Supplemental Nutritional Assistance Program Evaluation & Prediction

Kris Seekford

BNAL 515

Executive Summary

This project analyzed household participation in the Supplemental Nutrition Assistance Program (SNAP) using a structured workflow that moved from data preparation through descriptive exploration and predictive modeling. The dataset required extensive cleaning and restructuring to correct misclassified variables, standardize economic and demographic fields, and produce an analytically stable version of the data.

Once cleaned, the descriptive analysis revealed clear patterns in household income, composition, employment characteristics, and benefit usage, offering several meaningful insights into how SNAP-recipient households differ from non-recipient households. These findings informed the development of predictive models designed to classify households based on SNAP recipiency. Although the final models achieved relatively high overall accuracy, they performed poorly when identifying SNAP recipients themselves—a direct result of class imbalance and overlapping characteristics between the two groups.

As a result, the predictive component cannot be considered successful in its current form. Even so, the project delivered substantial value: it established a strong analytical framework, produced descriptive results that can guide future research, and clarified what types of transformations, variables, and modeling adjustments will be necessary for more accurate classification moving forward.

1. Introduction

This project investigates patterns of household participation in the Supplemental Nutrition Assistance Program (SNAP) and evaluates whether household characteristics can be used to predict SNAP recipiency. Because the dataset contained numerous structural inconsistencies and imprecisely coded fields, the analysis began by stabilizing and reorganizing the data to ensure that demographic and financial variables were consistent and analytically usable.

With a reliable dataset established, the next step was to explore household characteristics across SNAP and non-SNAP groups and determine how these differences could inform a predictive modeling approach. The goal was twofold: first, to document descriptive patterns that shed light on the structure of SNAP participation; and second, to test whether those patterns could be translated into a model capable of accurately identifying SNAP-recipient households. The following report outlines this full analytical sequence, focusing on the data cleaning decisions, exploratory findings, and ultimately the limits of the predictive modeling results.

2. Data Acquisition and Preparation

The data used for this project came from the Integrated Public Use Microdata Series (IPUMS) for the Current Population Survey (CPS). I originally intended to rely on both the monthly CPS data and the Annual Social and Economic Supplement (ASEC), believing the two together might give me the best balance of temporal breadth and variable richness. After examining both in SPSS Modeler, however, it became clear that the ASEC dataset contained all

of the socioeconomic and benefit-related detail necessary for the project. The monthly CPS, while excellent for tracking short-term labor market dynamics, ultimately added complexity without improving the substantive insight of the analysis. The short exploration phase comparing the two datasets still proved useful, as it helped illustrate the trade-off between temporal continuity and depth of information.

Once I committed to using ASEC exclusively, the next challenge involved reconciling the unit of analysis. Because SNAP participation is reported at the household level, I restricted the data to household heads (PERNUM = 1) and merged in household-level attributes—such as family income, household size, spouse earnings, and housing characteristics—using the household ID as the linking key. This approach allowed me to preserve relevant individual-level information about the householder while retaining the household-level context needed for understanding food stamp recipiency. While representing the entire household through the characteristics of the head is not perfect, it is consistent with SNAP eligibility criteria and with standard practices in CPS analysis.

A substantial portion of the preparation work involved dealing with user-defined numeric codes for missing or inapplicable values. ASEC does not use a single missing indicator; instead, each variable has its own set of codes—family income used values like 995–998, while SNAP benefit amounts used 996–999. SPSS Modeler struggled both with the dataset size and with the inconsistent missingness structure, so I moved this part of the cleaning process to Python. There, I reviewed each variable's documentation, replaced its user-defined missing values with NaN, and created a consistently cleaned file that I imported back into SPSS Modeler for analysis.

3. Descriptive Analysis

3.1 Overview and Initial Processing

The combined ASEC dataset for 2018–2023 contained 980,552 observations and 39 features. After filtering to household heads in Python, the sample was reduced to 375,551. During preprocessing, I also constructed a new feature identifying whether the household had single or dual income earners. This required combining marital status information with reported work hours for both the respondent and the spouse. Households where either spouse reported no or undefined work hours were classified as single-income; otherwise, they were labeled dual-income.

