Data-Driven Targeting and Customer Segmentation for High-Value Marketing

By - Shail Patel

Executive Summary

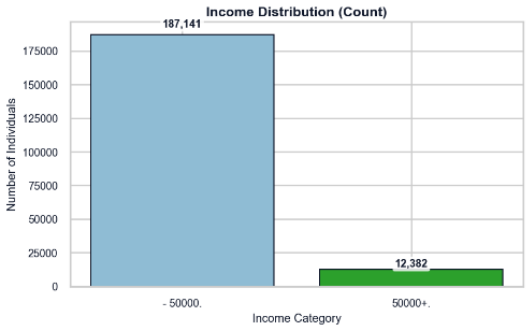
This report addresses two objectives aligned with your marketing strategy. First, it identifies individuals who are more likely to earn above $50,000 and therefore represent higher-value targets for outreach. Second, it segments the broader population into distinct customer groups that can be engaged with differentiated marketing approaches.

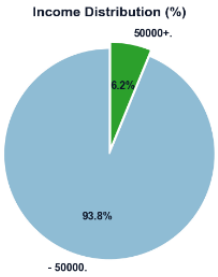
Using demographic, employment, and financial attributes, we developed a predictive income model to prioritize individuals with a higher likelihood of belonging to the high-income group, along with a complementary segmentation framework that explains how customer profiles differ beyond income alone. The income model supports more focused targeting decisions, while the segmentation model provides context for tailoring messaging, offers, and budget allocation across customer groups.

Together, these results enable a shift from broad, undifferentiated outreach toward a more targeted and data-driven approach. By concentrating effort on higher-value individuals and applying segment-specific strategies, you can improve conversion efficiency, reduce wasted marketing spending, and better align campaigns with underlying customer value.

Data Exploration

Data Overview and Target Distribution

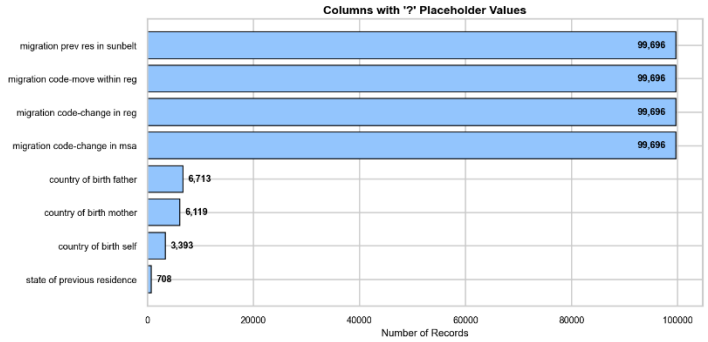




The dataset represents a broad cross-section of individuals described by demographic, employment, and financial attributes, with each record labeled by whether annual income falls above or below $50,000. A key characteristic of this population is the strong imbalance between income groups: the vast majority of individuals earn below $50,000, while a much smaller proportion earns above this threshold.

From a marketing perspective, this imbalance highlights that higher-income individuals represent a limited but disproportionately valuable segment. As a result, effective targeting requires precision rather than broad coverage. Identifying individuals with a higher likelihood of belonging to this group is therefore more valuable than maximizing overall classification accuracy.

