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

Your retail business tasked us with two major objectives. First, we needed to design a classifier capable of predicting whether an individual earns less than $50,000 or at least $50,000 annually. Second, we were asked to build a segmentation model that would divide the population into distinct, actionable groups for marketing purposes. The dataset provided originates from the Current Population Surveys (CPS) of 1994–1995 and includes 40 demographic and employment-related variables, a survey weight for each observation, and an income label.

Our approach emphasizes both methodological clarity and practical applicability. We ensured that each step of data processing, model training, and evaluation was reproducible and communicated in plain language. This report describes our workflow, findings, and recommendations for how these models can be applied in a real business context.

---

2. Data Exploration and Preprocessing

We began by examining the raw CPS dataset, which contained **199,523 records across 42 columns**. These included 40 demographic and employment variables, a survey weight, and an income label. The dataset mixes numeric and categorical variables: **12 integer columns, one float, and 29 categorical fields**.

Missingness

Nine columns exhibited missing data, encoded as `?`. For instance: - *Hispanic origin* had 874 missing values. - *State of previous residence* had 708. - Migration-related fields (e.g., *migration code-change in MSA*) had nearly **100,000 missing values each**, reflecting the fact that many individuals were not recent movers.

We treated these systematically: for categorical variables, `?` was replaced with nulls and imputed with the most frequent category; for numeric columns, invalid entries were coerced to NaN and replaced with the median. This ensured no data leakage while preserving population distributions.

Special Tokens

Several variables contained tokens such as “Not in universe,” which indicate that the question did not apply to that respondent (for example, work-related fields for children or retirees). We interpreted these as categorical categories rather than missingness, since they carry meaning.

Label Distribution

The income label was mapped so that values containing a `+` were coded as **1 (≥$50k)** and all others as **0 (<$50k)**. Using the provided weights, the class distribution was highly imbalanced: - **< $50k:** ≈ 325 million people (93.6% of weighted population) - **≥ $50k:** ≈ 22 million people (6.4%)

This imbalance motivated our choice of **ROC AUC** as the primary selection metric (threshold-independent and weight-aware) and guided our threshold tuning to balance recall vs. precision.

Preprocessing Pipeline

- **Numeric columns:** coerced to numeric, median-imputed, and standardized.
- **Categorical columns:** `?` replaced with nulls, imputed with the most frequent category, and one-hot encoded (ignoring unseen categories at inference).
- **Survey weights:** excluded from the feature matrix but consistently applied as `sample_weight` during training and evaluation.
- **Dimensionality reduction:** we applied **mutual information feature selection (top 50%)** followed by **Truncated SVD (100 components)** to reduce the variance introduced by one-hot encoding and improve runtime stability.

This careful exploration and preprocessing laid the foundation for reliable modeling, ensured representativeness through weights, and respected the quirks of the CPS dataset.

---

3. Classification Modeling

We selected three representative classification algorithms to balance interpretability and predictive performance: - Logistic Regression with L2 regularization, serving as a linear and interpretable baseline. - Random Forest with 300 estimators, capturing non-linear relationships through ensembles of decision trees. - XGBoost with 300 estimators, max depth of 6, and learning rate 0.1, chosen for its efficiency and ability to capture complex feature interactions.

Model selection was based on weighted 5-fold cross-validation using ROC AUC as the evaluation metric. This ensured that the class imbalance and survey weights were properly accounted for. The best-performing model was then evaluated on a held-out test set.

---

4. Classification Results

The final selected model achieved a **ROC AUC of approximately 0.956** on the test set, demonstrating strong ability to rank individuals by income likelihood.

At the default probability threshold of 0.50, the model yielded high accuracy but relatively modest recall. A confusion matrix of weighted counts highlighted this trade-off, with the model correctly identifying the majority of low-income individuals but missing a considerable portion of higher-income individuals.

By analyzing threshold-dependent performance, we found that lowering the threshold to **0.26** maximized the F1 score (≈ 0.633), improving recall to ≈ 0.686 while maintaining an accuracy of ≈ 0.949. This suggests that the optimal threshold should be chosen based on campaign cost and objectives: lower thresholds for broader reach, higher thresholds for precision-driven targeting.

Feature importance analysis revealed that **age, weeks worked in year, tax filer status, education level, and capital gains/dividends** were among the most influential predictors. These features align well with common socio-economic indicators of income, lending face validity to the model.

---

5. Segmentation Model

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

1. **Targeting Strategy:** Use the classifier's probability scores to rank individuals. Apply lower thresholds ($\approx$0.25–0.35) for broad awareness campaigns and higher thresholds ($\approx$0.50–0.65) for expensive or high-value campaigns.

2. **Segment-Specific Campaigns:** Tailor offers and creative messaging by segment. For instance, promote premium services to high-earning professionals while offering discount loyalty programs to retail/clerical segments.
3. **A/B Testing:** Validate real-world uplift through controlled experiments, comparing campaign outcomes with and without model-driven targeting.
4. **Fairness Monitoring:** Evaluate subgroup metrics to avoid disparate impact across protected attributes. Adjust thresholds or apply post-processing if necessary.
5. **Model Lifecycle:** Establish drift monitoring and retraining protocols, as the original CPS data is outdated. Recalibrate thresholds and re-cluster periodically with newer data.

---

7. Risks and Limitations

- **Data Recency:** The dataset is from 1994–1995; population and income patterns may have shifted substantially. Results should be validated on contemporary data.
- **Label Definition:** The $50,000 threshold may not align with current purchasing power. Consider inflation-adjusted thresholds or continuous income prediction.
- **Fairness Concerns:** Socio-demographic predictors could correlate with protected attributes. Care must be taken to ensure ethical and legal compliance in marketing use.

---

8. Implementation and Next Steps

The project deliverables include reproducible code, model artifacts, and explanatory plots. The trained pipeline can be wrapped in an API for integration into customer relationship management (CRM) systems. Going forward, we recommend: - Publishing a model card to document intended use, metrics, and risks. - Establishing regular retraining cycles as new data becomes available. - Extending the segmentation analysis to alternative clustering methods such as Gaussian Mixtures or hierarchical clustering for further refinement.

---

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 provide powerful tools for marketing optimization. The classifier offers strong predictive performance and clear trade-offs between recall and precision, while the segmentation model identifies coherent personas with actionable recommendations. By combining both approaches, the client can design campaigns that are both cost-effective and precisely targeted, enhancing customer engagement and overall ROI.