# Supervised Learning for a Proxy Economic Harm Outcome

Shaine Leibowitz
Community Insight & Impact

# Background: Community Vulnerability Core Metric

- The Community Vulnerability Index (CVI) aggregates data on the core metric, economic harm, impacted by the coronavirus pandemic
- Metric is a weighted combination of quantile-normalized variables
- Construction metric, including which variables to include and how to weight them relative to each other, was informed by a detailed review of relevant public health, social science, and urban planning literature
- The economic harm metric measures a community's risk of severe, negative economic impact due to COVID-19
- Provided at the county level for the entire United States

# Motivation

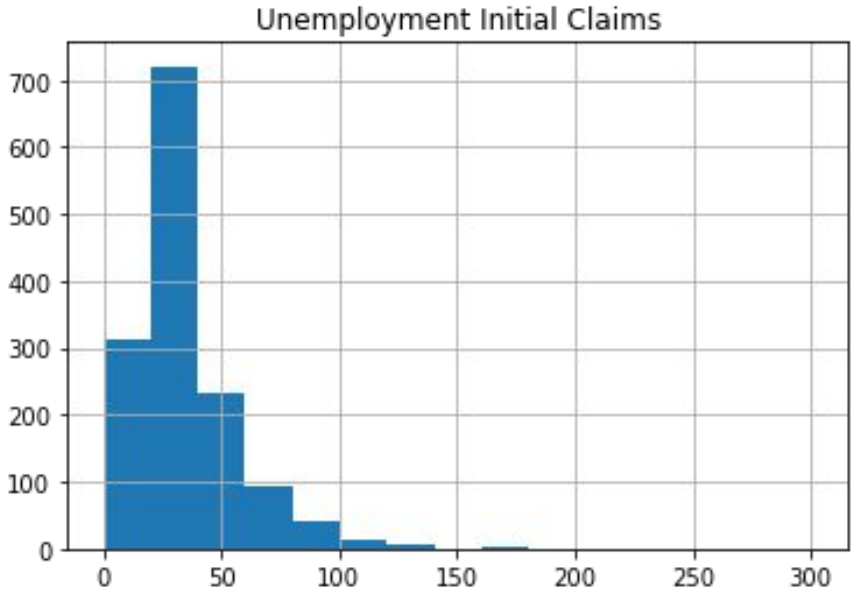
- Sought to assess our current feature weights, quantify the predictive power of the included variables, and discover any information gaps in the initial metric construction
- Inform the next iteration of CVI Economic Harm metric
- Implemented supervised learning models to predict two proxy economic harm outcomes: Number of unemployment initial claims per 100 people*
    - Measure of individual economic status
    - Differs from the pre-covid unemployment rates included in the original metric to estimate existing economic precarity in the county
- See related code in GitHub here

*accompanying paper and the Economic Tracker at https://tracktherecovery.org

# Original Features

| | | |
|---|---|---|
| % Below Poverty | Median Household Income | % No College Degree |
| Unemployment Rate | % Not in Labor Force | % Jobs in Leisure and Hospitality |
| % Part-time | % Self-Employed | |

