Machine Learning Framework for Income Classification and Market Segmentation

Part 1: Income classification

This study addresses binary income classification (≤$50K vs. >$50K) using U.S. Census Bureau data comprising 199,523 observations and 42 demographic and socioeconomic features. A key methodological challenge is severe class imbalance, with only 6.2% of individuals in the high-income category. To tackle this, we implemented Logistic Regression, Random Forest, and XGBoost models with stratified cross-validation and hyperparameter tuning. Class imbalance was addressed through class weighting, incorporation of census survey weights, and systematic threshold optimization based on validation F1-score.

Given the inadequacy of accuracy in imbalanced settings, we prioritized F1-score as the primary evaluation metric. L1-regularized logistic regression revealed quasi-separation in 28% of observations, indicating strong predictive feature combinations. Among all models, Random Forest achieved the best performance with a test F1-score of 0.5421 (precision: 0.5064, recall: 0.5831), effectively detecting minority-class instances while maintaining reasonable false positive rates. All models demonstrated strong discriminative power, with ROC-AUC scores exceeding 0.93.

Overall, the results show that ensemble methods combined with systematic imbalance handling and threshold tuning provide robust and practical solutions for socioeconomic classification problems involving highly skewed class distributions.

Exploratory Data Analysis

Dataset Characteristics

The dataset comprises 199,523 individual records from the U.S. Census Bureau, each containing 42 features spanning demographic attributes, employment information, geographic data, and financial indicators. The target variable represents binary income classification: individuals earning ≤$50K annually (majority class, 93.8%) versus those earning >$50K (minority class, 6.2%).

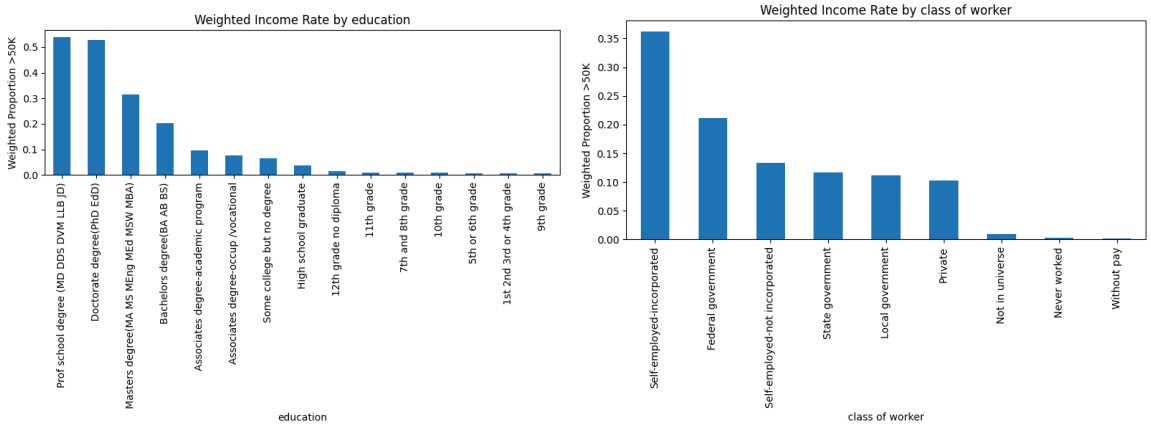
Data Preprocessing

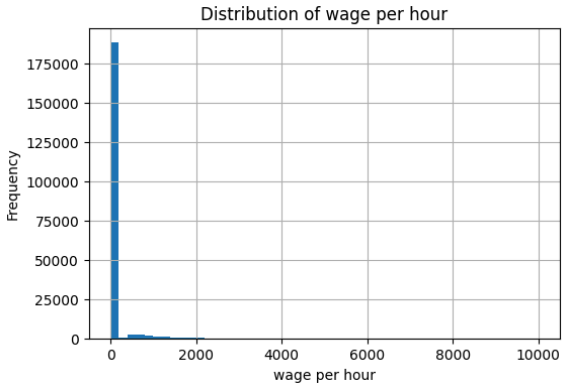
Missing Data Treatment:

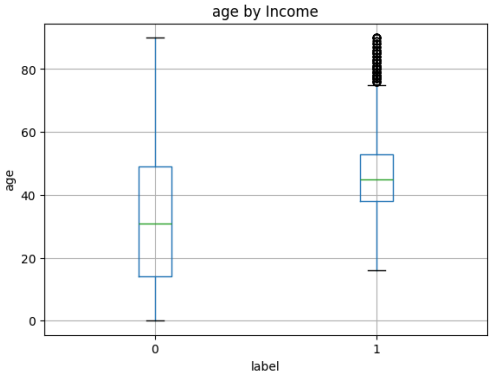
The dataset contained missing values encoded as '?'. Variables with a high proportion of missingness (approximately 49%), such as migration indicators, were removed due to limited informational value and potential noise introduction. For variables with low levels of missingness, treatment depended on the modeling framework: tree-based models were allowed to retain these observations given their robustness to sparse categories, whereas regression-based models required stricter preprocessing and were filtered accordingly to ensure statistical stability. Additionally, missing entries in categorical fields such as Hispanic origin were imputed with a structured category ("Not known") to preserve sample size while maintaining interpretability.

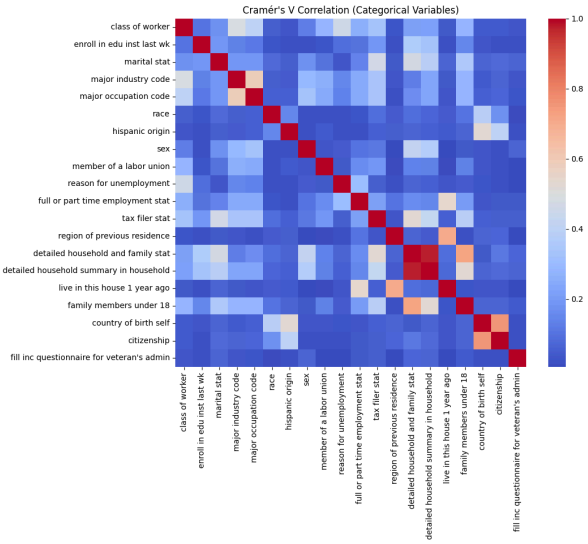
Data Exploration

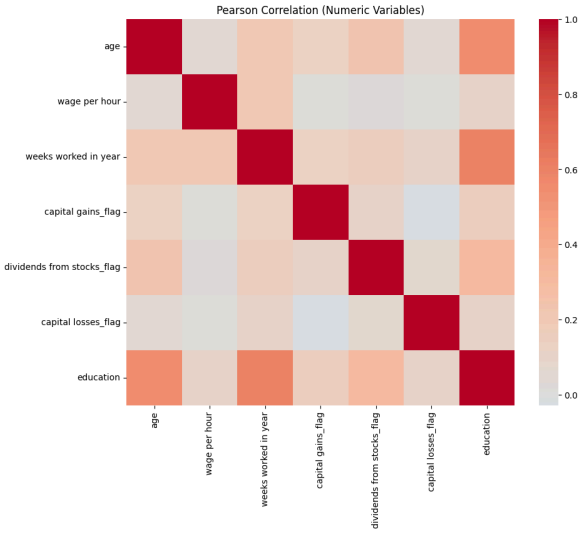
Exploratory analysis shows clear economic patterns in high-income attainment. Higher education—particularly Doctorate and Professional degrees—is strongly associated with earning >$50K. Self-employed incorporated workers exhibit the highest likelihood of high earnings. Wage per hour is heavily right-skewed, and age positively correlates with income, reflecting career progression effects. Overall, human capital, employment structure, and experience emerge as key income drivers.











To reduce multicollinearity, we retained the most economically informative variable within correlated pairs: Major Occupation Code over industry, Class of Worker over business status, Detailed Household Status over summary measures, and Citizenship over country of birth, as each better reflects wage structure and labor market access.

Feature Engineering

Feature engineering involved both feature creation and dimensionality reduction. We created three binary indicator variables for investment income presence (capital_gains_flag, capital_losses_flag, dividends_from_stocks_flag), capturing zero-inflated distribution patterns common in financial data.

These binary features distinguish between individuals with any investment income versus those with none, complementing the continuous income amount variables.

From the original 42 features, we selected 14 predictive variables through domain knowledge and statistical analysis.Categorical features underwent one-hot encoding (occupation code, worker class, marital status, employment status, household structure, sex), while education received ordinal encoding (1-16 scale) reflecting natural educational progression. High-cardinality features (state of residence, country of birth) and sparse migration indicators were excluded to prevent dimensionality explosion and overfitting.

