Supervised Learning for a Proxy Mobile Health Outcome

Shaine Leibowitz
Community Insight & Impact

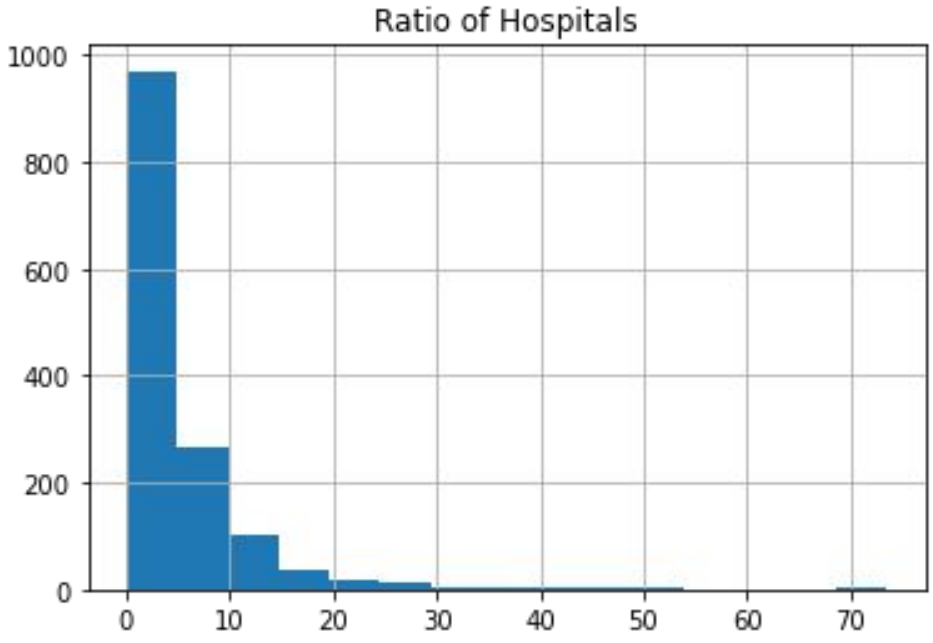
Background: Community Vulnerability Core Metric

- The Community Vulnerability Index (CVI) aggregates data on the core metric, mobile health needs, 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 Mobile Health metric measures community need for non-traditional healthcare delivery services
- 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 Mobile Health Needs metric
- Implemented a supervised learning model to predict a proxy mobile health outcome: Ratio of Hospitals per 100,000 population
    - Counties with a low number of hospitals have a higher need for mobile healthcare services
- See related code in GitHub here

Distribution of Proxy Outcome



Original Features

% Rural	% households wo car	% Adults 65 and Older
Primary Care Physicians Rate	% Without Health Insurance	% Non-white
% Limited English Proficiency	% Veterans in Civilian Adult Population	% Workers Commuting by Public Transit
% People with Disabilities	Opioid Death Rate	% Fair or Poor Health
Number of Hospitals*		

