Massachusetts
Institute of
Technology

Sloan School of Management

15.072 - Analytics Edge

# Time-Series Forecasting for COVID-19 Death Prediction and Policy-Making

*Authors*
Alexandre Berkovic
Max Petruzzi
Sri Reddy
Osho Yonzon

*Professor*
Bart Van Parys

December 9, 2022

## Abstract

At the turn of 2020, COVID-19 became responsible for infecting hundreds of millions, causing the death of millions and the downfall of economies across the world. In hindsight, there were mistakes made – policy-wise, at the hospital level, and at the individual level – that exacerbated this negative impact. This study proposes a class of analytical models to help predict COVID deaths at the hospital level in large populations, which in our case consists of the United States of America. Most importantly, the goal is to put in place adequate short to medium-term prediction models (7 days in advance) for forecasting the number of future deaths. The reason behind forecasting resides in optimizing the strategic planning of decision-makers in the public health domain, allowing them to pursue policies that limit the number of deaths, manage the number of hospitalized patients, and better distribute both human and material resources to alleviate medical workers.

In this paper, we propose forecasting models comprising seasonal autoregressive integrated moving averages (SARIMA), long short-term memory networks (LSTM), random forests, and optimal regression trees (ORT) to provide both high quantitative performance as well as highly interpretable results. These models are assessed for the accuracy of time series prediction of COVID deaths by using validation metrics (MAE, MAPE, MSE, RMSE) which are complemented by an explanation of our results. To propose models that do not overfit and generalize well, we perform an exhaustive data cleaning process and a thorough feature selection procedure by using variance inflation factor (VIF) in order to reduce the number of decision variables from 135 to 5. Additionally, to understand the variables and information we lacked, we also touch on how confounding variables, such as quarantine policies, could affect our model and results.

On the one hand, in terms of purely quantitative results, the devised LSTM with 7 days estimation period clearly captures the trends of the data and leads to an out-of-sample MSE of 0.114. The SARIMA performs with an MSE of 1.834; however, by properly accounting for both trend and seasonality, this model is still strong. In terms of qualitative results, the optimal regression trees use linear regression within the leaves to reach an out-of-sample $R^2$ of 0.803 and allow for clear comprehension of causes leading to a given number of COVID-related deaths. Furthermore, the RandomForest direct time-series forecasting model achieved an out-of-sample MAE of 0.028. It also allows for interpretation of the relative importance of features in predicting COVID deaths for that given week. We then used the Shapley value, which reveals a variable's contribution to a certain model, with XGBoost to verify our results of RandomForest and Optimal Regression Trees.

The aforementioned methods and results will be developed in this paper by following the hereinafter structure: Introduction 1, Exploratory Data Analysis 2, Methods & Results 3, Analysis & Insights 4, and Conclusion 5. These results are also shown on a live dashboard, which is linked in the references.

# 1 Introduction

Pandemics are predictable catastrophes that recurrently occur in history, with major worldwide outbreaks taking place about once a century. Despite this foreshadowing, the United States, and the world at large were woefully prepared for the COVID-19 pandemic. Before the coronavirus outbreak, Congress had allotted $42 billion to the NIH, which is only 5% of the $738 billion assigned to the Department of Defense.[1] In addition to funding additional research for preventative measures, there is an immense need for streamlining healthcare operations during the pandemic. Much of the devastation caused by COVID-19 originated from a lack of preparedness in medical infrastructures and supply chains. Some of these deficiencies manifested as a lack of sufficient supplies like ventilators and therapeutics, staff shortages, and hospitals operating above capacity. Because pandemic diseases are recurrent due to their inevitability and periodicity, avoiding these failures should have been at the forefront of public health concerns. A positive to take away from this tragedy is that the data recorded during the COVID-19 pandemic allows us to prepare for future pandemics more intelligently. Instead of simulating how deficiencies such as staff shortages or overcapacity impact pandemic outcomes, there are nearly three years of data including COVID deaths and these relevant features. Analyzing hospital data with the utilization of machine learning techniques will have a vital role in predictive and preventive healthcare for future outbreaks.

# 2 Exploratory Data Analysis

## 2.1 Data Overview

### 2.1.1 Data Description

This dataset from the U.S. government contains time series data from January 2020 to October 2022. Each observation is state-aggregated, meaning it contains data from all hospitals throughout a particular state, for a particular day. There are observations for every day; however, at the beginning of the dataset, not every state has an observation associated with it on a given day. With this raw dataset, we have 52,445 observations with 135 features. Since we are interested in the number of deaths due to COVID, our response variable is `deaths_covid`, representing the number of deaths recorded by the respective state each day. A graph of the COVID deaths by state over the years is shown in Fig. 19.

### 2.1.2 Data Cleaning

The key steps in the data-cleaning process are highlighted here. Features with over 50% of NaN values were removed from the dataset. Territories such as the Virgin Islands and Puerto Rico were removed from the dataset (not enough data). Features related to pediatrics were removed, as medically the pediatric population and adult population are considered distinct populations and should not be grouped together. The dataset was truncated for earlier dates that do not have data for all 50 states, preventing those from skewing the data. KNNs were used for data imputation to fill the remaining missing numeric values with nearest neighbors set to 5. The dataset was aggregated by date so that we were no longer stratifying data by states. Every column has a corresponding "coverage" column that represents how many hospitals reported that data (i.e. `deaths_covid` of 40 with `deaths_covid_coverage` of 3 indicates that 3 hospitals reported a total 40 deaths that day). Since we noticed the total number of reporting hospitals changes each day, every column was normalized by the number of hospitals reporting said column.

## 2.2 Feature Selection

### 2.2.1 Multicollinearity

After manually selecting and creating features, we examined the collinearity in the data, notably using Variation Inflation Metric (VIF). By using VIF we checked how much variance in the data was inflated which is caused by multicollinearity. Thus, we examined the VIF for each variable except our target, which shows us how inflated each variable is by the presence of the other variables [2]. The equation we employed is shown below:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{1}$$

Where $R_i^2$ represents the coefficient of determination for the $i^{th}$ variable against the other independent variables. By leveraging the coefficient of determination, we simply investigated if there were linear relationships between the independent variables. This allows us to find the variable that has the most relationships between other variables and hence has the most inflation. In short, a VIF of 1 represents zero correlation with the $i^{th}$ variable and the other variables. We calculated the initial VIF values, then we iteratively dropped variables rerunning the algorithm. The standard practice is to remove all variables until the highest VIF amongst the predictor variables is below 4. However, to target both interpretability and minimum multicollinearity, we rested on a much higher VIF [2]. These results including the initial and end VIF results are shown in Appendix A 5.

