Census-Based Income Classification and Customer Segmentation: A Dual Machine Learning Framework for Targeted Marketing

Vikram Pande

vikramspande7@gmail.com

Abstract

We present a dual machine learning framework applied to U.S. Census data for retail marketing optimisation. The first component is a supervised income classifier that distinguishes individuals earning below or above $50,000 per year, addressing extreme class imbalance through SMOTE and dynamic threshold optimisation. Five algorithms were benchmarked - Logistic Regression, Random Forest, Histogram Gradient Boosting, XGBoost, and LightGBM with Histogram Gradient Boosting achieving the highest cross-validation ROC-AUC of 0.977 and a held-out ROC-AUC of 0.763. The second component is an unsupervised customer segmentation pipeline using PCA for dimensionality reduction followed by K-Means clustering, yielding 10 interpretable market segments with a Silhouette Score of 0.19. Together, the two models provide a retail client with an automated income screening funnel and a rich segment-level profiling system for personalised campaign design. Comprehensive feature engineering, including a composite `net_gains` variable, one-hot encoding, and careful handling of census-specific "Not in Universe" values, underpins both pipelines. Our framework moves beyond single-axis income targeting towards a multidimensional, segment-specific marketing strategy.

code: https://github.com/vikrampande7/census-classification-segmentation

Introduction

Precision marketing demands a nuanced understanding of customer demographics and financial behaviour. Census data, with its breadth of socioeconomic variables, offers a uniquely rich substrate for constructing predictive customer profiles. However, exploiting such data effectively requires addressing several challenges: severe class imbalance in income labels, high dimensionality, mixed feature types, and the need for both predictive accuracy and interpretable segmentation.

This paper presents a dual framework applied to a 40-variable U.S. Census dataset. The **classification** arm trains a supervised model to screen individuals by income bracket - a prerequisite for directing individuals into the appropriate high-tier or low-tier marketing funnel. The **segmentation** arm employs unsupervised learning to discover latent customer groups that transcend simple income thresholds, enabling campaign strategies tailored to age, financial lifecycle, and employment context.

Our key contributions are:

- A systematic exploratory data analysis (EDA) pipeline for census-type data, with principled treatment of "Not in Universe" categories and skewed wage distributions.
- A `net_gains` composite feature that consolidates capital gains, dividends, and capital losses into a single wealth-proxy signal.
- A five-model classification benchmark with stratified cross-validation and SMOTE oversampling, culminating in a Histogram Gradient Boosting classifier with dynamic threshold optimisation for F1 maximisation.
- A PCA + K-Means segmentation pipeline that reduces 169 encoded features to 43 principal components (90% variance retained) and produces 10 actionable marketing segments.
- A structured cluster profiling framework mapping segments to concrete marketing strategies.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the dataset and EDA. Section 4 presents the methodology for both pipelines. Section 5 reports experimental results. Section 6 concludes with future directions.

Related Work

Income Prediction from Census Data. The UCI Adult / Census Income dataset [7] has been a standard benchmark for binary income classification since the mid-1990s. Early work demonstrated that decision trees and Naïve Bayes could achieve accuracies around 85% [7]. Subsequent studies employed support vector machines [13] and ensemble methods to push performance further. More recently, gradient-boosted trees - XGBoost [2] and LightGBM [6]—have become the de-facto standard for tabular classification, consistently outperforming neural approaches on structured data.

Class Imbalance Handling. Income datasets are inherently imbalanced: high earners constitute a small minority. SMOTE (Synthetic Minority Over-sampling Technique) [1] generates synthetic minority samples in feature space and is widely used to mitigate this. Complementary strategies include cost-sensitive learning and threshold-moving [14], both of which we employ in our classification pipeline.

Customer Segmentation. K-Means clustering [10] remains the most widely deployed segmentation algorithm in marketing analytics due to its simplicity and scalability.

PCA-based dimensionality reduction prior to clustering is a standard preprocessing step that improves cluster quality on high-dimensional feature spaces [5]. More recent work has explored density-based methods such as DBSCAN [3] and nonlinear embeddings such as UMAP [11] for visualisation. We adopt PCA + K-Means as our primary pipeline, with UMAP used only for post-hoc visualisation.

