Classifying High COVID Fatality Rates

Casey Durfee
University of Colorado
MSDS Data Mining Final Project
casey.durfee@colorado.edu
https://github.com/csdurfee/COVID_fatality_rates

ABSTRACT

Death rates from COVID-19 pandemic have varied widely between different parts of the United States. The introduction of vaccines a year ago further changed the dynamics of death rates from COVID.

A wide range of United States county-level data was collected and analyzed to determine which metrics predicted high COVID fatality rates. I found that those factors changed between year one and year two of the pandemic. Death rates in the second year had higher correlations with other factors than first year death rates, indicating they may have been more predictable.

I tested this by using five Machine Learning algorithms to classify whether a county would have higher than average COVID death rates. All of the algorithms had higher accuracy scores on year two death rates than they did on year one. An analysis using permutation importance found that 2016 Republican presidential vote share was the strongest factor used to classify high COVID death rates.

KEYWORDS

Machine Learning, Logistic Regression, COVID-19

INTRODUCTION

Coronavirus has killed over a million people in the United States over the past 2 years. We've all felt the burden of the disease and how it has changed our daily life.

But some communities have been hit much harder than others. The 25th percentile death rate for US counties is 2.5 per thousand. The 75th percentile rate is 4.6 per thousand, almost twice the rate. The US has a population of 330 Million people, so that represents a difference of 673,000 deaths, which is a massive range of outcomes.

What was the difference between counties that had lower death rates and those with higher death rates? I wanted to gather as wide a range of data as possible, rather than following my own biases about the cause of the variance.

All code used in this project is available at [Durfee].

RELATED WORK

One large and obvious cause of variance has been differing vaccination rates. One analysis found 163,000 deaths of COVID that could have been prevented by vaccines from June to December of 2021 [Ortaliza 2021].

There have also been investigations of COVID deaths and very specific demographic factors. One recent article [Lin 2022] identified higher access to broadband as a strong predictor of lower COVID fatality rates at the county level. [Tan 2021] looked at income inequality as a predictor of COVID fatality rates. [Millett 2020] found that COVID disproportionately affected minority communities during the early months of the pandemic. [Gao 2021] found a link between political leanings and COVID fatality rates.

However, I wasn't able to find any broad-spectrum analysis of community factors and COVID rates.

COMPLETED WORK

I obtained the COVID fatality data from [CSSEGISandData], a project of Johns Hopkins University. They cross-check official death totals with state and county-level agencies, and seemed to be the most complete. For vaccine rates, I used [vaccinetracking.us], a project of Georgetown University.

For the community metrics, I started with 2019 data from [countyhealthrankings.org], a project of the University of Wisconsin. They aggregate a wide variety of county-level metrics from different US Government sources and then produce rankings of counties in a wide range of categories (including Quality of Life, Health Behaviors, Health Outcomes, Social & Economic Factors, and Clinical Care). This project allows US residents to see how their

community compares to others. I used both their raw data and their calculated rankings in my analysis.

I also obtained community level data from the 2018 Social Vulnerability Index, a project of the CDC which tracks social factors that may affect the ability of a community to deal with pandemics or natural disasters, such as how many people live in mobile homes or do not have access to a car. The SVI provides 16 metrics in all.

For historical voting data, used [MEDSL], a project of the MIT Elections Data Science Lab. In particular, I looked at the share of votes for the Republican candidate for president in 2016, since political leanings were mentioned as a potential predictor in [Gao 2021]. Finally, for county-level economic data, I used [Wikipedia].

The FIPS (Federal Information Processing Standard) code was used to join datasets together. FIPS supplies a unique number for every county/borough in the United States.

Between all of these sources, I was able to gather 100 different county-level metrics. However, some of them have missing data, and some of the data from [countyhealthrankings.org] and [SVI] have already been imputed. Additionally, [countyhealthrankings.org] also provides aggregate metrics, which are a weighted and scaled combination of the raw data.

So I made sure to use as much of the raw data as possible, without imputing too many missing values. I threw out any columns with more than 5% missing values and used the median value for imputation. That left 70 metrics for 2,919 US counties. All data was scaled using sklearn's StandardScaler.

