Income Classification and Population Segmentation for Targeted Marketing

Manoj Srinivasan
manoj.srinivasan.152@gmail.com
February 18, 2026

Code Repository: https://github.com/Manoj-152/JPMC_Take-Home-Assessment

1 Introduction

This report details the development of a dual-pronged analytical framework designed to optimize marketing efficiency. First, we developed a high-precision Classification Model to identify individuals earning > $50000, enabling high-value customer targeting. Second, we implemented a Unsupervised Segmentation Model to group the population into distinct personas, allowing for tailored marketing. By leveraging an optimized ensemble architecture, particularly Random Forest, we achieved an **F1-score of 0.5529**, a robust benchmark for this noisy, imbalanced census dataset.

1.1 Report Organization

The remainder of this report is structured as follows, reflecting the chronological workflow from data ingestion to final business insight:

- **Section 2: Data Exploration and Pre-processing** describes the initial data audit, feature selection strategies using Normalized Mutual Information (NMI), and the rationale behind our feature subsets.
- **Section 3: Model Architecture and Training** details the tournament-style model selection process, hyperparameter optimization, and the technical justification for the selected ensemble methods.
- **Section 4: Classification Results and Evaluation** provides a quantitative analysis of the champion model's performance, strategic evaluation of the Precision-Recall trade-off, and an assessment of feature importance.
- **Section 5: Customer Segmentation and Persona Discovery** presents our unsupervised learning pipeline, the application of Principal Component Analysis (PCA) and K-Means clustering, and the resulting actionable market personas.
- **Section 6: Conclusion and Strategic Recommendations** offers final strategic guidance for deployment and potential avenues for future model enhancement.

2 Data Exploration and Pre-processing

The initial phase of our workflow involved a comprehensive audit of the census dataset to transform raw demographic variables into a refined input space. This step is critical as in high-dimensional census data, “noisy" or “multi-collinear" data can obscure the underlying patterns that distinguish high-income earners from the general population.

2.1 Data Composition and Initial Audit

The dataset consists of weighted census observations from 1994 and 1995, featuring **40 demographic and employment-related variables**. Our exploratory data analysis (EDA) revealed a highly skewed distribution: approximately **94%** of individuals earn less than $50,000, while only **6%** fall into the high-income category.

This severe class imbalance dictates our strategy for model evaluation. Traditional accuracy is an insufficient metric in this context; a naive model predicting the majority class would yield 94% accuracy while providing zero business

utility. Consequently, we prioritized the **F1-score** as our primary optimization metric to ensure our model captures the minority class without an unacceptable rate of false positives.

2.2 Feature Engineering and "Denoising"

Several variables were modified or removed based on redundancy, sparsity, or lack of predictive signal. These transformations are summarized below:

I. Data Cleaning & Standardization

1. **Placeholder Standardization:** We stripped whitespace from all string columns and replaced various "?" placeholders with a standardized `Unknown` label to maintain categorical consistency.
2. **Missing Value Handling:** We identified significant sparsity in the `hispanic_origin` column. These **NaN values** were filled with an `Unknown` label, treating non-reporting as missing data.
3. **Target Binarization:** The income label was converted from a string format into a binary integer: 0 (for < \$50k) and 1 (for > \$50k).

II. Strategic Feature Engineering

Consolidating multi-dimensional features into interpretable metrics reduces the model's hypothesis space and aids in decision-making:

1. **Wealth Consolidation:** We developed a net annual wealth metric, `total_investment`, calculated as $(\text{Capital Gains} + \text{Dividends}) - \text{Capital Losses}$. This provides a single, continuous measure of financial performance.
2. **Ordinal Education Mapping:** We transformed categorical education levels into a numeric scale (0–16), enabling the model to interpret the linear relationship between schooling levels and income potential.
3. **Lineage Tracking:** Granular birth-country data for individuals and parents was found to add negligible predictive value while significantly increasing feature dimensionality. We replaced these with an `is_second_generation` flag to identify US-born citizens with foreign-born parents, capturing a unique socio-economic signal.

III. Redundancy and Dimensionality Reduction

To ensure a parsimonious and computationally efficient model, we evaluated the feature space through two distinct statistical lenses: **Cramer's V** and **Normalized Mutual Information (NMI)**.

