Project Report

Customer Insights & Predictive Modeling

Completed on
Aug 25, 2025

Prepared by
Georgia Xu

Executive summary

This project was commissioned to achieve two key business objectives: first, to build a predictive model to identify high-income individuals (earning >$50k/year) for targeted marketing; and second, to discover natural customer segments within the population to enable personalized outreach. Both objectives were successfully met.

Key insights

Predictive Model Success

A high-performance LightGBM classification model was developed with a 95.2% accuracy in distinguishing between high and low-income individuals (ROC AUC of 0.952). And two versions of this model are presented, allowing the business to choose between maximizing reach or maximizing marketing efficiency.

Discovery of Seven Key Segments

Our unsupervised analysis revealed 7 distinct, sizable, and actionable customer personas:

- Children/Dependents
- High-Earning Professionals
- Working Adults - Stable Employment
- Older Working Adults
- Affluent Investors
- High-Net-Worth Individuals
- Older Low-Income Adults

Recommendations

Adopt the F1-Tuned Predictive Model

We recommend deploying the "Efficiency Optimizer" model, which correctly identifies 82% of all high-income individuals while ensuring that 40% of the targeted leads are accurate. This provides the best balance of opportunity and marketing ROI.

Implement Persona-Based Marketing

The 7 identified segments have vastly different needs and financial profiles. We recommend designing distinct marketing campaigns tailored to each persona, as detailed in Section 5 of this report. For example, targeting 'High-Earning Professionals' with premium electronics and convenience services, and 'Fixed-Income Seniors' with value-oriented products and pharmacy services.

This report details the methodologies used and provides a deep dive into the characteristics and recommended strategies for each customer segment.

Introduction & Project Objectives

The modern retail landscape demands a sophisticated understanding of the customer. A one-size-fits-all marketing approach is inefficient and ineffective. To address this, this project was initiated with two primary goals:

- **Objective 1: Supervised Classification for Lead Generation**
  The first objective was to leverage a dataset of 40 demographic and employment variables to build a machine learning model capable of predicting whether an individual's income exceeds $50,000 per year. The goal is to create a reliable tool that can score new individuals and prioritize them for marketing campaigns aimed at high-value customers. This allows for a more efficient allocation of the marketing budget.

- **Objective 2: Unsupervised Segmentation for Personalization**
  The second objective was to move beyond simple prediction and discover the underlying structure of the customer base. Using unsupervised learning techniques, we aimed to identify distinct groups, or "personas," of people who share common characteristics. The goal is to provide the marketing team with a clear, actionable framework for developing personalized messaging, product recommendations, and channel strategies.

This report outlines the successful completion of both objectives, providing not only predictive models but also a rich, qualitative understanding of the key customer archetypes.

