Income Classification & Customer Segmentation — Client Report

Author: Sree Harsha Koyi
Date: 2025-09-04
Audience: Retail business stakeholders (marketing, CRM), data science reviewers

---

1. Introduction

The retail business client presented us with two interrelated objectives that are fundamental to modern marketing. First, we were asked to design a classification model that predicts whether an individual earns less than $50,000 or at least $50,000 annually. This binary classification problem enables the client to distinguish between lower-income and higher-income segments of the population for targeted campaigns. Second, we were asked to develop a segmentation model that divides the population into coherent and interpretable clusters. These clusters would serve as actionable personas that the marketing team could address with tailored strategies.

The dataset provided originates from the **Current Population Surveys (CPS)** of 1994–1995, conducted jointly by the U.S. Census Bureau and the Bureau of Labor Statistics. This dataset includes 40 demographic and employment-related variables, a survey weight for each observation (to ensure representativeness of the U.S. population), and a label indicating whether the person earns above or below $50,000. The CPS is widely recognized as the most authoritative dataset on labor and income, which lends credibility to our analysis but also necessitates careful handling of weights and demographic complexity.

Our philosophy throughout this project was to balance technical rigor with practical applicability. We placed significant emphasis on making every step reproducible, transparent, and directly translatable into business impact. This report details our exploration of the data, the preprocessing pipeline we designed, the models we trained, the evaluation results, and finally our recommendations for how to use these models to drive real marketing outcomes.

2. Data Exploration and Preprocessing

Dataset Composition

The dataset contained **199,523 records** and **42 columns**. Of these, 40 were demographic and employment-related features, while the remaining two were the survey weight and the income label. The columns covered a wide variety of socioeconomic dimensions including age, education, marital status, occupation, industry, capital gains, capital losses, and hours/weeks worked.

We observed the following breakdown:

- **12 integer columns**: features like *weeks worked in year*, *age*, and *number of persons worked for employer*.
- **1 float column**: continuous variables like *wage per hour*.
- **29 categorical columns**: features such as *education*, *occupation*, *industry code*, and *tax filer status*.

Missingness Analysis

Nine columns exhibited missing data, with missingness encoded as ?. The extent of missingness varied:

- *Hispanic origin*: 874 missing values.
- *State of previous residence*: 708 missing values.
- Migration-related variables (e.g., *migration code-change in MSA*, *migration code-change in region*, *migration code-move within region*): each had nearly **100,000 missing values**. This reflects structural missingness, as many respondents had not recently migrated.

We addressed missingness by distinguishing between structural and incidental missingness. Structural missingness (like non-migrants in migration fields) was retained as categorical categories to preserve meaning, while incidental missingness (random omissions) was imputed using the most frequent category for categorical variables or median imputation for numeric variables.

Special Tokens

Several categorical variables included values such as "Not in universe." These indicated that the field was not applicable to the respondent (e.g., employment fields for children or retirees). Rather than treat these as missing, we preserved them as valid categories. This decision was important from a business perspective: "Not in universe" is itself meaningful, as it captures entire population groups like dependents or retirees who should be marketed differently.

Label Distribution

We mapped income labels containing a + to **1 (≥$50k)** and others to **0 (<$50k)**. Weighted by survey weights, the class distribution was highly imbalanced:

- **< $50k:** ≈ 325 million people (93.6% of population)
- **≥ $50k:** ≈ 22 million people (6.4% of population)

This imbalance revealed an immediate challenge: a naïve classifier could achieve 93% accuracy by always predicting < $50k. Thus, standard accuracy would be misleading. This insight shaped our evaluation methodology, leading us to prioritize ROC AUC, precision, recall, and F1 score.

