Predicting Consumer Spending Proportions per Income Level

Group Members

Group Members	Net ID
Seung Heon Brian Shin	shs522
Maria Del Mar Jaramillo	mdj308
Tofunmi Kupoluyi	jdk461
Berwin Gan	wqg203

* Member responsible for uploading submissions: shs522

Motivation and Description of Problem

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 spending capacity, quality of life, as well as income equality, our team aims to analyze and predict consumer spending proportions (i.e. where consumers spend their money most on) in order to determine whether UBI will be a useful tool to help alleviate poverty and improve income equality in the basic sense.*

* This study assumes rational spending of consumers and will focus on analyzing where the consumer will spend their UBI. It will assume no inflation that may be caused by UBI.

References

We plan on analyzing data from the Bureau of Labor Statistics (www.bls.gov) to determine consumer spending behavior. The Consumer Expenditure Survey dataset includes data about numerous kinds of consumer behaviour. We have chosen to work on the family interview (fmli) dataset which includes data on "consumer unit characteristics, income, weights, and summary level expenditures." We are particularly interested in seeing how spending patterns vary among income levels. We hope to do this by creating a linear model which maps various spending traits to consumer unit salary. By analyzing the weight distribution of the corresponding model, we can determine what individual features (spending behaviours) are most representative of high and low income earners.

Plans

Description of Methodology

The initial labels given from the Bureau of Labour Statistics on the feature of INC_RANK gives us a 0 to 1 label which ranks each reference on a cumulative scale. After reducing the 400+ features into 91 features, we first decided to run an unsupervised K-means clustering to see how each feature group together. We wanted to see how sparsely or densely the cluster would be, hoping to group them into a set number of groups. Furthermore, we have decided to standardize the data to make computations more feasible when input into our chosen algorithms.

Proposed Experiments

When we observe the data from the dataset, there are more than 800 different features being considered. Thus, in order to avoid overfitting, we will experiment with different forms of regularization to ensure that only the most important features are being considered. Due to the fact that Lasso regression is better for sparsity and we need to minimize the number of features, this would be the best form of regression to experiment with.
Feature normalization. To be able to compare the features accurately, we plan to carry out normalization. We will determine whether normalization or standardization is more relevant through experimentation.

Some Relevant Datasets:

The github link on the bottom is a modified and merged version of the data from www.bls.org. We have extracted the data regarding family income, expenditure (weekly, monthly, yearly), spending proportions, characteristics of individual, etc. of American families in 2018. We have cleansed the file by removing unnecessary columns/attributes and populating empty data with either zeros or the mean value (depending on necessity).

https://github.com/bri-shin/MachineLearning_ConsumerSpendingProject/blob/master/fmld_merged%20-%20fmld_merged.csv

Task
We want to determine whether consumer spending habits can be used to predict a person's income bracket.

4.1 Data

The initial data from the Bureau of Labour Statistics gave us around 400+ different columns. We noticed that numerous of these columns had missing data, were misleading in our experiment, and were also unnecessary. The following were characteristics of columns we chose to drop:

- Columns with a majority of missing data
- 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

At the end, we managed to narrow down our 400+ features to 58.

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

Aside from removing some of the columns, we filled nan values with the mean of the column or with 0, depending on which decision was more appropriate to the column name.

By analyzing the different data available, we noticed that numerou

The initial labels given from the Bureau of Labour Statistics on the feature of INC_RANK gives us a 0 to 1 label which ranks each reference on a cumulative scale. After reducing the 400+

The use of binary over scaler: To avoid having too many data columns, we cut down columns

JFSAMT -value of footstamp

REC_FS- yes or no boundaries

Food total, food takeout, food in house

Repetitive data: weekly vs monthly expenditure. As such, we have choosen to remove one. In the beginning, we tried reasoning (city vs outskirt) wether if any of these had further implications such as a person living in the
But after scrutinizing the entire list of labels,

(removing repetitive columns such as iterations and weekly vs monthly)

Spouse related:
Some of the data labels such as HORREF1 (Hispanic Origin of the reference person) which ends with a 1 usually have a coresponding label such as HORREF2 which is used for the reference person spouse. We have chosen to remove all such labels because we

Important but lacking data WELFRX (Earning from pubic assistance)

Removing income and indicitatibe of income such as
Except (investment, social welfare) → indicates something else

4.2 Methodology

Experiments:
* SVM, SVM with linear kernel and gausean, svm with rbf kernel and regularization constant linear kernel is not separable linearly
* svm with slack variables

gausean mixture models

k means clustering

Experiment: Description of methodology

We decided to analyze the data with both supervised and unsupervised learning. Through this, our aim was to observe what income classes (if any) our data could be separated into.

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 k=1-5 clusters.

Experiment: Explanations of Algorithms

Supervised Learning:

For supervised learning, we first attempted to fit a multiclass SVM to the data. Although we achieved good training error, our model did not generalize well to the test set. Given this, we focused on fitting a binary SVM and tinkering the threshold income rank, the income rank is a number betweeen 0 and 1 denoting income levels. We were We then analyzed the weights allocated to each feature. Given that our features were standardized, the features with the highest positive and negative weights gave us an indication of features that were strongly correlated to the respective classes.