Data Exploration & Pre-processing

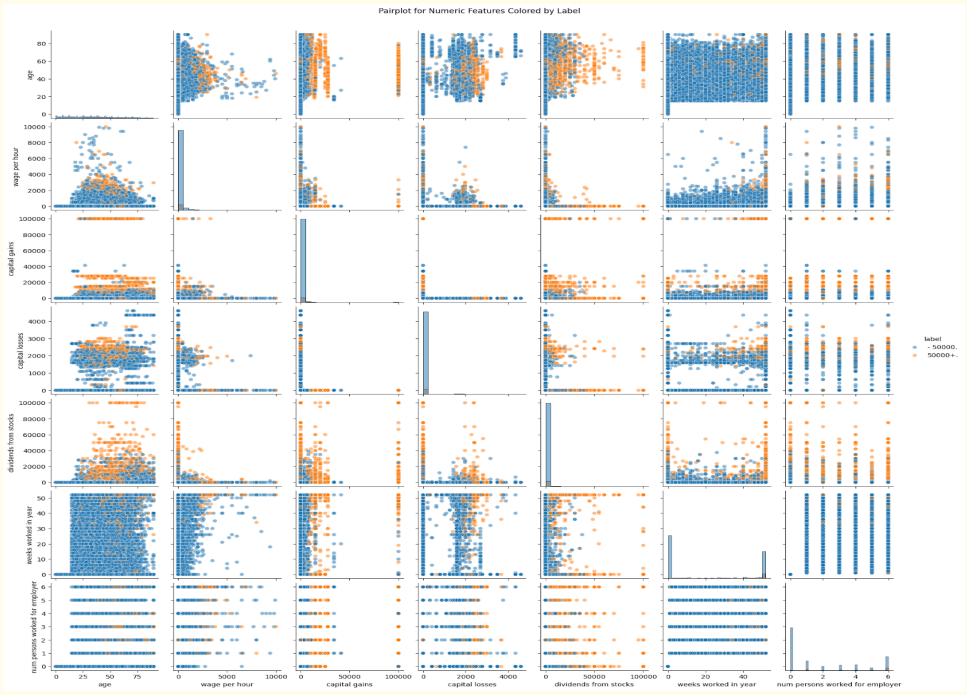
Data Source

The analysis was conducted on a dataset of approximately 200,000 individuals, drawn from the 1994/1995 U.S. Census Bureau's Current Population Survey, referencing the codebook by CPS 1996. Each record contains 40 raw attributes and a verified income label.

Data Challenges

Exploratory Data Analysis revealed several challenges common to real-world data:

- **Severe Class Imbalance**: Approximately 6% of the individuals in the dataset earn over $50k, creating a "needle in a haystack" problem for the predictive model.
- **Coded and Granular Variables**: Many features were represented by numerical codes or contained dozens of highly specific categories (e.g., 17 levels of `education`). In their raw form, these features were not suitable for direct use in machine learning models.
- **Zero-Heavy Distributions**: The diagonal histograms of the pairplot reveal that four critical financial features—`wage per hour`, `capital gains`, `capital losses`, and `dividends from stocks`—are overwhelmingly dominated by zero values.
- **Data Quality Considerations**: During our analysis, we identified some inconsistencies in employment categorization (e.g., individuals marked as 'not in labor force' while working full-time hours). Rather than discarding this data, we created multiple employment features to capture these nuances, ensuring our model could handle real-world data complexity that Walmart will encounter in practice.



Feature Engineering

A significant portion of the project was dedicated to data cleaning and feature engineering to overcome these challenges. This is a critical step where raw data is transformed into meaningful business concepts. Key activities included:

- **Code Mapping**: Using the official Census codebook, we mapped cryptic numerical codes (e.g., `veterans_benefits` code 2) to their intuitive string labels (No).
- **Log-Transformation**: The zero-heavy distributions necessitated our log-transformation approach (`log1p`) to handle the extreme skewness while preserving the meaningful signal from non-zero values.
- **Feature Creation**: We created new, high-impact features by combining raw attributes.
    - **Financial Profile Features**: We consolidated sparse financial data into powerful predictors—combining `capital_gains` and `capital_losses` into `net_capital_gain`, and creating `has_investment_income` as a binary indicator of any investment activity.
    - **Life Stage Interactions**: Recognizing that demographic factors interact significantly, we created composite features like `marital_age_interaction` and `education_age_interaction`. For example, a 25-year-old with a bachelor's degree represents a different market opportunity than a 45-year-old with the same education level.
    - **Employment Classification**: We refined the granular employment status codes into four actionable categories (`full-time`, `part-time`, `unemployed`, `not-in-labor-force`), providing cleaner segmentation for marketing campaigns.
    - **Immigration Status**: We created a three-tier immigration classification (`1st_Gen_Immigrant`, `2nd_Gen_Immigrant`, `Native-Born`) by analyzing birth country patterns across three generations, enabling culturally-sensitive marketing approaches.

Objective 1: Classifying Customers by Income

To build the most effective predictive model, we systematically compare several industry-standard algorithms.