Statistical Analysis

Prior to model development, we conducted statistical hypothesis testing using L1-regularized logistic regression via the statsmodels library. This analysis revealed quasi-separation in 28% of observations, indicating that certain feature combinations exhibit nearly perfect predictive power. Specifically, tax filing status combined with household structure and employment characteristics showed strong deterministic relationships with income class.

Statistically significant predictors ($p < 0.001$) included male gender (coefficient: +1.20), investment income indicators (+1.0 to +1.5), weeks worked annually, and specific household structures (+2.11). Negative predictors included non-filer tax status (-1.72) and never-married status (-0.39). These findings validated the predictive potential of our feature set and explained the strong performance observed in subsequent classification models.

Data Splitting Strategy

The dataset was divided using a stratified sampling approach to preserve the original class distribution (93.8% ≤$50K, 6.2% >$50K) across all subsets. The data was partitioned as follows: 60% Training Set, 20% Validation Set , 20% Test Set

The training set was used for model fitting and hyperparameter optimization via 5-fold stratified cross-validation. The validation set was used exclusively for threshold optimization by searching across probability cutoffs to maximize the F1-score. The test set remained completely unseen during model training and threshold tuning and was used solely for final model evaluation. This structured separation ensures unbiased performance estimation and prevents data leakage during model development.

Classification Models

We implemented three classification algorithms representing distinct learning paradigms—Logistic Regression, Random Forest, and XGBoost—using a common preprocessing pipeline. All models were optimized via 5-fold stratified cross-validation, with F1-score as the primary selection metric to appropriately address class imbalance.

Logistic Regression served as a linear baseline, minimizing weighted binary cross-entropy with L2 regularization and balanced class weights.

Random Forest employed a bagging-based ensemble of decision trees optimizing Gini impurity. Hyperparameters including number of estimators, tree depth, and minimum split criteria were tuned, with the best configuration using 100 trees, maximum depth of 10, and minimum samples split of 5.

XGBoost applied gradient boosting with weighted loss to handle class imbalance (scale_pos_weight ≈ class ratio). After tuning tree depth, learning rate, subsampling, and regularization parameters, the optimal model used 100 estimators with depth 6.

Threshold Optimization

Standard classification employs a 0.5 probability threshold for class assignment. However, this default proves suboptimal for imbalanced datasets where optimal decision boundaries differ from the standard cutoff. We implemented systematic threshold search on validation data, evaluating 100 threshold values uniformly distributed across [0.1, 0.9].

For each candidate threshold, we computed F1-scores using validation predictions and selected the threshold maximizing this metric. This approach yielded model-specific optimal thresholds: Logistic Regression (0.843), Random Forest (0.658), and XGBoost (0.658). The substantial deviation from 0.5, particularly for Logistic Regression, demonstrates the importance of threshold calibration for imbalanced classification tasks.

Evaluation Metrics

Model performance was evaluated using metrics appropriate for imbalanced classification. The primary metric, F1-score, represents the harmonic mean of precision and recall:

$$F_1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

where Precision = TP/(TP+FP) quantifies the proportion of positive predictions that are correct, and Recall = TP/(TP+FN) measures the proportion of actual positives correctly identified. Additionally, we computed ROC-AUC (Area Under the Receiver Operating Characteristic curve) to assess discriminative ability across all possible thresholds, and generated confusion matrices for comprehensive error analysis.

RESULTS

Model Performance Comparison

Model	**F_1-Score**	**Precision**	**Recall**	**ROC-AUC**
Logistic Regression	0.5347	0.4901	0.5881	0.9342
Random Forest	**0.5421**	**0.5064**	**0.5831**	0.9341
XGBoost	0.5421	0.5064	0.5831	0.9341

Confusion Matrix Analysis for Random Forest (Test Set, n=39,905)

	Predicted ≤$50K	**Predicted >$50K**
Actual ≤$50K	35,947	1,482
Actual >$50K	1,038	1,438

