Income Classification from CPS-Derived Census-Bureau Data

Technical Report for Marketing-Risk Use Case

Raghu Ram Sattanapalle

November 3, 2025

1 Data Understanding and Exploration

1.1 Dataset Overview

The analysis utilized the Census Income dataset, extracted from the United States Census Bureau database spanning the years 1994–1995. The raw dataset comprised 199,523 observations across 43 columns, encompassing both demographic and socioeconomic attributes. According to the data source documentation, the dataset exhibits substantial class imbalance, with approximately 93.8% of observations corresponding to incomes below \$50,000 and 6.2% at or above this threshold (Figure 1).

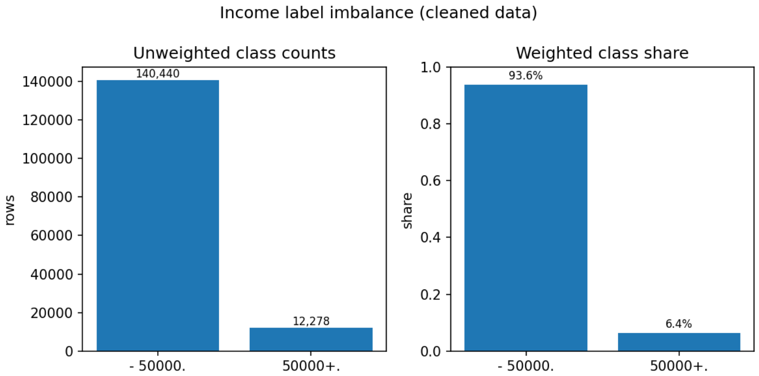

Figure 1: Class distribution showing severe imbalance: 93.8% negative class (¡ \$50K) vs 6.2% positive class ($\geq$ \$50K). This imbalance motivates the use of PR-AUC over accuracy as the primary evaluation metric.

The feature set includes 13 numeric attributes and 29 categorical attributes. Numeric features include continuous variables such as age, weeks worked in year, capital gains, capital losses, dividends from stocks, and wage per hour. Categorical features span demographic characteristics (education, marital status, sex, race, citizenship), employment attributes (class of worker, major industry code, major occupation code), and geographic/migration variables. Critically, the dataset includes a `weight` column representing survey sampling weights, which must be incorporated in all statistical analyses to ensure proper population-level inference.

1.2 Data Quality Assessment and Preprocessing

1.2.1 Missing Value Analysis

Initial inspection revealed two primary forms of data incompleteness. First, eight columns contained explicit missing value indicators coded as "?": migration code-change in msa, migration code-change in reg, migration code-move within reg, migration prev res in sunbelt, country of birth father, country of birth mother, country of birth self, and state of previous residence. These were systematically converted to proper NaN values for consistent handling.

Second, numerous columns contained "Not in universe" categorical values, indicating that a particular attribute was not applicable to certain population segments (e.g., labor force statistics for children, or unemployment reasons for employed individuals). Contrary to standard missing data imputation approaches, these values were **retained as distinct categories** rather than treated as missing data. This decision reflects the observation that "Not in universe" status is informationally significant—for instance, individuals not in the labor force (class of worker = "Not in universe") exhibit a weighted positive rate of only 0.9%, compared to 36.3% for self-employed incorporated individuals (Figure 3).

1.2.2 Duplicate and Conflict Resolution

A systematic integrity check revealed 53,499 duplicate records—observations with identical feature values appearing multiple times in the dataset. Additionally, 379 **conflicting records** were identified: instances where identical feature combinations mapped to different income labels. This inconsistency presents a fundamental challenge for supervised learning, as it creates irresolvable contradictions in the training data.

The preprocessing pipeline implemented the following protocol:

1. **Conflict removal**: All 379 conflicting records (both instances of each conflicting feature combination) were excluded from the modeling dataset. This conservative approach prioritizes data quality and model reliability over marginal sample size gains.