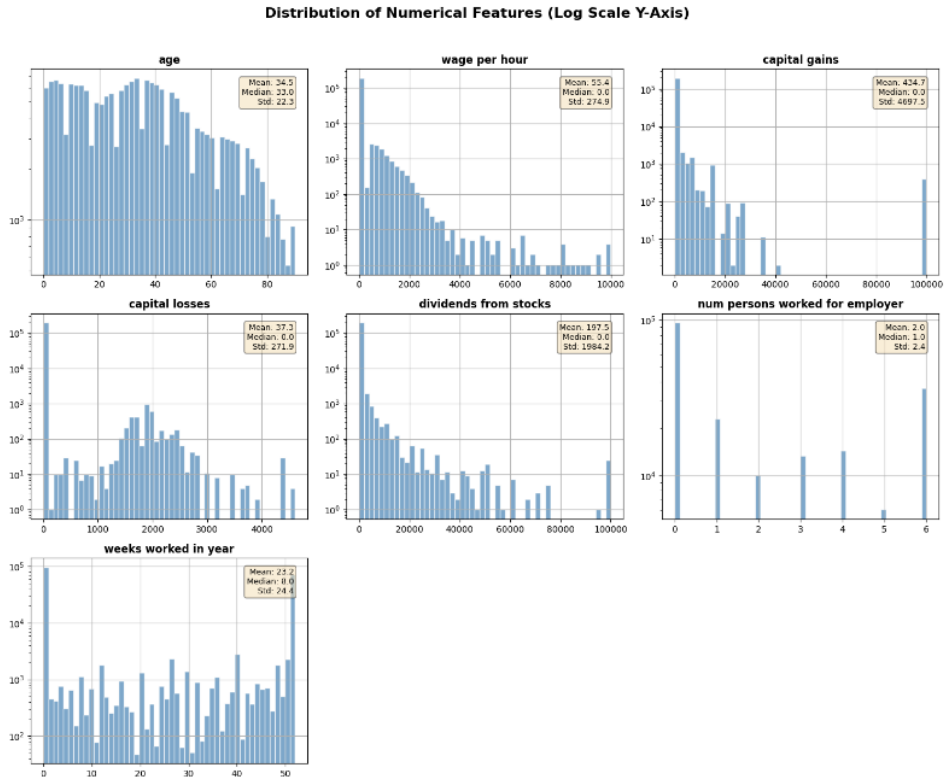
Missing Data Patterns



Several variables, particularly those related to migration and prior residence, contain many placeholder values. The concentration of these values is uneven across features, suggesting that the missingness reflects how certain questions were collected rather than random data loss.

This pattern informed downstream modeling decisions. Variables with structural missingness were handled explicitly to avoid introducing artificial signals that could distort predictions or segmentation outcomes.

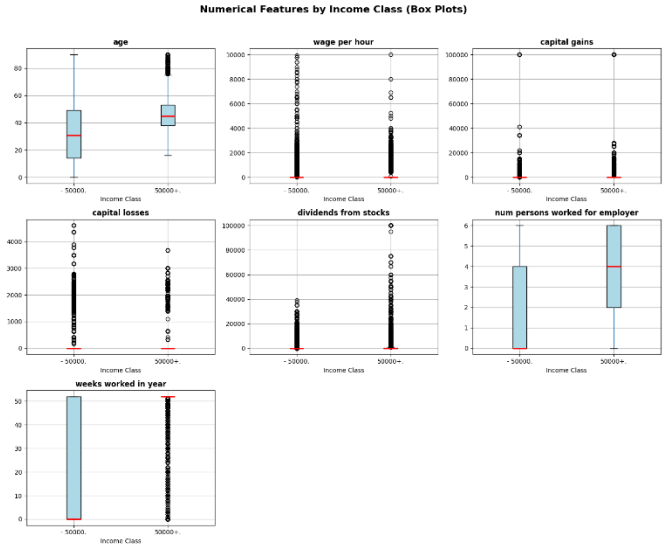
Numerical Feature Distributions



Many numerical attributes, including **capital gains, capital losses, dividends, and wage-related** measures, exhibit highly skewed distributions. For most individuals, these values are zero, while a small subset reports substantially higher amounts.

This structure indicates that income-related behavior is driven by threshold effects rather than smooth linear trends. As a result, modeling approaches capable of capturing nonlinear relationships are better suited for distinguishing higher-value individuals within this population.

Numerical Features by Income Group



Clear differences emerge when comparing numerical attributes across income groups. Individuals **earning above $50,000 tend to be older, work more weeks per year, and are significantly more likely to report non-zero capital gains**.

