Conclusion Write-Up

Yiying Wu (yw3996)

Note: All codes are available here: https://github.com/wuyiying2018/food-insecurity

1. Introduction

2. Methodology

2.1 Data sources

The dataset used in the report is the NHANES 2017-March 2020 dataset. This survey is conducted, and the data is collected by the National Center for Health Statistics (NCHS) from 2017 to March 2020.

The NHANES 2019-2020 field operations were suspended in March 2020 due to the COVID-19 pandemic, resulting in incomplete data collection that is not nationally representative. To address this issue and ensure national representativeness, the partially collected 2019-2020 data were merged with the complete 2017-2018 dataset, creating a combined dataset representative of the U.S. civilian non-institutionalized population before the pandemic. The dataset contains demographic, socioeconomic, dietary, and health-related information.

The target population is adults over 20 years old (includes 20) in U.S.

2.2 Weighting the Survey Data

In NHANES dataset weights are adjusted for selection probability, non-response, and post- stratification to match U.S. population figures. To obtain a valid statistical inference, a domain analysis for the adult subpopulation (age ≥ 20) are conducted by using the subset function in the survey package in R.

2.3 Multiple Imputation

To handle missing data in our NHANES dataset, we utilize Multiple Imputation (MI) to generate multiple datasets by imputing missing values repeatedly. This method retains the advantages of single imputation, such as consistent analyses and data collector's knowledge, while also accurately reflecting uncertainty and accounting for imputation error. We implemented it using the MICE (Multivariate Imputation by Chained Equations) package in R, which allows for flexible and efficient imputation of missing values. This ensures that the imputed values are plausible and improves statistical efficiency.

2.4 Logistic Regression

Given the binary outcome: food security (Yes/No), we chose logistic regression as our statistical model. Suppose there are n covariates $X_i, i = 1, \ldots n$, the model can be expressed as:

$$log(\frac{\pi}{1-\pi}) = X\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$$

- π is the risk
- $\frac{\pi}{1-\pi}$ is the odds
- β_0 is the log odds for $X_i's = 0$

- β_i is the log odds ratio per unit change of X_i, holding all other covariates fixed

In the logistic regression analysis for survey data, I have integrated multiple imputation and complex survey design in NHANES. Each imputed dataset was then individually analyzed with logistic regression models, incorporating essential survey design elements like stratification and weights via the survey package. Lastly, these results are aggregated to derive final estimates using Rubin's multiple imputation combining rule (Rubin, 2018).

2.5 Random Forest for Prediction

The function folds.svy() in the `surveyCV` package generates design-based fold IDs for K-fold CV, using any specified strata and clusters. For a stratified sample, each fold will contain data from each stratum. For a cluster sample, a given cluster's rows will all be assigned to the same fold. (Wieczorek, J., Guerin, C., & McMahon, T., 2022)

A cross-validation procedure is conducted to optimize the bin_size parameter of a Random Forest model tailored for survey data, particularly considering the clustering inherent to the dataset. The number of cross-validation folds are adjusted to the lesser of the initially intended folds or the number of unique clusters, ensuring an equitable representation of clusters across folds.

To navigate the parameter tuning, a range of bin_size values is established, and for each fold in the cross-validation, the dataset is split into training and testing sets based on fold IDs that account for the survey's clustering. The Random Forest model is then trained on these subsets with varying bin_size parameters, alongside considerations for survey weights, stratification, and clustering within the model. Predictions are generated for the test set, and the Area Under the Curve (AUC) metric is calculated for these predictions against actual outcomes, adjusted by survey weights to accurately reflect model performance across different bin_size settings.