2. **Duplicate aggregation**: For the remaining duplicates with consistent labels, records were aggregated by summing their associated survey weights while retaining the common label. This preserves population representation while reducing computational overhead.

Following these operations, the cleaned dataset comprised 152,718 unique feature-label combinations, representing a 23.5% reduction in row count while maintaining population coverage through weight preservation.

1.3 Exploratory Analysis: Weighted Univariate Associations

All exploratory analyses incorporated survey weights to produce population-representative statistics rather than sample-based estimates. For each categorical variable, weighted positive rates—the proportion of population-weighted observations with income $\geq$ \$50,000—were computed across categories, focusing on the top 20 categories by weighted population share for interpretability.

1.3.1 Education

Education demonstrated the strongest and most consistent gradient with income outcomes. The weighted analysis revealed a stark educational divide: professional school degrees (MD, DDS, JD) showed the highest positive rate at 53.9%, followed by doctorate degrees at 52.6%. Master's degrees exhibited 31.6%, while bachelor's degrees showed 20.3%. High school graduates showed only 3.8%, and the "Children" category exhibited 0.0% (Figure 2). This 50+ percentage point gradient underscores education as the primary determinant of earning capacity.

1.3.2 Employment Characteristics

The class of worker variable revealed substantial heterogeneity. Self-employed incorporated individuals exhibited the highest positive rate at 36.3%, more than three times the rate of private sector employees (10.2%). Government workers showed intermediate rates (federal: 21.1%, state: 11.7%, local: 11.2%). Categories indicating labor force non-participation ("Not in universe": 0.9%, "Never worked": 0.3%, "Without pay": 0.2%) demonstrated near-zero rates, confirming that labor force participation is necessary for substantial earnings (Figure 3).

Major occupation codes revealed a clear white-collar premium (Figure 4). Executive/managerial occupations led at 29.1%, followed by professional specialty at 25.4%. Manual and service occupations showed substantially lower probabilities (handlers: 1.8%, other service: 1.0%, private household services: 0.3%). This occupational stratification operates independently of education.

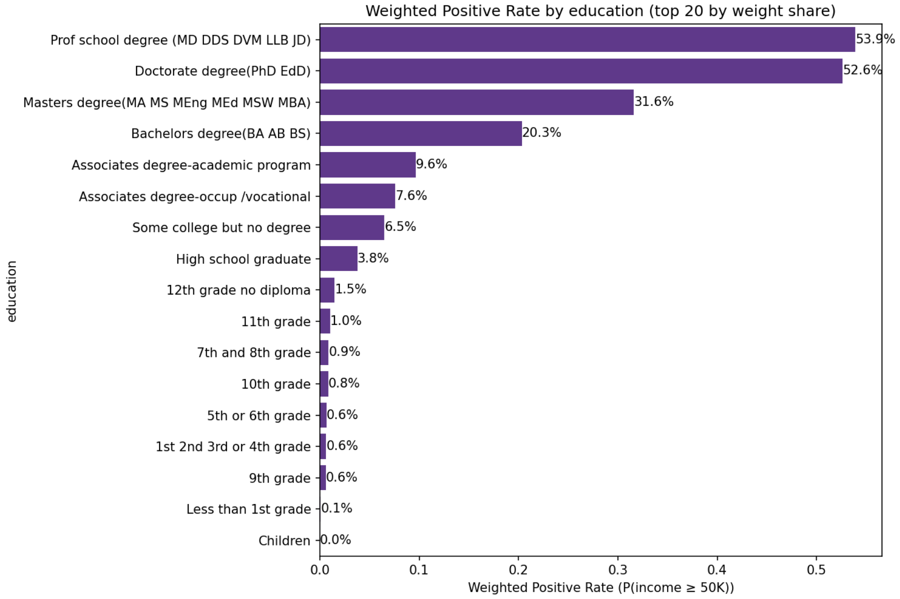

Figure 2: Weighted positive rates by education level (top 20 categories by population share).