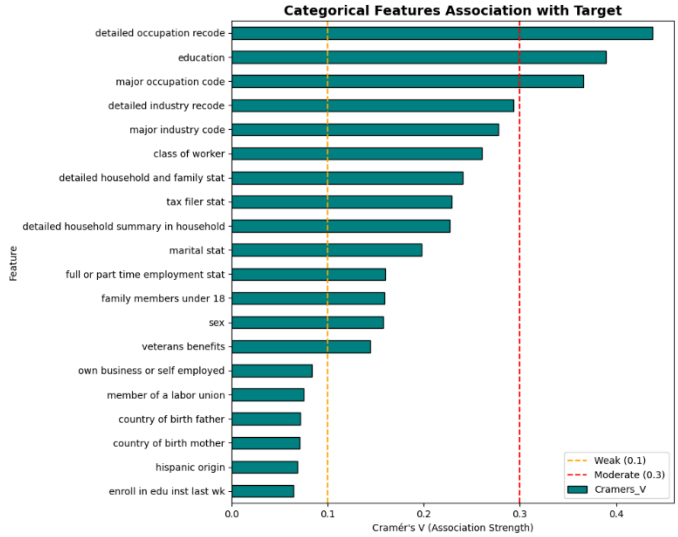
These patterns suggest that **income is influenced not only by wages, but also by employment stability and investment activity**. From a targeting perspective, these factors provide stronger signals of long-term customer value than hourly pay alone.

Categorical Feature Insights

Education and occupation exhibit some of the strongest relationships with income. The likelihood of earning above $50,000 **increases consistently with higher levels of educational attainment, particularly for professional, master's, and doctoral degrees**. Similarly, **professional, managerial, and executive occupations are substantially overrepresented in the higher-income group**.

These relationships are both intuitive and actionable, making them well suited for informing marketing strategies and customer segmentation.

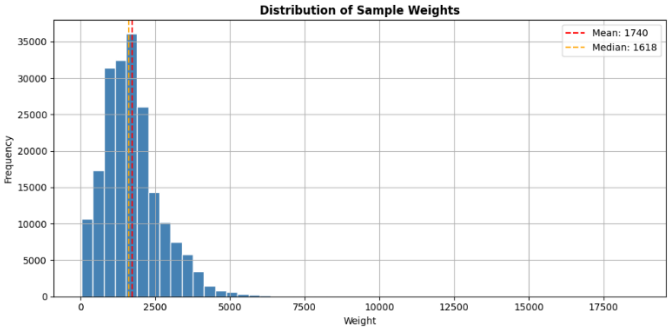
Feature Association Analysis



Association analysis **highlights occupation, education, industry, and employment status as the variables most strongly related to income outcomes**. Demographic attributes such as sex and marital status show weaker, but still meaningful, relationships.

Importantly, no single variable explains income on its own. Higher-income outcomes emerge from combinations of factors, reinforcing the need for multivariate targeting approaches rather than simple rule-based filters.

Population Representativeness and Weight Analysis



The dataset includes survey weights designed to reflect population-level representation. An examination of these weights shows **moderate skew but no extreme outliers, indicating that no small subset of records disproportionately influences population estimates**.

Comparisons between weighted and unweighted statistics reveal minimal differences across key variables, including age, employment intensity, and capital-related measures. Given the small magnitude of these differences, unweighted modeling was sufficient for predictive purposes, while weights remain useful for population-level interpretation and reporting.

Key Points

- Higher-income individuals represent a small but valuable segment, requiring precise targeting.
- Employment stability and investment-related signals are stronger indicators of income than wage measures alone.
- Education and occupation provide powerful, actionable levers for differentiation.
- Income outcomes are driven by combinations of factors rather than any single attribute.
- Careful handling of structurally missing variables is necessary to avoid misleading signals.

Modelling

Part 1: Supervised Income Classification

Business Objective

The objective of this modeling phase was to help the business identify individuals who are more likely to earn above $50,000 per year. This enables more efficient marketing and targeting by focusing effort on higher-value customer segments, rather than treating all individuals equally.

Because higher-income individuals represent a relatively small portion of the population, the model was designed to prioritize correctly identifying these individuals, even if it means accepting a limited number of false positives.

