Retail Client Project Report

(US Census Data 1994–1995 Income Prediction Project)

1. Report Summary

- Goal:
  This project analyzes 1994–1995 US Census demographic and employment data to predict whether an individual's annual income exceeds $50,000. And create a segmentation model to group the people represented in this data set to help the retail client better understand customer income segments for marketing strategy and risk assessment.

- Result Summary:
  1. For our best prediction model: ROC AUC: 93.3%, PR-AUC: 68.2%, F1-score: 0.627, Accuracy: 94.06%, and Balanced Accuracy: 85.5%.
  2. We can say this model can separate label 1 and label 0. (income greater than $50,000 is label 1)
  3. We also set the threshold by marking only the top 10 % of predictions as label 1. The resulting precision is 0.47 and recall is 0.54. And as long as each real customer is worth at least 1.12 times the cost of one false contact, this approach would yield a positive ROI.
  4. For the segmentation model, we use weighted features to create two indexes, one current income index and one potential index. We then use the K-means model to segment data into four groups:High-Income & High-Potential, High-Income & Low-Potential, Low-Income & High-Potential and Low-Income & Low-Potential.
- Recommendation: The model can support targeted marketing strategies by identifying more specific customer types and enabling different campaigns for each segment.

2. Data Description and Exploration

2.1 Data Source

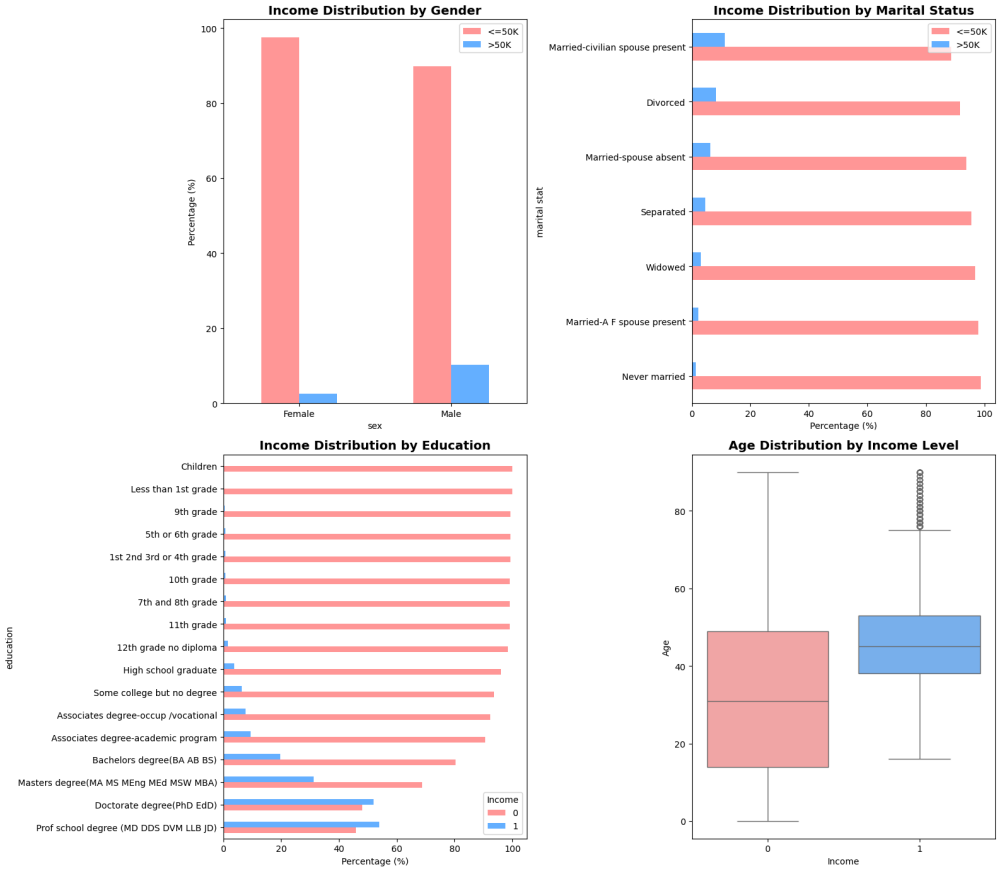
- US Census Bureau 1994–1995 dataset with income labeled by greater or less than $50K.
- Only 6.2% of the data have an income greater than $50K.
- There are 40 distinct features and 199522 data points. The data set also includes sampling weights.