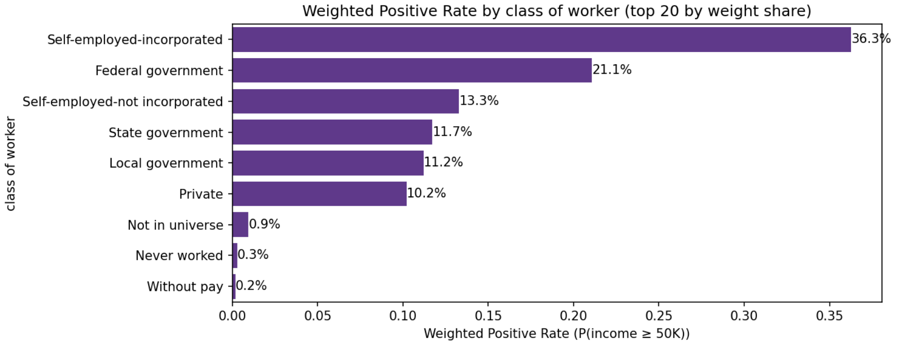

Figure 3: Weighted positive rates by class of worker.

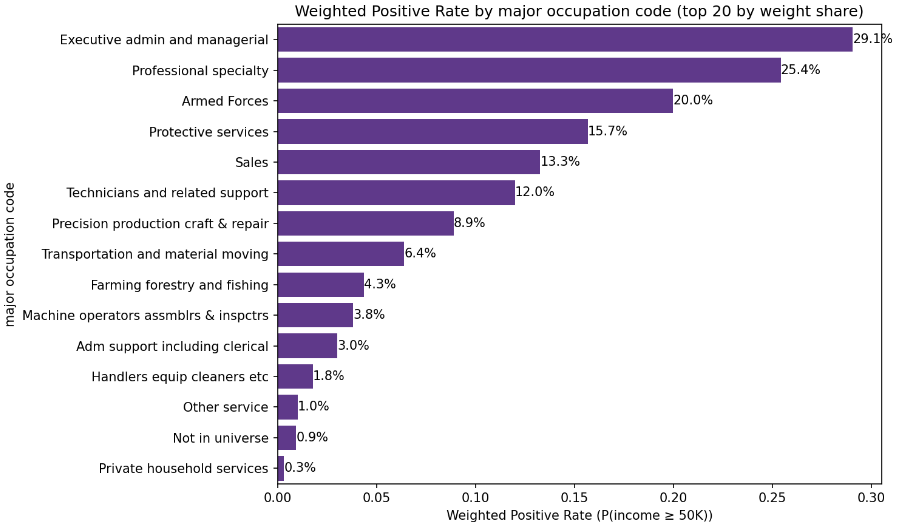

Figure 4: Weighted positive rates by major occupation code.

Additional weighted analyses for major industry codes, marital status, citizenship, and race revealed similar stratification patterns. Industry analysis showed professional services (23.6%) and communications (22.9%) at the high end versus retail trade (4.5%) and personal services (3.7%) at the low end.

Marital status confirmed married individuals with civilian spouse present (11.9%) earned substantially more than never-married individuals (1.4%).

1.3.3 Demographic Patterns and Fairness Considerations

Sex exhibited a substantial disparity: males showed a weighted positive rate of 10.3%, nearly four times the female rate of 2.7% (Figure 5). This 7.6 percentage point gap reflects well-documented gender wage disparities in the 1994–1995 labor market, encompassing both occupational segregation and within-occupation pay differences. Race analysis revealed similar disparities (Asian/Pacific Islander: 7.6%, White: 7.0%, Black: 2.5%), necessitating careful fairness auditing for deployment.

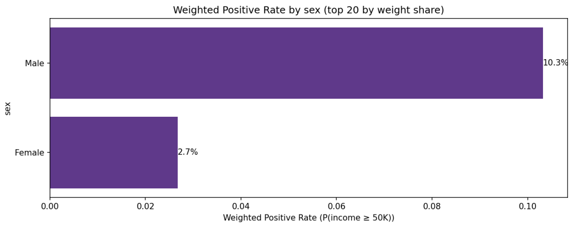

Figure 5: Weighted positive rates by sex, revealing a 4× gender gap.