Preprocessing Strategy

Our preprocessing pipeline included:

- **Numeric columns**: coerced to numeric, with invalid tokens converted to NaN. Missing values were imputed using the median. All numeric columns were standardized to zero mean and unit variance.
- **Categorical columns**: ? replaced with NA, imputed using the most frequent value, and then one-hot encoded. Unseen categories at inference were safely ignored.
- **Survey weights**: excluded from feature vectors but incorporated as sample_weight in all training and evaluation steps.
- **Dimensionality reduction**: high-dimensionality from one-hot encoding was managed by mutual information feature selection (top 50%) followed by Truncated SVD with 100 components. This reduced noise, improved training speed, and made the models more stable.

3. Model Architectures and Training Algorithms

We trained and compared two broad families of models: a supervised classifier for income prediction and an unsupervised segmentation model for persona identification.

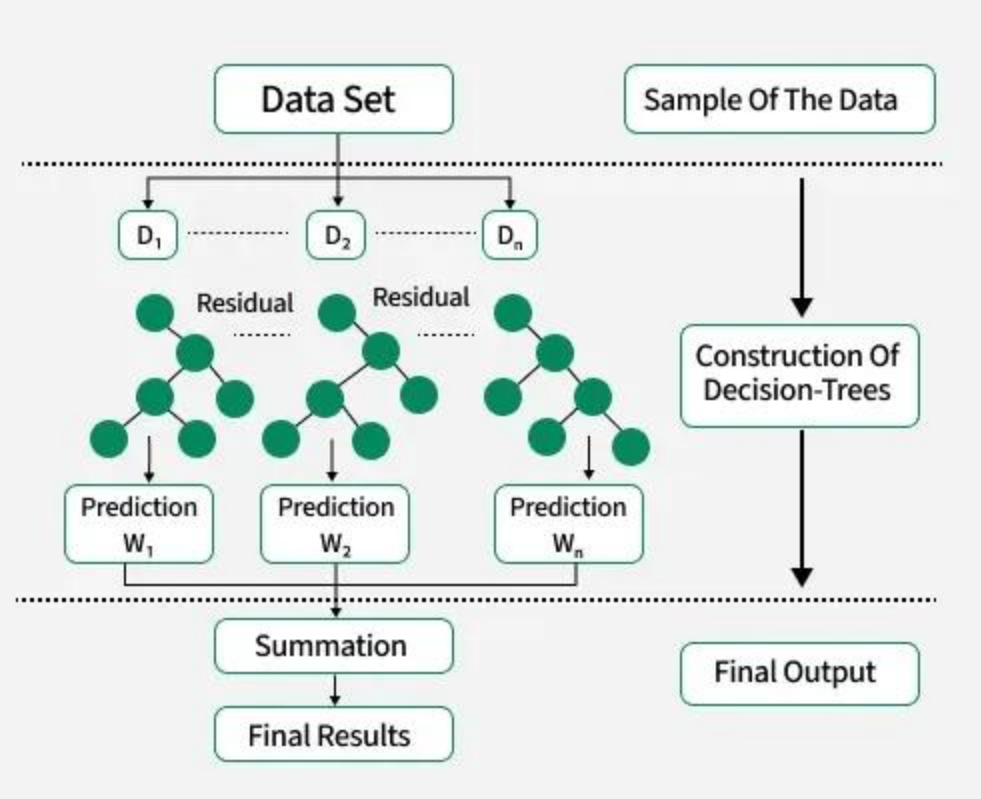
3.1 Classification Models

We implemented three classifiers as baselines:

- **Logistic Regression**: simple, interpretable, with L2 regularization.
- **Random Forest**: 300 estimators, unlimited depth, bagging for variance reduction.
- **XGBoost**: gradient boosting with 300 boosting rounds, max_depth=6, learning_rate=0.1, tree_method="hist". Regularization parameter λ=1.0, subsample=0.8, colsample_bytree=0.8.

Training Algorithm:
All classifiers were wrapped in pipelines including preprocessing steps. Training incorporated survey weights. XGBoost, the most complex model, was trained for 300 boosting rounds, each adding a shallow CART tree based on gradient and Hessian statistics for logistic loss. This was analogous to running multiple epochs until convergence. Subsampling and regularization mitigated overfitting.

