Supervised Learning for a Proxy Severity Outcome

Shaine Leibowitz
Community Insight & Impact

Background: Community Vulnerability Core Metric

- The Community Vulnerability Index (CVI) aggregates data on the core metric, case severity, 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 severity metric measures the risk of hospitalization for COVID-19
- Provided at the county level for the entire United States

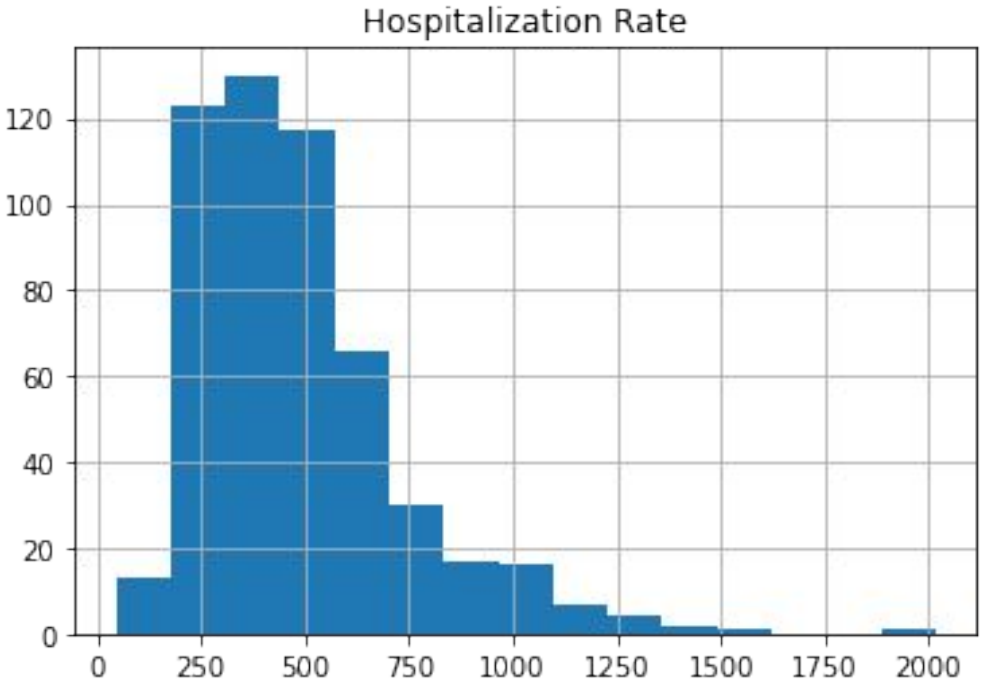
Motivation

- 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 Severity metric
- Implemented a supervised learning model to predict a proxy severity outcome: Hospitalization Rate per 100,000
    - Preferred base indicator for the Severity metric
- See related code in GitHub here

Proxy Outcome: Hospitalization Rate per 100,000

- The proxy outcome is the COVID-19 Hospitalization Rate per 100,000 population
- The hospitalizations were measured cumulatively from March 2020 to April 2021
- The following are the states from which we were able to collect data and links out to their data sources: Florida, Georgia, Tennessee, Virginia, and Wisconsin
    - These 5 states totaled 527 counties