* Not included in analysis as Number of Hospitals was used as proxy outcome

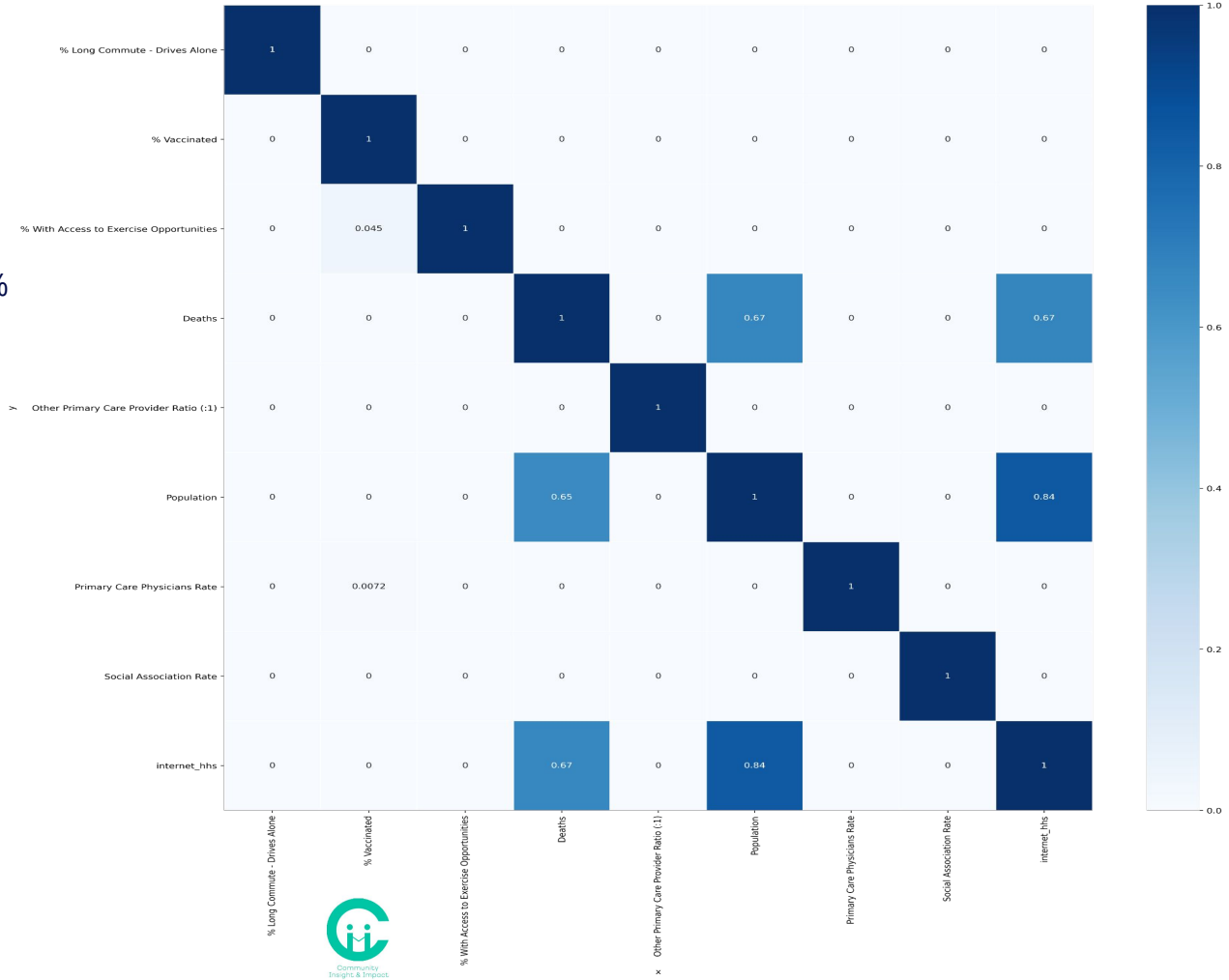
XGBoost Most Important Features

Starting with the full CVI dataset, we narrowed our comparison set by selecting most important features according to XGBoost (F-score > 50)

Primary Care Physicians Rate	Population	Social Association Rate
Deaths	% Vaccinated	% With Home Internet Access
% With Access to Exercise Opportunities	% Long Commute - Drives Alone	Other Primary Care Provider Ratio