Modelling Techniques Used

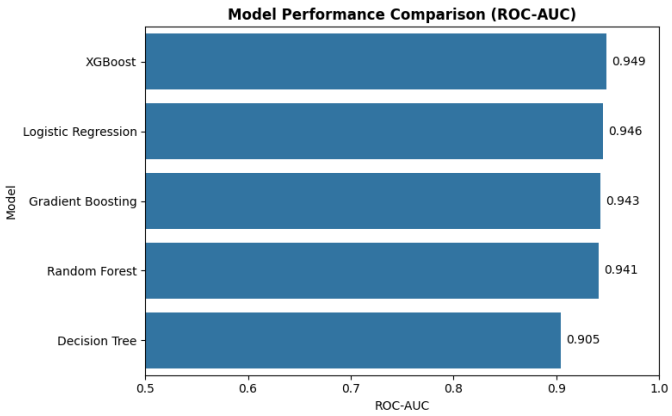
Several modeling approaches were tested to balance **performance, reliability, and ease of interpretation**:

- **Logistic Regression (baseline)**
- **Decision Tree (nonlinear reference)**
- **Random Forest (ensemble benchmark)**
- **Gradient Boosting**
- **XGBoost**

All models were trained using the same data preparation process to ensure a fair comparison. Since higher-income individuals are relatively rare, the training process explicitly accounted for class imbalance so the models would not default to predicting the majority group.

Results – Part 1

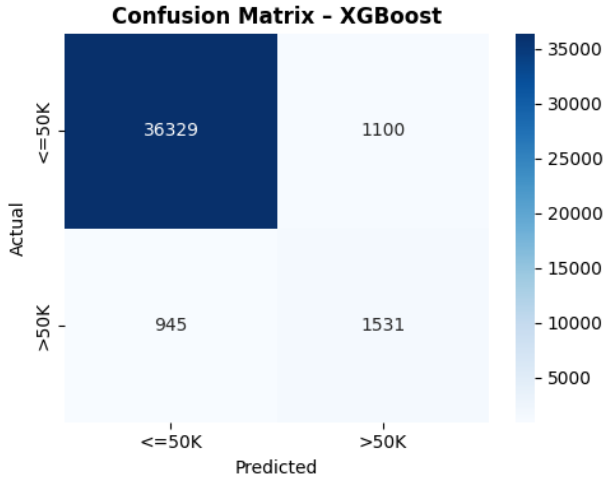
Overall Model Performance



Across all approaches, ensemble-based models significantly outperformed simpler methods. This confirms that income is influenced by **multiple interacting factors**, rather than any single variable.

Among all tested models, **XGBoost delivered the strongest overall performance**, with consistently high accuracy in distinguishing between income groups across different decision thresholds.

What Errors the Model Makes



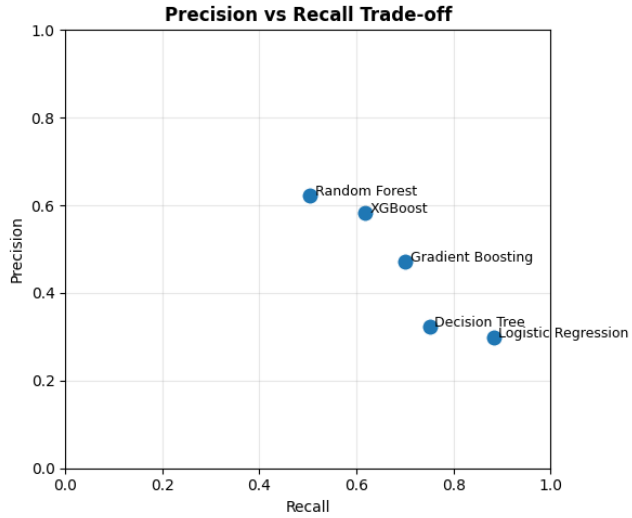
To understand how the model behaves in practice, outcomes were examined at the selected operating threshold. The results highlight three key behaviors:

- Most lower-income individuals are correctly identified, limiting unnecessary outreach
- A meaningful share of higher-income individuals is successfully flagged despite their small representation
- Some higher-income individuals are missed, reflecting an intentional trade-off to control outreach volume

Overall, the model achieves a balance between coverage and efficiency that aligns with practical marketing use cases.

Given the strong class imbalance and marketing use case, ROC-AUC was used as the primary model comparison metric, while recall for the high-income class was emphasized during threshold selection. Missing a high-value individual is costlier than sending a limited number of additional offers, making recall a more appropriate optimization objective than raw accuracy.