Distribution of Proxy Outcome

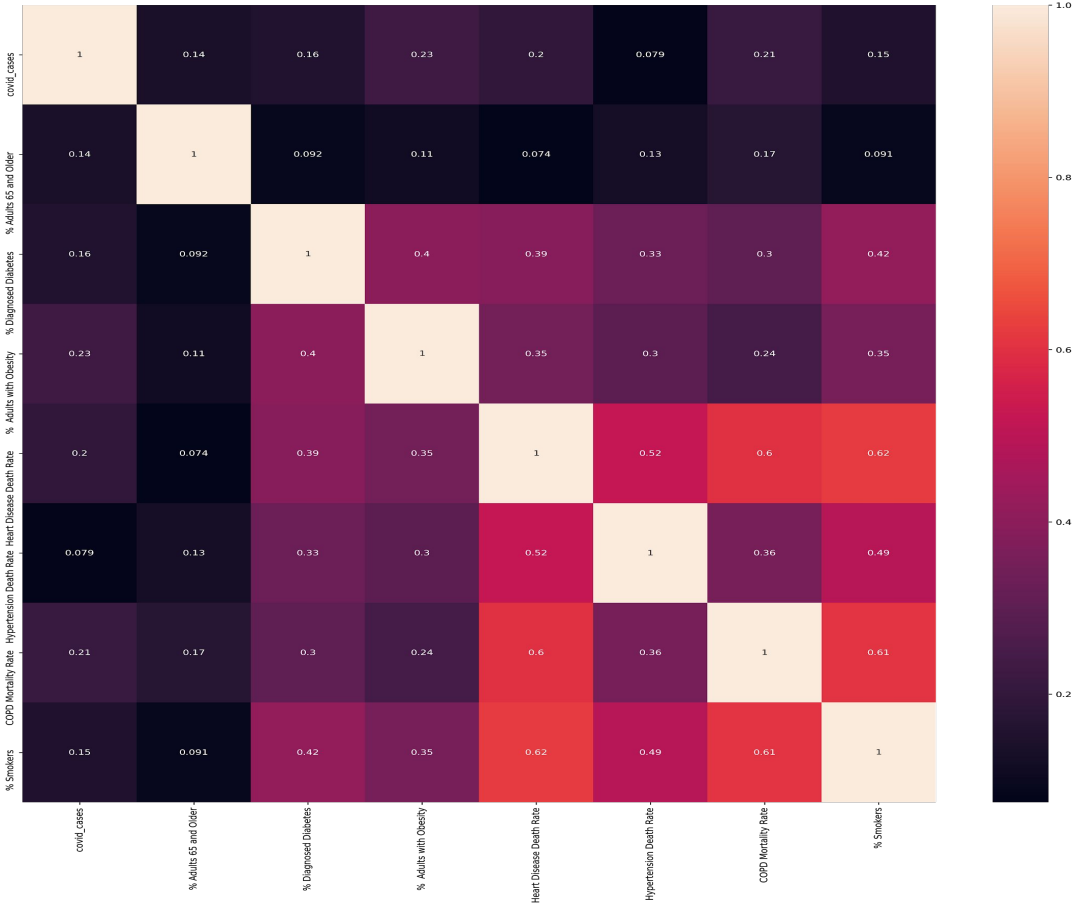


Original Features

Covid Cases	% Adults 65 and Older	% Diagnosed Diabetes
% Adults with Obesity	Heart Disease Death Rate	Hypertension Death Rate
COPD Mortality Rate	% Smokers	

Correlation: Original Features

- Due to a high correlation with three other variables, we considered removing '% Smokers"
- Chose not to as the removal did not improve performance on the validation data



