

**Client Report**

# **Income Prediction & Customer Segmentation Using CPS Data**

Harshit Kaushik

University of California, Berkeley

13 December 2025

## 1. Executive Summary:

The goal of this study was to predict individual income levels and identify high-value customer groups using the U.S. Census microdata for marketing purpose and product targeting by JPMorgan's Commercial & Investment Banking (CIB) division. We have leveraged machine learning, statistical analysis and clustering based segmentation techniques to identify these potential customer groups.

Our exploratory data analysis (EDA) revealed a highly imbalanced income label. This finding helped us decide several modeling steps, such as, class weighting, prioritizing the use of holistic metrics (ROC, AUC and Precision-Recall) instead of accuracy. Our findings suggest that **education, weeks worked in a year, occupation, marital status, and capital gains** displayed **strong correlation** with income.

In the second phase, a clustering framework was developed using k-means algorithm to segment the adult population into separate groups for marketing purpose. We identified four distinct customer segments. One cluster, characterized by higher education, full-year employment, professional occupations, married households and having substantial capital gains emerged as a high value segment, with an estimated **88%** probability of earning above \$50K. Our feature-importance analysis also confirmed that *work intensity* and *wealth* related attributes are the strongest determinants of high-income status.

Therefore, this study provides actionable business insights into income prediction, customer heterogeneity and financial product targeting, enabling a more efficient and precise client acquisition.

## 2. Data Understanding:

### 2.1 Dataset overview:

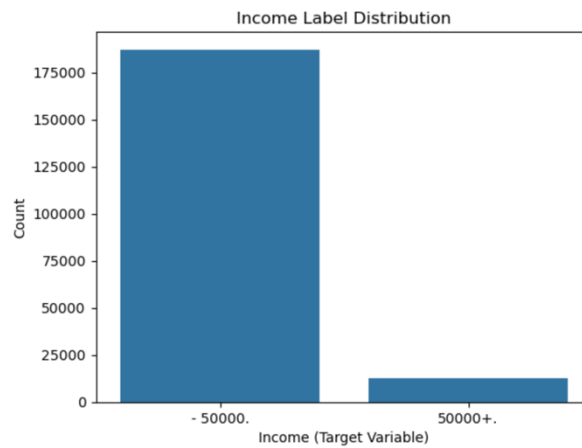
Our dataset [1] contains ~199,000 individual-level records and over 40 socioeconomic variables. The target variable indicates whether an individual earns more than \$50k annually, with a strong class imbalance. Initial descriptive statistics revealed that many categorical fields contain “*Not in universe*” values, wealth variables such as capital gains and dividends are highly right-skewed and work-related variables exhibit wide variation in labor force participation. These insights informed our data cleaning, feature engineering and modeling strategies used throughout the project.

### 2.2 Exploratory Data Analysis:

In this section, we identify drivers of income, tackle class imbalance and prepare a downstream preprocessing pipeline for modeling and segmentation. Additionally, the descriptive EDA statistics were computed with respect to survey weights to ensure accurate population representative insights.