### 2.2.2 Confounding Variables

Following feature selection, we wanted to understand our future models' true potential and flaws. Thus, we examined confounding variables: predictors that are not only associated with other independent variables but also the dependent variable. In short, these variables "confound the [true] relationship between two variables" [3]. Predicting COVID deaths using aggregated data from the United States does not go without assumptions and unaccounted confounders. To illustrate how a confounder such as quarantine policy impacts our model, we showcase a visual in Appendix H Fig. 18. This is not to say our model predictions and interpretations will fail. Instead, this is a tool to understand the limits of our model; we do not account for external factors that affect both the independent variables and dependent variables. To further examine this we must understand there are different confounding variables for each state. In Appendix H Fig. 19, there visibly exists different phases of COVID-19; though we are not limited to only the Delta and Omicron variants, the CDC reports these phases had the most impact on the United States. With these phases come different quarantine policies, rules for vaccinations, and other legislation. Though our models do not effectively account for this, our goal is to measure and explain the variance in the data causing the spikes in deaths solely using hospital data.

# 3 Methods and Results

## 3.1 Prediction

### 3.1.1 SARIMA

The time series nature of the data allows us to leverage auto-regressive methods. A strong modeling approach to time series data is Autoregressive Integrated Moving Average (ARIMA), which is able to forecast a target for the next specified time period. The standard procedure of this model is to include the variables $p$, $d$, $q$, where $p$ is the order of our autoregressive model, $d$ is the degree we difference, and $q$ is the order of our moving-average model. However, the standard ARIMA model does not account for seasonality, which are the trends that occur in certain time periods which is why we adopt Seasonal ARIMA (SARIMA). It does so by adding new variables $P$, $D$, $Q$ which are the same as ARIMA but now with respect to seasonality and by adding the variable $m$ representing the steps for seasonality. It is common practice to represent the SARIMA variables as $(p, d, q)x(P, D, Q, m)$, where the first group of variables represents the traditional variables and the second represents the SARIMA addition.

To implement SARIMA, we first took the weekly moving average of the time series data, as shown in Fig. 9 in Appendix F. With the moving average, we graphed the trend, seasonality, and residuals using seasonal decomposition in order to understand whether seasonality truly exists in our data. As shown in Appendix F, we see with the dip and rise of deaths during different seasons reveal seasonality exists. To cement our observations with metrics, we checked if the data was stationary (i.e. if the data is not affected by trends or seasonality), using the Dickey-Fuller test. This test's null hypothesis is that there is a unit root, which is evidence of non-stationary data. We failed to reject the null hypothesis for our data, leaving us to conclude our data is stationary and affected by trends and seasonality. Accounting for annual seasonality and a weekly trend, we were able to achieve stationary data.

Finally, auto-correlation and correlation plots are shown in Fig.12 13 of Appendix F, to find the best values for the four variables in the SARIMA model. This helps us understand the nature of seasonality and lags in the data. To solidify these results, we ran different combinations of possibilities for the model's parameters and chose the ones yielding the lowest Akaike Information Criterion (AIC). We found that (0,1,0)x(0,1,1,52) was the best combination of SARIMA variables yielding an AIC of 4. Shown below in Fig. 1 is how the model performs when predicting one week in advance.

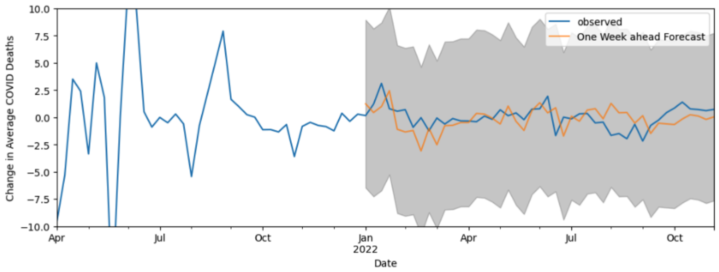


Figure 1: SARIMA Model Forecast

With a testing period starting in January 2022, we obtain a Mean Squared Error (MSE) of 1.83 and a Root Mean Square Error (RMSE) of 1.35. Additionally, an ARIMA model is built only by taking into account trends and not seasonality. However, we find this was naive, for it was not accounting for the yearly trends. This is shown in Appendix F.

### 3.1.2 Long Short-Term Memory Network (LSTM)

The task of predicting the number of COVID deaths at a given time period is inherently a time series regression problem. Considering the large number of variables our model has to work with even after implementing VIF, developing a multivariate LSTM (Long Short-Term Memory) network seems like an appropriate direction to take.

With the data at hand, we set aside the variable `deaths_covid` as our target variable. The aim here is to forecast one week in the future which represents 7 periods (as one period is equivalent to a day). Hence, for each training instance in our network, the model is given a *sequence* of observations. To do this, we implement `ShallowRegressionLSTM` in PyTorch with the standard regression objective of MSE (mean-squared error); the learning rate ($10^{-5}$) and the number of hidden units (32) are chosen arbitrarily to best suit our model but could also be optimized by using a GridSearch. Nevertheless, the number of epochs to prevent the model from over-fitting is chosen in order to minimize the test loss which occurs for 1309 as shown by Fig. 3 in the Appendix.
With the data and defined parameters, we can now run the model and obtain a predictive 1-week forecast as shown by Fig. 2. A scaled curve of the model applied on the test set can be seen in Fig. 4 of the Appendix.

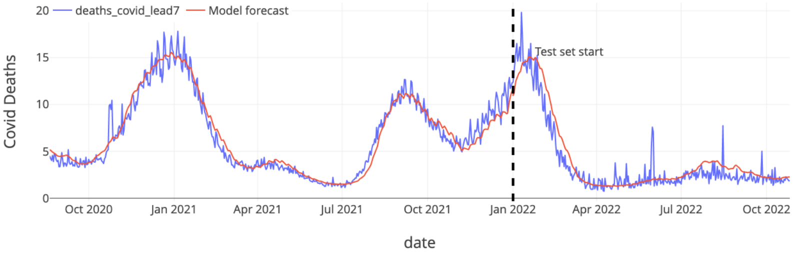


Figure 2: Average number of COVID deaths per hospital (in blue) and model forecast for 7 days lag (in red)

As pointed out by Tab. 3 in the Appendix, the model tends to overfit the training data but still performs well on the test data. We can see that the LSTM accurately captures the trends in the data but struggles to mimic extrema and abnormalities such as very abrupt increases or decreases. Nevertheless, it should be noted that on the test set, the model will often over-predict the average number of deaths which is better than the opposite. Indeed, in the case of prediction for epidemics to help policy-makers make decisions, underestimating the gravity of the disease's spread might lead to much more important consequences than the opposite. Additionally, we look into using the LSTM to predict the data 14 days in advance and present the results in Tab. 3. Although the model still captures the general trend, it is not as accurate and struggles to pick up on rapid changes in the slope.

