PREDICT COVID-19 DEATH RATES

James Perry
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
jp4dr@virginia.edu

Sharon Bryant
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
ssb7xx@virginia.edu

Christine Baca
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
cab8xd@virginia.edu

November 24, 2020

1 Abstract

To discover what social determinant most influences COVID death rates in Charlottesville, VA, the team ran tests on several subsets of data with different profiles from the Virginia Health Opportunity Index and Johns Hopkins COVID-19 Data (Virginia only). The methodology consists of running multiple subsets of the data through various regression algorithms and cross-validation to find the model with least error using Sci kit-Learn and Python on Google Colab. The success of the model decides which profile can help optimize the prediction of COVID death rates in Charlottesville. The team decided to use a Random Forest Regression Model tuned using both Random and Grid search; the regression performed best with the Health Opportunity Index, and data visualization showed the Wellness Disparity to most correlate with and possibly influence the death rates in Charlottesville, Virginia.

2 Introduction

There are many different factors that possibly influence COVID-19 death rates in counties within Virginia. Previous studies such as *Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing* by Shreshth Tuli, Shikhar Tuli, Rakesh Tuli, and Sukhpal Singh Gill from Science Direct, developed models that first used a Gaussian fit to predict COVID-19 infection and death rates through a World Health Organization dataset. After several months of data collection, a Generalized Inverse Weibull (GIW) Distribution fit better. After comparing the results, this group chose the Iterative Weibull over the non-iteratively weighted Weibull. Our team used the Virginia Health Opportunity Index calculated by the Virginia Department of Health Office of Health Equity to predict the rate of COVID-19 death rates in Charlottesville, Virginia by combining it with Johns Hopkins COVID-19 data. The Health Opportunity Index is made of four profiles: Economic Opportunity, Consumer Opportunity, Community Environmental, and Wellness Disparity. Finding the correlation between Health Opportunity Index and COVID-19 death may better inform Charlottesville, Virginia of the factors influencing the outbreaks and how to better predict death rates of a pandemic using machine learning. Using several machine learning regression algorithms and tuning methods, we found several factors that correlate with COVID-19 related death in Charlottesville, Virginia.

3 Method

The data was first cleaned and scaled. Since two different datasets were used, the team chose to incorporate both the Johns Hopkins datasets and the Health Opportunity Index dataset by finding entries that match by census tract for

Charlottesville, Virginia. The data was then split into multiple subsets corresponding to the profiles indicated by the Health Opportunity Index dataset and was also tested against data not utilizing the Health Opportunity Index. The subsets of data were made of the following: No HOI data, HOI calculation only, Community Environment Profile, Consumer Opportunity Profile, Economic Opportunity Profile, and Wellness Disparity Profile. For each subset of data, we used LinearRegression, DecisionTreeRegressor, and RandomForestRegressor and took note of the amount of error for each model. We chose the optimal regression model by finding the one with least rmse error after cross validation.

4 Experiments

We ended up choosing the Random Forest regression model because it had the lowest error from out cross validation. From there, we used a random search on a large set of values for both the estimators and features to narrow down the best options, and then performed a grid search using the best ranges. This gave us am optimal estimator which we again tested with cross validation to confirm the performance. After this we started testing on different subsets of our data set. Because the Health Opportunity Index consists of 4 indexes, we were able to split the data into subsets with different combinations of profiles and index calculations. Testing the random forest on all the subsets of data, we found that the Health Opportunity index calculation itself had the highest correlation with COVID deaths. Of the other profiles in the set, the Wellness Disparity Profile and its factors had the largest effect on our predictions. This profile is made up of a census tract's community diversity and access to care.

5 Results

After analyzing the data, training the model, and tuning its parameters we were able to successfully identify which health factors best indicate COVID risk and also help predict COVID deaths in Charlottesville. We found that the Health Opportunity index calculation itself had the highest correlation with COVID deaths. Of the other profiles in the set, the Wellness Disparity Profile and its factors had the largest effect on our predictions. This profile is made up of a census tract's community diversity and access to care. We were able to achieve a mean squared error of 1.476 from 10-fold cross validation, with a standard deviation of 1.361. These gave us confidence in the accuracy of our model, especially given the different data sets that we worked through and all the factors taken into account.

6 Conclusion

After training and testing a machine learning regression to predict COVID-19 deaths on several subsets of data with different profiles from the Virginia Health Opportunity Index and Johns Hopkins COVID-19Data (Virginia only), the experiment concludes that the regression performs best with the Health Opportunity index, and data visualization shows the Wellness Disparity to most correlate with death rates. Possible errors include the existence of bias and possible over fitting from training on a smaller data set; future extensions of this experiment could train and test the regressor on larger census data sets and possibly add on other related data sets to augment the scope and intuition of the algorithm.

7 Contributions

Christine Baca cleaned the the data, wrote the Abstract and Conclusion sections, and co-wrote the script for the video presentation. Sharon Bryant modelled the data, wrote the Introduction and Method sections, and animated the video presentation in its entirety. Jimmy Perry wrote the other sections as well as co-wrote and voiced the script for the video presentation.

References

[1] Virginia Health Opportunity Index. (n.d.). Retrieved October 01, 2020, from https://apps.vdh.virginia.gov/omhhe/hoi/

[2] Health Opportunity Index. (2020, September 21). Retrieved October 01, 2020, from https://data.virginia.gov/Family-Health/Health-Opportunity-Index/6q6u-dcz7

[3] Johns Hopkins COVID-19 Data (Virginia only). (2020, September 30). Retrieved October 01. 2020, from https://data.virginia.gov/Datathon-2020/Johns-Hopkins-COVID-19-Data-Virginia-only-/c62p-d8xr

[4] Tuli, S., Tuli, S., Tuli, R, Gill, S. (2020, January 01). Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing. Retrieved October 01, 2020, from https://www.medrxiv.org/content/10.1101/2020.05.06.20091900v1