Census Income Classification and Segmentation
JPMorgan Chase Data Science Challenge

Amir Taherkhani

November 5, 2025

Contents

1 Executive Summary

This analysis addresses a fundamental question for customer targeting: how can we identify high-income individuals and understand different customer groups? We developed two complementary approaches. The first approach uses a predictive model that identifies individuals earning at least $50,000 with 93.8% accuracy. The second approach groups customers into three segments based on their life stages, enabling targeted marketing strategies for each group.

The key finding is that effective prediction depends not just on model accuracy, but on choosing the right decision threshold. While the model can correctly identify patterns in the data, the question of where to draw the line between high and low income predictions has significant practical implications. We found that the commonly used default threshold optimizes statistical measures but may not align with business objectives. Different thresholds favor either reaching more potential customers or focusing on the most certain cases.

The customer segmentation reveals three distinct groups: older adults with reduced work participation, working professionals with stable employment, and younger individuals with limited labor market attachment. These groups show dramatically different income patterns, with high earners representing 13.3% of working professionals but virtually absent from the younger group.

2 Data Understanding and Quality

The analysis began with a dataset containing approximately 200,000 records from the 1994-1995 U.S. Census Bureau Current Population Surveys. Each record represents an individual with 40 different characteristics including age, education, occupation, work patterns, and financial information. The dataset also includes population weights, which account for how the survey was designed to represent the broader U.S. population rather than just the people who happened to be surveyed.

2.1 The Challenge of Missing Information

Working with real-world survey data meant confronting how missing or non-applicable information was recorded. The dataset used several different ways to indicate when a question did not apply to someone or when information was missing. For example, fields like "industry" and "occupation" would show "Not in universe" for people who were not employed, since questions about their workplace would not make sense. Birth country information showed "?" for 3-6% of records where this information was unknown. Other fields used the code 0 to indicate the question was not applicable.

This variety of missing data indicators created a challenge. If we treated "Not in universe" as just another category of response rather than as missing data, it could mislead the model. Someone who is unemployed is fundamentally different from someone employed in retail or manufacturing. The preprocessing needed to recognize these patterns and handle them appropriately.

2.2 The Class Imbalance Problem

Perhaps the most significant challenge was the distribution of income levels in the data. Of all individuals in the dataset, 93.8% earned less than $50,000 annually, while only 6.2% earned at or above this threshold. This creates what we call a 15-to-1 imbalance ratio.

This imbalance matters because a naive approach could simply predict that everyone earns less than $50,000 and achieve 93.8% accuracy without learning anything meaningful about what distinguishes high earners. Standard accuracy metrics become misleading in this context. We need methods that specifically account for the minority class we are trying to identify.
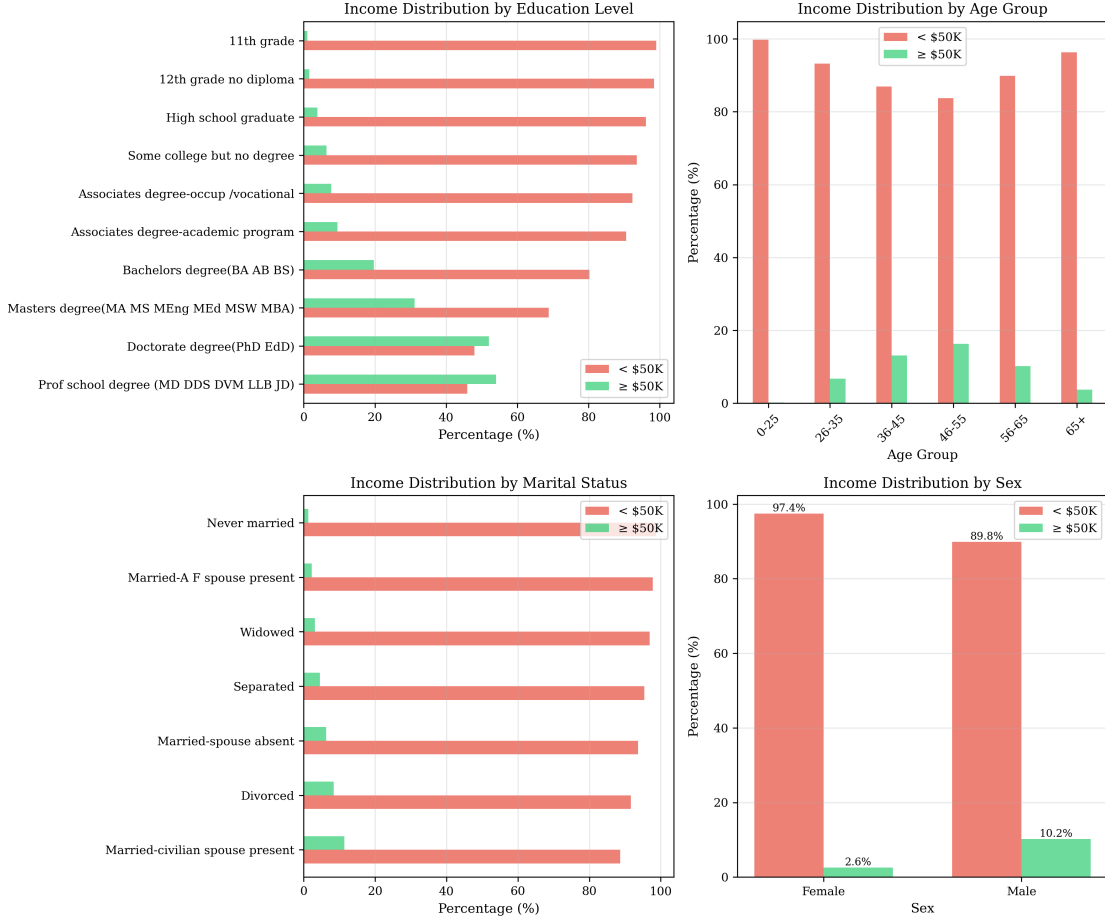
Figure 1: Income distribution across demographic factors. Education level, age, and marital status emerge as strong indicators of income potential.

