Logistic Regression Analysis of Variables in the Household Pulse Survey

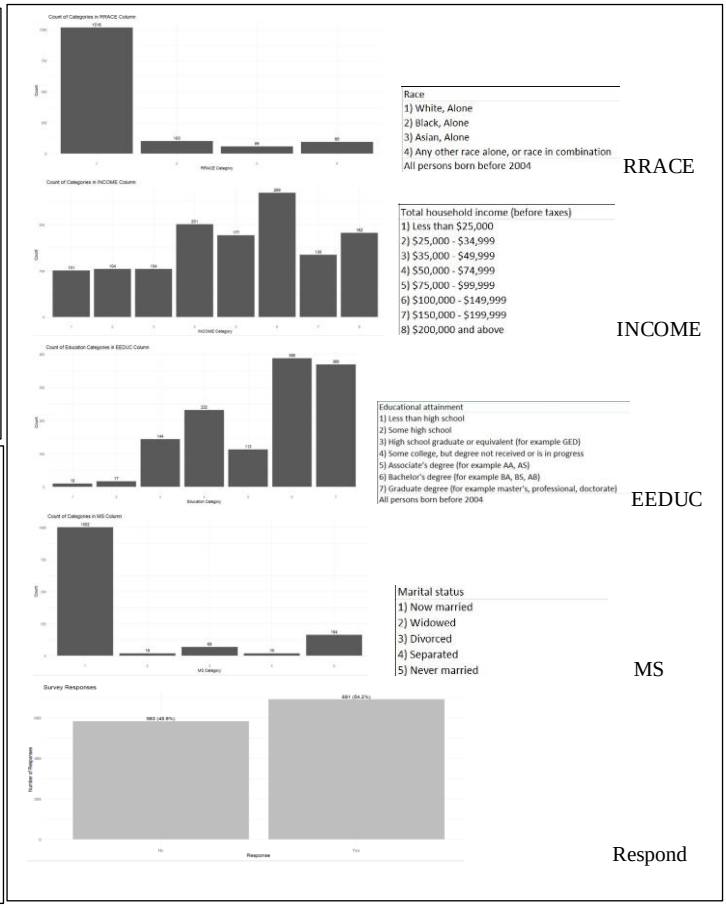
Heather Mello

Summary:

The project investigates whether race, income, education level, and marital status significantly indicate whether a household was affected by the infant formula shortage. Using logistic regression analysis, the study focuses on these variables with data preparation including cleaning and segregating into training and testing subsets. Results reveal that these factors are not strong predictors of impact from the shortage, with a model accuracy of 59.9% and an AUC of 61.32%. The study highlights limitations like potential biases and the need for a more diverse dataset, underscoring the complexity of real-world scenarios.

Introduction:

This project explores whether socio-demographic factors such as race, income, education level, and marital status can predict the impact of the infant formula shortage on households. Motivated by the critical need to understand and mitigate the effects of such shortages, especially on vulnerable populations. The research employs logistic regression analysis. This data science approach is chosen for its effectiveness in handling categorical data and its capability to provide insights into the influence of multiple independent variables on a binary outcome. The methodology includes rigorous data preparation, model training, and testing, ensuring a comprehensive analysis of the dataset's implications.



Planning:

Logistic regression analysis was the chosen machine learning algorithm because it is able to predict a binary output. In this case, the variable we wanted to analyze was whether a household was affected by the infant formula shortage, which is a binary response. The four explanatory variables that were chosen are as follows:

- Race (Categorical Nominal)
- Income (Categorical Ordinal)
- Education (Categorical Ordinal)
- Marital Status (Categorical Nominal).

Once variables were selected, the data was cleaned and prepared for analysis. All columns containing variables that were unrelated to the analysis were removed. All rows of data that had unanswered questions were also removed, resulting in 1273 responses to analyze. In addition to this, since the binary response variable (FRMLA_AFFCT) used 1/2 for yes/no, these values were changed to 0/1. The quality of the data was also assessed. It was determined that the source was credible and reliable, and the data was collected in an ethical manner; this is because the data was collected by the U.S. Census Bureau. Once the data was prepared for analysis, an exploratory data analysis was performed. The following insights were found:

