Census Income Analysis – Executive Summary

The Business Problem

A retail business aims to identify high-value customers those earning $50K or more annually to improve the efficiency of targeted marketing campaigns. Although the dataset contains over 40 demographic and employment variables, the real challenge is not prediction alone. The key questions are:

- What truly differentiates high-value customers from the rest of the market?
- How should customers be segmented to support real marketing decisions rather than superficial demographic grouping?

Answering these questions correctly directly impacts marketing ROI, market sizing accuracy, and campaign efficiency.

Framing the Insight

Early analysis uncovered a critical issue: the raw dataset significantly underrepresents high-value customers. Sample weights show that customers earning $50K+ have weights 2.3× higher than lower-income customers (mean: 2.16 vs 0.94, $p < 0.0001$).

Ignoring these weights would underestimate both the size and behavioral patterns of the target customer base by more than 100%. This discovery reshaped the analytical approach and ensured all conclusions reflect the true population.

What Actually Drives Income

Education alone does not reliably produce high-value customers. Twenty-eight percent of college-educated individuals still earn below $50K. Outcomes vary sharply depending on how education combines with occupation and work consistency.

For example, customers with a Bachelor's degree in professional or managerial roles show a ~72% likelihood of being high-value, while the same degree in service roles drops that likelihood to ~24%. Income emerges from the interaction of education, occupation, career stage, and work intensity, not from any single factor.

Career Stage and Income Dynamics

Across all education levels, income likelihood follows a consistent life-cycle pattern. Rates peak between ages 35–54 and decline after age 55. Even among Master's degree holders, the high-income rate drops from 42.4% to 31.6% post-55.

This contradicts the assumption that income increases continuously with age and indicates that mid-career customers represent the strongest marketing opportunity.

Key Model Insight

A Random Forest classifier was built to validate and quantify these patterns using sample-weighted data. The model achieved strong ranking performance with a ROC-AUC of 0.919 and recall of 86.2%.

Most importantly, weeks worked per year emerged as the single strongest predictor of income, accounting for 31.6% of feature importance—surpassing education and occupation individually. Consistent full-time employment is more decisive than credentials alone.

Classifier Limitations

Despite strong discrimination ability, precision is limited at 23.7%. This means that 3 out of every 4 customers flagged by the model are false positives.

While this represents a 3.8× improvement over random targeting (6.2% baseline), the model is not suitable for direct deployment without additional filtering.

Outcome-Driven Customer Segmentation

Six customer segments were defined based on income-producing mechanisms rather than demographic similarity:

Segment	Market Size	Profile	High-Income Rate	Business Priority
Credentialed Professionals	1.9%	Bachelor's+, Professional/Managerial, Age 35-55	~65-75% (est.)	TIER 1 - Highest conversion
Early Career Builders	13.0%	Bachelor's+, Professional roles, Age 25-35	~35-45% (est.)	TIER 1 - Largest viable
Educated Underemployed	13.1%	Bachelor's+ in non-professional occupations	~15-25% (est.)	TIER 2 - Upskilling offers
Service/Labor Workers	2.3%	Mixed education, service/labor occupations	~5-10% (est.)	TIER 3 - Low priority

Skilled Tradespeople	0.0% (73 people)	Low education, skilled trades, self-employed	~40-50% (est.)	Not viable - too small
Transitional/Other	69.6%	No clear income-producing pattern	~4-6% (baseline)	TIER 3 - Broad campaigns only

These segments reflect how income is produced, not surface-level demographics.

Strategic Recommendation

Primary Target: Early Career Builders

This segment offers the strongest balance of scale and income:
- 13% of the market with substantial weighted reach
- 5–7× higher likelihood of being high-value than baseline
- Large enough to drive meaningful revenue
- Well-suited for career-advancement messaging

Recommended channels include LinkedIn and mobile-first platforms.

Secondary Target

Credentialed Professionals represent a smaller but extremely high-conversion segment (~70%+). This group is best suited for premium or high-margin offerings.

Deployment Guidance

A staged deployment approach is recommended:
1. Use the classifier for initial filtering (probability >40%)
2. Apply segmentation to reduce false positives
3. Pilot with ~5,000 customers
4. Measure real conversion rates
5. Increase thresholds (60–70%) for higher-cost channels

Expected lift is 280–380% versus untargeted campaigns, with potential to reach 500–600% pending pilot validation.

Conclusion

High-value customers are defined by alignment and consistency across education, occupation, and work intensity not credentials alone. Understanding how income is produced enables

precise, defensible marketing decisions rather than broad demographic guessing.

References

1. U.S. Census Bureau (1994-1995). Current Population Survey Public Use Microdata Sample. https://www.census.gov/programs-surveys/cps.html
2. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
4. Scikit-learn Documentation. RandomForestClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.
6. Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8(1), 25.
7. Wedel, M., & Kamakura, W. A. (2000). Market Segmentation: Conceptual and Methodological Foundations (2nd ed.). Springer.
8. Anthropic Claude (2024). AI assistant for Python syntax reference and debugging support. https://claude.ai (Note: Used for technical coding assistance only; all analysis, insights, and strategic decisions are original work.)