PROJECT REPORT

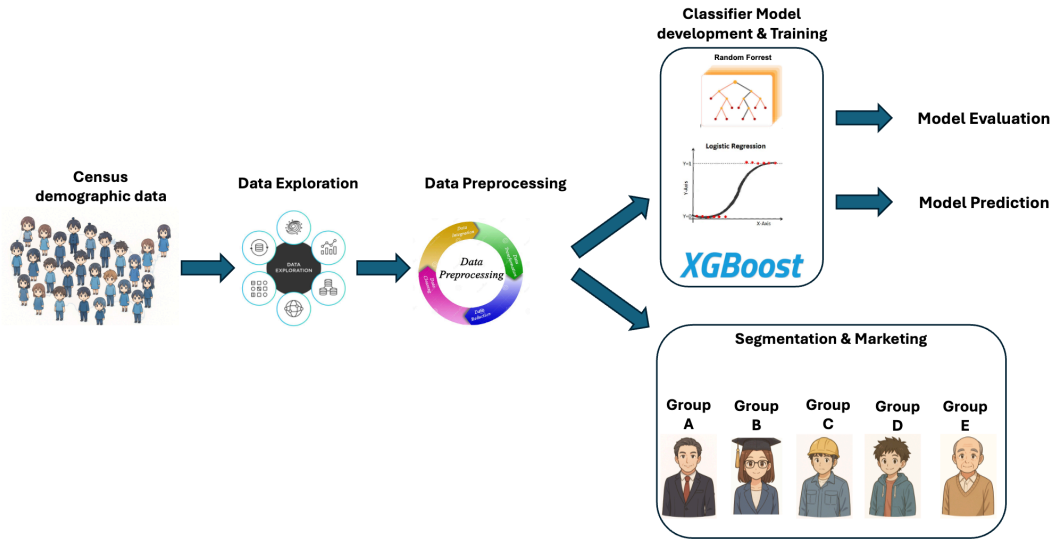
Title: Beyond the Paycheck: Uncovering People, Patterns, and Possibilities in Census Data

Chuanye (Charlie) Xiong

December 1, 2025

Summary: In this project, the author developed 2 complementary machine learning solutions to support the retail client's marketing strategy: (1) an income classification model to identify high-income and low-income individuals, and (2) a customer segmentation model to uncover distinct population groups for targeted marketing. As for the classifier, three ML models were adopted: (1) Logistic regression; (2) Random Forrest; (3) XGBoost. After full data exploration, preprocessing, and weighted performance evaluation, the Random Forrest model achieved the best overall results, getting **99.09%** weighted accuracy in predicting whether and individual earns more or less than $50,000 per year. As for segmentation for marketing, The author applied Principal Component Analysis (PCA) followed by K-Means clustering, identifying 13 meaningful clusters within the population. **One cluster exhibited a particularly high concentration of high-income individuals, with a weighted 78% earning more than $50,000**. Through further analysis, the author categorized the population into five major demographic groups: (1) high-income group; (2) Middle-class earners; (3) General Working population; (4) Young individuals; (5) Senior or retired individuals.

To gain increasingly deeper insights into the problem, the author structured our workflow around a four-stage process, shown as the Scheme 1 below: (1) Initial data exploration and data preprocessing (section 2); (2) Development and evaluation of income group classifiers (section 3); (4) Segmentation modeling and marketing recommendations (section 4). To close the report, Section 5 offers insights, reflections, and suggestions for future development.



Scheme 1. Architecture and data pipeline for this project

1.PROJECT INTRODUCTION

Nowadays, marketing driven by data insights is essential for maximizing customer engagement and business impact. In this project, the author partnered with a retail client to develop two complementary machine learning solutions aimed at identifying and understanding key consumer segments within the U.S. demographic data. Using weighted census data from the year 1994 and 1995, which includes 40 demographic and employment-related features, our objectives were twofold: (1) to build a classifier that predicts whether an individual earns more or less than $50,000 annually, and (2) to create a segmentation model that reveals distinct population groups for tailored marketing strategies.