2.3 Data Preparation Steps

We removed 3,229 duplicate records, representing 1.6% of the dataset. Duplicates artificially inflate model confidence by giving extra weight to identical cases.

For missing values in numeric fields like age and hours worked, we used the median value of all non-missing cases. The median is less affected by extreme values than the average, making it more reliable when dealing with skewed data. For categorical fields like education and occupation, we used the most common value. This approach preserves the dominant patterns in the data while filling gaps.

A key decision involved how to represent categorical information. Education has 17 different levels, from elementary school through doctorate degrees. Occupation has 15 categories. One common approach creates a separate binary column for each possible value, but this would generate over 200 columns for our 32 categorical features. Instead, we converted categories to numeric codes that preserve order where it exists. High school comes before bachelor's degree, which comes before master's degree. This keeps the data manageable while retaining meaningful relationships.

We split the data into training and testing sets using an 80/20 division. The training set of 157,035 records was used to teach the model patterns, while the test set of 39,259 records was held back to evaluate performance on data the model had never seen. We maintained the 15-to-1 income ratio in both sets to ensure consistent conditions. The population weights were preserved throughout to ensure results generalize to the full population rather than just our sample.

Finally, we standardized numeric features by converting them to a common scale where the average is 0 and most values fall between -3 and +3. This prevents features measured in different units from dominating the model simply because they have larger numbers.

2.4 Technical Considerations

The preprocessing pipeline involved several technical decisions that balance data quality with model performance:

- **Duplicate Detection**: Used exact row matching across all 40 features to identify duplicates. The 1.62% duplicate rate is typical for large administrative datasets.

- **Missing Value Strategy**: Applied median imputation for continuous variables and mode imputation for categoricals after converting implicit missing indicators ("Not in universe", "?", 0) to explicit null values.

- **Encoding Method**: Label encoding was chosen over one-hot encoding to manage dimensionality. With education having 17 levels and occupation 15 categories, one-hot encoding would create 200+ sparse columns. Label encoding preserves ordinal relationships where they exist.

- **Stratified Sampling**: The train-test split used stratification to maintain the 93.8%/6.2% income distribution in both sets, preventing evaluation bias from random sampling variations.

- **Sample Weights**: Population weights from the Census Bureau's stratified sampling design were preserved and applied during model training, ensuring predictions generalize to the full population rather than the survey sample.

- **Standardization**: Applied z-score normalization to continuous features only (age, weeks worked, wages, capital gains, losses, dividends), leaving encoded categorical variables on their natural scale.

3 Classification Model Development

With clean data prepared, we turned to building a model that could predict whether someone earns at least $50,000 based on their demographic and employment characteristics. The goal was not just accuracy, but understanding what patterns drive high income and how confidently we can make predictions in different cases.

3.1 Model Selection and Training

We selected XGBoost, a machine learning method that builds its predictions through a series of decision trees. Each tree learns from the mistakes of previous trees, gradually improving the overall prediction. This approach handles mixed data types well, working with both categorical features like occupation and numeric features like age without requiring extensive transformation.