#### 2.2.1 Income Distribution & Class Imbalance:

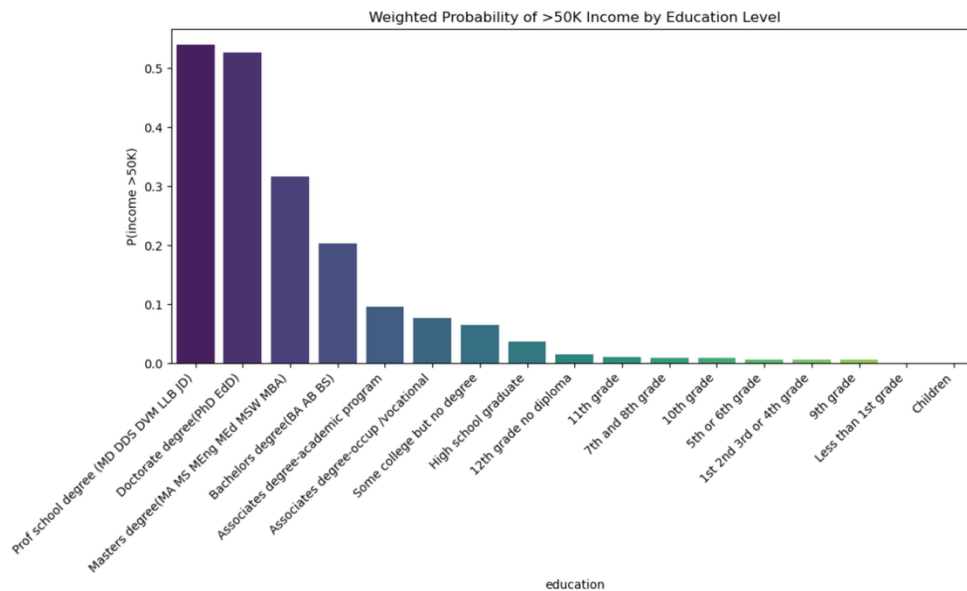
Weighted estimates revealed that only a small minority of individuals in the population earn more than \$50,000 annually. This weighted probability of high income was significantly lower than the raw dataset proportion, which confirmed a severe class imbalance. To tackle this, we used synthetic minority oversampling technique (SMOTE), weight scaling in XGBoost and evaluated model performance using ROC-AUC and Precision/Recall over accuracy. **Fig 1.** depicts this class imbalance.



**Figure 1.** Bar plot to visualize label class imbalance

### 2.2.2 Education as a Primary Driver of Income:

In Fig. 2, we can observe a correlation between educational attainment and income. Individuals with advanced degrees showed weighted probabilities exceeding **50%–55%**, while those with only high school or lower education had probabilities closer to **0–10%**. Hence, education is one of the strongest socioeconomic predictors of income.



**Figure 2.** Weighted Probability of >\$50K income by Education Level

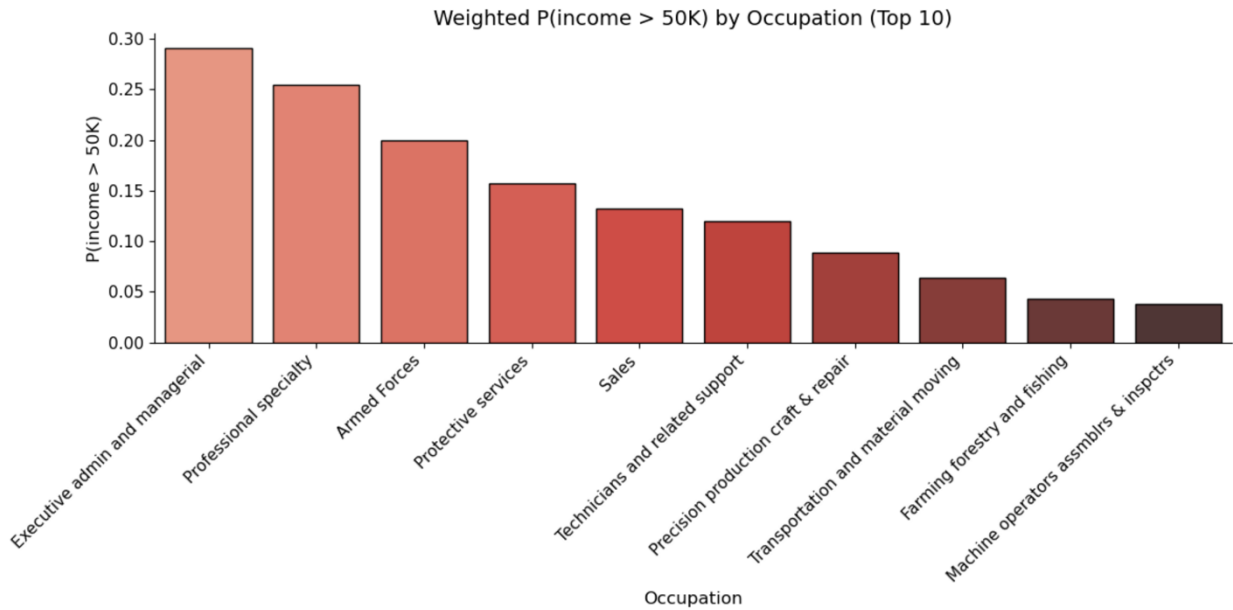
Our main learning from this chart was to conduct an **Ordinal Encoding** of education column during feature modeling and inclusion of education level as a core clustering feature. This could prove a strong anchor for customer segmentation.