I approached the task through a structured, four-stage process: initial data exploration, data preprocessing, classification modeling, and segmentation analysis. For classification, I tested Logistic Regression[1], Random Forest[2], and XGBoost[3] models, ultimately finding that Random Forest yielded the highest performance with 99.09% weighted accuracy. For segmentation, the author applied PCA and K-Means clustering to uncover 13 meaningful clusters, five of which were further profiled for strategic targeting.

2. DATA EXPLORATION & PREPROCESSING

The data exploration stage aims to understand the structure, quality, and distribution of the features provided in the census demographic dataset. This step is critical for designing an appropriate preprocessing pipeline and ensuring that downstream machine learning models receive clean and meaningful inputs. I began by examining the data types of all 42 columns. As shown in Figure 1(a), the dataset is composed primarily of object-type categorical variables (69%), representing fields such as education, class of worker, marital status, and citizenship. A smaller portion of the features are numeric—int64 (28.6%) and float64 (2.4%)—corresponding to values such as age, capital gains, and weeks worked in year.

To prepare the dataset for modeling, categorical variables were converted to ordered categorical encodings. This choice was made deliberately: although one-hot encoding is common, it dramatically increases dimensionality for features with many unique values, reducing interpretability and complicating feature selection. The ordinal approach provides a more compact representation while maintaining model performance. Next, I handled missing values. As for the category features, missing entries were imputed with an “Unknown” category. While in numeric features: missing values were imputed using the median of each feature to avoid distortion from skewed distributions.

A weighted analysis was then conducted using the survey-provided population weight variable. Because each record represents a proportion of the broader U.S. population, unweighted plots would distort the underlying demographic reality. Using weighted counts, Figure 1(b) illustrates the population distribution across income categories. Approximately 93.59% of individuals earn less than $50,000, while only 6.41% earn more, confirming that high-income earners are a minority within the dataset. This imbalance highlights the importance of using class weighting and sample weights during classifier training to avoid bias toward the majority class. I further examined the weighted distribution of age,

shown in Figure 1(c). The population exhibits two notable peaks: (i) Around age ~15, reflecting a large youth population; (ii) Around age ~35, representing the core working-age group.

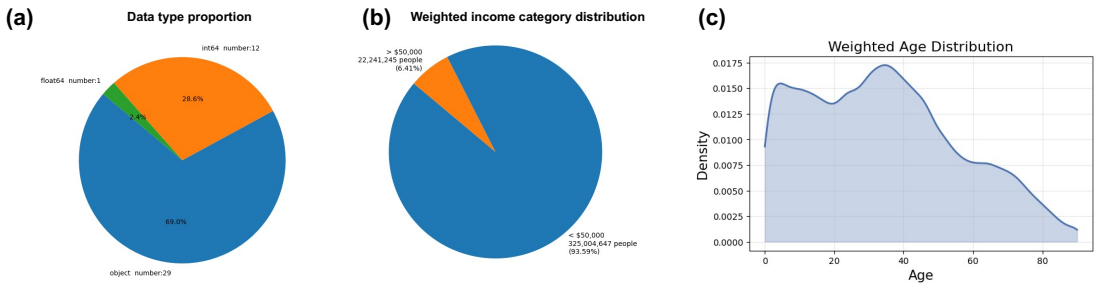


Figure 1. (a) Data type proportion in census demographic dataset; (b) Weighted income category distribution (< $50,000 and >$50,000); (c) Weighted Age Distribution

I also summarize the weighted distributions of education level and class of worker in the census dataset. As shown in Figure 2(a), the largest educational group is High School Graduate (26.45%), followed by Children (22.78%), reflecting the natural demographic structure of the population. Only 10.29% of individuals hold a bachelor's degree, while the remaining categories—such as vocational programs, associate degrees, or graduate degrees—appear in smaller proportions. Figure 2(b) shows that nearly 49.21% of the population is labeled "Not in Universe," meaning they are not part of the labor force (e.g., minors, retirees, homemakers). Among employed individuals, the private sector is the dominant category, representing 37.20% of the population, with smaller proportions working in government or self-employment roles. These distributions provide important context for understanding later income classification and segmentation patterns.