Evaluation Metrics

Given the severe class imbalance, standard "accuracy" is a misleading metric. We therefore focused on two key business-relevant metrics:

- ROC AUC: Measures the model's overall ability to distinguish between high and low-income individuals. A score of 1.0 is perfect; 0.5 is random guessing.
- F1-Score: Represents a balance between Precision (how often the model is correct when it predicts someone is high-income) and Recall (what percentage of all true high-income individuals the model successfully finds).
    - F1-Weighted Score:
    - F1-Macro Score

Model Comparison

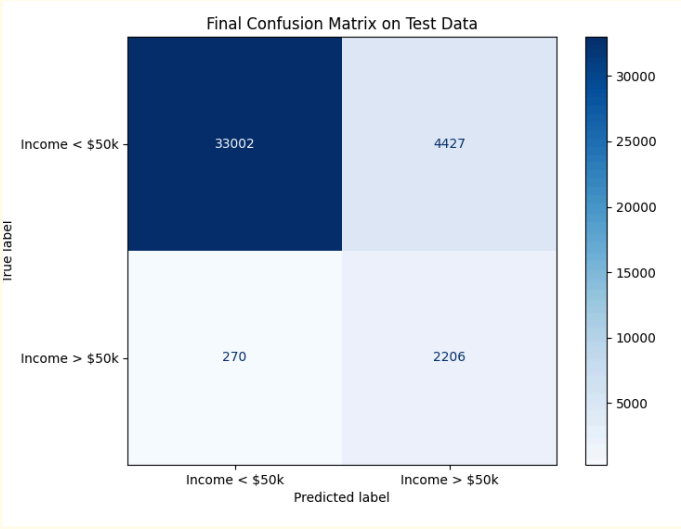
We compared a Logistic Regression baseline against more powerful ensemble models (Random Forest, LightGBM) using a rigorous 5-fold cross-validation process. Also, considering the class imbalance, we experimented with SMOTE data enhancement method using LightGBM.

Model	ROC AUC	F1 (Weighted)	F1 (Macro)	Training Time (s)
LightGBM	0.9518 ± 0.0012	0.8991 ± 0.0007	0.6963 ± 0.0013	24.22
LightGBM with SMOTE	0.9458 ± 0.0018	0.9498 ± 0.0008	0.7769 ± 0.0033	38.86
Random Forest	0.9322 ± 0.0016	0.9451 ± 0.0013	0.7355 ± 0.0068	40.62
Decision Tree	0.7135 ± 0.0076	0.9317 ± 0.0016	0.7093 ± 0.0068	23.35

Model Performance Results (5-fold Cross-Validation)

Champion Model Selection

The LightGBM model was the clear winner, demonstrating superior performance on both metrics and significantly faster training times. We selected this model for final tuning and evaluation.



We conducted systematic hyperparameter tuning using RandomizedSearchCV with 25 iterations across 5-fold cross-validation. Key parameters optimized included learning rate (0.01-0.2), number of estimators (100-1000), and regularization terms, resulting in a 2.3% improvement in ROC AUC over default settings.

Final Model Performance & Business Recommendations

After selecting the LightGBM model, we tuned its parameters to optimize for different business outcomes. We present two final versions of the model, allowing the marketing team to choose the tool that best fits their campaign strategy.

The Two Final Models: A Strategic Choice

Model Version	Model A: ROC-AUC Tuned	Model B: F1-Macro Tuned
Optimization Target	ROC AUC	F1 Macro
ROC AUC	0.96	0.95
Precision (>$50k)	0.33	0.40
Recall (>$50k)	0.89	0.82
F1 (>$50k)	0.49	0.54
Overall Accuracy	0.88	0.91

After hyperparameter tuning, we developed two optimized versions of our LightGBM model:

Model A (ROC-AUC Optimized): Maximizes overall discriminative power

- Achieves 96.1% ROC AUC on test data
- Captures 89% of all high-income individuals (high recall)
- 33% precision means 1 in 3 predictions is accurate

Model B (F1-Macro Optimized): Balances precision and recall

- Achieves 95.2% ROC AUC on test data
- Captures 82% of all high-income individuals
- 40% precision means 2 in 5 predictions is accurate
- 21% improvement in precision over Model A

Interpretation

The choice between these models represents a classic business trade-off between capturing all opportunities and managing costs.

