Income Prediction & Customer Segmentation Analysis

Prepared for: Retail Business Client
Prepared by: Phani Tarun M
Date: February 2026

1. Executive Summary

This report presents two complementary machine learning solutions developed to support targeted marketing and strategic segmentation initiatives. First, a supervised classification model was built to identify individuals earning more than $50,000 annually. Second, an unsupervised segmentation model was developed to uncover structurally distinct customer groups.

The classification model enables high-efficiency marketing campaigns by prioritizing high-income prospects, while the segmentation model supports differentiated messaging and optimized budget allocation.

Metric	Value
Selected Model	XGBoost (Tuned)
ROC-AUC	0.95
Accuracy	0.9
Precision (> 50K)	0.37
Recall (> 50K)	0.85
F1 Score (> 50K)	0.52
Precision @ Recall ≥ 0.4	0.8

Final Classification Model Performance Summary

segment	Avg Age	Weeks Worked	Wage/hr	Capital Gains	Dividends	High-Income Rate (%)	Share of High-Income Population (%)	Segment Size
0	24.6	16.3	29.3	169	28	2.06%	2.51%	15,073
1	45.5	32.4	80.9	649	315	9.35%	88.75%	117,512
2	39.2	26.4	48.6	402	99	5.57%	8.72%	19,380
3	8.6	1.2	3.5	3	2	0.00%	0.02%	47,558

Final Segmentation Summary Table

2. Business Objectives

2.1 Income Classification Objective

The objective of the classification model is to distinguish between individuals earning less than $50,000 annually and those earning more than $50,000. Income level is used as a proxy for purchasing power and potential customer value.

Broad marketing campaigns often distribute resources uniformly, leading to inefficient spending. A predictive model enables targeted outreach toward individuals more likely to belong to the higher income segment.

Business objectives:

- Improve marketing efficiency through focused targeting
- Reduce campaign waste and acquisition cost
- Increase conversion likelihood for premium offerings
- Allocate marketing budget more effectively

The model is designed to balance coverage and precision by identifying a meaningful portion of high-income individuals while maintaining strong targeting accuracy.

2.2 Customer Segmentation Objective

Beyond income prediction, the segmentation model identifies distinct population groups based on demographic and economic characteristics. While classification tells us who to prioritize, segmentation tells us how different customer groups differ structurally.

The segmentation framework aims to:

- Identify meaningful and interpretable customer clusters
- Measure income concentration within each segment
- Highlight segments with higher strategic value
- Support tailored marketing strategies by segment

Business impact:

- More relevant campaign messaging
- Better alignment between products and customer profiles
- Improved strategic allocation of marketing resources

Therefore, classification allows accurate targeting, while segmentation allows a well-structured market strategy.

3. Data Understanding & Preprocessing

3.1 Dataset Description

The dataset consists of 40 demographic and employment-related variables along with survey weights reflecting stratified sampling. Survey weights were incorporated into supervised model training to ensure population-level validity.

3.2 Missing Value Handling

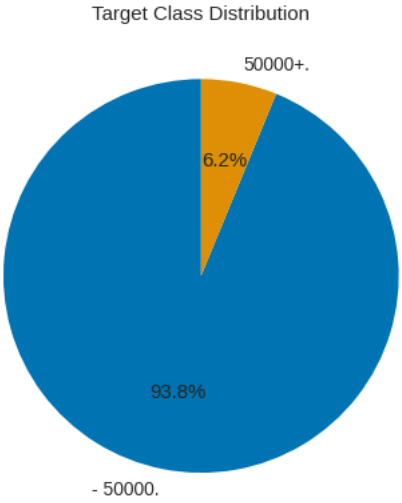
Missing values were handled using median imputation for numerical features and mode imputation for categorical features to preserve distributional properties.

3.3 Feature Engineering Strategy

Financial variables were log-transformed to address skewness. An income proxy feature was engineered to approximate total earning power, combining wage, capital gains, and dividends.

