Problem 1: Income Classification

The first problem can be categorized as an income classification problem, and the goal is to use different demographic and employment-related variables to predict the outcome. Since the income is given as a label, this essentially falls into the supervised learning category. The problem is the dataset is really imbalanced, where the ratio of people who earn an income of less than $50,000 to those who earn more than $50,000 is 15:1. Regarding that, I use SMOTE (Synthetic Minority Oversampling Technique) for training data, which creates synthetic examples to balance the training data between both income categories. I also use Stratified Split to ensure that the testing set has the same proportion as the original dataset.

I implemented three different classification algorithms to compare performance. The first is Logistic Regression with balanced class weights and liblinear solver, which is suitable for binary classification problems. The second is Random Forest with 100 estimators and balanced class weights, which can handle non-linear relationships in the data. The third is XGBoost with 100 estimators and "scale_pos_weight" parameter set to 15 to address class imbalance.

For training, I used SMOTE to resample the training data, creating synthetic examples of the minority class, in this case people who earn more than $50,000, to achieve balanced training data. The models were trained on this balanced dataset and then evaluated on the original imbalanced test set to simulate real-world conditions.

Data Approach and Model Selection

For data approach decisions, I chose to sample the large dataset to 50,000 observations for computational efficiency while maintaining representativeness. This decision balances processing speed with data quality.

For model selection, I chose Random Forest as the best model because it achieved the highest F1-score (85.47%) and ROC-AUC (93.88%), indicating good performance on both majority and minority classes.

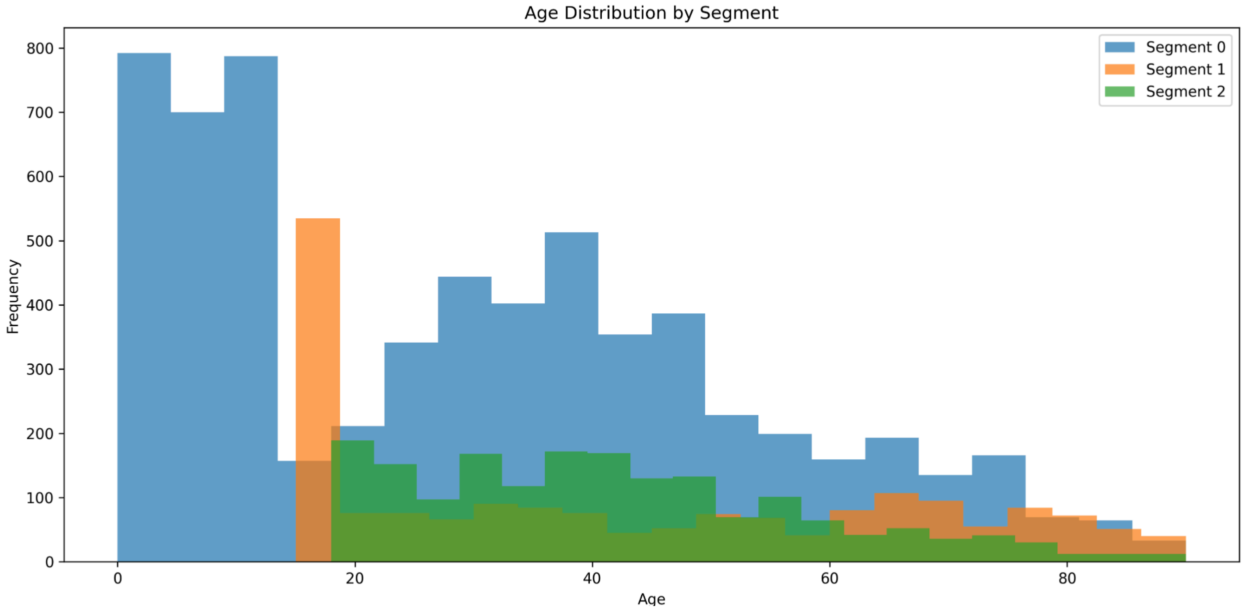
Problem 2: Customer Segmentation

The second problem is to group customers into meaningful segments for marketing purposes. This involves using unsupervised learning techniques to identify natural groupings in the customer base based on demographic characteristics.

I used K-means clustering with 3 segments as the primary algorithm. K-means is suitable for this problem because it can identify natural groupings in the customer base based on demographic similarities. The algorithm was configured with random initialization and multiple runs to ensure stable results.

Feature extraction and preprocessing were applied to prepare the demographic variables for clustering. Categorical variables such as education, occupation, marital status, race, and sex were encoded using label encoding to convert them into numerical values that the clustering algorithm could process. Numerical variables like age were standardized to ensure all features contributed equally to the segmentation process.

Here is the result:



The model identified three distinct customer segments:

Segment 0: Diverse age distribution with peaks in very young (0-15 years) and middle-aged (35-45 years) groups, suggesting a “family-oriented” segment.

Segment 1: Highly concentrated in young adults (15-20 years), representing young people.

Segment 2: Even distribution across working-age adults (20-50 years).

Segment 1 enables targeted campaigns for young adults, while Segments 0 and 2 allow for family-focused and professional marketing strategies respectively.

Reference:

1. ChatGPT 5
2. Chawla, N. V., et al. "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research, 2002.
3. Chen, T., & Guestrin, C. "XGBoost: A Scalable Tree Boosting System." KDD, 2016.