JP Morgan Take Home Project Report

Income Classification and Customer Segmentation for Retail Marketing

Tanmay Sharma | December 2025

Executive Summary | Business Objective

The retail client aims to improve marketing efficiency by identifying individuals who earn more than $50,000 versus those who earn $50,000 or less, using 40 available demographic and employment-related variables.
Using a labeled population dataset, the goal is to develop:

1. A **classifier** to prioritize individuals most likely to earn >$50K
2. A **segmentation model** to group people into actionable marketing audiences.

Together, these models enable targeted outreach and tailored messaging while controlling marketing costs.

Objective 1: Income Classification

Recommendation: Build a supervised classification model to predict whether an individual earns more than $50,000, while treating the model's probability outputs as a prioritization mechanism for marketing decisions instead of a hard classification rule.

Methodology and rationale

Trained and validated multiple classification models and selected a tuned LightGBM classifier due to its strong performance on imbalanced data and ability to model non-linear interactions among demographic and employment variables.

Rather than using a default probability cutoff, we treated the decision threshold as a business control to balance outreach volume and hit-rate.

Supporting evidence

- Multiple classification models were tested.
- A tuned LightGBM model delivered the strongest and most stable ranking performance on imbalanced data.

Metric	Value
Threshold	0.87
Precision	0.6157
Recall	0.6089
F1-score	0.6123
PR-AUC	0.6753

How can this be used to make marketing decisions

- High-cost channels (e.g., personalized discounts) target the top-ranked group.
- Lower-cost channels expand reach to lower-propensity bands.
- Thresholds can be adjusted as campaign economics change.

- Production deployment would require agreement on operating thresholds, channel-level costs, and fairness review.

Objective 2: Customer Segmentation

Recommendation: Adopt a five-segment customer model based on demographic, employment, and household characteristics. The following segments reflect clear differences in workforce attachment, income potential, and life stage. These groups are stable and interpretable, supporting consistent marketing use.

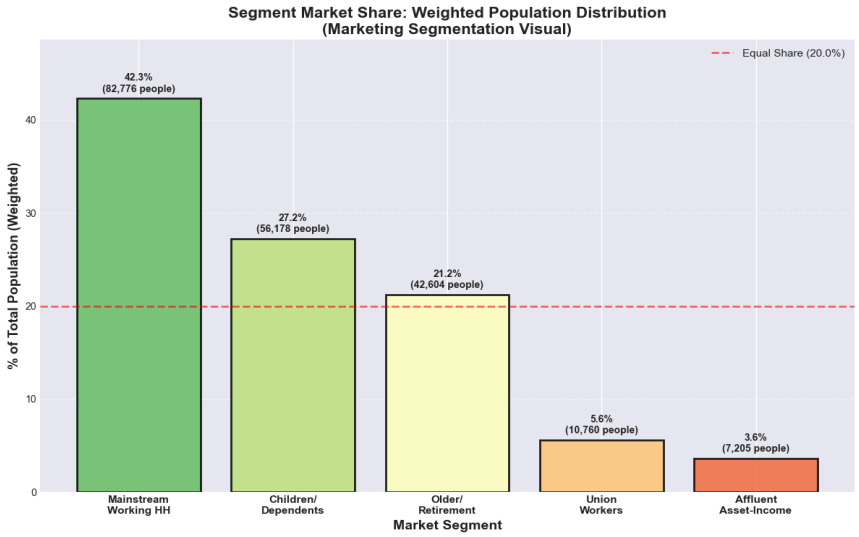


Fig 1: Segments for marketing

How can the segmentation be used for marketing purposes

- Affluent households: Premium, high-touch campaigns.
- Mainstream workers: Broad, value-focused messaging.
- Union-connected workers: Stability- and benefit-oriented offers.
- Older individuals: Trust-based and essential product campaigns.
- Children and dependents: Excluded from individual-level outreach.

Risks and mitigation plan

- **Bias:** Managed by using segments as high-level personas, not individual decision rules.
- **Data:** Census data reflects mid-1990s conditions and may not represent current markets
- Use segments strictly as **high-level marketing personas**, and require review of segment usage for compliance and fairness.

