Open Exploration Report — STATS 415 Final Project

Michael Williams, Nikhil Dayal, Samuel Tan, Sarah Lim

I. Introduction

In the United States, there exists a diverse spectrum of economic circumstances, ranging from individuals below the poverty line to those with low incomes, and those that do not fall into either category. The disparity between income levels results from an intricate interplay of various factors, such as educational achievements, number of members in a household, and health outcomes. Thus, investigating the factors that wield substantial influence on a household's poverty level is paramount in comprehending the diverse economic landscape.

This report strives to answer two questions regarding the complexities of economic well-being: 1) which demographic and health-related predictors do not influence one's income level and 2) what the key determinants that significantly contribute to disparities in poverty status are. Our response variable will be INDFMPIR (ratio of family income to poverty) and income_level, which is a categorical variable we defined. We will explain more about our response variables in the next section.

We achieve this through the use of three different statistical methods–lasso regression, random forest trees, and cross-validation–on public demographics data collected by the CDC. Specifically, the first question will be answered via lasso regression's ability to facilitate variable selection, which determines the less influential predictors. The second question will be explored with random forest trees that help identify the most significant influences in prediction.

II. Data

The data used in this exploration is the 2017-2018 demographic information, sourced from the National Health and Nutrition Examination Survey (NHANES)[1]. Specifically, the dataset consists of a variety of variables related to participants' characteristics, such as age, gender, race/ethnicity, education, and household income. The original dataset contained several missing values, represented by '.' and information which participants were not willing or failed to disclose, represented by '9,' '7,' '99,' or '77' depending on variables; for the purpose of this exploration, all the variables with more than 1250 missing records were omitted.

Further processing the data, we converted all categorical features into factor variables, being mindful not to convert quantitative variables. We also scaled all numerical variables to have a mean of 0 and a variance of 1, because all the numerical variables in the dataset are on different scales. We also dropped variables that would have no correlation with the response, such as RIDSTATR ("Interview/Examination Status"), as these variables tend to be housekeeping data. The full list of these housekeeping variables we removed are: "SEQN",

[1] https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DEMO_J.htm#DMDHHSZE

"RIDSTATR", "RIDEXMON", "SIAPROXY", "SIAINTRP", "FIAPROXY", "FIAINTRP", "WTINT2YR", "WTMEC2YR", "SDMVPSU", and "SDMVSTRA." There is one housekeeping variables which we decided to keep, namely SIALANG, this variable represent the language that was used to carry out the survey. We believe that these two variables are representative of the interviewees' English skills, which might be relative to their income, so we decided to keep these two variables.

We also want to note that we won't be using INDHHIN2 and INDFMIN2 in our statistical analysis. Our response variables are highly dependent on these two variables, and Exploratory Data Analysis showed a very high correlation between these two variables and our response variable ratio of family income to poverty. Exploring these two variables gave us a good impression of the distribution of family and household income in our dataset, but our questions are concerned about other factors that can affect the family income. Therefore, we have removed these two variables from the dataset, as we won't be using these variables in the following sections.

To better understand the different income groups in our dataset and facilitate assigning participants into different income groups, we have also created an abstraction of the variable INDFMPIR (ratio of family income to poverty) to use as our response variable. To do this, we created another column in the dataset called income_level, and had this column take the value 0, 1, or 2 when INDFMPIR was less than or equal to 1 (poverty), between 1 and 2 (low income), and greater than 2 (sustainable income), respectively. After data cleaning and adding an extra column, we are left with 7164 observations of 17 variables.

Below are visualizations of some variables in our dataset:

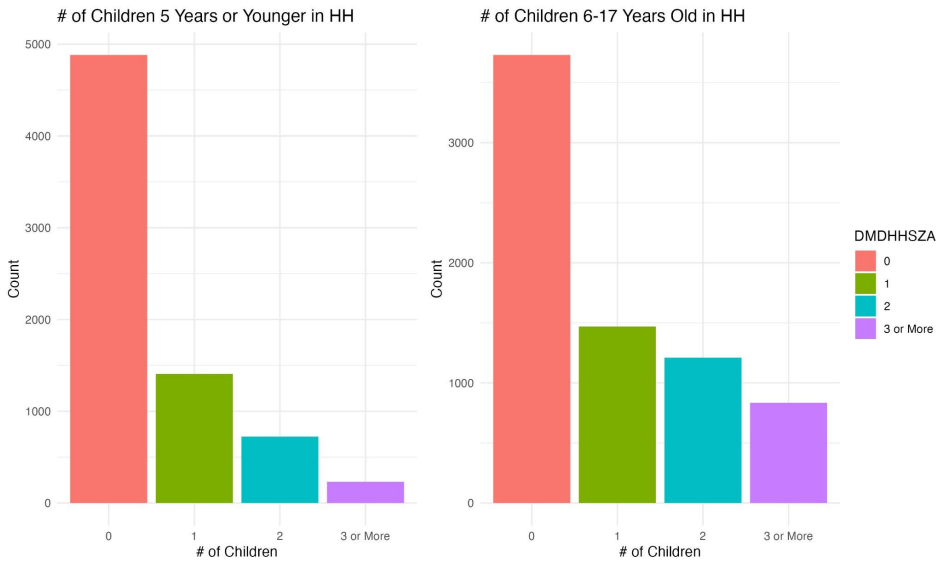

Figure 1: Number of Children Per Household, by Age Group

We can see from the plots above that most of the participants in our dataset have no children in their households, either 5 years or younger or 6-17 years old, followed by households with 1 child of 5 years or younger or 6-17 years old, then 2 children and 3 or more children. This gives us a basic overlook of the general family structure of the participants in the dataset.

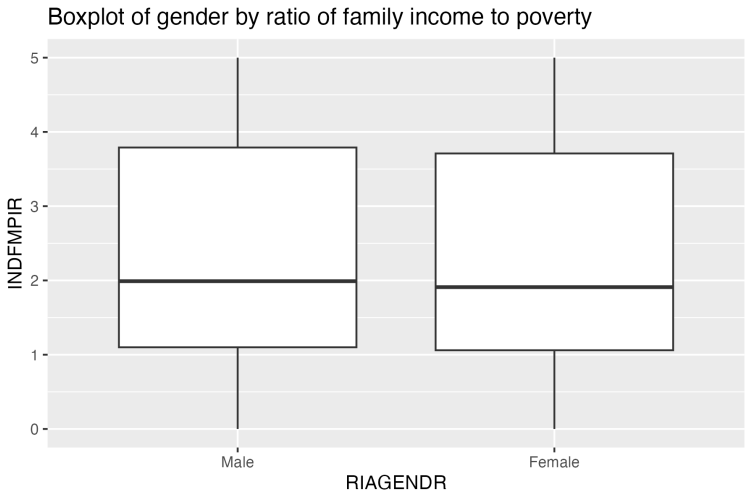

Figure 2: Boxplot of gender by ratio of family income to poverty

The presented box plot illustrates the distribution of the ratio of family income to poverty categorized by gender. Notably, the median, 1st quartile, and 3rd quartile values pertaining to the aforementioned ratio are observed to be nearly indistinguishable between the male and female groups. This parity implies that gender may not exhibit a statistically significant influence on an individual's income level within the context of this analysis. We would be exploring this relationship further in the following sections.

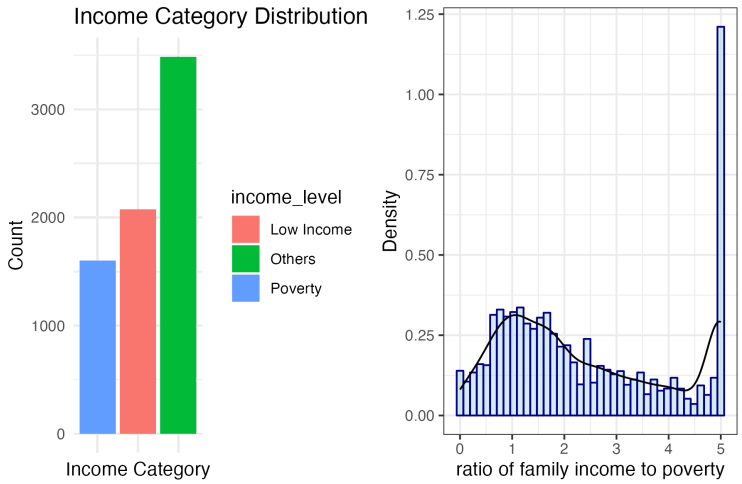

Figure 3: Income Level Distribution

The visual representation above depicts our two response variables. On the left, most participants fall into the sustainable income category, with a smaller proportion in low-income and poverty. This distribution is influenced by our definition, classifying participants with a family income-to-poverty ratio above 2.0 as sustainable income. The right-sided plot indicates a concentration of participants with a family income-to-poverty ratio of 5, as all values equal to or exceeding 5 were set to 5 in the dataset. To facilitate model fitting, we propose excluding high-income participants, defined as those with a ratio of 5 or higher, despite potential information loss. Our focus is on understanding factors impacting low-income and impoverished groups. After removing high-income participants, a slight rightward skewness is evident in the distribution of the family income-to-poverty ratio.