## 3.2 Interpretation

### 3.2.1 RandomForest

From the perspective of policymakers and hospital administrators during a global pandemic, time-series prediction models are much less actionable if they are a black box without interpretability. On the other hand, if one is also provided a list of factors deemed most important in generating that week's prediction, these factors can be acted upon directly. To this end, direct time-series forecasting was utilized using RandomForest (RF) models to predict one week into the future, from 10/26/22 to 11/2/22. Using the 7 days as lag, the entire COVID data set prior to 10/26/22 was used to train 7 RandomForest models. Each of the 7 RF models was used to predict one of the days' COVID deaths within the 7-day horizon. The model forecasted the COVID deaths (Fig. 8 in Appendix E) with a Mean Absolute Percent Error (MAPE) of 12.2% and an RMSE of 0.03 relative to the true COVID deaths for that week.
RandomForest was used because, although less interpretable than CART, the decision-tree structure allows for the straightforward ranking of feature importance. Because this forecasting method utilizes one RF model for each day in the prediction horizon, we obtain 7 variable importance outputs for our 7-day horizon. It should be noted that, due to the 7-day lag, the feature importance on a given day `X` in our horizon represents how well the values of this feature over the past 7 days can predict COVID deaths on day `X`. Tab. 4 in Appendix E depicts the top-5 features based on average importance rating, averaged among each of the 7 days in the horizon. Interestingly, regardless of which day of the horizon one is predicting, `is_critical_staffings_shortage_expected` and `number_of_inpatient_beds_used` are always the top-2 features, respectively. Further, `is_critical_staffings_shortage_expected` is substantially more important in predicting COVID deaths than any other variable, including the current week's number of COVID deaths, which typically ranks $3^{rd}$ in order of importance.
It is important to clarify that a feature's ability to predict COVID deaths does not definitively imply a causal relationship between the two. For example, perhaps spikes in COVID transmission could cause both more staff to miss work due to COVID infection and cause more overall deaths. However, there are several reasons why we believe that this confound does not eliminate the potential for a prominent causal relationship between staffing shortages and COVID deaths.
First, the correlation coefficient between `anticipated_staff_shortages` and a particular day's COVID deaths is -0.06; it is also only 0.32 with the `number_of_inpatient_beds_used` for COVID patients. This means that anticipated staffing shortages are unlikely to be merely a function of spikes in COVID transmission, as proposed earlier. Second, as mentioned previously, anticipated critical staffing shortages are more important in predicting the following week's COVID deaths than even the current week's COVID deaths. Lastly, although certainly not foolproof, there are plenty of precautions taken to prevent staff from being infected with COVID. So, we would not intuitively expect staff shortages to be strongly

correlated with spikes in COVID transmission. Unfortunately, although staff shortages during COVID have been well-documented, their impact on the number of COVID deaths has not yet been well-quantified by literature [4]. We argue that this is something worth investigating further in future studies.

### 3.2.2 Optimal Regressive Trees

Similar to CART, Optimal Regressive Trees is an ensemble learning method that enables regression tasks to be performed with an increased level of interpretability. However, here, instead of building the decision tree using a recursive approach based on a greedy heuristic, the tree is built at once by using Mixed Integer Optimization. Once again, we set aside the variable `deaths_covid` as being our target variable and perform a GridSearch to determine the optimal parameters to use such as the tree's maximum depth, the minimum number of points per leaf and the coefficient of complexity.

Firstly, regular Optimal Regressive Trees were leveraged in order to determine the number of daily COVID deaths in hospitals. Doing so allows policymakers to understand how certain variables and the synergies between them lead to higher or lower death counts. An initial ORT is computed with a GridSearch and leads to an in-sample $R^2$ of 0.905 and an out-of-sample $R^2$ of 0.765. Following that, it was decided to implement linear predictions in the leaves of the tree to increase the performance of the model, which led to an in-sample $R^2$ of 0.935 and an out-of-sample $R^2$ of 0.803, representing a 5% improvement. This tree is highly interpretable, as shown in Fig. 5 in Appendix I, we can see that if the utilization of ICU beds for adults is above 14.8% and critical shortage within the week is anticipated, then the model estimates about 11 deaths per hospital per day. Moreover, ORTs with Hyperplanes are also trialed but their performance does not improve the model's $R^2$ (out-of-sample $R^2$ of 0.802) and their output is not interpretable as it uses multiple variables with coefficients to perform the splits. The computed ORT-H can be viewed in Fig. 6 in the Appendix.

## 4 Analysis and Insights

Although the task at hand was treated as a regular time series forecasting problem, the truth is that modeling the spread and impact of a disease is much more complex. Indeed, although the seasonality (the period between peaks and valleys) might be the same, the scale of those extrema is very variable as those increase drastically at the beginning of the epidemic and then decrease once the peak has been reached (due to confounding factors). Therefore, factors such as the number of susceptible individuals to be contaminated, and the number of infected and recovered individuals, are essential variables to create a better-performing model. With more specific data, more specialized models could have been implemented.

SIR models are the most common models to describe the temporal dynamics of an infectious disease in a population and account for all the variables mentioned earlier by setting ordinary differential equations that describe the *change* in counts at an instance of time. The advantage of SIR is that by taking parameters such as transmission coefficient, it can effectively model policies like stay-at-home which tends to diminish that parameter. An example of SIR curves can be seen in Fig. 7 of the Appendix. It should be noted that, although too sparse to create a full model, the dataset could be leveraged to develop a discrete-time Markov Chain model relying on SIR parameters as explained in a paper on Generalized Markov Models of Infectious Disease Spread by Yaesoubi et al..

As noted in the prior sections, predictive strength and interpretability are both relevant to the applicability of these models. SARIMA and LSTM models provide accurate forecasting of COVID deaths in advance. In the short term, this forecasting can directly connect to forecasting of resources required, such as personal protective equipment, hospital beds and staff, and other supplies, and thus improve logistical decision-making in the day-to-day.

However, in regard to long-term policy and decision-making, interpretable models are much more important. Through RandomForest, we highlight critical staffing shortages and the number of inpatient beds used as important variables in the forecasting of COVID deaths. While we note uncertainty that there is a causal relationship present, our analysis indicates that we certainly cannot exclude a causal relationship and that at a minimum, these two factors be considered in future studies. Moreover, we recommend that these two factors be highly considered in policy-making to improve preparedness during pandemics.