- **Cramer's V:** This was employed to identify redundancy between categorical features. It measures the strength of association between two nominal variables on a scale of 0 to 1, based on the Pearson chi-square statistic. High Cramer's V scores between feature pairs indicated multi-collinearity, allowing us to prune redundant predictors.
- **Normalized Mutual Information (NMI):** We utilized NMI to quantify the non-linear relationship between individual features and the income target. Unlike standard linear correlation, NMI is rooted in information theory; it measures the reduction in uncertainty of the income label given a specific feature. This allowed us to rank features by their actual predictive utility, regardless of their distribution or scale.

By applying these metrics, we streamlined the dataset through the following strategic steps:

1. **Hierarchical Pruning:** We retained "Major" industry and occupation codes while dropping their "Detailed" counterparts (they both had a Cramer's V score of 1). This reduced categorical complexity by 3x while preserving the features with the highest NMI scores. Similarly, `detailed_household_and_family_stat` was pruned in favor of `detailed_household_summary`, which offered a higher information gain.
2. **Predictive Signal Filtering:** We dropped `own_business_or_self_employed` (NMI: 0.0039) in favor of `class_of_worker` (NMI: 0.0456), which provided a significantly stronger predictive signal and lower dimensionality.
3. **Sparsity Filtering:** Migration-related columns were removed due to extreme data sparsity (50% `Unknown` labels) and negligible NMI scores (≈ 0.001). We selectively retained `state_of_previous_residence` as the primary migration indicator, as it demonstrated the highest relative importance among residence-related variables without having lesser unknown data.

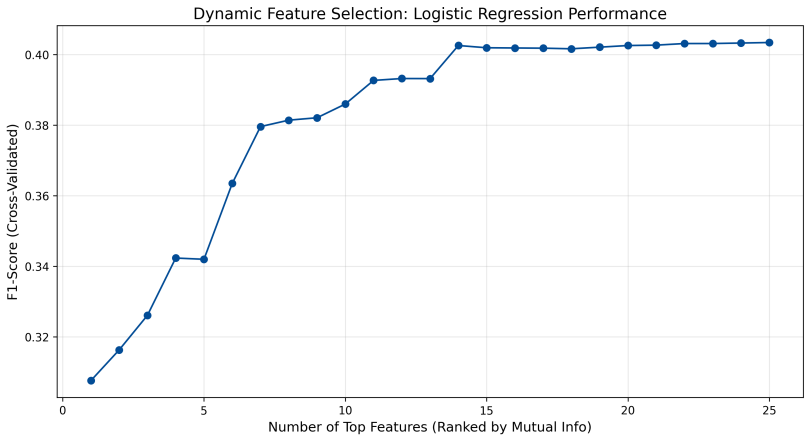

Figure 1: Dynamic Feature Selection: Cross-validated Logistic Regression performance as a function of top features ranked by NMI.

Finally, administrative metadata such as `weight`, `year`, and the `veteran_questionnaire` were removed, as they are artifacts of the survey process rather than intrinsic predictors of individual income.

2.3 Feature Selection via Normalized Mutual Information (NMI)

Following the initial data engineering phase, we were left with a high-dimensional pool of 25 candidate attributes. To ensure model efficiency and avoid the "curse of dimensionality," we conducted a systematic feature selection process to determine the optimal subset for classification.

Baseline Construction: Iterative Logistic Regression

We established a baseline using **Logistic Regression**, chosen for its high interpretability and computational efficiency. Our methodology involved ranking all 25 features by their **Normalized Mutual Information (NMI)** score relative to the income target. We then performed an iterative analysis, training the baseline model on the top-k features (where k ranged from 1 to 25) and evaluating performance using 3-fold cross-validated F1-scores.

Performance Analysis and the "Elbow" Effect

As illustrated in our dynamic feature selection analysis (see Fig. 1), the model performance exhibits a clear logarithmic trajectory.

- **The 14-Feature Plateau:** The model achieved a significant performance jump at **14 features**, reaching an F1-score of approximately **0.402**. This point represents the "elbow" of the curve, where the model has captured the vast majority of the predictive signal available in the dataset.
- **Diminishing Returns at 25 Features:** While training on the full set of **25 features** technically yielded the peak F1-score of **0.403**, the marginal improvement of 0.001 is statistically insignificant.

