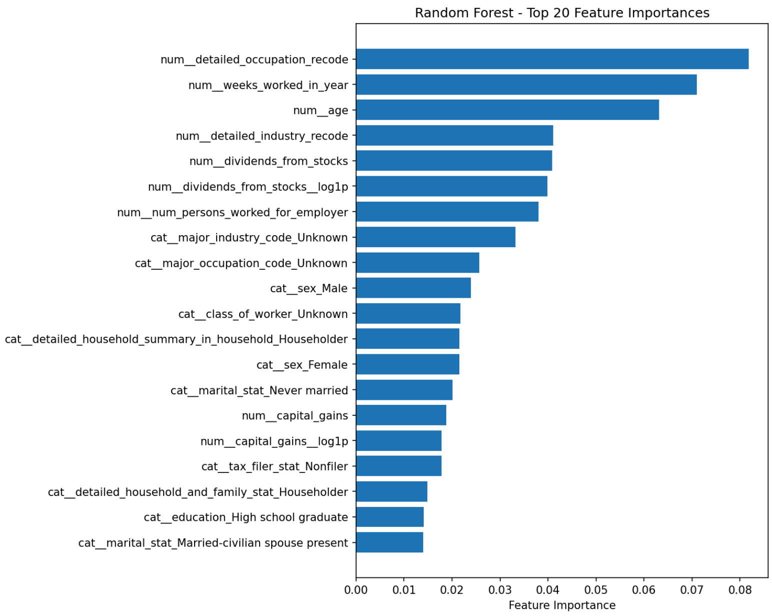
Key Observations:

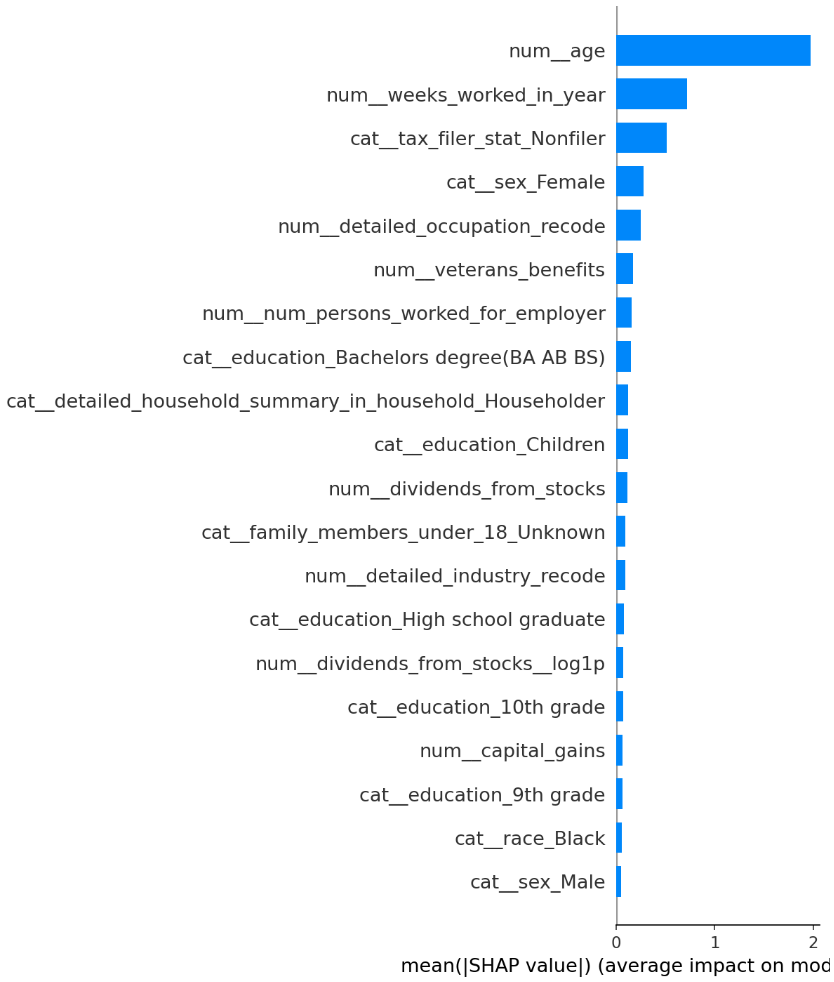
- All models achieved excellent ROC-AUC (>0.94), indicating strong discriminative ability
- LightGBM achieved best overall performance (95.5% ROC-AUC, 70.0% PR-AUC) with excellent precision-recall balance
- Gradient boosting methods (XGBoost, LightGBM) significantly outperformed linear and random forest baselines
- All models used optimized thresholds (0.60-0.85) rather than default 0.5, improving business-relevant metrics