Furthermore, ORTs provide a visually interpretable way for decision-makers to understand the impact of certain factors on predicting COVID deaths. Beyond variable importance, ORTs provide thresholds where splits occur, and the numeric value of these thresholds helps policymakers understand not just where to improve, but the level of improvement necessary.

## 5 Conclusion

The different models developed in this study allow us to conclude that when predicting COVID deaths, LSTMs outperform other predictive methods such as SARIMA, and should be used as the primary method for prediction. However, with more diverse data in the future, SARIMA could outperform LSTMs, for seasonality and trends in the data may prove to be more potent. We also conclude that the most important variables that contribute to hospital COVID deaths are the inpatient beds used and critical staffing shortage, which is shown through both the RandomForest and Optimal Regression Tree models. Though these results may seem obvious when looking back at the height of the pandemic, they cement the fact that hospitals were unprepared for a worldwide outbreak. In the future, these models can be performed on a per-state basis. This way, hospitals in particular regions can understand precisely what their weaknesses are, improving on them to prepare for unforeseen circumstances.

# Appendix A: VIF

Table 1: VIF Values before VIF Feature Selection

| Feature | VIF |
|---|---:|
| critical_staffing_shortage_today_not_reported | 5.653332e+07 |
| critical_staffing_shortage_anticipated_within_week | 5.651664e+07 |
| inpatient_beds_utilization_utilization | 4.223471e+04 |
| inpatient_beds_used | 4.125496e+04 |
| inpatient_beds | 3.512014e+04 |
| adult_icu_bed_utilization_utilization | 1.690436e+04 |
| staffed_adult_icu_bed_occupancy | 1.139998e+04 |
| total_staffed_adult_icu_beds | 8.717810e+03 |
| critical_staffing_shortage_anticipated_within_week | 7.401827e+03 |
| total_adult_patients_hospitalized_confirmed_covid | 7.015022e+03 |
| inpatient_bed_covid_utilization_utilization | 6.811808e+03 |
| total_adult_patients_hospitalized_confirmed_and | 6.517941e+03 |
| critical_staffing_shortage_today_no | 5.777808e+03 |
| staffed_icu_adult_patients_confirmed_and_suspected | 5.063853e+03 |
| staffed_icu_adult_patients_confirmed_covid | 4.597697e+03 |
| percent_of_inpatients_with_covid_utilization | 4.408997e+03 |
| previous_day_admission_adult_covid_confirmed | 1.947843e+03 |
| inpatient_beds_used_covid | 1.410311e+03 |
| adult_icu_bed_covid_utilization_utilization | 1.068796e+03 |
| critical_staffing_shortage_anticipated_within_week | 1.067663e+03 |
| critical_staffing_shortage_today_yes | 7.705780e+02 |
| previous_day_admission_adult_covid_confirmed_60-69 | 6.961388e+02 |
| previous_day_admission_adult_covid_confirmed_70-79 | 5.727656e+02 |
| previous_day_admission_adult_covid_suspected | 5.207502e+02 |
| previous_day_admission_adult_covid_confirmed_50-59 | 4.231249e+02 |
| previous_day_admission_adult_covid_suspected_60-69 | 3.237052e+02 |
| previous_day_admission_adult_covid_suspected_70-79 | 2.963050e+02 |
| previous_day_admission_adult_covid_confirmed_80+ | 2.858520e+02 |
| previous_day_admission_adult_covid_suspected_80+ | 2.407732e+02 |
| previous_day_admission_adult_covid_suspected_50-59 | 2.111917e+02 |
| previous_day_admission_adult_covid_confirmed_30-39 | 2.107643e+02 |
| previous_day_admission_adult_covid_confirmed_40-49 | 2.072935e+02 |
| previous_day_admission_adult_covid_suspected_40-49 | 1.293927e+02 |
| previous_day_admission_adult_covid_suspected_20-29 | 1.118458e+02 |
| previous_day_admission_adult_covid_suspected_30-39 | 1.083590e+02 |
| previous_day_admission_adult_covid_confirmed_20-29 | 9.016365e+01 |
| hospital_onset_covid | 4.228266e+01 |
| previous_day_admission_adult_covid_confirmed_unknown | 2.159156e+01 |
| previous_day_admission_adult_covid_suspected_18-19 | 1.816665e+01 |
| previous_day_admission_adult_covid_suspected_unknown | 6.605897e+00 |
| previous_day_admission_adult_covid_confirmed_18-19 | 1.303630e+00 |

Table 2: VIF Values after VIF Feature Selection

| Feature | VIF |
|---|---:|
| inpatient_beds_used_covid | 113.996285 |
| total_adult_patients_hospitalized_confirmed_and... | 78.190769 |
| adult_icu_bed_covid_utilization_utilization | 48.977481 |
| hospital_onset_covid | 17.429036 |
| critical_staffing_shortage_anticipated_within_w... | 7.738703 |
| deaths_covid | 7.660202 |