There is some overlap between metrics, for instance several different estimates of household income from different sources. I left these potentially redundant sources of data in, with the goal being to let the models decide which factors are more important, rather than potentially injecting bias by hand-selecting metrics.

I have defined the first year as March 19, 2020 to March 19, 2021. This corresponds to the year after COVID was declared a pandemic by the World Health Organization. The second year, March 20, 2021 to March 19, 2022, roughly corresponds to the wide availability of vaccines in the United States. By March 13, 2021, over 100 million vaccinations had been administered in the United States [CDC 2022]. So it roughly represents the period of time when we'd expect vaccines to have an impact on COVID fatality rates. Comparing exact one year periods avoids potential biases due to seasonality of COVID.

I looked at the correlations between community factors and COVID death rates and how they were different between year one and year two.

I then trained a wide range of models (including Random Forest, SVC, AdaBoost and Logistic Regression) to classify when a county would have a higher or lower than average COVID fatality rate. I identified the best model using accuracy score with k-fold cross validation. I then used permutation importance to determine which factors most strongly predicted high fatality rates. I compared the rankings of importance between models and found some factors that were consistently important across very different types of models.

Finally, I performed the same analysis on counties with a population over 50,000. This added 10 more factors that had too many missing values for the smaller counties. There are 932 of these bigger counties. 87% of all people in the US live in one of these bigger counties. This led to a higher accuracy model.

EVALUATION

I found that correlations with year two death rates were stronger than those with year one death rates. On the surface, this is surprising, because the county-level data is all from 2018 and 2019, so they should correlate more strongly with 2020 than with 2021.

The top factors for year one were:

Factor	r^2 (year one)
Age-Adjusted Mortality (Hispanic) (CHR)	0.376
Teen Birth Rate (CHR)	0.351
% Disconnected Youth (CHR)	0.338
MV Mortality Rate (CHR)	0.337
% No HS Diploma (SVI)	0.321
Homicide Rate (CHR)	0.32
Years of Potential Life Lost Rate (CHR)	0.311
Child Mortality Rate (CHR)	0.305
Age-Adjusted Mortality (CHR)	0.304
% Physically Inactive (CHR)	0.299

The top factors for year two were:

Factor	r^2 (year two)
Age-Adjusted Mortality (CHR)	0.495
Life Expectancy (CHR)	-0.49
% Disabled (SVI)	0.482
Years of Potential Life Lost Rate (CHR)	0.467

Median Family (Wikipedia)	-0.457
Physically Unhealthy Days (CHR)	0.453
Teen Birth Rate (CHR)	0.45
% Some College (CHR)	-0.443
COVID Complete Coverage	-0.44
Median Household (Wikipedia)	-0.434

There are several interesting things here. Right off the top, we see that some of the top factors in year two had negative correlations, whereas none did in year one.

Second, correlations were much stronger across the board in year two than in year one. There were zero factors that had an r^2 value of 0.4 or higher for year one, while there were 22 for year two.

Finally, there are clearly several measures that are similar to each other showing up. Life Expectancy, Age-Adjusted Mortality, and Years of Potential Life Lost Rate (abbreviated below as “YPLL”) are all basically estimating the same thing in slightly different ways. If we look at the strongest correlated factor with each one, we see that they’re not complete duplicates of each other, though:

Column	Highest Correlated With	r^2
YPLL (CHR)	Age-Adjusted Mortali	0.96
Teen Birth Rate (CHR	Age-Adjusted Mortali	0.724
Physically Unhealthy	Age-Adjusted Mortali	0.714
% Physically Inactiv	Age-Adjusted Mortali	0.656
Median Household (Wi	Median Family (Wikip	0.962
% Some College (CHR)	Median Family (Wikip	0.671
Life Expectancy (CHR	Median Family (Wikip	0.632
COVID Complete Cover	Median Family (Wikip	0.445
Median Family (Wikip	Median Household (Wi	0.962
% Disabled (SVI)	Physically Unhealthy	0.617
% No HS Diploma (SVI	Teen Birth Rate (CHR	0.711
% Disconnected Youth	Teen Birth Rate (CHR	0.592
Mortality (Hispanic) (CHR)	Teen Birth Rate (CHR	0.497
Age-Adjusted Mortali	YPLL (CHR)	0.96
Child Mortality Rate	YPLL (CHR)	0.723
MV Mortality Rate (C	YPLL (CHR)	0.658
Homicide Rate (CHR)	YPLL (CHR)	0.631

I then trained a set of models on combined year one and two fatality rates, then on each year separately. The goal of these models is to do a binary classification to predict whether the COVID fatality rate for a county will be higher or lower than the mean value.