Once imported into SPSS Modeler, the features were assigned as continuous, nominal, ordinal, or flag variables and unnecessary variables were removed through a Filter node. This streamlined dataset formed the basis for the descriptive analysis.

3.2 Main Findings from Exploratory Analysis

The first major pattern became evident in the household income variable. The initial distribution was extremely right-skewed, with a handful of observations reaching more than three million dollars in reported income. *Figure 1* visualizes this initial skew, which SPSS flagged with over 13,000 outliers. After removing these extreme cases, the distribution took on a much more interpretable shape (*Figure 2*), making the inverse relationship between income and SNAP participation more visually apparent.

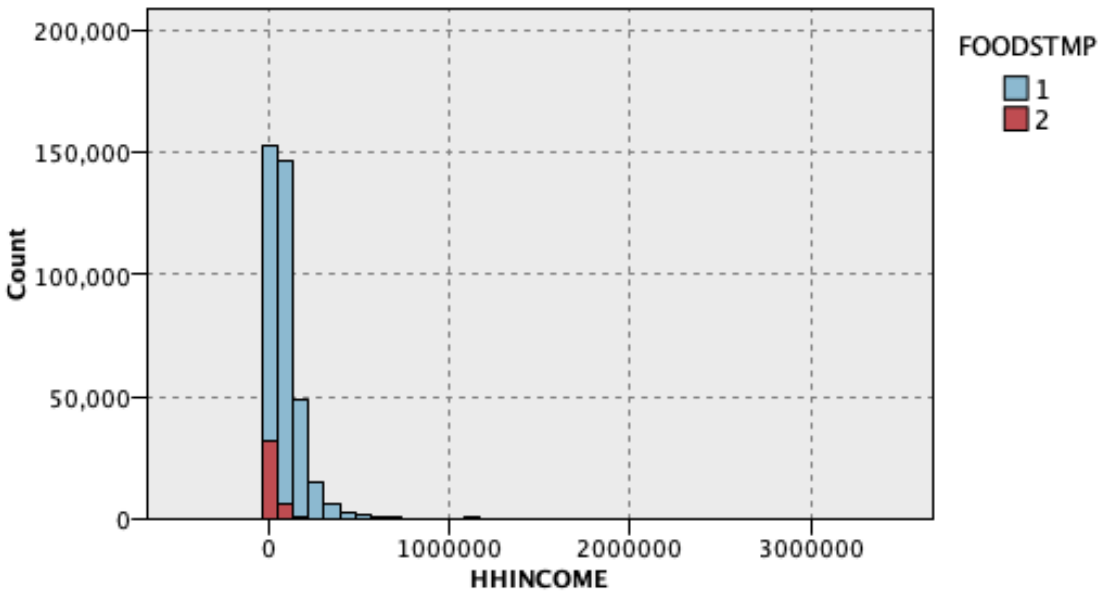

Figure 1: HHINCOME Histogram with Outliers

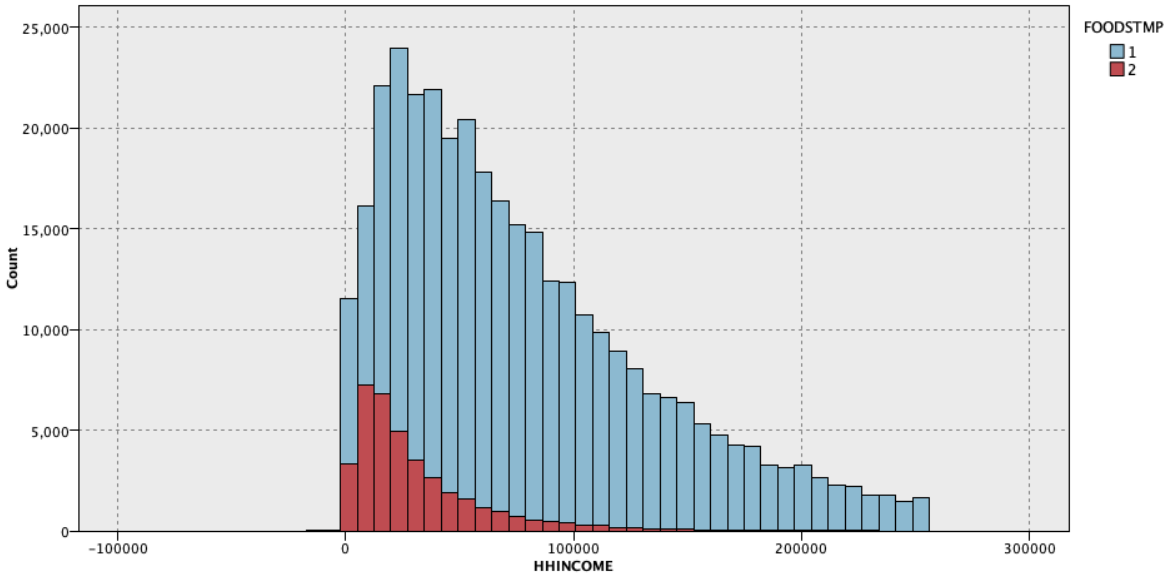

Figure 2: HHINCOME Histogram without Outliers

