Project Name: CT Covid Death Cases Prediction

Prepared by Team E3 (A-Z): Henry Zhang, Jason Smith,
Monica Torrijos Ronquillo, Rachel Kallely, Xiasun Huang

MAY 2, 2022

Purpose of This Report

On a website called Towards Data Science, Mashinchi started a Covid-19-death-cases-in-the-US prediction competition with CNN, using an ARIMA model in Python. He employed various predicting variables, like confirmed cases, demographic features of the population, and some economic statistics in his model. He achieved a prediction score of an underestimation of just 500 cases using CNN's prediction as a benchmark (Mashinchi, 2020). Even though Mashinchi considered diversified variables in the time series forecasting, he did not include more recent COVID-19 events such as vaccination information and covid variants in his model. In addition, many studies are conducted to investigate the relationship between geographic features and Covid-19 seasonality. For example, Christophi et al.'s study suggested that an increase by 1 degree Celsius in temperature is associated with a 6% lower COVID-19 mortality rate (Christophi et al., 2021). This research shines a light on the role of temperature in relation to COVID-19 death cases prediction.

This project is inspired by various research relating to the Pandemic and aims to predict Covid-19 death cases in Connecticut by incorporating multiple factors, including newly emerging events like the spread of Covid variants. To walk through our procedures, first, we performed missing value imputations and smoothened predicting variables in our model using 7-day moving averages[1]. Second, we employed machine learning algorithms to select top features that were ranked in terms of their coefficients to the target variables. Third, we took advantage of SAS's time series forecasting functionalities to train and select the best model from our model "basket" after addressing the non-stationarity problems underlying the target variable trend. At the end of this project, our team provided constructive suggestions to the users of this model regarding the future counter-Covid strategies.

This report serves to explain the procedures we took throughout the project and analyze and justify key variable movements, variable correlations, and the prediction effectiveness of our model. While concluding that two of the variables, the infections of people eighty years older and those of people

[1] A note from the editor: please note that all the predicting variables and the target variable are smoothened using 7-day moving averages. For some variables for EDA purposes only, 31-day moving averages are used. Pleases refer to Appendix III for more details of our procedures and scripts.

less than ten years old have the strongest predicting powers, we point out limitations of our model for future death cases prediction in the state.

An Introduction to Our Data

When collecting precise and trustworthy data, we give credits to the official governmental website, CT Data Government, which discloses Covid-19 related data for public decision-making. Our primary dataset (dataset1) contains variables like confirmed cases, deaths, and confirmed cases by age. To examine the vaccine's efficacy in saving people's lives, we also collected daily vaccination records (dataset2), including the records specific to different age groups, from the same source. In addition to the daily data, the cumulative vaccination rates (dataset3) of the state were also pooled from ctpublic.org, added to the vaccination category. For the temperatures in Connecticut (dataset4), we obtained raw data from the National Centers for Environmental Information, which was taken good care of before being processed in Python. Most of the temperature records were provided by the station, Hartford Bradley International Airport, CT, US. If the station did not observe the temperature for some period, we reached out to another station, Hartford Brainard Field, CT US. For the missing values of the temperature data, we used 3-day moving averages for the imputation after pooling data from both the observation station. Please note that our temperature data were the data points recorded when they were observed. Our Covid variant data (dataset5) was contributed by outbreak.info, on which the data was already preprocessed using 7-day rolling averages.

The three datasets, the primary dataset, the temperature data, and the cumulative vaccination records, were merged into a grouped one in Python. We took three approaches to address the missing value confusion on the grouped dataset level. For the missing values in the primary dataset (containing death cases), we use 3-day moving averages for the imputation. For the missing values in the cumulative vaccination rates, we filled in the blanks with the valid value on the date closest to the time when the missing value occurred. Null values in the daily number of instances of vaccinations (dataset2) were dealt with in a separate file by the filling-in-zeroes function in Python.