These are the accuracy score results, calculated with 5-fold cross validation:

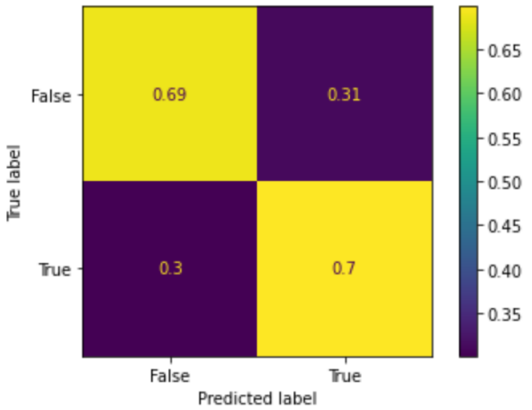
Classifier	Year 1	Year 2	Both	Total
RandomForestClassifier	0.636	0.74	0.725	2.101
SVC	0.652	0.756	0.725	2.133
LogisticRegressionCV	0.631	0.758	0.741	2.131
RidgeClassifierCV	0.637	0.751	0.715	2.103
AdaBoostClassifier	0.623	0.726	0.695	2.045
BaggingClassifier	0.624	0.704	0.683	2.011
GradientBoostingClassifier	0.648	0.734	0.723	2.105
LinearSVC	0.633	0.757	0.719	2.109

All classifiers had higher accuracy scores on year 2 than they did on year one, or on both years combined. All other things being equal, we would expect the models trained on both years combined to have a higher accuracy score than either year individually, because a two year long sample of deaths will have lower variance than a single year sample. This suggests that year 2 deaths were more predictable than year 1 deaths.

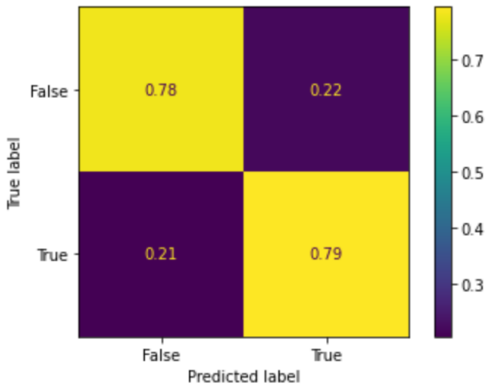
LogisticRegressionCV had the best performance across all 3 sets of data, with SVC close behind. BaggingClassifier was consistently the worst across all 3 datasets.

LogisticRegressionCV, also has a couple appealing qualities. It applies regularization, which means a penalty is applied for overly complex models. This seemed like a good choice to understand which factors are most important among a bunch of metrics that have a tangled web of correlations with each other. The algorithm also tunes its own hyperparameters, saving us from having to do that. (It is possible that other algorithms could outperform LogisticRegressionCV with a grid search of their tuning parameters.)

I then trained a LogisticRegressionCV model with an 80% train/test split. The confusion matrix for year one fatality rates was:



For year two, it was:



I then used permutation importance to figure out which factors had the most influence on the decision of the classifier. Permutation importance scrambles the inputs to a classifier and records whether the classification changes. It's a good technique to be able to determine what is important to different classes of machine learning algorithms. Here's year one:

Factor	Importance (Year One)
% Over 65 (SVI)	0.033
Life Expectancy (CHR)	0.026
Physically Unhealthy Days (CHR)	0.02
% Disabled (SVI)	0.019
% Physically Inactive (CHR)	0.018
Median Household (Wikipedia)	0.018
% Rural (CHR)	0.017
Mentally Unhealthy Days (CHR)	0.017
2016 Repub Vote Share	0.016
% Screened (CHR)	0.013

