Client Report on

Income Prediction & Customer Segmentation Using CPS Data

Harshit Kaushik

University of California, Berkeley (Masters)

13 December 2025

Role: Applied ML/AI Associate at JP Morgan

1. Executive Summary:

In this take-home project we have applied ML and clustering based segmentation techniques to predict individual income levels and identify high-value customer groups using the U.S. Census microdata. The goal was to evaluate the feasibility of using specific attributes to identify potential high-income clients for JPMorgan's Commercial & Investment Banking (CIB) division.

We started by conducting an exploratory data analysis revealing a **highly imbalanced income distribution**. This finding helped us decide several modeling decisions, such as use of class weighting and prioritizing the use of holistic metrics such as ROC, AUC and Precision-Recall curves over accuracy. In my research, variables such as education, weeks worked in a year, occupation, marital status, and capital gains (major attribute) displayed **strong correlation** with income. This eventually guided my preprocessing and feature engineering strategy later in the study.

To complement the supervised machine learning, an unsupervised clustering framework was developed to segment the adult population into meaningful socioeconomic groups. Using K-Means on curated features, 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 professional segment, with an estimated **88%** probability of earning above $50K. Feature-importance analysis confirmed that work intensity and wealth related attributes are the strongest determinants of high-income status

In conclusion, these analyses provide actionable insights into income prediction, customer heterogeneity, and financial product targeting, ultimately enabling more efficient and precise client acquisition within JPMorgan's CIB context.

2. Data Understanding & Assumptions:

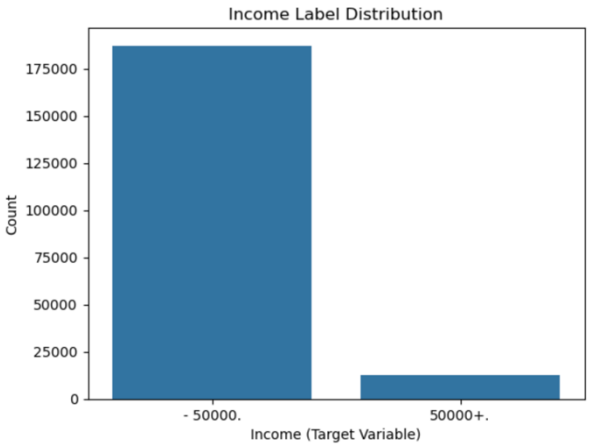
2.1 Dataset overview:

The dataset of this study 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 several important characteristics of the dataset: many categorical fields contain "*Not in universe*" structural categories, income-related 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 the cleaning, feature engineering and modeling strategies used throughout the project.

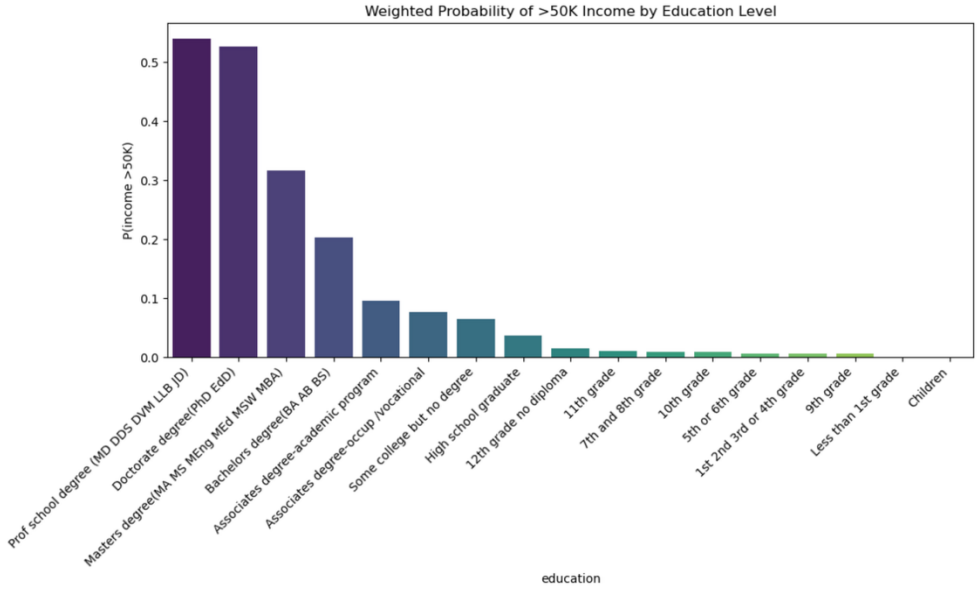
2.2 Exploratory Data Analysis – Key takeaways:

In this section, I specifically work to identify drivers of income, tackle class imbalance and prepare a downstream preprocessing pipeline for modeling and segmentation. The interesting fact about this section is that *all EDA statistics (descriptive) were computed using survey weights, ensuring population representative insights.*

2.2.1 Income Distribution & Class Imbalance: During EDA, 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. This finding influenced multiple modeling choices, including the use of scale_pos_weight in XGBoost and prioritizing metrics such as ROC-AUC and Precision/Recall over accuracy. **Figure 1.** shows a bar plot explaining this class imbalance.



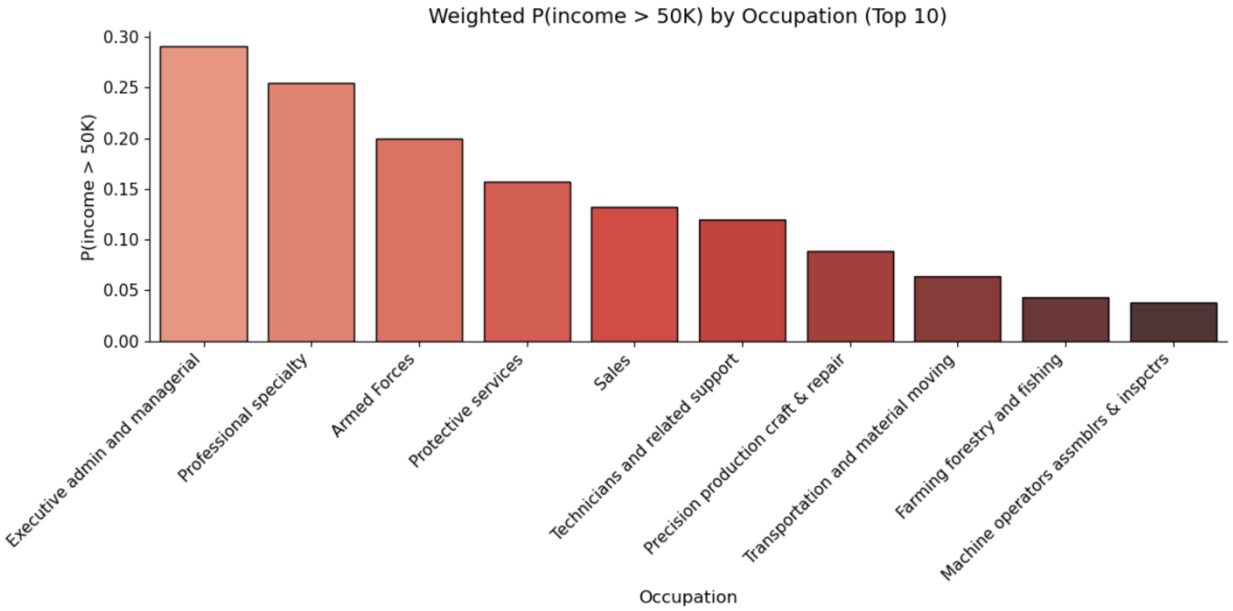
2.2.2 Education as a Primary Driver of Income (Figure 2): We observe the weighted Probability of >$50K income by Education Level. We can see a monotonic and nonlinear relationship between educational attainment and income. Individuals with advanced degrees showed weighted probabilities exceeding **50%–55%**, 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.



Important: My main observation from this chart for feature modeling was to conduct an **Ordinal Encoding** of education during clustering (reflecting the true educational hierarchy) and Inclusion of education level as a core clustering feature. This creates a strong anchor for customer segmentation.