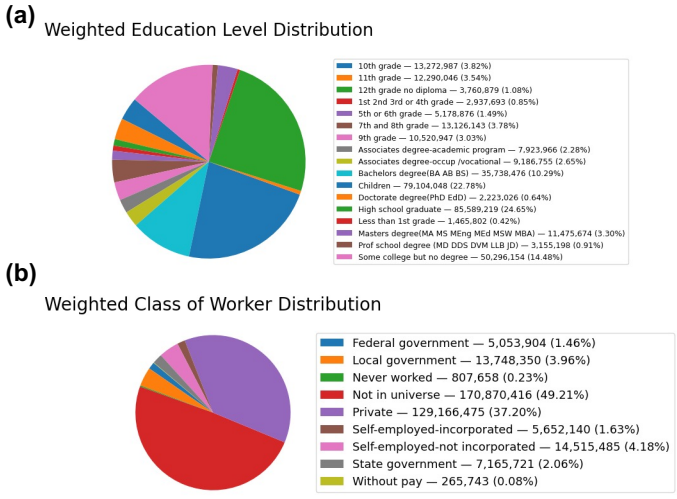


Figure 2. (a) Weighted Education level distribution; (b) Weighted Class of Worker Distribution

3. DEVELOPMENT AND EVALUATION OF INCOME GROUP CLASSIFIERS

3.1 Classifier Method

Once the data were preprocessed and encoded using an ordinal encoder, I applied feature scaling to stabilize model training and performed a standard train–test split. To evaluate income group prediction performance, I selected three complementary machine-learning models: (1) Logistic Regression, (2) Random Forest, and (3) XGBoost. The author intentionally chose models from three fundamentally different families—linear, bagging-based, and boosting-based—so that their strengths complement each other and provide a robust benchmark comparison. Specifically, logistic Regression serves as a simple and interpretable baseline, helping us understand the linear separability of the income classes. Moreover, Random Forest represents a high-variance-reducing ensemble model that handles non-linear relationships and interactions well. Furthermore, XGBoost, a gradient-boosting algorithm, is known for achieving state-of-the-art performance on structured/tabular data due to its ability to sequentially correct the previous decision-tree errors.

3.2 Model Performance

To evaluate the performance of the three classifiers, I report 4 important metrics: (a) Weighted Accuracy, (b) Precision, (c) Recall, and (d) F1-score, summarized in Table 1. Because the dataset is highly imbalanced (only ~6% of individuals earn >$50,000), accuracy alone is insufficient; therefore, I emphasize metrics that better capture minority-class performance.

Weighted accuracy accounts for the census survey sample weights as is shown in equation (1) and let TP: true positives, TN: True negatives, FP: False Positives, FN: False Negatives.

$$Weighted\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision measures how many predicted high-income cases are correct—a key metric when over-targeting high-income consumers is costly, as shown in equation below:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall captures the model's ability to identify true high-income individuals, which is crucial because the high-income group is very small. The equation to calculate recall has shown below:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-Score balances precision and recall, offering a single measure of minority-class performance.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

As it's shown in the Table 1 below, the Random forest is the best among all the three models, it not only has the best accuracy, considering that the highly imbalanced high income >$50,000 group, the precision and recall are very crucial to the model. As is shown in Table 1, the Random forest has prevision as high

as 0.97 and recall 0.88, even better than the XGBoost. The superior performance of Random Forest stems from its bagging ensemble strategy, which reduces variance, mitigates overfitting, and captures nonlinear interactions between demographic features.