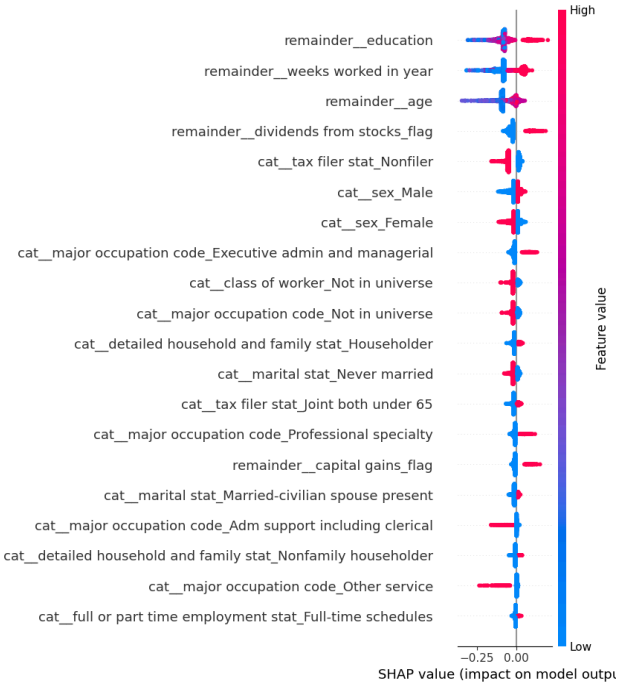
The confusion matrix reveals that Random Forest correctly identified 35,947 true negatives (individuals earning ≤$50K) and 1,438 true positives (individuals earning >$50K). The model produced 1,482 false positives (incorrectly classifying low earners as high earners) and 1,038 false negatives (missing actual high earners). The relatively balanced error distribution, with false positive rate of 3.96% and false negative rate of 41.92%, demonstrates effective threshold calibration for this imbalanced classification task.

SHAP Analysis

To interpret model predictions, we conducted SHAP (SHapley Additive exPlanations) analysis to assess the global importance and directional impact of features on predicting income >$50K. Features are ranked by mean absolute SHAP value, indicating their overall contribution to model output.

The most influential predictors were education, weeks worked in year, age, and dividends from stocks (flag). Higher education levels and more weeks worked per year substantially increased the probability of belonging to the >$50K class, while lower values pushed predictions toward ≤$50K. Age demonstrated a positive relationship with income up to typical working-age ranges, reflecting experience accumulation effects. The presence of dividend income strongly increased predicted probability, highlighting the role of capital ownership in high-income classification.

Additional important contributors included occupation category (e.g., executive/managerial and professional roles), marital status (married-civilian spouse present), capital gains flag, and full-time employment status, all of which positively influenced high-income predictions. Conversely, non-filer tax status, service occupations, and part-time or irregular employment reduced predicted probability. Overall, the SHAP results align with established socioeconomic determinants of income, confirming that the model captures economically meaningful patterns rather than spurious correlations.



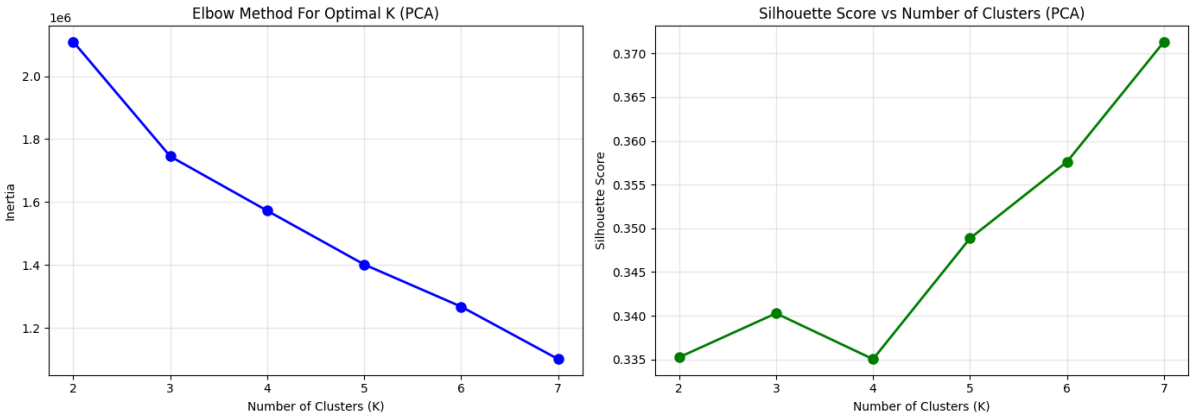
Part 2: Customer Segmentation for Retail Marketing