2.2.3 Occupational Segmentation of Income (Figure 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%.

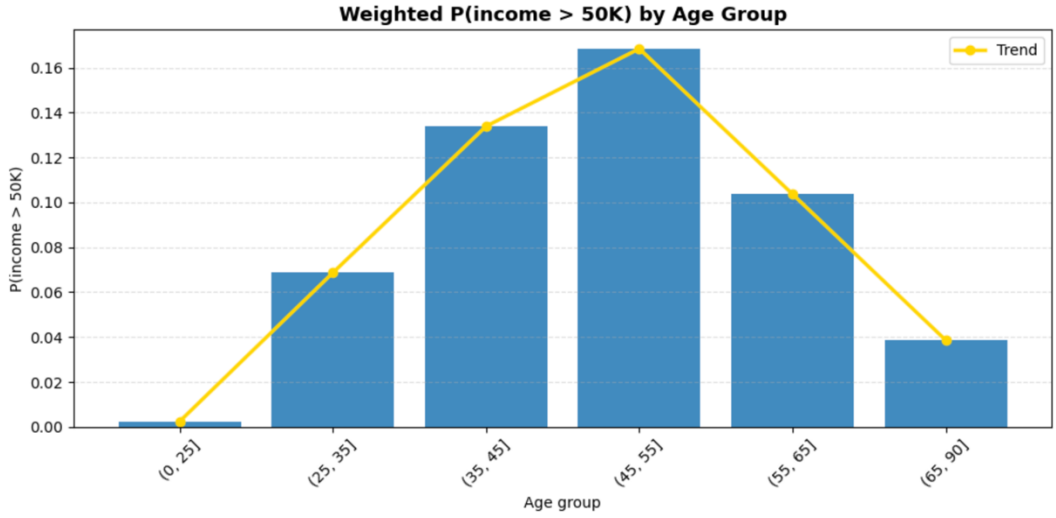


Important: These findings influenced my modeling strategy in two ways. First, it encouraged me retaining occupational codes despite their high cardinality and second, prioritizing **one-hot encoding** instead of collapsing categories, preserving important economic distinctions.

2.2.4 Age as a Nonlinear Predictor of Income (Figure 4): Age and income patterns demonstrated a nonlinear lifecycle effect, as it can be seen with the trendline:

1. Young individuals (under 25) had near-zero probability of high income.
2. Income probability climbs steadily through ages **35 to 55**, peaking around **17%**.
3. It declines sharply after age 60, consistent with retirement trends.

Important: This plot validated the use of the feature, age, as a continuous predictor and justified not binning age for modeling. Additionally, this pattern helped interpret clusters, for example, young & entry-level cluster vs peak-earnings cluster.



2.2.5 High level summary of statistical patterns: 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), labor supply (weeks worked in that year) and the career type (occupation of the population group), with wealth indicators.*

These insights were directly applied into:

1. Feature engineering decisions (ordinal encodings, numeric scaling)
2. Modeling priorities (focusing on high-signal socioeconomic features during feature selection)
3. Cluster design (including education level, weeks worked, age, marital status and capital gains as drivers of segmentation)

4. Data Cleaning and Preprocessing:

4.1 Handling Missingness and Structural Categories: This dataset 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)

Moreover, categorical 'NaN' was replaced with 'unknown' and numerical 'NaN' were replaced with median of the column, which is robust to the skew of outliers.

4.2 Feature Standardization for Outlier Awareness: Numeric features exhibited substantial variation in scale. For example, capital gains peaked at 99,999, while wage per hour reached 9,999. The model XGBoost is tree-based (and thus insensitive to scale), standardization was not applied for supervised modeling. However, standardization was crucial for K-Means clustering, which relies on Euclidean distances.

4.3 Encoding Categorical Variables:

Ordinal Encoding (Education, Marital Status):

During my experiment, 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. 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.)

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

4.3 Addressing Class Imbalance: SMOTE vs Class Weighting:

Synthetic Minority Oversampling Technique (SMOTE) artificially generates new minority-class samples by interpolating between existing high-income individuals. However, Since CPS data weights reflect real demographics, generating artificial individuals can break the statistical design. Therefore, SMOTE was used **only** for baseline algorithms such as Logistic regression and Random Forest, however, for final XGBoost algorithm class weighting was used.