PPScore: XGBoost Most Important Features

- Removed Population and % With Home Internet Access



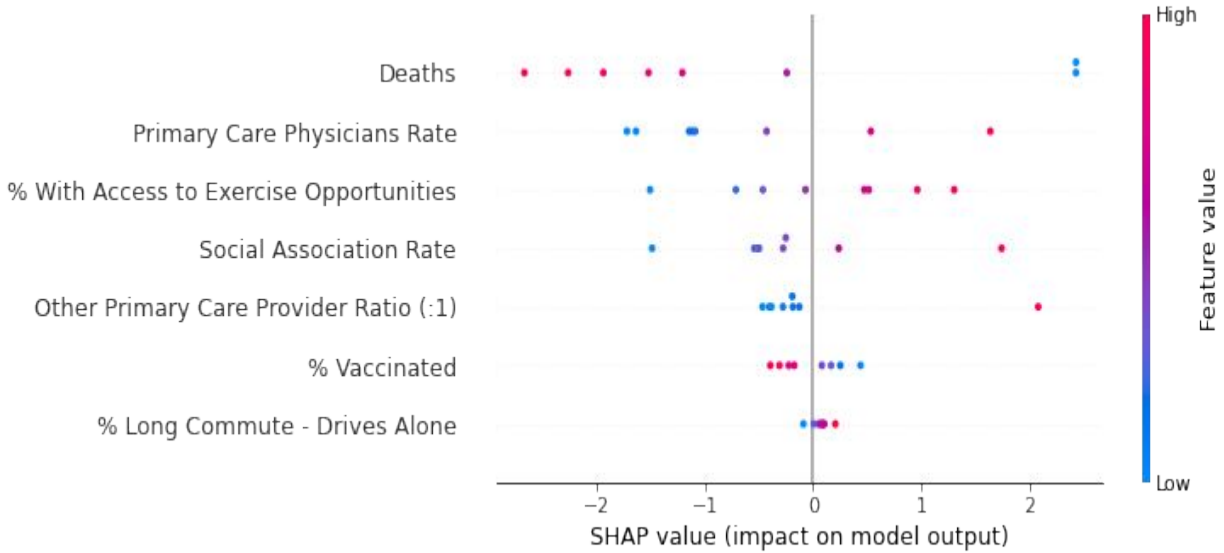
Comparison of Different Feature Sets

With a Root Mean Squared Error (RMSE) of 4.29 on the test data set, the XGBoost Most Important Features performed better than the original features as its RMSE was 4.97. Multilayer Perceptrons were applied for these models.

Dummy Baseline	Original Features	XGBoost Most Important Features
106.9	4.97	**4.29**

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



Comparison of Different Feature Sets

With a Root Mean Squared Error (RMSE) of 4.29 on the test data set, the XGBoost Most Important Features performed better than the original features as its RMSE was 4.97. Multilayer Perceptrons were applied for these models.

Dummy Baseline	Original Features	XGBoost Most Important Features
106.9	4.97	**4.29**

Comparison of Different Feature Sets

The RMSE on the test data set decreases by simply adding "Deaths".

Dummy Baseline	Original Features	XGBoost Most Important Features	Original Features + “Deaths”
106.9	5.0	4.3	4.3

Discussion

- Only two of the original metric features, ‘Primary Care Physicians Rate’ and ‘% With Home Internet Access’, were identified by XGBoost
  - Currently ‘% With Home Internet Access’ is included as an overlay rather than directly incorporated into the metric
- ‘Primary Care Physicians Rate’ has a large SHAP value, supporting it being one of the highest (negatively) weighted variables in the original metric
- ‘Other Primary Care Provider Ratio’, which describes access to non-physician-based care, such as nurse practitioners or physician assistants, was also identified by XGBoost and has a large SHAP value
  - We plan to incorporate this variable into future versions of the metric but require additional study on the interplay with ‘Primary Care Physicians Rate’ and poverty indicators as a majority of non-physician-based care in an area can still be indicative of shortcomings of the local healthcare infrastructure

Discussion

- Interestingly, the '% Vaccinated' SHAP value is inversely correlated with the proxy outcome
    - This is possibly due to many COVID-19 vaccination clinics being setup in pharmacies, community centers, and other other 'pop-up' locations, rather than solely in hospitals, however this requires further study
- 'Access to Exercise Opportunities' (percentage of population with adequate access to locations for physical activity including sidewalks, parks, and gyms) and 'Social Association Rate' (number of membership associations per 10,000 population) also have large SHAP values
- Both variables describe access to social and physical infrastructure that enable healthy behavior and improved health outcomes
- It is possible that a lack of these resources also indicated a need for additional mobile healthcare resources, in which case including these variables would improve our Mobile Health metric by augmenting the more traditional variables of 'Number of Hospitals' and 'Primary Care Physicians Rate'
    - Needs additional literature review on the variables' backgrounds and their causal impact on community well-being

Discussion: 'Deaths'

- 'Deaths' (daily COVID-19 death counts) was the XGBoost identified feature with the largest SHAP value
- The SHAP value is negatively correlated with the proxy outcome, which intuitively makes sense as areas with fewer healthcare resources have higher rates of poor health and were more likely to overwhelm existing healthcare infrastructure during a COVID outbreak
- However,the COVID-19 death rate in different counties is also highly dependent on public health policy implementation and adherence, so we will not include 'Deaths' in future versions of the Mobile Health Needs metric

Last Updated: 05/18/2021