Architecture Diagram:



3.2 Segmentation Model

We used KMeans clustering on the same preprocessed feature space. After testing different k values, we chose **k=6**, which balanced interpretability and separation. Each record was assigned to one of six clusters. Clusters were then profiled into business-ready personas using weighted means and dominant categories.

---

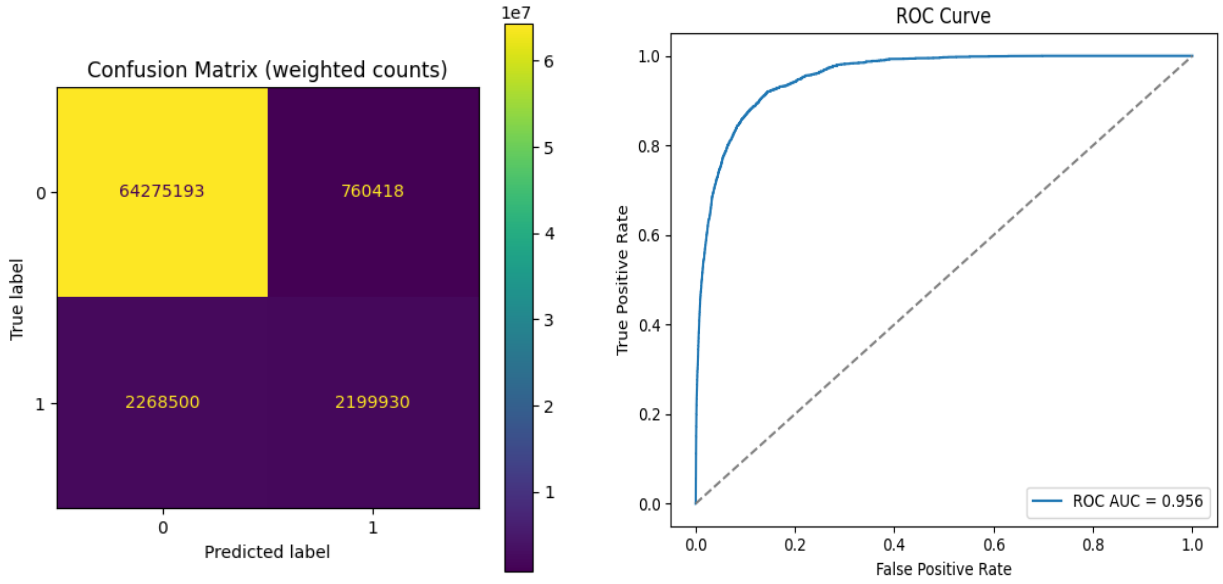
4. Evaluation Procedure and Metrics

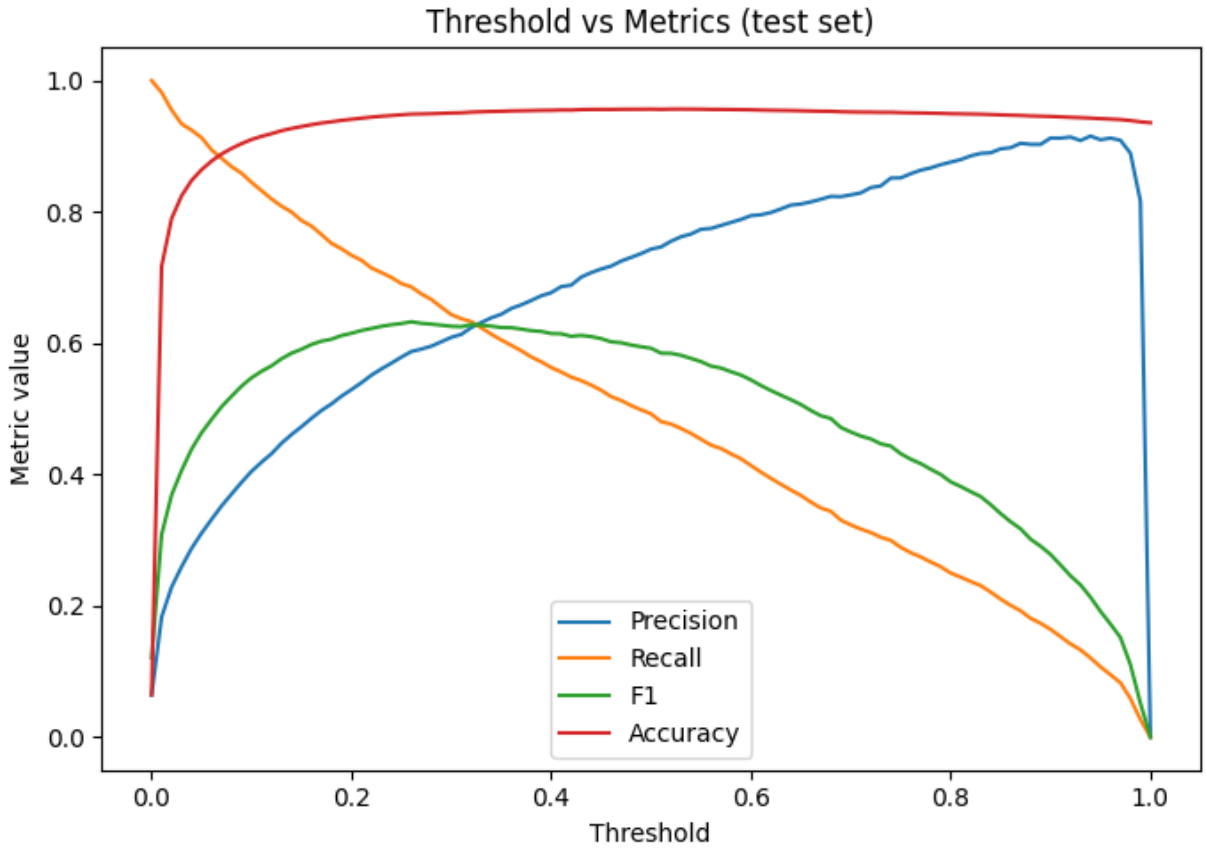
4.1 Classification Evaluation

We evaluated models using a rigorous suite of metrics:

- **ROC AUC:** XGBoost achieved ≈0.956 on test data. ROC AUC was chosen as the primary metric because it is insensitive to thresholds and directly reflects the model's ability to rank higher-income individuals above lower-income ones.
- **Precision and Recall:** At the best-F1 threshold (≈0.26), precision was ≈0.59 and recall ≈0.69. This balance ensured we did not miss too many high-income individuals while keeping false positives manageable.
- **F1 Score:** ≈0.63 at the optimized threshold, a strong balance metric.

- **Accuracy:** ≈0.95 at τ=0.26. We explicitly de-emphasized accuracy because of class imbalance.
- **Confusion Matrix:** At τ=0.50, many high-income individuals were missed (false negatives). At τ=0.26, recall improved significantly.





Business judgment: We concluded that adjusting the threshold is more impactful than seeking marginal improvements in the model algorithm itself. For awareness campaigns, lowering the threshold captures more potential customers. For expensive campaigns, a higher threshold ensures precision.

4.2 Segmentation Evaluation

- **Silhouette Score:** ≈0.19 for k=6. Though modest, this is acceptable in high-dimensional sparse data.
- **Interpretability:** Segments were coherent and interpretable. Weighted averages revealed meaningful distinctions in age, occupation, education, and labor force status.
- **Cluster Profiles:** Business relevance was emphasized in profiling, not just statistical separation.

---

5. Interesting Findings and Exploration

Classification Insights

- **Permutation Importance:** The most influential features for classification included *age*, *weeks worked in year*, *tax filer status*, *education*, *capital gains*, *dividends from stocks*, and *sex*. These aligned well with intuitive socio-economic drivers.
- **Threshold Analysis:** Adjusting τ from 0.50 to 0.26 increased recall dramatically while retaining respectable precision and accuracy. This trade-off is central to campaign design.