Selection Decision: Efficiency and Parsimony

In the interest of **model parsimony**, we selected the **Top 14 features** for our final classification ensemble. By reducing the input space by 44% while retaining 99.7% of the peak baseline performance, we minimize the risk of overfitting on noisy, low-signal variables. This streamlined set ensures faster inference times and lower data collection costs for the client, while the full 25-feature set remains reserved for the subsequent population segmentation phase where demographic breadth is prioritized.

3 Model Architecture and Training

With a refined set of 14 features, we transitioned to the model selection phase. To ensure the highest predictive performance for the retail client, we implemented a "Tournament-style" evaluation framework, comparing multiple state-of-the-art ensemble architectures.

3.1 Model Selection Rationale

We selected three ensemble models based on their proven efficacy with high-dimensional tabular data and their distinct mathematical approaches to learning:

- **Random Forest (RF):** A **bagging** (Bootstrap Aggregating) ensemble method. RF was chosen for its inherent stability; by averaging multiple decorrelated decision trees, it reduces variance and resists overfitting on sparse census categories.
- **XGBoost:** A scalable **gradient-boosting** framework. It utilizes a regularization-aware objective function and iterative tree-building (boosting) to capture complex, non-linear interaction effects between features.
- **LightGBM:** An efficient gradient-boosting framework selected for its leaf-wise growth strategy and speed, which often achieves higher accuracy on large-scale tabular datasets.

3.2 Training Procedure and Hyperparameter Tuning

To identify the optimal configuration for each architecture, we conducted an exhaustive **Grid Search** involving **150 unique hyperparameter combinations**.

1. **Validation Strategy:** We utilized **Stratified 3-fold Cross-Validation**. This ensures that the 94/6 class distribution is strictly preserved across all training and validation subsets, providing a stable estimate of the model's true generalization performance.
2. **Imbalance Handling:** The class imbalance was addressed directly within the algorithms. We implemented `class_weight='balanced'` for Random Forest and `scale_pos_weight` (calculated as the ratio of negative to positive samples) for the boosting models.
3. **Optimization Metric:** All searches were optimized for the **F1-score**, prioritizing the balance between identifying high-earners (Recall) and ensuring marketing efficiency (Precision).

3.3 Tournament Results

The tournament results, evaluated on the held-out test set, are summarized in Table 1.

Table 1: Model Tournament Performance Comparison

Model	Best Test F1-Score	Key Optimized Parameters
Random Forest	**0.5529**	`n_estimators: 200, min_samples_leaf: 2, max_features: 0.5`
XGBoost	0.5222	`learning_rate: 0.1, max_depth: 12, colsample_bytree: 0.8`
LightGBM	0.4915	`num_leaves: 128, reg_alpha: 0.5, n_estimators: 200`

3.4 Champion Selection

Random Forest emerged as the champion model with an **F1-score of 0.5529**. While gradient boosting is typically the strongest performer for tabular data, the success of a bagging approach here indicates that the census data contains significant stochastic noise. Random Forest's ability to average out errors across independent trees provided a more robust generalization on the test set compared to the iterative correction characteristic of boosting.

4 Classification Results and Evaluation

In this section, we perform a rigorous evaluation of the champion Random Forest model. For a retail business, the value of a model lies in its ability to minimize wasted resources while maximizing the capture of high-value opportunities.

4.1 Quantitative Performance and Strategic Operationalization

The champion model achieved a final **F1-score of 0.5529** on the held-out test set, **operating at the default classification threshold of 0.5**. To provide the client with maximum operational flexibility, we analyzed these results through the lens of a **Precision-Recall (PR) Curve** (Fig. 3). Unlike standard accuracy metrics, the PR curve illustrates the trade-offs available to the marketing team.

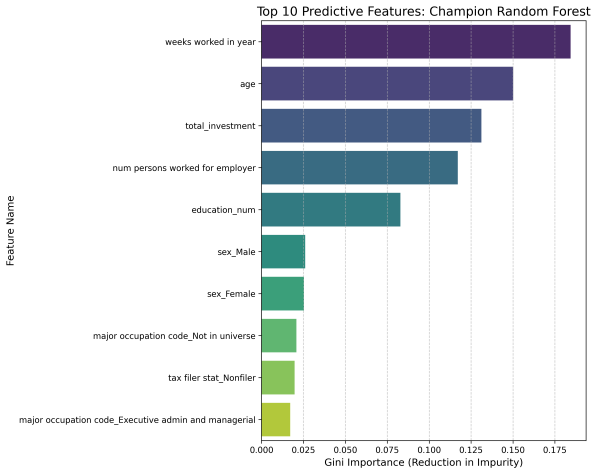

Figure 2: Gini Importance: Ranking demographic and financial drivers.