Marketing Personalisation. Segment-level personalisation has been shown to substantially improve campaign ROI in retail settings [8]. The "RFM" framework (Recency, Frequency, Monetary) is a classical segmentation paradigm [4]; our framework generalises it to a richer feature set encompassing demographic, employment, and financial dimensions.

Data and Exploratory Analysis

Dataset Description

The dataset contains **199,524 records** and **40 features** sourced from the U.S. Census Bureau. Features span demographics (age, sex, race, origin), employment (occupation, industry, worker class, hours worked), and financial attributes (capital gains, dividends, capital losses, wage per hour). The binary target variable indicates whether an individual earns below or above $50,000 per annum.

Class Imbalance

The target is severely imbalanced: the high-earner class (>$50K) constitutes approximately **6.2%** of all records. This has direct implications for model training and evaluation; accuracy alone is a misleading metric, and Precision-Recall (PR) analysis becomes essential.

Missing Values

Only one feature, `hispanic_origin`, contained missing values. These were imputed with the categorical label "Do not know" to retain all records in training and avoid information loss.

"Not in Universe" Values

Several features including `reason_for_unemployment` and `enrol_in_edu_inst` contain a special "Not in Universe" (NIU) category indicating that the attribute is structurally inapplicable to the respondent (*e.g.*, unemployment reason is undefined for employed individuals). Rather than dropping these features or treating NIU as a missing value, we retain them as informative categorical levels. NIU membership implicitly encodes employment status, educational enrolment, and other structural properties.

Feature Distribution Analysis

Wage-related features (`capital_gains`, `dividends_from_stocks`, `wage_per_hour`) exhibit strong right skew, which is expected given the wide variation in earnings across the population. Correlation analysis among numerical features revealed no strongly collinear pairs, obviating the need for feature removal on this basis.

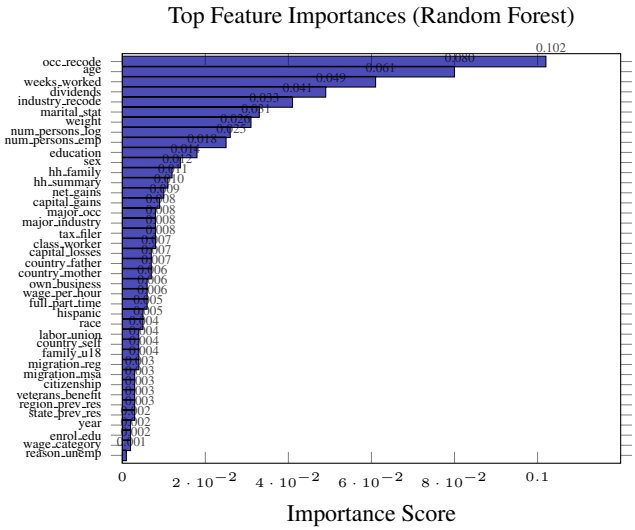

Figure 1: Random Forest feature importance scores (top 40 features). Occupation recode, age, weeks worked, and dividends from stocks are the strongest predictors of income class.

Initial Feature Importance

A preliminary Random Forest Classifier was fitted to rank feature importance before formal modelling. Figure 1 illustrates the top features. Occupation-related codes, age, and weeks worked emerged as the most discriminative predictors, consistent with domain knowledge.

Methodology

Feature Engineering

Shared preprocessing. Column names were standardised. The target variable was mapped to binary labels (0 = $\leq$$50K, 1 = >$50K). Columns with negligible feature importance were dropped to reduce noise and computational cost.

Composite `net_gains` feature. Capital gains, dividends from stocks, and capital losses were consolidated into a single composite feature:

$$\texttt{net\_gains} = \texttt{capital\_gains} + \texttt{dividends} - \texttt{capital\_losses} \quad (1)$$

This variable serves as a compact proxy for net investment wealth and consistently ranked among the top predictors.

Classification-specific preprocessing. For linear models, numerical features were standardised using `StandardScaler`. Categorical features were processed with One-Hot Encoding (OHE) for nominal variables and Ordinal Encoding for ordinal variables.