The model architecture includes several important features for our specific problem. First, it directly addresses class imbalance by giving extra weight to the minority class during training. We set this weight to 14.85, matching the ratio of low to high earners in our data. This prevents the model from simply predicting everyone as low income. Second, the model incorporates sample weights from the Census Bureau, ensuring our predictions reflect population patterns rather than just our sample. Third, it optimizes for F1-score, which balances catching as many high earners as possible while maintaining reasonable precision.

Training used 20 iterations of parameter optimization to find the best configuration. Rather than testing every possible combination, we used an efficient search method that learns from previous attempts to intelligently select the next parameters to try. The optimization used 3-fold cross-validation, meaning we split the training data into thirds and trained on two-thirds while validating on the remaining third, rotating through all combinations.

The final model configuration emerged from this process: trees can grow up to 10 levels deep, providing enough complexity to capture patterns without overfitting to noise. The learning rate

of 0.076 controls how much each new tree adjusts the predictions. We use 500 trees total, finding this balances accuracy with training time. Two regularization parameters prevent overfitting by randomly sampling 72% of data points and 73% of features for each tree, ensuring trees learn different aspects of the patterns.

3.2 Understanding Model Performance

The model achieved strong differentiation ability, with an ROC-AUC score of 0.938. This metric measures how well the model separates high and low earners across all possible threshold settings, with 1.0 being perfect and 0.5 being random chance. The PR-AUC score of 0.630 provides a complementary view focused on performance with imbalanced classes.

However, these overall scores mask an important reality. At the default threshold of 0.5, where the model classifies anyone with a predicted probability above 50% as a high earner, precision drops to 47.8%. This means that if we contacted 100 people the model classified as high earners, only 48 would actually earn at least $50,000. The remaining 52 would be false positives.

At the same time, recall reaches 67.5%, meaning the model successfully identifies about two-thirds of all actual high earners in the data. These two metrics pull in opposite directions. We can increase precision by being more selective, but this means missing more actual high earners. We can increase recall by being more inclusive, but this means more false positives.
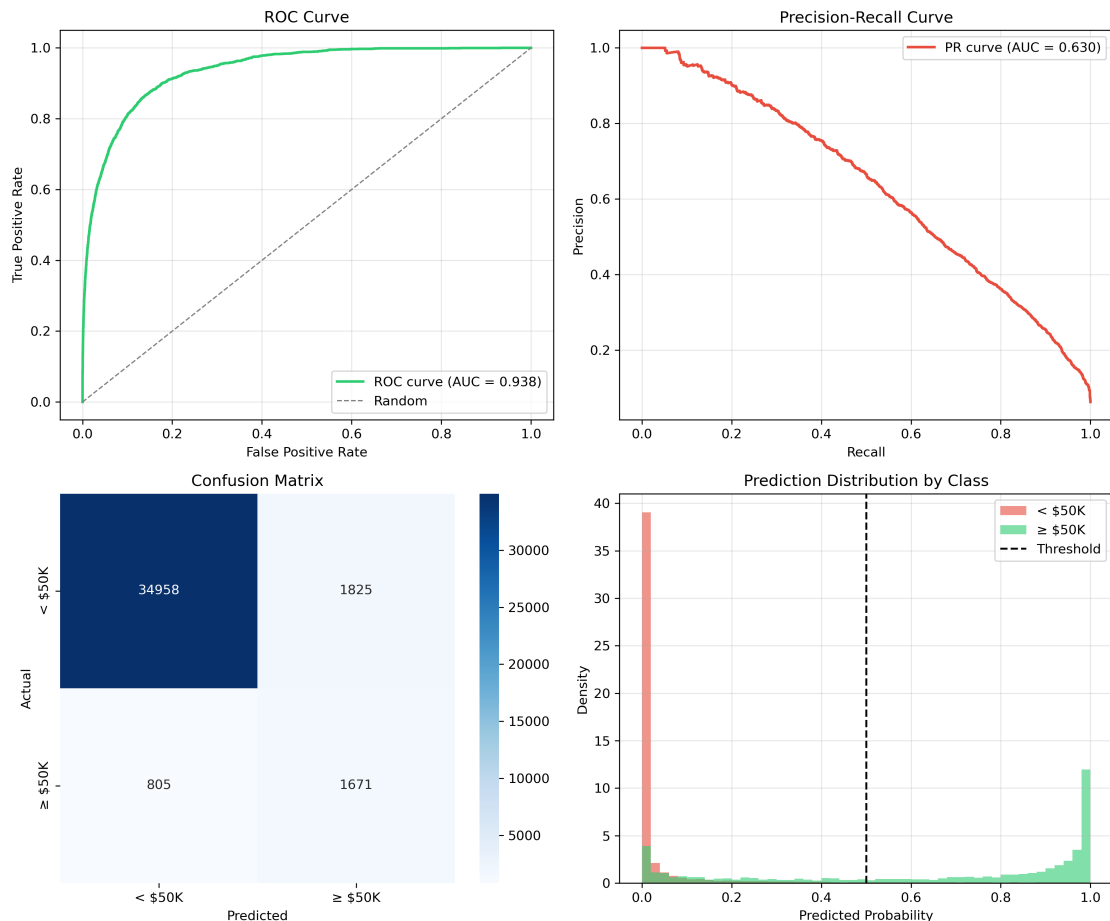

Figure 2: Model performance curves showing strong differentiation ability with ROC-AUC of 0.938.