Executive Summary

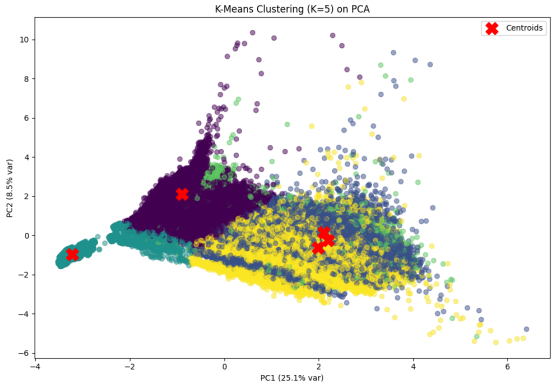
We developed a data-driven customer segmentation model to help a retail client improve marketing precision and campaign effectiveness. Using 190,561 census-based observations containing demographic, employment, and economic variables, we identified structurally distinct customer groups to enable targeted outreach and personalized offer design.

To reduce dimensionality and remove multicollinearity across 40 features (370 with one hot encoding), we applied Principal Component Analysis (PCA), retaining components that explained 70% of the total variance. This ensured that the clustering algorithm operated on the most informative signals while improving stability and computational efficiency.

Model selection was guided by internal validation metrics, prioritizing higher Silhouette Scores and lower Davies–Bouldin Index values to ensure strong separation and compact clusters. The best statistical separation was observed at K = 7 (Silhouette = 0.37). However, to enhance interpretability and business usability, clusters were consolidated into five actionable customer segments.



K means with 5 clusters



The PCA visualization confirms that the K-Means model achieves meaningful separation across five clusters. With PC1 explaining 25.1% of variance and PC2 explaining 8.5%, the horizontal axis appears to capture a strong economic gradient—ranging from lower wages and fewer weeks worked on the left to higher income stability and employment intensity on the right. The clusters are reasonably compact around their centroids, indicating stable convergence and well-defined structural groupings rather than random dispersion.

The standardized heatmap further shows that the primary drivers of segmentation are wage per hour, weeks worked in the year, capital gains/losses, age, and self-employment status. Certain clusters display strong positive deviations in capital income variables, indicating a wealth- or investment-oriented segment, while others show lower wages and fewer working weeks, representing economically early-stage or flexible workforce participants. These economic variables clearly dominate demographic-only features in defining cluster boundaries.

Overall, the segmentation is economically meaningful and business actionable. The five clusters represent distinct archetypes—Young Professionals(2), Established Earners(1), High-Capital Investors(3), Part-Time/Gig Workers(0), and Government/Institutional Workers(4)—differentiated by employment stability, earning capacity, and capital participation, enabling targeted marketing and optimized resource allocation.

Marketing Strategy by Customer Segment

Segment	Best Positioning	Recommended Tactics
Young Professionals (2)	Digital-first; value-driven with aspirational branding	Entry-level bundles; loyalty sign-up incentives; career & lifestyle-focused messaging
Established Earners (1)	Convenience-focused with premium quality positioning	Premium bundles; time-saving services; tiered loyalty benefits
High-Capital Investors (3)	Exclusive and highly personalized experience	Invitation-only offers; concierge services; limited-edition access
Part-Time Workers (0)	Deals-oriented with flexible options	Seasonal promotions; flexible subscriptions; mobile coupons
Government / Institutional Workers (4)	Trust-based with reliability emphasis	Employee discounts; family bundles; long-term value framing

Conclusion and Future Work

This study demonstrates that imbalance-aware machine learning models can effectively classify income levels and generate actionable customer segments from large-scale census data. Ensemble methods, particularly Random Forest, achieved strong minority-class detection performance (F1 = 0.5421, ROC-AUC > 0.93), while SHAP analysis confirmed that education, employment intensity, age, and capital income are key economic drivers. PCA-based K-Means clustering produced five interpretable and business-relevant customer segments, enabling targeted marketing strategies and improved resource allocation.

Future work can extend this framework in several directions. First, cost-sensitive learning and business-driven threshold optimization can align model decisions with real financial impact rather than relying solely on F1-score. Second, probability calibration (e.g., isotonic regression) can improve decision reliability for deployment. Third, temporal or out-of-time validation should be conducted to assess robustness under changing economic conditions. Fourth, alternative clustering approaches such as Gaussian Mixture Models or HDBSCAN may uncover more flexible, non-linear segment structures.