2.2 Feature Overview

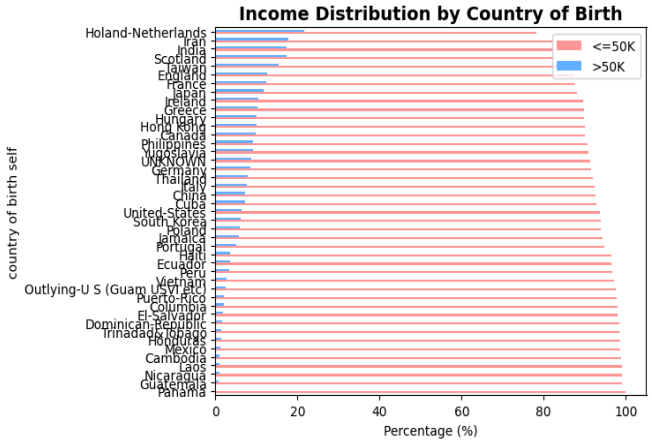
- Numerical features: age, wage per hour, capital gains, etc.
- Categorical features: education, marital status, industry, occupation, sex, etc.
- Target variable: *Income* (0 = ≤ $50 K, 1 = >$50 K).

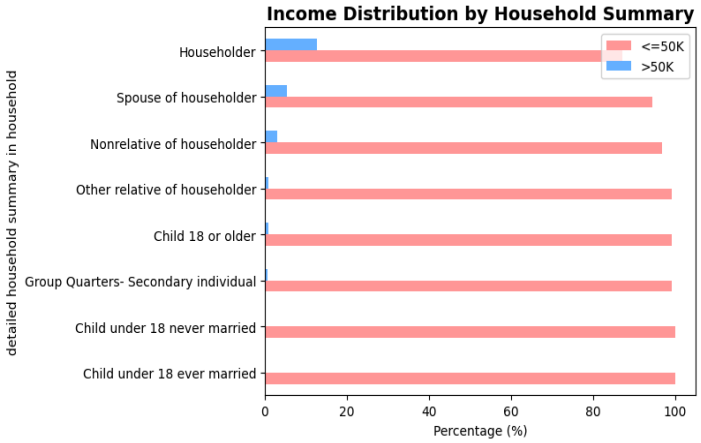
2.3 Exploratory Data Analysis

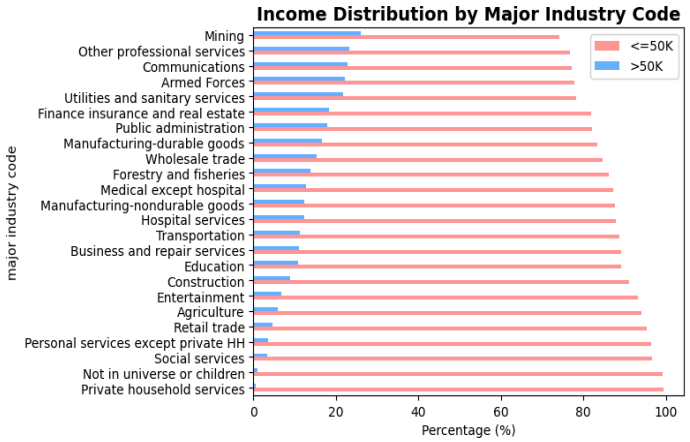
The first thing we did was to check the ratio of the outcomes of our target variable among the data set. We found that the data set was significantly imbalanced, with only 6.2% of the datapoints corresponding to income label 1.