Our data for the modeling contained records of 744 rows and 34 predicting variables. The variables we used for the forecasting can be categorized into confirmed cases, confirmed cases by age group, cumulative vaccination rates, cumulative vaccination rates by age group, temperature, and Covid variant proportions. As part of the data preprocessing, we also did variable smoothening for the convenience of data visualization. To better understand our data, we performed exploratory data analysis (EDA) to obtain a general understanding of the Pandemic.

1. Seasonal “Viral Strikes”

Flatting the canvas of time, we drew the confirmed cases in Connecticut by age group. Figure 2.1 shows that the virus “revives” in Springs after being “deactivated” for some time during the Summers. After each “hibernation” of the virus, another wave of infections follows in just several months on a

much larger scale. In the third wave of the Pandemic, the cases of infection can "rocket" to a ceiling of around 14K confirmed cases. Also, it appears that the "period of silence" decreased drastically, leading to the concerning situation that the fourth wave of the attack may already be signaled in early April of 2022. Zooming in for more details, we can see that the confirmed cases of young children less than ten years old are much higher in the third wave than those in the second wave (Figure 2.2 & Figure 2.3).

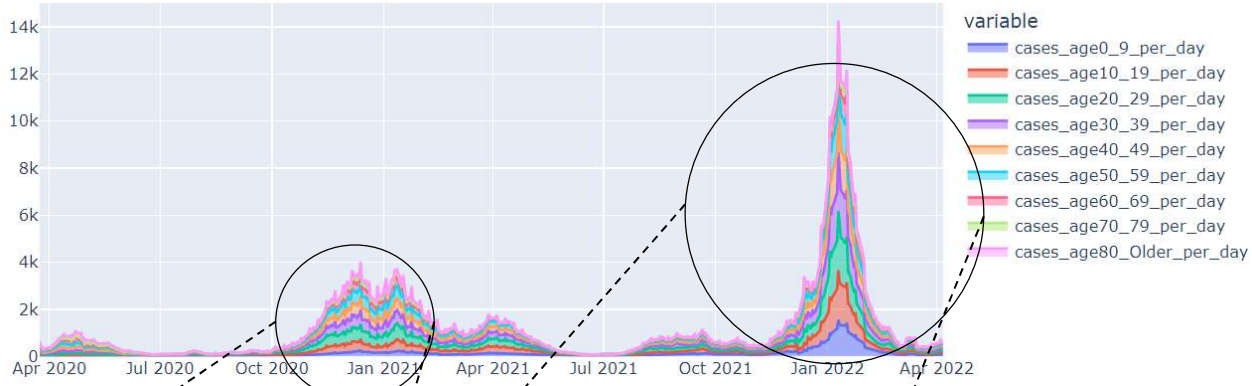


Figure 2.1 Confirmed cases by age group

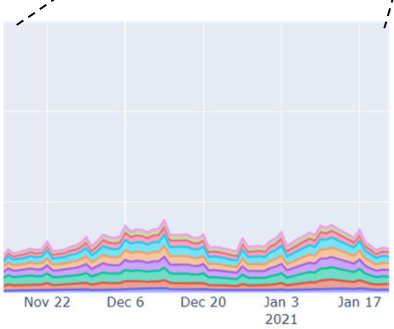


Figure 2.2 Zoomed in-1

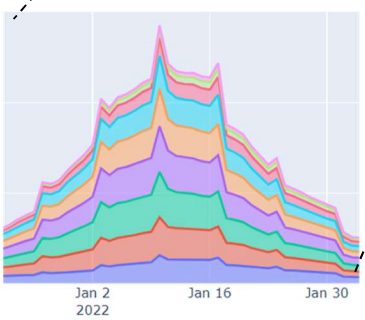


Figure 2.3 Zoomed in-2

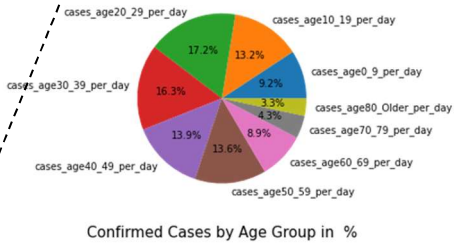


Figure 2.4

Figure 2.4 adds information about the infections by age group at a point in time. As of April 6, 2022, young adults (20-29 years old) have accounted for the highest proportion of the infected population, whilst older people (80 years older) take the smallest share.