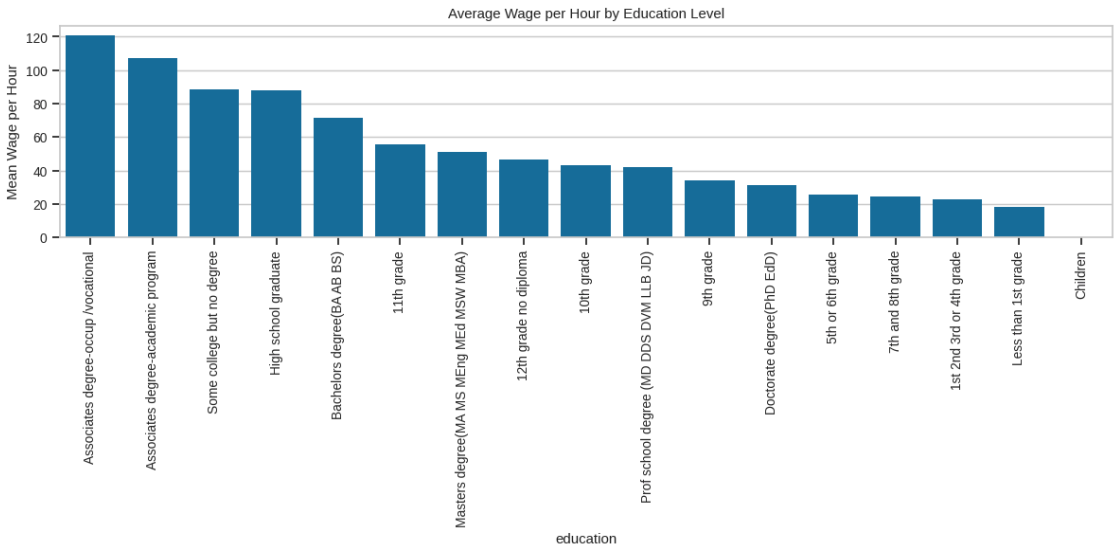
3.3 Key Exploratory Insights

Insight 1: Income is highly imbalanced



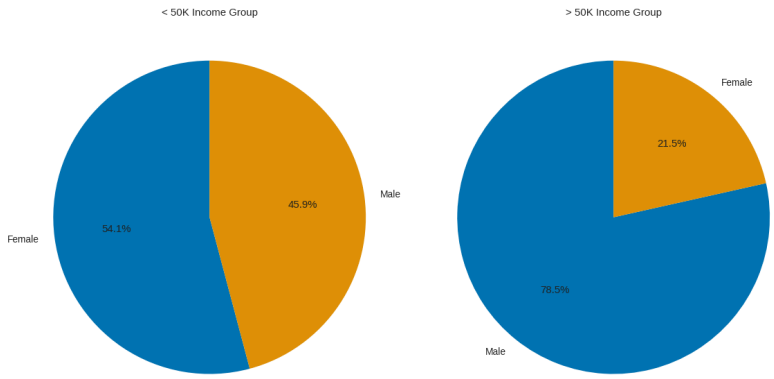
The dataset is highly imbalanced, with most individuals earning less than $50,000. This implies that overall accuracy alone would be misleading, and the model must prioritize precision and recall to effectively identify the smaller high-income segment for targeted marketing.

Insight 2: Income correlates strongly with education and wage



Average wage increases significantly with higher levels of education, with professional and advanced degree holders earning substantially more than individuals with lower educational attainment. This reinforces education as a strong economic signal and a meaningful predictor of income for targeted marketing strategies.

Insight 3: Gender Distribution Differs Significantly Across Income Groups



The >$50K income group is disproportionately male (approximately 79%), while the <$50K group has a higher female representation. This suggests that income-based targeting may naturally skew toward male customers, highlighting the importance of monitoring balance and inclusivity in campaign design.

Note: To see all the plots for more insights, look into EDA section in the jupyter notebook or [here]

4. Classification Modeling Approach

4.1 Model Benchmarking Strategy

Three candidate models were evaluated: Logistic Regression as an interpretable baseline, Gradient Boosting Machine as a nonlinear ensemble model, and XGBoost as a high-capacity boosting framework. The objective was not simply to achieve the highest overall accuracy, but to select a model aligned with marketing economics and campaign efficiency.

Model selection was guided by the following criteria:

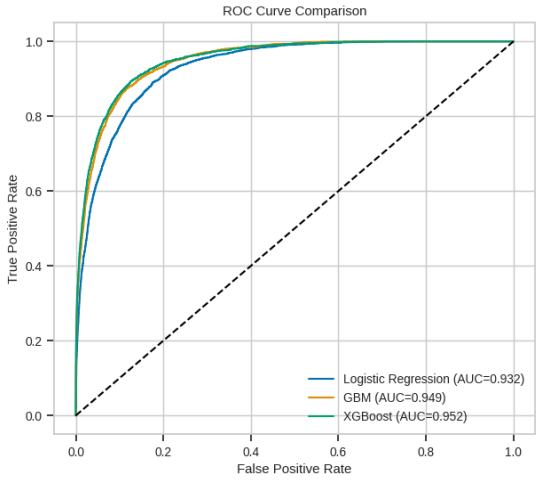
- Maximize precision while maintaining a minimum recall threshold of 40% for the high-income segment
- Ensure a meaningful portion of high-value individuals is identified
- Minimize unnecessary outreach costs driven by false positives
- Avoid relying on overall accuracy, which can be misleading in imbalanced datasets

By incorporating a recall constraint and precision optimization, the final model attains a balance with campaign reach and cost efficiency. This strategy of benchmarking enables the model to measure marketing return on investment instead of just relying on statistical outcomes for optimization.