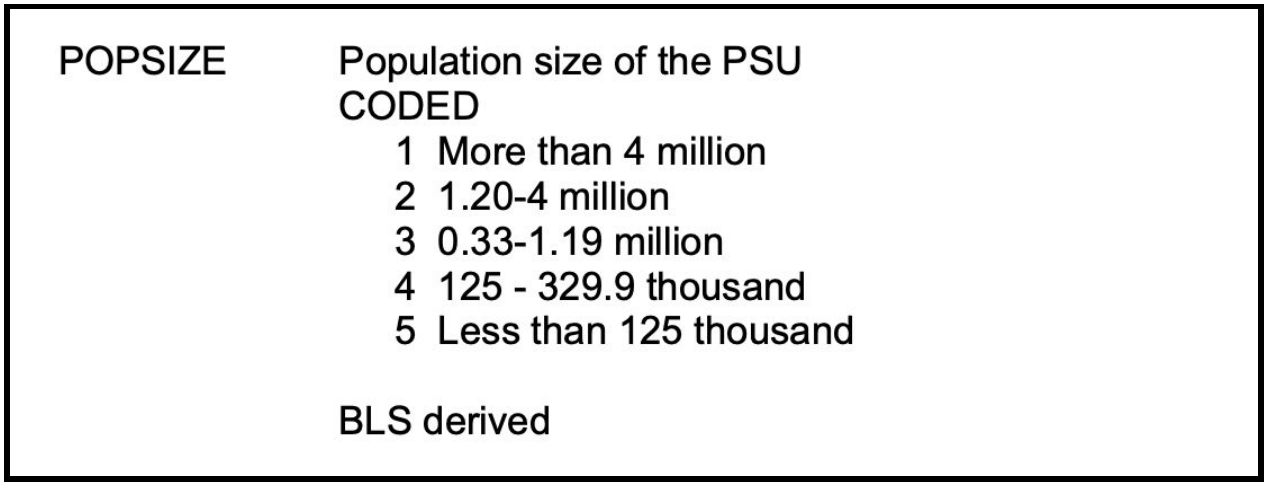
Supervised Learning

Here, we attempted to find if the data can be clustered directly according to income groups based on the expenditures. We clustered with k-means and then looked at the average income rank in each group. A good clustering for us would have

As an aside, we also attempted to use the elbow method to determine the appropriate number of clusters based on the distortion.

CUTENURE	Housing tenure CODED 1 Owned with mortgage 2 Owned without mortgage 3 Owned mortgage not reported 4 Rented 5 Occupied without payment of cash rent 6 Student housing BLS derived

Example of non-linear symbolic data



Example of linear symbolic data

Linear SVM
Split at 5:1

Threshold	Training Accuracy	Test Accuracy	Highly correlated items	Related words
0.5	0.7786	0.7710	[4, 8, 10, 14, 15, 16, 19, 44, 54, 56]	EDUC_REF FGVX HRSPRWK1 NO_EARNR PERSLT18 PERSOT64 REC_FS FOODAWAY INTRDVXM HISP_REF
0.6	0.7990	0.75036	[4, 8, 9, 10, 12, 14, 19, 44, 54, 56]	EDUC_REF FGVX FIRAX HRSPRWK1 JGROCYMV NO_EARNR REC_FS FOODAWAY INTRDVXM HISP_REF
0.7	0.829	0.7828	[4, 8, 10, 12, 14, 19, 24, 44, 54,	EDUC_REF FGVX

			56]	HRSPRWK1 JGROCYMV NO_EARNR REC_FS VEHQ FOODAWAY INTRDVXM HISP_REF
0.8	0.8662	0.84638	[4, 8, 10, 12, 14, 19, 24, 44, 54, 56]	EDUC_REF FGVX HRSPRWK1 JGROCYMV NO_EARNR REC_FS VEHQ FOODAWAY INTRDVXM HISP_REF
0.9	0.91	0.954	[4, 9, 10, 11, 14, 19, 44, 49, 54, 56]	EDUC_REF FIRAX HRSPRWK1 JGRCFDMV NO_EARNR REC_FS FOODAWAY PERSSERV INTRDVXM HISP_REF
0.95	0.9462	0.937961595273 2644	[1, 4, 9, 11, 14, 15, 19, 33, 54, 56]	BLS_URBN EDUC_REF FIRAX JGRCFDMV NO_EARNR PERSLT18 REC_FS EGGS INTRDVXM HISP_REF

Top ten items that are positively-correlated to an increase in ranking prediction,

The largest spending factors that had the highest correlation to

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

Attributes that appear throughout:

4 - EDUC_REF: **Education of reference person**

10 - HRSPRWK1: **Number of hours usually worked per week by reference person**

14 - NO_EARNR: **Number of CU members reported as income earners**

19 - REC_FS: (binary 1=yes and 2=NO)
Have any members of your CU received any Food Stamps, during the past 12 months?

44 - FOODAWAY: Food away from home

54 - INTRDVXM: Amount of income received from interest and dividends, mean of the iterations

56 - HISP_REF: Hispanic origin of reference person

12 (from 0.6 to 0.8)
8 (except last)-FGVX: Amount of government retirement deducted from last pay

Split int N groups	Train Accuracy	Test Accuracy
5	0.9202	0.4401772525849335
4	0.9128	0.4667651403249631
3	0.924	0.6159527326440177
2	0.9462	0.7651403249630724

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.