Missing values varied dramatically across features. The variable with the highest missingness—visa status—was missing for 97.8% of respondents, but this made sense upon inspection because the question only applied to non-citizens. For the descriptive phase, I retained variables with high missingness or outliers if they provided context about the population or survey design, though I later excluded them from predictive modeling.

The descriptive comparisons between demographic variables and SNAP recipiency revealed several notable patterns. Veterans made up 9.3% of the sample, yet only 5.9% received food stamps, compared with 11.6% of non-veterans. *Figure 3* illustrates this disparity. Gender differences were even more striking: households headed by women had nearly double the SNAP participation rate of male-headed households—14.4% vs. 7.5%—shown in *Figure 4*.

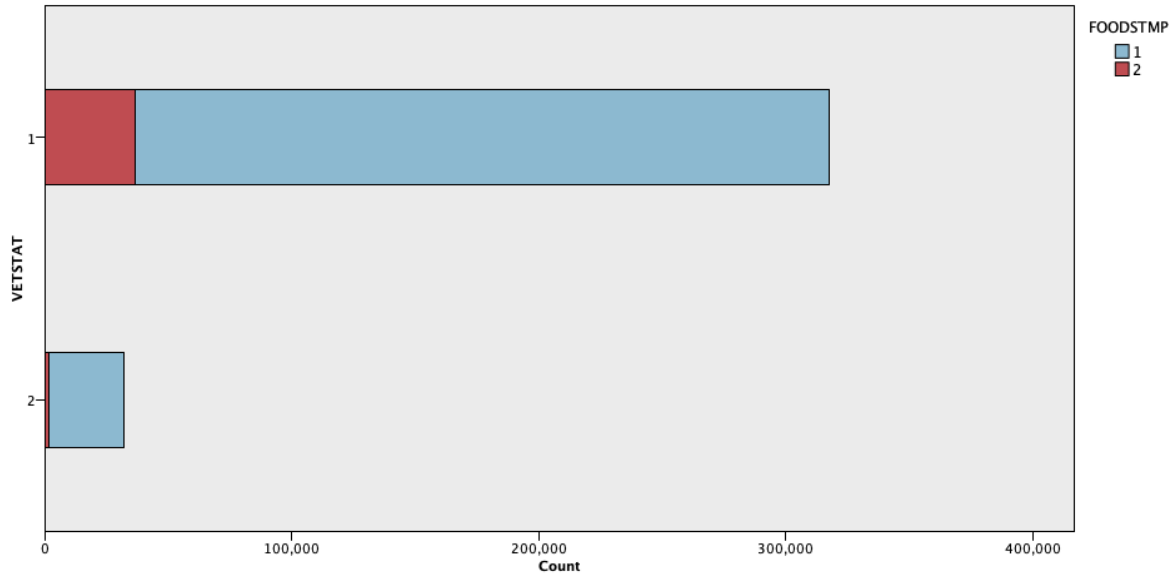

Figure 3: Veteran Status (1 = Non-Veterans, 2 = Veterans) Distribution with SNAP Recipiency Overlay (red = SNAP Recipient)