3. Results

3.1 Descriptive Statistics

Table 1: Summary of Dataset by Food Security

Characteristic	**food insecurity**, N = 3,163[1]	**food security**, N = 5,380[1]
gender		
male	1,490 (47%)	2,639 (49%)
female	1,673 (53%)	2,741 (51%)
age	48 (33, 62)	54 (38, 67)
race		
Mexican American	507 (16%)	478 (8.9%)
Other Hispanic	446 (14%)	419 (7.8%)
Non-Hispanic White	778 (25%)	2,244 (42%)
Non-Hispanic Black	1,038 (33%)	1,224 (23%)
Non-Hispanic Asian	208 (6.6%)	798 (15%)
Other Race - Including Multi-Racial	186 (5.9%)	217 (4.0%)
not_born_in_us	997 (32%)	1,418 (26%)
Unknown	3	0
bmi	30 (25, 35)	28 (25, 33)
Unknown	216	471
education		
Less than 9th grade	402 (13%)	250 (4.6%)
9-11th grade	542 (17%)	414 (7.7%)
High school graduate/GED or equivalent	928 (29%)	1,132 (21%)
Some college or AA degree	993 (31%)	1,786 (33%)
College graduate or above	292 (9.2%)	1,796 (33%)
Unknown	6	2
marital_status		
Married/Living with Partner	1,547 (49%)	3,340 (62%)
Widowed/Divorced/Separated	842 (27%)	1,161 (22%)
Never married	769 (24%)	876 (16%)
Unknown	5	3
family_income	1.27 (0.78, 2.06)	3.28 (1.76, 5.00)
Unknown	328	387
hbp	1,252 (40%)	2,034 (38%)
Unknown	6	6
diabetes	547 (17%)	771 (14%)
Unknown	1	1
ckd	154 (4.9%)	196 (3.6%)
Unknown	5	9

insurance	2,383 (76%)	4,807 (89%)
Unknown	11	5

[1] n (%); Median (IQR)

The provided table summarizes characteristics of individuals divided into two groups based on food security status, with 3,163 facing food insecurity and 5,380 enjoying food security. A closer analysis reveals a similar gender distribution across both groups, a slightly younger median age in the food insecurity group (48 years) compared to the food security group (54 years), and marked differences in racial composition, notably with a higher percentage of Non-Hispanic Black individuals in the food insecurity group. A substantial proportion of the food insecure were not born in the U.S. (32%), which is higher than in the food secure group. Educational attainment varies significantly, with individuals having less education predominantly in the food insecurity group, while higher education levels are more common among those with food security. Marital status also differs, with a greater proportion of married individuals or those living with a partner in the food security group. Economic disparity is evident in the family income levels, with the food security group reporting higher median values. Health-wise, instances of high blood pressure and diabetes are slightly more prevalent in the food security group, although chronic kidney disease remains low in both. Finally, insurance coverage is more widespread in the food security group (89%) compared to the food insecurity group (76%), highlighting a possible link between economic stability and access to healthcare.

3.2 Regression Analysis Results

Table 2: Summary of Regression Coefficients with Odds Ratios and Confidence Intervals

coef names	**OR**	**CI (95%)**	**p value**
Intercept	0.0866	(0.0455, 0.1646)	$< .001$
gender			
male	ref	-	-
female	0.8883	(0.7938, 0.9940)	0.0389
age	1.0243	(1.0165, 1.0320)	$< .001$
race			
Mexican American	ref	-	-
Other Hispanic	0.6945	(0.4787, 1.0074)	0.0547
Non-Hispanic White	1.6626	(1.0990, 2.5154)	0.0161
Non-Hispanic Black	1.1298	(0.7313, 1.7453)	0.5824
Non-Hispanic Asian	1.7311	(1.1106, 2.6983)	0.0154
Other Race - Including Multi-Racial	0.7474	(0.3991, 1.3995)	0.3630
not born in U.S.	1.1156	(0.8023, 1.5512)	0.5156
bmi	0.9851	(0.9753, 0.9950)	0.0032
education			
Less than 9th grade	ref	-	-
9-11th grade	1.3439	(0.9765, 1.8497)	0.0697
High school graduate/GED or equivalent	1.6826	(1.2053, 2.3488)	0.0022
Some college or AA degree	2.1287	(1.6313, 2.7776)	$< .001$
College graduate or above	3.7149	(2.7390, 5.0384)	$< .001$
marital status			
Married/Living with Partner	ref	-	-
Widowed/Divorced/Separated	0.7186	(0.5344, 0.9662)	0.0287
Never married	0.8817	(0.6729, 1.1553)	0.3611
family income	2.0250	(1.8245, 2.2474)	$< .001$
high blood pressure,	0.8546	(0.6849, 1.0664)	0.1642
diabetes	0.7961	(0.5730, 1.1061)	0.1741
chronic kidney disease	0.8876	(0.5684, 1.3859)	0.5999
insurance	1.2391	(0.9339, 1.6440)	0.1372

