Income Classification & Customer Segmentation

1. Executive Summary

The machine learning solution is designed to help improve the marketing efficiency and return on investment (ROI). The machine learning model was trained and tested with the weighted census data from 199,523 individuals and 40 demographic and employment features. The following models were developed:

1. **A Classification Model** to predict whether an individual earns more or less than $50,000 annually.
2. **A Segmentation Model** to group customers into distinct personas for targeted marketing strategies.

Key Findings & Business Takeaways:

- **Classification Performance:** The top-performing model, an XGBoost classifier, achieves a high ROC-AUC score of **94.81%**, making it highly effective at identifying high-income prospects despite a severe class imbalance (only 6% of the data are high earners). Key predictors of high income are weeks worked in year, tax filer status, education, and sex.
- **Customer Segments:** The best clustering model selected based on the Silhouette score is KMEANS algorithm. The machine learning model identified two distinct customer clusters using K-Means clustering. These segments, "Working Class/Mainstream" and "Young Professionals," exhibit significantly different behaviors and financial profiles.
- **Business Impact:** This framework enables you to:
    - Prioritize marketing efforts on high-income prospects predicted by the model.
    - Tailor product offerings (premium vs. value-focused) to the distinct needs of each customer segment.
    - Identify groups that would most benefit from financial impact and wealth-building programs.

2. Business Problem Overview

Objective:

To improve marketing precision by understanding the customer base at a deeper level. This project addresses two core problems:

1. **Targeted Prediction (Classification):** Develop a model to accurately classify individuals into two income brackets: <=50K and >50K. This promotes efficient allocation of marketing resources toward prospects with higher potential value.
2. **Customer Understanding (Segmentation):** Create an unsupervised model to segment the customer population into meaningful groups based on their demographic and employment characteristics. This provides a framework for tailoring marketing strategies and product development to specific customer personas.

Dataset:

The analysis used a weighted census dataset derived from the 1994-1995 Current Population Surveys. It contains 199,523 observations with 40 demographic and employment-related features. The dataset is highly imbalanced, with approximately 94% of individuals earning <=50K and only 6% earning >50K.

3. Data Exploration, Preprocessing & Feature Analysis

Before modeling, a thorough exploratory data analysis (EDA) and preprocessing phase was conducted to ensure data quality and model reliability.

Key Findings from Exploratory Data Analysis (EDA):

- **Class Imbalance:**
  The dominance of the <=50K class as shown in Figure 1 is a critical challenge. As a result, standard accuracy would be a misleading metric. Therefore, I prioritized the ROC-AUC score for model evaluation on test dataset and used the SMOTE (Synthetic Minority Over-sampling Technique) algorithm during model training to balance the classes.

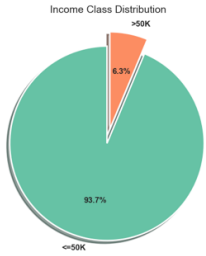

Figure 1: Class Distribution (Class Imbalance)

- **Missing Data:**
  Missing values were represented as '?' or 'Not in Universe'. They were handled by:
  - Removing columns with >70% missing data.
  - Imputing numerical features with the median to reduce the impact of outliers.
  - Imputing categorical features with the mode (most frequent value).
  - Proportion of missing values for top 20 feature are shown in Figure 2

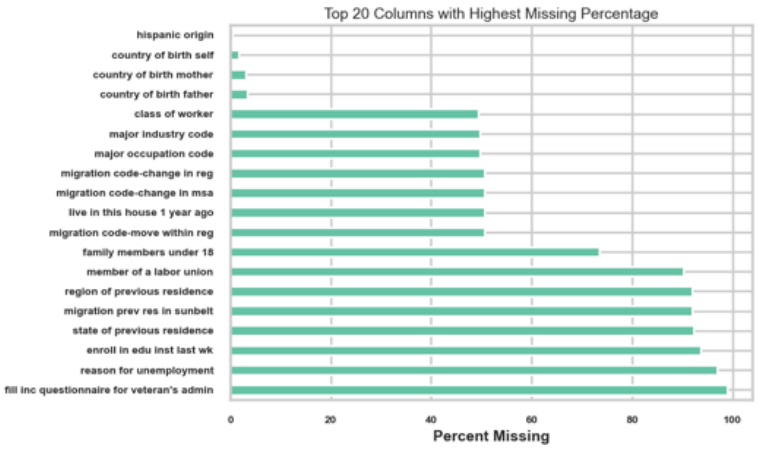

Figure 2: Proportion of missing values by features

- **Feature Insights:**
  - **Demographics:**

    The charts in Figure 3 and Figure 4 reveal clear demographic disparities. Income distribution varies by race and sex, with a noticeable gap between male and female high earners. This underscores the importance of these features in the model but also highlights potential bias that must be monitored.