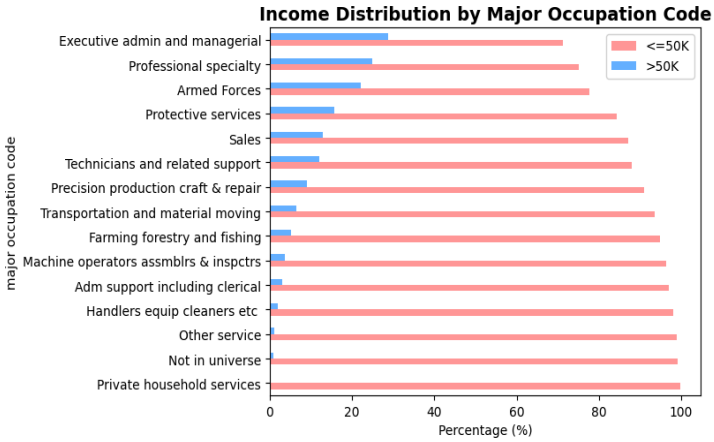


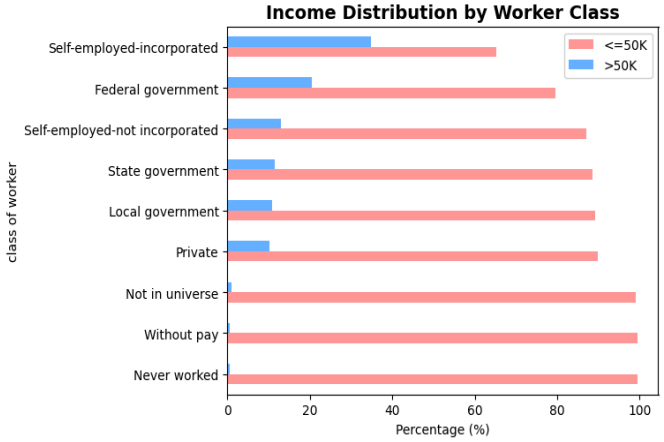
We continued our data exploration by plotting various histograms that demonstrate how the potential traits in each column are associated with our target variable (we show some of these above and below). Although most of this data was unsurprising, we were surprised to see a positive correlation between being divorced and income, between states of previous residence and income, and between the countries of birth of the individual and their parents and income.

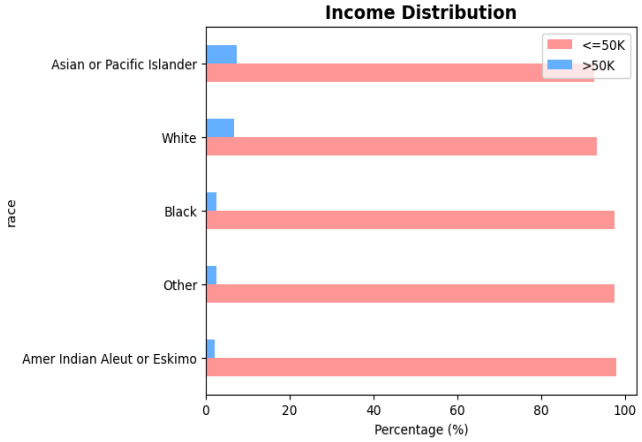












Interestingly, there does not seem to be a clear cut relationship between income and the regions of the associated familial births. One can see that European, middle eastern and asian countries are intermixed in the first chart with no clear broader regional hierarchy.

3. Data Pre-Processing

The first thing we did was check for missing data. We converted “?” and NaN into “UNKNOWN” as a new category. We separated the data into numerical and categorical. We decided early on to use the XGBoost classifier which natively can handle categorical data. We also made sure to not include the weights in our numerical data, as this is a characteristic of the sampling; we will use a weighted XGBoost classifier to account for this data.

The next thing we did was account for the fact that the very vast majority of children do not have an income. In particular, we dropped the rows for individuals under 18, all of which except for two outliers had incomes under 50k. We randomly split the data into 80% training data and 20% testing data, stratified by the income label. To treat the imbalance in our income data, we applied a higher positive weight. We avoided down-sampling to allow the resulting model to be more generalizable to real-world data, and apply our marketing strategy.

4. Model Design and Training

We created three separate models, all three utilizing the weighted XGBoost classifier. In all three models we utilize an AUC-PR evaluation metric due to imbalance in our target data. In all three of these we use XGBoost’s native categorical data handler with a logistic regression for binary classification objective function.

First Model:
For the first model, we dropped children from the data set, but we did not apply a higher positive weight. We mainly constructed this model to serve as a baseline. We evaluated this first model by calculating the ROC AUC: 80.2% and PR-AUC: 39%. We also test various thresholds that produced the best results for three metrics: F1-score: 0.473, Accuracy: 92.2%, and Balanced Accuracy: 77.1%.