Class Weighting: This technique maintains *true data geometry* while penalizing misclassification of the minority class. This was implemented as below:

scale_pos_weight = (negative samples / positive samples)

5. Predictive Modeling: This section explains the modeling strategy used to identify groups of people earns more or less than $50K annually. My goal was to develop interpretable models that could support business decision-making, especially in scenarios involving customer segmentation.

5.1 Model Architecture & Reasoning:

1. Logistic Regression (Baseline): Logistic regression provides a transparent, linear decision boundary with coefficients that map directly to feature importance. I chose this as it's easy to audit and also aligns with financial regulatory expectations around explainability.

Architecture: In this model I leveraged L2 regularization along with maximum iterations of 2000 for convergence. Also, since it was a baselined, it was trained on SMOTE balanced and standardized numeric features

2. Random Forest (Baseline): Random Forests can model complex interactions, for example, age x occupation x education, which are common in socioeconomic data. It is also robust to noise and provide feature importances useful for business insights.

Architecture: For hyperparameters, I leveraged 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, like Logistic Regression.

3. XGBoost (Enterprise Grade Gradient Boosting Model): My 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: 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.

5.2 Training algorithm Pipeline:

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

1. Split $\mathcal{D}$ into training and test sets. $(X_{\text{train}}, y_{\text{train}})$ and $(X_{\text{test}}, y_{\text{test}})$ 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 $(X_{\text{train}}, y_{\text{train}})$ only for XGBoost, additionally set a class weight w_+ 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.

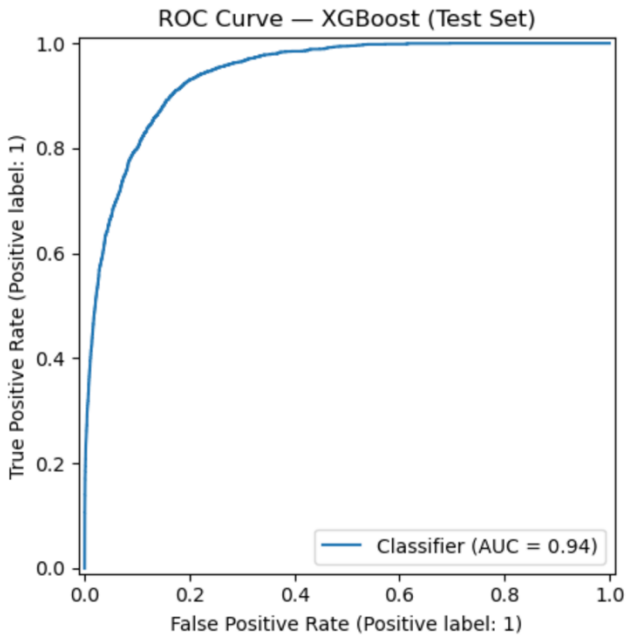
5.3 Model Evaluation and Business Insights: The table 1 presents a comprehensive analysis of the results obtained through number of classification models.

Model	AUC	Minority Recall	Minority Precision	Notes
Logistic Regression	~0.926	~0.46	~0.54	Strong interpretability; linear boundaries insufficient
Random Forest	~0.937	~0.53	~0.63	Best balance; strong feature insights
XGBoost	**~0.941**	~0.34 (at t=0.5)	~0.21	Highest AUC; threshold tunable for business needs

If **interpretability & compliance** matter most → choose **Logistic Regression** because it is ideal for regulated financial decisions (credit scoring, loan underwriting).

If the goal is **maximizing correct identification** of high-value customers → choose **XGBoost** because it has best discriminatory power (Fig. 5) and therefore catches subtle nonlinear signals that logistic regression cannot. If JPMC wants to **rank customers** by likelihood of being high-income (for targeted offers, wealth management leads or customer scoring), XGBoost is the strongest candidate.

If the goal is a balance between **power & interpretability** → choose **Random Forest** because it is easier to interpret than boosting models yet performs very competitively.