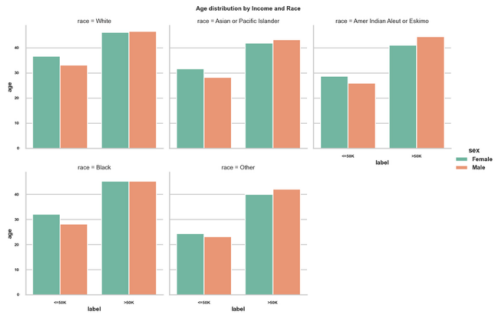

Figure 3: Age Distribution by income class, race, and sex.

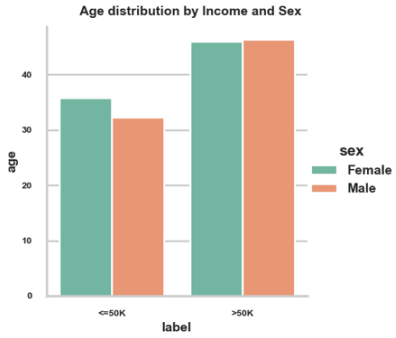

Figure 4: Age Distribution by income class and sex

  - **Financial Indicators:**

    The distribution of weeks worked annually as shown in Figure 5 is a strong indicators of the >50K class, which is consistent with financial common sense.

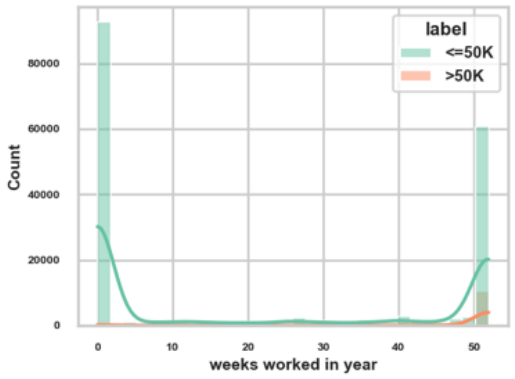

Figure 5: Work week Distribution by income level

Preprocessing Pipeline:

- **Feature Encoding:** Categorical features were transformed using OneHotEncoder.
- **Feature Scaling:** Numerical features were standardized to ensure they contribute equally to the models like K-Means and logistic regression.
- **Data Splitting:** The data was split into 70% training and 30% test sets, using 5-fold cross-validation on the training set for robust model selection and hyperparameter tuning.

4. Classification Model: Predicting High-Income Earners

I trained and compared three classification models: Logistic Regression, Random Forest, and XGBoost. The goal was to find the most accurate and reliable model for predicting the >50K class.

Models: Logistic Regression | RandomForest | XGBoost

- The models were trained and evaluated using GridSearch cross validation on a balanced stratified 5-fold sets of datasets.
- The evaluation metric chosen for each model was ROC-AUC scores. The metric generalizes as a good metric for the classification, because it encompasses other evaluation metrics like accuracy, precision, and recall.
- Hyper-parameter tuning of each model parameters indicated that the optimal parameters of the models demonstrated a high ROC-AUC score of above 94%.
- XGBoost model evaluated higher than the other evaluated models using the ROC-AUC score. The ROC-AUC score is 94.82% as shown in Table 1.

Table 1: Model Metrics comparison

Model	ROC-AUC (Test)	Accuracy	Precision (≥$50K)	Recall (≥$50K)	Selected
Logistic Regression (Baseline)	94.46%	84.84%	28%	89.77%	Baseline
Random Forest	94.25%	94.91%	60.86%	54.02%	—

Gradient Boosting ✓	94.82%	95.09%	62.24	56.47%	★ **BEST**

Table 2: Model Perfoormance Indicators

Model	ROC-AUC Score	Key Observation
XGBoost	**94.82%**	Best performer, robust to class imbalance.
Random Forest	>90%	Strong performance, but slightly lower than XGBoost.
Logistic Regression	>90%	Solid baseline, but lower recall on minority class.