### 2.2.3 Occupational Segmentation of Income (Fig 3):

This was an interesting insight because the occupational patterns revealed clear stratification in earning potential. Findings are listed below.

- Executive and Professional specialty roles showed the highest probabilities of earning >\$50K.

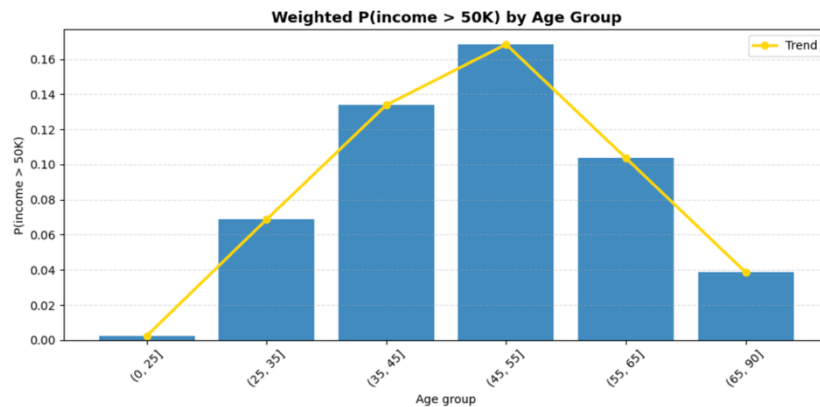
- Sales and Service roles formed a middle tier with moderate probabilities (~10%–20%).
- Low skill occupations, such as Handlers, Machine Operators, farming etc., registered very low probabilities and often below 3%.



**Figure 3.** Weighted probability of >\$50K income by occupation level

These findings influenced my modeling strategy in two ways. First, it encouraged me retaining occupational codes despite their high cardinality and second, prioritizing One-Hot Encoding instead of collapsing categories, preserving important economic distinctions.

**2.2.4 Age as a Nonlinear Predictor of Income:** Age and income patterns demonstrated a nonlinear lifecycle effect, as it is seen with the trendline in **Fig. 4**. Young individuals (under 25) had near-zero probability of high income, whereas income probability climbs steadily through ages **35 to 55**, peaking at ~17%. However, it declines sharply after age 60, consistent with retirement trends. This plot validated the use of age, as a continuous predictor and justified not binning age for modeling.



**Figure 4.** Trend analysis of Age with respect to Income

### 2.2.5 Summary of Statistical Findings:

In the above plots, we observed that across features like education, occupation and age, a consistent theme emerged: *high income is driven by human capital (education), wealth indicators (capital gains), specific age bands, and the career type (occupation of the population group).*

## 3. Data Cleaning and Preprocessing:

**3.1 Handling Missingness and Structural Categories:** Categorical ‘NaN’ values were replaced with ‘unknown’ and numerical ‘NaN’ were replaced with median of the column, because it is robust to the skew of outliers. This dataset also includes responses such as “*Not in universe*”, reflecting individuals not eligible for certain labor or migration questions. Rather than treating these as traditional missing values, these responses were preserved as meaningful categorical indicators, especially for:

1. **Employment-related attributes** (for example, reason for unemployment, labor union status)

2. **Family/household fields** (for example, household summary, migration status)

### 3.2 Feature Standardization:

Numeric features exhibited substantial variation in scale. Capital gains peaked at 99,999, but wage per hour reached 9,999. To tackle this, we applied gaussian scalers achieving mean 0 and standard deviation 1 on the dataset. This step is also crucial for segmentation because K-Means clustering relies on Euclidean distance.