And here's year two:

Factor	Importance (Year Two)
2016 Repub Vote Share	0.038
% Some College (CHR)	0.018
Teen Birth Rate (CHR)	0.018
% Insufficient Sleep (CHR)	0.016
% Hispanic (CHR)	0.016
% Diabetic (CHR)	0.015
% Limited English (SVI)	0.015
% Rural (CHR)	0.015
% Fair/Poor (CHR)	0.013
Clinical Care Percentile (CHR)	0.012

This just shows importance to the classifier, not in which direction. In year one, "Life Expectancy", "Median Household" and "% Screened" were correlated with lower COVID fatality rates, and the rest with higher rates. ("% Screened" measures the % of women who have had mammography screenings recently.)

In year two, "% Some College", "% Hispanic", "% Limited English", and "Clinical Care Percentile" were correlated with lower fatality rates, and the others were correlated with higher rates.

I then fit the year two data with several of the runner up models: LinearSVC, Gradient Boosting, Ridge Regression, and Random Forest. These all had fairly good performance, and represent 4 very different approaches to classifying data. (Although SVC had the second highest accuracy score, calculating permutation importance on it is prohibitively slow.)

I calculated the rankings of permutation importance for each of the four models and took the mean. Here is how they compare to the rankings from the LinearRegressionCV ("LRCV") model:

Factor	LRCV Rank	Mean Rank
2016 Repub Vote Share	1	1.5
Teen Birth Rate (CHR)	2	16.25
% Some College (CHR)	3	10.25
% Hispanic (CHR)	4	23.125
% Insufficient Sleep (CHR)	5	39.125
% Diabetic (CHR)	6	9.25
% Limited English (SVI)	7	18.875
% Rural (CHR)	8	37.875
% Fair/Poor (CHR)	9	20.5
Clinical Care Percentile (CHR)	10	16.5
% Obese (CHR)	11	26.875
% Severe Housing Problems (CHR)	12	36.625
% Frequent Physical Distress (CHR)	13	27.25
Median Family (Wikipedia)	14	25.5
% Uninsured (CHR)	15	27.625
% No HS Diploma (SVI)	16	21.5
Physically Unhealthy Days (CHR)	17	32.5
Income Ratio (CHR)	18	28.625
Life Expectancy (CHR)	19	38.5
COVID Complete Coverage	20	9.75

"2016 Repub Vote Share", "% Diabetic", and "% Some College" appear to be the strongest common factors across all five models.

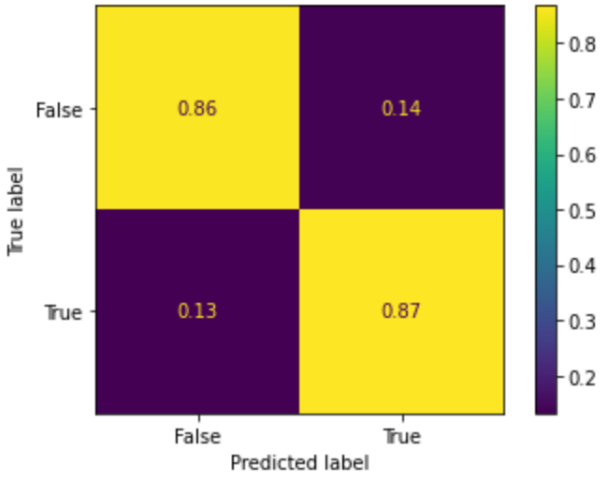
Finally, I restricted the data to just counties with over 50,000 people. Because of the imputation being done by the providers of the data I'm using, I wanted to limit it to higher population counties, where the data should be more accurate.

