

Income Classification & Customer Segmentation

A Machine Learning Approach to Retail Marketing Targeting
Edison E Chukwuemeka

199,523	40	94.81%	2
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
Total Records

6.21 %
Earn > \$50K
(Minority Class)

94.81%
Best Model
ROC-AUC

2
Distinct Marketing Segments

Objective:

- Improve marketing efficiency by utilizing machine learning models to identify the following class:
 - Individual earnings $\leq 50K$ annually
 - Individuals earning $> 50K$ annually
- Segmentation framework that tailor marketing strategies to distinct customer groups

Business Impact:

- Marketing efforts should prioritize high-income prospects predicted by the model.
- Segmentation model differentiated the cluster group that can be impacted by the client's products (premium vs value-focused options)
- Segmentation model identified cluster group that will benefit from impact programs of the client.

Key Results:

- **Classification Model Summary**
 - All models exceed 90% ROC-AUC
 - **Selected Model:** Gradient Boosting Classifier (XGBoost)
 - Efficient with class imbalance
 - Test ROC-AUC: 94.81%
 - Precision:
 - Recall
 - Class imbalance (15:1) handled with oversampling technique (SMOTE)
- **Segmentation Model Summary**
 - Clustering Model: KMEANS
 - Cluster size: k = 2
 - Silhouette Score: 0.29
 - Cluster 0: Working Class/Mainstream
 - Cluster 1: Young Professional

Business Problem Overview

Problem Statement:

The retail client seeks to:

- Improve marketing precision
- Improve campaign ROI
- Reduce outreach cost impact
- Understand the customer structure

This is a binary classification problem with severe class imbalance (~6% high income earners) combined with a unsupervised customer segmentation problem.

Dataset Summary: 199523 weighted census observations with 40 demographics and employment features.

Target Summary:

- People earning $\leq 50K$ annually (~94%)
- People earning $> 50K$ annually (~6%)

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 & Preprocessing

Class Imbalanced:

- The dataset was imbalanced. Only 6% of the target class (high income earners) was available in the dataset.
- Metrics of evaluation: Accuracy is not sufficient for evaluation. ROC-AUC metric is prioritized.
- The imbalanced class was stratified using SMOTE algorithm during model training.

Missing Data:

- '?' and 'Not in Universe' are missing values
- Numerical features were imputed with median values
- Categorical features were imputed with most-frequent data (mode)
- Removed columns with $> 70\%$ missing values
- Preprocessing were applied during cross-validation steps of the training

Feature Encoding:

- Numeric features were standardized
- Categorical features were encoded with the OneHotEncoder

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 $\geq \$50K$ — 15:1 ratio. Accuracy alone is misleading; ROC-AUC and balanced weighting are essential.

Impact of Education:

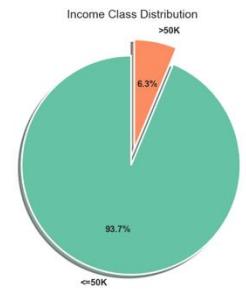
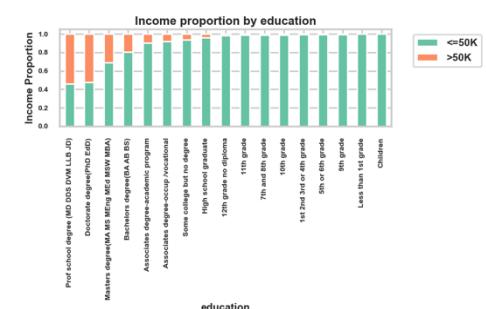
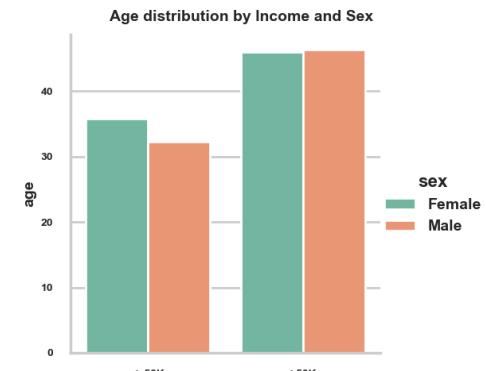
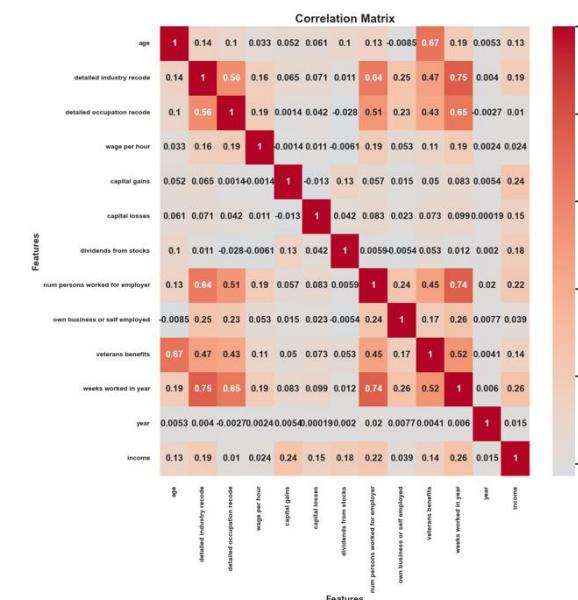
Professional/PhD degree holders earn $\geq \$50K$ at 50%+. High school grads under 5%. Strongest categorical predictor.