2. The Death Cases Follow

High infection rates have serious consequences. Directly relating to the infections, death cases follow a very similar pattern of ups and downs (Figure 2.5). The good news is that the overall trend of death cases is decreasing as waves of strikes landed the state. Unlike the concerning infection rate in early April, death cases fall steadily to the floor, giving hope that vaccinations are working to save more lives. But is it true?

Figure 2.5 Death cases

3. Getting Vaccinated

Connecticut has been achieving an exciting vaccination rate in its population (Figure 2.6); even though the government did not disclose information about young children regarding this matter, the public witnessed a high first and second dose vaccination rate among people over 16 years old. As of April 6, the proportions of old people receiving boosters are much higher than those of the younger.

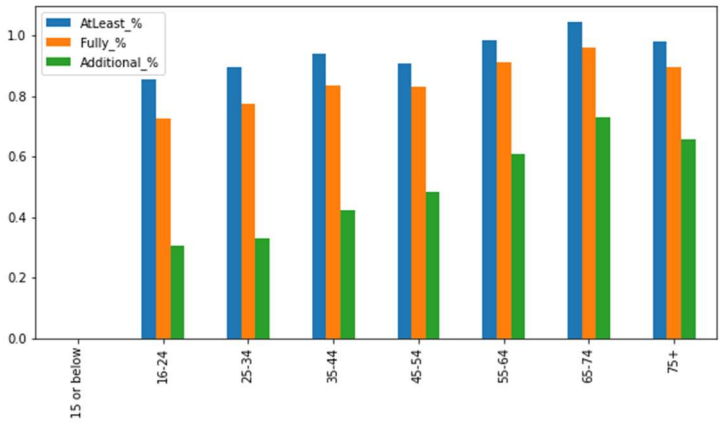


Figure 2.6 Vaccination rates by age group

Figure 2.7 shares the information about the first-dose vaccination trend among the population. People are passionate about receiving their fist-dose in FEB 2022. The first Pfizer Covid-19 vaccine administered for Governor Ned Lamont on FEB 16, 2022, had been motivating more and more people to get vaccinated (CT Govt, 2021). From the end of APRIL 2022 onwards, the pace of vaccination decreased but suddenly escalated on AUG 12, 2022 (Figure 2.7). Data.ct.govt explained that the first-dose vaccinations could be overstated when the state approved the booster administration in AUG 2021, as it happened that vaccine administration records for individuals could not be linked because of differences in how names or date of birth are reported (CT Govt, 2022). This led to the abrupt but temporary event we see in Figure 2.8.