Second Model:
In this model, we applied a higher positive weight based on the data ratio of our target variable. We also implemented a k-fold cross validation with early stopping. Rather than using XGBoosts higher level classifier, in order to make use of the k-fold cross validation, we constructed a DMatrix manually so that we could specify the optimal number of boost iterations (i.e. the number of trees built by the model) determined by the cross validation process. This resulted in significant improvements in comparison to the base model. In particular, our new metrics are: ROC AUC: 92.8%, PR-AUC: 65.9%, F1-score: 0.604, Accuracy: 93.29%, and Balanced Accuracy: 85.3%.

Third Model:
Using XGBoosts build in importance score evaluation, we decided to try dropping some low importance input features. Otherwise, everything is the same as in the second model. We also grouped the education column and the household status column in a more simplified manner. This resulted in minor, yet still notable improvements. In particular, our metrics yielded: ROC AUC: 93.3%, PR-AUC: 68.2%, F1-score: 0.627, Accuracy: 94.06%, and Balanced Accuracy: 85.5%.

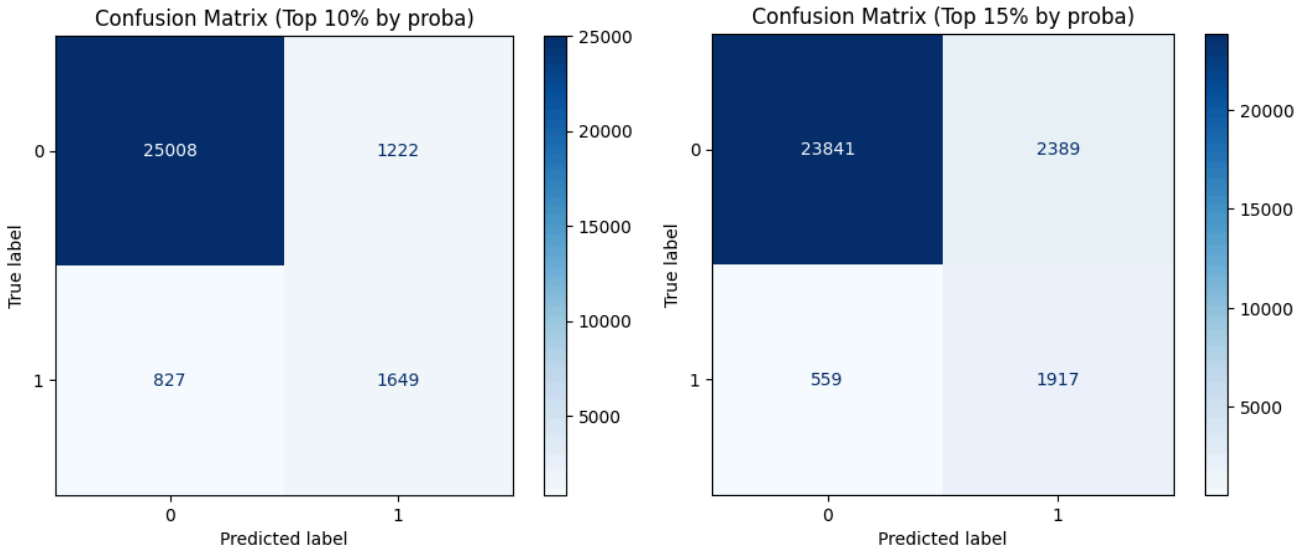
5. Results and Interpretation

Although the gains to the third model are modest compared to the second model. However, the fact that we are able to obtain better results from less data is a significant improvement, especially pertaining to real world applications. The less features and categories within those features requires, equates to shorter surveys being necessary. In particular, not every survey may contain all forty of these variables, so understanding which of these variables are truly necessary for determining one's income and which are noise is of high practical importance.

Based on the Accuracy, Balanced Accuracy, ROC AUC and PR-AUC score we can say that the model is able to separate labels 1 and 0 well.

Suppose we wish to employ a marketing strategy where we aim to target people whose income label is 1. When choosing the prediction threshold, we need to balance precision and recall for label 1.
Below, we set the threshold by marking only the top-K predictions (K = 10% and 15%) as label 1, and examine their corresponding precision, recall, and estimated ROI (Return on Investment). In practice once the true ratio of True Positive/False Positive

cost is known. Then we would have a better idea on whether choosing Top 10% is better or choosing Top 15% is better to define our threshold to get better ROI.