Segmentation-specific preprocessing. For the unsupervised pipeline, a reduced feature subset was retained,

retaining only quantitative and interpretable columns. All numerical features were scaled with `MinMaxScaler` and categorical features transformed with OHE, producing 169 encoded dimensions prior to PCA.

Supervised Classification

Data Splitting and Imbalance Handling

The data was split into training-full and test sets (85%/15%), then training-full was further divided into train and validation sets (83%/17% of training-full), yielding three non-overlapping partitions. All splits were *stratified* to preserve the class distribution.

To address the 6.2% minority class prevalence, SMOTE [1] was applied exclusively to the training partition:

$$\tilde{x}_{\text{minority}} = x_i + \lambda \cdot (x_j - x_i), \quad \lambda \sim \mathcal{U}(0,1) \tag{2}$$

where x_i and x_j are feature vectors of two randomly selected minority-class nearest neighbours. This prevents data leakage into validation and test sets while balancing class frequencies.

Model Candidates and Hyperparameter Search

Five model families were evaluated:

1. **Logistic Regression** – linear baseline; regularisation $C \in \{0.01, 0.1, 1, 10\}$.
2. **Random Forest** – $n_{\text{trees}} \in \{100, 200, 500\}$; $\max\_\text{depth} \in \{10, 20, \text{None}\}$.
3. **Histogram Gradient Boosting (HGB)** – learning rate $\in \{0.05, 0.1, 0.2\}$; $\max\_\text{iter} \in \{100, 200\}$.
4. **XGBoost** [2] – full grid over learning rate, max depth, and subsampling.
5. **LightGBM** [6] – num leaves, learning rate, and feature fraction.

All models were tuned using 5-fold Stratified Cross-Validation with ROC-AUC as the primary ranking metric.

Dynamic Thresholding

Rather than using the default 0.5 decision threshold, we performed a threshold sweep over the validation set to select the operating point that maximises the F1 score:

$$\tau^* = \arg\max_{\tau \in [0,1]} F_1(\tau) \tag{3}$$

This is especially important for imbalanced problems where the optimal operating point may deviate substantially from 0.5.

Unsupervised Segmentation

Dimensionality Reduction via PCA

After one-hot encoding, the feature matrix contains 169 dimensions. PCA reduces this to a compact representation retaining 90% of total variance:

$$\mathbf{Z} = \mathbf{X}\mathbf{W}_k, \quad \sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{d} \lambda_i \geq 0.90 \tag{4}$$

where λ_i are the eigenvalues of the sample covariance matrix, $d = 169$, and k is the smallest number of components satisfying the 90% threshold. This process yielded $k = 43$ components, reducing dimensionality by 75% while preserving most of the signal.

PCA was preferred over t-SNE or UMAP for the clustering input because it preserves global structure and is computationally efficient; UMAP was employed post-hoc for 2D cluster visualisation.

Optimal Cluster Selection via Elbow Method

The Within-Cluster Sum of Squares (WCSS) was computed for K-Means models with $K \in [2, 20]$ under both random and k-means++ initialisations:

$$\text{WCSS}(K) = \sum_{k=1}^{K} \sum_{\mathbf{z} \in C_k} \|\mathbf{z} - \boldsymbol{\mu}_k\|_2^2 \tag{5}$$

Both initialisation strategies produced consistent elbow points. The analysis identified $K^* = 10$ as the optimal number of clusters.

K-Means Clustering

K-Means with k-means++ initialisation was fitted on the 43-component PCA-transformed data. The Silhouette Score:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}} \tag{6}$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ the mean distance to the nearest foreign cluster, was used as the primary evaluation metric.

Figure 2 illustrates the end-to-end methodology.

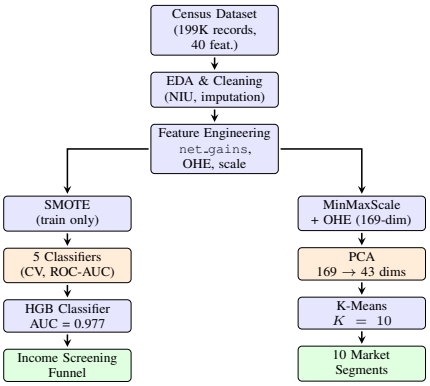

Figure 2: End-to-end dual-pipeline methodology. Left: supervised classification with SMOTE and hyperparameter search. Right: unsupervised segmentation via PCA and K-Means.