# Appendix B

## Loss as a function of Number of Epochs

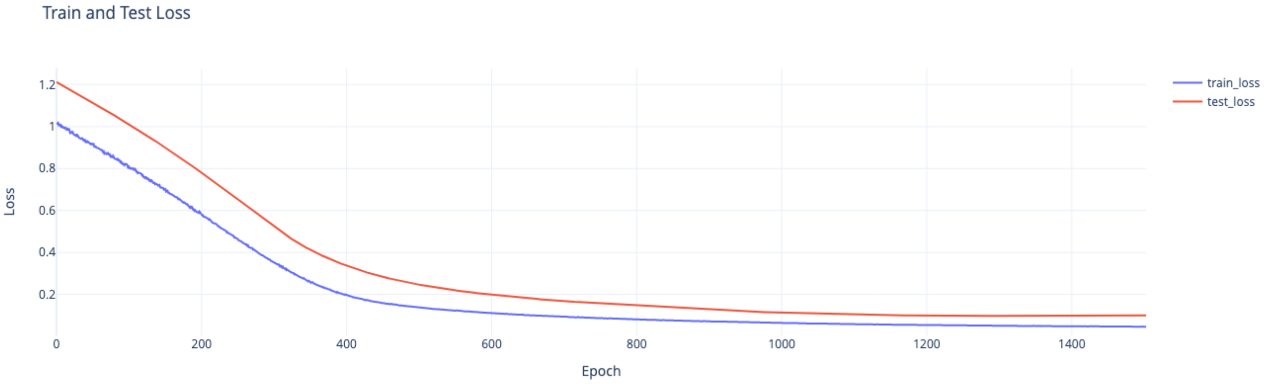


Figure 3: Train and test sets loss as a function of number of epochs

## Scaled Forecast Curve on Test Set

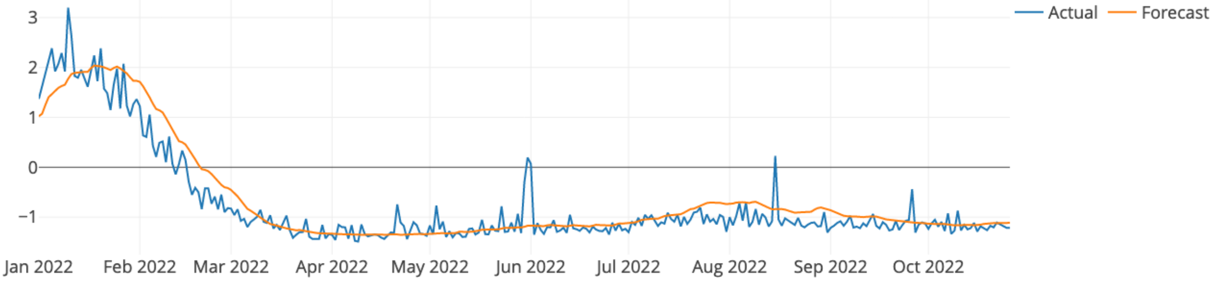


Figure 4: Scaled actual and forecast curves on test set as a function of time

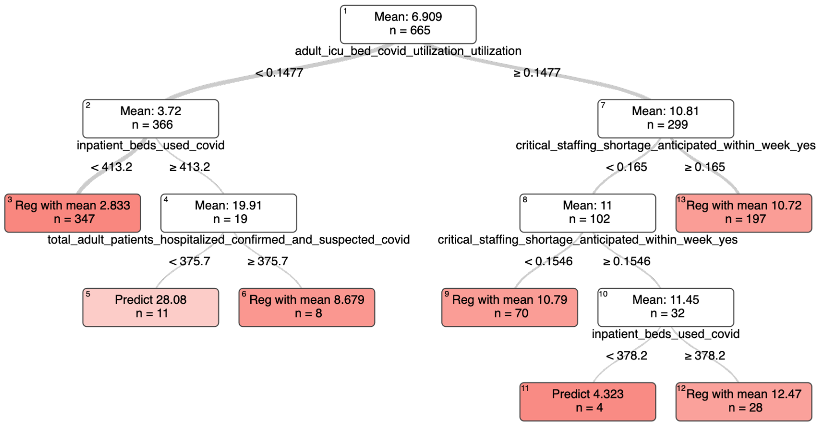
## Validation Metrics for LSTM Model

Table 3: Validation Metrics of the LSTM Network for 7 Days Lag and 14 Days Lag

| Lag | Metric | MAE | MAPE | MSE | RMSE |
|---|---|---|---|---|---|
| 7 days | Train set | 0.172 | 82.4% | 0.054 | 0.231 |
| 7 days | Test set | 0.231 | 44.3% | 0.114 | 0.337 |
| 14 days | Train set | 0.227 | 182.2% | 0.089 | 0.299 |
| 14 days | Test set | 0.219 | 76.4% | 0.142 | 0.377 |

# Appendix C

## Interpretable Decision Tree for Analyzing Feature Impact on COVID Deaths



1 Mean: 6.909 n = 665
adult_icu_bed_covid_utilization_utilization
< 0.1477
≥ 0.1477
2 Mean: 3.72 n = 366
inpatient_beds_used_covid
< 413.2
≥ 413.2
7 Mean: 10.81 n = 299
critical_staffing_shortage_anticipated_within_week_yes
< 0.165
≥ 0.165
3 Reg with mean 2.833 n = 347
4 Mean: 19.91 n = 19
total_adult_patients_hospitalized_confirmed_and_suspected_covid
< 375.7
≥ 375.7
8 Mean: 11 n = 102
critical_staffing_shortage_anticipated_within_week_yes
< 0.1546
≥ 0.1546
13 Reg with mean 10.72 n = 197
5 Predict 28.08 n = 11
6 Reg with mean 8.679 n = 8
9 Reg with mean 10.79 n = 70
10 Mean: 11.45 n = 32
inpatient_beds_used_covid
< 378.2
≥ 378.2
11 Predict 4.323 n = 4
12 Reg with mean 12.47 n = 28