III. Methods

As previously mentioned, the first statistical method used in this exploration is the L1 regularization, or LASSO, which can simultaneously perform parameter estimation and variable selection by introducing a penalty term. We will be fitting LASSO with the original INDFMPIR values (excluding values that are equal to 5) as the response, and all the other variables except the variable income_level as the predictor. There are some categorical variables in our dataset, but we have changed it into factors in the data section part of the report, so dummy variables will be used to fit the LASSO model. Since the original INDFMPIR values are continuous quantitative variables, we can fit a LASSO model to predict the values. As seen in ISLR2[2], the LASSO coefficients minimize the quantity

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

The penalty term, controlled by a tuning parameter λ, induces shrinkage of some coefficients towards zero and encourages sparsity in the model. This feature facilitates variable selection, allowing us to focus on a subset of the variables with the most substantial impact in predicting the response variable, which can in turn help us to answer the questions we proposed in the beginning. Since the original dataset contains more than 15 predictors, it is important to know which variables are significant in predicting the response variable (ratio of family income to poverty) and excluding less influential variables.

Another useful statistical tool used in this exploration is random forest. By leveraging an ensemble of decision trees, random forests can effectively capture complex relationships and interactions among predictors. For the random forest, we will be using the categorical variable we created, income_level, as response, and all the other variables except INDFMPIR, the original poverty ratio values, as predictors. Recall that income_level is a categorical variable

[2] G. James, D. Witten, T. Hastie, R. Tibshirani (2013). An Introduction to Statistical Learning with Applications in R. Second Edition. Springer.

with three levels: “impoverished,” “low-income,” and “sustainable,” predicting income_level based on other variables will be a classification problem, which is the perfect fit for random forest. The nature of this method enables us to rank the predictors based on their importance in contributing to the predictive power of the model. The variables that consistently contribute to accurate predictions across different trees are assigned higher importance scores. This is what makes random forests a powerful tool, as it not only facilitates the identification of key predictors but also offers insights into the relative impact of variables on the response variable.

Cross-validation is employed to select the regularization parameter in lasso. In the context of Lasso regression, cross-validation assists in identifying the optimal λ value that balances model complexity and predictive accuracy. It allows for a comprehensive assessment of the model’s ability to generalize to new data, helping prevent overfitting and ensuring that the selected parameter is well-suited.

IV. Results

We used a 5-fold cross-validation to select a λ of 0.001754996. After running the lasso regression on the data, notably, two features had its coefficient dropped to 0, RIAGENDR2 (dummy variable for female gender) and RIDAGEYR (age of participant).

