Explainable machine learning models to understand determinants of COVID-19 mortality in the United States

Authors: Piyush Mathur, MD, FCCM[1]; Tavpritesh Sethi, MBBS, PhD[2]; Anya Mathur[3]; Kamal Maheshwari ,MD, MPH[1] ; Jacek B Cywinski, MD, FASA[1] ; Ashish K Khanna, MD, FCCP, FCCM[4,5]; Simran Dua[6] ;Frank Papay, MD[7]

Affiliations: [1]Department of General Anesthesiology, Anesthesiology Institute, Cleveland Clinic, Cleveland ; [2]Indraprastha Institute of Information Technology, Delhi, India ; [3]Brecksville Broadview Heights Middle School, Ohio; [4]Department of Anesthesiology, Section on Critical Care Medicine, Wake Forest School of Medicine, [5]Outcomes Research Consortium, Cleveland, OH; [6]Lambert High School, Suwanee, Georgia; [7]Dermatology and Plastic Surgery Institute , Cleveland Clinic, Cleveland.

Corresponding author: Piyush Mathur MD,FCCM. Anesthesiology Institute,Cleveland Clinic,E3-205,9500 Euclid Avenue,Cleveland, Ohio 44195(mathurp@ccf.org)

Abstract

COVID-19 mortality is now the leading cause of death per day in the United States,ranking higher than heart disease and cancer.Multiple projection models have been built and used to understand the prevalence of disease and anticipated mortality.These models take into account various epidemiologic factors of disease spread and more recently some of the mitigation measures.The authors developed a dataset with many of the socioeconomic, demographic, travel, and health care features likely to impact COVID-19 mortality.The dataset was compiled using 20 variables for each of the fifty states in the United States.We subsequently developed two independent machine learning models using Catboost regression and random forest.Both the models showed similar level of accuracy.CatBoost regression model obtained R^2 score of 0.99 on the training data set and 0.50 on the test.Random forest model similarly obtained a R^2 score of 0.88 on the training data set and 0.39 on the test set. To understand the relative importance of features on COVID-19 mortality in the United States,we subsequently used SHAP feature importance and Boruta algorithm.Both the models show that high population density, pre-existing need for medical care and foreign travel may increase transmission and thus COVID-19 mortality whereas the effect of geographic, climate and racial disparities on COVID-19 related mortality is not clear.Location based understanding of key determinants of COVID-19 mortality, is needed for focused targeting of mitigation and control measures.Explanatory models such as these are also critical to resource management and policy framework.

Introduction

The COVID-19 pandemic exhibits an uneven geographic spread which leads to a locational mismatch of testing, mitigation measures and allocation of healthcare resources (human, equipment, and infrastructure).[1] In the absence of effective treatment, understanding and predicting the spread of COVID-19 is unquestionably valuable for public health and hospital authorities to plan for and manage the pandemic.

While there have been many models developed to predict mortality, the authors sought to develop a machine learning prediction model that provides an estimate of the relative association of socioeconomic, demographic, travel, and health care characteristics of COVID-19 disease mortality among states in the United States(US).

Methods

State-wise data was collected for all the features predicting COVID-19 mortality and for deriving feature importance (eTable 1 in the Supplement).[2] Key feature categories include demographic characteristics of the population, pre-existing healthcare utilization, travel, weather, socioeconomic variables, racial distribution and timing of disease mitigation measures (Figure 1 & 2).

Two machine learning models, Catboost regression and random forest were trained independently to predict mortality in states on data partitioned into a training (80%) and test (20%) set.[3] Accuracy of models was assessed by R^2 score.

Importance of the features for prediction of mortality was calculated via two machine learning algorithms - SHAP (SHapley Additive exPlanations) calculated upon CatBoost model and Boruta, a random forest based method trained with 10,000 trees for calculating statistical significance [3-5].

Results

Results are based on 60,604 total deaths in the US, as of April 30, 2020. Actual number of deaths ranged widely from 7 (Wyoming) to 18,909 (New York).CatBoost regression model obtained an R^2 score of 0.99 on the training data set and 0.50 on the test set. Random Forest model obtained an R^2 score of 0.88 on the training data set and 0.39 on the test set.