1.3.4 Investment and Capital Income

Numeric features related to capital income—capital gains, capital losses, and dividends from stocks—exhibited extreme right-skewed distributions, with the vast majority of observations recording zero values. However, the presence of *any* non-zero capital or dividend income served as a strong positive indicator. This suggests these variables function as binary flags for investor status or wealth accumulation rather than capturing incremental income effects.

1.4 Multivariate Relationships

Weighted Pearson correlation analysis between numeric features and the binary income target confirmed that weeks worked in year, dividends from stocks, and age ranked among the strongest linear associations. Spearman rank correlation among numeric features identified moderate relationships (e.g., age and weeks worked: $\rho = 0.29$), with work-related variables showing substantial intercorrelation (weeks worked exhibited correlations above 0.7 with detailed industry, detailed occupation, and num persons worked for employer).

The univariate analyses collectively suggest that high-income outcomes are associated with the simultaneous presence of multiple favorable characteristics: advanced educational attainment, white-collar occupations, professional industries, and full-year employment. This pattern of co-occurring predictive features motivates the use of tree-based models capable of capturing interaction effects, rather than purely additive linear approaches that assume independent feature contributions.

1.5 Implications for Modeling

The exploratory analysis yielded several key insights that informed subsequent modeling decisions:

- **Class imbalance**: The severe imbalance (93.8% negative class) necessitates evaluation metrics robust to imbalance, specifically precision-recall AUC rather than accuracy or standard AUC-ROC, which can be misleadingly optimistic for imbalanced datasets.

- **Survey weights**: All model training and evaluation must incorporate sample weights to ensure population-level generalizability rather than sample-specific fit.

- **Categorical cardinality**: Features such as detailed industry recode (52 levels) and country of birth (43 levels) require modeling approaches that handle high-cardinality categoricals efficiently.

One-hot encoding would generate over 200 binary features, creating computational burden and potential overfitting.

- **Non-linear interactions**: The univariate analyses reveal that favorable characteristics (high education, white-collar occupation, full-year work) co-occur in high-income cases, suggesting gradient boosting or tree ensemble methods will substantially outperform linear models by capturing these synergistic effects.

- **"Not in universe" informativeness**: Retaining NIU categories as distinct values rather than treating them as missing data is empirically justified. For instance, "Not in universe" in class of worker (0.9% positive rate) is far more informative than a missing value would be, clearly signaling labor force non-participation.

- **Fairness considerations**: The substantial disparities observed across sex (10.3% vs 2.7%) and race (7.6% to 2.5% range) indicate that models trained on this data will encode historical inequities. Deployment requires fairness audits and consideration of whether such disparities should be perpetuated or mitigated through debiasing techniques.

2 Model Development and Evaluation

2.1 Baseline Model: Logistic Regression

To establish a transparent performance benchmark, we initially implemented a regularized logistic regression model. This baseline served dual purposes: providing interpretable feature coefficients for validation against exploratory findings, and quantifying the performance gains achievable through more sophisticated modeling approaches.

The preprocessing pipeline comprised one-hot encoding for categorical features and standardization (zero mean, unit variance) for numeric features via `StandardScaler`. The resulting feature matrix contained over 200 dimensions due to the high cardinality of categorical variables. An L2-regularized logistic regression classifier was trained on this encoded representation, with sample weights incorporated through the `sample_weight` parameter to maintain population-level representativeness.

Performance on the 20% holdout validation set yielded a weighted PR-AUC of 0.601 and weighted ROC-AUC of 0.940. While these metrics appear strong superficially, the PR-AUC indicates substantial room for improvement given the class imbalance.