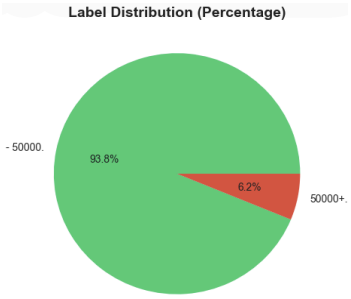
Controls and mitigations:

- Restrict use to marketing prioritization only.
- Conduct fairness testing before production deployment.
- Monitor model performance and feature drift over time.

Data Exploration: Data Quality, Feature Types, and Key Findings

Target label analysis

Income distribution is highly skewed toward lower-income individuals, making naive targeting inefficient. As a result, model evaluation and targeting decisions must prioritize precision–recall trade-offs rather than overall accuracy to avoid wasted outreach spend.



- Model evaluation should prioritize **precision/recall tradeoffs** and **PR-AUC**, which better reflect performance on rare positive classes.
- The prediction threshold should be treated as a **business control lever** to balance outreach volume and hit-rate.

Overall data quality assessment

Most demographic and employment variables are stable, interpretable, and suitable for marketing use at scale. However, special codes such as “Not in universe” represent meaningful life-stage states (e.g., children, retirees) and must be preserved to avoid bias targeting only actively employed adults. Also, missing and ‘?’ marks will be considered as ‘unknown’ category values. This decision improves downstream interpretability and prevents biasing the model toward only actively employed adults.

- **Financial variables with extreme values (wages and investment-related income) were normalized to prevent a small number of high-income outliers from disproportionately influencing targeting decisions.** This ensures more stable and fair prioritization across the population.
- **Categorical attributes were handled in a way that preserves interpretability while remaining operationally scalable.** Simple categories were kept transparent for stakeholder review, while complex categories were summarized to avoid overfitting and operational complexity.
- **No meaningful redundancy was found among numeric variables, allowing multiple income-related signals to be used together.** This supports a multi-factor view of customer value rather than reliance on any single metric.

Industry and occupation codes

Compared detailed numeric recodes versus major categories for industry and occupation. Major industry/occupation categories provide better interpretability because they are already binned and meaningfully named, whereas detailed codes lack usable labels in the provided dictionary. Therefore, major categories was used to ensure models and segments can be explained and used by marketing stakeholders.

Household, Family Structure, and Geographic Context for Objective 2

Observed household composition, family structure, and geographic mobility variables to understand how broader life context relates to income outcomes and marketing relevance.

- Household role and labor-force attachment emerge as strong, interpretable signals that naturally separate individuals **into meaningful life-stage groups, such as dependents, working-age adults, and retirement-age individuals**. These groupings align closely with income likelihood and purchasing capacity, making them particularly valuable for marketing segmentation.
- Survey weights further enable segments to be translated into population-level estimates, supporting campaign sizing and reach planning.

Geographic segmentation is poorly aligned with practical marketing geotargeting and offers limited lift relative to demographic and employment-based segmentation.

- Regional and migration-related variables provide limited incremental value for segmentation. The majority of records are coded as “Not in universe” (approximately 92%) for prior state or region fields, and these variables describe previous residence and mobility rather than current location.
- Among the small subset of individuals with valid migration data, category counts are sparse and income differences across regions are modest.

Deep dive: Objective 1 Income Classification for Marketing Targeting

Pre-processing Approaches, Model Architecture, Training Algorithm, and Evaluation Procedure

Pre-processing Strategy (Informed by Data Exploration)

1. **Preserved meaningful population groups:** “Not in universe” values were retained as valid life-stage indicators (e.g., children, retirees), preventing the model from over-focusing on actively employed adults and improving representativeness for marketing decisions.
2. **Ensured realistic performance assessment:** Model training and evaluation preserved the true income distribution in the population, ensuring reported results reflect real-world deployment conditions.
3. **Stabilized income signals:** Extreme-value financial variables were normalized to prevent outliers from disproportionately influencing predictions.
4. **Maintained consistency and integrity:** All data preparation steps were applied consistently across training and validation to avoid leakage and ensure reliable performance estimates.