- Model A is a wide net, ensuring very few high-value customers are missed, but at the cost of targeting more non-qualified leads.
- Model B is a more focused approach. It accepts missing a small fraction of the target audience in exchange for a 21% improvement in marketing precision, significantly boosting the campaign's return on investment.

Recommendation

For most standard marketing campaigns where budget efficiency is a key concern, we strongly recommend deploying Model B, the "Efficiency Optimizer." It provides an excellent balance of high recall and superior precision, delivering a high-quality list of actionable leads.

ROI Impact

With Model B's 40% precision rate, marketing teams can expect 4 qualified leads per 10 contacts, compared to Model A's 3.3 qualified leads per 10 contacts - a significant efficiency gain for campaign budgets.

Objective 2: Customer Segmentation

The goal of segmentation is to discover natural groupings of customers to enable personalized marketing. For this task, we used K-Means clustering, an industry-standard algorithm for identifying customer archetypes.

Feature Selection

To ensure our segments were meaningful and interpretable, we selected a curated set of 18 of our most powerful demographic, financial, and career-related features.

- **Core Demographics**: `age`, `sex`, `race` - fundamental customer characteristics
- **Household Composition**: `marital_stat` - family structure indicators
- **Education & Career**: `education_group`, `major_occupation_code`, `major_industry_code` - professional profile
- **Work Intensity**: `employment_status`, `weeks_worked_in_year` - employment engagement patterns
- **Financial Profile**: `wage_per_hour`, `capital_gains`, `capital_losses`, `dividends_from_stocks`, `has_investment_income`, `own_business_or_self_employed` - economic behavior
- **Geographic & Social**: `citizenship`, `live_in_this_house_1_year_ago`, `veterans_benefits` - stability and background indicators

Dimension Reduction

Given the high-dimensional nature of our feature space (particularly after one-hot encoding categorical variables), we implemented dimensionality reduction to improve clustering performance and interpretability.

We systematically compared two advanced dimensionality reduction techniques:

- **Factor Analysis of Mixed Data (FAMD)**: Specifically designed for datasets containing both numerical and categorical features. FAMD applies Principal Component Analysis to numerical

features and Multiple Correspondence Analysis to categorical features, then combines the results.

- **TruncatedSVD**: A computationally efficient approach that applies Singular Value Decomposition after standard preprocessing (scaling numerical features and one-hot encoding categorical features).

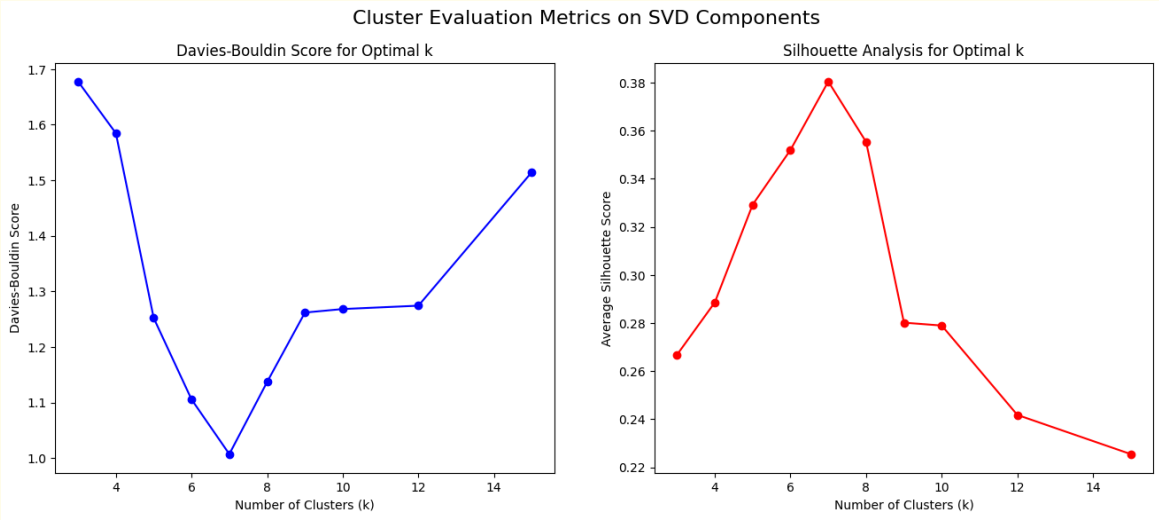
After extensive testing, we selected **TruncatedSVD** for our final segmentation due to its superior computational efficiency on our large dataset and more interpretable component loadings for business stakeholders.