Figure 5: Optimal Regression Tree with Linear Regression in the Leaves

## Optimal Regressive Tree with Hyperplanes

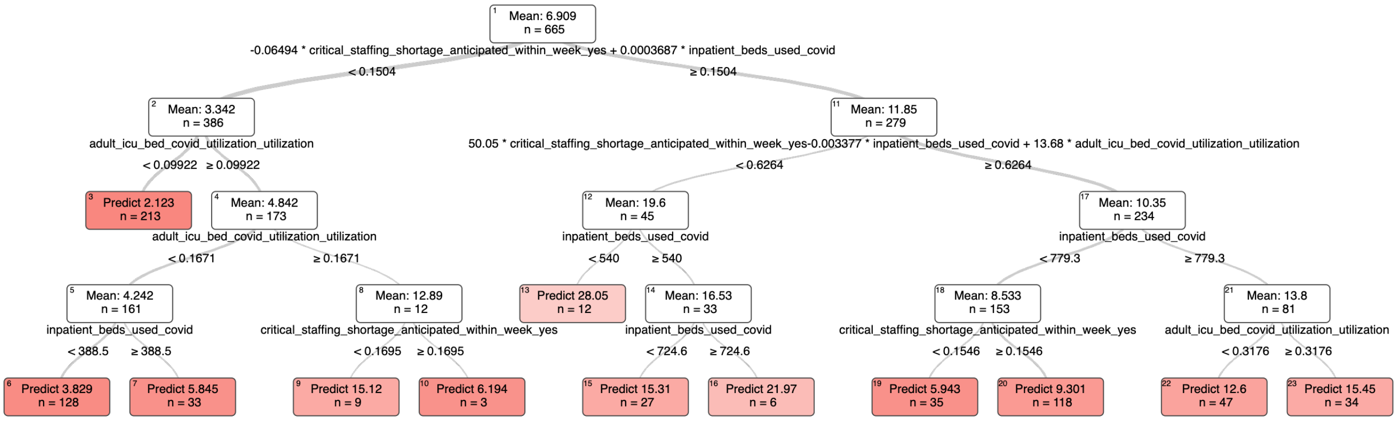


Figure 6: Optimal Regressive Tree with Hyperplanes

# Appendix D

## SIR curves

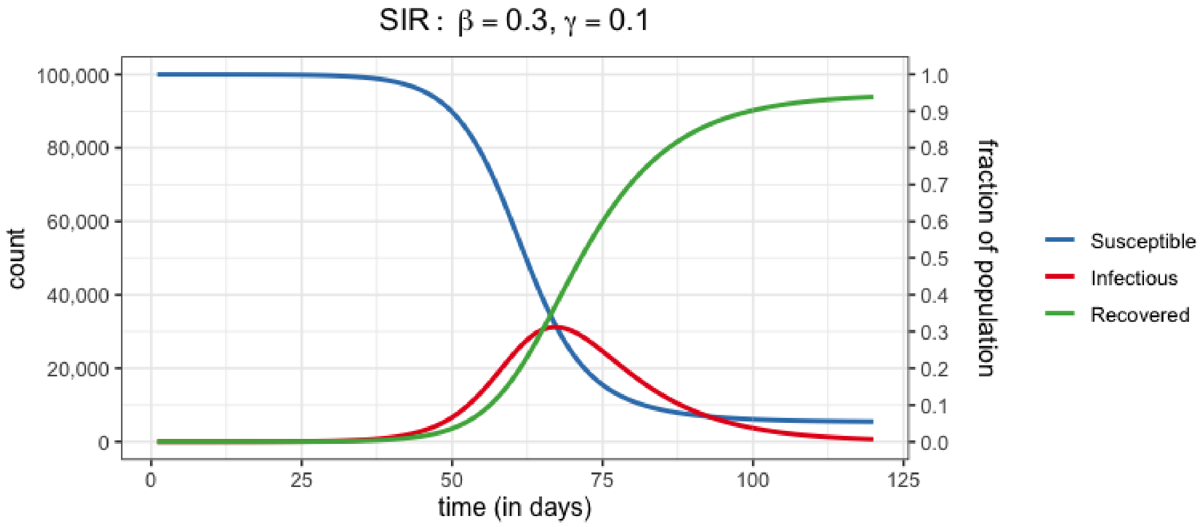


Figure 7: SIR curves from the continuous time model that specifies a set of three ordinary equations

# Appendix E

## RandomForest Interpretable Forecasting

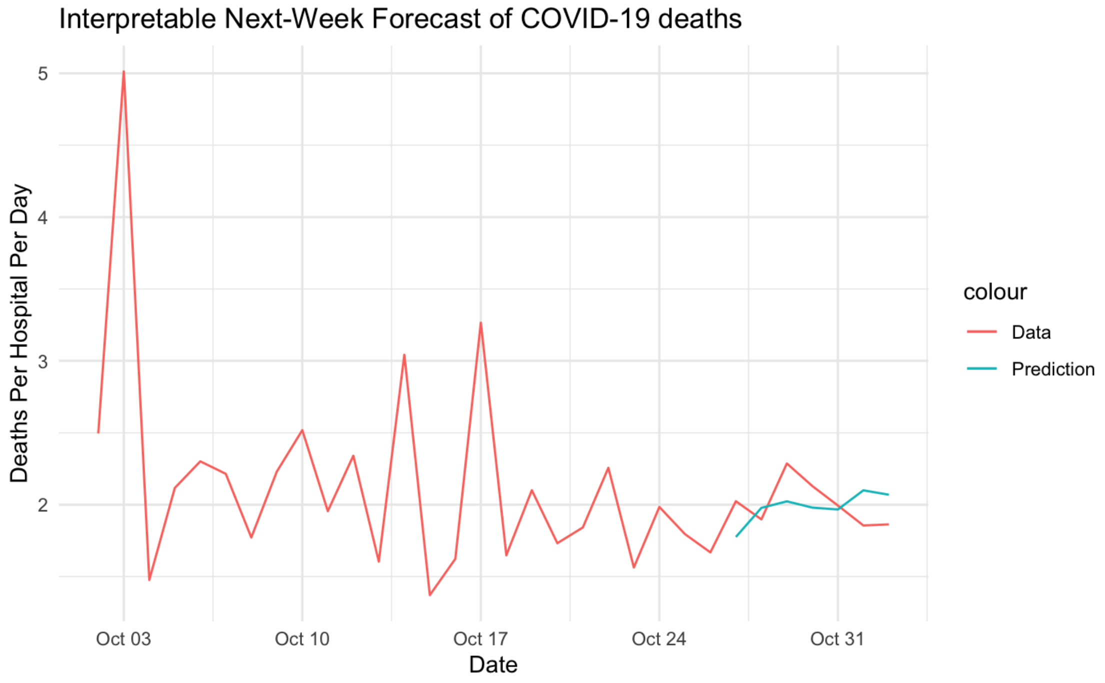


Figure 8: Next-week forecasted deaths using seven RandomForest models, one for each day in the prediction horizon (10/26/22 to 11/2/22)

| Feature Name | Average Importance Rating | Average Rank |
|---|---|---|
| Is Critical Staffing Shortage Expected | 494.36 | 1 |
| Number of Inpatient Beds Used | 14.4 | 2 |
| Number of COVID Deaths | 8.50 | 3.14 |
| Adult COVID ICU Bed Utilization | 2.52 | 4.71 |
| Previous Day's Admission of Suspected COVID (80+ yrs old) | 0.85 | 6 |

Table 4: Top five features based on importance rating, averaged among the 7 RandomForest models used to predict COVID deaths for each day in the prediction horizon (10/26/22 to 11/2/22)

