PRESENTATION SCRIPT

SLIDE 1 – TITLE PAGE

SNAP ELIGIBILITY PREDICTION BY GEOGRAPHIC REGION

Presented by Trish Girmus

SLIDE 2 - OVERVIEW

Good afternoon! My name is Trish Girmus, an aspiring future data scientist, and today I am going to talk about SNAP eligibility prediction by geographic region. This presentation will begin with a background about SNAP, how I decided to work with data for this project, discussion of data preparation prior to modeling, the modeling I used for prediction, and the conclusion of my analysis.

SLIDE 3 – PROJECT TOPIC ORIGINATION

I was born and raised in a small farming community located in south central Nebraska. Yes, indeed Nebraska is the Good Life! Located in the Midwest, Nebraska is known for its agriculture and food production. Eating three meals a day was something I did not give a second thought to. The cupboards in my home were never bare, and I never went hungry. I have taken for granted my entire life the fact that food has always been accessible to me. Fortunately, I have always had the financial resources to purchase food. Not everyone is this fortunate, especially within the past year during the pandemic where unemployment and uncertainty have been at an all-time high. While I researched for a project topic, hunger and food sustainability really stood out to me, and I wanted to learn more about SNAP, and explore the topic further.

SLIDE 4 – WHAT IS SNAP?

SNAP is an acronym that stands for the Supplemental Nutrition Assistance Program. I will be referring to this acronym throughout the presentation. SNAP helps families in need by providing supplemental benefits to purchase healthy and nutritious food. It also gears families towards self-sufficiency. Examples of food that be purchased with SNAP benefits include: dairy products, meat, poultry, fish, fruits and vegetables, breads, cereals, snack foods as well as seeds and plants if families desire to grow their own food. Thankfully, food assistance has existed in the United States as early as 1939 to the present. As of June 2021, nearly 21,900,000 households receive SNAP benefits. This total is a 6% increase over the past 2 years.

While this total only captures the first 6 months of the year, I presume that it will continue to increase given the current circumstances with the COVID-19 pandemic and uptick with the delta variant.

SLIDE 5 – PROBLEM STATEMENT

As I conducted more research, I was curious how SNAP benefits were distributed to families based on where they resided. For example, using an arbitrary monthly SNAP benefit of $1,500, would that amount be sufficient for a household in New York City vs. a location such as Omaha, NE? I therefore concluded that the problem statement for this project is to predict how SNAP eligibility accurately portrays a families’ need for food based on geographic region. The 4 geographic regions used in this project are: the West, Midwest, Northeast, and South regions.

SLIDE 6 – FOODAPS SURVEY

On the U.S. Department of Agriculture’s (USDA) website, the Economic Research Service (ERS) houses interesting information which determines how decisions are made. Topics include the environment, rural America, food, and agriculture. From April 2012 to January 2013, a survey was conducted by the USDA, called the Food Acquisition and Purchase Survey (FoodAPS). FoodAPS was the first survey performed to gather information about how food was purchased. It was also created to review and learn more about households in America regarding the interrelationships of acquiring food, what determines food demand, and the well-being in those households. The survey was also designed to understand the stores where foods were purchased to identify how accessible these stores were, as well as the preference for food purchased, and the well-being of individuals (primarily health and obesity). The survey had a representative sample of 4,826 households. with four types of sub-groups of households. These four sub-groups were:

* Households who did not participate in SNAP (income was equal or greater than 185% of the poverty guideline)
* Households who did not participate in SNAP (income at or above 100% but less than 185% of poverty guideline)
* Households who did not participate in SNAP (income below poverty guideline)
* Households who did participate in SNAP

SLIDE 7 – DATA OVERVIEW

I selected a dataset using households who participated in the FoodAPS survey, which is located on the ers.usda.gov website. There was a total of 11 data files included for this survey; however, for the relevancy of this project, I felt the households dataset was the most applicable. The dataset contains 4,826 rows, which is each household who participated in the survey, and 279 variables.

SLIDE 8 – MATH SIPP+ MODEL

To determine SNAP eligibility estimates conducted for the FoodAPS survey, the Microanalysis of Transfers to Households (MATH) SIPP+ Microsimulation Model was used. This was for collecting data during the initial and final interviews of primary survey respondents.

The MATH SIPP+ microsimulation model was constructed by Mathematica Policy Research in part to simulate the impacts of proposed new policies or regulations that affect SNAP eligibility or the dollar amount of the SNAP monthly allotment.

SNAP eligibility is based on Federal and State rules. Eligibility requirements are based on rules including assets, income, deductions, employment requirements, special rules for elderly and disabled individuals, and immigrant eligibility.

These estimates resulted in the 4 target variables used with the prediction model.

The technical report published by Mathetmatica Policy Research is included in my references for further information on how the model was created.

SLIDE 9 – SNAP ELIGIBILITY RESULTS OF 4 MODELS

The table below shows the SNAP eligibility results using the MATH SIPP+ model. The simulation model found of range from 47.9% to 55.1% of Households that were eligible for SNAP benefits. As a result, 4 target variables were determined with the model being run 4 times using the same equation.