Table 1. Model performance of the 3 Classifier models

Model	Weighted Accuracy	Precision	Recall	F1
Logistic Regression	0.947	0.706	0.274	0.395
Random Forrest	0.991	0.973	0.882	0.925
XGBoost	0.966	0.832	0.541	0.656

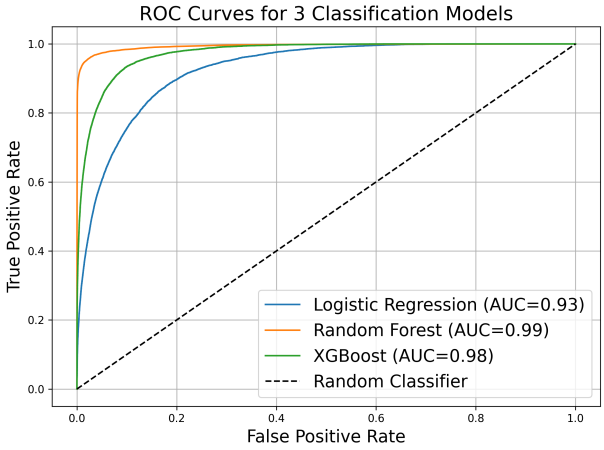


Figure 3. The ROC Curve of the Logistic regression (blue), Random Forest (Yellow), and XGBoost (Green)

The ROC curves in Figure 3 further supports these findings. The Random Forest ROC curve rises the steepest toward the top-left corner, achieving the highest ROC-AUC (0.994), which demonstrates superior discriminative capability across all classification thresholds. In contrast, logistic regression shows weaker separation, while XGBoost sits between the two. The diagonal dashed line represents a random classifier, and the random forest achieves the greatest margin.

In a nutshell, both the tabular metrics and ROC visualization highlight that Random Forest is the most effective model for predicting income groups, offering strong accuracy, robust minority-case detection, and excellent overall classification power.

3.3 The feature importance ranking from the Random forest.

We further examined the feature importance produced by the Random Forest classifier, and the top ten predictors are listed in Table 2. Age emerges as the most influential factor, followed by dividends from stocks, detailed occupation recode, capital gains, and education. These results highlight that both numerical financial indicators (e.g., capital gains, dividends) and human-capital attributes (e.g., age, education) play central roles in determining income level. Several encoded Census variables such as detailed occupation and industry recode also appear among the top contributors; however, because their encoded values do not directly map back to interpretable semantic categories, we do not analyze those recoded features in detail in this report. Aiming at segmentation analysis in the next section, we select a subset of interpretable and high-impact features—**namely age, capital gains, dividends from stocks**

(numeric), and class of worker and education (categorical)—as the primary variables for characterizing and differentiating population groups.

Table 2. the top 10th important features achieved by the random forest

Rank	Feature name	importance	Rank	Feature name	importance
1	Age	0.112	6	detailed industry recode	0.059
2	dividends from stocks	0.096	7	num persons worked for employer	0.046
3	detailed occupation recode	0.095	8	weeks worked in year	0.041
4	capital gains	0.090	9	major industry code	0.041
5	education	0.060	10	Major occupation code	0.035

4. SEGMENTATION MODELING AND MARKETING RECOMMENDATIONS

In this section, the author develops a segmentation model to identify meaningful population groups that can support targeted marketing strategies. Because the original dataset contains a mix of categorical and numerical variables with relatively high dimensionality, the author first applies Principal Component Analysis (PCA) to project the data into a lower-dimensional space while retaining most of the variance. After dimension reduction, the author employed the K-means clustering algorithm to partition all the people in the dataset into distinct clusters. To determine an appropriate number of clusters, the author evaluates within-cluster dispersion using the Elbow Method, which reveals the point at which additional clusters yield diminishing improvements.

4.1 Segmentation Model

As shown in Figure 4, the author first applied Principal Component Analysis (PCA) to determine a suitable reduced dimensionality for clustering. The cumulative explained variance curve (Figure 4a) indicates that approximately 20 principal components are sufficient to retain over 80% of the information from the original feature space. This provides a compact yet informative representation of the data for downstream clustering. Next, the elbow method was employed on the PCA-transformed features to identify an appropriate number of clusters (Figure 4b). The Sum of Squared Errors (SSE) decreases rapidly between K = 2–13 and then flattens, indicating diminishing improvement beyond K = 13. Therefore, 13 clusters are selected as the optimal balance between model complexity and within-cluster cohesion.