### 3.3 Data Encoding:

#### Ordinal Encoding (Education, Marital Status):

The column education was mapped to an ordered 1–16 scale, indicating increasing academic attainment. This also improved clustering and modeling accuracy by allowing the algorithm to learn income gradients associated with educational hierarchy (correlation shown in EDA section). However, marital status was categorized into simplified ordinal values representing household stability:

- 2 = Married
- 1 = Divorced/Separated
- 0 = Never married

These mappings reflect socioeconomic patterns relevant to financial behavior and income stability.

#### One-Hot Encoding (Occupation, Industry, Citizenship, etc.)

Categorical variables were one-hot encoded to avoid introducing false ordinal relationships and to preserve fine-grained occupational distinctions.

### 3.4 Tackling Class Imbalance:

*Synthetic Minority Oversampling Technique (SMOTE)* [3] artificially generates new minority-class samples by interpolation. SMOTE features were used for Logistic regression (Baseline) and Random Forest (Model 1). For the XGBoost algorithm positive class weighting was used.

*Class Weighting:* This technique maintains true data geometry while penalizing misclassification of the minority class. Its logical intuition is displayed in Eq. 1:

$$pos\_weight = \frac{negative\_samples}{positive\_samples} \quad (1)$$

In our case, positive weight came up to be 15.11 units.

#### 4. Predictive Modeling:

This section explains the modeling strategy used to identify groups of people earns more or less than \$50K. The goal is to develop an interpretable and explainable model that could support business decision-making, along with enough evidence to justify the results, especially in scenarios involving customer segmentation. We started with a simple baseline and later moved to other complex architectures.

##### 4.1 Model Architectures:

**1. Logistic Regression (Baseline):** Logistic regression provides a transparent and linear decision boundary with coefficients that map directly to feature importance through a Logit function. We chose this because it is faster to implement, easy to audit and therefore aligns with financial regulatory expectations around explainability.

**Architecture configuration:** We leveraged L2 regularization along with maximum iterations of 4000 for convergence. Also, since it was a baseline, it was trained on SMOTE balanced and standardized numeric features

**2. Random Forest (Model 1):** Random Forests can model complex interactions and learn better correlations, for example, how age aligns with occupation due to a better education. It is also robust to noise and provides feature importances useful for business insights.

**Architecture configuration:** We used 300 trees ( $n\_estimators$ ) during training, with a default depth parameter (a full-grown tree without pruning) and trained it on a SMOTE balanced data.

**3. XGboost (Model 2):** The final model was XGBoost, which is widely used in modern financial modeling because it handles class imbalance (through the parameter:  $scale\_pos\_weight$ ) and understands complex nonlinear relationships.

**Architecture configuration:** XGBoost's architecture has regularization (L1 and L2 penalties) to control model complexity, shrinkage via learning rate, and column and row subsampling to reduce overfitting and improve generalization. In this model, learning was controlled using a  $learning\_rate = 0.07$  with a fixed ensemble size of 250 trees, allowing the model to incrementally refine predictions while reducing the risk of overfitting.

#### 5. Training algorithm Pipeline:

**Input:** Dataset  $\mathcal{D} = (X, y)$ , where  $y \in \{0,1\}$

1. Split  $\mathcal{D}$  into training and test sets.  $(X_{train}, y_{train})$  and  $(X_{test}, y_{test})$  using stratification to preserve class proportions.
2. Fit a scaler on  $X_{train}$  and apply it to both  $X_{train}$  and  $X_{test}$ .
3. Handle class imbalance by applying SMOTE to  $(X_{train}, y_{train})$  only for Baseline and Random Forest, additionally set a class weight  $w_+$  for XGBoost, based on the imbalance ratio.

4. Train the model  $\mathcal{M}$  on the balanced training data.
5. Evaluate  $\mathcal{M}$  on  $(X_{\text{test}}, y_{\text{test}})$  using ROC-AUC, precision, recall, F1-score, the confusion matrix, and probability-based ROC analysis.