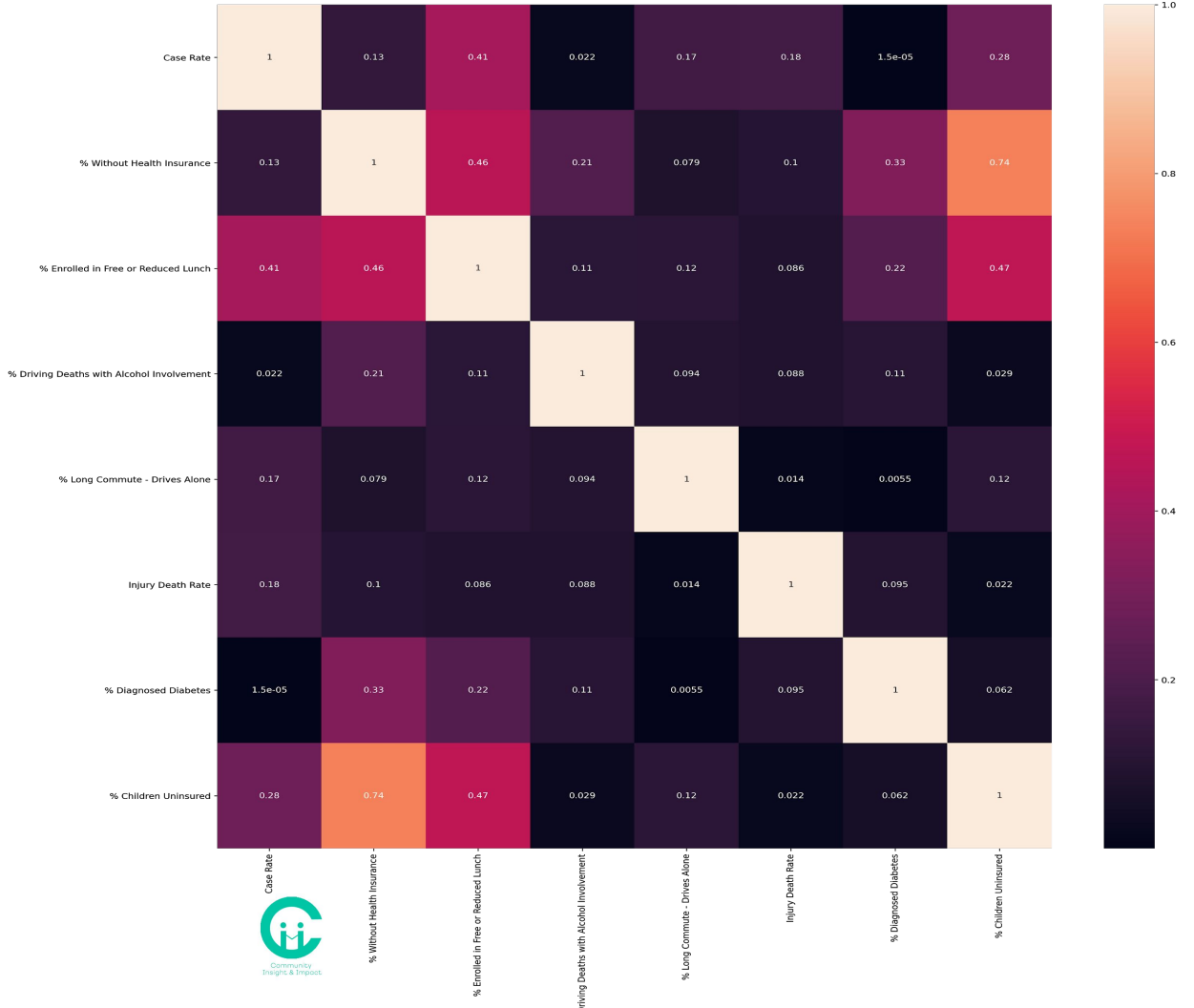
XGBoost Most Important Features

- Due to accessibility and accuracy constraints, the analysis was limited to 5 states totaling 527 counties
- 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

Case Rate	% Without Health Insurance	% Enrolled in Free or Reduced Lunch
% Driving Deaths with Alcohol Involvement	% Long Commute - Drives Alone	Injury Death Rate
% Diagnosed Diabetes		% Children Uninsured

Correlation: XGBoost Most Important Features

- Removed ‘% Children Uninsured’ due to its unsurprisingly high correlation with '% Without Health Insurance' and '% Enrolled in Free or Reduced Lunch'