3.3 The Threshold Question

This brings us to a fundamental question: where should we set the threshold? The default value of 0.5 assumes equal cost for two types of errors, but this rarely reflects real situations. In practice,

the costs and benefits of correctly identifying a high earner versus incorrectly targeting a low earner are different.

We analyzed performance across different threshold values to understand the tradeoffs. At a threshold of 0.71, we achieve balanced performance with 58% precision and 58% recall. This represents the point where the F1-score, a combined measure of precision and recall, reaches its maximum.

Moving to a more selective threshold of 0.81 increases precision to 64.9% while recall decreases to 51.8%. This means we contact fewer people overall, but a higher percentage of them are actually high earners. We successfully reach about half of all high earners while maintaining nearly two-thirds accuracy in our positive predictions.

An even more conservative threshold of 0.87 pushes precision to 70.1% but drops recall to 46.4%. At this point, we are highly confident in the people we target, but we miss more than half of potential high earners.
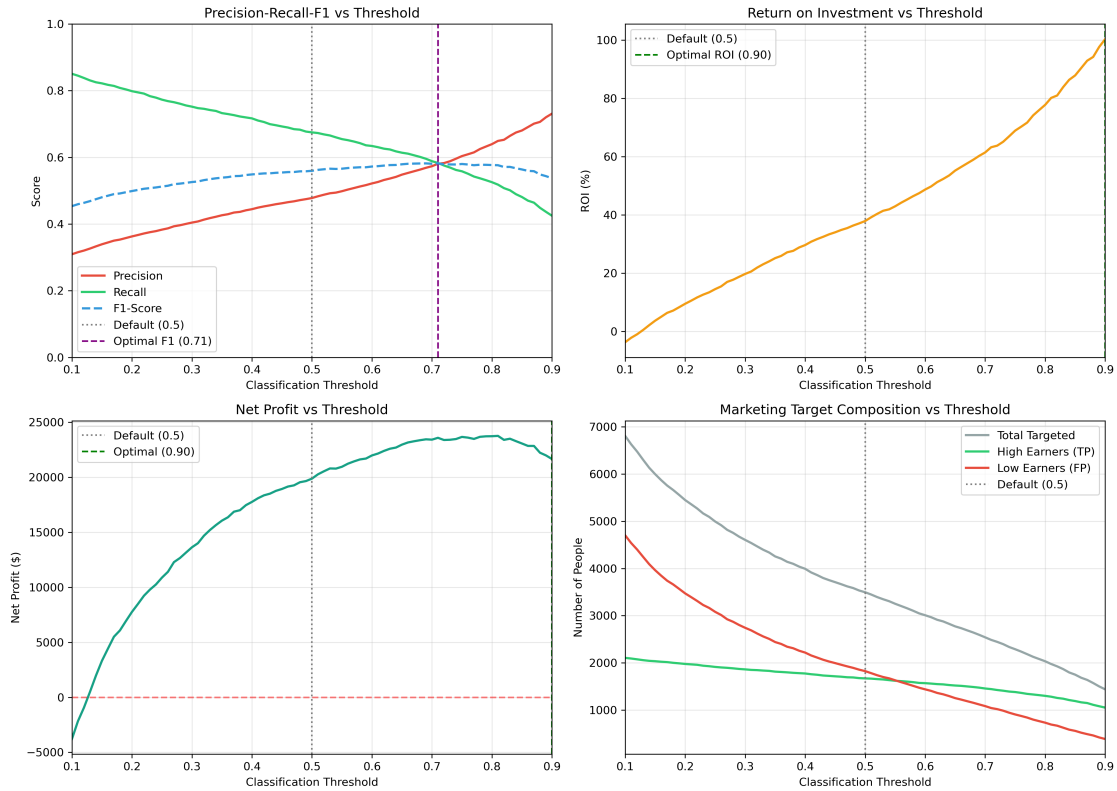

Figure 3: Performance across different thresholds. The optimal choice depends on the relative importance of precision versus recall for the specific application.