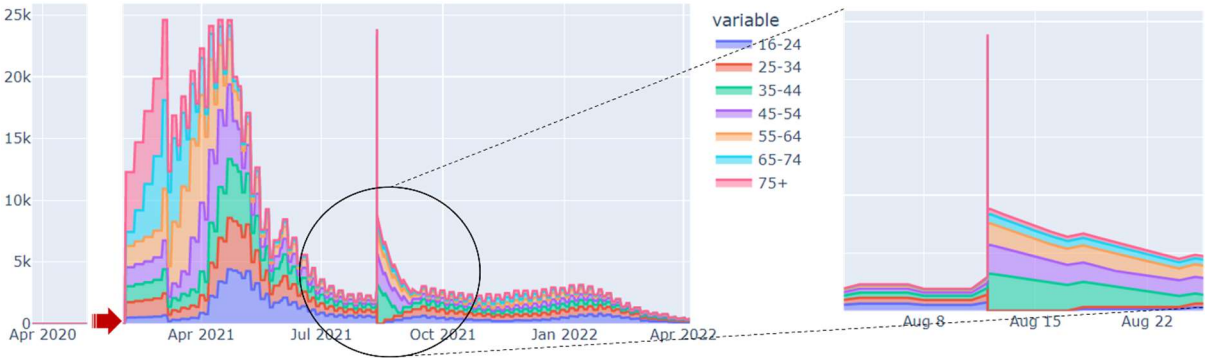


Figure 2.7 First-dose vaccinations per day **Figure 2.8 First-dose vax zoomed-in**

Figure 2.9 summarizes the vaccination performance of the state in terms of cumulative vaccination rates. As of April 6, 2022, the state had achieved a promising 95% vaccination rate of people receiving at least one jab. Stratifying the vaccination records and using daily vaccination instances (Figure 2.10), we can see a very similar trend to Figure 2.7, except that an abrupt and permanent event (the approval of boosters) occurs in JAN 2022.