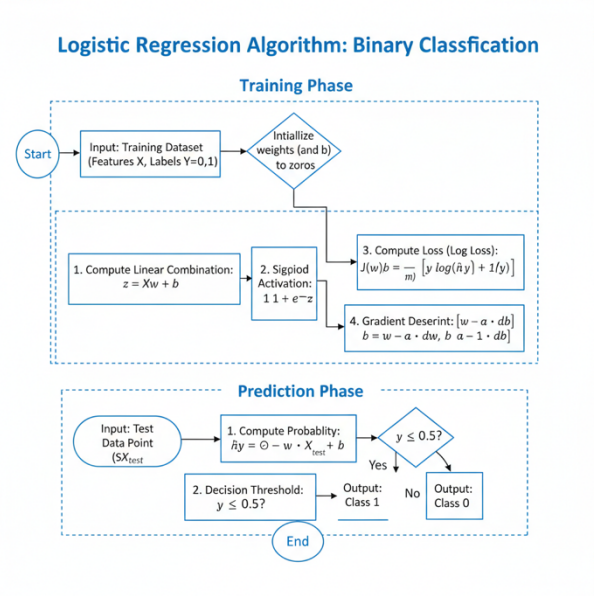

Figure 6: Logistic Regression Model

Model Performance & Evaluation:

The confusion matrices in Figure 7 provide more detailed view of the performance of each model:

- **XGBoost:** Correctly identified **53,901** of the $\leq 50K$ class and **2,007** of the >50K class. It has a low false positive rate (2.16%) but misses a significant portion of actual high earners (false negatives).
- **Random Forest:** Shows a similar pattern, with very high accuracy on the majority class but a slightly higher false positive rate than XGBoost.
- **Logistic Regression:** Exhibited a much higher false negative rate, misclassifying a larger proportion of high-income individuals, making it less suitable for this task.

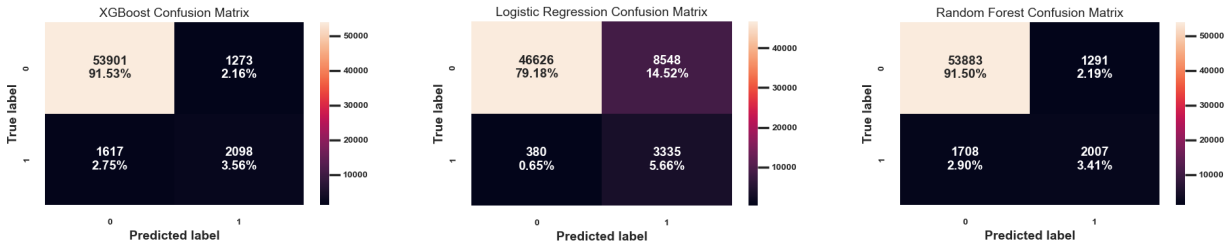

Figure 7: Confusion Matrix

Selected Model: XGBoost

Based on its superior ROC-AUC score and better handling of the minority class, the XGBoost classifier was selected as the final model. Its ability to manage class imbalance and capture complex feature interactions makes it ideal for this business problem.

Feature Importance:

Understanding why the model makes its predictions is crucial for business trust and insight.

- Figure 8 shows that the most important feature by a large margin is weeks worked in year.
- Other highly influential features include:
    - tax filer stat_Nonfiler
    - sex_Male / sex_Female
    - dividends from stocks
    - major industry code_Retail trade
- This tells us that employment stability, filing status, and investment income are the strongest drivers of the model's decision. The inclusion of sex as a top feature confirms the income disparity observed in EDA.