All of the models performed better on the subset:

Classifier	Year 1	Year 2	Both	Total
RandomForestClassi	0.731	0.798	0.778	2.307
SVC	0.739	0.806	0.783	2.328
LogisticRegression	0.731	0.811	0.791	2.333
RidgeClassifierCV	0.718	0.79	0.747	2.254
AdaBoostClassifier	0.718	0.796	0.783	2.297
BaggingClassifier	0.687	0.768	0.737	2.192
GradientBoostingCl	0.764	0.801	0.784	2.349
LinearSVC	0.701	0.768	0.755	2.224

Once again, LogisticRegressionCV performed the best on Year 2 data, although GradientBoosting performed better overall.

Here is the confusion matrix for the year 2 large county model:



And here are the most important factors:

Factor	Importance
2016 Repub Vote Share	0.025
Segregation index (CHR)	0.016
Income Ratio (CHR)	0.015
% Over 65 (SVI)	0.015
% Female (CHR)	0.014
Clinical Care Percentile (CHR)	0.014
% Insufficient Sleep (CHR)	0.013
% Rural (CHR)	0.013
Infant Mortality Rate (CHR)	0.012
COVID Booster Coverage	0.011

Because it showed up in every single model, it's worth looking at what 2016 Republican vote Share correlates with:

Name	r^2 with Repub vote
COVID Partial Coverage	-0.692
COVID Complete Coverage	-0.688
COVID Booster Coverage	-0.594
% Severe Housing Problems (CHR)	-0.564
% Multiunit Housing (SVI)	-0.529
% Non-Hispanic White (CHR)	0.52
% Minority (SVI)	-0.518
% Rural (CHR)	0.469
% Homeowners (CHR)	0.467
% Asian (CHR)	-0.441

DISCUSSION

In retrospect, not doing feature reduction made the cross-model comparisons less valuable than they could be. For instance, there are 3 different metrics for life expectancy (Years of Potential Life Lost, Life Expectancy, Age-Adjusted Mortality); additionally Life Expectancy is broken down by race.

Three different models could each decide a different one of those life expectancy metrics is the best one, which makes comparing the rankings of different classifiers the way I did it a bit artificial. If I had more time, I would have tried to group the metrics into general buckets such as health, money, vaccination status, and cultural influences, and tried to pick the strongest predictors from each. That would help elucidate which of the bigger general themes are the most important to COVID fatality rates.

The LogisticRegressionCV algorithm appeared to do a good job of selecting non-overlapping features. For instance, it only chose one of the many metrics we had for family income. So another approach would be to re-run the other models on just the top 20 features that LogisticRegressionCV found most important and see how that affected their accuracy.

There are also more sophisticated techniques for feature reduction, such as recursive feature elimination, that could be fruitful to explore.

I wasn't able to complete a bunch of work I did on how correlations changed over time during the pandemic. Some correlations stayed steady from year one to year two while others changed, sometimes dramatically. Correlations could change just because of chance, so I wanted to pin down how unlikely certain changes in correlation were to occur due to chance alone.

CONCLUSION

I was able to train a fairly accurate machine learning model to classify high COVID fatality rates in year 2 of the pandemic. The LogisticRegressionCV algorithm achieved an accuracy of .755 on all US counties.

Year two fatality rates were much more predictable than year one rates. The best model of year one fatalities achieved an accuracy of only .652. Correlations were also higher in year two than in year one.

A model trained on the smaller subset of counties with over 50K residents achieved an even higher accuracy score of .811, likely due to less uncertainty in the underlying data.

COVID vaccination coverage wasn't the factor most highly correlated with COVID fatality rates in year two of the pandemic. That was age-adjusted mortality (and its twin, life expectancy). In other words, places where people tended to live a long time before COVID tended to have fewer fatalities because of COVID – healthwise, the rich got richer. Money was also a big factor, with Household Income being strongly correlated with fatality rates.

Vaccination coverage also didn't end up being the most important factor used by the models for prediction.

2016 Republican vote share was the one factor that five different machine learning models identified as being a top predictor of COVID fatality rates. "% Diabetic" and "% Some College" were also highly ranked as predictors. The models didn't simply select the mostly highly correlated features for classification purposes, so it's a different look at how the data is connected to the outcomes beyond linear correlation.