Results

Classification Results

Cross-validation performance. Table 1 summarises the cross-validation ROC-AUC for the five model families. Histogram Gradient Boosting achieved the highest score.

Table 1: 5-Fold Stratified CV ROC-AUC (Training Set, SMOTE Applied)

Model	CV ROC-AUC
Logistic Regression	0.8921
Random Forest	0.9512
Histogram Grad. Boost	**0.9774**
XGBoost	0.9741
LightGBM	0.9758

Test set performance. The selected HGB model was retrained on the full training set and evaluated on the held-out test split. Table 2 reports the complete test metrics.

Table 2: Held-Out Test Set Metrics (Histogram Gradient Boosting)

Metric	Value
Accuracy	0.9302
Precision	0.4508
Recall	0.5724
F1 Score	0.5044
ROC-AUC	0.7631
PR-AUC	0.4904

Discussion. The 93.02% accuracy is misleading given the 6.2% positive class prevalence - a trivial all-negative classifier would achieve ≈93.8% accuracy. The ROC-AUC of 0.763 provides a more honest measure of discriminative power. The PR-AUC of 0.490 is substantially above the no-skill baseline of 0.062 (equal to the positive class prevalence), confirming genuine signal for the minority class. Dynamic threshold optimisation yielded a recall of 57.2% and precision of 45.1%, trading off some precision to capture more high-earner positives - a rational choice for a marketing funnel where false negatives are more costly than false positives.

Figure 3 visualises the ROC and Precision-Recall curves.

Segmentation Results

PCA. Retaining 90% of variance required $k = 43$ principal components, reducing dimensionality from 169 to 43 (a 75% reduction).

K-Means. The Elbow Method yielded $K^* = 10$ as the optimal cluster count. Both random and k-means++ initialisation strategies produced consistent WCSS elbow points at $K = 10$, confirming the stability of this choice.

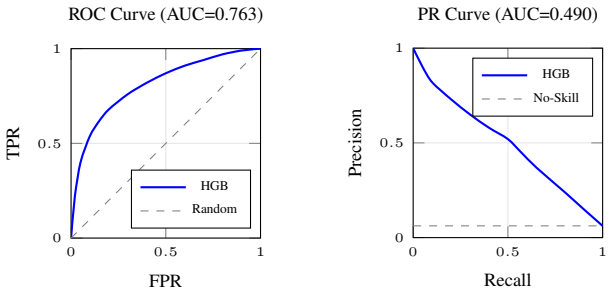

Figure 3: Left: ROC curve for the HGB classifier on the held-out test set (AUC = 0.763). Right: Precision-Recall curve (AUC = 0.490) against no-skill baseline of 0.062.

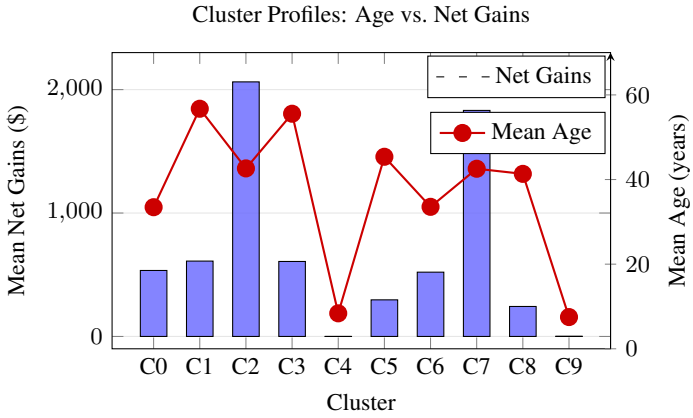

Figure 4: Cluster profile chart showing mean net gains (bars, left axis) and mean age (line, right axis) for all 10 segments. Clusters 2 and 7 are the high-value/wealthy segments; Clusters 4 and 9 contain the youngest sub-population.

