Project Report

Income Classification and Customer Segmentation for Retail Marketing

Manan Shah (mananshah1703@gmail.com)

1. Executive Summary

Objective

The objective of this project:

1. Build a supervised classification model to predict whether an individual earns **more than or less than $50,000 annually**.
2. Develop a **customer segmentation model** to help the retail business better understand and target different groups within the population.

Business Value

- The **classification model** enables targeted marketing, personalized offers, and more efficient allocation of marketing spend.
- The **segmentation model** provides insight into distinct population groups, enabling differentiated messaging, product positioning, and campaign strategies.

2. Data Overview

Dataset Description

The dataset is derived from the **1994–1995 U.S. Census Current Population Survey**, consisting of:

- **40 demographic and employment-related variables**
- A **binary income label** (`<=50K`, `>50K`)
- A **sampling weight** representing population distribution
- Demographics: age, sex, race, marital status, citizenship etc
- Work / Employment: class of worker, occupation, industry, full/part-time status, weeks worked per year
- Economic: wage per hour, capital gains, capital losses, dividends from stocks

Target Variable

- **Income Class**:
  - `<= $50,000`
  - `> $50,000`

3. Data Exploration and Preprocessing

Exploratory Data Analysis

Key observations from exploration:

- The target feature is extremely **imbalanced**, with a higher proportion of individuals earning ≤ $50K (93.7% vs 6.3%).
- The numerical features exhibit strong right skew and heavy concentration near zero, with a small number of large outliers, indicating the need for scaling and transformation.
- Many categorical variables are high-cardinality and skewed, with long-tailed category distributions indicating a small number of dominant categories and many infrequent levels, reflecting strong population concentration in common demographic groups.

Data Cleaning

- Dropped duplicates (1.6%) to prevent bias and ensure data integrity.
- Placeholder missing-value indicators were converted to NaN to ensure uniform missing-data processing.
- Features with high missingness, extreme class imbalance, excessive cardinality, redundancy, or weak predictive signal were removed to improve model stability and generalization.
- High-cardinality features were consolidated to reduce dimensionality and improve model robustness.

Feature Encoding

- **Numerical features** were standardized using scaling to ensure comparable feature magnitudes.
- **Categorical features** were one-hot encoded to handle nominal variables.
- **Ordinal features** were encoded using a predefined ordering to preserve their inherent rank structure.

Train–Test Split

- **Data split:** The dataset was divided into training (60%), validation (20%), and test (20%) sets.
- **Stratification:** Stratified sampling was used to preserve income-class proportions across all splits.

4. Classification Model

Models Evaluated

Three supervised models were trained and compared to predict whether an individual earns more than $50K: **Logistic Regression** as a strong linear baseline, **XGBoost** as a high performance gradient boosted tree model, and **CatBoost** as a gradient boosted model designed to handle categorical features effectively and perform well with limited feature engineering.

Why threshold tuning was necessary

This is an imbalanced classification problem, where the positive class (income greater than $50K) is meaningfully smaller than the negative class. In this setting, using a default probability threshold of

0.50 often under detects the minority class because the model can achieve high overall accuracy while still missing many true high income individuals.

To address this, the probability threshold was explicitly tuned to align the classifier with the business objective, rather than relying on a fixed default cutoff.

Threshold tuning strategy 1: Optimize for F1 on the minority class

Goal

The first tuning strategy focused on improving performance on the minority class by selecting the probability threshold that maximizes **F1 score**. F1 is appropriate here because it balances **precision** (how many predicted high income are truly high income) and **recall** (how many true high income individuals are successfully identified). This provides a balanced operational point when the client wants both decent targeting quality and meaningful coverage.

Best threshold and results

Using F1 optimization, the best threshold was: **0.298**.

At this tuned threshold, the selected final model was **CatBoost**, with the following performance on the test set:

- **Class 0 (income ≤ $50K)**
  Precision: **0.97**, Recall: **0.97**, F1: **0.97**
- **Class 1 (income > $50K)**
  Precision: **0.60**, Recall: **0.62**, F1: **0.61**
- **Overall accuracy: 0.95**
- **ROC AUC: 0.952**

Interpretation in business terms

- The model is highly reliable for the majority population, correctly classifying most individuals earning ≤ $50K.
- For the high income group, the model achieves a workable balance:
  - **Recall 0.62** means it captures around 62 percent of true high income individuals.
  - **Precision 0.60** means around 60 percent of the people flagged as high income are truly high income.
- The ROC AUC near **0.95** indicates strong ranking ability, meaning the model is effective at ordering people from most likely to least likely to be high income even if the exact threshold changes.

Why CatBoost was selected

Final selection was based on **overall accuracy** plus **minority class F1**, because accuracy alone can be misleading in imbalanced problems. CatBoost delivered the strongest combined result, producing the best minority class performance while preserving excellent overall accuracy.

Threshold tuning strategy 2: Optimize for Youden's J to maximize high income capture

Why a second tuning strategy

After choosing CatBoost as the best overall model, a second threshold optimization was performed for an alternate business use case.

If the retail client's strategy is to focus on not missing high income individuals, for example a premium offer campaign where coverage matters more than efficiency, then maximizing recall for the high income class becomes the priority. This requires accepting more false positives.

Youden's J statistic, defined as **TPR - FPR**, selects a threshold that maximizes separation between true positives and false positives across classes.

Best threshold and results

Using Youden's J optimization, the best threshold was: **0.0656**.

At this threshold, performance becomes recall heavy for class 1:

- **Class 0 (income ≤ $50K)**
  Precision: **0.99**, Recall: **0.86**, F1: **0.92**
- **Class 1 (income > $50K)**
  Precision: **0.31**, Recall: **0.90**, F1: **0.46**
- **Overall accuracy: 0.86**

Additional operational counts on the test set:

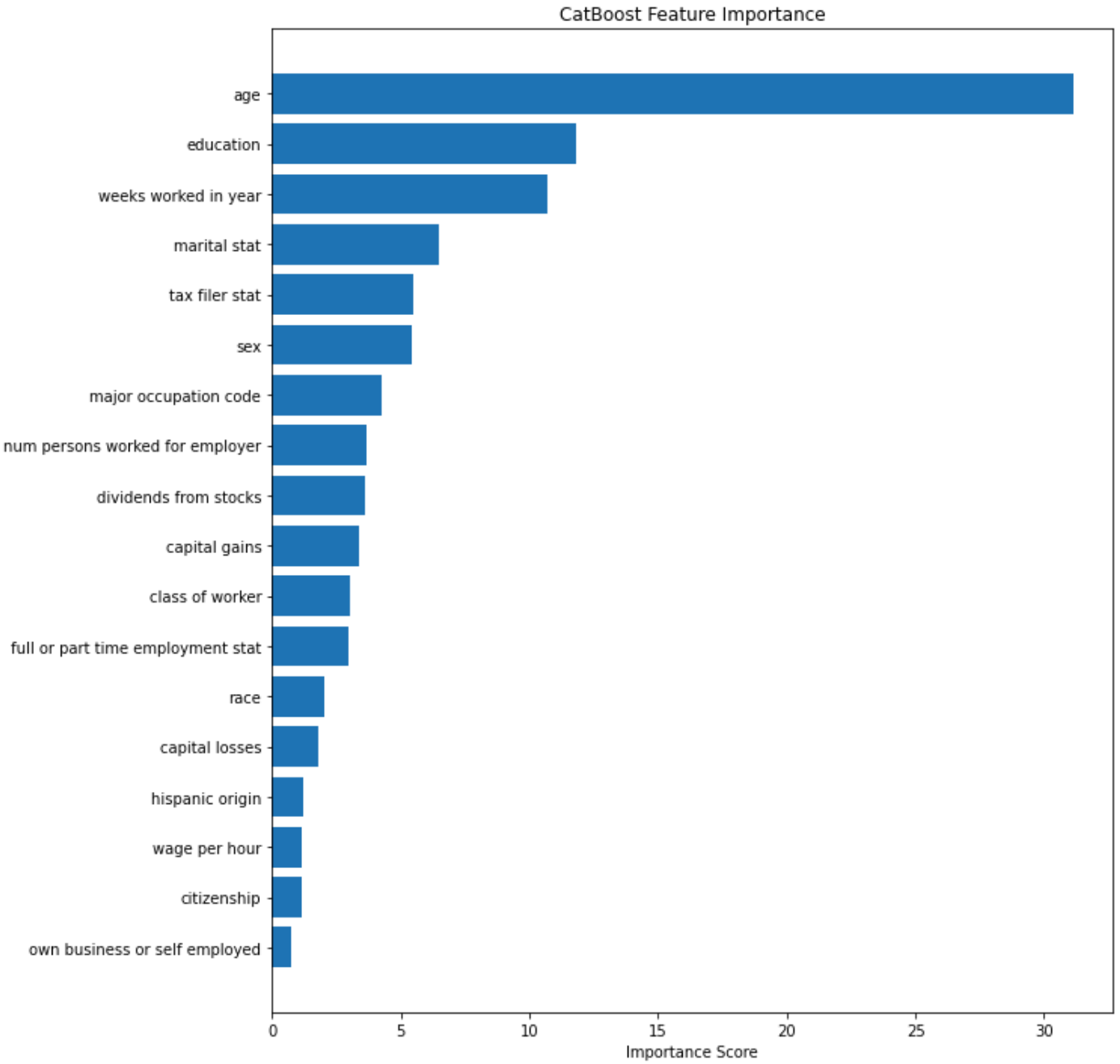
- Total test samples: **39,259**
- True positives: **2,216**
- False positives: **5,125**
- False positives as a fraction of the full test set: **0.1305**

Interpretation in business terms

- **High income recall increases to 0.90**, meaning the model captures about 90 percent of truly high income individuals.
- The tradeoff is **precision drops to 0.31** for the high income class. In practice, that means a larger share of contacted customers will not actually be high income.
- False positives are about **13.05 percent** of the entire tested pool, presenting a manageable trade-off for targeting the high income individuals.

So this configuration is best suited for campaigns where the cost of sending an offer is low, and the cost of missing a high value customer is high.

Feature Importance



- The model shows that age is the strongest predictor of earning over $50K, suggesting income likelihood increases significantly with career progression and experience.
- Education level and weeks worked in a year are the next most important features, highlighting that both human capital and consistent labor play a major role in crossing the $50K threshold.
- Demographic and job-structure variables such as marital status, tax filer status, sex, occupation, and class of worker contribute moderately, indicating socioeconomic and employment-type effects.
- Investment-related variables (capital gains, dividends) add some predictive value but are secondary
- Factors like citizenship, race, and self-employment have relatively low importance, meaning they add limited incremental signal once core labor and education variables are accounted for.

Recommended way to use the model in production

1. **Balanced targeting mode (recommended default)**
   Use F1 tuned threshold Catboost when the client wants a reasonable balance of reach and efficiency. This is suitable for most general marketing efforts where the cost of contacting the wrong person is non trivial.
2. **High income capture mode**
   Use Youden's J tuned threshold Catboost when the client wants to maximize the likelihood of reaching most high income individuals, accepting a larger volume of false positives. This is suitable for premium acquisition funnels.

5. Segmentation

Objective

The goal of segmentation is to uncover **naturally occurring groups** within the population that differ in demographics, employment, and income-related behaviors.

Data consistency and deduplication

As with the income classification task, duplicate records were removed to ensure that each observation represents a unique individual and does not bias the segmentation results.

Feature retention strategy

Most features removed during income classification were also excluded from segmentation. However, several demographic and household-level variables were retained to preserve richer signals relevant for customer grouping, **including**:

country of birth, education enrollment, household composition, industry code, residential stability

These features provide additional context around family structure, mobility, cultural background, and employment environment, which are critical for meaningful and interpretable segmentation. Where necessary, categories were consolidated to improve stability and interpretability.

Feature Engineering

Income and capital-related features

An **annualized wage** variable was constructed under the assumption of a standard 40-hour work week:

- Annualized wage = wage per hour × weeks worked per year × 40

A **net capital income** feature was also created to capture non-wage financial activity:

- Net capital income = capital gains − capital losses + dividends from stocks

Life-stage feature

To capture demographic patterns related to age without treating age as a strictly linear variable, individuals were grouped into **age bands** representing common life stages:

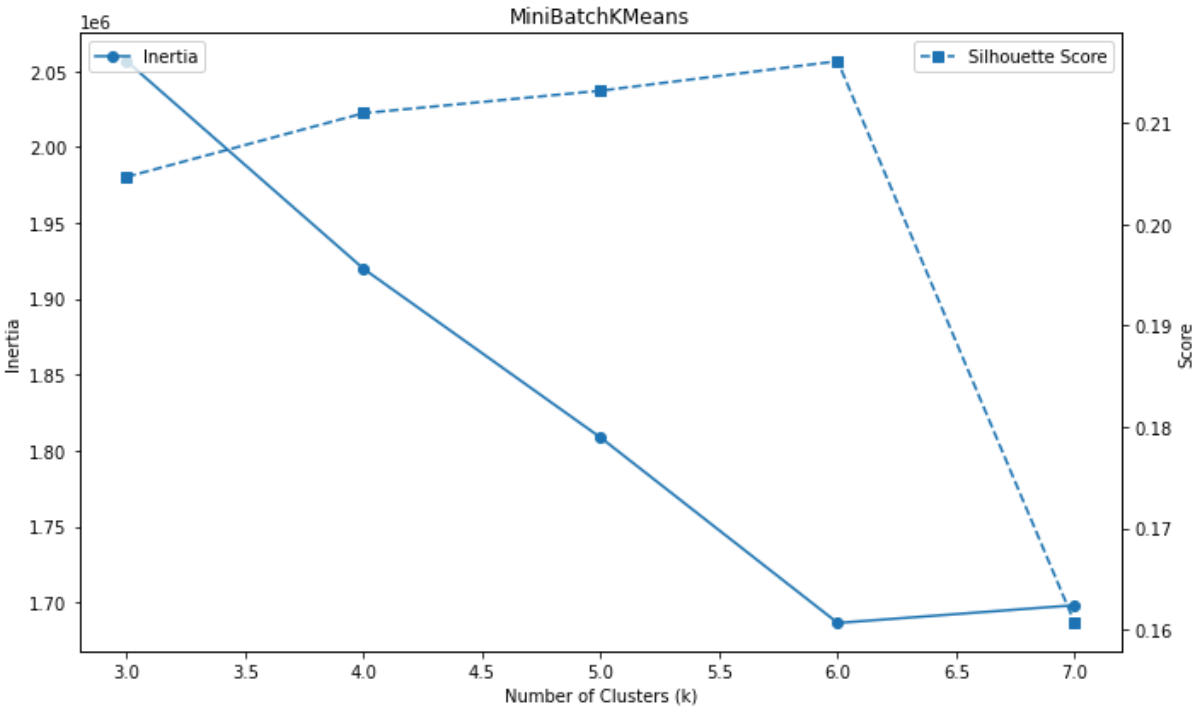
- Child (<18)
- Young Adult (18–25)
- Early Career (26–35)
- Mid Career (36–50)
- Late Career (51–65)
- Senior (65+)

Raw income and capital-related variables, including hourly wage, weeks worked, capital gains, capital losses, dividends, and the income label, were removed after feature engineering.

Segmentation model and cluster selection

For segmentation, **MiniBatch K-Means** was selected due to its computational efficiency and scalability on large datasets, while still producing stable and interpretable clusters.

Choice of number of clusters (k)



The optimal number of clusters was determined by jointly evaluating **inertia** and **silhouette score** across values of *k* ranging from 3 to 7.

Key observations:

- **Inertia** consistently decreased as *k* increased up to 6, indicating improved within-cluster compactness, after which gains diminished.
- **Silhouette score** increased steadily up to *k* = *6*, suggesting better cluster separation, but dropped sharply beyond this point.

Based on the combined elbow behavior in inertia and the peak in silhouette score, **k = 6** was selected as the optimal number of clusters.

Cluster Descriptions

- **Cluster 1: Senior Retirees (22.7 %)**
  Retired seniors living on Social Security, pensions, and modest investments. They have lower educational attainment and worked primarily in blue-collar or mid-level positions. Predominantly married couples with traditional joint tax filing.
- **Cluster 2: Young Dependents (27.3%)**
  Children and young dependents, likely including infants through teenagers. They have no independent income, are claimed as dependents, and are either in K-12 education or too young for school.
- **Cluster 3: Working-Class Commuters (25.1%)**
  Young to early-career workers in blue-collar and transportation jobs. High school educated workforce engaged in manufacturing, logistics, and trade sectors. Likely includes truck drivers, warehouse workers, factory employees.
- **Cluster 4: Upper Middle-Class Families (3.1%)**
  Married couples filing jointly with the highest wages in the population. Mix of skilled blue-collar and white-collar workers in manufacturing and professional roles. Suburban family-oriented lifestyle with traditional household structures.
- **Cluster 5: Older Government & Public Sector Workers (1.1%)**
  Late-career and recently retired government workers, including teachers, civil servants, public sector administrators. Many likely have pensions from government employment. Higher proportion of previously married individuals suggests life transitions.
- **Cluster 6: Educated Professional Class (20.7%)**
  Highly educated professional class with advanced degrees working in healthcare, education, and professional services. Highest net capital income suggests substantial investment portfolios, real estate holdings, or business ownership. Married professionals with dual incomes and strong financial planning. Doctors, lawyers, executives, consultants, and successful business owners.

Key Cluster Characteristics

Metric	Leader
Highest Earners (Wage)	Cluster 4 (Suburban Upper Middle)
Highest Net Worth (Capital)	Cluster 6 (Educated Professionals)
Youngest	Cluster 2 (Children: 8.8 years)
Oldest	Cluster 1 (Seniors: 59.5 years)
Most Educated	Cluster 6 (67% bachelor's or higher)
Least Educated	Cluster 2 (Children/Students)
Most Likely Married	Cluster 6 (75.1%)
Highest Self-Employment	Cluster 6 (16.1%)

Interpretation

These clusters represent distinct socioeconomic segments spanning the full lifecycle from childhood through retirement, with clear differences in earnings strategies, education levels, and wealth accumulation patterns. The analysis reveals:

- **Lifecycle Stages**: Clear progression from dependents (Cluster 2) through working years (Clusters 3, 4, 6) to retirement (Clusters 1, 5)
- **Income Diversity**: Different paths to financial security - high wages (Cluster 4) vs. capital income (Cluster 6)
- **Education-Income Relationship**: Strong correlation between education level and both wage earnings and capital accumulation
- **Employment Sectors**: Distinct clustering by industry and occupation type, reflecting different career paths and economic niches

How the Client Can Use These Segments

These clusters enable the retail client to design **segment-specific marketing strategies** rather than relying on income alone.

- **Exclude or deprioritize non-targetable segments** such as *Young Dependents (**Cluster 2**)* and *Senior Retirees (**Cluster 1**)* for income-driven campaigns, reducing wasted marketing spend.
- **Target Working-Class Commuters (Cluster 3)** with value-oriented offers, practical products, and messaging focused on reliability and affordability.
- **Engage Upper Middle-Class Families (Cluster 4)** with premium household products, family-focused bundles, and long-term loyalty programs.
- **Prioritize Educated Professionals (Cluster 6)** for high-margin products, financial services, subscriptions, and premium experiences where lifetime value is highest.

6. Future Scope

- **Evaluate additional classification models** such as neural networks, stacked ensemble methods, and support vector machines.
- **Improve handling of class imbalance** through cost-sensitive learning, adaptive resampling strategies, or alternative loss functions.
- **Incorporate dimensionality reduction techniques** such as PCA or UMAP prior to clustering to reduce redundancy, improve cluster separation, and enhance visualization.
- **Experiment with alternative clustering algorithms** including DBSCAN, hierarchical clustering, or Gaussian Mixture Models to capture non-spherical and overlapping population structures.
- **Enrich the feature space with behavioral or transactional data**, which would enable more actionable segmentation and improve downstream marketing effectiveness.

7. References

1. U.S. Census Bureau. *Current Population Survey (CPS), 1994–1995*. https://archive.ics.uci.edu/dataset/117/census+income+kdd
2. MiniBatchKMeans https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html
3. XGBoost https://xgboost.readthedocs.io/en/stable/
4. CatBoost https://catboost.ai/docs/en/
5. Logistic Regression https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html