# Distribution of Proxy across Counties



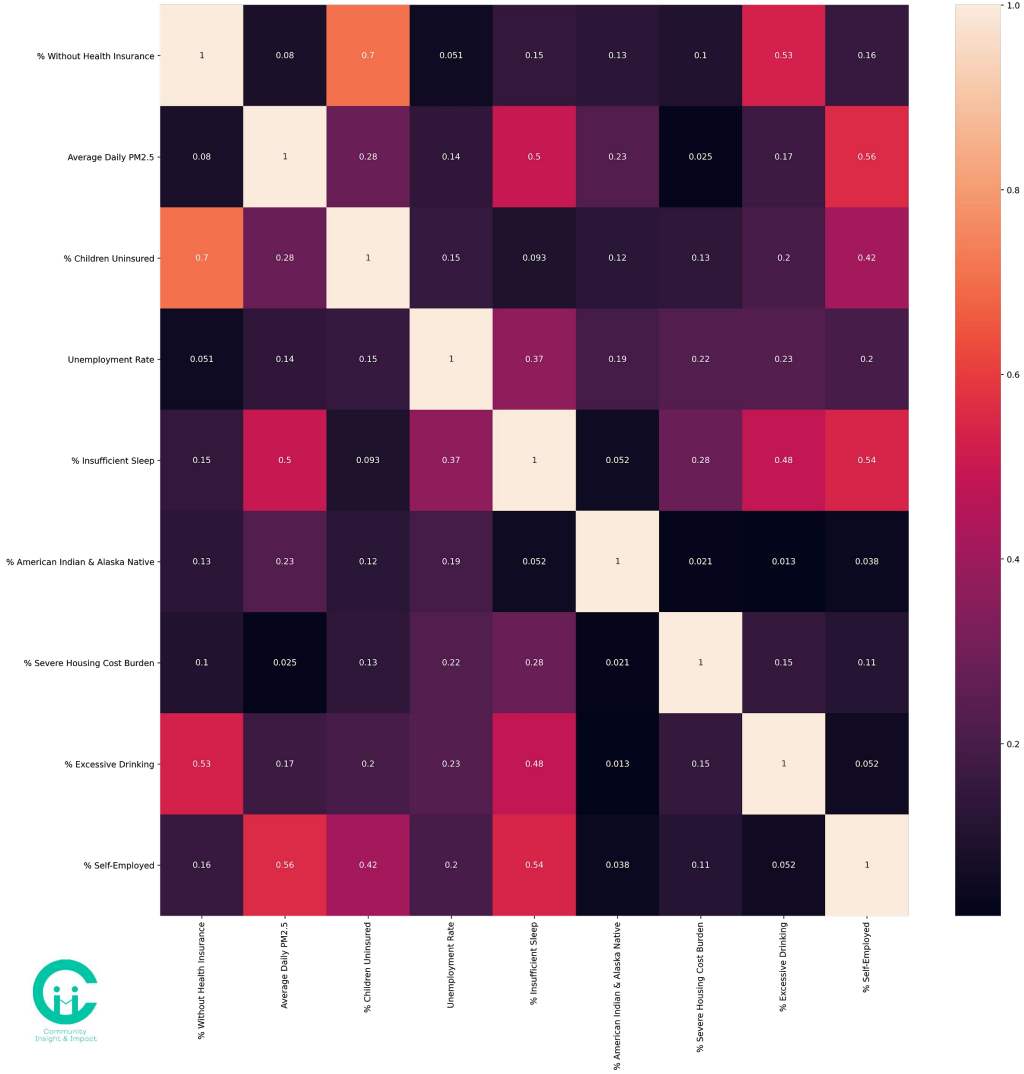
# XGBoost Most Important Features

- Starting with the full CVI dataset, we narrowed our comparison set by selecting most important features according to XGBoost (F-score > 350) which underwent hyperparameter tuning with k=5 cross-validation

| | | |
|---|---|---|
| % Without Health Insurance | % Insufficient Sleep | Average Daily PM2.5 |
| Unemployment Rate | % Children Uninsured | % Severe Housing Cost Burden |
| % American Indian & Alaska Native | % Excessive Drinking | % Self-Employed |

# Correlation: XGBoost Most Important Features

'% Children Uninsured' was removed from the XGBoost most important features due to its unsurprisingly high correlation with '% Without Health Insurance'