As you can see, a run of BENEST2\_HH had the lowest percent with half of the households eligible at 47.9%. BENEST4\_HH was almost half of all families that were eligible for benefits.

SLIDE 10 – TARGET VARIABLES USED FOR ANALYSIS

The previous slide showed the 4 target variables which resulted from the microsimulation model. The monthly estimate calculated by the MATH SIPP+ model resulted in 4 target variables which I used for the analysis. The variable names stand for Benefits Estimated in (Groups 1, 2, 3, 4) HouseHold.

These include:

BENEST1\_HH: Sum of estimated monthly SNAP benefits for all eligible SNAP units in HH model run 1

BENEST2\_HH: Sum of estimated monthly SNAP benefits for all eligible SNAP units in HH model run 2

BENEST3\_HH: Sum of estimated monthly SNAP benefits for all eligible SNAP units in HH model run 3

BENEST4\_HH: Sum of estimated monthly SNAP benefits for all eligible SNAP units in HH model run 4

The unit is defined as an entire household, or a subset of individuals living in a household.

SLIDE 11 = TARGET VARIABLE BENEST1 VISUALIZATION

Since I have discussed my target variables used for the analysis, I thought this would be a good Segway into the exploratory data analysis visualizations of the target variables BENEST1, BENEST2, BENEST3, and BENEST 4. I used the Matplotlib library in Python for these visualizations. I created a bar graph and a scatterplot for each region.

Starting with target variable BENEST1\_HH, the bar graph shows a breakdown of the average monthly snap benefits in U.S. dollar amounts by region. There are four regions, as I mentioned earlier in the presentation.

Region 1 is Northeast

Region 2 is Midwest

Region 3 is South

Region 4 is West

The Midwest and West regions show the highest monthly snap benefits, while the Northeast shows the lowest monthly snap benefits.

For the scatterplot, each datapoint is 1 household surveyed in the dataset. As you can see, the highest monthly income total of $25,000 means zero or no participants received SNAP benefits. The microsimulation model deemed them ineligible for these benefits. On the right end of the scatterplot, you will see that 5 participants received around $1,500 a month for SNAP benefits and have an average monthly income total of $0 to $5,000.

SLIDE 12 = TARGET VARIABLE BENEST1 VISUALIZATION

For BENEST2, the bar graph shows that again the Midwest and West regions shows the highest monthly SNAP benefits, while the Northeast shows the lowest monthly SNAP benefits.

On the scatterplot, the findings are very similar to BENEST1. There are 5 respondents who earn $0 to slightly above a $5,000 average monthly income also receive around $1,500 a month for benefits. One respondent received monthly benefits around $1,650, earning shy of $5,000 for total monthly income.

SLIDE 13 – TARGET VARIABLE BENEST3 VISUALIZATION

For BENEST 3, the bar graph shows that the West region receives the highest monthly SNAP benefits, while again the northeast region shows the lowest monthly SNAP benefits received.

On the scatterplots, the findings again are very similar to BENEST 1 & 2. There were fewer respondents again who received the highest monthly SNAP benefits, with a total monthly income of $0 to $5,000. There were 2 respondents who received around $1,700 to $1,750 who had a total monthly income around the $5,000 mark.

SLIDE 14 – TARGET VARIABLE BENEST4 VISUALIZATION

For BENEST 4, the bar graph shows the Midwest Region receives the highest monthly SNAP benefit, while the Northeast region again shows the lowest monthly SNAP benefits received. On the scatterplot, there are nearly a dozen respondents who earn a monthly income between $0 to around $5,000 who receive around $1,250 to nearly $1,650 a month for SNAP benefits.

Drawing some conclusions from the bar graphs indicate the South region did not earn the highest or lowest for monthly SNAP benefits, but was around the middle, or average amongst the 4 regions. The Northeast region received the lowest monthly SNAP benefits amongst all 4 target variables.

On the scatterplots, on average, households who earned $0 to $5,000 a month were the largest group receiving SNAP benefits amongst the 4 target variables. Fewer respondents received the largest monthly SNAP benefits amongst the 4 target variables.

SLIDE 15 – TARGET VARIABLE BENEST1 & BENEST2 VISUALIZATION USING R

For my next visualizations, I used the ggplot library in R. These are both scatterplots comparing family size of each household survey in the dataset again using the calculated monthly SNAP benefits in total U.S. dollars. This slide shows the target variables BENEST1 & BENEST2.

Each datapoint is colored according to what region they were listed as in the survey. As you can see on this slide, all 4 regions are represented here. This is not meant to be a statistical analysis, but merely a birds eye view.

The Family size ranges from 1 to 14 members. The larger the family size, fewer households participated as you can see. Those households do receive higher monthly SNAP benefits. Typically, we don’t see family sizes in that range. The larger family sizes suggest the model takes into account the family size as part of the SNAP eligibility amount.

SLIDE 16 – TARGET VARIABLE BENEST3 & BENEST4 VISUALIZATION USING R

The same conclusions can be made on this slide for target variables BENEST3 & BENEST4.