Nine out of twenty variables were significantly higher than the maximum variable importance achieved by the shadow dataset in Boruta regression (Figure 2).Both models showed the high feature importance for pre-existing high healthcare utilization reflective in nursing home beds per capita and doctors per 100,000 population. Overall population characteristics such as total population and population density also correlated positively with the number of deaths.Notably, both models revealed a high positive correlation of deaths with percentage of African Americans. Direct flights from China, especially Wuhan were also significant in both models as predictors of death, therefore reflecting early spread of the disease.

Associations between deaths and weather patterns, hospital bed capacity, median age, timing of administrative action to mitigate disease spread such as the closure of educational institutions or stay at home order were not significant. The lack of some associations, e.g., administrative action may reflect delayed outcomes of interventions which were not yet reflected in data.

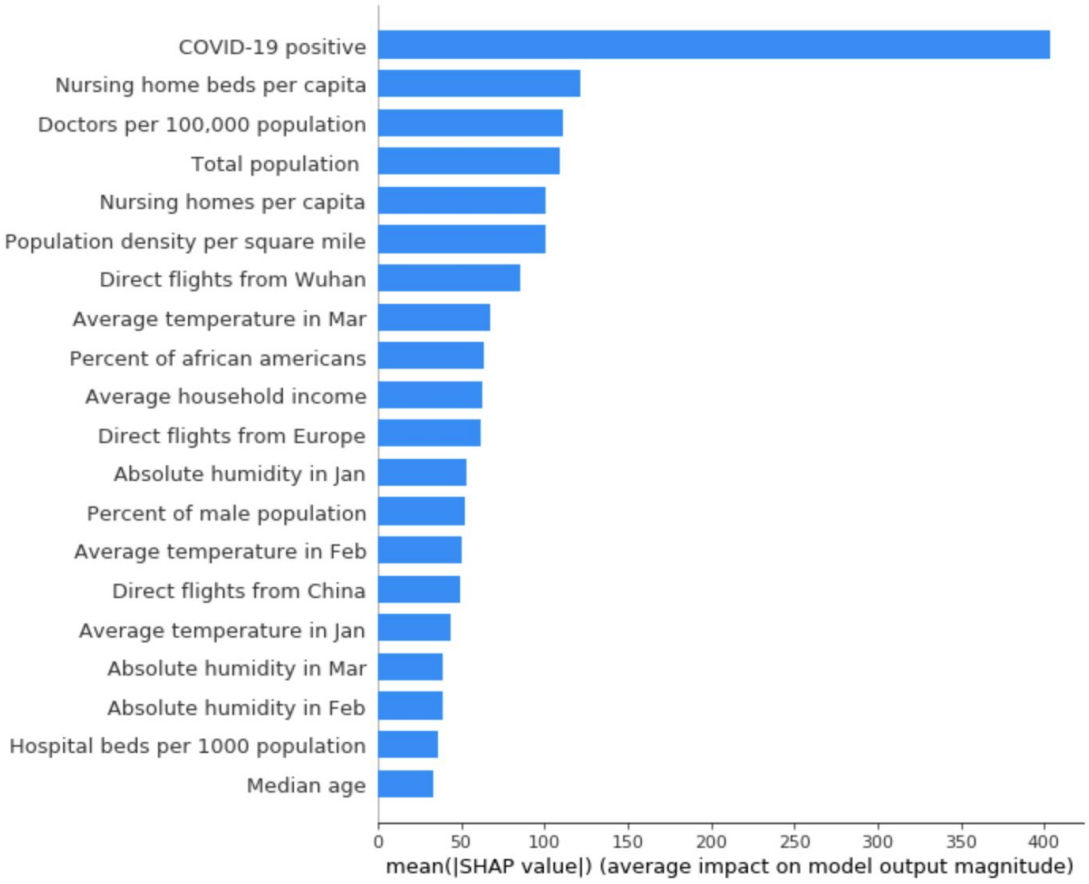

Figure 1a. SHAP feature importance model results for COVID-19 deaths in the US. Feature importance plot lists the most significant features in descending order. The top features contribute more to the model than the bottom ones and thus have high predictive power [5].