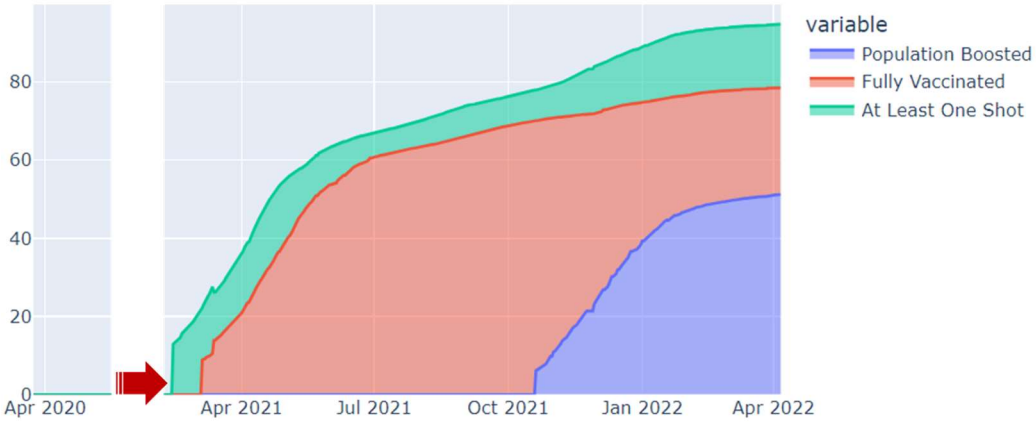


Figure 2.9 Cumulative vaccination rates

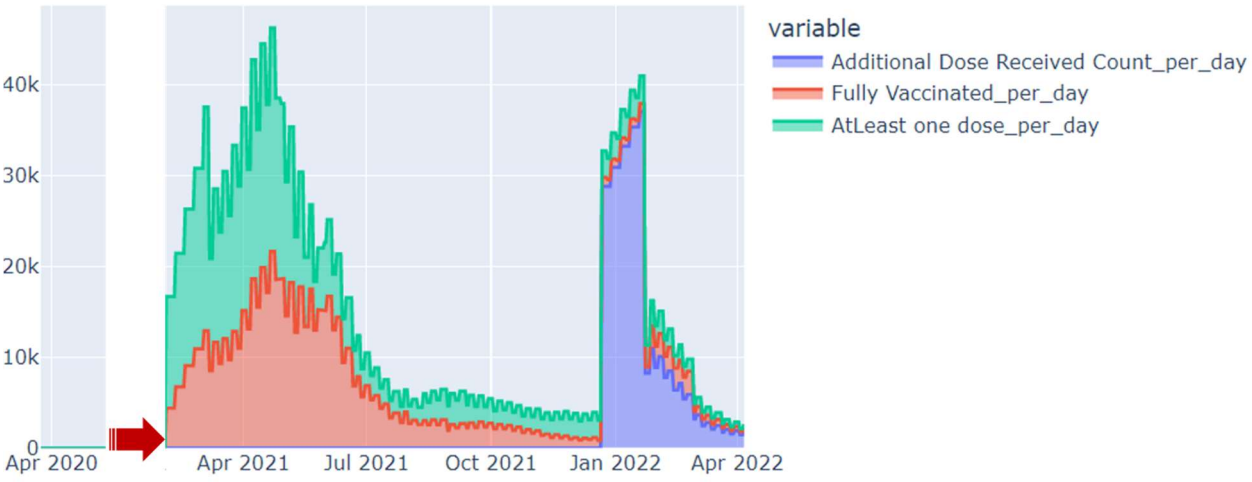


Figure 2.10 Vaccination administered per day

4. Covid-19 Variant “Iteration”

The approvals of Covid-19 vaccinations enhance the morale among healthcare workers, but the virus seems to always find an adaptive way to survive. In the Summer of 2021, Delta landed and dominated the viral landscape (Figure 2.11), as other variants like Alpha, Gemma, and Beta became weaker, and numbers of vaccinations decreased (Figure 2.10). When the boosters staged in JAN 2022, Omicron thrived in the Winter and boomed the confirmed cases. Even though the number of failed “testing” variants has decreased, the world becomes worried about the next-coming variant that may become part of human society.

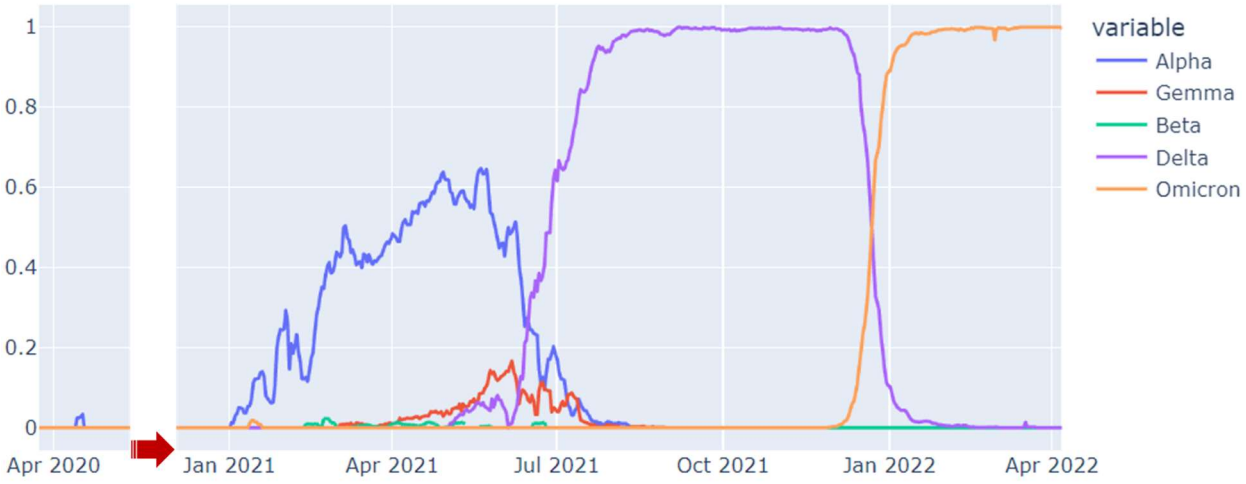


Figure 2.11 Covid variant proportions to confirmed cases

Our Modeling Process

After the data preprocessing stage, our data is ready for modeling. This section will consider the importance of the features, addressing the non-stationarity problem of the target variable time series, and appending other time series modeling variables, like autoregressive, curve trend, and seasonal dummies to our models.