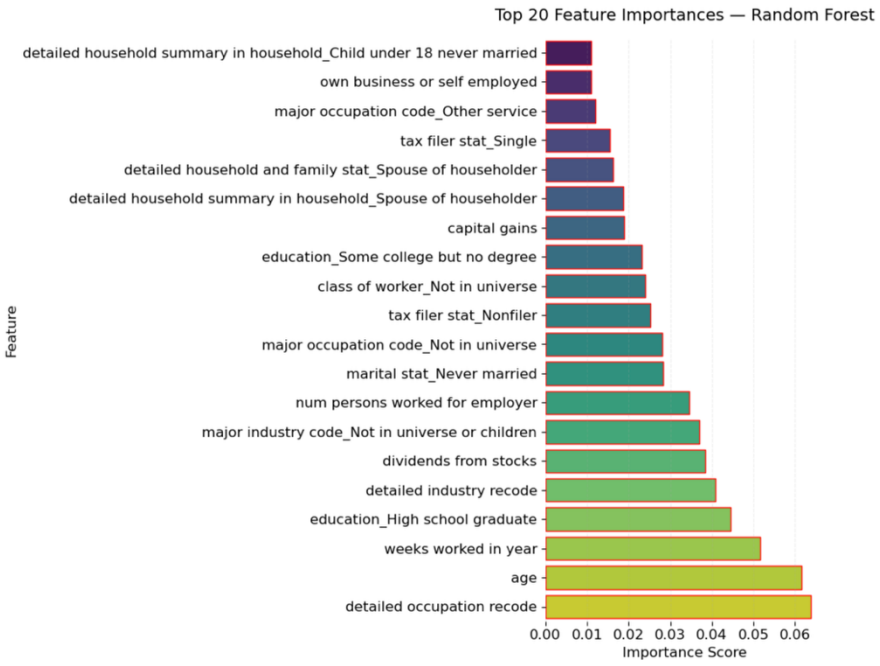


Model usage recommendations:

- 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.

5.4 Additional Insights: Feature importances from Random Forest and XGBoost aligned strongly with weighted EDA for example, education level, occupation type, age group, capital gains. This reinforces the consistency of our data-driven approach.

Important (ref Fig 6): The importance results affirm our patterns revealed during weighted EDA in previous section. High-ranking features such as **capital gains**, **weeks worked per year**, **age** and **occupation** align with well-established drivers of income. These continuous variables exhibit strong nonlinear relationships with earnings, and tree-based models are naturally effective at capturing such interactions.



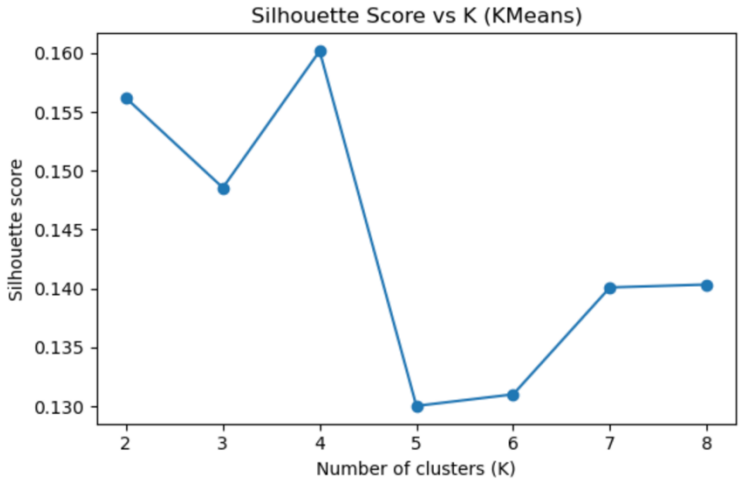
6. Customer Segmentation:

To complement the predictive modeling workflow, an unsupervised customer segmentation model is developed for marketing purposes. **K-Means clustering** was selected due to its effectiveness on large, structured tabular data and its interpretability which is an important consideration for downstream business decision-making in financial services.