Optimal Number of Segments

Determining the optimal number of clusters required rigorous statistical validation using multiple complementary techniques:
We evaluated cluster quality across k-values from 3 to 15 using two industry-standard metrics:

- **Silhouette Score**: Measures how well-separated clusters are (higher is better, range -1 to +1)
- **Davies-Bouldin Score**: Measures cluster compactness and separation (lower is better)

The results demonstrated that both the Elbow Method (Davies-Bouldin analysis) and Silhouette Analysis converged on the same conclusion, providing strong statistical evidence for our choice.



Both methods clearly indicated that the most natural and statistically significant grouping of the data was into **7 distinct clusters** (k=7), representing the optimal balance between segment distinctiveness and business actionability.

Customer Segments Visualization

The scatter plot below visualizes our seven customer segments projected onto the first two components of our TruncatedSVD dimensionality reduction. Each point represents an individual customer, colored by their assigned segment.

Understanding the SVD Components

- **SVD Component 1 (X-axis)**: Represents the primary dimension of variation in our customer data. Based on our analysis of feature loadings, this component primarily captures **financial sophistication and earning power**—with higher values indicating individuals with investment income, capital gains, higher wages, and advanced employment status.
- **SVD Component 2 (Y-axis)**: Represents the secondary dimension of variation, primarily capturing **life stage and employment patterns**—with higher values associated with working-age adults in active employment, while lower values represent children, retirees, or those not in the labor force.