# Appendix F: SARIMA Results

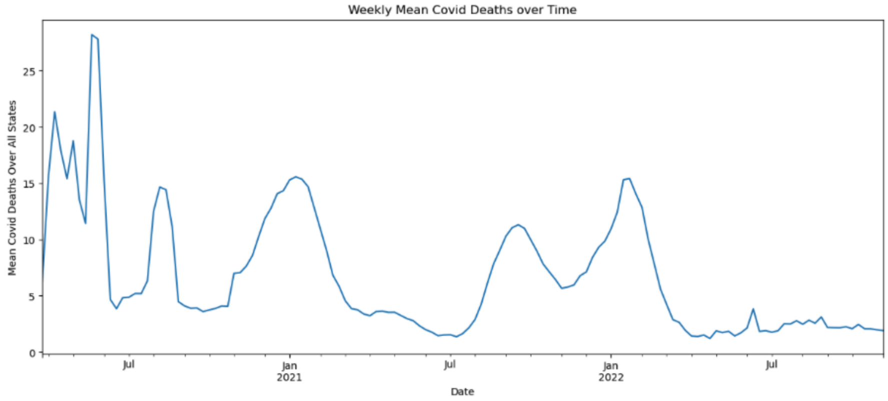


Figure 9: Weekly COVID Deaths Average

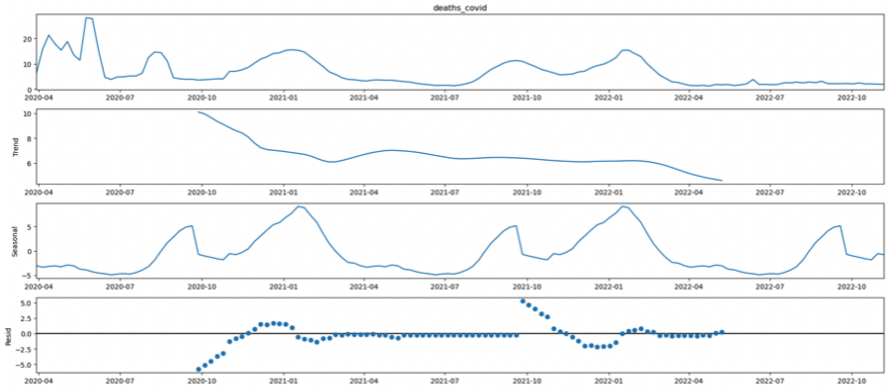


Figure 10: Decomposition of Weekly COVID Deaths Average

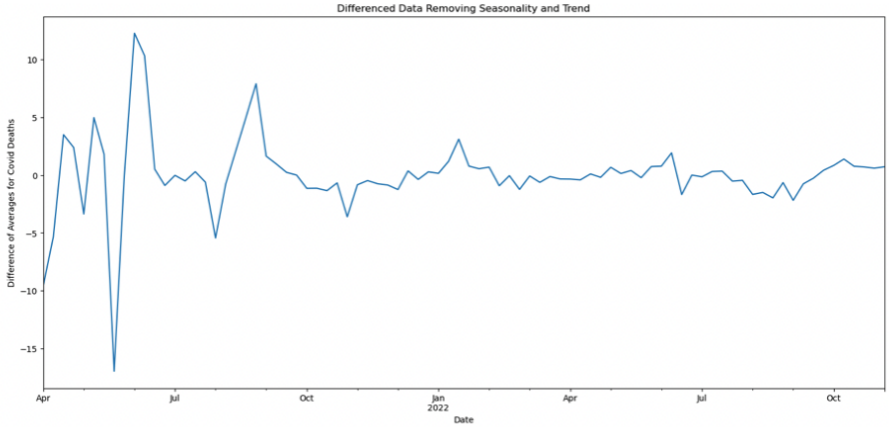


Figure 11: Weekly COVID Deaths Average Accounting for Weekly Trend and Yearly Seasonality