Coefficient analysis validated the exploratory findings: education level, weeks worked in year, and capital income variables exhibited the strongest positive associations with high income. Notably, the coefficient for `sex_Female` was negative, reflecting the historical gender wage gap documented in Section 2. This baseline confirms that standard linear approaches can capture first-order effects but may struggle with high-cardinality categoricals and non-linear interactions.

2.2 Primary Model: CatBoost Gradient Boosting

2.2.1 Algorithm Selection and Rationale

Based on the exploratory analysis findings, particularly the high-cardinality categorical features, evidence of feature interactions, and the need for sample weight support, we selected CatBoost (Categorical Boosting) as the primary modeling framework. This choice reflects several technical and practical considerations:

- **Native categorical handling**: CatBoost processes categorical features directly without requiring one-hot encoding, avoiding the dimensionality explosion that would occur with 52-level industry codes and 43-level country-of-birth features. The algorithm employs ordered target statistics and random permutations to compute categorical feature splits while mitigating target leakage.

- **Sample weight integration**: CatBoost natively supports instance weights through the `Pool` data structure, enabling proper population-weighted training and evaluation essential for survey data.

- **Ordered boosting**: CatBoost's ordered boosting mechanism reduces overfitting compared to standard gradient boosting by using different random permutations of the training data when computing residuals, addressing the prediction shift problem inherent in classical boosting.
- **Missing data robustness**: The algorithm handles missing values internally without requiring imputation, treating them as informative signals during tree construction.

2.2.2 Model Architecture and Training

The final CatBoost architecture employed the following hyperparameters:

- **Tree depth**: 7 (balancing model complexity and interpretability)
- **Learning rate**: 0.045 (moderate step size for stable convergence)
- **L2 leaf regularization**: 3.0 (penalizes extreme leaf values)
- **Iterations**: 3,000 maximum (with early stopping)
- **Early stopping rounds**: 200 (halts training if no improvement over 200 iterations)
- **Loss function**: Logloss (binary cross-entropy)
- **Evaluation metric**: PRAUC (precision-recall AUC, aligned with business objective)

Categorical features were identified by object dtype and passed to CatBoost via the `cat_features` parameter. Training and validation `Pool` objects incorporated sample weights to ensure all optimization and evaluation remained population-representative.

2.2.3 Evaluation Strategy and Results

Model performance was assessed through a multi-tiered validation approach to ensure robustness and generalizability:

Single Holdout Validation. Initial training on an 80% stratified split (maintaining class balance) with 20% holdout validation yielded a weighted PR-AUC of 0.6887 and weighted ROC-AUC of 0.9519. The model converged at iteration 2,047 based on the early stopping criterion, indicating stable learning dynamics.

5-Fold Cross-Validation. To quantify performance variance and assess generalization, we conducted stratified 5-fold cross-validation with fixed random seeds for reproducibility. Results across folds:

- Fold 1: weighted PR-AUC = 0.7023
- Fold 2: weighted PR-AUC = 0.6947
- Fold 3: weighted PR-AUC = 0.6976
- Fold 4: weighted PR-AUC = 0.6945
- Fold 5: weighted PR-AUC = 0.6976

The mean weighted PR-AUC of **0.6973 ± 0.0032** demonstrates consistent performance across data partitions, with low standard deviation indicating a stable, generalizable model. This represents a **16% relative improvement** over the logistic regression baseline (0.601 vs 0.697).

Out-of-Fold Predictions. Aggregating predictions across all validation folds into a single out-of-fold (OOF) prediction set, where each observation is predicted by a model that never saw it during training, yielded an OOF weighted PR-AUC of 0.6969 and weighted ROC-AUC of 0.9556. The close alignment between mean CV score (0.6973) and OOF score (0.6969) confirms the absence of overfitting.