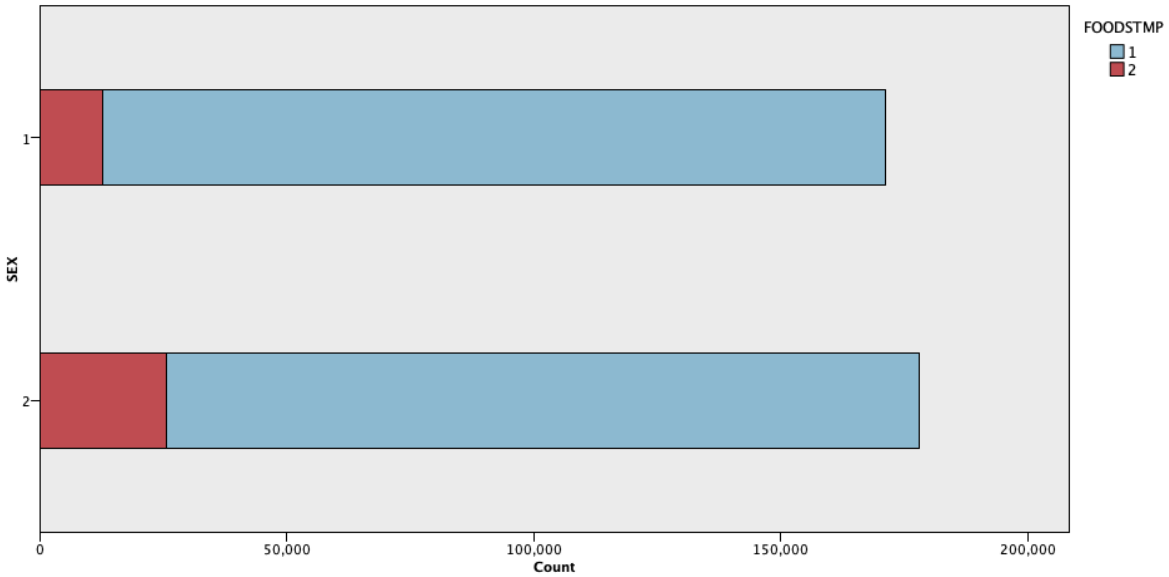

Figure 4: Sex Distribution (1 = Male, 2 = Female) with SNAP Recipiency Status Overlay (red = SNAP Recipient)

Income structure also played a meaningful role. Single-income households participated at higher rates (8.1%) than dual-income households (3.1%), a pattern displayed in *Figure 5*. Age, by contrast, showed no significant association with recipiency and was therefore not emphasized further.