Precision vs Recall Tradeoff

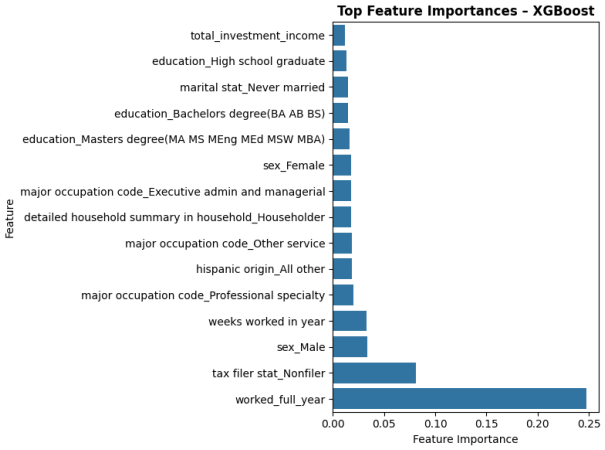


Different models emphasize different trade-offs:

- Some models prioritize finding as many high-income individuals as possible but generate many false positives.
- Others are more conservative but miss too many valuable customers.

XGBoost provides a balanced middle ground, identifying a strong share of high-income individuals while maintaining reasonable precision. This makes it suitable for real-world deployment where both budget and reach matter.

What Drives the Predictions?



Consistent patterns are observed across XGBoost feature importance and SHAP analysis. The model relies most heavily on:

- Employment stability (e.g., full-year employment, weeks worked)
- Tax filing behavior
- Education and occupation
- Investment-related income signals as secondary contributors

Feature influence is distributed rather than dominated by a single variable, increasing confidence that predictions are driven by meaningful patterns rather than shortcuts.

Why was XGBoost Selected?

XGBoost was selected as the final model because it consistently provides the most reliable separation between higher-income and lower-income individuals. Performance remains stable across a range of targeting thresholds, allowing outreach strategies to be adjusted without degrading model behavior.

From a marketing perspective, the model achieves a balanced error profile that supports efficient targeting, identifying a meaningful share of higher-value individuals while limiting unnecessary outreach. In addition, the model scales well to high-dimensional customer data and offers clear insight into the factors driving predictions, supporting transparency and informed decision-making.

Rather than relying on a fixed classification cutoff, targeting thresholds can be tuned based on campaign budget and desired reach. This flexibility allows coverage and precision to be traded off dynamically without retraining the model, making the approach well-suited for ongoing marketing use.

Part 2: Customer Segmentation

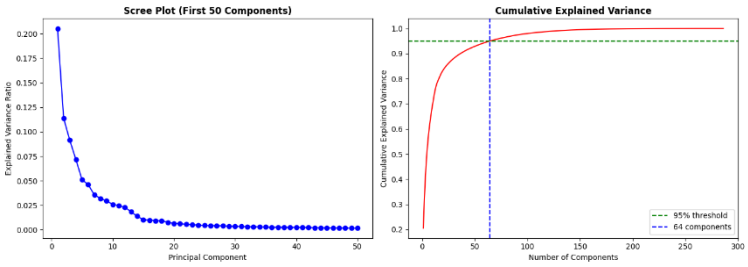
Business Objective

While the income prediction model identifies *who* is more likely to earn above $50,000, segmentation is used to understand *how* different groups within the population differ in meaningful ways. The objective of this analysis is to group individuals into distinct, interpretable segments that can support differentiated marketing strategies, messaging, and budget allocation.

Rather than treating all individuals within an income group as homogeneous, segmentation provides additional context on employment patterns, financial behavior, and life stage, enabling more nuanced and effective engagement.

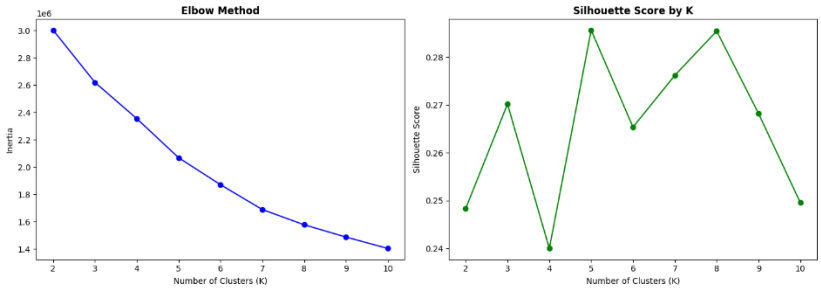
Modelling Techniques Used

Dimensionality Reduction



To reduce noise and improve clustering stability, Principal Component Analysis (PCA) was applied prior to segmentation. A reduced set of components captures the majority of variation in the data, allowing clustering to focus on underlying behavioral patterns rather than redundant features.