The choice between these thresholds depends on the relative importance of reaching more potential customers versus focusing resources on the most certain cases. There is no single "correct" answer, only tradeoffs that align differently with different objectives.

3.4 What Drives the Predictions

Understanding which features most influence the model's predictions provides insights into income patterns in the data. The number of weeks worked per year dominates at 16% of total feature importance. This makes intuitive sense as full-year employment strongly signals stable income.

Occupation ranks second at 6.7%, indicating that what you do matters more than many demographic characteristics. The type of work provides information about income level beyond what we know about age, education, or other factors.

Capital gains (4.9%) and dividend income (3.2%) serve as wealth indicators. People with investment income tend to have higher overall earnings. Sex appears fourth at 4.5%, reflecting wage gap patterns in the 1990s data.

Education appears seventh at 2.6% despite being one of the strongest single predictors of income when examined alone. This occurs because much of education's effect works through other variables already in the model. Higher education leads to professional occupations, which leads to more weeks worked and higher wages. Once these intermediate factors are included, education's direct additional contribution becomes smaller.
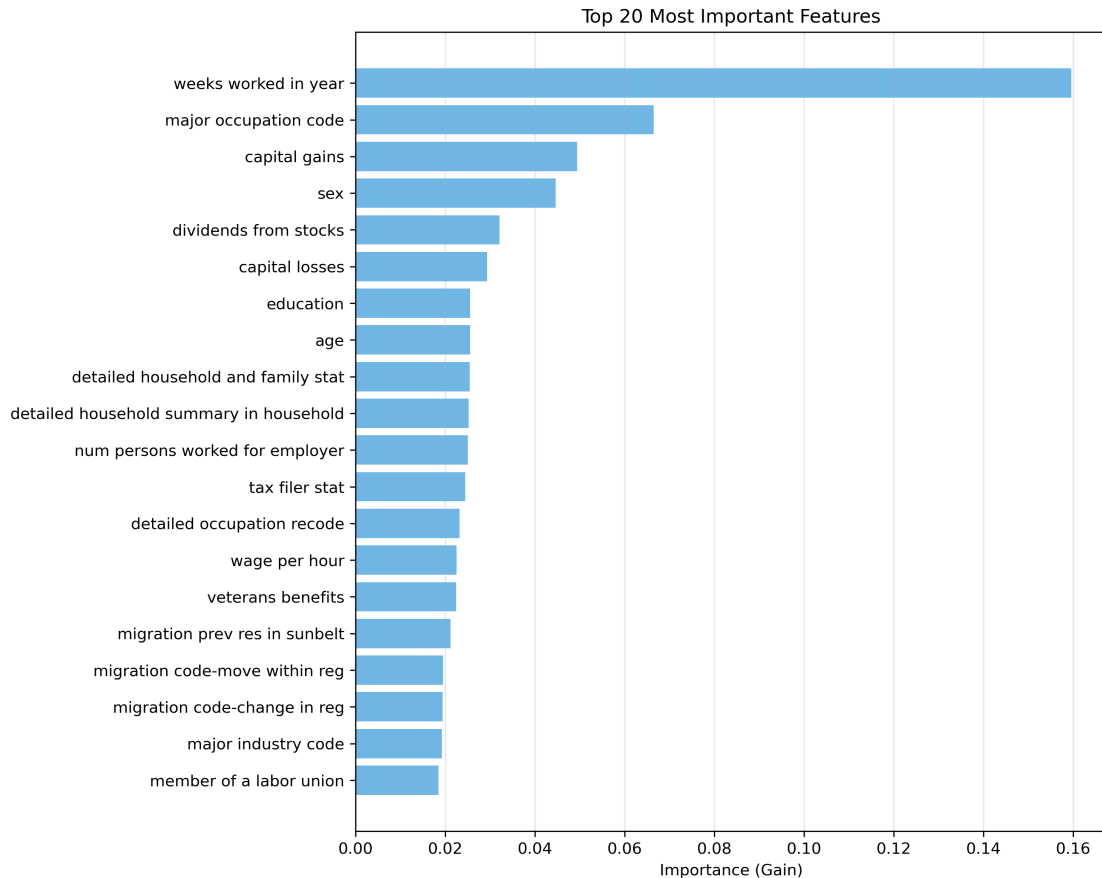

Figure 4: Features ranked by their contribution to model predictions. Work engagement and occupation dominate, while demographic factors play supporting roles.

Cross-validation showed consistent performance across different subsets of the training data, with F1-scores ranging from 0.54 to 0.58. Test set performance matched training performance, indicating the model generalizes well to new data without overfitting.