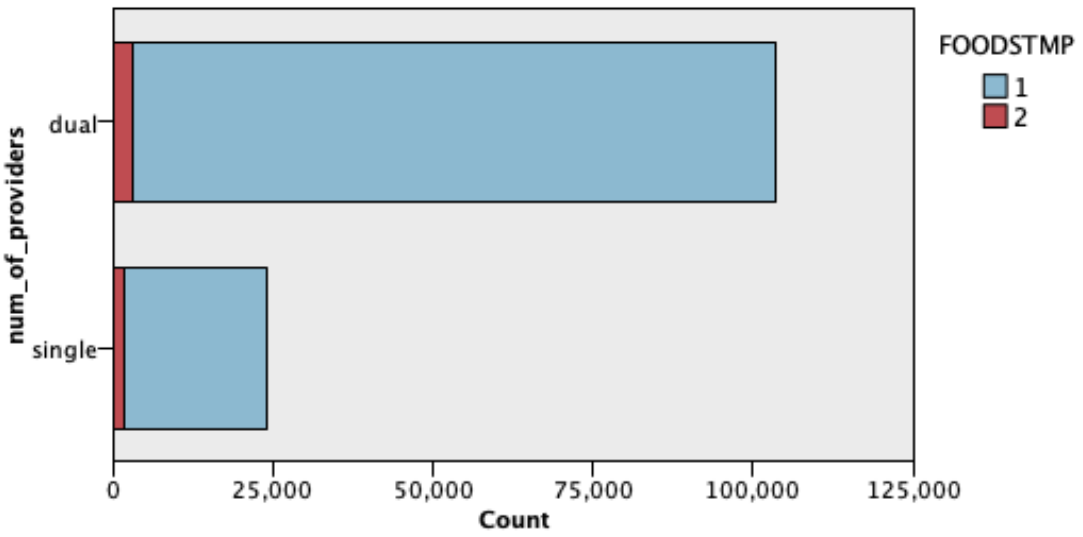

Figure 5: Number of Income Providers per Household with SNAP Recipiency Overlay (red = SNAP Recipients)

Because the number of observations varied by year, balancing was required before modeling, and *Figure 6* documents the initial year distribution. Another notable pattern involved unemployment: when the head of household had been unemployed for at least one consecutive week, SNAP recipiency rose to 23.4%, as shown in *Figure 7*.

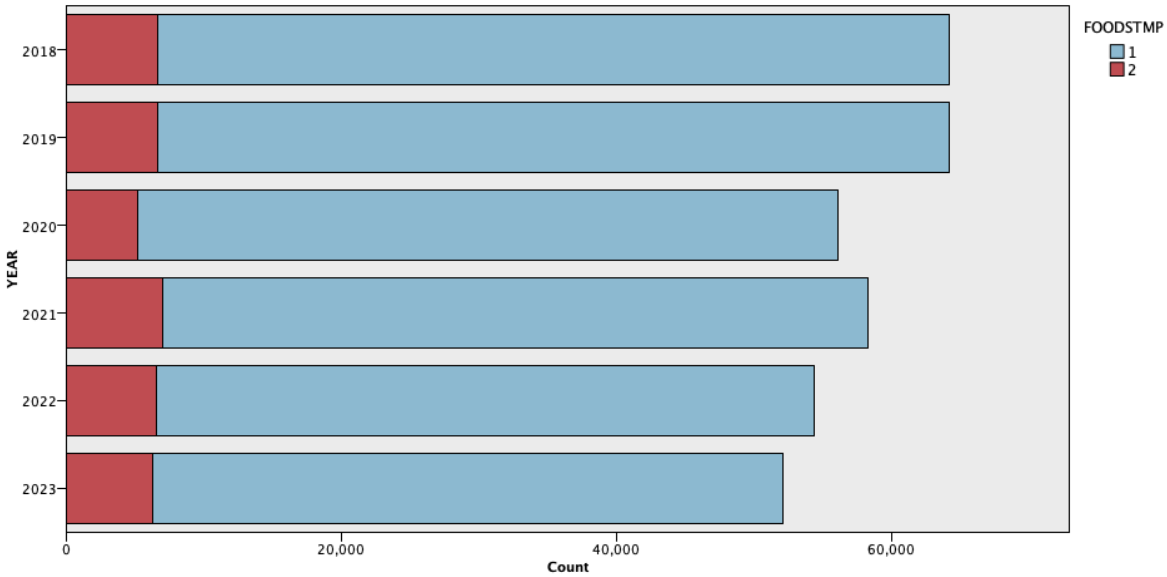

Figure 6: Distribution of Observations per Year with SNAP Recipiency Status Overlay (red = SNAP Recipients)