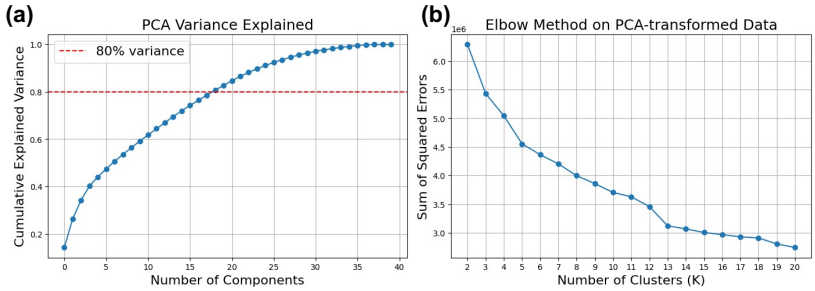


Figure 4. (a) cumulative explained variance changes with number of components; (b) Elbow method on PCA-transformed Data

Table 3. Weighted high-income rate, cluster population share, and assigned group for each cluster

Cluster index	Weighted High-income Rate	Weighted Cluster Population Share	Group
0	0.036	0.030	(3)
1	0.000	0.124	(4)
2	0.035	0.030	(3)
3	0.127	0.213	(2)
4	0.124	0.175	(2)
5	0.002	0.050	(4)
6	0.000	0.113	(4)
7	0.052	0.026	(3)
8	0.312	0.019	(2)
9	0.016	0.101	(5)
10	0.782	0.001	(1)
11	0.016	0.009	(5)
12	0.056	0.033	(3)

To further get what each of cluster people like, the author calculated the weighted high-income rate and cluster population share of each cluster listed in table 3. Interestingly, the cluster 10 has up to 78% of high-income people. Only the high-income cluster only take 0.001 percent weight in the total population. By observing the distribution of the important features mentioned in section 3 of each cluster. We combine the 13 cluster into 5 major groups: (1) high-income group; (2) Middle-class earners; (3) General Working population; (4) Young individuals; (5) Senior or retired individuals, the corresponding group assign is also listed in Table 3.

4.2 Characterization of the 5 group of people

Figures 5 and 6 show the distributions of key features such as class of worker, education, age, capital gains, and dividends from stocks. From these profiles, the five population groups can be characterized as follows**: High-income group:** Wide age range, high educational attainment, and substantial income from capital gains and stock dividends. **Middle-class earners:** Mostly ages 20–50, moderate to higher education, and primarily employed in private-sector jobs. **General working population:** Similar ages to middle class but lower education (often high school) and mainly private-sector or manual occupations. **Young individuals:** Limited work engagement, often students or early-career individuals with low income. **Senior/retired individuals:** Older adults with low labor-force participation and modest income sources.

4.3 Market suggestion

High-income group: Promote premium products, investment services and personalized financial solutions such as trust funds and managed financial porfolios. **Middle-class earners:** focus on insurance and mortgages products. **General working populations:** discounts and low budget products and services. **Young individuals:** offer them education-related services, entry-level financial products; **Senior retired group:** offer healthcare insurance plan, and retirement financial funds.