3.5 Technical Considerations

Several technical aspects of the modeling process deserve attention:

- **Algorithm Selection**: XGBoost was chosen over logistic regression and random forests for its superior handling of mixed feature types, built-in class imbalance weighting, and native sample weight support. The gradient boosting framework iteratively improves predictions by learning from residual errors.

- **Hyperparameter Optimization**: Used Bayesian optimization with 20 iterations and 3-fold cross-validation. This approach is more sample-efficient than grid search, particularly important with 157,000 training samples. Optimized for F1-score rather than accuracy due to class imbalance.

- **Final Parameters**: max_depth=10 controls tree complexity, learning_rate=0.076 sets gradient step size, n_estimators=500 provides 500 boosting iterations, subsample=0.72 and colsample_bytree=0.73 provide stochastic regularization, min_child_weight=10 prevents overfitting on minority class, gamma=0.5 adds additional regularization.

- **Class Imbalance Handling**: scale_pos_weight=14.85 directly addresses the 15:1 class imbalance by upweighting positive class loss during training. Without this adjustment, the model would achieve high accuracy by predicting the majority class.

- **Evaluation Metrics**: ROC-AUC measures differentiation ability across all thresholds, PR-AUC focuses on performance with imbalanced classes, F1-score balances precision and recall, and confusion matrices show actual prediction counts at specific thresholds.

- **Feature Importance**: Calculated using gain-based importance, measuring the average gain of splits using each feature across all trees. Education's lower direct importance despite strong univariate correlation reflects its mediated effect through occupation and work patterns.

4 Customer Segmentation Analysis

While the classification model predicts individual income levels, segmentation takes a different approach. Instead of asking whether someone is high or low income, it asks what natural groups exist in the customer base and how these groups differ from each other. This provides a complementary view for developing marketing strategies.

4.1 Methodology and Segment Selection

We applied K-means clustering, a method that groups similar individuals together based on 15 characteristics. These characteristics span demographics, employment patterns, financial behavior, and household structure. The algorithm iteratively assigns people to clusters and adjusts cluster centers to minimize the total distance between individuals and their assigned cluster.

A key question was how many segments to create. We evaluated solutions with three, four, and five clusters using multiple quality metrics. While five clusters achieved slightly better statistical separation, we selected three clusters for two important reasons.

First, the five-cluster solution contained redundant groups. Two segments showed nearly identical income distributions at 10.9% and 12.4% high earners. Another pair showed 0.03% and 1.8% high earners. These similarities suggested we were splitting natural groups unnecessarily.

Second, three segments align with recognizable lifecycle stages that marketing teams can operationalize. Five segments with subtle statistical differences would be harder to distinguish and target in practice.

The silhouette score of 0.163 indicates weak statistical separation between clusters. This falls below the commonly cited threshold of 0.25 for "reasonable" clustering. However, statistical separation is not the primary goal. What matters is whether the segments show meaningful business differences.

