Income Classification & Customer Segmentation

A Machine Learning Approach to Retail Marketing Targeting

Edison E Chukwuemeka

1

199,523	40	95.23%	3
Census Records	Feature Variables	Best ROC-AUC	Market Segments

Machine Learning Team | February 2026 |

Agenda

1. Executive Summary
Key Findings and Business Takeaways

2. Business Problem Overview
- Objective, Dataset and Project Scope

3. Data Exploration (EDA) & Preprocessing
- Feature Analysis,
- Distributions and Insights
- Missing Values and Class Imbalance
- Categorical Encoding

4. Classification Model
- **Training**
- **Evaluation**
- **Model Comparison**
- **Feature Importance**

5. Segmentation Model
- **PCA**
- **Model Evaluation**
- **Model Comparison**
- **Optimal K selection**

6. Business Recommendations
- **Model deployment for marketing**

7. References
- **Data Sources and technical references**

Executive Summary

199,523	6.21 %	94.81%	2
Total Records	Earn > $50K (Minority Class)	Best Model ROC-AUC	Distinct Marketing Segments

Classification Model Summary

- All models exceed 90% ROC-AUC
- Selected Model: Gradient Boosting Classifier (XGBoost)
  - Efficient with class imbalance
  - ROC-AUC: 94.81%
  - Class imbalance (15:1) handled with oversampling technique (SMOTE)

Segmentation Model Summary

- Clusters Identified: 2 cluster group
  - Silhouette Score: 0.2924
  - Clustering Model: KMEANS
- Cluster 0: Working Class/Mainstream
- Cluster 1: Young Professional

Business Problem Overview

Problem Statement:

A retail client needs to identify two groups for targeted marketing campaign

- People earning $\leq 50K/year$
- People earning $> 50K/year$

The client has access to 40 demographic and employment features datasets from the 1994-1995 census bureau.

Deliverables

- **Classification Model**
  - Predict the income class of any individual given demographic and employment features
- **Segmentation Model**
  - Develop a machine learning pipeline to group the customer population.

Exploratory Data Analysis and Preprocessing

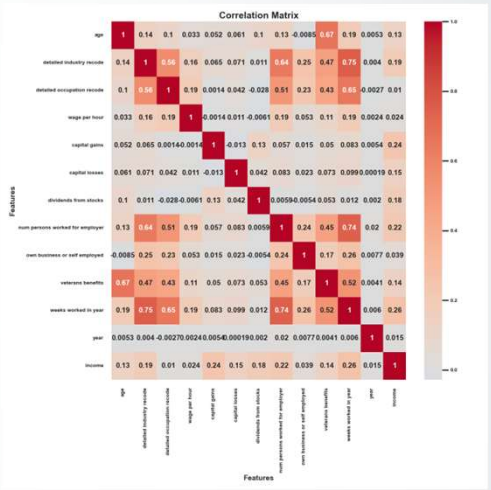
Data Preprocessing

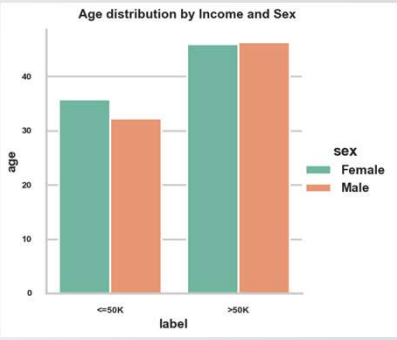
Missing Data:
- '?' and 'Not in Universe' are considered to be missing values
- Numerical features were imputed with median values
- Categorical features were imputed with most_frequent data (mode)
- Dropped columns with missing values > 70%

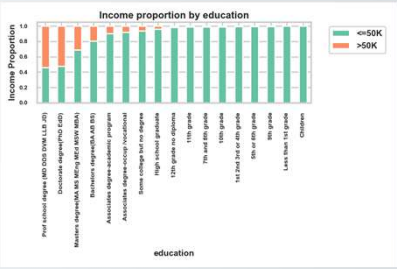
Feature Encoding:
- Numeric features were standardized
- Categorical features were encoded with the OneHotEncoder

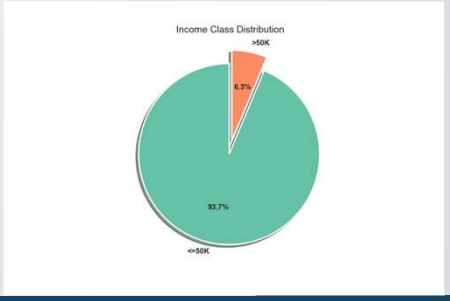
Class Imbalanced:
- The dataset was imbalanced. Only 6% of the target class was available in the dataset.
- The imbalanced class was stratified using SMOTE algorithm during model training.