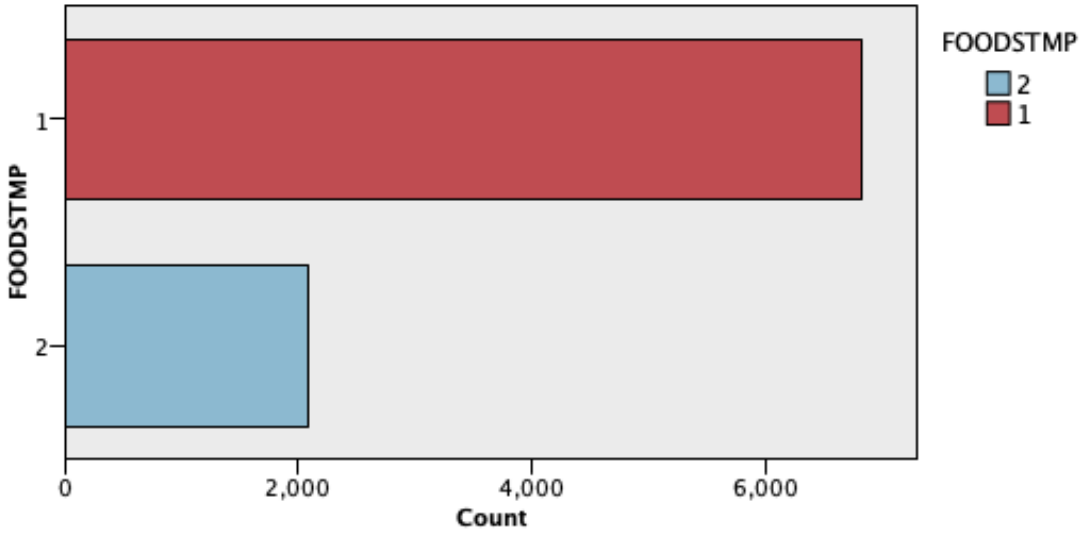

Figure 7: SNAP Recipiency Distribution for Head of the Household Unemployed More than a Week

Racial differences also emerged. SNAP participation rates were 9.3% among White householders, 7.5% among Asian/Pacific Islander householders, but substantially higher among

Black (20.1%) and American Indian/Aleut/Eskimo (21.6%) householders. *Figure 8* visualizes this contrast.

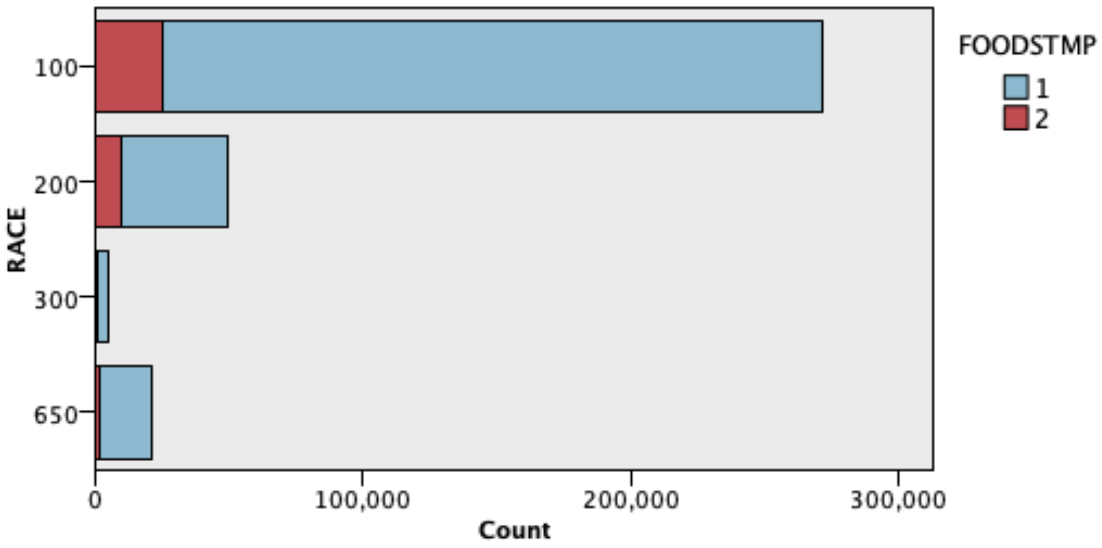

Figure 8: Distribution of Race (100 = White, 200 = Black, 300 = American Indian/Aleut/Eskimo, and 650 = Asian/Pacific Islander) With SNAP Recipiency Overlay (red = SNAP Recipients

Overall, the descriptive analysis revealed meaningful socioeconomic and demographic patterns that helped structure the modeling stage and identified several features likely to contribute predictive power.