## 6. Model Evaluation:

**Logistic Regression** demonstrated strong overall accuracy (94%) and solid discriminatory ability (ROC-AUC  $\approx 0.926$ ). However, performance on the **high-income** class was modest and only with a precision of **0.54** and recall of **0.46**. This means the model correctly identifies less than half of high-income individuals, though it rarely misclassifies low-income customers.

**Random Forest** offered the most balanced performance across metrics, achieving **95%** accuracy and a ROC-AUC of **0.937** (**Fig. 5**). It improved precision for high-income customers (**0.63**) while maintaining similar recall to logistic regression (0.46). The model captures nonlinear relationships between demographic and labor-market variables, and its feature importance outputs remain relatively interpretable. This makes Random Forest especially useful for *segmentation and marketing use cases* where predictive strength and explainability coexist.

**XGBoost** achieved the highest ROC-AUC (**0.941**) and captured almost all high-income individuals, with a recall of **0.95**. However, its precision for the high-income class was low (**0.21**), meaning it generates many **false positives**, and it will flag numerous low-income customers as high-income. This behavior is expected from boosting models that are optimized for high recall. In my opinion, XGBoost is therefore ideal for top-of-funnel marketing, wealth-management outreach or early-stage lead expansion where “missing a high-value customer” is more costly than reviewing extra candidates.

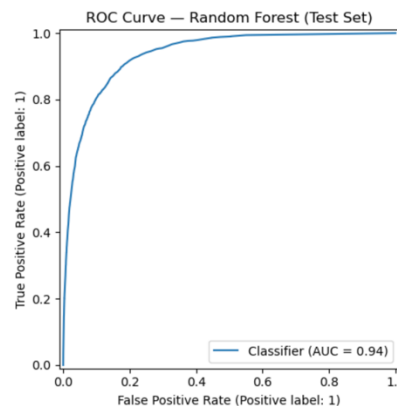
### Model usage recommendations:

- Start with **XGBoost** for ranking and prospect/customer scoring.
- Use **Random Forest** to identify key drivers for customer segmentation and marketing strategy.
- Use **Logistic Regression** where transparency and auditability are required.

Table 1 presents a comprehensive analysis of the results obtained through number of classification models. *More detailed classification report numbers are in the GitHub file.*

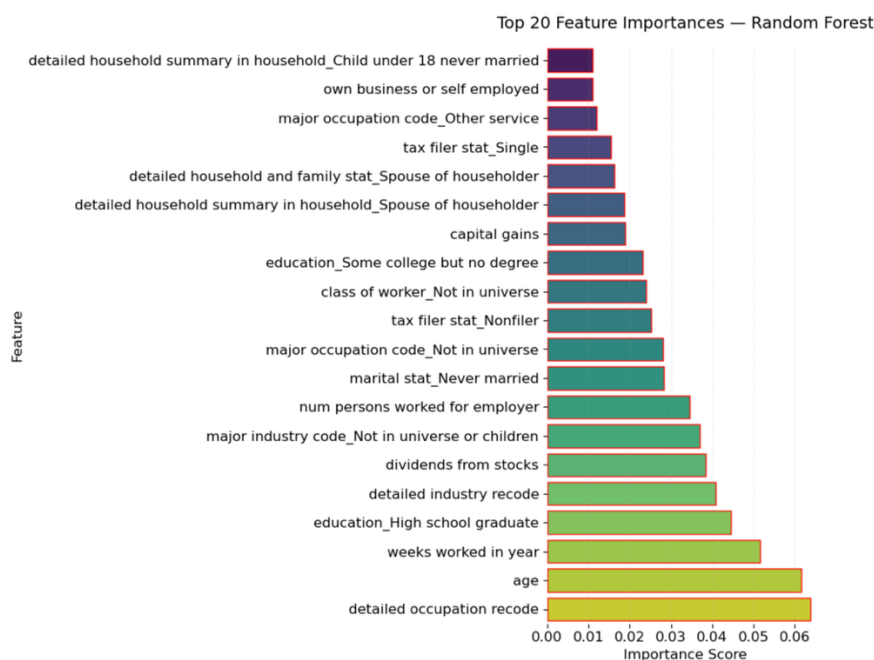
**Table 1.** Evaluation metrics for Classification models

