Mapping Spending Habits to Income Brackets

Tofunmi Kupoluyi(jdk461) Berwin Gan(wqg203) Brian Shin(shs522) Maria Jaramillo(mdj308)

Introduction

One-quarter of American workers make less than $10 per hour. That creates an income below the federal poverty level. Individuals on top of the income bracket are constantly generating more wealth, as those on the lower income brackets are struggling to meet their ends. With recent political candidates putting forth the concept of Universal Basic Income (UBI) as an essential tool to improve people's quality of life, our team aims to map consumer expenditure to income levels in order to identify the avenues of expenditure a UBI could facilitate for lower income levels.

Model - Description of Methodology

We decided to analyze the data with both supervised and unsupervised learning. Generally, our aim was to observe what income classes (if any) our data could be separated into i.e. if it was possible to get a good classifier for our data. Particularly in the case of supervised learning we were also curious to see if a linear svm could be fitted to the data so the corresponding weights of features could be analyzed.

Supervised Learning For supervised learning, we experimented with both binary and multi-class SVM using linear, rbf and polynomial kernels.

Unsupervised Learning For unsupervised learning, we experimented with k-means clustering over 1 to 5 clusters.

Model - Explanation of Algorithm

Supervised Learning – Determining Number of Income Group Division
We repeated the following process over various divisions of income groups to see which resulted in the most generalizable model. We performed a grid search over various kernels and regularization parameters to obtain the optimal svm configuration.
With the optimal svm configuration attained, we then checked the accuracy of the model on our test set.
A model with high training and test precision and accuracy, indicated that the data could be separated correctly into our predefined groups.

Supervised Learning – Analyzing Weights of the Linear SVM
We performed a grid search over regularization constants from 1-100 to determine the optimal binary linear svm that could be fitted to our data. Since we standardized our data, we analyzed the weights of the features. The highest positive weights indicated which features were most correlated to the higher income earners while the most negative weights indicated the features most correlated to lower income earners.

Unsupervised Learning – KMeans
We first ran k-means clustering over 1..5 clusters and graphed the distortion to determine the best number of clusters to split the data (Elbow Method for K-Means).
We also performed k-means clustering over k=1..5 clusters and tried to see if the average income in each cluster was representative of income groupings.

References

www.bls.gov

Experiments - Descriptions of Datasets

The initial dataset consumed 400+ different columns. We were able to reduce this to 58 columns by removing the following sets of columns:

- Columns with a majority of missing data (80% or more)
- Columns pertaining to information about spouses, as missing data from single users would skew the results
- Repetitive data. If there was a column about weekly and monthly expenditure, we kept the monthly.
- Columns pertaining to income and taxation
- Columns pertaining to aggregate data. For example, if a column C was the summation of A and B, we removed C and left A and B.
- Columns pertaining to government subsidies
- Columns pertaining to income, expect for columns pertaining social welfare and income from investments

Finally, we used ScikitLearn's standard scaler to standardize our data.

Experiments - Explanation of Results

Unsupervised Learning – KMeans
For k above 2 clusters, the clusters generated generally did not correspond to income groupings. With k=2 clusters, the data was separated into clusters with average income ranks: 0.64958097 0.46266122. The income rank is the percentile of income for a respondent.

Supervised Learning - SVM
Based on our findings, it was determined that we best split the data into two groups. Using SVM with a linear kernel, we tested threshold values from 0.5 to 0.9. For example, a 0.6 threshold will give the bottom 60% a 0 label and the top 40% a 1 label.Through our experiments, we were able to identify that the 9:1 division between the "poor" and "rich" gave the highest accuracy of 95.4%, while a 6:4 division gave the lowest accuracy of 75.0%. The factors that had the highest correlation to wealth accumulation are:

- Higher Education Levels

- Number of Hours Worked per Week

- Number of household earning members

- Amount of Interest from Dividends

Discussion - Evaluation of Findings

Our findings indicate that based on expenditure types collected, our data is best separated into two income groups. This could indicate that expenditure generally does not differ amongst lower and middle-income earners.
Further, higher income earners generally had higher e

Discussion - Possible Next Steps

Though promising, we believe our findings are not yet conclusive. More varied expenditure data needs to be collected in order to conduct more meaningful experiments. Also, the data collected majorly corresponded to middle to low-income earners. A more varied sample could also help provide more conclusive results.