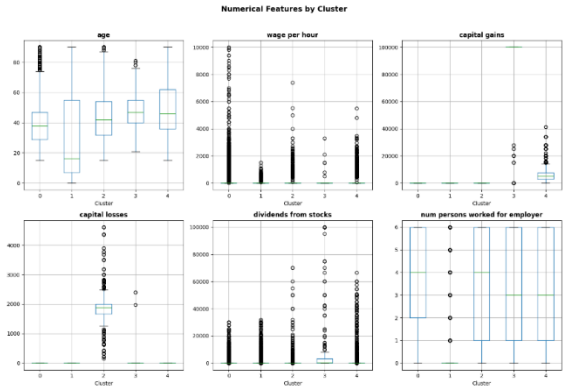
Clustering Approach

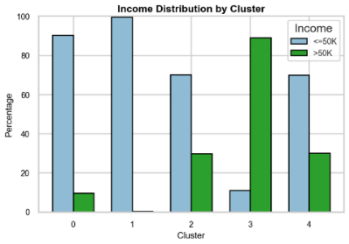


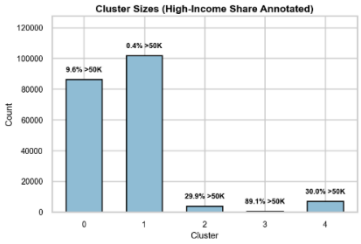
Following dimensionality reduction, K-Means clustering was used to group individuals with similar demographic, employment, and financial characteristics. Both the elbow method and silhouette analysis indicate that five clusters provide a strong balance between interpretability and separation.

Results – Part 2

Cluster Characteristics



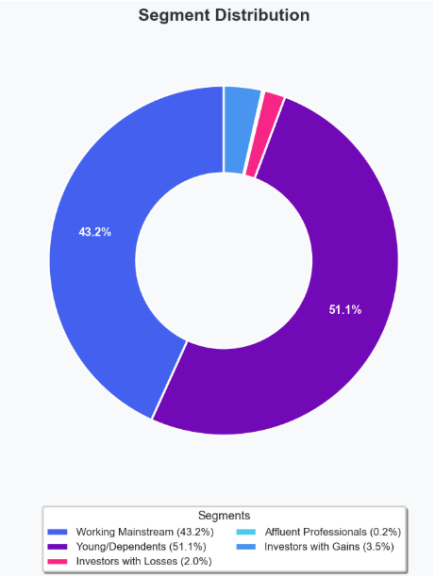




The resulting clusters exhibit clear differences across key characteristics such as age, employment intensity, wages, and investment-related variables. These differences confirm that the segmentation captures meaningful economic and behavioral distinctions rather than arbitrary groupings.

Analysis of income composition across clusters shows substantial variation in the proportion of individuals earning above $50,000. Notably, cluster size does not directly correspond to customer value. Some smaller clusters contain a disproportionately high concentration of higher-income individuals, while larger clusters are dominated by lower-income earners.

Segment Interpretation

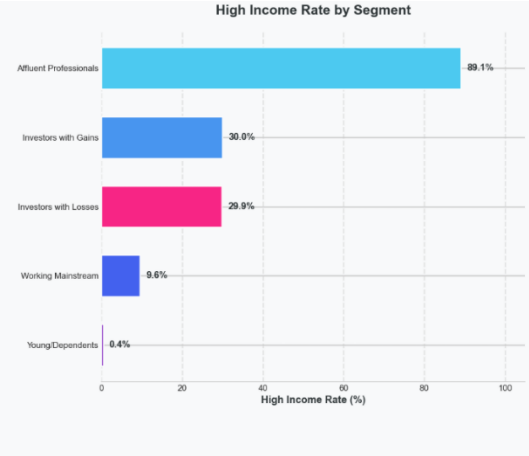


To support practical use, clusters were translated into business-friendly segments based on their dominant characteristics:

- **Young / Dependents**: Younger individuals with limited employment history and minimal financial activity
- **Working Mainstream**: The largest segment, characterized by steady employment and moderate income levels
- **Investors with Losses**: Individuals with notable investment activity but lower overall income
- **Investors with Gains**: A smaller segment with significant investment income and moderate employment stability
- **Affluent Professionals**: A high-value segment defined by stable full-year employment, higher education, and professional occupations

These labels provide a shared language for aligning marketing, messaging, and budget decisions.