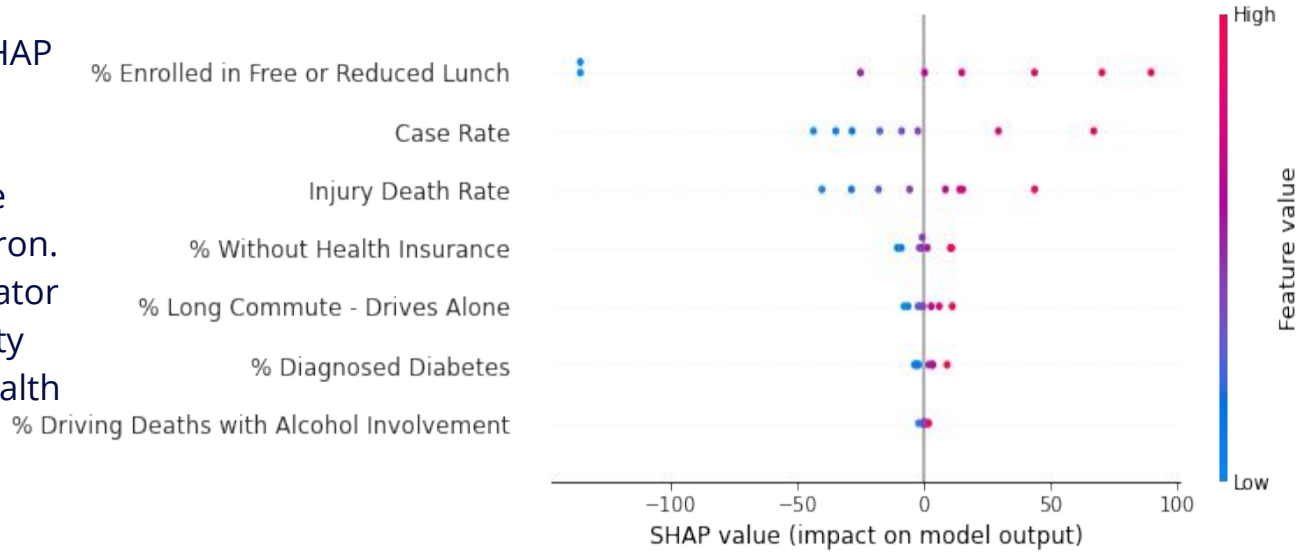
Comparison of Different Feature Sets

With a Root Mean Squared Error (RMSE) of 183.1 on the test data set, the XGBoost Most Important Features performed better than the original features as its RMSE was 228.2. Multilayer Perceptrons were applied for these models.

Dummy Baseline	Original Features	XGBoost Most Important Features
2358.6	228.2	**183.1**

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.
- The top feature is an indicator of poverty. Those in poverty are more likely to delay health care which would result in more severe cases.



Comparison of Different Feature Sets

The RMSE on the test data set decreases by simply adding "% Enrolled in Free or Reduced Lunch".

Dummy Baseline	Original Features	XGBoost Most Important Features	Original Features + "% Enrolled in Free or Reduced Lunch"
2358.6	228.2	183.1	198.3

How do the original feature set perform with indicators of poverty other than '% Enrolled in Free or Reduced Lunch'?

Original Features	Original Features + "% Enrolled in Free or Reduced Lunch"	Original Features + “% Below Poverty”	Original Features + "Unemployment Rate"	Original Features + "% Children in Poverty"
228.2	198.3	221.7	231.9	229.6

Discussion

- Public health policy and community health measures vary widely across geographic regions
  - Plan to seek additional data sources and scale up this study before making final adjustments to the Severity metric
  - Used a limited dataset for this initial model with only 5 states, all located in the South or Midwest
- Only two of the original metric features, ‘Case Rate’ and ‘% Diagnosed Diabetes’, were identified by XGBoost
  - ‘Case Rate’ is the most powerful predictor of severe COVID-19 cases in the supervised model suggesting the prevalence of COVID-19 in an area is more important than the prevalence of pre-existing comorbidities
  - Consider increasing the relative weight of ‘Case Rate’ in future iterations of the Severity metric
- Interestingly, only one of the well-studied COVID-19 comorbidities, ‘% Diagnosed Diabetes’ is identified by XGBoost
- ‘% Without Health Insurance’ was identified and studies have demonstrated an association between lack of insurance and increased mortality or disease severity due to the other prominent COVID-19 comorbidities: heart disease, hypertension, and COPD
  - Including ‘% Without Health Insurance’ in the Severity metric captures additional relevant information: disease prevalence combined with access to healthcare are indicative of severe disease and mortality

Discussion: '% Enrolled in Free and Reduced Lunch'

- Initially, we postulated that this served as a measurement of poverty, as higher rates of COVID-19 deaths are associated with poorer counties
- However, including other traditional measurements of poverty did not reduce the RMSE
  - Including '% Children in Poverty' did not reduce the RMSE despite focusing on the child recipients of Free and Reduced Lunch programs
- In addition to being an indicator of poverty, Free and Reduced Lunch programs are an effort to mitigate childhood food insecurity
  - Often indicative of family-level food insecurity
    - Not always associated with living below the poverty line and can have additional infrastructure and food access causes
- Amongst adults, food insecurity is associated with increased rates of and complications due to chronic diseases (including many comorbidities of COVID- 19)
- Conclude that the variable '% Enrolled in Free and Reduced Lunch' captures important information related to the causes and severity of key COVID-19 comorbidities and other relevant health impacts
  - Will include it in future iterations of the Severity metric

Last Updated: 05/18/2021