Meggie Baron

Professor Amy Handlan

ECON 1680: Machine Learning, Text Analysis, and Economics

March 12, 2023

Predicting SNAP Non-Participation Using Machine Learning Tools

1. Introduction

In 2021, the Supplemental Nutrition Assistance Program (SNAP) was utilized by about 1 in every 8 Americans, equating to 41.5 million people (Schanzenbach 2023). In 2020, Rhode Island alone over $280.7 million dollars were issued as part of the program (Service 2022). It is the only social assistance program that is available to all low-income individuals, regardless of work status, making it a critically important social support system in the US (Schanzenbach 2023). However, in 2018 only 82% of SNAP eligible recipients participated in the program (Agriculture 2018). In the same year, Rhode Island had a 93% among those eligible, which is a notable improvement upon the national level, yet it still signifies that about 76,000 eligible households were left without essential nutrition assistance (Agriculture 2018).

This discrepancy raises questions regarding which demographic groups are not redeeming SNAP benefits and how policies may be updated to incentivize participation by these groups. In a 2019 Rhode Island SNAP report, the population of participating households was shown graphically divided by race, education level, age, and town (Rhode Island Department of Human Services 2019). This report revealed the diverse nature of those who utilize SNAP benefits in Rhode Island and increased curiosity around which of these and other characteristics are most highly correlated with SNAP non-participation. The ability to predict SNAP non-participation based on personal characteristics would provide useful data when constructing and modifying SNAP related policies such that assistance may be accessed by all who demonstrate a need. In this research project, I intend to discover which demographic characteristics are most predictive of SNAP non-participation in Rhode Island using machine learning methods. I will employ multiple different regression and analysis methods to discover how allowing for non-linearity may affect the predictive capabilities of a model and determine which model most accurately predicts SNAP non-participation.

1. Data Sources and Descriptions

To perform my research, I will be using data from the American Community Survey (ACS) from Rhode Island in the year 2021. I accessed this data through the IPUMS USA database (Ruggles et. al. 2023). I will be considering the following ACS variables in my analysis: region, metropolitan area status, indigenous peoples homeland status, household income, family size, sex, age, marital status, Hispanic identity, U.S. citizenship, years spent in the USA, English speaking ability, veteran status, employment status, education level, and a variety of disability measuring variables which I will consolidate into one binary variable. I will also consider interactions among race and sex, sex and age, age and sex, and metropolitan area status and family size. All these regressors will provide me with a complete and nuanced representation of each household when using models to predict the output variable: SNAP participation.

INSERT TABLE HERE WITH ALL VARS (CODE ABBREVIATIONS AND MEANING).

1. Methodology

I used many different machine learning methods to assess the possibly non-linear relationship between SNAP participation and key demographic factors. I began by cleaning and normalizing the data such that all variables were numerical with a mean of zero and equal weight. I then divided the data sample into a training set and a testing set so that I was able to check for out-of-sample accuracy.

Using the previously mentioned regressors, I performed an OLS regression on the training data as a benchmark with which to compare the other estimators. I then performed a LASSO regression and a Ridge regression on the training data to account for the possibility of overfitting. I compared the resulting coefficients across the three regressions and calculated the mean squared error for each regression using the testing group to check for out-of-sample accuracy. Using the feature selection benefits of the LASSO regression, I also created a list of the variables that were determined to be most predictive, meaning they did not go to zero in the regression.

Finally, I intend to train a neural network on the training group, apply it to the testing group, and calculate its mean squared error to compare it with the other machine learning methods I am employing. I will interpret the results and determine the most suitable method based on the minimum mean squared error and human reasoning. This methodology will allow me to observe how the use of different linear and non-linear estimators may model SNAP data differently. It will also allow me to discover which demographic variables are most correlated with SNAP non-participation in Rhode Island and how they are correlated with the output so that I can determine which groups are most likely to be underserved by the Supplemental Nutrition Assistance Program in Rhode Island.

1. Results/Expected Results

In my research thus far, I have found that the Ridge regression has produced the smallest mean squared error equal to about 2.60e-26. The LASSO regression gives a slightly less accurate out-of-sample prediction with a mean squared error of about 2.75e-26. The LASSO regression selected level of education, age, the age and race interaction variable, marital status, employment status, residence on an indigenous homeland, veteran status, living in a metropolitan area, citizenship, and the sex and race interaction variable to be the ten variables most highly correlated with SNAP participation. More specifically, low levels of education, old age, and living in an indigenous homeland designated area are highly correlated with SNAP non-participation. I intend to further interpret and explain the other notable coefficients before the final draft using plots and graphs created in Python.

The OLS regression was the least accurate out of sample, although only by a small margin. Its mean squared error was about 2.65e-25. I found that coefficients across the OLS regression were generally larger in absolute value compared to those of the Ridge regression, which is fitting with the design of the model.

INSERT TABLE WITH ALL COEFFICIENTS FOR 3 REGRESSIONS (data below)

A screenshot of a computer

Description automatically generated with low confidence

While I have yet to perform the neural network, I expect the Ridge regression to be the most accurate out-of-sample model considering the neural networks are often prone to overfitting. Given this information, I will need to compare the features selected by the LASSO regression with those most heavily weighted in the RIDGE regression to make a final conclusion about which demographic features are most highly correlated with SNAP non-participation. I will show the differences in out-of-sample estimation by the four models using line plots (below). I will also include a table showing the mean squared errors side-by-side for easy comparison.

INSERT LINE PLOT WITH LINE TRACKING ACTUAL TEST DATA AND LINES FOR EACH MACHINE LEARNING ESTIMATOR (TESTING SET).

INSERT TABLE WITH ALL MEAN SQUARED ERRORS.

1. Conclusion

Through my coursework in economics at Brown, it has become clear that economists are habitual and slow to adopt modern tools, especially those related to computer science, a field in which many economists do not have a robust background. However, through my research into demographic factors correlated with the Supplemental Nutrition Assistance Program, it has become clear that machine learning methods deserve a place in economics research. Using the Ridge regression, I was able to develop a more accurate predictor of SNAP participation based on economic factors. I was also able to determine which factors are most predictive of SNAP participation using the LASSO regression.

From this I found people who (insert key demographic predictors here) are least likely to utilize SNAP benefits for which they are eligible. While it is important to note that this research does not identify why these groups of people are less likely to participate in SNAP, it provides a starting point for further research into what policies or social phenomena may be preventing these groups from accessing their benefits. It also provides intuition around which groups need more attention from SNAP, which can guide policy decisions. Although I chose to focus specifically on Rhode Island, this same methodology can be applied in any state to address SNAP non-participation. This would be especially useful in states with participation rates lower than the national average. It also could be applied to national samples with the addition of state as a categorical regressor.

Bibliography

Schanzenbach, Diane Whitmore. 2023. “Understanding SNAP: An Overview of Recent Research.” Food Policy 114 (January). EBSCOhost.

Rhode Island Department of Human Services. 2019. Rhode Island SNAP Report.

RI Department of Human Services. 2022. “Eligibility & How to Apply.” State of Rhode Island Department of Human Services. <https://dhs.ri.gov/programs-and-services/supplemental-nutrition-assistance-program-snap/supplemental-nutrition-0>.

Steven Ruggles, Sarah Flood, Matthew Sobek, Danika Brockman, Grace Cooper, Stephanie Richards, and Megan Schouweiler. 2023. IPUMS USA: Version 13.0 ACS 2021. Minneapolis, MN. <https://doi.org/10.18128/D010.V13.0>

Supplemental Nutrition Assistance Program State Activity Report: Fiscal Year 2020. 2022. N.p.: Food and Nutrition Service.

U.S. Department of Agriculture. 2018. “SNAP Participation Rates by State, All Eligible People | Food and Nutrition Service.” USDA Food and Nutrition Service. https://www.fns.usda.gov/usamap.