Baseline Model Architecture and Motivation

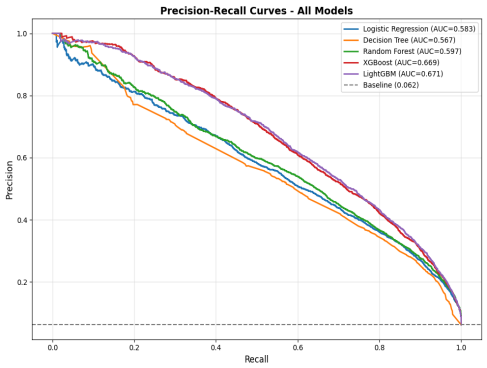
- **Interpretable baseline models (Logistic Regression, Decision Tree):**
  Simple classification models were evaluated first to establish a transparent performance baseline and to validate that income patterns are directionally sensible and explainable to non-technical stakeholders.

- **Advanced ensemble models (Random Forest, XGBoost, LightGBM):**
  Tree-based ensemble models were then assessed to capture complex, non-linear interactions across occupation, education, household, and employment variables, which are critical drivers of income differences.

While baseline models provided interpretability, ensemble methods offered materially stronger performance by leveraging interacting signals without manual feature engineering, making them better suited for **scalable marketing prioritization**.

Baseline Results and Evaluation Metrics

Results showed a clear progression in performance. Linear and single-tree models achieved limited F1-scores, indicating that income outcomes are not well captured by simple linear boundaries or shallow rules. Ensemble methods consistently improved performance, with boosting-based models (XGBoost and LightGBM) achieving the strongest test-set results across most metrics.



Feature Importance and Stability Analysis

- Income differentiation is driven by a small set of intuitive factors: occupation, education level, and weeks worked are the strongest and most consistent signals, with asset-related income(capital gains/dividends)providing incremental value beyond wages alone. This aligns well with how marketing teams might reason about customer value.
- More advanced models provide greater stability and reliability: ensemble models distribute influence across multiple drivers rather than relying on a few dominant variables, reducing volatility and improving generalizability. This makes them better suited for consistent, repeatable marketing prioritization.

Model Refinements and Final Design Choices

Several enhancements were evaluated to improve performance and practical usability. **Synthetic oversampling(SMOTE)** and **population weighting(weights feature in dataset)** were tested but did not improve real-world performance, and were therefore excluded to avoid unnecessary complexity.

Threshold tuning, however, materially improved business alignment. By evaluating multiple operating points instead of relying on a default cutoff, the model was calibrated to balance outreach efficiency and audience coverage via precision - f1 score tuning. A single, stable operating threshold was selected as a general-purpose configuration, allowing marketing teams to adjust targeting depth as **campaign costs and priorities change**.

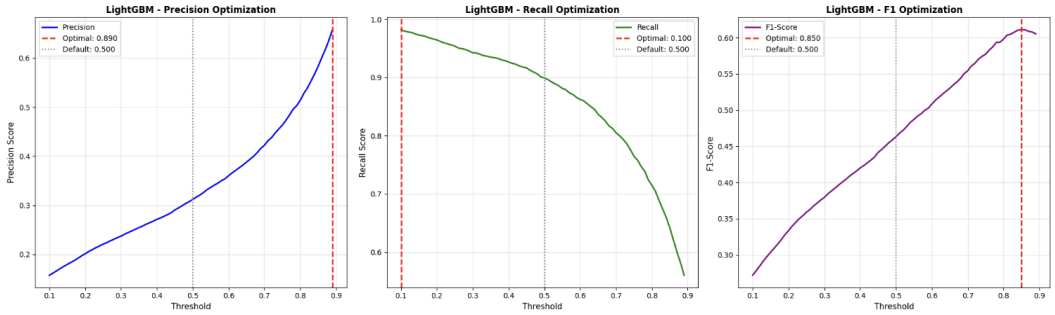


Fig-2: Threshold tuning: F-1 tuning gave the best result

Final Model Selection and Interpretation

The final model was refined through a targeted **hyperparameter search**, informed by earlier baseline and enhancement experiments, to identify a configuration that delivers a stable balance between targeting precision and audience coverage. This ensured that additional model complexity translated into meaningful performance gains rather than overfitting.