Train/Test Split:
- The dataset was split in 70% Training and 30% Test
- Cross validation fold= 5 was applied during model training.









Class Imbalance:
Only 6.21% earn ≥$50K — 15:1 ratio. Accuracy alone is misleading; ROC-AUC and balanced weighting are essential.

Impact of Education:
Professional/PhD degree holders earn ≥$50K at 50%+. High school grads under 5%. Strongest categorical predictor.

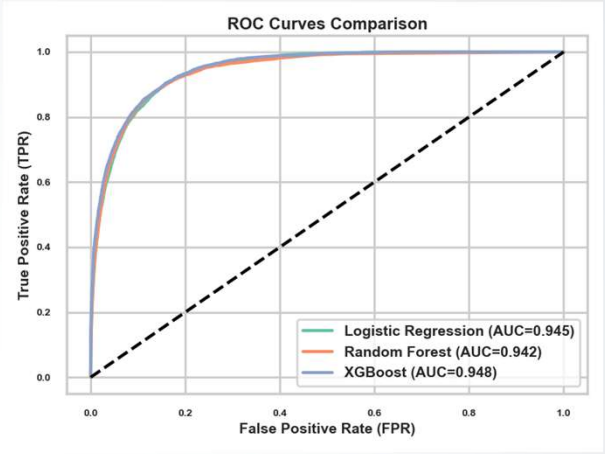
Gender Imbalance:
Males earn ≥$50K at 10.1% vs females at 2.5%, reflecting 1994-95 labor market inequalities.

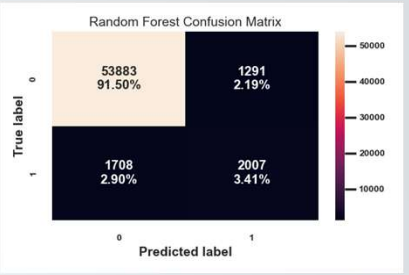
Work Commitment
High earners work nearly 52 weeks/year. Low earners are dispersed across part-time and seasonal work.

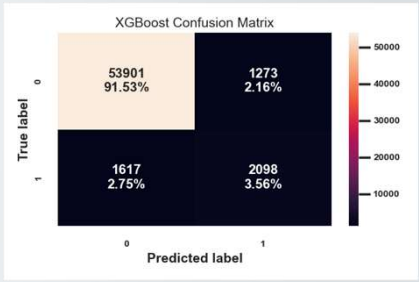
Classification Model: Training and Evaluation

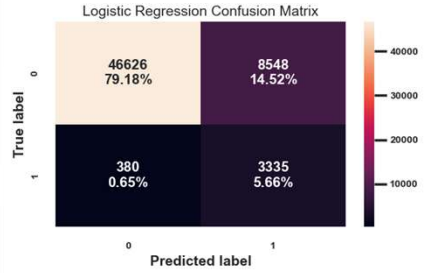
Models: Logistic Regression | RandomForest | XGBoost

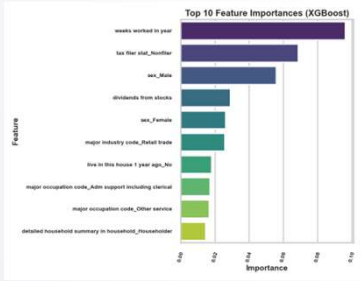
- Models were trained and evaluated using a balanced stratified 5-fold sets of dataset.
- Evaluation metric was ROC-AUC scores. The metric generalizes as a good metric for classification because it encompasses other metrics accuracy, precision, recall.
- Hyper-parameter tuning of the model parameters indicated that the optimal parameters of the models demonstrated a high ROC-AUC score of above 94%.
- XGBoost model evaluated higher than other models using the ROC-AUC score. The ROC-AUC score is 94.82%