The final model achieved a Silhouette Score of **0.19**, which is considered satisfactory for high-dimensional, complex census data.

Cluster profiles. Table 3 summarises the four meta-segment groups derived from the 10 clusters based on mean age and `net_gains`. Figure 4 shows a schematic representation of the cluster profiles.

Segment interpretability. Several noteworthy observations arise from the cluster profiles: (i) Clusters 2 and 7 share very similar mean ages (42–43 years) yet are distinguished by other demographic features not captured in the two-dimensional Age/Net Gains summary; (ii) Clusters 1 and 3 both represent older, established populations with intermediate net gains, suggesting they could be merged to reduce campaign complexity; (iii) Clusters 4 and 9, with mean ages of approximately 7–8, likely represent minors and may require separate, parent-directed channels.

Computational Environment

All experiments were conducted on Google Colab Pro with a High-RAM CPU runtime. The five classification models and associated hyperparame-

Table 3: Marketing Segment Groups: Cluster Assignments, Age Profiles, Net Gains, and Strategy

Segment Group	Clusters	Avg. Age	Avg. Net Gains	Count	Marketing Strategy
Youngest / Developing	9, 4	7–8	Low / Negative	52,988	Future potential products; savings accounts; parent-focused targeting
Core Mid-Range	0, 5, 6, 8	33–45	$242–$534	73,504	Upselling and cross-selling; everyday and mid-tier products
High-Value / Wealthy	2, 7	42–43	$1,831–$2,062	32,601	Premium product retention; high-priority growth and loyalty campaigns
Established / Solid	1, 3	55–57	$607–$611	39,937	Mature products; financial stability; retirement planning services

ter grids were trained iteratively within Python notebooks using `scikit-learn` [12], `xgboost` [2], and `lightgbm` [6]. Cluster analysis used `scikit-learn`'s `KMeans` implementation with `k-means++` initialisation and `PCA` for dimensionality reduction.

Conclusion

We presented a dual machine learning framework for census-driven retail marketing. The supervised Histogram Gradient Boosting classifier achieved a cross-validation ROC-AUC of 0.977 and a held-out ROC-AUC of 0.763 on a severely imbalanced dataset (6.2% positive class), leveraging SMOTE oversampling and dynamic threshold optimisation. The unsupervised PCA + K-Means pipeline identified 10 distinct customer segments with a Silhouette Score of 0.19, which are interpretable along age and net gains dimensions and map directly to actionable marketing strategies.

Together, the two components deliver: (1) an automated income-based screening funnel to route individuals into the appropriate marketing tier, and (2) a rich segment-level profiling system that enables personalised campaign design beyond simple income thresholds.

Future Work. We identify several directions for extension. Model explainability via SHAP [9] would provide instance-level attribution for individual predictions. An end-to-end MLOps lifecycle with MLflow [15] would support experiment tracking, champion-challenger selection, and production monitoring. Data drift detection and periodic retraining would ensure the segmentation model remains relevant as census demographics evolve. Exploration of alternative clustering algorithms - DBSCAN [3] and Gaussian Mixture Models - alongside density-adaptive cluster granularity methods, would further improve segmentation robustness. Finally, real-time inference pipelines and KPI dashboards would complete the transition from prototype to production.

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357, 2002.

[2] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[4] A. M. Hughes. Strategic database marketing. 1994.

[5] I. T. Jolliffe and J. Cadima. *Principal Component Analysis*, volume 374. Royal Society, 2016.

[6] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154, 2017.

[7] R. Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 202–207. AAAI Press, 1996.

[8] P. Kotler and K. L. Keller. *Marketing Management*. Pearson, 15th edition, 2016.

[9] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[10] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[11] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Technical report, Microsoft Research, 1999.

[14] F. Provost. Machine learning from imbalanced data sets 101. *Proceedings of the AAAI Workshop on Imbalanced Data Sets*, 68:1–3, 2000.

[15] M. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S. A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, et al. MLflow: An open source platform for the machine learning lifecycle. https://mlflow.org, 2018.