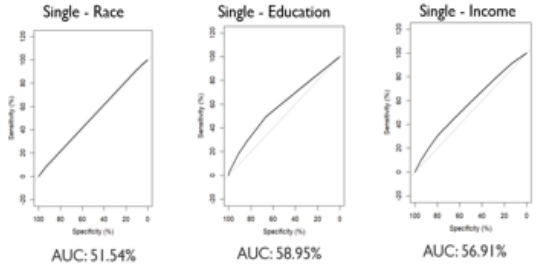
- For RRACE category, the majority identify as "White, Alone," outnumbering other groups significantly. This implies a lack of racial diversity in the dataset or a focus on predominantly White areas.
- For INCOME category, responses span various income levels, but most respondents report household incomes between $100,000 - $149,999, suggesting a predominantly high-income demographic in the survey.
- For EEDUC category, the highest number of respondents have a bachelor's degree, followed by those with a graduate degree. This suggests that the survey population is relatively well-educated.
- For MS category, For Marital Status Distribution: The "Now married" category has the highest count, which might reflect the demographic's life stage or societal norms within the population.
- For Responded category, shows a fairly even split in opinions, with a slight majority leaning towards "Yes." This information is valuable for gauging the overall sentiment or stance of a group on a particular issue or question.

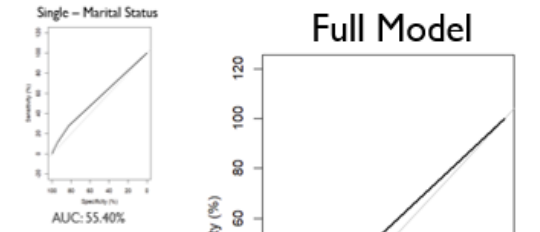
Implementation

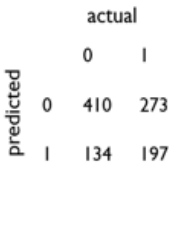
First, the data was segmented into training and testing data with an 80/20 split, respectively. Because race and marital status are categorical nominal variables, they both needed to be defined as factor/categorical variables in R using the as.factor() function. This was not necessary for income and education level since both of these values can be ordered. Next, single regression analyses were performed for each of the explanatory variables using the glm() function in R. Through the single logistic regression model for race, it was determined that race is not a statistically significant factor in the model since the reported p-values were all greater than 0.05. However, all other explanatory variables were deemed significant, so only race was excluded from the final model. Finally, the full logistic regression model was created with marital status, income, and education as the explanatory variables. An example of the R code and raw output can be seen below. This same process was repeated for each of the logistic regression models.

```r
#individual logistic regression for race, including ROC  and AUC
m1 = glm(train$FRMLA_AFFCT~race, family = "binomial", data=train)
summary(m1)
p1 = predict(m1, data='train', type = 'response')
head(p1)
hist(p1)
r = multiclass.roc(train$FRMLA_AFFCT, p1, data = 'train', percent = TRUE)
roc = r[['rocs']]
r1 = roc[[1]]
plot.roc(r1)
auc = auc(r1)
auc
```

```
Call:
glm(formula = train$FRMLA_AFFCT ~ race, family = "binomial",
    data = train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.14648    0.07059  -2.075    0.038 *
race2        0.22652    0.24167   0.937    0.349
race3       -0.24298    0.29129  -0.834    0.404
race4       -0.05419    0.23556  -0.230    0.818
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1400.3  on 1013  degrees of freedom
Residual deviance: 1398.5  on 1010  degrees of freedom
AIC: 1406.5

Number of Fisher Scoring iterations: 3
```







Validation

To validate the model, the ROC curve was plotted for each variable, and their respective AUC values were calculated. The ROC curves for each variable can be seen to the left. The ROC curves provide another example of why race was not included in the final model, since the individual regression analysis model is only 1% better at predicting than guessing at random. The ROC curve for the full logistic regression model can be seen to the left, along with the confusion matrix. The values in the confusion matrix indicate that the model is 59.9% accurate, and the AUC for the full model is 61.32%.

Interpretation

The logistic regression model used in this study had an accuracy of 59.9% and an AUC of 61.32%, indicating a moderate predictive capability. These results suggest that addressing the formula shortage requires a more nuanced understanding of the contributing factors, possibly including variables not considered in this study. The findings also highlight the complexity of real-world scenarios. The logistic regression model used is relatively simple. The limitations of using certain socio-demographic data for predictive analysis. Also, it can related dataset might not be large or diverse enough to generalize the findings.

Learning Take-aways

Key lessons include the importance of considering potential biases in data collection and the limitations of a model's predictive power. The process involved selecting relevant variables, preparing data, performing exploratory analysis, and validating the model. A significant insight was the need for a more diverse dataset, especially in racial representation. Additionally, a deeper exploration of data biases and their impact on results would be beneficial.

Conclusion

This study concludes that socio-demographic factors like race, income, education, and marital status are not significant predictors of households affected by the infant formula shortage. This finding underscores the complexity of such crises and the limitations of using limited socio-demographic data for predictive analysis, highlighting the need for more diverse data in addressing real-world problems.