On this criterion, the three-segment solution succeeds. The segments show a 292-fold difference in high-income rates, from 0.05% to 13.3%. Work engagement varies systematically across segments. Investment behavior shows clear patterns. These practical differences justify the segmentation despite modest statistical separation.
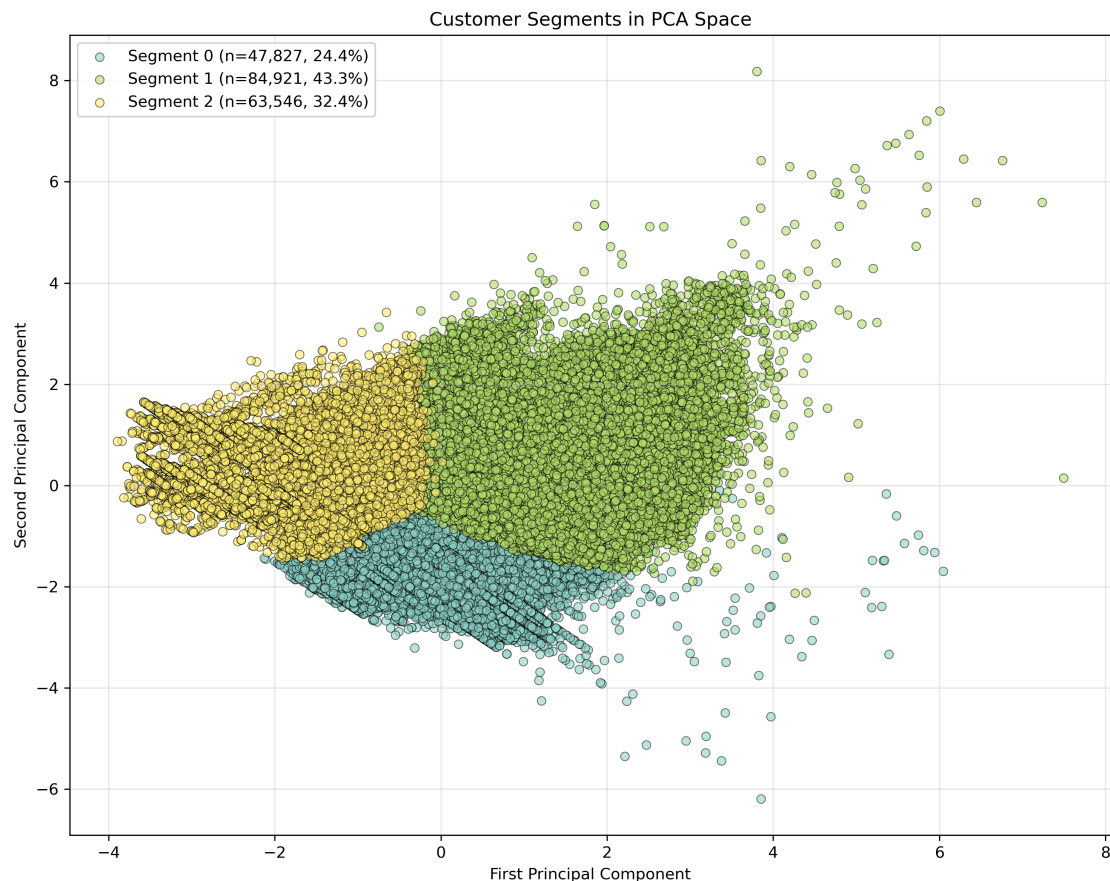
Figure 5: Visualization of customer segments using principal component analysis. The first two components explain 27.9% of variation, indicating high-dimensional complexity that cannot be fully captured in two dimensions.

The visualization uses principal component analysis to project the 15-dimensional clustering space onto two dimensions for viewing. The first component explains 18.1% of variance, the second explains 9.8%, for a total of 27.9%. The relatively low explained variance indicates the data has high-dimensional structure that cannot be simplified to two dimensions without information loss. The segments overlap in the visualization but maintain distinct centers, reflecting that they represent tendencies rather than hard boundaries.

4.2 Segment Profiles

Segment 0: Older Adults and Retirees (47,827 people, 24.4%)
This segment shows the characteristics of reduced labor market participation combined with accumulated financial assets. Only 2.1% earn at least $50,000, the lowest rate across segments. Work engagement sits at 71% below average based on weeks worked per year.

Age is significantly elevated, with the average person being one standard deviation older than the overall population. Investment patterns are notable: 10.5% receive dividend income compared to 3.4% with capital gains. This inverted ratio suggests buy-and-hold investment strategies typical of retirees rather than active trading.

The profile fits individuals who have withdrawn from full-time employment but maintain financial resources through retirement accounts, social security, pensions, and long-term investments. While current employment income is low, total financial resources may be more substantial than income alone suggests.

Segment 1: Working Professionals (84,921 people, 43.3%)
This segment represents the core of stable employment and highest income potential. At 13.3%,

9

the high-income rate is more than six times the next highest segment. These individuals work full-year schedules without seasonal gaps or part-time status, with work engagement exactly at the average.

Investment activity sits at moderate levels, with 6.5% showing capital gains and 8.1% receiving dividend income. Age trends slightly above average, suggesting mid-career positioning rather than entry-level roles.

The segment includes employed individuals across occupations: mid-career professionals, managers, skilled tradespeople, and business owners. They show up consistently, work full schedules, and are building financial assets alongside income. This large segment likely includes a smaller premium subgroup with substantially higher income rates and investment activity, but the three-cluster solution merges them into the broader working professional category.

Segment 2: Early Career and Low Attachment (63,546 people, 32.4%)

This segment shows minimal current income potential, with high earners representing just 0.05% of the group. Work engagement is 81% below average. Investment activity is nearly absent, with only 0.3% showing capital gains and 0.4% receiving dividends.

Age is significantly younger, sitting one standard deviation below the population average. The combination of young age and low work engagement signals limited labor market attachment.

This profile includes students, part-time workers, people between jobs, and those early in their careers who have not yet established stable employment. The near-zero high-income rate indicates minimal current purchasing power, though some individuals in this segment will transition to higher-earning segments as they age and develop careers.

4.3   Technical Considerations

The segmentation methodology involved several technical decisions:

- **Algorithm**: K-means clustering was selected for its interpretability and efficiency with large datasets. The algorithm minimizes within-cluster sum of squared distances through iterative assignment and centroid updates.