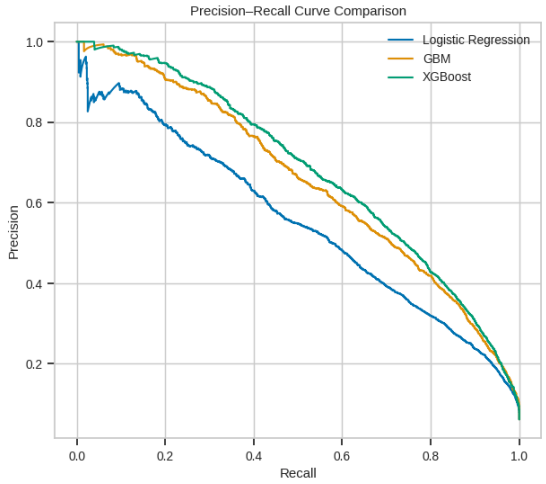
4.2 Model Evaluation Metrics

index	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
0	Logistic Regression	0.83	0.25	0.89	0.39	0.93
1	GBM	0.96	0.76	0.41	0.53	0.95
2	XGBoost	0.9	0.37	0.85	0.52	0.95

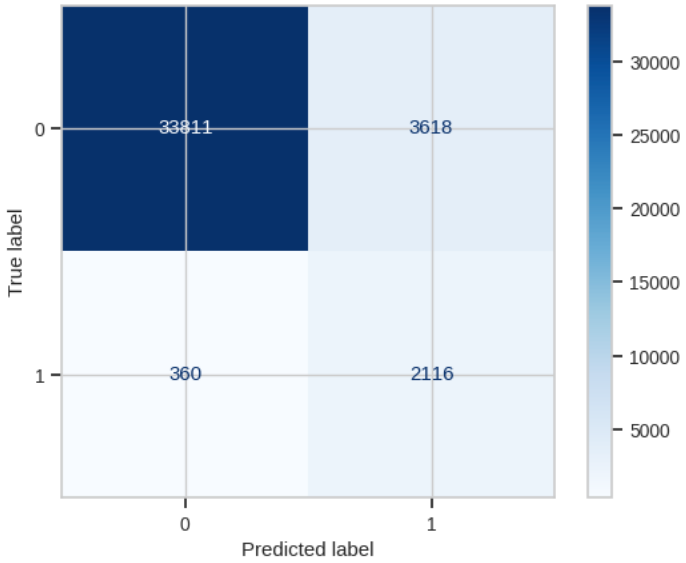
Model Performance Comparison Table



ROC Curve



Precision-Recall Curve



Confusion matrix of best model (XGBoost)

4.3 Final Model Selection Rationale

The final model was selected to maximize precision while ensuring that at least 40% of high-income individuals are captured. This ensures the campaign reaches a meaningful portion of valuable customers without excessively increasing outreach costs through false positives. By balancing reach and targeting accuracy, the model aligns directly with marketing ROI objectives rather than purely statistical performance. (i.e., Precision @ Recall ≥ 0.4)

5. Customer Segmentation Analysis

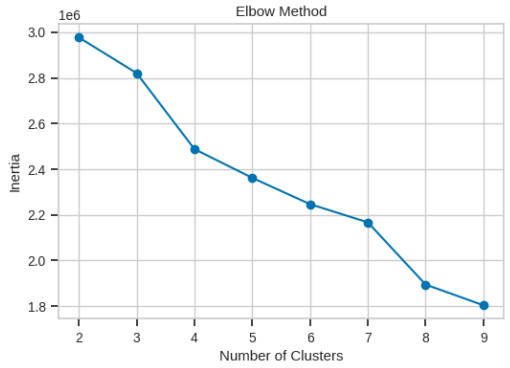
5.1 Feature Space Definition

Segmentation focused on economically meaningful features including demographics, employment characteristics, and income proxies.

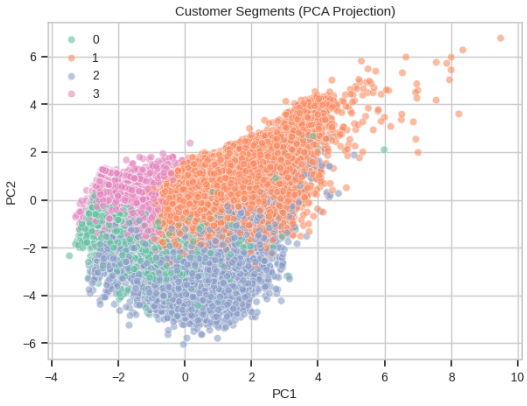
5.2 Clustering Methodology

KMeans clustering was applied after feature scaling. The number of clusters was determined using the Elbow method and silhouette analysis, with four clusters selected to balance statistical fit and business interpretability. While both four and eight clusters were plausible, a four-cluster solution was chosen for operational simplicity.