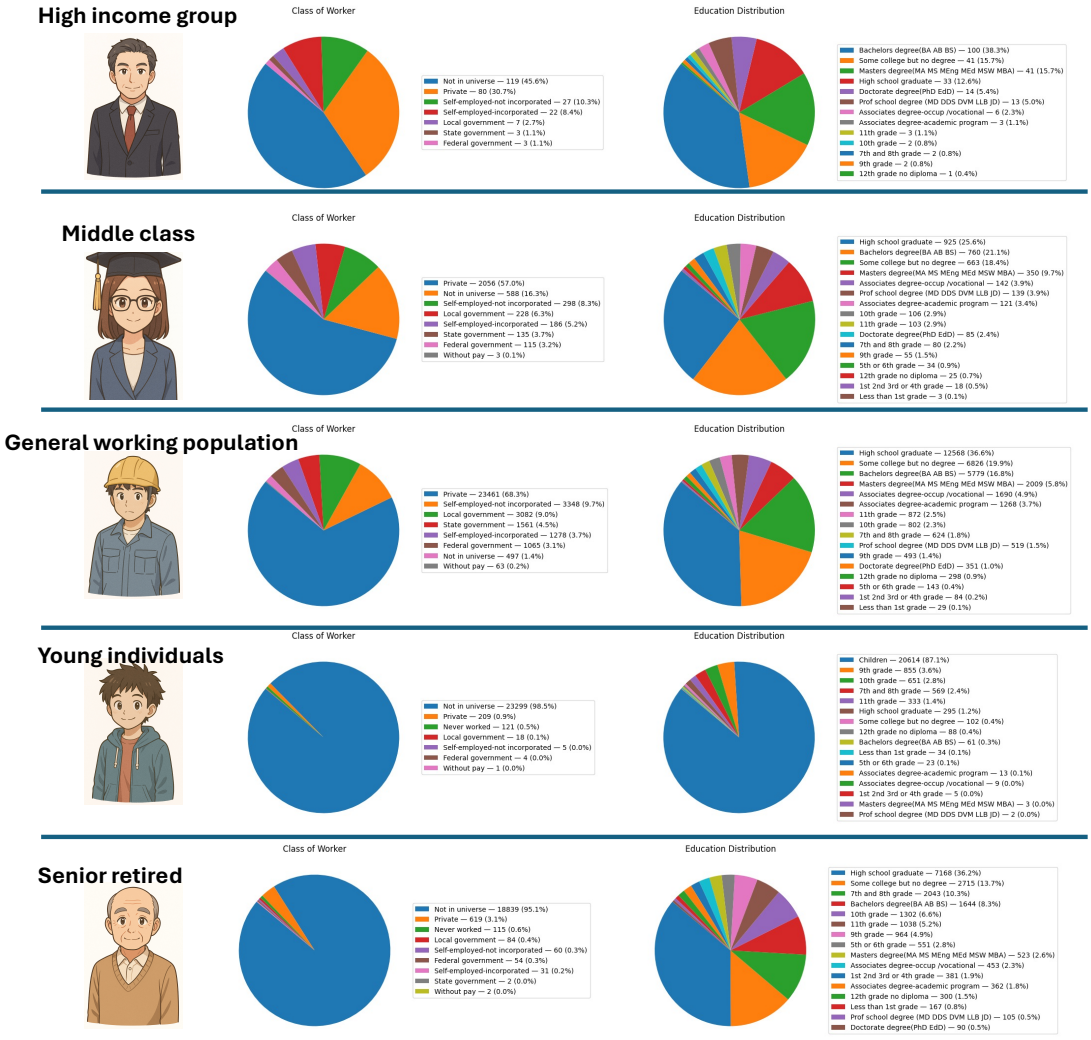


Figure 5. Class of worker (middle) and education (right) distribution of the five group of people

Figure 6. Age, capital gains, and dependent from stock distribution of the 5 group of people

5. CONCLUSION AND COMMENTS

In this project, we developed the classifier and segmentation model for marketing propers, in the classifier, random forest has the best performance and in the segmentation model, 5 main group of people are found the high-income group 78% of people gains more than $50,000 and most of fundings are from stock and capital gains, corresponding marketing plan are made with corresponding group. As for the comments and future outlook, the model-agnostic tools like SHAP values may further strengthen the insights of the dataset and adaptive learning techniques may be utilized to continuously update the model as new demographic data becomes available.

Reference:

(1) LaValley, M. P. Logistic Regression. *Circulation* **2008**, *117* (18), 2395–2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658.

(2) Sekhar, Ch. R.; Minal; Madhu, E. Mode Choice Analysis Using Random Forrest Decision Trees. *Transp. Res. Procedia* **2016**, *17*, 644–652. https://doi.org/10.1016/j.trpro.2016.11.119.

(3) Zhang, P.; Jia, Y.; Shang, Y. Research and Application of XGBoost in Imbalanced Data. *Int. J. Distrib. Sens. Netw.* **2022**, *18* (6), 15501329221106935. https://doi.org/10.1177/15501329221106935.