In addition, **decision threshold tuning** was applied to align model outputs with marketing priorities. Instead of relying on a default cutoff, the operating threshold was selected to balance outreach efficiency and reach. Based on overall performance, stability, and interpretability, a tuned LightGBM classification model with optimized thresholding was selected as the final solution.

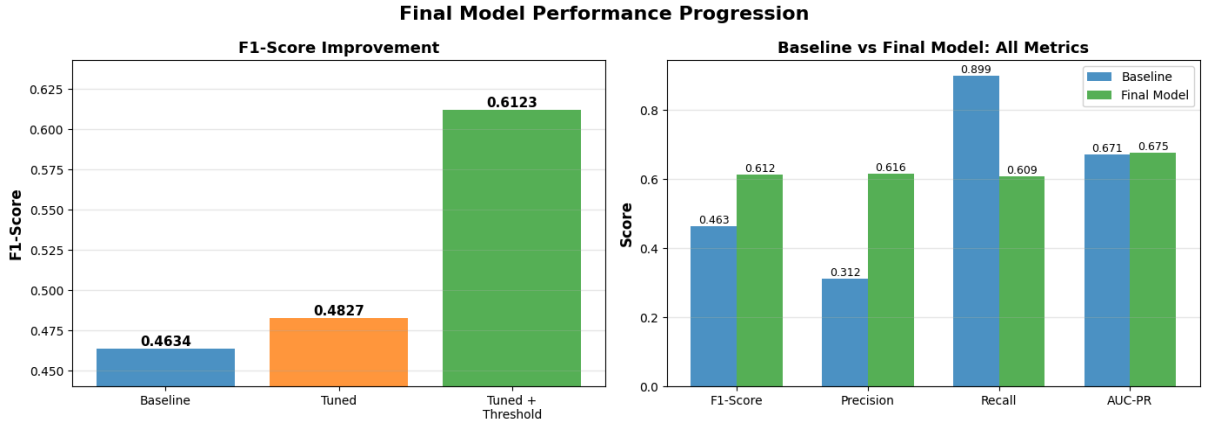


Fig-3: Final fine tuned LightGBM model

The deployment threshold was selected by optimizing the F1-score to balance precision and recall in a marketing context. This choice controls wasted outreach spend while maintaining sufficient coverage of high-income prospects, aligning model behavior with typical campaign efficiency and reach objectives.

Business Interpretation

The model performs most reliably for individuals with strong labor-force attachment, higher education or skilled occupations, and asset-related income signals. Individuals whose profiles are dominated by “Not in Universe” indicators consistently fall into the lower-income category and can be deprioritized for premium marketing, helping reduce inefficient spend.

Model Usage Recommendation

While the solution is built as a binary income classifier, it should be operationalized as a ranking and prioritization tool. Marketing teams can define probability-based audience tiers to tailor outreach intensity by allocating high-touch or higher-cost campaigns to top-ranked individuals, while using lower-cost channels or limiting outreach for lower-probability groups.

Assumptions and Open Questions

Addressing the following questions would allow further refinement of threshold selection and model usage in a production setting:

- Information about the client’s specific product or pricing structure: current approach assumes that the $50,000 income threshold serves as a reasonable segmentation boundary for differentiating higher- and lower-value customer groups.
- Explicit campaign cost or revenue data: model decisions were guided by a precision-leaning objective, with the assumption that incorrectly targeting lower-income individuals is generally more costly for retail marketing than missing some higher-income prospects.
- Practical deployment: how outreach costs vary by channel, whether targeting decisions should be made at the individual or household level.
- Regulatory, compliance, and fairness constraints apply to the use of demographic and employment-related features.

Deep dive: Objective 2 Customer Segmentation for Marketing Strategy

Segmentation Framework and Data-Driven Selection

- **Multiple segmentation approaches were evaluated** to align with retail marketing objectives, including demographic, geographic, psychographic, and behavioral.
- **Geographic segmentation was deprioritized** due to limited practical value: approximately 92% of location and migration fields are non-informative and primarily reflect prior residence rather than actionable targeting signals.
- **Demographic and economic variables provide the strongest and most interpretable separation.** Age, education, household composition, labor-force attachment, occupation, industry, and asset-income indicators align closely with life stage and purchasing capacity.