High-Income Concentration by Segment



The **Affluent Professionals** and **Investors with Gains** segments exhibit the highest concentration of higher-income individuals despite representing a smaller share of the population. In contrast, the **Working Mainstream** segment contains many individuals but a substantially lower proportion of high-income earners.

This distinction enables prioritization of high-value segments without sacrificing awareness of the broader customer base.

Justification of the Segmentation Approach

This segmentation approach provides clear business value:

- It converts a large, heterogeneous population into **interpretable customer groups**
- It enables **segment-specific marketing strategies**, rather than uniform targeting
- It distinguishes between **high-volume** and **high-value** segments
- It complements the income prediction model by adding **context and interpretability**

Importantly, the segmentation is stable, scalable, and can be reused as new customer data becomes available.

Risks and limitations

The results in this report should be interpreted with the following considerations in mind:

- The dataset reflects historical census data and may not fully capture recent shifts in labor markets, income structures, or employment patterns.
- Income is modeled as a binary outcome, which simplifies interpretation but may mask meaningful variation within income groups.
- Observed relationships reflect existing socioeconomic patterns and should be used to inform targeting strategies responsibly.
- Model performance and segment composition should be monitored over time, with periodic retraining recommended as population characteristics evolve.

This ensures that the models remain accurate, relevant, and aligned with business objectives as conditions change.

Closing Note and Recommended Next Steps

The analysis provides a clear, data-driven foundation for improving marketing effectiveness through both targeted outreach and segment-based engagement.

Key takeaways:

- Higher-income individuals can be reliably prioritized using a combination of employment stability, education, occupation, and financial activity signals.
- Customer value is unevenly distributed across segments, with smaller groups such as *Affluent Professionals* and *Investors with Gains* contributing disproportionately high value.
- A single, uniform marketing strategy is unlikely to perform equally well across all customer groups.

Recommended next steps:

- Use the income prediction model to prioritize outreach toward individuals with a higher likelihood of belonging to the high-income group.
- Apply segment-specific messaging strategies, particularly for professional and investor-oriented segments.
- Adjust targeting thresholds dynamically based on campaign budget, desired reach, and performance feedback.
- Monitor model performance and segment distributions over time to ensure continued effectiveness.

By operationalizing these insights, marketing efforts can become more focused, efficient, and aligned with underlying customer value.

References

- **U.S. Census Bureau.** *Current Population Survey (1994–1995).*
  Source dataset used for modeling and segmentation.
- **Chen, T., & Guestrin, C. (2016).** *XGBoost: A Scalable Tree Boosting System.*
  This paper introduces XGBoost, the model used for income prediction in this report.
- **Lundberg, S. M., & Lee, S. I. (2017).** *A Unified Approach to Interpreting Model Predictions (SHAP).*
  Provides the methodology behind the interpretability analysis of model drivers.
- **Jolliffe, I. T. (2002).** *Principal Component Analysis.*
  Classic reference for PCA, used here to support dimensionality reduction prior to clustering.
- **Benjamin, M. (2025).** *Customer Segmentation and Behavior Analytics.* ResearchGate.
  A study exploring segmentation and behavior analytics in marketing contexts that supports the segmentation approach used in this report.
- **Kasem, M. S. E., Hamada, M., & Taj-Eddin, I. (2023).** *Customer Profiling, Segmentation, and Sales Prediction using AI in Direct Marketing.*
  Demonstrates the combination of profiling, segmentation, and predictive modeling in a marketing setting similar to this project.
- **Kasem, M. S. E. (2024).** *Customer Profiling, Segmentation, and Sales Prediction using Clustering Methods. Neural Computing and Applications.*
  Applies clustering for customer segmentation and validates cluster usefulness, corroborating the segmentation findings.