KMeans was selected due to its scalability and ability to generate clear, interpretable customer groups suitable for marketing action. The resulting segments show strong economic separation, as confirmed by lift analysis and income concentration metrics, indicating that the clustering structure is meaningful and business-relevant.



Elbow plot



PCA Cluster visualization

5.3 Segment Profiling & Interpretation

segment	Avg Age	Weeks Worked	Wage/hr	Capital Gains	Dividends	High-Income Rate (%)	Share of High-Income Population (%)	Segment Size
0	24.6	16.3	29.3	169	28	2.06%	2.51%	15,073
1	45.5	32.4	80.9	649	315	9.35%	88.75%	117,512
2	39.2	26.4	48.6	402	99	5.57%	8.72%	19,380
3	8.6	1.2	3.5	3	2	0.00%	0.02%	47,558

Segmentation Summary Table

The segments show clear economic and demographic separation, particularly across age, work intensity, and income composition. Segment 1 represents mature, full-time working individuals with the highest wages and investment income, making it the primary target for premium offerings. Segment 2 reflects mid-career earners suitable for growth-oriented products, while Segments 0 and 3 represent early-career or low-income groups better suited for entry-level or long-term engagement strategies. The structured differences across age and income indicators confirm that the segmentation provides actionable marketing guidance.

segment	High Income Rate	Lift vs Population
0	0.021	0.33
1	0.094	1.51
2	0.056	0.9
3	0.0	0.0

Lift Analysis

The lift values vary significantly across all four segments, ranging from 0.00 to 1.51, indicating clear economic differentiation. Segment 1 shows a strong overrepresentation of high-income individuals, Segment 2 is close to population average, while Segments 0 and 3 are underrepresented. This distribution confirms that the segmentation model effectively separated customers into economically distinct groups rather than forming uniform or random clusters.

segment	Segment Size	High Income Rate	Expected Buyers	Expected Revenue ($)
0	15073	0.021	16.0	7775.0
1	117512	0.094	549.0	274725.0
2	19380	0.056	54.0	27000.0
3	47558	0.0	0.0	50.0

Revenue Simulation Table

The revenue simulation estimates potential buyers by multiplying segment size, high-income rate within the segment, and an assumed marketing conversion rate of 5%. Expected revenue is then calculated using an average product value of $500 per converted customer. These assumptions provide a conservative, scenario-based estimate of potential financial impact across segments.

segment	Income proxy
0	3.6%
1	87.5 %
2	8.2 %
3	0.7 %

Income Concentration Table

Income concentration is heavily skewed toward Segment 1, which accounts for approximately 87.5% of total income proxy in the dataset. Segments 0, 2, and 3 contribute only marginal shares, indicating significant economic inequality across clusters. This confirms that Segment 1 represents the dominant economic value pool and should be prioritized for high-revenue marketing strategies.

6. Strategic Marketing Recommendations

The segmentation results reveal clear economic stratification across customer groups, particularly in income concentration, work intensity, and age distribution. Marketing strategy should therefore prioritize segments based on both revenue potential and customer life stage. Rather than applying a uniform campaign approach, differentiated allocation of budget and messaging can significantly improve return on marketing investment.

Recommended Marketing Focus by Segment:

Segment	Economic Profile	Strategic Priority	Recommended Approach
1	High wage, high investment income, mature working population	Highest Priority	Premium product targeting, wealth-oriented offerings, high-touch campaigns
2	Mid-career, moderate wage and investment activity	Growth Opportunity	Cross-sell, upsell, career-growth aligned financial products
0	Younger, lower work intensity, lower income concentration	Long-Term Development	Entry-level products, brand-building campaigns, loyalty programs
3	Very low income and minimal economic activity	Low Immediate Priority	Low-cost digital outreach, minimal premium allocation

7. Limitations & Future Enhancements

The analysis is based on historical census data and does not incorporate temporal dynamics or behavioral data. Future work may include fairness auditing, model calibration analysis, and integration with transactional data. The current implementation is developed within a Jupyter Notebook environment for analytical exploration. For production deployment, the codebase should be modularized into Python scripts, with trained models serialized and versioned to enable scalable inference and integration into downstream systems. Incorporate interpretability visualizations for the final XGBoost model to highlight which features contribute most to the prediction outcomes. Presenting feature importance and contribution insights strengthens transparency, supports business decision-making, and ensures that model-driven actions are grounded in clear economic rationale rather than treated as a black-box outcome.

8. References

[1] Logistic regression model

[2] GBM model

[3] XGBoost model

[4] Machine learning with xgboost