- **These dimensions support stable, actionable marketing personas** that can be consistently applied across campaigns.
- **The final segmentation strategy prioritizes demographic and economic factors**, supplemented by limited psychographic indicators where available, while intentionally excluding geographic segmentation due to insufficient signal and interpretability.

Features and Population Representation using Survey Weights

The customer segmentation model was built using a curated set of demographic, household, and economic variables selected for their interpretability, data completeness, and direct relevance to marketing decisions. Feature selection was informed by exploratory analysis and income modeling, ensuring segments align with life stage, workforce attachment, and purchasing capacity rather than abstract statistical patterns.

Key feature groups used in segmentation include:

- **Demographic and life-stage indicators:** age, marital status, education level, household and family structure, number of dependents under 18, recent educational enrollment, and citizenship status.
- **Economic and employment indicators:** major occupation and industry, class of worker, full- or part-time status, weeks worked per year, self-employment status, employer size, labor union membership, and income-related variables such as wages and investment income.

Together, these variables capture **workforce participation, income stability, and asset ownership**, which are core drivers of consumer purchasing behavior and marketing responsiveness.

Survey weights were applied when profiling final segments to ensure population-representative market sizing. While clustering was performed on unweighted data for stability, weights were used during interpretation to support accurate estimates of campaign reach and resource allocation.

Clustering Methodology and Segment Selection

Customer segments were designed to group individuals with similar demographic and economic profiles into clear, actionable marketing personas. A scalable and interpretable clustering approach was used to ensure the segments are stable, easy to explain, and practical for campaign execution.

Multiple segment configurations were evaluated to balance insight and usability. While a six-segment structure provided strong differentiation, one segment was too small to be operationally meaningful and was merged with its closest peer. The final result is a five-segment framework that:

- Differentiates customers by life stage, workforce attachment, and income potential
- Supports meaningful campaign sizing and resource allocation
- Avoids over-fragmentation that adds complexity without incremental business value

This structure ensures the segmentation is both analytically sound and fit for real-world marketing use.

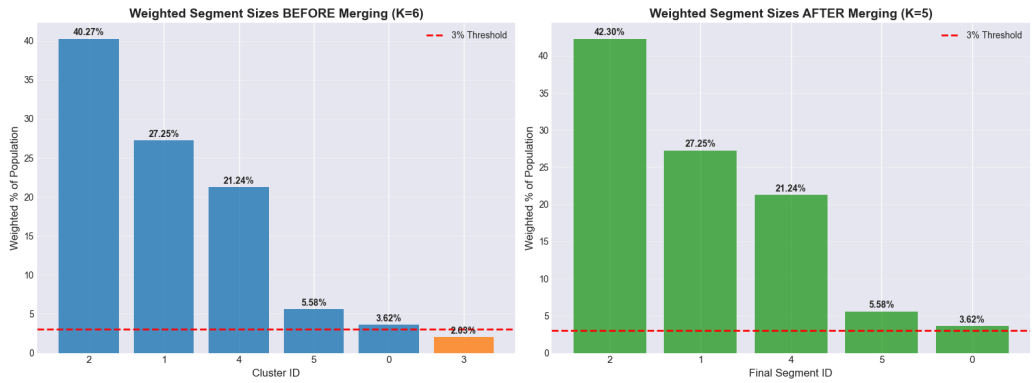


Fig-4: Left: 6 segments from K-means; Right: 5 Clusters after combining

Population-Weighted Segment Overview

Final segment profiles were constructed using survey-weighted statistics to reflect population-level composition. The market is highly concentrated, with three segments accounting for approximately **91% of the population**—representing mainstream working households, dependents/children, and older individuals with limited workforce attachment.

Income opportunity, however, is **unevenly distributed across segments**. A small, high-value segment shows a disproportionately high likelihood of earning more than $50,000, while a much larger segment offers moderate income opportunity at scale. Across all segments, **workforce attachment is the primary differentiator**, followed by occupation, education, and asset-income indicators.