For top 10% threshold cutting, we have the following metrics regarding the target population: Precision: 0.5744, Recall: 0.6660, and F1: 0.6168.

For top 15% threshold cutting, we have the following metrics regarding the target population: Precision: 0.4452, Recall: 0.7742, and F1: 0.5653.

6. Future Improvements and Directions

There are many different directions one could attempt to improve the performance and practicality of this model, which we are unable to do at this time due to time constraints. We list some of these here.

First, as we noticed with the third model we were able to obtain improved performance while requiring less data columns and in the case of some columns coarser overall data. As mentioned previously, being able to reduce the information needed to accurately predict one's income has many obvious real work benefits (e.g. certain information individuals may be reluctant to share). Given more time, it would be good to see what other low importance features could be dropped without impacting model performance. It would also be useful to attempt various groupings of columns that are not dropped to see if we can reduce the complexity of the data set without sacrificing performance.

We also note that we only used the XGBoost Classifier. It would be of interest to compare the performance of our model with other state of the art classifier models such as random forest, LightGBM, or neural networks.

Finally, we think it would be interesting to see how well this data generalizes to more recently collected data. The given data set was collected in the 90's, so it is very possible that trends have shifted since this time.

7. Problem 2 Methodology

We want to run a segmentation model to classify the people in the data set. Since my client is in the retail industry, they might be interested in advertising high quality more expensive products to the people who can afford them and advertising some cost-effective and necessary products for the people who have lower income.
From the first problem, we already built a classifier to predict people's income, we could also use our classifier to capture people who are more likely to be high income. However, in doing so we may ignore some potential groups, for example: some PhD students right now are still at school so their income is clearly less than 50K in general, but once they graduate these people may have a higher probability of consuming the high quality products in the near future. For this group my retail client may want to advertise to them during their current lifestage in order to establish consumption habits and build trust and customer loyalty.
Our model also ignores that even among the high income people, some of them might not be willing to buy unnecessary or luxury items. Although their income is high, this only means that they can afford higher quality and more expensive products. If we just use the problem 1 classifier for our marketing strategy, then we might not get the highest ROI (return on investment).
Given these discrepancies, my natural thoughts are that I want to attempt to classify the people in our dataset into 4 groups: High-Income & High-Potential, High-Income & Low-Potential, Low-Income & High-Potential and Low-Income & Low-Potential. In this case, we will no longer drop the data from people under the age of 18.

To achieve this goal we need to classify our features into two groups and build two indexes for classification. One group of features will indicate people's current economic capacity and the other group of features will indicate people's future growth potential.

Step 1: We do some column engineering. In particular, we convert age, education, occupation, marital status, type of industry, type of house hold, and mobility to numerical data as our previous model from problem 1 showed that these categories have high importance scores with regards to determining income. In order to convert these into numerical data, we will have to assign some numerical order. We based our orderings on the hierarchies that we found during data exploration.

Step 2: Normalize the scores. For some skewed financial data like capital gains, dividends from stocks and wage per hour, we normalized these columns with a log transform. Other chosen columns' scores use the general normalization by distribution.

Step 3: Create the current income index with weights. The weights are based on the importance scores of the features that we obtained in the first problem:

Occupation: 30%, Industry: 25%, Capital Gains: 15%, Dividends: 15%,
Worker Class: 8%, Marital status: 5%, Wage per hour: 2%.