2.2.4 Hyperparameter Tuning

Systematic hyperparameter search was conducted over a grid spanning:

- Tree depth: $\{6, 7, 8\}$
- Learning rate: $\{0.03, 0.045\}$
- L2 leaf regularization: $\{3, 5, 7\}$

Using fixed-fold cross-validation to ensure fair comparison across configurations, the original parameters (depth=7, learning_rate=0.045, l2_leaf_reg=3) emerged as optimal, with no configuration achieving statistically significant improvement. An additional experiment testing normalized "Not in universe" representations yielded no performance gain, confirming that retaining NIU as distinct categories was the correct preprocessing decision.

2.2.5 Feature Importance Analysis

CatBoost's feature importance analysis, computed via gain-based attribution, revealed the relative contribution of each feature to model predictions (Figure 6). The top five predictors align closely with the exploratory univariate analyses:

1. **Family members under 18**: Highest importance, capturing household structure and lifecycle stage
2. **Age**: Reflects earnings lifecycle patterns
3. **Num persons worked for employer**: Proxy for employer size and job stability
4. **Dividends from stocks**: Captures investor status and wealth effects
5. **Education**: Confirms educational attainment as a primary driver

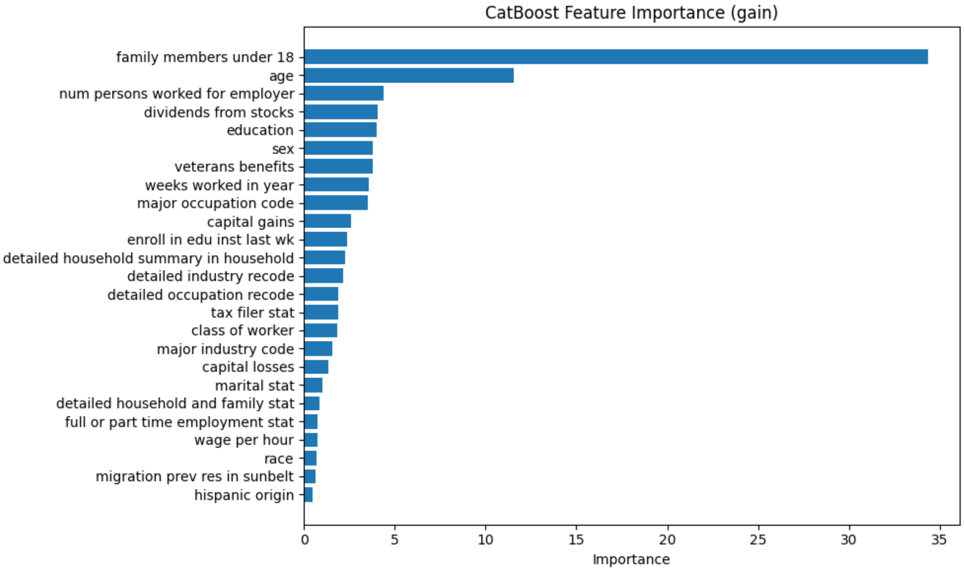

Figure 6: CatBoost feature importance (top 25 features by gain). The model prioritizes socioeconomic fundamentals (household structure, age, employment characteristics, education) over administrative variables (migration codes, year), indicating focus on substantive rather than artifactual patterns.

This importance ranking avoids the spurious one-hot artifacts observed in the logistic baseline (e.g., rare country-of-birth categories receiving inflated coefficients due to sample size effects). The model's emphasis on household structure, human capital, and employment variables over administrative or geographic features suggests it has learned generalizable socioeconomic patterns rather than dataset-specific artifacts.

2.3 Classification Threshold Optimization

For deployment in imbalanced settings, the default classification threshold of 0.5 is suboptimal. We identified the threshold maximizing weighted F1-score (harmonic mean of weighted precision and recall) using precision-recall curves on sample-weighted validation predictions.

The optimal threshold of **0.3513** achieves a weighted F1-score of 0.6207. This lower-than-default threshold reflects the need to maintain reasonable recall on the minority class. In production, threshold calibration should align with business-specific utility functions (e.g., marketing campaigns may tolerate higher false positive rates to maximize coverage, while credit applications require higher precision).