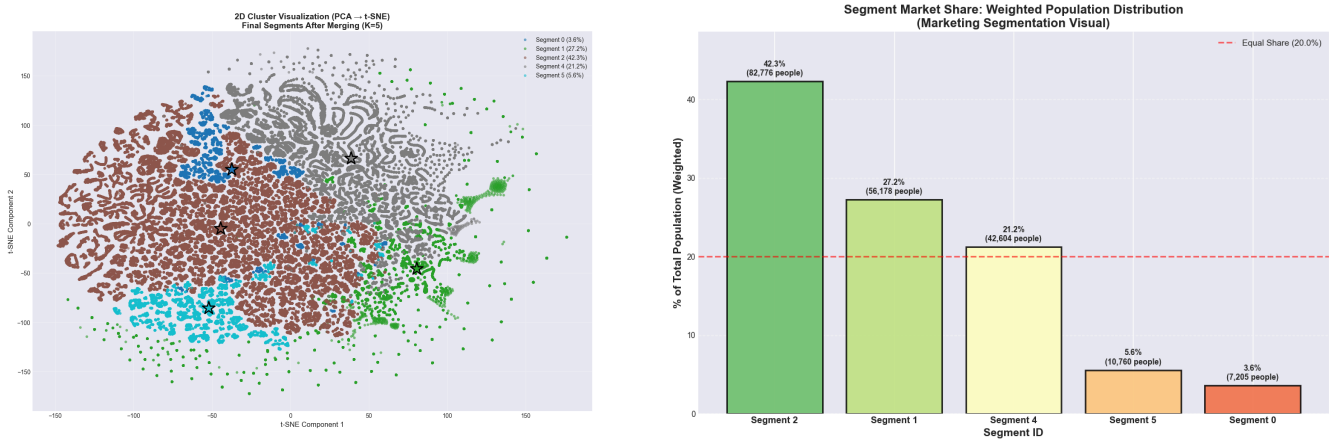


Fig-6: Left: Clusters t-SNE map; Right: Segments relative to population weights

Marketing Segment Personas

Segment 2 - Mainstream Working Households (Retail/Admin Core): This is the largest segment (42.3%), characterized by steady employment (approximately 44 weeks worked per year), concentration in retail and administrative occupations, and moderate income opportunity (~10.9% earning more than $50,000). This segment represents the highest-volume target for broad-based marketing campaigns.

Segment 1 - Children and Dependents: This segment has a very low average age, near-zero workforce participation, and virtually no high-income likelihood. It should be excluded from income-based targeting and considered only in household-level marketing strategies.

Segment 4 - Older, Retirement-Age, Low Workforce Attachment: An older population with minimal labor-force participation and low income potential. This group is a low priority for premium marketing but may be relevant for necessity- or trust-oriented messaging.

Segment 5 - Union-Connected Hourly Workers: A smaller but distinct segment (5.6%) with full-time employment, strong retail/admin representation, and elevated labor union membership. This segment supports tailored messaging focused on value, benefits, and stability.

Segment 0 - Affluent Asset-Income Households: A small but high-value segment (3.6%) with strong asset-income signals and the highest likelihood of earning more than $50,000 (~33.6%). This group is ideal for premium offerings and high-ROI campaigns.

Assumptions and open questions

Given limited information about the retail client's specific products or marketing channels, the segmentation model is constructed using demographic and economic variables from the U.S. Census Bureau dataset (1994-1995) that are commonly associated with life stage, purchasing capacity, and labor-force attachment. The resulting segments are designed to be product-agnostic personas that can support a range of retail marketing strategies.

Resources

1. **Dataset:** U.S. Census Bureau Current Population Survey (1994–1995),
2. **Models:** Supervised classification models and K-means clustering for segmentation.
3. **Reference:** Investopedia — *Market Segmentation*. https://www.investopedia.com/terms/m/marketsegmentation.asp

Use of AI Tools

AI-assisted tools were used selectively to support this project: Structuring analysis, refining written explanations, and improving clarity of presentation. All modeling decisions, data preprocessing steps, feature selection, experiments, and interpretations were designed, implemented, and validated by the author, Tanmay Sharma.