Model	ROC-AUC	Precision (High-Income)	Recall (High-Income)	F1-Score (High-Income)	Overall Accuracy	Key Strength
Logistic Regression	0.926	0.54	0.46	0.5	0.94	Highly interpretable, regulatory-friendly
Random Forest	0.937	0.63	0.46	0.53	0.95	Strong accuracy + feature interpretability
XGBoost	0.941	0.21	0.95	0.34	0.77	Best recall and has excellent ranking power



**Figure 5.** ROC curve for Random Forest on unseen Test set (AUC = 0.94)

Feature importances (**Fig. 6**) from Random Forest (best performing model) aligned strongly with our EDA. For example, **education level, occupation type, weeks worked in a year and age**. These results reinforce the consistency of our data-driven EDA. These variables exhibit strong correlation with earnings, and tree-based models are naturally effective at capturing such interactions.



**Figure 6.** Feature importance score by Random Forest

## 7. Customer Segmentation Model:

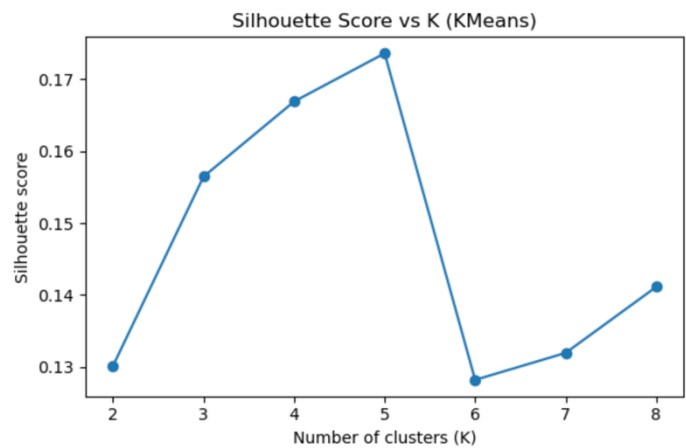
To support the retail banking client's requirement a segmentation model was developed using the K-Means clustering algorithm. K-Means was chosen because of its strong performance on large, structured tabular datasets and its interpretability. My objective was not only to cluster the dataset statistically, but to create **stable, economically meaningful customer groups** that the JPMC teams could operationalize.

Before running the clustering model, few preprocessing steps were applied to ensure that the resulting segments reflected behavioral difference. **First**, individuals classified as “Children” (age <18) were **removed**. This group consistently formed its own cluster with zero weeks worked and no income. Including them artificially distorted cluster boundaries and consumed one full cluster regardless of the choice of  $k$ . Their removal ensured that the segmentation focused on economically active adults, which aligns with real-world business applications. **Second**, rather than performing clustering on the full set of 80+ census variables, the model used a small number of high-signal features derived from earlier weighted EDA. Variables such as *age*, *weeks worked per year*, *education level*, *capital gains*, and *marital status* demonstrated strong associations with income. We avoided the curse of dimensionality, which occurs when distance measures lose significance in very high-dimensional space.

**Optimal number of cluster (K=4 instead of 5):** The Silhouette Score was chosen because it measures both cluster cohesion and separation and showed that  $k = 5$  in **Fig. 7** achieved the highest numerical value ( $\approx 0.173$ ).

However, after examining the cluster assignments, the five-cluster solution produced two very similar mid-career, moderate-income groups that differed only slightly in work intensity. These groups offered no meaningful differentiation for marketing or product strategy. In contrast,  $k = 4$  produced cleaner, economically distinct, and more actionable segments, separating (1) young low-income workers, (2) prime-age high-income professionals, (3) near-retirement individuals with limited work hours, and (4) stable mid-career workers with moderate earning potential. Although  $k=5$  showed a marginally higher silhouette score,  **$k = 4$  was selected as the final model because it delivered superior interpretability** which is the central goal of segmentation in a financial-services context.