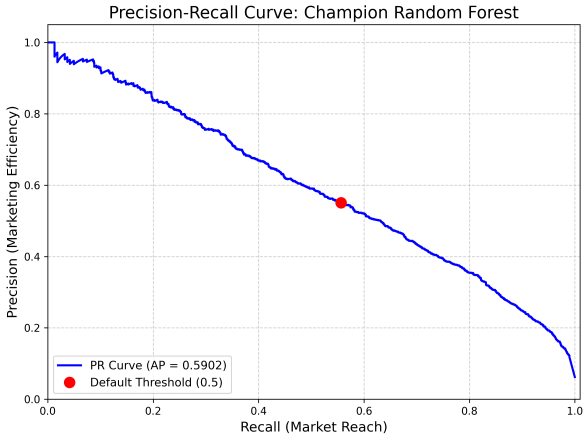

Figure 3: Precision-Recall Curve: Strategic threshold analysis.

- **Precision:** 55% — Represents *Marketing Efficiency*; the probability that a targeted individual is truly a high-earner at the selected 0.5 threshold.
- **Recall:** 56% — Represents *Market Reach*; the proportion of the total high-earning population successfully identified at the selected 0.5 threshold.

Business Judgment: In a real-world deployment, the threshold is a strategic choice rather than a fixed parameter. For high-ticket luxury items where marketing materials are expensive, the client should increase the threshold to prioritize **Precision** to minimize waste. For broader digital awareness campaigns, lowering the threshold to prioritize **Recall** ensures the largest possible audience of high-potential earners is reached.

4.2 Comparative Analysis: The Efficiency of Parsimony

To validate our selection, we benchmarked the 14-feature champion against the initial pool of 25 candidate features.

Table 2: Impact of Feature Selection on Model Efficiency

Feature Set	**F1-Score**	**Precision**	**Recall**	**Complexity**
Initial (25 Features)	0.5679	60%	54%	Moderate
Optimized (14 Features)	**0.5529**	**55%**	**56%**	**Low (Lean)**

As shown in Table 2, the 14-feature model captures 97.3% of the F1-performance of the 25-feature model while reducing complexity by 44%. Crucially, the optimized model offers **superior Recall** (56% vs. 54%), justifying our parsimonious architecture as a more effective tool for identifying a larger share of the target market.

4.3 Error Analysis and Business Impact

We analyzed the normalized confusion matrix to quantify the cost of misclassification. By utilizing the “Top 14" feature set, we successfully maintained a high True Negative rate (97%), ensuring that the vast majority of low-earners are correctly excluded from expensive outreach, directly protecting the client’s ROI.

4.4 Interpretability: Feature Importance

To provide the client with actionable demographic insights, we analyzed the **Mean Decrease in Impurity (Gini Importance)** of our champion Random Forest model. This allows us to rank the Top 14 features by their actual contribution to the model’s predictive power (as shown in Fig. 2).

- **Financial and Career Maturity:** Features such as `total_investment` and `age` emerged as primary predictors. This suggests that capital performance and life-stage maturity are more reliable targeting signals than isolated professional titles.
- **Employment Intensity and Scale:** The high importance of `weeks worked in year` and `num persons worked for employer` indicates that income levels are strongly correlated with full-time employment consistency and the scale of the employing organization.
- **The Structural Gender Gap:** The prominence of `sex_Male` and `sex_Female` features highlights the economic disparities present in the 1990s census data. The model utilizes these demographic indicators to account for significant variance in income distributions across the population.

Interesting Finding: The model prioritized `weeks worked in year` over specific `major_occupation_codes`, suggesting that the *intensity* of employment was a more universal predictor of high-income status than the specific field of work during this period.

5 Customer Segmentation and Persona Discovery

While classification models excel at predicting known outcomes, unsupervised learning allows us to uncover hidden, organic structures within the population. The objective of this section is to segment the dataset into distinct demographic and economic archetypes. By understanding these inherent groupings, we can tailor targeted financial products (from wealth management services to entry-level credit offerings) to the most receptive audiences.