3.3 Descriptive Analysis Discussion

The descriptive phase made clear that the dataset contained both rich information and structural limitations. Income, unemployment, number of earners, race, and gender all showed strong associations with SNAP participation. At the same time, the extreme skewness of income, the substantial missingness in certain variables, and the survey's complex demographic distributions required careful preprocessing before building any predictive models.

Some of the disparities—such as the lower SNAP participation among veterans and the unusually low rate of school lunch subsidy use among unemployed SNAP recipients—suggest potential gaps in access, awareness, or administrative coordination across federal programs. While descriptive statistics cannot establish causality, they do highlight areas where policy interventions or further research may be warranted.

4. Predictive Analysis

After the descriptive phase, I shifted to predictive modeling with the goal of identifying households likely to receive SNAP benefits. Before modeling, I addressed the data quality issues revealed earlier. Extreme income outliers were removed, racial categories were consolidated to the four most populated groups, and only complete cases for the selected predictors were retained. The dataset was further balanced across survey years to prevent time periods with larger samples from dominating the models. These steps resulted in a finalized modeling dataset containing 306,806 observations and 16 predictors.

4.1 Cluster Analysis

I began with an unsupervised approach using the Auto-Cluster node to determine whether natural groupings in the data aligned with SNAP participation. The initial models—TwoStep, KMeans, and Kohonen—produced low Silhouette scores, and even after removing weak predictors, the best Silhouette (0.384 from TwoStep) did not correspond to meaningful segmentation. SNAP recipients were nearly evenly distributed across clusters, and no cluster captured a group uniquely associated with food stamp participation. This indicated that unsupervised clustering was not well-suited to this problem.

4.2 Decision Tree

I next turned to supervised learning, beginning with a C5.0 decision tree. The first model produced an eight-layer tree with an overall accuracy of 91.46%. However, much of this accuracy came from correctly classifying non-SNAP households, which made up the majority of the sample. The predictor importance plot (*Figure 9*) showed that two variables—CAIDLY (Medicaid recipiency) and HHINCOME—dominated the model. Removing features with zero importance did not materially improve performance, and simplifying the feature set slightly reduced accuracy. Although the tree performed well overall, it struggled to correctly identify SNAP recipients.

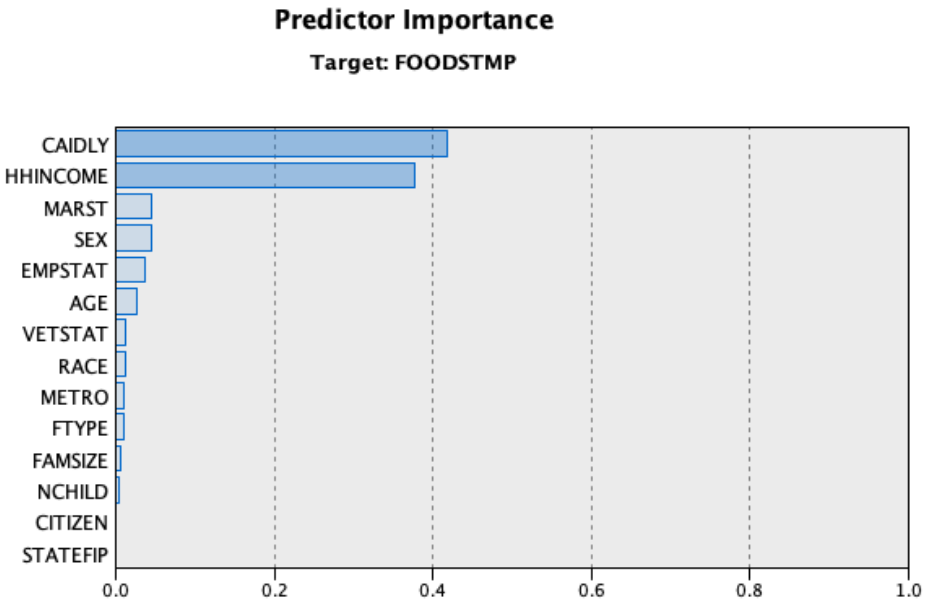

Figure 9: Predictor Importance Plot for Initial Decison Tree

4.3 Logistic Regression

Logistic regressions produced similar challenges. The initial model achieved 90.6% accuracy but correctly identified only 33.7% of SNAP recipients, reflecting the severe class imbalance. Even after selecting only the strongest predictors, sensitivity to the minority class declined slightly.