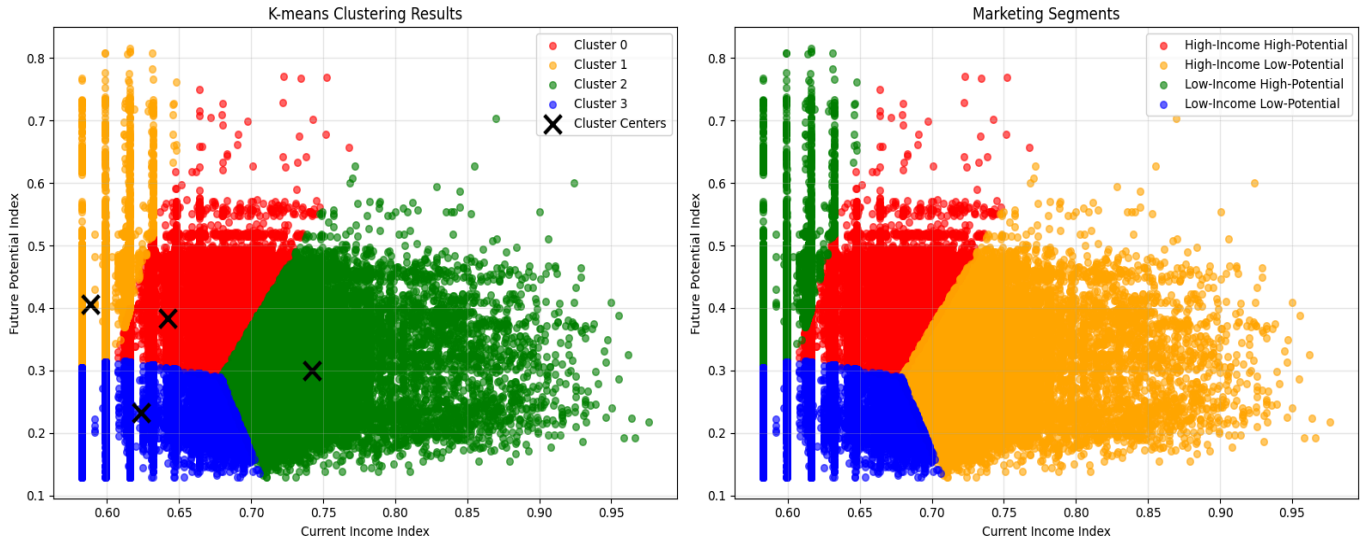
Step 4: We create a future potential index with weights. The weights of geographic mobility and household flexibility are based on the importance scores. And the educational level we adjusted by age.

Education(age-adjusted): 35%, Age Potential: 25% (younger = higher),
Geographical Mobility: 25%, Household Flexibility: 15%.

Step 5: We apply K-means clustering to find natural segmentation.

Step 6: We assign the marketing label. Convert K-means clusters to business segments: Cluster 0: High-Income High-Potential Cluster 1: Low-Income High-Potential
Cluster 2: High-Income Low-Potential Cluster 3: Low-Income Low-Potential

Step 7: Create comprehensive visualizations



Step 8: Evaluation of the K-means Segmentation. Using three scores: Silhouette score, Davies-Bouldin index and Calinski-Harabasz index. Here are more explanations and results for this model.

1. Silhouette Score: measures cluster cohesion vs separation.
   Our model's Silhouette Score: 0.461

2. Davies–Bouldin Index: measures average cluster similarity.
   Our model's Davies-Bouldin Index: 0.859

3. Calinski–Harabasz Index: measures between-cluster variance vs within-cluster variance.
   Our model's Calinski-Harabasz Index: 212158.375

8. Problem 2 Future Improvements and Directions

1. Given more time, we would have liked to try incorporating more features and experimented more with how to assign the potential and income scores, hopefully leading to higher cluster cohesion.
2. If we have more time, we would like to try an alternative approach to solve problem 2: A Causal Inference Approach (GMM + Causal Forest).

   To precisely identify the key factors that cause a customer to become high-income, we propose a future analysis of using causal inference. Since we

cannot conduct real-world experiments(A/B test), this method will simulate them statistically. Here is our proposed method:

1. GMM: First, we use a Gaussian Mixture Model(GMM) for clustering. GMM will provide a probability for each customer belonging to every segment. (e.g.: Person A belongs to group 1 has probability 0.7, Person A belongs to group 2 has probability 0.2, Person A belongs group 3 has probability 0.1).
2. Implement Causal Forest: Next we will use a Causal Forest model. It treats the GMM probability vectors and some other features as input covariates. The model will allow us to define a hypothesis intervention. For example, the Causal Forest could answer the questions like: If a person's probability of belonging to group 1 increases from 0.3 to 0.8, how much does their probability of being high-income change? It allows us to estimate the causal importance of the underlying customer characteristics.

9. References:

https://xgboost.readthedocs.io/en/stable/python/python_api.html
https://en.wikipedia.org/wiki/Silhouette_(clustering) (also used wikipedia to look up some other metrics)
https://matplotlib.org/