Before clustering, a few preprocessing steps were applied to ensure meaningful segment formation. First, **children (age < 18)** were removed because they form a degenerate cluster with zero weeks worked, no income, and no labor-force attributes. Their inclusion artificially inflated within-cluster variance and consistently dominated one cluster during experimentation. Second, we selected a subset of **high signal features** identified from EDA sections: *age, weeks worked per year, education level (ordinal), capital gains, marital status (binary encoded)*. These variables capture the demographic and economic behaviors most strongly associated with

income. Scaling using StandardScaler ensured that all features contributed proportionally and avoided dominance by variables measured on larger numeric ranges.

How did I find cluster number? To determine the optimal number of clusters, I evaluated K-Means under multiple values of *k* and computed the **Silhouette Score**. It is a robust measure of cluster cohesion and separation. Unlike the within-cluster sum of squares, which always decreases with more clusters, the Silhouette Score penalizes overlapping groups. In this dataset, ***k = 4*** produced the highest Silhouette values (**Fig 7**), indicating that four distinct, reasonably well-separated customer groups exist in the underlying population.



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

7 Business Recommendation:

Cluster 0: This cluster represents younger adults who are early in their careers or not yet fully involved in the labor market. The extremely low work intensity (0 working weeks median) combined with low educational attainment translates into almost nonexistent earning power.

- **Business Insight:** This segment resembles **emerging or dependent customers** which is an ideal target for entry-level financial products (secured credit cards, student-friendly offerings), but **not credit-risk-efficient for lending products** due to lack of stable income indicators.

Cluster 1: This is the **highest-value segment**, consisting of experienced professionals with continuous employment and strong educational backgrounds. This aligns perfectly with EDA findings where **education level** and **professional occupations** had the highest weighted probability of earning >50K.

- **Business Insight:** This is a premium customer segment with strong revenue potential. They are eligible for investment products, premium credit cards, wealth advisory, Low risk for credit lending, High long-term customer value.

Cluster 2: It represents **older individuals likely past retirement age** or with extremely low labor force participation. Low earnings are expected not because of poor job prospects but due to lifecycle stage.

- **Business Insight:** This group may be stable in deposits but will not qualify for income-dependent lending products. However, they may be candidates for, retirement financial planning, Estate and insurance services etc.

Cluster 3: This segment consists of individuals fully engaged in the labor market, often in skilled roles but typically with **lower formal education** than the high-earning Cluster 1 group. This explains why even with full-year work effort, their income probability remains modest.

- **Business Insight:** These customers are ideal for, auto loans, personal loans, mid-tier credit cards and financial literacy resources.

8. Limitations:

8.1. **Use of Historical CPS Data**: While valuable for pattern discovery, these distributions may not fully represent current market dynamics or evolving consumer behaviors, especially in periods of economic change.

8.3. **Limited Behavioral Signals**: This dataset contains demographic and labor-market variables but lacks transactional and digital features that typically enhance segmentation.

8.4. **Model Interpretability Constraints**: While tree-based models and XGBoost perform well, they still abstract underlying causal structures. Without longitudinal data, we cannot isolate whether variables *cause* income differences or merely correlate with structural socioeconomic patterns.

9. Future Work:

9.1 Causal Modeling: As a future study I would love to incorporate causal inference to understand *why* certain demographic patterns drive income or use uplift models to evaluate which customer groups benefit most.

9.2 Fairness and Bias Evaluation: Income based models may reinforce socioeconomic biases. Incorporating fairness metrics and bias-mitigation strategies would be an important next step before deployment in a production environment.

10. Conclusion:

In this study, weighted EDA showed that education level, occupation type, and work intensity are the strongest predictors of earning potential, patterns that were later confirmed through Random Forest and XGBoost feature importance. Predictive modeling (classification models) achieved strong performance, highlighting the feasibility of using structured data to estimate income tiers and support downstream decisioning.

The final segmentation analysis revealed four distinct customer clusters aligned with life stage, employment stability, and earning capacity, offering clear implications for targeted marketing and personalized product strategy. For JPMC, these insights can support more accurate customer qualification, improved cross-sell

targeting and tailored financial advisory strategies. While additional bank-specific behavioral data would further enhance accuracy, this analysis provides a robust foundation for data-driven customer understanding and segmentation.