Feature Importance (XGBoost)

Feature	Importance
weeks worked in year	0.10
tax filer stat_Nonfiler	0.07
sex_Male	0.06
dividends from stocks	0.03
sex_Female	0.03
major industry code_Retail trade	0.03
live in this house 1 year ago_No	0.02
major occupation code_Adm support including clerical	0.02
major occupation code_Other service	0.02
detailed household summary in household_Householder	0.01

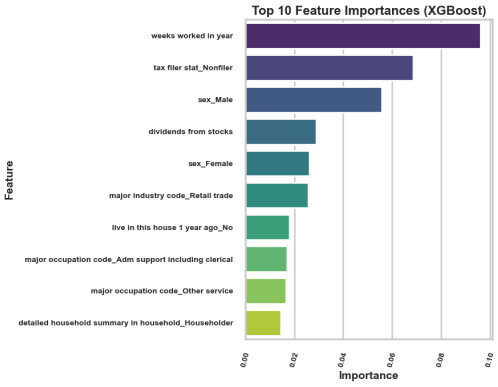

Figure 8: Feature Importance

5. Segmentation Model: Defining Customer Personas

To understand the underlying structure of the customer base, I applied three different clustering algorithms to the mixed numerical and categorical data: K-Means (with PCA on one-hot encoded data), K-Prototypes, and Gower + K-Medoids.

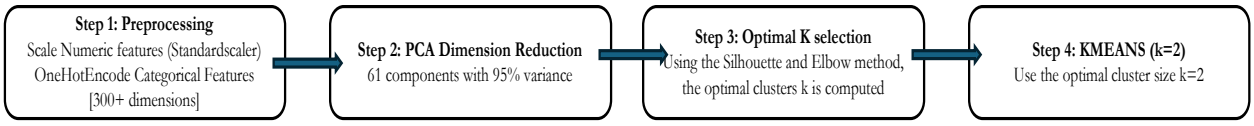

Figure 9: Kmeans Clustering Algorithm

Model Selection:

The models were evaluated using the Silhouette Score, which measures how well-separated and cohesive the clusters are.

- As shown in Figure 10, K-Means with k=2 achieved the highest silhouette score of 0.29. While this indicates moderate cluster separation, it was the best-performing model among those tested.

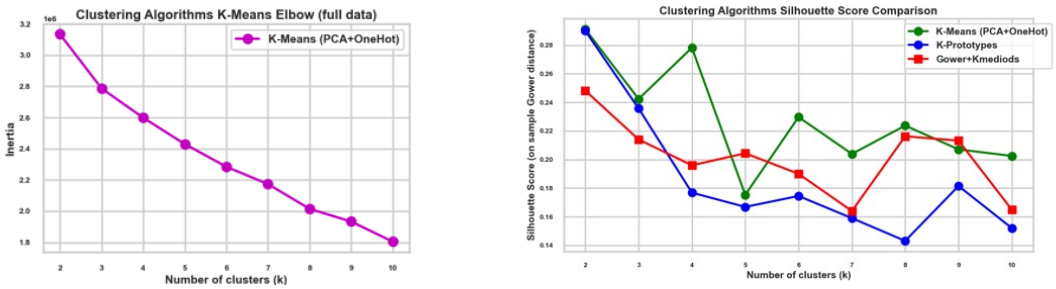

Figure 10: Kmeans Elbow Curve and Silhoette Score Comparison.

The Two Customer Segments:

Based on the analysis of the clusters (pca_clusters_best.png, age_distribution_by_persona.png, num persons worked for employer_distribution_by_persona.png), we have defined the two personas as follows:

Table 3: Cluster Nomenclatures

Feature	Cluster 0: Working Class / Mainstream	Cluster 1: Young Professionals
Average Age	~30 years	~38 years
Avg. Weeks Worked/Year	~1 week	~45 weeks
Avg. Wage per Hour	~$1/hr (likely underemployment)	~$109/hr
Income Profile	Predominantly <=50K	Mixed, but includes the majority of >50K
Key Characteristics	Lower income, unstable or no employment, younger	High earners, stable, full-time employment

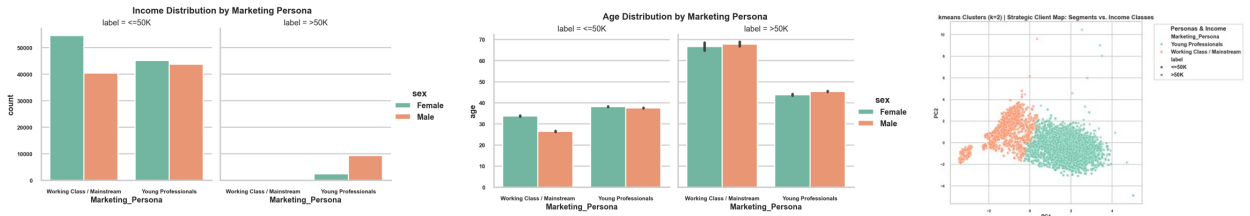

Figure 11: Customer Segmentation and Distribution

Business Application:

The segmentation provides a clear path for targeted marketing:

1. **Target "Young Professionals" (Cluster 1) for Premium Products:** This group has high disposable income and stable jobs. They are the ideal audience for premium services, investment products, and luxury goods.
2. **Target "Working Class/Mainstream" (Cluster 0) for Value & Impact Programs:** This group is more price sensitive. Marketing should focus on value, essential services, and financial literacy or impact programs that help build long-term stability and wealth.

6. Business Recommendations & Deployment Strategy

Based on the analysis conducted, the following actions are recommended:

1. **Deploy the XGBoost Model for Prospect Targeting:** Integrate the model into your marketing pipeline to score new prospects. Prioritize outreach to individuals predicted to be in the >50K class, because they represent the highest potential value for premium offerings.
2. **Apply Segmentation for Tailored Campaigns:**
    - **Cluster 1 (Young Professionals):** Design A/B tests for premium product campaigns. Emphasize more on quality, time saving, and wealth building.
    - **Cluster 0 (Working Class/Mainstream):** Design A/B tests for this group that focuses on the value of the products and financial wellness programs. Messaging context should focus on affordability, reliability, and long-term growth.
3. **Address Income Disparity in Impact Programs:** The model's reliance on features like sex and race highlights existing income disparities. The model should be used to identify and measure the participation of underrepresented groups in the "building wealth and legacies" programs to ensure equitable impact.
4. **Establish a Monitoring Framework:** Continuously monitor the model for data drift and model drift. Retrain the model periodically with new census data to maintain its accuracy and relevance as economic conditions change.

7. Limitations & Future Improvements

Limitations:

- **Data Age:** The dataset is from 1994-1995 and may not accurately reflect current income distributions, job markets, or demographic trends.
- **Moderate Cluster Separation:** A silhouette score of 0.29 suggests the clusters are not perfectly distinct. Further refinement could lead to more defined personas.
- **Model Bias:** The model has learned societal biases present in the data (e.g., gender income gap). It should be used responsibly and subjected to fairness audits.

Future Improvements:

- **Data Refresh:** Retrain the model with more recent Census Bureau data to improve its real-world applicability.

- **Threshold Optimization:** Work with your marketing team to define the cost of a false positive (marketing spend on a low-value prospect) vs. a false negative (missed opportunity). This threshold should be used to optimize the model's decision threshold.
- **Explore Advanced Techniques:**
    - Test non-linear dimensionality reduction techniques like UMAP for potentially better cluster separation.
    - Evaluate models using Precision-Recall AUC (PR-AUC), which is often more informative for imbalanced datasets than ROC-AUC.
    - Conduct fairness and stress testing to quantify and mitigate model bias across different demographic groups.

8. Conclusion

The project developed a robust machine learning framework to address the core business needs proposed. The XGBoost classifier provides a powerful tool for identifying high-income prospects, while the K-Means segmentation model offers a clear, actionable view of the customer base into "Working Class/Mainstream" and "Young Professional" personas.

By deploying these models, it can significantly enhance marketing precision, improve campaign ROI, and help to develop more empathetic and effective customer outreach strategies. As a result, I recommend proceeding with the deployment plan and initiating the process for a data refresh to ensure the models' long-term value.

Reference

- Huang, Z.: Clustering large data sets with mixed numeric and categorical values, Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore, pp. 21-34, 1997. [HUANG98]
- Scikit-Learn Kmeans clustering documentation: https://scikit-learn.org/stable/modules/clustering.html#k-means
- Scikit-learn supervised learning documentation: https://scikit-learn.org/stable/supervised_learning.html
- Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y, Yuan J, Cortes D (2026). *xgboost: Extreme Gradient Boosting*. R package version 3.3.0.0, https://github.com/dmlc/xgboost.
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly.