Income Prediction and Customer Segmentation for Targeted Retail Outreach

Michael Michelini

Introduction

Marketing strategy plays a critical role in the success of retail businesses. One of the most important questions facing marketing teams is “Which customers should we target?” It is neither efficient nor cost effective to send marketing campaigns to all consumers. Instead, firms must identify customers who are most likely to respond positively to a marketing campaign. With the increasing availability of big data on consumer behavior and background, machine learning techniques now enable data driven marketing strategies.This project aims to develop a machine learning model to predict whether an individual’s income exceeds $50,000. Accurately identifying higher income individuals allows the marketing team to focus campaigns on customers more likely to qualify for premium products. This report outlines the data preprocessing, model selection and development, model evaluation and business implications.

Data Sources

Our analysis relied on one primary data source that describes the general population of interest.

1. **Current Population Surveys (US Census 1994-1995)**: The dataset contains demographic and employment related attributes for a representative sample of the U.S. population. Each record includes a binary label indicating whether annual income exceeds $50,000.

Data Limitations

This analysis uses U.S. Census data collected between 1994 and 1995. Because the data is over 30 years old, the income patterns and workforce trends reflected here may not fully match the economic environment today. The results should therefore be viewed as insights into relationships within this dataset rather than direct predictions of current consumer behavior.

Income Prediction Modeling

Exploratory Data Analysis

An exploratory data analysis was conducted to better understand class distribution, feature behavior and relationships between demographic variables and income.

Class Distribution
The dataset shows noticeable class imbalance, with around 94% of groups earning less than $50,000 compared to just 6% earning above $50,000.

Age vs. Income and Gender
Figure 1 shows the proportion of individuals earning more than $50,000 across different age groups. The likelihood of earning above $50,000 increases steadily from early adulthood through peak working years, reaching its highest levels around ages 40 to 60. After age 60, the proportion declines as individuals approach retirement. When separated by gender, the overall age pattern remains the same for both males

and females. However, males exhibit a substantially higher proportion of incomes above $50,000 during prime working years when compared to females.

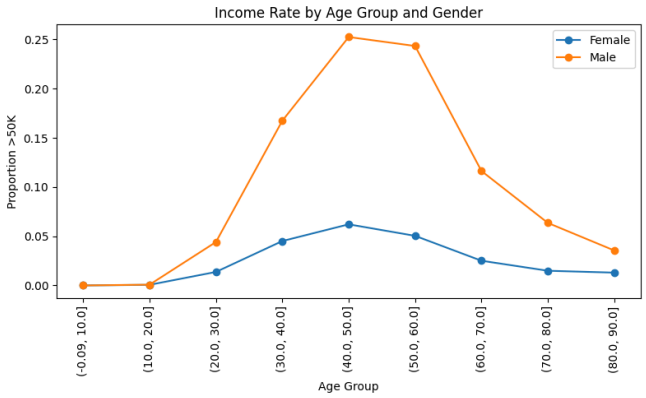

Figure 1: Income Rate by Age and Gender

Education and Income

Figure 2 demonstrates a strong positive relationship between education level and income. Higher levels of education achieved correspond with a substantially greater likelihood of individuals earning above $50,000. In contrast, individuals achieving a high school education or less are much more unlikely to earn above $50,000.

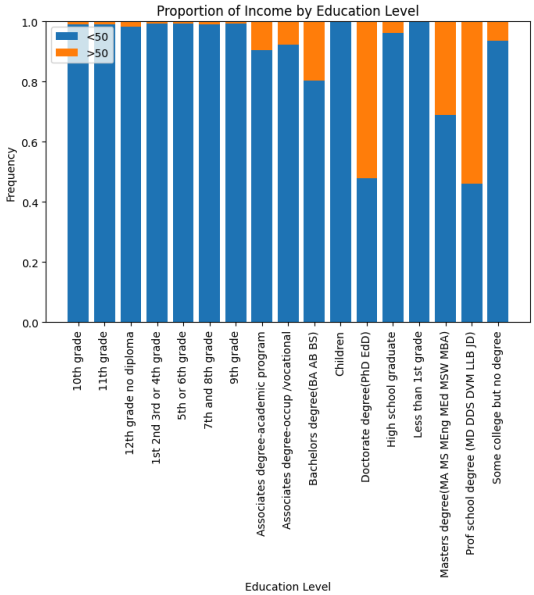

Figure 2: Income Rate by Education Level

Financial Activity and Income

Figure 3 compares income distribution between those with zero financial activity and those reporting any financial activity. Financial activity includes non zero values in any of capital gains, capital losses, dividends, or hourly wage income. Among those with no reported financial activity, nearly all fall into the less than $50,000 class. In contrast, those reporting any non zero value of these financial measures show a higher proportion of incomes above $50,000. This suggests a clear relationship between reported financial activity and higher income levels.

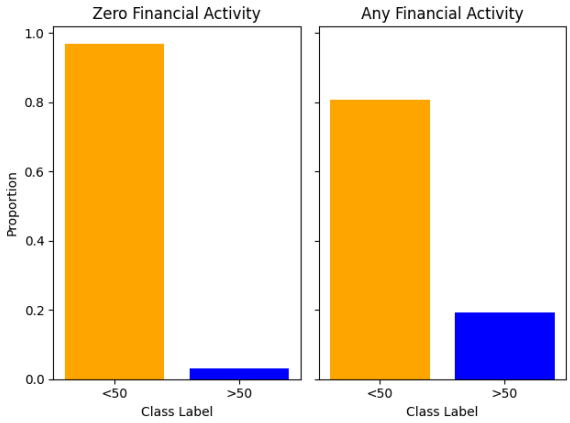

Figure 3: Income Rates by Financial Activity

Data Preprocessing Pipeline

The objective of the preprocessing pipeline is to convert the raw census data into a format suitable for classification models. The pipeline consists of the following stages:

1. **Age Filtering**
   The dataset was filtered to include only individuals 20 years or older. There were zero individuals under the age of 20 that were earning more than $50,000 annually. Including them would further the class imbalance issue and would provide little predictive signal, potentially biasing the model.
2. **Handling Missing or Non Applicable Values**
   Certain columns contained high frequencies of values such as “Not in universe” or “?” representing cases where the attribute did not apply. If 40% or more of a column’s values fell into this category, the column was dropped. Retaining these variables would require large scale imputation and could introduce noise rather than signal.
3. **Removing Redundant Features**
   Highly correlated or redundant columns were removed to improve interpretability and avoid duplicating information. If columns were found to be duplicating information, only one representative variable was retained.
4. **One Hot Encoding**
   All categorical variables were transformed using one hot encoding. Each category becomes a binary indicator variable, enabling numerical based models to process categorical data.
5. **Numerical Variable Scaling**
   All numerical features were standardized to mean 0 and standard deviation 1. This prevents numerical features with large magnitudes from disproportionately influencing the model coefficients.

Feature Engineering

Beyond standard preprocessing, several transformations were applied to variables to improve model performance and interpretability.

1. **Category Consolidation**
   Several categorical variables contained values that were too specific to learn any signal from and as a result were grouped into broader, more meaningful categories:
   - Education levels were grouped into four tiers: high school or less, some college, college and advanced degree
   - Marital status categories were collapsed into never married, married or previously married
   - Employment status was simplified into full time, part time and unemployed
   - Citizenship was consolidated into US citizen vs non citizen
   - Worker class categories were simplified to reduce redundancy
2. **Rare Category Handling**
   For industry and occupation codes, categories representing less than 5% of the dataset were grouped into an "other" category. Since there were several unique occupations, the model had the potential to overfit to occupations that rarely occurred in the dataset. Therefore, this helped to reduce noise from those rare classes and improved model generalization.
3. **Financial Activity Indicator**
   Most financial variables including capital gains, capital losses, dividends and wage per hour were zero for the majority of individuals, with only a small subset of the population reporting positive values. Instead of modeling these highly skewed variables directly, a binary indicator was created to capture whether an individual reported any non zero financial activity. This keeps the relevant signal while reducing the influence of outliers on the model.

Modeling Approach

Since the goal of this project is to predict the binary outcome of income greater or less than $50,000, two different classifier models were evaluated that are capable of handling a supervised learning task like this.

Logistic Regression
Logistic regression is a binary classification model that estimates the probability that an individual belongs to a particular income group based on their characteristics. It evaluates how each feature such as education level or age contributes to increasing or decreasing the likelihood of earning above $50,000. One key advantage of logistic regression is interpretability. It allows us to clearly see which variables have the strongest influence on predictions, making it easier to explain results to business stakeholders.

Decision Tree Classifier
A decision tree model was also evaluated. A decision tree works by repeatedly splitting the data into groups based on feature values. For example, the model might first split on education level, then age, then hours worked, creating a series of decision rules that lead to a final classification. Decision trees can capture complex relationships between variables that logistic regression may not be able to capture.

Evaluation Metrics

Model performance was evaluated using precision, recall and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). Because the dataset contains a high proportion of individuals earning less than $50,000, accuracy alone would not provide a complete picture of model performance. The model could simply predict less than $50,000 each time and achieve deceptively high accuracy, which is why it is not used. Below are brief descriptions of each evaluation metric.

Precision
Precision measures how often the model is correct when it predicts that an individual earns more than $50,000. For example, a precision of 0.50 would mean that 50% of individuals predicted to be in the higher income group actually belong to that group. In the context of marketing, this reflects how efficiently we are able to reach the target group.
Recall
Recall measures how many of the true high income individuals the model successfully identifies. For example, a recall of 0.50 would indicate that the model captures 50% of all individuals who actually earn more than $50,000.
ROC-AUC
ROC-AUC measures how well the model distinguishes between groups of people who do and do not have an income greater than $50,000 across different probability thresholds. A score of 0.50 means the model is randomly guessing which class each point belongs to while a score of 1 means the model is able to perfectly predict which class each point belongs to.

Given the marketing objective of this project, precision on the minority class of people with incomes greater than $50,000 was emphasized over recall to prioritize the quality of outreach and reduce unnecessary outreach costs.

Base Model Comparison

Both logistic regression and decision tree classifiers were first evaluated using their default hyperparameters to establish baseline performance. The scores for precision and recall below are representative of the model's score only on the greater than $50,000 class.

Model Type	Precision	Recall	ROC-AUC
Logistic Regression	0.69	0.35	0.90
Decision Tree	0.40	0.39	0.67

Figure 4: Model Performance Comparison

Logistic regression demonstrated more stable and consistent performance across all evaluation metrics when compared to the decision tree classifier. Given the comparable performance and higher interpretability, logistic regression was selected for further tuning.

Model Tuning

Below are several methods that were tested to assess their effect on the overall performance of the logistic regression model.

Class Imbalance Methods

Class imbalance techniques were evaluated to determine whether model performance could be improved. Two approaches were tested, random oversampling and random undersampling. Random oversampling increases balance by duplicating existing minority class observations, while random undersampling reduces the proportion of the majority class by removing a portion of its observations. However, neither approach improved performance for the model. Both methods resulted in a substantial decline in precision for the minority class, with values decreasing to around 0.32. Because the primary goal of this model was to prioritize precision in identifying higher income individuals, these techniques were not adopted in the final model.

Hyperparameter Tuning

Hyperparameter tuning was performed using grid search with cross validation. Grid search tests different combinations of model hyperparameters to find the settings that produce the best model performance. In this project, precision was used as the scoring metric to align with the goal of minimizing false positives (targeting a low-income individual for a premium product) when identifying high income individuals. The set of parameters that achieved the highest average precision during cross validation was selected for the final model. The following metrics were evaluated:

- **Regularization Strength**: 0.01, 0.1, 1, 10
- **L1 Ratio**: 0, 0.5, 1
- **Class Weighting**: None, Balanced

The top scoring model was able to achieve a precision of 0.70 and ROC-AUC of 0.90 with the following parameters:

- **Regularization Strength** = 0.01
- **L1 Ratio** = 1
- **Class Weight** = None

Threshold Tuning

By default logistic regression classifies observations using a probability threshold of 0.50. This means an individual is predicted to earn more than $50,000 only if the model assigns a probability greater than 50%. However, in different applications, the optimal threshold depends on the goals of the specific project. In this case we already determined we would like to identify high income individuals while minimizing wasted outreach. To assess this tradeoff, model performance was tested across a range of thresholds between 0.30 and 0.90. As the threshold decreases we receive the following results on precision and recall:

- Recall increases meaning we capture more total high income individuals
- Precision decreases, increasing the likelihood of incorrectly targeting lower income individuals and losing valuable marketing time and resources

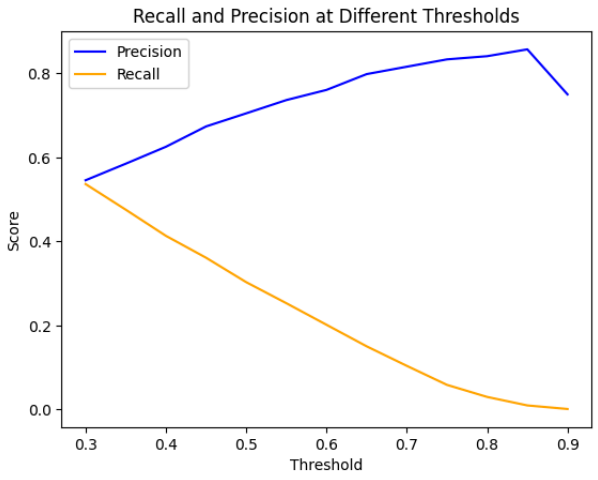

Figure 5: Precision vs Recall at Different Thresholds

After testing several thresholds, we determined the optimal value to be 0.55. This resulted in a precision of 0.74 and recall of 0.25. This was chosen to again emphasize targeting quality over quantity. From a business perspective, the cost of a false positive includes not only wasted marketing spend and resource allocation but also a potential degradation of brand sentiment due to irrelevant offer targeting. By accepting a lower recall, we ensure that the individuals identified by the model have a significantly higher probability of conversion, thereby protecting the marketing budget and ensuring smoother customer experience for those flagged as high income.

Final Model Results

Using the selected probability threshold of 0.55, the tuned logistic regression model achieved the following performance on the held out test set:

- **Precision**: 0.74
- **Recall**: 0.25
- **ROC-AUC**: 0.90

The model prioritizes precision, meaning the individuals predicted to earn above $50,000 are highly likely to belong to that income group. While recall is lower due to the higher threshold, this tradeoff aligns with the objective of minimizing unnecessary marketing outreach.

Feature Importance

To improve interpretability, logistic regression coefficients were exponentiated to obtain odds ratios. An odds ratio greater than 1 indicates a higher likelihood of earning above $50,000, while a value less than 1 indicates a lower likelihood. A value of exactly 1 has no effect.

The most influential features associated with higher income included:

Variable	Coefficient
Obtaining an advanced degree	2.75
Number of weeks worked per year	2.59
Major Occupation Code - Executive Admin	2.45

Figure 6: Variables Most Associated with High Income

The most influential features associated with lower income included:

Variable	Coefficient
Completing high school or less	0.36
Gender - Female	0.42
No financial activity	0.52

Figure 7: Variables Most Associated with Lower Income

Based on the coefficients above, individuals with an advanced degree have 2.75 times the odds of earning above $50,000 compared to those without an advanced degree, holding other factors constant. Individuals who completed high school or less have approximately 64% lower odds of earning above $50,000

compared to those with higher levels of education, holding other factors constant. For continuous variables such as number of weeks worked per year a log odds of 2.59 means for every one unit increase in weeks worked per year, the odds of obtaining an income over $50,000 increase by a factor of 2.59. For example, someone who worked 21 weeks has 2.59 higher odds to have an income greater than $50,000 than someone who worked 20 weeks.

Income Prediction Model Business Impact

The supervised income model gives the marketing team a more focused way to decide who to target. By setting the probability threshold at 0.55 and prioritizing precision, the model emphasizes quality over quantity. In practice, this means that most individuals flagged as earning above $50,000 are likely to truly fall into that category, helping reduce wasted outreach and improving the efficiency of marketing spend.

The feature importance results also provide useful context. Factors such as holding an advanced degree, working more weeks per year, and being in executive or managerial roles were strongly associated with higher odds of earning above $50,000. These insights can help refine targeting strategies and shape messaging for campaigns aimed at higher income customers.