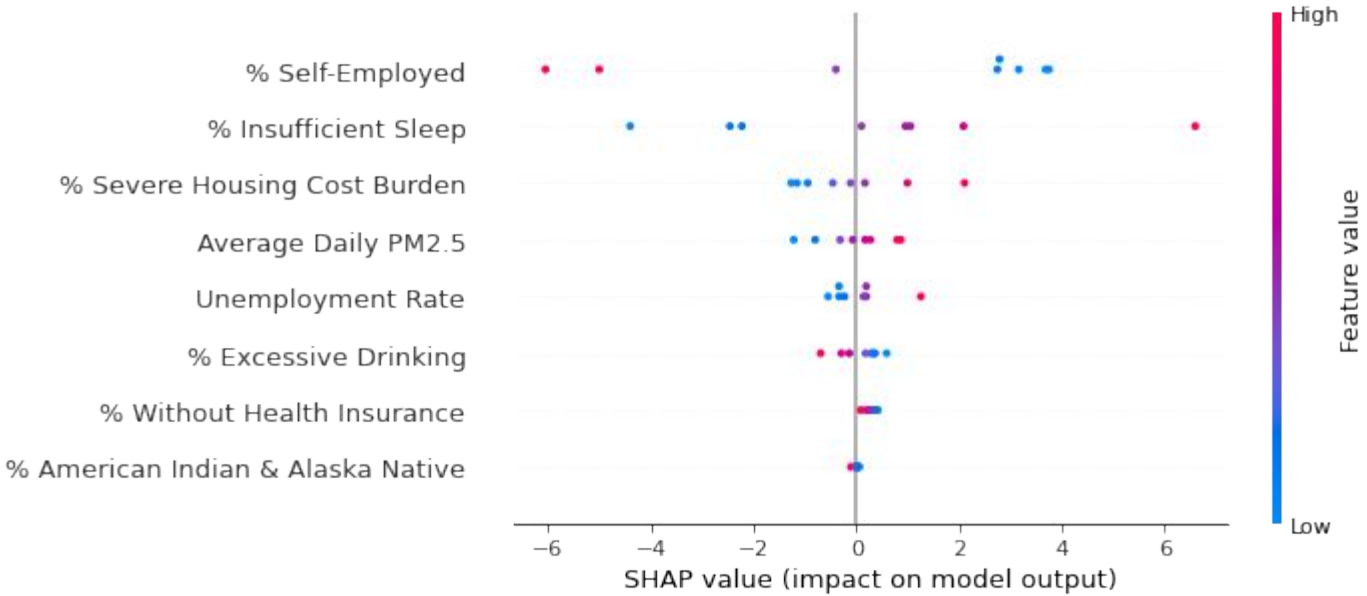
# Comparison of Different Feature Sets

With a Root Mean Squared Error (RMSE) of 14.7 on the test data set, the XGBoost Most Important Features performed better than the original features as its RMSE was 22.2. Multilayer Perceptrons were applied for these models.

| Dummy Baseline | Original Features | XGBoost Most Important Features |
|---|---|---|
| 338.7 | 22.2 | **14.7** |

# SHAP Feature Importance

From the XGBoost Most Important Features, the SHAP values demonstrate which features were the most important according to the trained Multilayer Perceptron



# Discussion: '% Self Employed'

- Only two of the original metric features: 'Unemployment Rate' and '% Self Employed' were identified by XGBoost
- Both have impactful SHAP scores, providing initial indication that they should be highly weighted in the Economic Harm metric
- Interestingly though, '% Self Employed' is inversely correlated with the proxy outcome
- We postulate that this is because although self employed workers were able to file for unemployment after the CARES Act was passed in March 2020, many who continued to work but with reduced hours were instead eligible for a different type of aid: Self-Employment Income Support Scheme grants which is not captured by the proxy outcome variable
    - This demonstrates a shortcoming of our selected proxy variable as it is well documented that self employed individuals were disproportionately impacted by COVID-19

# Discussion: '% Severe Housing Cost Burden'

- Identified by XGBoost
- '% Severe Housing Cost Burden' (the percentage of households in a county paying more than 50% of their income on housing makes sense as a possible causal predictor of economic harm
- Households with severe housing cost burdens are more likely to forgo healthcare and are less likely to have savings or emergency funds, making them more vulnerable to the economic and health impacts of the COVID-19 pandemic
- We will consider including '% Severe Housing Cost Burden' in future iterations of the Economic Harm metric

# Discussion: '% Insufficient sleep'

- Identified by XGBoost
- Insufficient sleep has been previously studied as an economic indicator due to the impact lack of sleep has on school and labor market success and public health
  - One study estimates that insufficient sleep amongst the US working population costs the economy up to $411 billion per year due to decreased productivity and missed workdays
- Additional analysis of how sleep behavior has changed during the COVID-19 pandemic and the impact of insufficient sleep on local economies is necessary to decide if '% Insufficient Sleep' should be included in the Economic Harm metric
- Including it with the original metric features did not substantially reduce the RMSE of the supervised model

# Discussion: 'Average Daily PM2.5'

- Although air pollution, specifically the average daily density of fine particulate matter has been causally connected to decreased lung function and adverse pulmonary effects, and is known to increase premature death risk, we are unable to find a possible causal link between this variable and local economic impact
- Will not be included in future iterations of the Economic Harm metric

Last Updated: 05/18/2021