SLIDE 17 – MONTHLY INCOME BY REGION 1 & 2

This slide is a scatterplot of regions 1 (Northeast) and 2 (Midwest) again using the ggplot library in R. These visualizations are a breakdown of each surveyed household and their reported monthly income per respective region. The index is a representation of the households which are the dots. The monthly snap benefits are along the x axis.

Both of these scatterplots show once again that the majority of the households who participated in the survey earn a monthly income of $0 to around $5,000.

SLIDE 18 – MONTHLY INCOME BY REGION 3 & 4

This slide shows the visualization with regions 3 (South) and 4 (West). Another takeaway from this slide shows that once again almost the majority of all survey respondents in the South region have the lowest monthly income earned of $0 to $5,000.

The West region does show a little more distribution of respondents averaging an income of $5,000 to $17,000 of monthly income, with a few of those respondents earing the most monthly SNAP benefit around $1,000 and slightly higher.

SLIDE 19 – RURAL vs. NONRUAL REPORTED MONTHLY INCOME

This slide also used the ggplot library in R for a scatterplot. These visualizations compare the reported household monthly income between households in rural areas versus households in nonrural areas. As one might presume and can see here, the nonrural reported monthly income has more respondents, with the average monthly income ranging from $0 to $5,000.

SLIDE 20 – PREPROCESSING DATA

The steps for preprocessing my dataset are straightforward. My first goal with the size of this dataset was to remove features that were not relevant to the data. I went through the household codebook PUF.pdf and read through all 279 features and dropped those. Once I did that, I had 74 features. The next step was to find negative values in the survey and remove those. The negative values in the survey were coded as missing values, so it was necessary to remove them. Once I did this, I had 4,777 rows and 74 columns. My next step was visualization, which I have shown in previous slides. After this point, I converted my data to a .csv file which I used in R for other visualization, which was shown in previous slides.

My next step was creating a feature matrix, to determine my independent features and dependent feature. I made 4 copies of my Jupyter notebook file so that each copy was referencing the 4 target variables BENEST1, 2, 3 and 4. After this, I converted the initial quantitative variables into categorical variables. I also had to use dummy encoding. After doing this step, I had 4,032 rows and 140 columns.

I then did a correlation feature and used a threshold of .75 to drop highly correlated features. There were 40 features that were highly correlated, so I dropped those. This left 4,032 rows and 99 features. I noticed after this step that there were 2 features with invalid or no responses. I ended up removing those features as well. This brought the number of features to 97.

I also created a heatmap to visually see the correlations of the independent features.

SLIDE 21 – LINEAR REGRESSION

* For modeling, I used the scikit-learn library which is one of more popular machine learning packages in Python. As this was a large set, I decided to split the data into a test set and training set to prevent overfitting. This was performed for each target variable. Within scikit-learn, there is a built-in linear regression model that I used for this analysis. The training set was used to fit the model. The test set was used to predict using the same model.
* To define linear regression, I found a definition on O’Reilly.com that explains it best. Simple linear regression formulates the relationship between a single continuous independent variable and a single continuous independent variable. A simple regression is used to show the extent that changes in a dependent variable which is shown on the y-axis can be attributed to changes in an explanatory variable shown on the x axis.
* I also used ordinary least squares (OLS) model for further analysis. I also used statsmodels package in python. I used the OLS results to report my findings which are on the proceeding slides.

SLIDE 22 – REGRESSION ANALYSIS RESULTS BENEST1\_HH

This visualization was used in Python. This linear regression model compares the predicted values which is the blue line against the test values, which are the blue dots. I plotted the graph to visually see if there is a strong relationship with predicted data and original target values in the test set. All 4 target variables showed a strong relationship with the test and predicted values.

The x axis is the range of data points available. The y axis is the target variable BENEST1\_HH.

The OLS Regression results indicated an R-squared score of 0.727. Although this could be considered a good score, it could be improved. The region variable was dummy encoded and therefore now only has 3 p values. While I do recognize while there were originally 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.

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.

SLIDE 23 – REGRESSION ANALYSIS RESULTS BENEST2\_HH

This visualization also indicated the linear regression model shows a strong relationship with the predicted data and test values.

The OLS Regression results indicated an R-squared score of 0.637, which overall is the lowest score over all 4 target variables. This score is okay but could be greatly improved.

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.

SLIDE 24 – REGRESSION ANALYSIS RESULTS BENEST3\_HH

This visualization also indicated the linear regression model shows a strong relationship with the predicted data and test values.

The OLS Regression results indicated an 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.

SLIDE 25 – REGRESSION ANALYSIS RESULTS BENEST4\_HH

This visualization also indicated the linear regression model shows a strong relationship with the predicted data and test values.

The OLS Regression results indicated an 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.

SLIDE 26 – CONCLUSION

* Unfortunately, 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’s1, 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 could determine whether region is a strong predictor of snap benefits or not

This project was a great learning experience for me. I believe I chose a topic that required a lot of research to really understand the meaning behind how these survey results were recorded. Choosing a dataset with over 200 variables was a challenge for me as well. Unfortunately, the results of my models could not predict whether region affected SNAP eligibility requirements by household, but as the results from the ordinary least squares indicate other coefficients could be used for prediction.

Thank you for your time today!