2.4 Model Comparison and Selection

Table 1 summarizes the performance of both modeling approaches across key metrics.

Table 1: Model Performance Comparison on Validation Data

Model	Weighted PR-AUC	Weighted ROC-AUC	CV Std Dev
Logistic Regression	0.601	0.940	—
CatBoost (single split)	0.689	0.952	—
CatBoost (5-fold CV)	0.697 ± 0.003	0.956	0.0032

The CatBoost model demonstrates clear superiority across all metrics, with robust performance validated through cross-validation. The 16% relative improvement in weighted PR-AUC (0.601 to 0.697) translates to substantially better precision-recall tradeoffs in the imbalanced regime. The low variance across folds ($\sigma = 0.0032$) and tight alignment between cross-validation (0.6973) and out-of-fold (0.6969) scores confirm generalization capability and absence of overfitting.

Beyond quantitative performance, CatBoost offers practical deployment advantages: native categorical handling eliminates preprocessing complexity, built-in missing value support reduces data quality sensitivity, and interpretable feature importance facilitates model validation and stakeholder communication. We therefore recommend CatBoost as the production model, with the understanding that deployment requires careful fairness auditing given the historical biases encoded in the 1994–1995 training data.

3 Business Segmentation and Targeting Strategy

3.1 Motivation for Model-Aligned Segmentation

While the CatBoost classifier provides individual-level income predictions, business applications typically require actionable population segments. Standard unsupervised clustering is suboptimal here: the population contains substantial non-working segments (children, retirees), causing K-means to allocate clusters to non-target groups rather than high-value prospects.

To address this limitation, we implemented **model-aligned segmentation**: a hybrid approach that incorporates the trained CatBoost model's predictions as a clustering feature alongside interpretable demographic and employment attributes. This strategy ensures that at least one segment captures the high-income-probability population while maintaining interpretability through business-relevant features.

3.2 Segmentation Methodology

The segmentation pipeline comprised three stages:

Stage 1: Score Generation. The full cleaned dataset (152,718 observations) was scored using the trained CatBoost model to generate predicted probabilities of income $\geq$ \$50,000. These scores serve as a learned summary of each individual's income potential based on the full feature set.

Stage 2: Feature Engineering. Business-interpretable binary flags were constructed:

- `worked_full_year`: 1 if weeks worked ≥ 50, else 0
- `worked_some`: 1 if weeks worked > 0, else 0
- `high_edu`: 1 if bachelor's degree or above (BA, master's, professional, doctorate)
- `is_married`: 1 if marital status contains "Married"
- `is_public_sector`: 1 if employed by federal/state/local government
- `any_invest_inc`: 1 if capital gains > 0 or dividends > 0

Stage 3: K-Means Clustering. A 10-dimensional feature matrix was constructed comprising the model score, the six binary flags, and three numeric attributes (age, weeks worked in year, family members under 18). This matrix was standardized (zero mean, unit variance) and subjected to K-means clustering with $K = 5$ clusters, using 20 random initializations to avoid local minima (`n_init=20`, `random_state=42`).

The inclusion of the model score as a clustering feature ensures that high-income-probability individuals concentrate in a single segment, while the interpretable features enable straightforward segment characterization for business stakeholders.

3.3 Segment Profiles and Characteristics

Weighted segment statistics (computed using census survey weights) reveal five distinct population groups, ordered by descending model score (Table 2).

Table 2: Weighted Segment Profiles (Population-Representative Statistics)

Seg.	Profile	Pop %	Avg Score	Income Rate	Full-Year	High Edu	Invest Inc
3	High-value	7.1%	0.56	57.4%	88.9%	81.5%	73.5%
0	Public sector	6.9%	0.09	8.5%	75.8%	37.6%	16.3%
1	Working class	30.9%	0.05	4.7%	83.7%	12.5%	10.0%
2	Retired/inactive	18.3%	0.01	1.2%	0.0%	12.3%	19.2%
4	Dependents	36.7%	0.001	0.1%	0.0%	1.7%	1.1%

Segment Characterizations. **Segment 3 (High-Value, 7.1%)**: 57.4% income rate aligned with 0.56 model score. Characteristics: 81.5% bachelor's+ degree, 88.9% full-year work, 73.5% investment income, avg age 45. **Primary targeting** for premium products and credit.