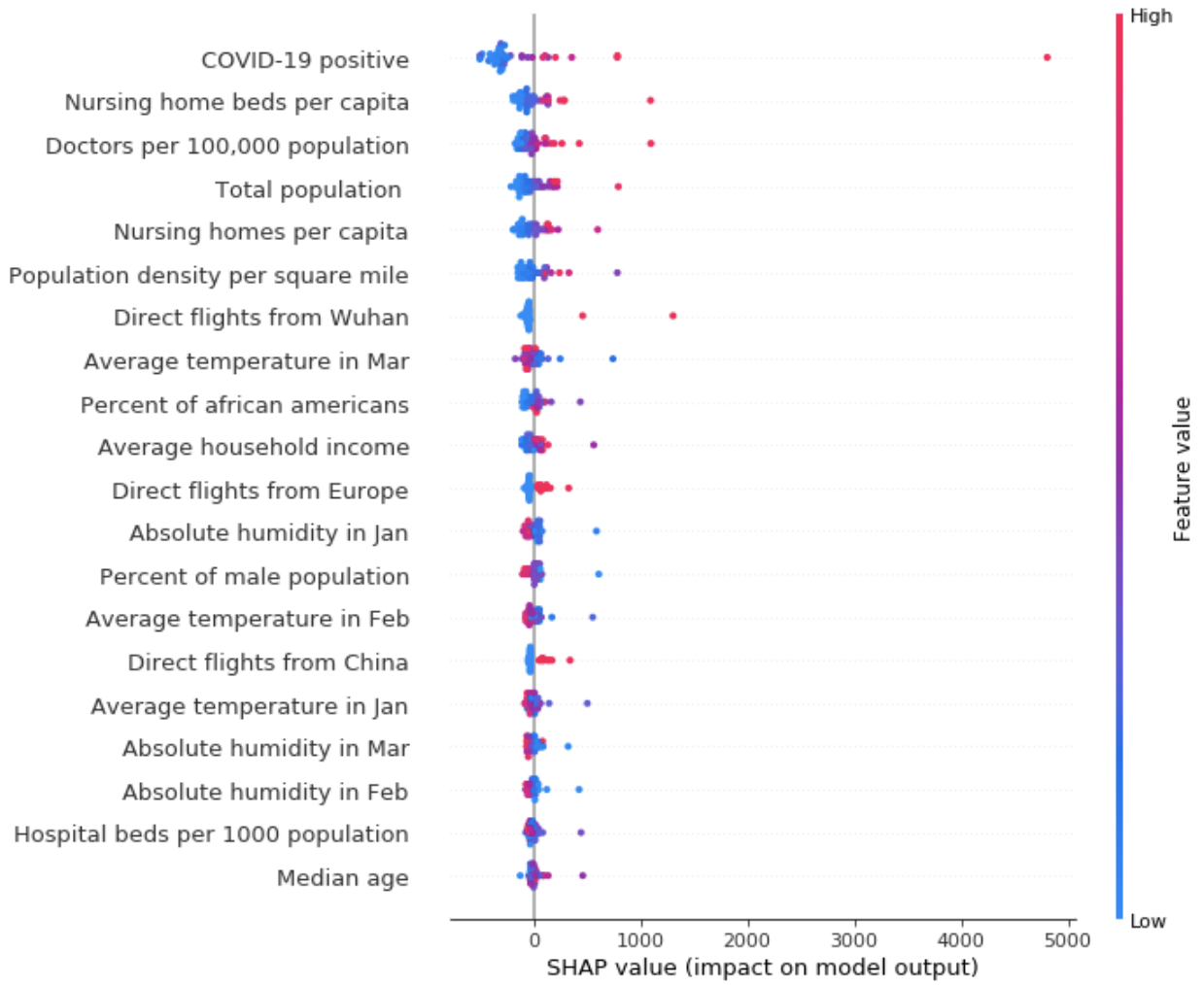

Figure 1b. SHAP feature importance distribution of importance amongst each state for COVID-19 deaths in the US. The vertical location of the dots show what feature they are depicting.Color shows whether that feature value was high or low for that row of the dataset.Horizontal location shows whether the effect of that value caused a higher or lower prediction.