Gender Imbalance:

Males earn $\geq \$50K$ at 10.1% vs females at 2.5%, reflecting 1994-95 labor market inequalities.

Work Commitment

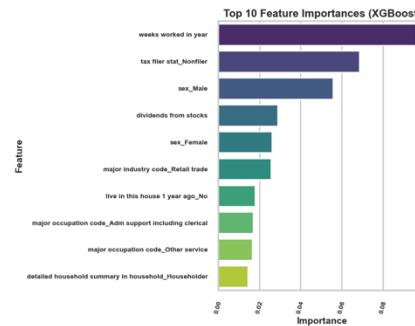
High earners work nearly 52 weeks/year. Low earners are dispersed across part-time and seasonal work.



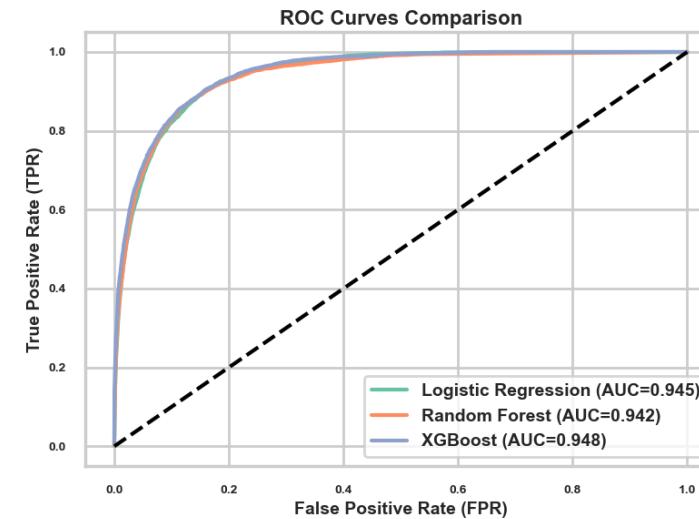
Classification Model: Training and Evaluation

Models: Logistic Regression | RandomForest | XGBoost

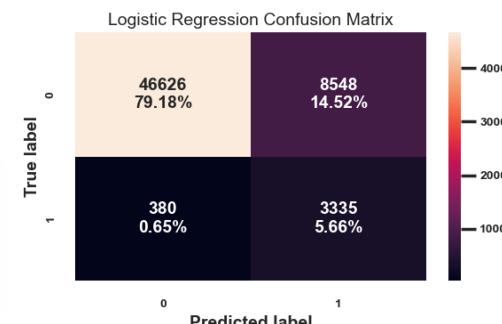
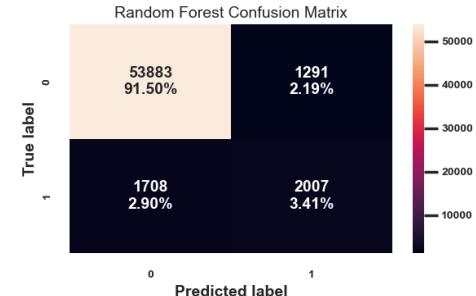
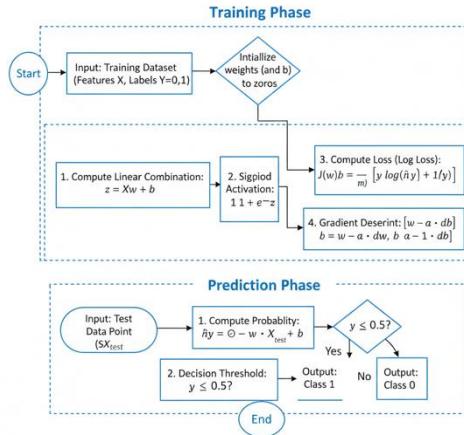
- Models were trained and evaluated using a balanced stratified 5-fold sets of dataset.
- Evaluation metric was ROC-AUC scores. The metric generalizes as a good metric for classification because it encompasses other metrics accuracy, precision, recall.
- Hyper-parameter tuning of the model parameters 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 ($\geq \$50K$)	Recall ($\geq \$50K$)	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