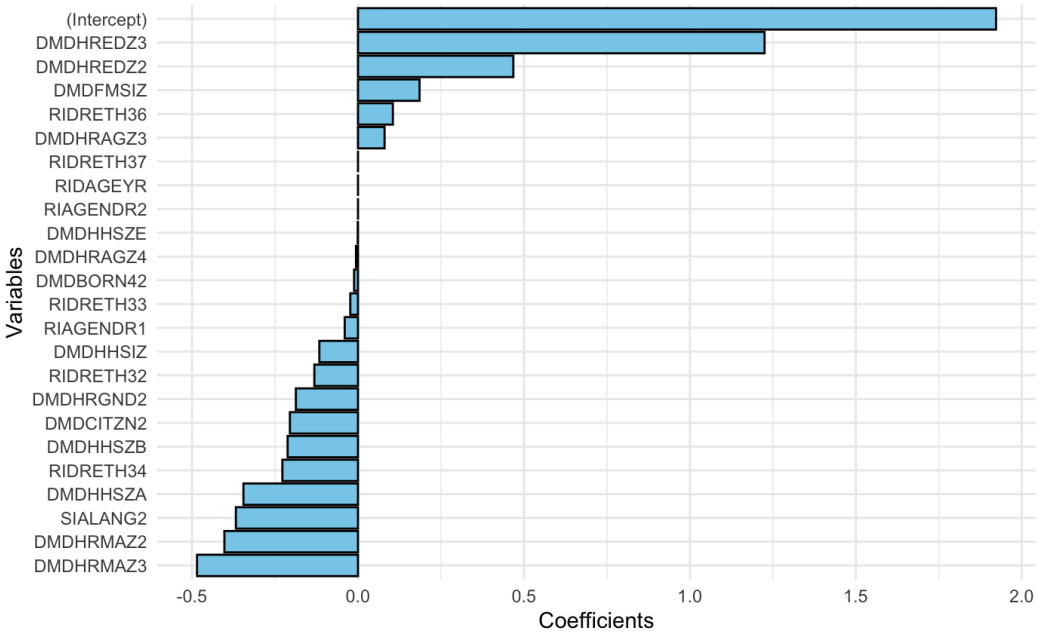

Figure 4: Penalized Coefficients for each variable in LASSO Regression

While only the dummy variable for female gender’s coefficient was set to zero, the dummy variable for male gender was also very low, which is inline with the relationship we explored in the Exploratory Data Analysis, presented as Figure 2. Thus, we chose to drop RIAGENDR entirely. We likely would have seen more features get dropped had we not cleaned

the data in our preprocessing phase. The low lambda value represents a weaker penalty, which may have hurt our feature selection as the lambda may not have been sufficiently large enough to impose a heavier penalty and filter out more features. This left us with the variables RIDAGEYR (age at time of screening), RIDRETH3 (Race/Hispanic Origin w/ Non-Hispanic Asian), DMDBORN4 (Country of Birth), DMDCITZN (Citizenship Status), SIALANG (Language of Sample Person Interview), DMDHHSIZ (Total Number of People in Household), DMDFMSIZ (Total Number of People in Family), DMDHHSZA (# of children 5 Years of Younger in Household), DMDHHSZB (# of Children 6-17 Years Old in Household), DMDHHSZE (# of Adults 60 Years or Older in Household), DMDHRGND (Household Reference Person's Gender), DMDHRAGZ (Household Reference Person's Age in Years), DMDHREDZ (Household Reference Person's Education Level), & DMDHRMAZ (Household Reference Person's Marita Status) as having the most impact on INDFMPIR.

We then ran a random forest model on the remaining variables. The figure below shows the order of variables in terms of their importance in predicting the household poverty level.

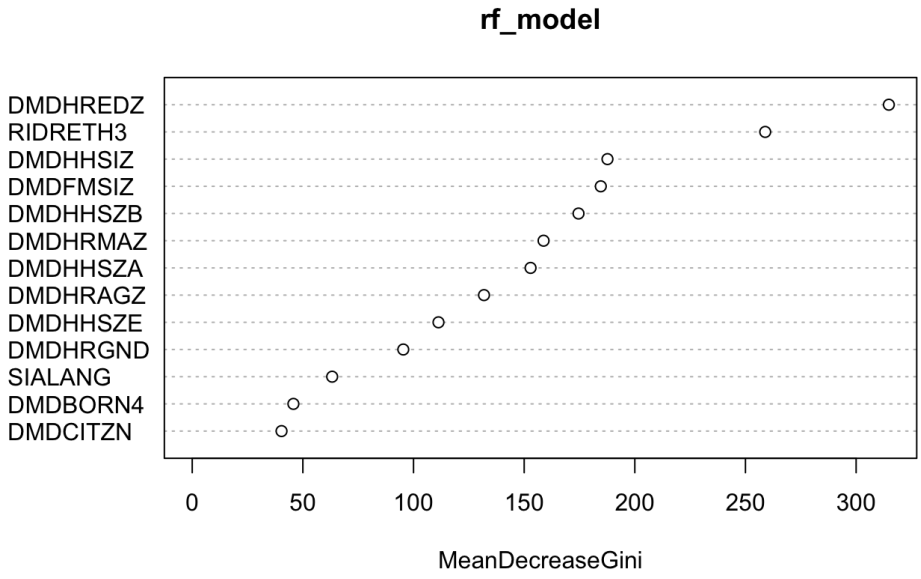

Figure 5: Rank of Variable Importance in Predicting Household Poverty Level

The most important variable selected by both the Random Forest and LASSO model is education level. It is a common perception that individuals with higher levels of education tend to have better job opportunities, hence a higher level of income. Therefore, it makes sense that one's education level has a significant impact on their poverty level. While our main goal is interpretation of the model, we are also interested in how our model performs with regards to predicting the family income status. Below is a confusion matrix, which represents the accuracy of our random forest model:

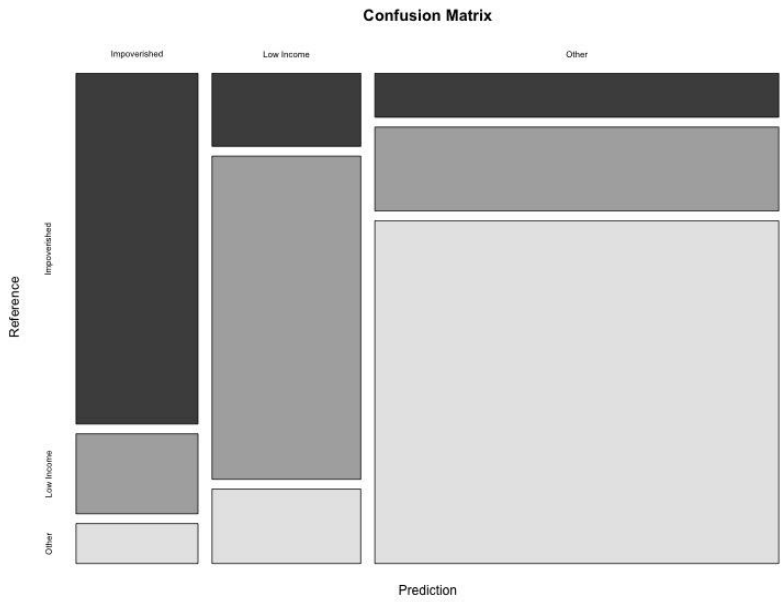

Figure 6: Random forest confusion matrix

The three large boxes positioned diagonally represent the correct predictions. Those boxes are the largest out of all 9 boxes, which means our model performed fairly well in predicting the correct household poverty level. In fact, the model showed a 0.86 accuracy in its predictions[3]. This suggests that our model (and thus its respective Gini Coefficients) are accurate.

V. Conclusion

In this comprehensive analysis, our objective was to discern potential influences on a household's poverty level. The ensuing key findings and insights derived from our meticulous analysis are succinctly summarized below.

At the outset of this report, we posited two overarching questions guiding our analysis:

1. **Identification of Non-Influential Predictors**: In addressing the first question, pertaining to demographic and health-related predictors that do not sway family income levels, we employed a LASSO model detailed in Part IV. Notably, the coefficient for RIAGENDR (Gender of the participant) and RIDAGEYR (age at time of screening) were set to zero, signifying their insignificance in predicting the Ratio of family income to poverty. Thus, we contend that gender and age emerge as demographic predictors with negligible influence on one's family income level.
2. **Key Determinants of Poverty Disparities**: The second question delved into the key determinants contributing significantly to disparities in poverty status. In the same section

[3] See attached code for calculations.

of the report (Part IV), a random forest model was fitted to predict participants' income status. The inclusion of a mean decrease Gini plot, which gauges the contribution of each variable to the homogeneity of nodes and leaves in the resulting random forest, revealed that DMDHREDZ (education level at the time of screening) scored the highest Gini values. This underscores its paramount importance in predicting family income status, suggesting that education level helps to predict disparities in poverty status within our dataset. Thus, we contend that education level contributes significantly to disparities in poverty status.

It is incumbent upon us to acknowledge the limitations intrinsic to our analysis. One such limitation arises from the dataset's collection methodology. The data, sourced from the CDC, introduced a constraint in its gender classification by utilizing a binary framework (male and female). Consequently, our insights are circumscribed, lacking nuance for individuals outside the traditional gender binary. Additionally, the temporal constraint of our data, collected during 2017-2018, necessitates consideration. The onset of the COVID-19 pandemic in 2019 renders some of our conclusions potentially obsolete.

VI. Contributions

- Michael: Fitted the model and tuned the parameters
- Nikhil: Interpreted the model and wrote out the report
- Samuel: Visualization and analysis of the data
- Sarah: Experimenting on models and wrote out the report

VII. Reproducibility

This section outlines the steps and details required to reproduce the analysis presented in this report. The analysis was conducted using R, and the primary libraries include randomforest (for random forest model), glmnet (for LASSO model), and tidyverse (for visualization of the data). The raw dataset was obtained from CDC's "Demographic Variables and Sample Weights 2017-201", and data cleaning was performed as described in the "Data" section above.

To achieve result reproducibility, a random seed of 42 was used throughout the analysis. Following the outlined steps in the "Methods" section above will produce the same key results, such as the best lambda for lasso chosen by cross validation and the results of random forest. The complete code is in an attached rmd file for easy replication. Running the rmd file should reproduce the results mentioned earlier, ensuring the report's reproducibility.