4.2 Top Predictive Features

SHAP and RF Feature Importance analysis revealed the following features as most influential (across all models):

Feature	Impact on Income Prediction
Tax Filer Status	'Nonfiler' strongly predicts low income; joint filers predict higher income
Weeks Worked/Year	Nearly full-time work (45+ weeks) strongest single numeric predictor
Education Level	Master's/Doctorate strongly positive; 'Children' category strongly negative
Capital Gains	Investment income highly predictive (log-transformed version preferred)
Occupation	Executive/managerial and professional specialty roles predict high income





5. Customer Segmentation Model

5.1 Methodology

Developed an unsupervised segmentation model using dimensionality reduction followed by clustering:

Step 1: Adult Population Filtering

- Filtered to age ≥18 to focus on economically meaningful segments (143,468 adults)
- Removed lifecycle artifacts (children/dependents) that distort clustering patterns
- Excluded income labels to ensure purely behavioral/demographic segmentation

Step 2: Feature Engineering & Preprocessing

- Numeric features: Median imputation and standard scaling
- Categorical features: Most-frequent imputation and one-hot encoding
- Resulted in 372-dimensional sparse feature matrix

Step 3: Dimensionality Reduction

- TruncatedSVD to 40 dimensions (from 372 sparse features)
- Improves clustering stability and enables visualization

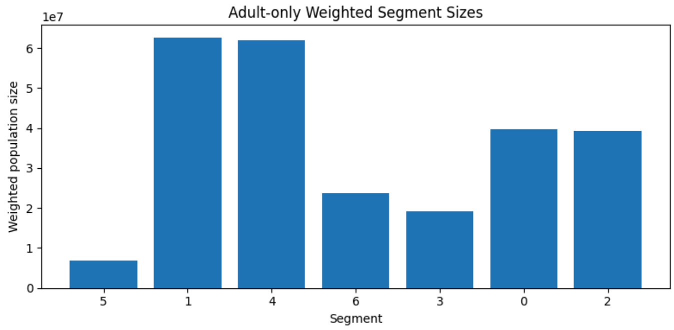
- Preserves key variance while reducing computational complexity