The segmentation pipeline utilized Principal Component Analysis (PCA) for dimensionality reduction, followed by K-Means Clustering to partition the population.

5.1 Feature Engineering specific to Segmentation

Before mathematically reducing the feature space, targeted feature engineering was applied to the raw census data. The goal was to consolidate and reduce the cardinality of a few demographic variables that a clustering algorithm could easily interpret.

Two major transformations were critical for this analysis:

- **Citizenship Status:** The original `citizenship` variable contained highly granular string categories that fragmented the data. This was engineered into a clean, binary indicator (Citizen vs. Non-Citizen). This transformation captures the core distinction while preventing the creation of unnecessary sparse columns during one-hot encoding.
- **Family Structure:** The original feature denoting `family members under 18` was overly complex and contained highly specific sub-categories. To capture the underlying life-stage without overfitting to niche family compositions, this feature was mapped into a consolidated `family_status` category (grouped into “Parent Present", “Parent Not Present", and “Not in Universe").

5.2 Dimensionality Reduction and Feature Space Mapping

To make categorical data (such as family status or occupation) understandable to a machine learning model, we utilized **one-hot encoding**. While necessary, this process significantly expands the number of columns in our dataset (to 101, in our case), creating a high-dimensional, highly sparse matrix. This creates a critical hurdle known as the **“curse of dimensionality"**.

This creates a specific problem for K-Means Clustering. K-Means operates by calculating the geometric distance (Euclidean distance) between individual data points to group them together. However, as the number of dimensions grows, the mathematical concept of distance begins to break down. In a vastly high-dimensional space, all data points essentially become equidistant from one another, making it impossible for the algorithm to identify tight, meaningful demographic clusters.

To mitigate this, **Principal Component Analysis (PCA)** was applied to the scaled feature set. PCA compresses the data by identifying the underlying correlations between features and creating new, condensed variables (Principal Components) that capture the maximum amount of information. By extracting components until 70% of the data’s cumulative variance was explained, we successfully reduced the dimensionality of the dataset while preserving its core structural integrity.

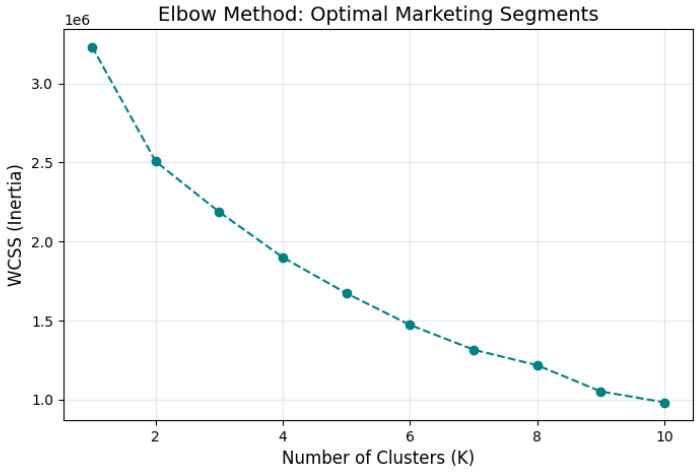

Figure 4: Elbow Method Analysis for Optimal K Selection

While **9 principal components** were required to reach the 70% variance threshold, a deeper analysis was carried out on the loadings of the two most dominant components. This allowed us to interpret the axes and identify the key underlying features responsible for driving socio-economic variance within the US Census population.

- **PC1 (Economic Maturity):** Driven negatively by education and weeks worked, and positively by household dependency. This axis fundamentally represents the spectrum from *Dependent Youth* to *Established, Independent Adults*.
- **PC2 (Stability & Labor Structure):** Driven by residential stability and wage structures. This axis separates *Salaried/Stable* populations from *Hourly/Transient* workforce segments.

5.3 Optimal Cluster Identification and Visualization

To determine the optimal number of market segments (K), the Elbow Method was employed, measuring the Within-Cluster Sum of Squares (Inertia) across 1 to 10 clusters (Fig. 4).