The final clusters (**Table 2**) reveal clear differences in age, labor participation, education level, and income patterns. These differences translate directly into distinct customer personas with different needs, behaviors, and product affinities—forming the basis for targeted marketing, differentiated service models, and more accurate customer value assessments.



**Figure 7.** Silhouette Score for different  $k$  values

**Table 2.** Cluster analysis based on input features and probability of income >50K

cluster	age_mean	age_median	weeks_worked_mean	weeks_worked_median	education_mode	marital_status_mode	occupation_mode	income_binary_mean
0	26.225777	24.0	9.436359	0.0	High school graduate	Never married	Not in universe	0.003116
1	46.869231	46.5	47.458974	52.0	Bachelors degree(BA AB BS)	Married-civilian spouse present	Professional specialty	0.882051
2	68.766165	69.0	2.723369	0.0	High school graduate	Married-civilian spouse present	Not in universe	0.021713
3	39.638089	39.0	49.564557	52.0	High school graduate	Married-civilian spouse present	Professional specialty	0.132385

8. Business Recommendations:

**Cluster 0 (Emerging Workforce / Low Income Potential):** This cluster is composed of very young adults with minimal work experience and limited educational attainment. Their low labor-force engagement (median 0 working weeks) results in extremely low earning capacity. From a banking perspective, these individuals function as emerging customers. They are unsuitable for income-dependent lending but represent an important

entry point for long-term relationship building. **The teams at JPMC can target this group with beginner-friendly products such as secured credit cards and student accounts.**

**Cluster 1 (High-Income Professionals / Core Profit Segment):** This is the most economically attractive group, defined by higher education, full-year employment and professional occupations. Also, consistent with the EDA findings, this cluster shows the strongest likelihood of earning >50K. These customers have both the financial capacity and behavioral profile associated with long-term profitability. **They are prime candidates for premium credit cards, investment products and low-risk lending,** making this cluster the centerpiece for revenue growth and cross-sell strategies.

**Cluster 2 (Older Adults / Low Workforce Participation):** This cluster captures older individuals or retirees who work infrequently and predictably show low annual income. While they may not be strong candidates for traditional lending, **they represent a stable deposit base and are ideal for retirement planning, annuities, insurance products and estate management services.**

**Cluster 3 (Full-Time Workers with Moderate Income Potential):** These customers are fully active in the labor market and frequently employed year-round but typically possess lower formal education than Cluster 1. **This group aligns well with financial offerings such as auto loans, personal loans and mid-tier credit cards.** With targeted guidance and product design, they can transition into higher-value segments over time.

## 9. Future Work:

As a future extension of this work, I believe incorporating causal inference would help uncover *why* specific demographic patterns drive income outcomes. Additionally, evaluating fairness and bias is essential as income-based models risk reinforcing existing socioeconomic disparities. Integrating fairness metrics and mitigation strategies would be a necessary step before any real-world deployment to ensure the model supports equitable and responsible financial decision-making.

## 10. Conclusion:

In summary, this report demonstrates that structured demographic and employment attributes can effectively predict income tiers and reveal actionable customer segments. Weighted EDA consistently highlighted education level, occupation and work intensity as the strongest drivers of earning potential, insights that were later validated through Random Forest and XGBoost feature importance. The predictive models achieved strong performance, showing that income estimation can be operationalized to support downstream decisioning such as lead scoring, risk assessment, and product eligibility. The segmentation model further uncovered four clear customer groups differentiated by life stage, employment stability and earning capacity, providing a practical framework for targeted marketing, tailored product design, and more efficient cross-sell strategies. While this is not a complete customer-360 solution, it establishes a solid analytical foundation. This study should be viewed as a scalable starting point that can be expanded into a more comprehensive customer intelligence system for JPMC.

## 12. References:

- [1] <https://www.census.gov/programs-surveys/cps.html>
- [2] <https://scikit-learn.org/stable/>
- [3] <https://www.jair.org/index.php/jair/article/view/10302>
- [4] <https://www.jpmorganchase.com/ir/annual-report/2024/ar-ceo-letters>
- [5] Grus, Joel. *Data science from scratch: first principles with python*. O'Reilly Media, 2019.