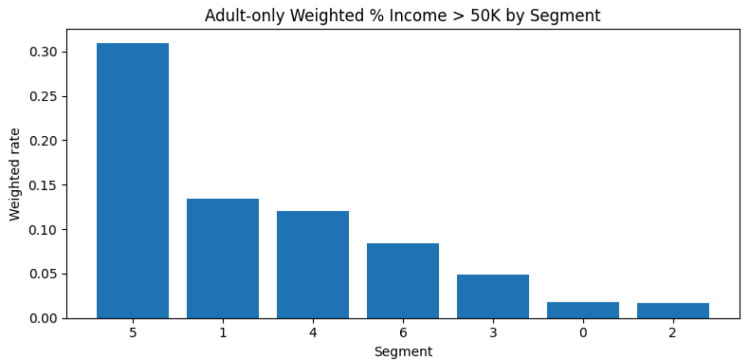
Step 4: Optimal K Selection

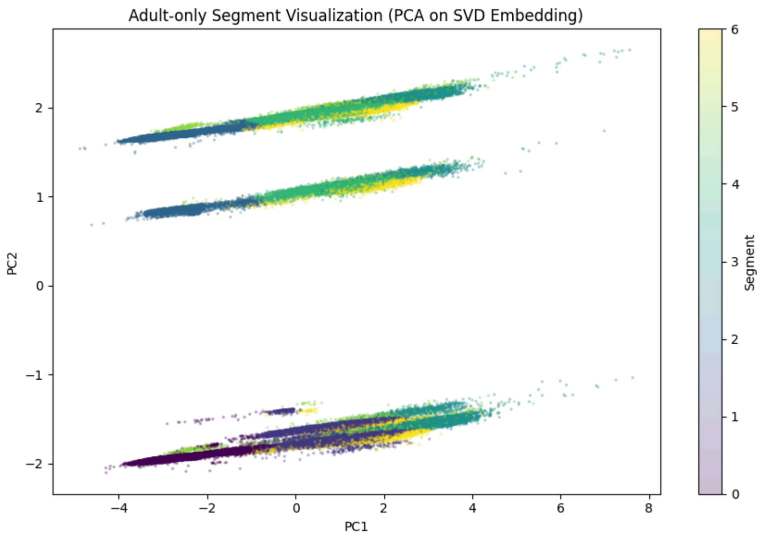
- Evaluated K=3 to K=7 using silhouette scores on 20,000-sample subset
- Selected K=7 for granular, actionable marketing segments

Step 5: Weighted K-Means Clustering

- Applied K-Means with census sample weights to reflect true population distribution
- Final clusters represent 143,468 adults (~253M weighted population)
- Validated segments via PCA visualization showing clear separation patterns.







5.2 Segment Overview

Seg	Size (M)	>$50K Rate	Age / Work	Description
5	**6.8M**	**31.0%**	44 yrs / 42 wks	**High-Income Professionals**
1	62.7M	13.4%	40 yrs / 46 wks	Working Adults - High Reach
4	62.0M	12.0%	40 yrs / 45 wks	Working Adults - Similar to Seg 1
6	23.6M	8.4%	34 yrs / 44 wks	Mid-Tier Emerging Workers
3	19.2M	4.9%	37 yrs / 46 wks	Lower-Income Working Adults
0	39.7M	1.8%	56 yrs / 3 wks	Older/Low Work Attachment
2	39.3M	1.6%	56 yrs / 3 wks	Older/Low Work Attachment

5.3 Detailed Segment Profiles

Segment 5: High-Income Professionals (PREMIUM TARGET)

Profile:

- **Income Rate:** 31.0% earn >$50K (highest by far)
- **Size:** 6.8M (small but high-value)
- **Demographics:** Mid-40s age, strong work attachment (42 wks/year)
- **Occupation:** Professional specialty, executive/managerial roles
- **Education:** Bachelor's degree, high school graduate mix
- **Financial:** Strong capital gains signal

Marketing Strategy:

- **Priority:** Highest precision targeting for premium products
- **Products:** Premium electronics, luxury goods, investment products, exclusive memberships
- **Messaging:** Quality, exclusivity, investment value, professional benefits
- **Channels:** Premium credit card offers, executive programs, private events
- **Expected ROI:** Highest conversion rate but limited reach

Segments 1 & 4: Working Adults (HIGH REACH + GOOD VALUE)

Profile:

- **Income Rate:** 13.4% and 12.0% respectively
- **Size:** 62.7M and 62.0M (massive reach potential)
- **Demographics:** Late-30s, nearly full-time employment (45-46 wks/year)

- **Occupation:** Administrative support, professional, managerial roles
- **Industry:** Retail trade, manufacturing, education
- ***Note:*** *Segments 1 and 4 are very similar - treat as twin segments*

Marketing Strategy:

- **Priority:** Primary volume driver - balance scale and conversion
- **Products:** Mid-tier appliances, electronics, home improvement, family packages
- **Messaging:** Value for money, family focus, career advancement, convenience
- **Channels:** Email campaigns, loyalty programs, seasonal promotions
- **Expected ROI:** Strong conversion at massive scale - core revenue driver