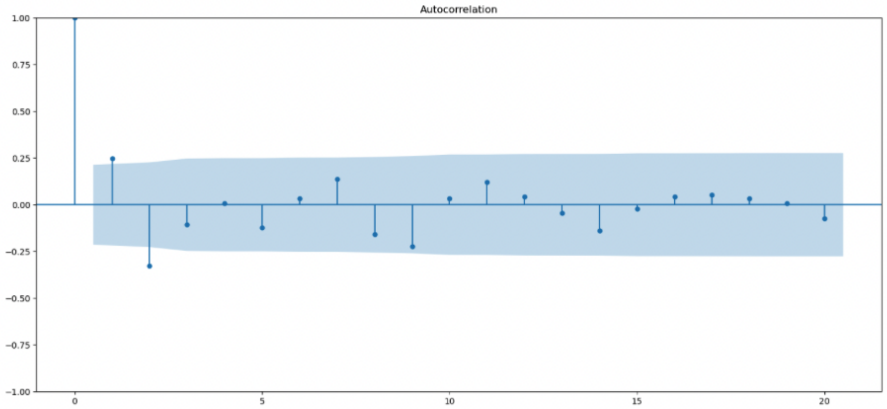


Figure 12: Auto Correlation Depiction

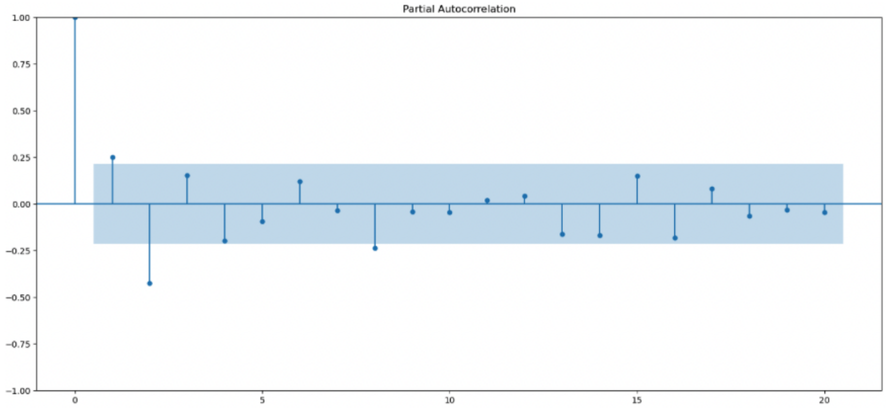


Figure 13: Partial Auto Correlation Depiction

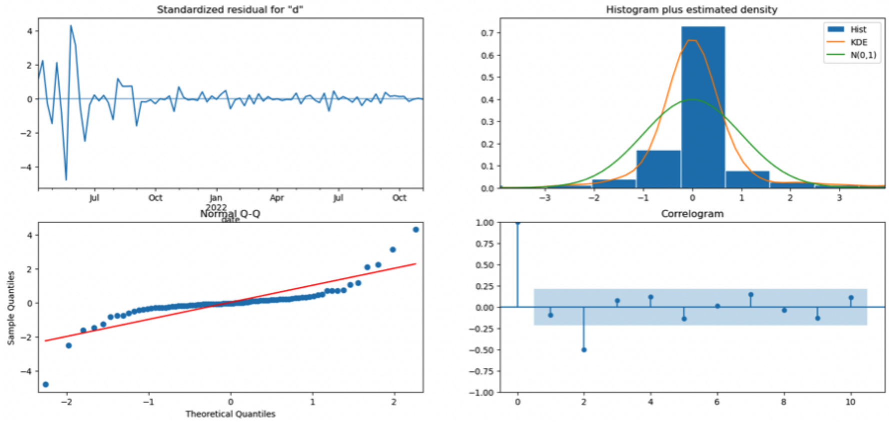


Figure 14: Model Results for SARIMA Accounting for Seasonality and Trend

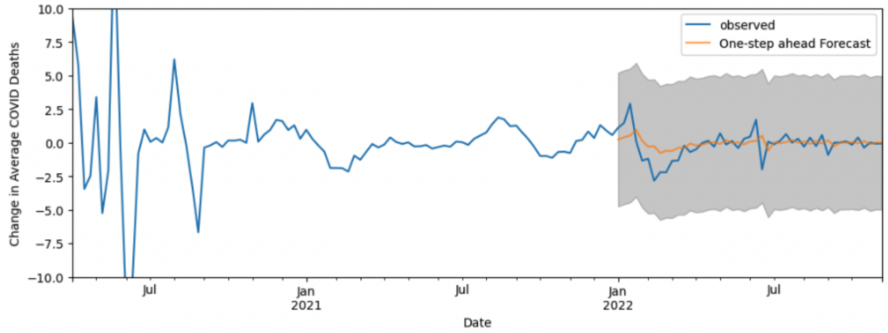


Figure 15: Model Results for ARIMA Accounting for Trend

# Appendix G: Shapley for Interpretation

To strive for interpretability, we leveraged the Shapley value while running a XGBoost. Specifically, the Shapley value is an idea adopted from game theory, and it represents the "average expected marginal contribution of one player after all possible combinations have been considered" [5]. Additionally, the Shapley value can be used to measure a variable's contribution to a certain model. In particular, researchers in [6] show that Shapley can be used for both helping interpretation of a model and for feature selection. Thus, we ran XGBoost model to predict which variables contribute to COVID-19 Deaths and use Shapely value to interpret the contribution of each feature.

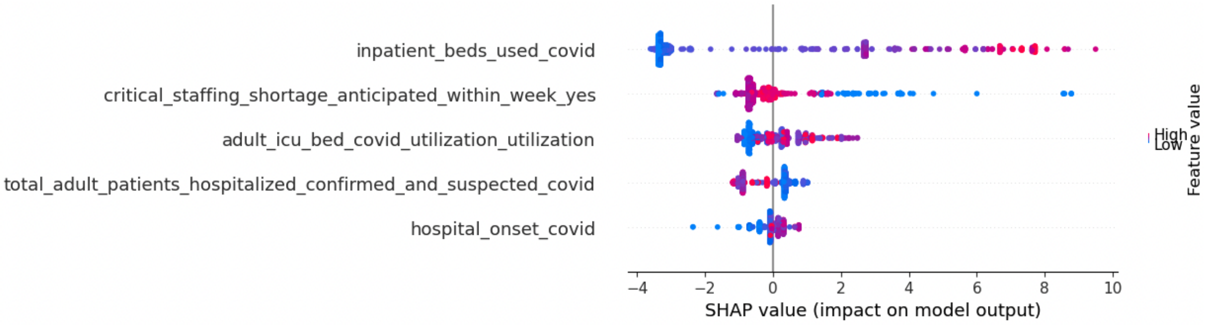


Figure 16: Shapley Value XGboost Variable Impact

In Figure 17 above, we see the impact of each variable. This follows our analysis using Optimal Regressive Trees as the first two variables, "inpatient_beds _used_covid" and "critical_staffing_shortage_anticipated_within_week_yes", are used in splits that result in a large amount of deaths shown in 6 and 5. To understand the overall contribution of each variable, we show the mean absolute value of the Shapley variable contributions shown below in 17.

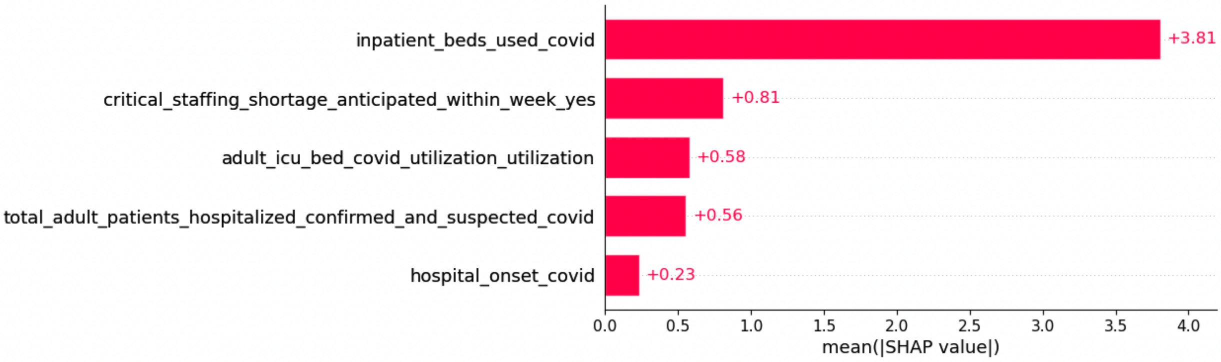


Figure 17: Mean Absolute Value of Shapley XGBoost Variable Contributions

We do not report a testing metric for our XGBoost model, but instead we rely on this model to confirm our results about the interpretation of variables affecting COVID deaths. In a further study, we could rely on Shapley rather than VIF for feature selection following the practice shown in [6].

# 6 Appendix H: Confounding Variables and Impacts on COVID Deaths

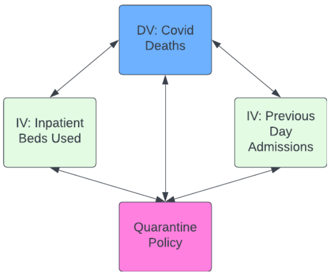


Figure 18: COVID Policy Confounding on Independent and Dependent Variable

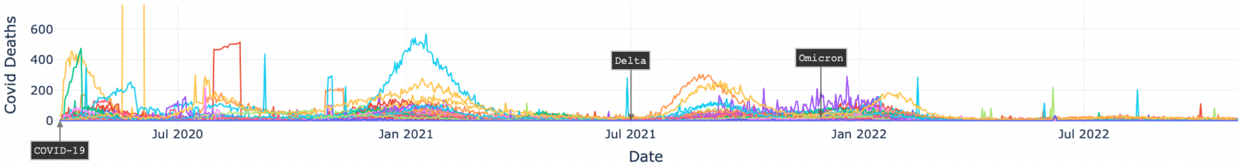


Figure 19: Covid Deaths by State

# References

[1] Goldstein G Hartung W Royer C Sundberg E van der Ziel C Van Elzakker M Roberts R King J, Goldenberg D. Congressional budget responses to the pandemic: Fund health care, not warfare. *Am J Public Health*, 111:200–201, 2021.

[2] 10.7 - detecting multicollinearity using variance inflation factors — STAT 462.

[3] Confounding Variables.

[4] Douglas M Sloane Rachel French Brendan Martin Kyrani Reneau Maryann Alexander Matthew D McHugh Karen B Lasater, Linda H Aiken. Chronic hospital nurse understaffing meets covid-19: an observational study. *BMJ Quality Safety*, 30:639–647, 2021.

[5] Will Kenton. Shapley value. *Investopedia*, Sep 2021.

[6] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. The shapley value in machine learning. *arXiv*, May 2022.