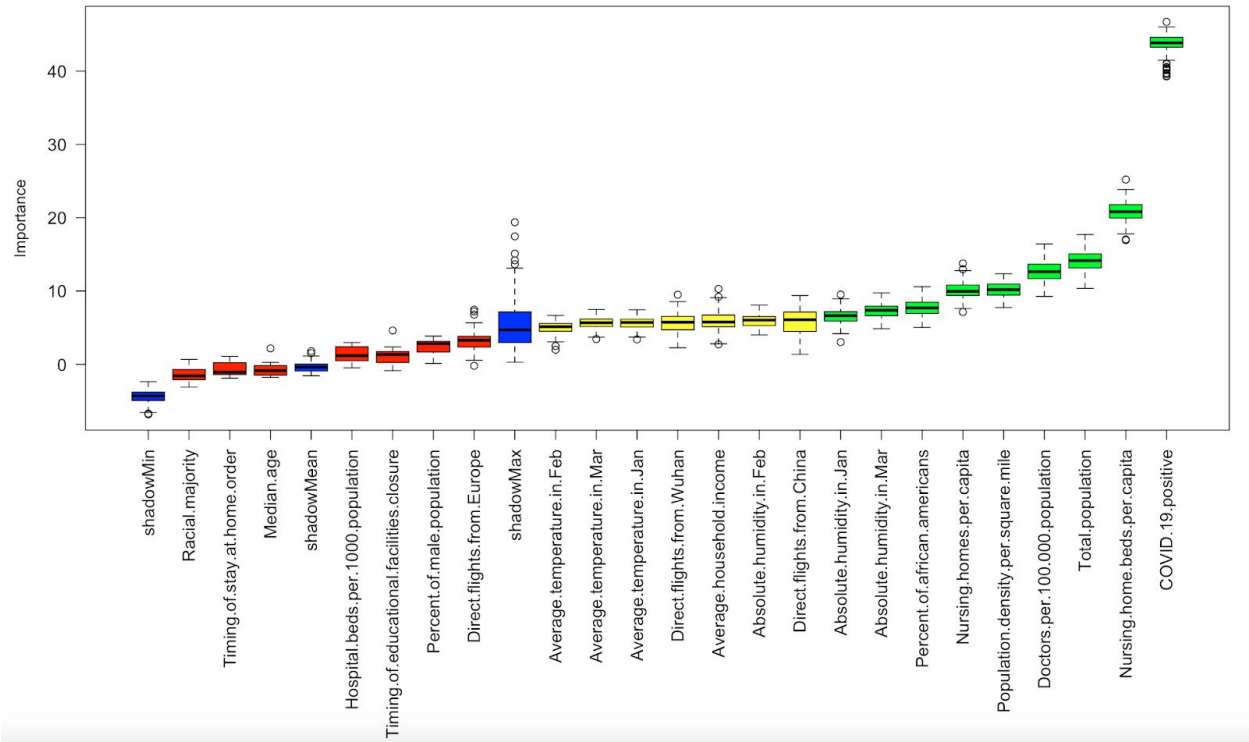

Figure 2. Statistical significance of feature importance computed through Boruta algorithm for COVID-19 deaths in the US[4]. Boxplot distribution of importance, highlights features (green, yellow) whose median importance is significantly higher than the shadow importance (blue).

Discussion

COVID-19 disease has varied spread and mortality across communities amongst different states in the US. While our models show that high population density, pre-existing need for medical care and foreign travel may increase transmission and thus COVID-19 mortality, the effect of geographic, climate and racial disparities on COVID-19 related mortality is not clear.

The purpose of our study was not state-wise accurate prediction of deaths in the US, which has already been challenging.[6] Location based understanding of key determinants of COVID-19 mortality, is critically needed for focused targeting of mitigation and control measures.

Risk assessment-based understanding of determinants affecting COVID-19 outcomes, using a dynamic and scalable machine learning model such as the two proposed, can help guide resource management and policy framework.

References

1. Schuchat A, Team CC-R. Public Health Response to the Initiation and Spread of Pandemic COVID-19 in the United States, February 24-April 21, 2020. MMWR Morb Mortal Wkly Rep. 2020;69(18):551-6.

2. Gupta S, Raghuwanshi GS, Chanda A. Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020. The Science of the total environment. 2020;728:138860.

3. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Proceedings of the 32nd International Conference on Neural Information Processing Systems; Montréal, Canada: Curran Associates Inc.; 2018. p. 6639–49.

4. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. 2010. 2010;36(11):13.

5. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 4768–77.

6. Jewell NP, Lewnard JA, Jewell BL. Caution Warranted: Using the Institute for Health Metrics and Evaluation Model for Predicting the Course of the COVID-19 Pandemic. Annals of Internal Medicine. 2020.