Segment 6: Mid-Tier Emerging Workers (GROWTH OPPORTUNITY)

Profile:

- **Income Rate:** 8.4% earn >$50K
- **Size:** 23.6M (moderate reach)
- **Demographics:** Younger (mid-30s), strong work attachment (44 wks/year)
- **Occupation:** Professional, administrative, sales roles
- **Profile:** Economically active, potentially upwardly mobile

Marketing Strategy:

- **Priority:** Growth segment - invest in long-term relationship
- **Products:** Entry-level premium, financing options, career development tools
- **Messaging:** Career growth, lifestyle improvement, future planning
- **Channels:** Digital marketing, social media, mobile-first experiences
- **Expected ROI:** Lower immediate return but high lifetime value potential

Segment 3: Lower-Income Working Adults (VALUE SEGMENT)

Profile:

- **Income Rate:** 4.9% earn >$50K (despite full work attachment)
- **Size:** 19.2M
- **Demographics:** Mid-30s, high weeks worked (46 wks/year)
- **Occupation:** Clerical, service, production/craft roles
- **Key Insight:** Economically active but lower wage positions

Marketing Strategy:

- **Priority:** Value-based offerings and retention focus
- **Products:** Essentials, bulk savings, budget-friendly options, financing
- **Messaging:** Affordability, savings, practical solutions, loyalty rewards
- **Channels:** Direct mail, retail partnerships, discount programs
- **Expected ROI:** Lower margin but important for scale and retention

Segments 0 & 2: Older/Low Work Attachment (LIMITED TARGET)

Profile:

- **Income Rate:** 1.8% and 1.6% (very low)
- **Size:** 39.7M and 39.3M combined (~79M total)
- **Demographics:** Mid-50s, minimal work (3 wks/year)
- **Tax Status:** 'Nonfiler' predominant

- **Profile:** Retirees, economically inactive, low labor force participation

Marketing Strategy:

- **Priority:** Low priority for income-based targeting
- **Products:** Essential services, health products, senior-focused offerings
- **Messaging:** Simplicity, affordability, stability, convenience
- **Channels:** Traditional media, community partnerships, low-cost acquisition
- **Expected ROI:** Very low conversion; deprioritize unless specialized products

6. Business Recommendations

Recommended Production Model

Deploy **LightGBM** for production use based on:

- Best overall performance: 95.5% ROC-AUC, 70.0% PR-AUC
- Excellent precision-recall balance (62.1% precision, 65.5% recall at optimal threshold)
- Fast inference speed and memory efficiency for production deployment
- Robust to production edge cases with built-in regularization

Threshold Configuration

Use optimized threshold of **0.85** (vs default 0.50) to maximize F1-score. This threshold achieves:

- 62.1% precision: 621 out of 1000 predicted high-earners will actually earn >$50K
- 65.5% recall: Captures nearly 2/3 of actual high-earners in the population
- Optimal balance between marketing efficiency (precision) and market coverage (recall)

7. References

The following resources were consulted during project development:

I. *Scikit-learn Documentation.* Machine Learning in Python. https://scikit-learn.org
II. *XGBoost Documentation.* Scalable and Flexible Gradient Boosting. https://xgboost.readthedocs.io
III. *LightGBM Documentation.* Light Gradient Boosting Machine. https://lightgbm.readthedocs.io
IV. *Lundberg, S. M., & Lee, S. I. (2017).* A Unified Approach to Interpreting Model Predictions. NIPS 2017.
V. *Chawla, N. V., et al. (2002).* SMOTE: Synthetic Minority Over-sampling Technique. JAIR 16, 321-357.
VI. *U.S. Census Bureau.* Current Population Survey Technical Documentation. https://www.census.gov
VII. *Pandas Documentation.* Data Analysis Library. https://pandas.pydata.org
VIII. *SHAP Documentation.* SHapley Additive exPlanations. https://shap.readthedocs.io