To address the imbalance directly, I used the Balance node to adjust the ratio of SNAP to non-SNAP households. When re-weighting the classes to a 75%/25% split and rerunning the full model, sensitivity increased to 57.8%, and the predicted SNAP rate came closer to the true value. Increasing the minority class further to 33.3% pushed sensitivity to 66.2%, though overall accuracy fell to 82.2%, demonstrating the inevitable trade-off between sensitivity and total accuracy. *Figures 10, 11, and 12* document these logistic regression performance changes across iterations.

Classification

	Predicted		
Observed	1	2	Percent Correct
1	266416	6415	97.6%
2	22392	11361	33.7%
Overall Percentage	94.2%	5.8%	90.6%

Figure 10: Classification Table from the Logistic Regression with the Original 89/11 Percent Class Distribution

Classification

Observed	Predicted 1	Predicted 2	Percent Correct
1	266192	6470	97.6%
2	22684	11094	32.8%
Overall Percentage	94.3%	5.7%	90.5%

Figure 11: Classification Table from the Logistic Regression with a 75/25 Percent Class Distribution

Classification

Observed	Predicted 1	Predicted 2	Percent Correct
1	94214	7095	93.0%
2	14242	19541	57.8%
Overall Percentage	80.3%	19.7%	84.2%

Figure 12: Classification Table from the Logistic Regression with a 66.7/33.3 Percent Class Distribution

4.6 Predictive Analysis Conclusion

Across all modeling approaches—unsupervised and supervised—the results pointed to the same conclusion: predicting SNAP participation from ASEC survey variables is extremely difficult. Clustering methods failed to reveal meaningful structure, decision trees relied heavily on just a few variables and offered limited minority-class performance, and logistic regression required substantial class rebalancing to achieve acceptable sensitivity. Even then, improvements in identifying SNAP households came at the expense of overall accuracy.

These challenges reflect both the underlying imbalance in SNAP participation and the limited predictive scope of the available survey variables. Household SNAP use is driven by a complex mix of economic conditions, program eligibility rules, administrative factors, and household characteristics, many of which are not fully captured in ASEC. Future work would benefit from more detailed administrative or transactional data, richer behavioral or economic features, or more advanced modeling techniques that can better handle rare event classification.

Nonetheless, the modeling exercise offered valuable insights into the complexity of predicting benefit participation and highlighted the strengths and limitations of different modeling approaches when applied to large-scale social survey data.

5. Conclusion

The results of this project highlight a clear divide between its descriptive and predictive outcomes. The descriptive analysis produced several meaningful insights regarding how household income, employment status, age structure, and family composition relate to SNAP participation. These findings offer a strong foundation for deeper policy-oriented analysis and can help shape more refined modeling efforts.

The predictive modeling stage, however, demonstrated the challenge of classifying SNAP-recipient households. While the overall accuracy of the final model was relatively high, its ability to correctly identify SNAP households was weak, indicating that the model did not successfully capture the distinction between recipients and non-recipients. This limitation reflects both class imbalance and substantial overlap in demographic characteristics across groups.

Even so, the project represents a partial success. It delivered a fully cleaned, structured dataset; a clear descriptive profile of SNAP participation; and a functional modeling framework that can be expanded or improved in future work. By clarifying which variables, modeling strategies, and data adjustments require further refinement, the project establishes a solid starting point for building more accurate predictive tools in subsequent analyses.

Citations

Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Etienne Breton, Grace Cooper, Julia A. Rivera Drew, Stephanie Richards, David Van Riper, and Kari C.W. Williams. IPUMS CPS: Version 13.0 [dataset]. Minneapolis, MN: IPUMS, 2025. https://doi.org/10.18128/D030.V13.0

All other material was learned in class.