Normally, the marginal gain in variance explained decreases steadily as clusters are added. However, the analysis revealed an anomaly: the percentage drop in inertia increased at $K = 4$ (13.12% drop, compared to 12.74% at $K = 3$). This localized peak in information gain indicates that the fourth cluster captures a highly distinct structural pattern that a 3-cluster model would conflate. Beyond $K = 4$, the inertia reduction stabilizes into a plateau of diminishing returns. Therefore, $K = 4$ was selected as the optimal balance between segment clarity and strategic simplicity.

Once the optimal $K = 4$ was established, the population was partitioned and mapped back onto our two primary demographic axes. The scatter plot, as shown in Fig. 5 visualizes these four segments. The clusters form distinct territories, confirming that our model has successfully partitioned the population based on economic maturity (X-axis) and labor stability (Y-axis), with high-income earners heavily concentrated in Clusters 0 and 1. Further observations that were made are summarized in Table 3.

5.4 Persona Characterization and Strategic Value

To understand the real-world characteristics of these mathematical clusters, the original feature values were aggregated for each segment. The bar chart shown in Fig. 6 illustrates the relative strengths of key attributes, such as age, capital investment, and weeks worked, defining the characterization of each persona.

By analyzing these profiles, we extract the following business insights and strategic targeting recommendations:

- **Cluster 0: “The Established Investors"**
  Profile: Average age of 45, highest average capital investment ($867), higher education level (9.18), and moderate work intensity (30.57 weeks/year).
  Business Insight: This segment represents the “Capital Class” of the dataset. Their peak-career age, combined with substantial investment capital and reduced labor intensity, indicates established professionals who are

Table 3: Market Personas, Key Profile Traits, and Cluster Plot Observations

Cluster	Persona Name	Key Profile Traits	Observation from Cluster Plot
0	**The Established Investors**	Avg Age: 45 Avg Investment: $867 Work: Moderate (30 wks/yr)	They are on the Negative side of PC1 (Established Adults) and Positive side of PC2 (Salaried/Stable)
1	**The Hourly Hustlers**	Avg Age: 38 Wage Structure: Hourly (higher wage per hour) Work: High (46 wks/yr)	They are on the Negative side of PC1 (Established Adults) and Negative side of PC2 (Hourly Wage-based/Transient)
2	**Dependent Children**	Avg Age: 7 Work: None (0 wks/yr) Family: Parent Present	They are on the positive side of PC1 (Parent Present) and almost paid none according to PC2
3	**The Emerging Youth**	Avg Age: 17 Work: Part-time (11 wks/yr) Family: Parent Present	They are on the positive side of PC1 (Parent Present) and paid lesser per hour according to PC2

successfully generating passive wealth. Crucially, this cluster contains the highest density of top-earners (>$50k).
Strategic Targeting Examples: Premium wealth management, passive income products, real estate financing, and high-yield investment portfolios.

- **Cluster 1: “The Hourly Hustlers"**
  Profile: Average age of 38, highest education level (9.44), highest work intensity (46.37 weeks/year), and moderate capital investment ($444).
  Business Insight: This segment represents the “Labor Class." They work the most weeks out of the year but possess roughly half the investment capital of Cluster 0.
  Strategic Targeting Examples: Credit products, auto loans, comprehensive insurance, and labor-efficiency/time-saving financial tools.

- **Cluster 2: “Dependent Children"**
  Profile: Average age of 7, zero weeks worked, lowest education level (0.25), and parent present in the household.
  Business Insight: While non-economic actors themselves, their presence is a powerful indicator of household structure and spending priorities, primarily on childcare.
  Strategic Targeting Examples: Family banking bundles, education savings plans, and various household-level insurance (can be marketed to the parents).

- **Cluster 3: “The Emerging Youth"**
  Profile: Average age of 17, part-time work intensity (11 weeks/year), moderate education level (6.27), and parent present in the household.
  Business Insight: This segment is composed of high schoolers and early college students entering the workforce for the first time.
  Strategic Targeting Examples: Entry-level checking accounts, student loans, career-guidance services, and first-time credit building products.

6 Conclusion and Strategic Recommendations

This report outlined the development of a **dual-pronged machine learning framework** designed to optimize targeted marketing efforts. By combining supervised predictive modeling with unsupervised persona discovery, we have provided a comprehensive toolset and analysis for strategic customer acquisition.