Segment 0 (Public Sector, 6.9%): 8.5% income rate, 75.8% full-year, 37.6% high education. Concentrated government/education workers. **Secondary target** for stability-focused products.

Segment 1 (Working Class, 30.9%): 4.7% income rate, 83.7% full-year work but 12.5% high education. **Low priority** for high-value targeting; suitable for reach campaigns.

Segments 2 & 4 (Retired/Dependents, 55%): Minimal work participation and income (¡1.2%). **Exclude** from employment-based targeting.

3.4 Business Recommendations

The segmentation analysis yields three concrete strategic recommendations:

1. **Prioritize Segment 3 for high-value targeting.** With a 7.1% population share but 57.4% income rate, nine times the overall 6.2% baseline, this segment offers exceptional ROI for resource allocation. The 81.5% high-education rate and 73.5% investment income presence indicate receptiveness to premium financial products, advanced educational services, and wealth management offerings.

2. **Develop tier-appropriate products for Segment 0.** The public sector professional segment exhibits stable employment and moderate income, suggesting potential for products emphasizing security over high returns: government-backed savings vehicles, moderate-premium insurance, and educational programs for career advancement.

3. **Calibrate marketing spend based on segment size vs. conversion rate tradeoffs.** While Segment 1 represents 30.9% of the population (4.3× larger than Segment 3), its 4.7% income rate is 12× lower. For campaigns targeting high-income individuals, concentrating resources on Segment 3 yields superior precision, while broad-reach campaigns may justify inclusion of Segment 1 despite lower per-capita returns.

3.5 Segmentation Validation

The segmentation validates EDA findings: Segment 3's 81.5% high-education and 88.9% full-year work rates align with univariate predictors identified in Section 1, while the co-occurrence of favorable characteristics confirms multivariate interactions. The close alignment between model score (0.56) and actual income rate (57.4%) confirms calibration and demonstrates successful population stratification along the income axis.

4 Conclusion

This analysis developed an income classification system achieving 0.697 weighted PR-AUC, representing a 16% improvement over logistic regression baselines. The CatBoost model's feature importance rankings (family structure, age, employment characteristics, education) align with exploratory analyses, confirming that the model learned generalizable socioeconomic patterns rather than dataset artifacts.

The model-aligned segmentation identified a high-value target segment (7.1% of population, 57.4% income rate) characterized by advanced education, full-year employment, and investment income. This segment offers 9× higher conversion rates than the population baseline, enabling precision-targeted resource allocation.

4.1 Limitations and Future Directions

While the model demonstrates strong performance, several limitations warrant consideration. The 30-year temporal gap limits direct applicability; labor market dynamics and wage structures have evolved substantially, requiring retraining on contemporary data. The model encodes historical disparities (male income rate 4× female, 7+ point racial gaps), and deployment requires fairness-aware techniques (reweighting, adversarial debiasing) to avoid perpetuating inequities. Additional feature engineering (education×occupation interactions, geographic adjustments, industry cycles) may enhance predictive power.

Successful deployment requires threshold calibration for specific business contexts (0.35 optimizes F1), fairness auditing of protected group outcomes, and periodic retraining as labor markets evolve. With appropriate monitoring and fairness controls, this system provides actionable intelligence for marketing, credit risk, and strategic planning applications.

References

[1] CatBoost Development Team. (n.d.). CatBoostClassifier — Python API Reference. Retrieved from https://catboost.ai/docs/en/concepts/python-reference_catboostclassifier

[2] IPUMS CPS. (n.d.). FAQ: What does ”universe” mean in the variable descriptions? Retrieved from https://cps.ipums.org/cps-action/faq.do#ques8