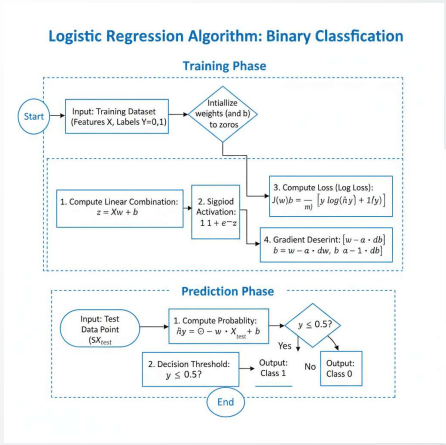












Model	ROC-AUC (Test)	Accuracy	Precision (≥$50K)	Recall (≥$50K)	Selected
Logistic Regression (Baseline)	94.46%	84.84%	28%	89.77%	Baseline
Random Forest	94.25%	94.91%	60.86%	54.02%	—
Gradient Boosting ✓	94.82%	95.09%	62.24	56.47%	★ BEST

Segmentation Model: Clustering

KMEANS CLUSTERING ALGORITHM

Step 1: Preprocessing
Scale Numeric features (Standardscaler)
OneHotEncode Categorical Features
[300+ dimensions]

Step 2: PCA Dimension Reduction
61 components with 95% variance

Step 3: Optimal K selection
Using the Silhouette and Elbow method, the optimal clusters k is computed

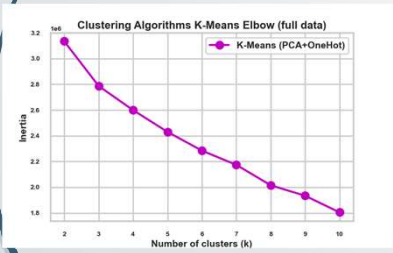
Step 4: KMEANS (k=2)
Use the optimal cluster size k=2

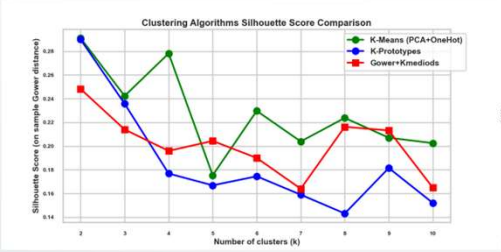
Models: kmeans | kprototypes | kmediods + gower

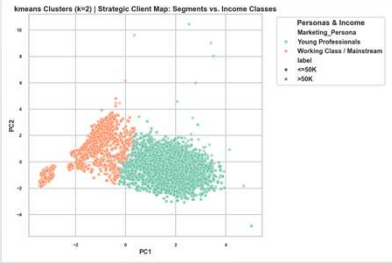
- The datasets is a mixed datasets. Three models were evaluated for the clustering.
- The models were evaluated by the silhouette scores such that the model with the highest silhouette scores is selected.
- KMEANS with cluster size=2, had the maximum Silhouette score of 0.2911.

Result:

- Average Age of Cluster 0: Working class /Mainstream is 30yrs and cluster 1: Young Professional is 38yrs
- Average weeks worked for cluster 0: Working class/Mainstream is 1 week/year and cluster 1: Young Professional is 45weeks/year.
- Average wage per hour for cluster 0: 1hr and cluster 1: $109/hr.







Business Recommendation

Recommendation

- Deploy XGBoost model to predict high earning customers for targeted marketing.
- To improve customer's financial stability and resilience, the model can predict customers who will benefit from the impact program.
- The model determined income disparity in gender and race. It can be used to determine the gender and race of participants in the building wealth and legacies program of the client.
- Prioritize cluster 1: Young Professional customer for targeted marketing.
-

Future Improvements

- Threshold Optimization using key business cost parameters
- Investigate the regulatory compliance concerns of using the ML model.
- Due to lack of recent dataset for training, I would recommend the ML be trained with more current census bureau dataset.
-

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.