The logistic regression analysis presented in the table identifies several significant predictors of food security, coded as 1 for food secure and 0 for food insecure individuals. Females are less likely to be food secure compared to males, with an odds ratio (OR) of 0.8882700, while age

shows a positive association, with each additional year increasing the odds of food security (OR = 1.0242605). Racial disparities are evident, with Non-Hispanic Whites more likely to be food secure (OR = 1.6626392) compared to the reference group of Mexican Americans. Education level is a strong predictor, with higher education correlating to increased food security, particularly for college graduates or above (OR = 3.7148680). Marital status impacts food security, with married individuals or those living with a partner being the reference group; widowed, divorced, or separated individuals have lower odds of food security (OR = 0.7185600). Family income also plays a significant role; higher family income are associated with increased odds of food security (OR = 2.0249656). Health conditions like high blood pressure, diabetes, and chronic kidney disease are associated with lower odds of food security. Notably, having insurance is linked to higher odds of being food secure (OR = 1.2391050). These findings highlight the multifaceted nature of food security, influenced by a complex interplay of demographic, socioeconomic, and health factors.

3.3 Random Forest Results

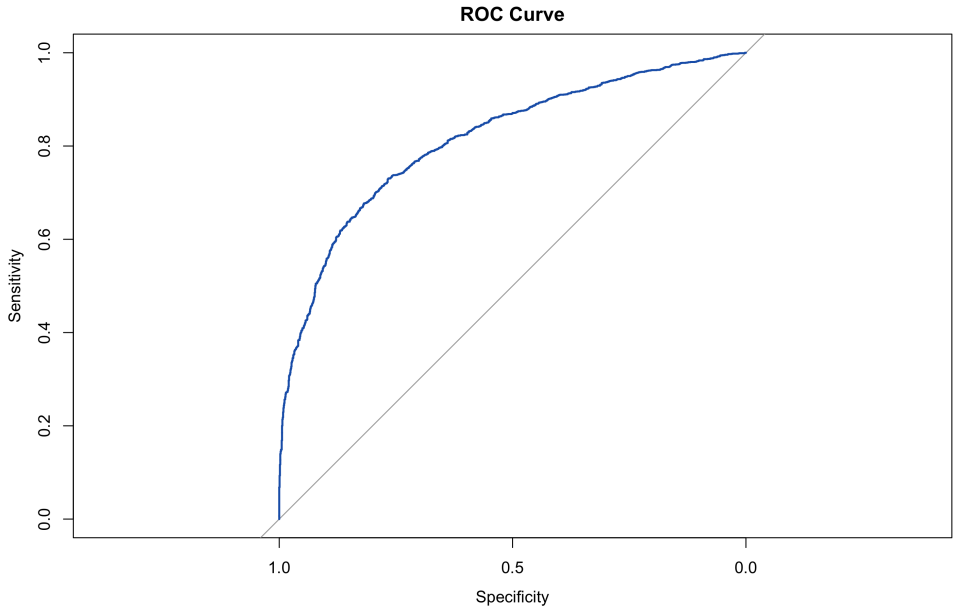
Table 3: AUC Values Across Different Bin Sizes for Random Forest Model

bin sizes	AUCs
10	0.8312
20	0.8401
50	0.8393
100	0.8400
250	0.8391
500	0.8358

From the table, it is apparent that the highest AUC value is achieved at a bin size of 20 with an AUC of 0.8401. This suggests that, of the bin sizes tested, a bin size of 20 provides the best model

performance in terms of the Area Under the Receiver Operating Characteristic (ROC) Curve, which is a common measure of the accuracy of a predictive model.

The ROC curve of the final model is



The ROC curve plot further illustrates the model's diagnostic ability. The curve shows the trade-off between sensitivity (true positive rate) and specificity (1 - false positive rate) across different thresholds. A perfect model would follow the left-hand border and then the top border of the ROC space, which would correspond to an AUC of 1. In this case, the curve is closer to the top left corner, suggesting a good level of discrimination.

The strong AUC value suggests that the combination of these variables (Gender, age, race, Country of birth, BMI, education level, marital status, poverty status, High blood pressure (HBP), diabetes, chronic kidney disease (CKD), Health Insurance) provides substantial information that the model can use to accurately identify patterns that differentiate between individuals who are food secure and those who are not.

4. Conclusion

Reference

Rubin, D.B. (2018). Flexible Imputation of Missing Data, Second Edition. Chapman and Hall/CRC.

Wieczorek, J., Guerin, C., & McMahon, T. (2022). K-fold cross-validation for complex sample surveys. Stat, 11(1). https://doi.org/10.1002/sta4.454