More work in the future is needed to understand why 2016 voter preferences ended up being such a strong predictor of deaths in 2021. Vaccination status is probably a big part of it: higher Republican vote share was strongly correlated with lower vaccination rates, and lower vaccination rates were correlated with higher COVID fatality rates. So it's somewhat of a proxy for vaccination status. Yet the machine learning models chose vote share as the main factor to use for classification, rather than vax rates directly.

Beyond that, it isn't obvious. 2016 Republican vote share isn't strongly correlated with other health factors that would put people at high risk for COVID (such as "% Diabetic", or "% over age 65").

This suggests to me that vote share explains other factors besides vaccination status that may also affect COVID fatality rate. For instance places with a higher Republican vote share may have practiced less masking or social distancing. More data and analysis on this front is needed.

REFERENCES

[CDC] CDC Museum COVID-19 Timeline. (2022). cdc.gov. Retrieved June 7, 2022 from https://www.cdc.gov/museum/timeline/covid19.html#Early-2021

[countyhealthrankings.org] County Health Rankings. (2019) *2019 Reports.* Countyhealthrankings.org. Retrieved May 24, 2022 from https://www.countyhealthrankings.org/reports

[CSSEGISandData] CSSEGIS. (2022) *Time Series COVID19 Deaths - US.* CSSEGISandData github page. Retrieved May 24, 2022 from https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_US.csv

[Durfee] Durfee, C. (2022) *COVID Fatality Rates*. csdurfee github page. Retrieved June 18, 2022 from https://github.com/csdurfee/COVID_fatality_rates

[Gao 2021] Gao, J., & Radford, B. J. (2021). Death by political party: The relationship between COVID-19 deaths and political party affiliation in the United States. *World medical & health policy*, *13*(2), 224–249. https://doi.org/10.1002/wmh3.435

[Lin 2022] Lin Q, Paykin S, Halpern D, Martinez-Cardoso A, Kolak M. Assessment of Structural Barriers and Racial Group Disparities of COVID-19 Mortality With Spatial Analysis. *JAMA Netw Open.* 2022;5(3):e220984. doi:10.1001/jamanetworkopen.2022.0984

[MEDSL] MIT Election Data and Science Lab. (n.d.) *County Presidential Election Returns 2000-2016.* github.com/MEDSL. Retrieved May 24, 2022, from https://github.com/MEDSL/county-returns/blob/master/countypres_2000-2016.csv

[Millett 2020] Millett A. et al. Assessing differential impacts of COVID-19 on black communities, *Annals of Epidemiology*, Volume 47, 2020, Pages 37-44, ISSN 1047-2797. https://doi.org/10.1016/j.annepidem.2020.05.003

[Ortaliza 2021] Amin K., Ortaliza J., Cox C., Michaud J., Kates J. (2021). *COVID-19 mortality preventable by vaccines.* Peterson-KFF Health System Tracker. Retrieved May 24, 2022, from https://www.healthsystemtracker.org/brief/covid19-and-other-leading-causes-of-death-in-the-us/

[SVI] CDC (2019). *CDC/ATSDR Social Vulnerability Index.* www.atsdr.cdc.gov. Retrieved May 24, 2022, from https://www.atsdr.cdc.gov/placeandhealth/svi/index.html

[Tan 2021] Tan AX, Hinman JA, Abdel Magid HS, Nelson LM, Odden MC. Association Between Income Inequality and County-Level COVID-19 Cases and Deaths in the US. *JAMA Netw Open.* 2021;4(5):e218799. doi:10.1001/jamanetworkopen.2021.8799

[vaccinetracking.us] Vaccinetracking.us (2022). https://www.vaccinetracking.us/. Retrieved June 20, 2022 from https://www.vaccinetracking.us/data.html

[Wikipedia 2022] Wikipedia (2022). *List of US Counties by Per Capita Income.* Retrieved May 24, 2022 from https://en.wikipedia.org/wiki/List_of_United_States_counties_by_per_capita_income

[Wood 2021] Wood, D., Brumfiel, G. (2021). *Pro-Trump counties now have far higher COVID death rates.* Wbur.org. Retrieved May 24, 2022, from https://www.wbur.org/npr/1059828993/data-vaccine-misinformation-trump-counties-covid-death-rate