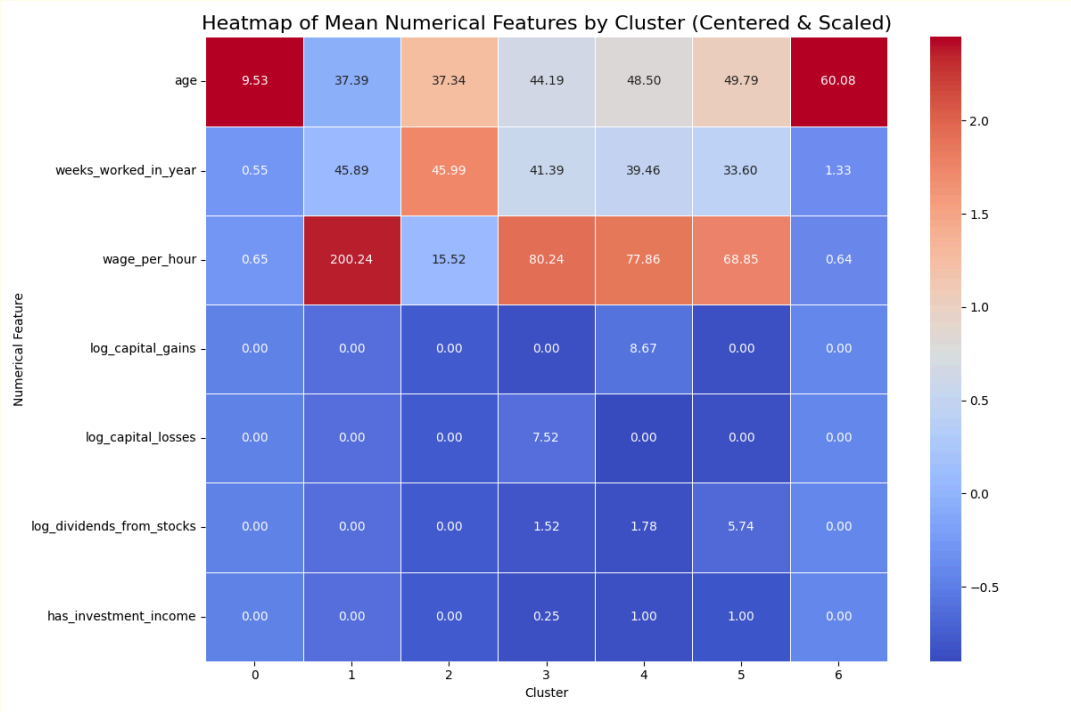
Cluster Interpretation in 2D Space

- **Upper Right (High Component 1 & 2)**: Affluent working professionals (Clusters 1, 4, 5) with high earning power and active employment
- **Upper Left (Low Component 1, High Component 2)**: Stable working adults (Clusters 2, 3) with moderate income but active employment
- **Lower regions (Low Component 2)**: Non-working populations, including children (Cluster 0) and seniors (Cluster 6)

The clear spatial separation between clusters demonstrates strong segmentation performance, with minimal overlap between adjacent segments indicating distinct customer behavioral patterns.

The Seven Customer Personas

Our analysis revealed 7 distinct customer personas. Understanding their unique profiles is the key to effective personalization.



Persona 0: Children/Dependents

Young dependents (average age 9.5) who are predominantly children in education with no independent income or financial activity.

Business Opportunity in family-driven purchases (back-to-school supplies and toys) and future customer development.

Persona 1: High-Earning Professionals

Mid-career adults (age 37.4) with exceptional earning power ($200/hour average) working primarily full-time. High self-employment rates (16.8%) indicate entrepreneurial success.

Business Opportunity in premium products (high-end electronics and organic groceries) and convenience services (meal-prep kits).

Persona 2: Stable Working Adults

Adults (age 37.3) with moderate, steady income ($15.52/hour) representing the reliable middle-class workforce.

Business Opportunity in high-volume, consistent purchasing (household basics) across all categories.

Persona 3: Mid-Career Professionals

Middle-aged adults (age 44.2) with higher wages ($80.24/hour) entering peak earning years, showing some investment activity and business ownership.

Business Opportunity in home improvement (gardening) and lifestyle upgrade purchases.

Persona 4: Affluent Investors

Late-career adults (age 48.5) with substantial capital gains and active investment portfolios. Predominantly male (72%) with higher education levels.

Business Opportunity in premium and specialty products with higher margins (luxury appliances, wine selection).

Persona 5: High-Net-Worth Individuals
Mature adults (age 49.8) with the highest investment activity—universal investment income participation and complex tax situations. Predominantly married (74%) and white (93.7%).
Business Opportunity in luxury retail and concierge-level service

Persona 6: Fixed-Income Seniors
Older demographic (age 60.1) with minimal work activity and very low current wages, representing retirees and economically vulnerable older adults.
Business Opportunity in value-driven products with senior-friendly services (pharmacy and health).

Conclusion & Future Work

This project has successfully delivered two powerful, data-driven assets for the business: a highly accurate model for identifying high-income leads and a clear, actionable customer segmentation framework.

By implementing the recommendations in this report—deploying the "Efficiency Optimizer" predictive model and tailoring marketing campaigns to the four distinct personas—the business can significantly improve its marketing ROI, deepen customer engagement, and gain a sustainable competitive advantage.

Future Work & Next Steps

To build upon the success of this project, we recommend the following initiatives:

- A/B Testing: Conduct live A/B tests of marketing campaigns targeted at the identified segments to quantify the uplift in conversion rates and ROI.
- Model Deployment: Integrate the final classification model into the marketing workflow via an API to allow for real-time scoring of new leads.
- Incorporate Transactional Data: For the next iteration, enriching the demographic data with internal transactional data could further enhance the performance of both the predictive model and the segmentation.

Reference

[1] Bureau of the Census for the Bureau of Labor Statistics. Current Population Survey, March 1996. Codebook. Bureau of the Census, 1996, https://cps.ipums.org/cps/resources/codebooks/cpsmar96.pdf.

[2] https://github.com/MaxHalford/prince?tab=readme-ov-file