On the other hand, characteristics linked to lower odds of exceeding the income threshold suggest where premium product marketing may be less effective. Recognizing these differences allows the retail client to allocate resources more thoughtfully and align offerings with customers' likely purchasing power. Together, the predictive model and its interpretable features provide both a targeting tool and a clearer understanding of the customer base.

Segmentation Model

Modeling Approach

To complement the income prediction model, unsupervised learning was used to identify customer segments within the same preprocessed and cleaned dataset. The objective was to uncover meaningful demographic and employment based groupings that could inform future marketing strategies. Because the dataset contains many categorical and numerical features, Principal Component Analysis (PCA) was applied prior to clustering. PCA was used to reduce the number of variables by summarizing related features into a smaller set of components. This keeps most of the important information while making the data easier to cluster.

Two clustering methods were evaluated:

- **K-Means**: Groups individuals into k number of clusters by minimizing differences within each group.
- **DBSCAN**: Density based method that identifies clusters based on areas of high data concentration and can detect outliers.

Evaluation Metrics

For evaluation of clustering, we tested three metrics: Silhouette Score, Calinski-Harabasz Index and Davies-Bouldin Index. Together they help to describe the general separability, similarity and compactness of clusters. Silhouette score will be the main metric of choice as it provides the most robust evaluation of clustering results. These three metrics are the most suited for our problem as they do not require ground truth labels:

- **Silhouette Score**: Measures how well individuals fit within their assigned cluster compared to other clusters. Higher values indicate better separation
- **Calinski-Harabasz Index**: Evaluates how distinct clusters are relative to their compactness. Higher values suggest better defined clusters
- **Davies-Bouldin Index**: Measures overlap between clusters. Lower values indicate clearer separation

Results and Model Selection

Across the tested configurations, the best performing model was a 3 cluster K-Means model reduced to 12 principal components. The scores for the best model in each category are below.

Model	Silhouette	Calinski Harabasz	Davies Bouldin
K-Means	0.26	5360	1.37
DBSCAN	0.59	2572	0.64

Figure 8: Comparison of Clustering Methods

Although DBSCAN achieved stronger values in two of the three evaluation metrics, this performance was driven by labeling a large proportion of observations as noise. By excluding a substantial share of the data from cluster assignment, the model effectively reduced the complexity of the clustering task, resulting in artificially stronger separation among the remaining points. Because the objective of this analysis was to segment the majority of the population into actionable groups, DBSCAN was not selected as the final model. Figure 8 illustrates this distinction, showing that DBSCAN concentrates clustering on a limited subset of observations, whereas K-Means produces broader and more stable groupings.

Figure 9: Comparison of KMeans and DBSCAN in PCA Space

Looking at the cluster averages from our K-Means model we can infer the below characteristics of each of the groups.

Cluster Label	Description
0	This group consists largely of older individuals who are no longer actively participating in the workforce, as reflected by fewer weeks worked per year. It includes a higher share of previously married individuals and is majority female
1	This segment consists primarily of middle aged individuals with strong workforce participation, most of whom are employed full time in the private sector. Compared to other groups, they show the highest level of financial activity and the largest share of advanced degrees, along with meaningful representation of self employed individuals. The group is predominantly married and has a slightly higher proportion of males
2	This segment is the youngest overall and is characterized by lower workforce stability and minimal financial activity. It includes the highest proportion of unemployed individuals and a larger share of non citizens compared to other groups. The cluster is also predominantly female.

Figure 10: K-Means Cluster Descriptions

Segmentation Model Business Impact

While the supervised model helps identify customers who are more likely to earn above $50,000, the segmentation model provides additional insight into how those customers differ from one another. Rather than treating high income customers as a single group, the clustering results highlight clear differences in life stage, employment patterns, and economic engagement.These differences can inform how marketing campaigns are designed. Established working professionals may be more responsive to premium product offerings. Customers later in their careers may prioritize value or long term quality. Younger or less economically established individuals may respond better to lower priced products or promotions.

By using both models together, the retail client can first identify likely high income customers and then tailor messaging, pricing strategy, and product positioning to better align with each segment. This approach supports more focused marketing efforts and more relevant customer engagement.