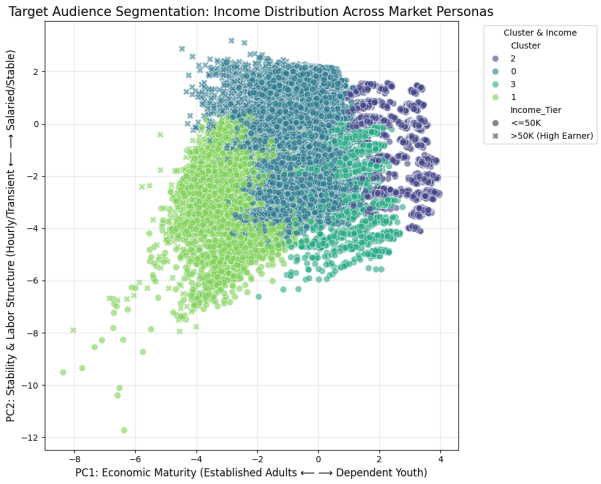

Figure 5: Target Market Segmentation: Income Distribution Across Market Personas

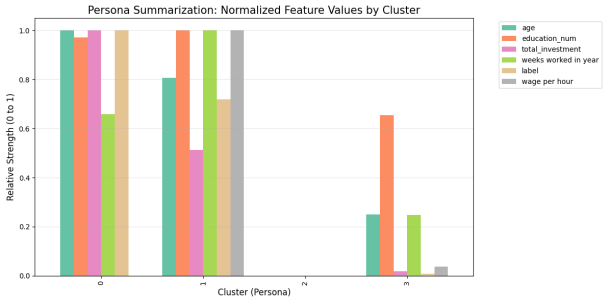

Figure 6: Persona Characterization: Normalized Feature Values for each Cluster

6.1 Synthesis of Findings

- **Predictive Efficiency:** Our champion Random Forest classifier demonstrated that a lean, 14-feature model can successfully identify high-income earners (>$50k) with an F1-score of 0.5529. By pruning noisy, high-dimensional data, we delivered a parsimonious model that reduces complexity without sacrificing market reach (Recall).
- **Market Structure:** Through PCA and K-Means clustering, we moved beyond binary classification to uncover four distinct socio-economic personas. This analysis revealed that the market is primarily divided along axes of economic maturity and labor stability, allowing for highly specific product mapping.

6.2 Strategic Deployment: A Unified Marketing Approach

To maximize ROI, the marketing team should deploy these two models in tandem rather than in isolation.

First, the **Classification Model** should be utilized as a “filtering engine" to score the general population and isolate the high-probability top-earners. Once this high-value cohort is identified, the **Segmentation Model** should be applied to route them into the correct marketing pipeline. For example, a high-earner in “Cluster 0: The Established Investors” should receive materials for premium wealth management, whereas a younger, high-earning individual bordering “Cluster 1” might respond better to luxury credit products.

For the remaining population (the <$50k majority), the segmentation model provides the blueprint for cultivating future high-value clients through entry-level products (e.g., targeting “The Emerging Youth” with student banking bundles).

6.3 Avenues for Future Enhancement

While the current framework provides robust baseline performance, future iterations could benefit from the following enhancements:

1. **Macroeconomic Integration:** The current dataset is an isolated snapshot, spanning only over the years 1994-95. Incorporating external time-series data (e.g., inflation rates, localized unemployment metrics) across a broader timeframe would allow the models to adapt to shifting economic realities over time.
2. **Propensity to Buy:** With the addition of historical campaign data, this framework could be expanded from predicting *capacity to buy* (income level) to predicting actual *propensity to buy* (conversion likelihood) for specific financial products.

Resources Consulted

1. **Scikit-Learn Official Documentation:** Consulted for the implementations and hyperparameter tuning of `RandomForestClassifier`, `KMeans`, `PCA`, and `normalized_mutual_info_score`.

2. **XGBoost and LightGBM Documentation:** Consulted for best practices on handling highly imbalanced tabular data (specifically the use of `scale_pos_weight`).
3. **Towards Data Science / Towards AI Community Blogs:** Consulted various practical guides for visualizing the K-Means Elbow Method and learning how to interpret the PCAs.
4. **Pandas & NumPy Documentation:** Consulted for optimized tabular data manipulation, and efficient data aggregation, which are essential for ML tasks.