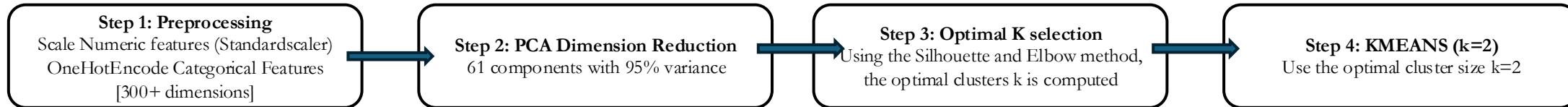


Logistic Regression Algorithm: Binary Classification



Segmentation Model: Clustering

KMEANS CLUSTERING ALGORITHM



Models: kmeans | kprototypes | kmmedioids + gower

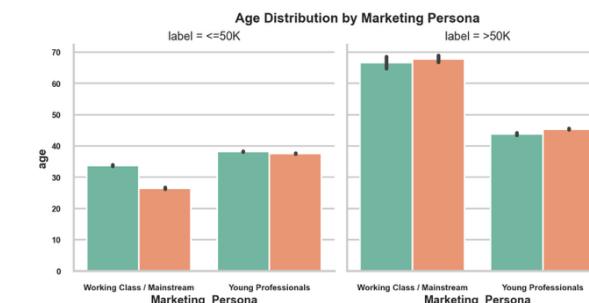
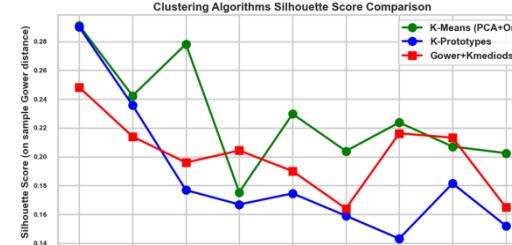
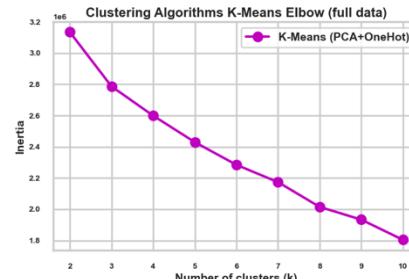
- The datasets is a mixed datasets. Three models were evaluated for the clustering.
- The models were evaluated by the silhouette scores such that the model with the highest silhouette scores is selected.
- KMEANS with cluster size=2, had the maximum Silhouette score of 0.29.

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/Mainsstream 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 Application:

- Premium products marketing should target cluster 1
- Value added product marketing should target cluster 0
- Apply A/B testing by cluster group



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.
- Monitor the model for model and data drift over time.
- Retrain the model with updated census data

Limitations

- Census data (1994 – 1995) may not reflect income structures
- Silhouette scores indicates a moderate cluster separation
- Economic shifts may impact the stability of the model
- Performance of the model depends on accurate inputs.

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.
- Explore cost-sensitive learning
- Evaluate PR-AUC explicitly
- Conduct fairness stress testing
- Explore UMAP for non-linear cluster segmentation.

Conclusion

The machine learning framework identifies:

- Two customer segmentation clusters for target marketing
- The best classification models for the dataset as XGBoost
- Business threshold optimization

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.

THANKS