- **Features**: Used 15 features capturing demographics (age, sex, race, marital status, education), employment (weeks worked, occupation, industry, class of worker), financial behavior (capital gains, losses, dividends, wage per hour), and household structure (family status, children under 18).

- **Cluster Number Selection**: Evaluated k=3, 4, and 5 using three metrics:

    - Silhouette score: 0.163 for k=3 (measures cluster cohesion and separation)
    - Calinski-Harabasz index: 26,116 (higher is better, measures ratio of between-cluster to within-cluster variance)
    - Davies-Bouldin index: 2.12 (lower is better, measures average similarity between clusters)

- **Standardization**: All features were z-score normalized before clustering to prevent features with larger scales from dominating distance calculations.

- **Initialization**: K-means used k-means++ initialization to select starting centroids intelligently, reducing sensitivity to random initialization.

- **Convergence**: Algorithm ran until centroids moved less than 0.0001 or maximum 300 iterations reached.

- **Dimensionality Reduction**: PCA was used purely for visualization. Clustering was performed in the original 15-dimensional space to preserve full information. The 27.9% variance explained by two principal components indicates the high-dimensional nature of customer differences.

5 Conclusions and Recommendations

This analysis demonstrates that effective customer targeting requires both prediction and segmentation working together. The classification model excels at identifying high-income individuals with 93.8% differentiation ability, but its practical value depends critically on threshold selection. The segmentation model identifies natural customer groups that align with lifecycle stages, enabling targeted approaches even when individual income prediction is uncertain.

5.1 Key Findings

The threshold analysis reveals that default settings optimize statistical metrics rather than practical outcomes. At the standard 0.5 threshold, the model achieves 93.3% accuracy but may waste resources on false positives. Moving to 0.81 drops accuracy to 89.7% while substantially improving precision. This illustrates why understanding business context matters as much as technical performance.

The segmentation demonstrates that weak statistical separation does not preclude practical value. A silhouette score of 0.163 would typically signal poor clustering, yet the 292-fold income difference and systematic variation in work patterns justify the three-segment approach. Marketing teams can work with lifecycle-based segments more easily than statistically optimal but conceptually unclear alternatives.

Feature importance analysis shows that work engagement dominates income prediction at 16%, followed by occupation at 6.7%. Education appears surprisingly low at 2.6% despite being a strong individual predictor, because much of its effect operates through occupation, work hours, and financial behavior captured elsewhere in the model.

5.2 Practical Applications

The classification model enables targeted identification of high-income prospects. Different threshold choices serve different purposes. A threshold of 0.81 balances precision and recall, reaching about half of high earners with nearly two-thirds accuracy. A threshold of 0.87 achieves 70% precision for budget-constrained situations where only the most certain targets make sense. A threshold of 0.71 provides balanced 58% precision and recall for broader coverage.

The segmentation provides a framework for differentiated strategies. Segment 1 working professionals warrant primary focus given their 13.3% high-income rate and stable employment. Segment 0 older adults show low current income but may have accumulated wealth, suggesting different product offerings. Segment 2 early career individuals have minimal current income but represent future potential, justifying brand-building investment.

The combined approach manages risk better than either method alone. Pure classification risks over-targeting the same high-probability individuals. Segmentation ensures coverage across diverse customer types. If classification predictions prove unreliable, segmentation still provides differentiated strategies based on observable characteristics.

5.3 Limitations and Considerations

Several limitations require attention when applying these findings:

Data Currency: The 1994-1995 source data is now three decades old. Income distributions, wage gaps, education returns, and financial behaviors have evolved substantially. The gender wage gap has narrowed. Technology has changed how people work and invest. The model patterns may not fully reflect current conditions.

Protected Characteristics: The model uses sex and race as predictive features. While these improve predictions, their inclusion creates potential regulatory concerns. Deployment would require careful legal review and consideration of fair lending and discrimination laws. Alternative approaches might exclude these features or apply fairness constraints.

5.4   Next Steps

Organizations considering this approach should follow a staged implementation:

1. **Validate on current data**: Test model performance on recent customer data to assess whether 1990s patterns still hold.

2. **Conduct pilot test**: Deploy to a small customer subset to measure actual response rates and validate threshold choices.

3. **Review legal considerations**: Assess regulatory requirements around protected characteristics and fair lending.

4. **Refine segments**: Validate whether the three-segment structure matches actual customer behavior and refine as needed.

5. **Monitor performance**: Track ongoing results to detect degradation and identify when retraining becomes necessary.

The framework provides a foundation for data-driven customer targeting. Success depends on adapting these methods to specific business contexts, validating assumptions through testing, and maintaining awareness of both technical capabilities and practical limitations.