Segmentation Insights

To complement classification, we developed a segmentation model using KMeans clustering on the same preprocessed feature space. After evaluating multiple cluster counts, we selected **six clusters**, striking a balance between interpretability and statistical separation (silhouette score ≈ 0.19, which is acceptable in high-dimensional spaces).

The resulting clusters represent distinct personas:

- **High-Earning Professionals (Cluster 2):** Mid to late career individuals with strong educational backgrounds and significant capital gains. Suitable for premium product and financial service targeting.
- **Prime Working Age Retail/Clerical (Clusters 0 & 4):** Middle-aged individuals in retail or clerical roles, skewed male and married. Best addressed through value-driven offers and loyalty programs.

- **Full-Time Hourly Workers (Cluster 5):** Individuals working long weeks in manufacturing or manual labor. Effective targets for practical goods, tools, and durable products.
- **Older Adults Not in Labor Force (Cluster 3):** Predominantly older, retired, or not working. Marketing should focus on healthcare, retirement, and leisure services.
- **Dependents/Children (Cluster 1):** Non-earners who should not be directly targeted, but relevant via household-level campaigns (e.g., family products).

These segments provide clear business levers, allowing marketing teams to tailor messaging, channels, and offers to maximize relevance and ROI.

---

6. Business Recommendations and Business Judgment

Our modeling choices were guided not only by technical performance but also by business implications:

- **Model Selection:** Logistic Regression and Random Forest were trained as baselines. XGBoost outperformed both on ROC AUC and F1, so we selected it as the final model. This decision balances predictive lift with manageable complexity.
- **Transparency vs. Performance:** Logistic Regression remains valuable for explainability, but XGBoost provides the highest ROI by identifying more high-income prospects accurately.
- **Threshold Tuning as a Lever:** Threshold adjustment is a business knob. A lower τ maximizes reach, while a higher τ conserves budget.
- **Segmentation as a Complement:** Personas derived from clustering translate directly into creative strategies. For instance, premium offers for professionals, loyalty programs for clerical workers, and healthcare promotions for retirees.

Recommendations

1. Deploy the XGBoost classifier with calibrated thresholds tailored by campaign type.
2. Use segmentation to personalize messaging and channels per cluster.
3. Conduct A/B testing to quantify campaign uplift and optimize cost per acquisition.
4. Monitor fairness across demographic subgroups, recalibrating thresholds if disparate impacts arise.
5. Retrain models periodically with updated census or customer data to prevent drift.

---

7. Risks and Limitations

- **Outdated Data:** CPS 1994–1995 may not reflect today's demographics. Results should be validated on more recent datasets.
- **Fixed Threshold Definition:** The $50,000 income threshold is outdated in real terms. Inflation-adjusted thresholds would improve relevance.
- **Fairness Risks:** Socio-demographic features risk encoding bias. Ethical use requires subgroup audits.

---

8. Implementation and Next Steps

Deliverables include code, metrics, plots, and artifacts. The classifier can be deployed as an API. We recommend:

- Publishing a model card documenting intent, metrics, and risks.
- Establishing retraining cadence.
- Exploring advanced clustering methods like Gaussian Mixtures for softer, overlapping segments.

---

9. References

- scikit-learn User Guide (pipelines, preprocessing, model evaluation)
- Hastie, Tibshirani, Friedman — *The Elements of Statistical Learning*
- Kaufman & Rousseeuw — *Finding Groups in Data* (silhouette and clustering)

---

10. Conclusion

The classification and segmentation models together provide a dual strategy for marketing optimization. The classifier achieved strong discriminatory power (ROC AUC ≈0.956) and offers a flexible threshold to balance reach and precision. The segmentation revealed six actionable personas, each with distinct marketing strategies. By combining these two approaches, the client gains a powerful toolkit to maximize campaign ROI, engage customers more personally, and manage marketing budgets effectively.