Income Classification & Customer Segmentation: Predictive Modeling Report for Retail Marketing

02.19.2026

Amrit Mahajan

1. Summary

This project uses the 1994–1995 U.S. Census Bureau dataset with weighted observations sampled from ~350k population and 40 variables to build two marketing models. An income classifier predicts whether an individual earns above or below $50,000, using LightGBM selected after comparing five algorithms with Bayesian hyperparameter tuning, achieving a ROC-AUC of 0.94. A segmentation model identifies six customer personas using K-Means clustering to support differentiated targeting. Both models account for population weights, class imbalance, and fairness across gender, with SHAP-based explainability for business interpretation.

2. Data Understanding (EDA)

EDA was structured in 3 progressive layers - Univariate, Bivariate, and Multivariate. All analyses are driven by population weight rather than row counts to reflect true demographic proportions

2.1 Univariate Analysis

The first step was profiling each variable independently. For the target variable, weighted bar charts confirmed the 93.6% vs 6.4% income split. For categorical variables, weighted horizontal bar charts were generated for each column to understand the population distribution of each category - for example, identifying that the majority of the weighted population falls under Private sector workers in class of worker, or that most individuals are Not in universe for migration-related variables, flagging them as low-signal features early. For continuous variables, bar plots and boxplots were built to understand spread, skew, and outlier extent. Capital gains, capital losses, and dividends from stocks showed extreme right skew with the vast majority at zero, which directly informed the outlier capping strategy applied during preprocessing.

2.2 Bivariate Analysis

Each variable was then examined against the income label to assess predictive signal. For categorical variables, weighted count plots split by income class revealed clear separations in features like marital status, tax filer status, class of worker, and education. For continuous variables, weighted violin plots showed that higher-income individuals concentrated at higher ages, more weeks worked, and significantly higher capital gains and dividends. A geographic analysis was also conducted, mapping weighted population counts by state of previous residence using a Plotly choropleth to identify regional migration patterns. Cross-tabulations were explored for semantically related variable pairs such as race and hispanic origin, and reason for unemployment and class of worker, to detect redundancy and understand interaction effects before feature selection.

2.3 Multivariate Analysis

A Pearson correlation heatmap on all seven continuous variables plus the binary income target confirmed low inter-feature correlation, with weeks worked and age showing the strongest positive relationship with income. For categorical variables, one-hot encoding was applied across all categorical columns and a full correlation analysis against the income target was run, identifying the top 20 most predictive features by absolute correlation.

3. Data Preprocessing

All Feature Engineering was validated through a baseline Random Forest model before selecting the best method. Due to imbalance in the dataset, weighted ROC-AUC was considered to be the evaluation metric to justify separation of high and low income labels.

This ensured that decisions on imputation, encoding, and transformation were grounded in actual model performance rather than convention.

3.1 Data Cleaning/ Handling Missing Values (“?”)

Missing values appeared as "?" strings across only categorical columns in 104,393 rows (~ 35%) of the dataset. Dropping rows was ruled out immediately given the scale of missingness.

Four strategies were tested:

1. Mode imputation, 2. KNN imputation, 3. random forest imputation 4. **introducing a new Unknown category.**

KNN imputation was noted as theoretically unsuitable due to high dimensionality. Introducing Unknown as a distinct category outperformed all alternatives on the held-out

weighted ROC-AUC and was selected. This approach is well-suited to tree-based models, which can learn that missingness itself.

4. Feature Engineering and Selection

Feature selection was treated as a formal experiment with two competing approaches evaluated under the same conditions : the same train-test split, same encoding, and same baseline Random Forest to ensure a fair comparison.

The education column was ordinally ranked (1-16) based on the level of degree/education and was treated as a continuous variable.

Current **Baseline is at AUC: 0.9363** after imputing missing values with an extra category.

4.1 Approach: Weight of Evidence (WOE)

WoE and IV were computed across all variables on the training set. Continuous variables were binned into deciles before WoE calculation, while categorical variables were assessed directly at the category level. Standard rules were applied: bins with fewer than 5% of observations were flagged, zero-event or zero-non-event bins were handled to avoid division errors, and monotonicity of WoE across bins was checked.

Features with an IV above 0.02 were considered as Important Features for further variable selection.

WoE-transformed features were also tested as inputs to both Logistic Regression (AUC: 0.9113) and Random Forest (AUC: 0.8804), replacing the original encoded values with their WoE scores.

This transformation produced lower weighted ROC-AUC than the baseline encoding and thereby discarded

4.2 Feature Selection

Two approaches used -

1. WOE important features with Information Value over 0.02 (AUC: 0.9285)
2. **Feature Importance** functionality by Random Forest (AUC: **0.9362**)

Top 25 Features from Random forest model gave better AUC score and was selected.

4.3 Encoding

Categorical Variables -

- Variables with less than 10 -15 categories were One hot encoded

- Variables with high cardinality were Label Encoded based on the ranking of percentage high income people in the respective group

Numerical/Continuous Variables -

- Variables were outlier-capped using an upper bound of Q3 + 2 x IQR, applied only where the cap value was non-zero.
- A VIF check was performed post-encoding to assess multicollinearity

4.4 Approach: PCA Principal Component Analysis

- Tested as an additional dimensionality reduction step.
- StandardScaler was applied followed by PCA retaining 95% of variance, but the resulting ROC-AUC (0.9144) was lower than the baseline without PCA. PCA was therefore dropped for the classification pipeline.

4.5 Sampling

15:1 Imbalance present

Two Sampling techniques used and tested on hold out set

- Under Sampling (via Stratified Sampling)
    - Undersampled to equal (1:1) proportion of minority and majority class
    - Undersampled to (1:7) proportion of minority and majority class
- Over Sampling (via SMOTE) -
    - Oversampled minority class to (1:7) proportion of minority and majority class

Under Sampling to 1:7 proportion resulted in best AUC (**0.9405**) and was selected as the final strategy, striking a better balance between exposing the model to enough majority class variation while correcting for imbalance.

5. Classification Model Experiments

5.1 Baseline Models tried

Five models were trained on the 1:7 undersampled set with population weights applied: Logistic Regression, Random Forest, SGD Classifier, CatBoost, XGBoost, and **LightGBM**

Logistic Regression and SGD were trained on StandardScaler-normalized features. CatBoost was passed raw categorical indices directly without OHE. All tree-based models used the weighted class imbalance ratio via scale_pos_weight.

LightGBM achieved the highest ROC-AUC (**0.9522**), a decent F1 score (0.5) and high recall (0.89) for high income class with LightGBM selected as the final model.

5.2 Hyperparameter Tuning on LightGBM

- Grid Search with Cross Validation (AUC:0.9528)
- Bayesian Optimization via HyperOpt (AUC:0.9534)

5.3 Final Model

```
LGBMClassifier
LGBMClassifier(colsample_bytree=0.5187618128294202,
               learning_rate=0.023098069259371082, max_depth=7,
               min_child_samples=np.int64(75),
               min_split_gain=0.17421591065929348, n_estimators=np.int64(810),
               num_leaves=np.int64(45), random_state=141,
               reg_alpha=0.06067184360325228, reg_lambda=0.4253819203894115,
               scale_pos_weight=np.float64(6.802484033819776),
               subsample=0.5653125958019853, verbose=-1)
```

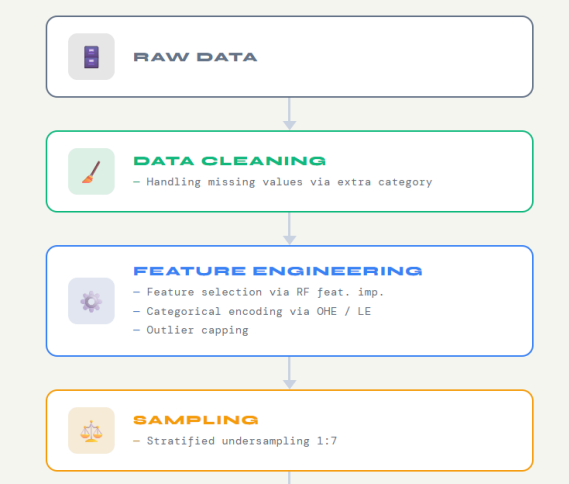
5.4 Fairness Analysis & Explainable AI

Fairness was evaluated using Fairlearn across two sensitive attributes: sex and race to ensure discrimination free classification.

SHAP values were computed on the first 5,000 samples of the held-out test set for better explainability of the blackbox LightGBM model.

The global summary plot reveals not just which features matter most but how their values interact with the prediction: for example, high Age/education pushes predictions strongly toward the positive class while low values cluster near zero contribution.

Model Development Pipeline



6. Segmentation Model

6.1 Approach

6.1.1 Preprocessing and Feature Construction

The same cleaning and encoding pipeline from the classification task was applied- Unknown category imputation, OHE for low-cardinality categoricals, and target encoding for high-cardinality columns

Four derived features were engineered on top. All features were then standardized using StandardScaler before dimensionality reduction.

6.1.2 Dimensionality Reduction

PCA was applied retaining 85% of explained variance, reducing the high-dimensional OHE feature space to a compact set of principal components. A cumulative explained variance plot was used to confirm the 85% threshold. The PCA-reduced matrix was used as input to all clustering algorithms, improving both computational efficiency and cluster separation by removing noise dimensions.

6.1.3 Algorithm

Two algorithms were evaluated across the number of clusters (k) from 2 to 11.

- K-Means was assessed using the elbow method on inertia and silhouette score, with population weights in consideration.
- Gaussian Mixture Models were evaluated using BIC, AIC, and silhouette score across the same values of k

K-Means with k=6 was selected based on a combination of elbow curve inflection point, silhouette score, and interpretability of the resulting segments. Clusters were visualized in 3D using a separate 3-component PCA projection, plotted both by cluster assignment and by income label to visually validate that the segments captured meaningful variation.

6.2 Segment Profiles

Segment 2 - " Next Generation" (22.2% of population)

Description: Average age of 7.5 years. These are children. 89.8% are children under 18, zero income, zero labor force participation, 0% high income. Predominantly US-born at 95.4%.

Marketing use: They are not buyers themselves but they are the reason their parents buy. This segment tells the retailer which households have children and can be used to target the parents of this segment with: Children products/supplies, back to school promotions, or family based loyalty programs

Segment 1 - "Retired Wealthy" (13.5% of population)

Description: Oldest segment at average age 51.2. Almost nobody works (97.7% not in the labor force), avg weeks worked just 3.4. Yet they have the highest dividends of all segments at $345.They are living off investment income and savings. 45.2% don't even file taxes. Low high-income rate of 1.7% because they have no wage income, but their dividend income suggests accumulated wealth. Mostly married, US-born, householders.

Key insight: Low wage income does not mean low wealth. These are mostly retirees.

Marketing use: Premium leisure products, health and wellness, financial product, avoid advice related to career goals

Segment 0 - "Married Homemakers" (30.7% of population / LARGEST segment)

Description: Average age 45.9, married (46.9% file joint taxes), educated at some college level (10.4). They show wage income of $75.86/hr but only work 31 weeks a year suggesting part-time or seasonal work. 33.9% are not in the labor force at all while 10.1% work in admin and 9.3% in professional roles. They are primarily householders (51.8%) and spouses (29.7%). High income rate is 8.5% not wealthy themselves but their joint household income is likely higher given married filing status. Capital gains of $592 and dividends of $292 suggest some household investment activity.

Key insight: This is the largest segment and represents the primary household purchasing decision maker like the married people managing the household.

Marketing use: Largest segment for retail, Household consumables, Family healthcare and wellness products, Loyalty programs with family rewards

Segment 3 - "Prime Earners" (25.4% of population / Highest high income ratio)

Description: The most economically powerful segment. Average age 38.7, working full-time (82.4%), earning the highest wage at $113.48/hr, highest capital gains at $810, working nearly a full year at 45.3 weeks. Education is highest at 11.0. Occupations are professional - admin/clerical 15%, professional specialty 14.4%, executive and managerial 12.9%. Industries span retail, manufacturing, and education. 56.5% file joint taxes (mostly married). Highest high-income rate at 12.4%. Most >$50k earners are concentrated. Both male and female, US-born predominantly.

Key insight: This is a premium marketing target. They have income, they are active consumers, they are in their prime spending years with families.

Marketing use: Premium/luxury products, financial products, investment and wealth products.

Segment 5 - "The Young Adults" (8% of population)

Description: Youngest adult segment at average age 27.6, never married. Work about 25 weeks a year. Wage of $52.47/hr but low weeks worked suggests periodic breaks. Education is low at 8.16. 44% not in the labor force, but 8.1% in service occupations and 8% in admin. 37.6% are householders living independently. 25.8% are still children under18. High income rate just 3.9%. Mix of nonfilers (36.8%), joint filers (30.8%), and singles (25.3%) suggesting some are still part of a family unit.

Key insight: Young people starting out - low income now but building their consumer habits and brand loyalties for the future.

Marketing use: young adult discount programs, affordable options, entry level fashion/tech

Segment 4 -"Recent Immigrants" (0.26% of population / niche)

Description: Tiny segment but demographically distinct. Average age 28.2, 65.7% are foreign-born non-citizens and only 25.1% are native US-born. Low wage of $24.81, low weeks worked at 14.7. High income rate just 3.4%. Household structure is mixed with some householders, some still children in the household, some spouses.

Key insight: Recent immigrant community, likely undocumented or early-stage residents, low income, cash economy likely given high nonfiler rate.

Marketing use: Value products and essentials, money transfer services, bilingual marketing

7. Conclusion

Data & Evaluation Integrity

- Population weights were applied consistently across all EDA, evaluation metrics, and cluster fitting to ensure every result reflects true population behaviour,
- Most of the decisions throughout the pipeline were result oriented (AUC based).
- The held-out test set was locked away before any sampling, encoding, or feature selection decisions were made, and they represented the population.
- 1:7 undersampling was a deliberate choice since 1:1 over-corrects and produces unrealistic class distributions that hurt probability calibration on the real population.

Deployment & Targeting Strategy

- Combine classifier scores with segmentation personas for tiered targeting like within Segment 3, flag predicted high earners (>0.8 threshold) for luxury tier offers, while the remaining (>0.5 threshold) get premium-but-accessible messaging.
- Deploy the classifier as a probability score instead of a binary label and define targeting tiers based on marketing ROI thresholds, not a fixed 0.5 cutoff. Playing with thresholds lets the client trade off precision versus reach per campaign.
- Monitor input feature distributions of incoming scoring data against training distributions using PSI (Population Stability Index)
- Given the data is from 1994-1995, any deployment on current population data should be treated as an immediate retraining trigger since structural income drivers have shifted significantly over 30 years.

Future Work

SHAP contribution profiles can be used beyond explainability and can be used to group individuals by their per-feature SHAP values to create buckets for refining personas or identifying actionables.

Expanding fairness analysis to intersectional groups (race and sex jointly) for a more complete disparity picture

Cost-sensitive learning can be applied after understanding the weightage between false positives and false negatives.

8. References

LightGBM: https://lightgbm.readthedocs.io/en/stable/

CatBoost: https://catboost.ai/docs/en/

Hyperopt: https://hyperopt.github.io/hyperopt/

Fairlearn: https://pypi.org/project/fairlearn/

SHAP : https://shap.readthedocs.io/en/latest/

https://medium.com/data-reply-it-datatech/explainable-ai-shap-values-1c7128ef06c2

Imbalanced-learn: https://imbalanced-learn.org/stable/user_guide.html#user-guide

GMM: https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/modules/mixture.html