SNAP ELIGIBILITY BY GEOGRAPHIC REGION – Technical Report

Trish Girmus

8/14/2021

Introduction/Background

For this project a dataset was selected from a survey containing 4,826 households conducted by Mathematica Policy Research, on behalf of the USDA's (U. S. Department of Agriculture) Economic Research Service (ERS) from April 2012 to January 2013 (https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/background/). The survey was called the Food Acquisition and Purchase Survey (FoodAPS). The purpose of the survey was to understand how food purchases were made. The survey included four types of sub-groups who participated.

SNAP eligibility of the households who participated in the survey was researched further to learn more. SNAP is an acronym that stands for the **S**upplemental **N**utrition **A**ssistance **P**rogram. To determine SNAP eligibility estimates conducted for the FoodAPS survey, the Microanalysis of Transfers to Households (MATH) SIPP+ Microsimulation Model was used (https://mathematica.org/publications/technical-working-paper-creation-of-the-2012-baseline-of-the-2009-math-sipp-microsimulation-model-and-database). The problem statement was determined as to can it be predicted whether SNAP benefit eligibility accurately portrays a families' need for food based on geographic region?

Methods

The majority of coding written for this project was conducted in Python. R was used for other visualizations. A dataset was selected using households who participated in the FoodAPS survey, which is located on the ers.usda.gov website. A total of 11 data files were included for this survey; however, for the relevancy of this project, the households dataset was selected as it was the most applicable. The dataset contains 4,826 rows, which is each household who participated in the survey, and 279 variables.

The process began by reviewing the faps_household_puf.pdf Household Codebook to determine what features were relevant. After reading the file and saving into a dataframe, 205 features were dropped. The target variables (for supervised learning) were identified for this project which consisted of four:

BENEST1_HH: this is a sum of estimated monthly SNAP benefits for model run 1

BENEST2_HH: this is a sum of estimated monthly SNAP benefits for model run 2

BENEST3_HH: this is a sum of estimated monthly SNAP benefits for model run 3

BENEST4_HH: this is a sum of estimated monthly SNAP benefits for model run 4

(1_Household Codebook PUF.pdf)

Data preprocessing was performed to remove negative values that were filled in the target variables to identify missing values of respondents who took the survey. Visualizations were then conducted using the matplotlib library in Python. Bar plots and scatter plot

visualizations were created to review at a high level. A .csv file was also created to build visualizations in R.

After more data preprocessing to filter out more data with invalid or missing responses, a feature matrix was made. X was the independent features and y was the dependent feature. Each target variable was saved in separate Jupyter notebook file and analysis was performed on each target variable individually.

Initial quantitative variables were converted into categorical variables into a string. Once this was complete, a copy of those columns were dummy encoded. A correlation threshold of 0.75 was then set to determine highly correlated features that could be dropped during feature selection. Pearson Correlation was also used to determine highly correlated features. A heatmap was created using the seaborn library. More features were dropped after this step.

A train and test split was then created to prevent overfitting. The train_test_split package was imported from sklearn. A linear regression package was also imported from sklearn. More scatterplot visualizations were created to review the targeted dataset vs. the predicted group. A multiple regression graph was then created to analyze the y_test values and predicted values. Finally, stasmodels.api was imported so that Ordinary Least Squares (OLS) results could be used for reviewing statistical analysis.

Results

The multiple linear regression graph indicated a strong relationship between the two values for all four target variables. When using the multiple linear model, the following results occurred:

BENEST1_HH - R-squared score of 0.727

The region variable was dummy encoded and therefore now only has 3 p-values. Originally there were 4 region variables. Region 1, the Northeast region, was dropped during feature selection as it was highly correlated. Regions 2, 3, and 4 are used for the remaining target variables BENEST2_HH, BENEST3_HH, and BENEST4_HH.

The p-value for region 2 is 0.654, region 3 is 0.867, and region 4 is 0.227, which is not statistically significant.

BENEST2_HH - R-squared score of 0.637 which overall is the lowest score over all 4 target variables.

The p-value for region 2 is 0.762, region 3 is 0.801, and region 4 is 0.251 which is not statistically significant.

BENEST3_HH - R-squared score of 0.741. This score is higher than the results of BENEST1_HH & BENEST2_HH but could also be improved.

The p-value for region 2 is 0.545, region 3 is 0.823, and region 4 is 0.202, which is not statistically significant.

BENEST4_HH - R-squared score of 0.664. This score is slightly higher than the results of BENEST2_HH but could also be improved.

The p-value for region 2 is 0.617, region 3 is 0.689, and region 4 is 0.222, which is not statistically significant.

Discussion/Conclusion

The regression analysis is currently not predicting SNAP eligibility by region based on the p-values received by the 4 individual target variables BENEST_HH's 1, 2, 3, and 4. However, strong significant p values in BENEST1_HH were found from other features such as whether they paid property tax (0.043 [expproptax_r]), how much they spend on their electric bill (0.011[expelectric_r]), how much they paid in health insurance (0.022 [exphealthins_r]), and whether they've been evicted in the past 6 months (0.001 [evicted6mos]). This feature is exactly why SNAP benefits are important in this case, to help eat nutritious meals and promote self-sufficiency. In BENEST4_HH, the rural p-value was 0.044 which does indicate it could be a good predictor. By possibly using other machine learning models it could be determined whether region is a strong predictor of snap benefits or not.

Acknowledgments

Learning is a journey, and this one has been like no other. I would like to take this opportunity to thank those who have been with me along the way for this project.

I would like to thank my work colleagues, my family, and my friends who have been my biggest cheerleaders and supporters. I appreciate the distractions of life with your calls, texts, and friendly faces via technology while working and learning from home for the past 16 months.

References

1_Household Codebook PUF.pdf

About ERS. USDA ERS - About ERS. (n.d.). https://www.ers.usda.gov/about-ers/.

Ahmad, I. (n.d.). *40 algorithms every programmer should know*. O'Reilly Online Learning. https://www.oreilly.com/library/view/40-algorithms-every/9781789801217/fa0a5d03-e2a7-4f01-bd9c-53c3f54a7da1.xhtml.

Background. USDA ERS - Background. (n.d.). https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/background/.

Documentation. USDA ERS - Documentation. (n.d.). https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/documentation/.

Figure 2. Census regions of the United States. Figure 2. Census Regions of the United States | Bureau of Transportation Statistics. (n.d.). https://www.bts.gov/archive/publications/america_on_the_go/us_business_travel/figure_02.

https://fns-prod.azureedge.net/sites/default/files/resource-files/34SNAPmonthly-6.pdf

https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/. (n.d.). *A Short History of SNAP*. USDA. (n.d.). https://www.fns.usda.gov/snap/short-history-snap. *SNAP Data Tables*. USDA. (n.d.).

https://mathematica.org/publications/technical-working-paper-creation-of-the-2012-baseline-of-the-2009-math-sipp-microsimulation-model-and-database. (n.d.).

Supplemental nutrition Assistance Program (SNAP). USDA. (2021, April 23). https://www.fns.usda.gov/snap/supplemental-nutrition-assistance-program.