1. Feature Selection

To fill up our top feature pool, we used the filter method to pick up the key variables that had solid explanatory power in predicting our target. By employing the Python embedded package, xgboost, we evaluated the coefficients of the variables within the machine learning algorithms and ranked these coefficients to generate a list of top features. Figure 3.1 presents the ranking of the explanatory power of those single variables. From left to right, the ranking of the single variables slides exponentially to a level of around 0.04. We included those 19 variables on the plateau in our pool of top features. We summarized the correlation among these features in Figure 3.2. We can see that many variables are dependent on one another, meaning we should remain alert when interpreting these variables in our model. We will also compare the performance of these variables with a benchmark model to see if they can make a difference.

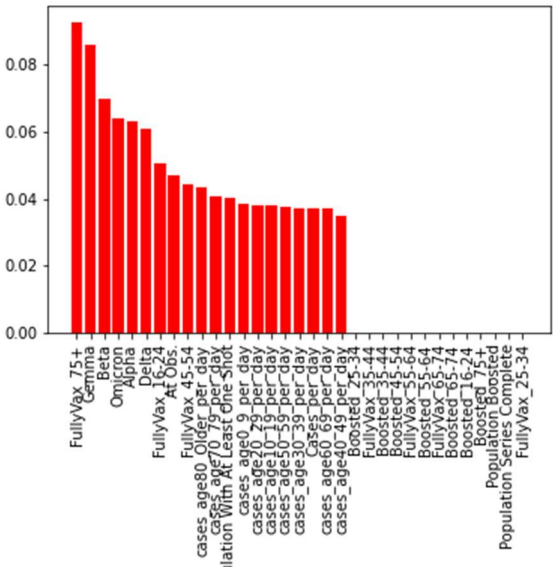


Figure 3.1 Coefficients ranking

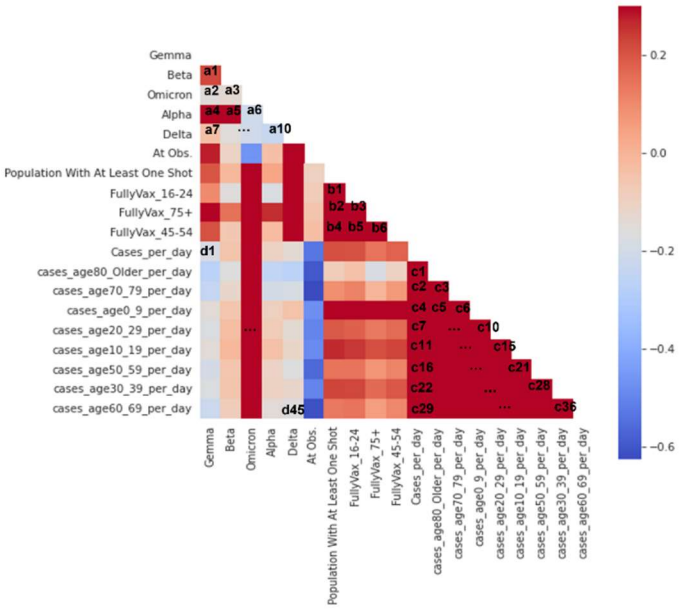


Figure 3.2 Correlation matrix

2. Stationarity Diagnostics

Before training models for the prediction, we should also consider whether the time series of the death cases were stationary. Previously, Figure 2.5 shows that the time series of death cases has regular up-and-downs during the Pandemic. It seemed that it took around five months for the death cases to rise again since the last time it arrived at the peak. Seasonality, in this case, will obviously jeopardize the assumption of non-stationarity of the time series forecasting model and should be the first thing on the to-do list on which the team sets goals to make sure the assumption is held. SAS becomes handy when it comes to stationarity diagnostics by providing various visualizations of relevant testing.

The autocorrelation plot (Figure 3.3) indicates a strong autocorrelation of the time series, which decays slowly over the periods. We did not observe regular fluctuations in the autocorrelation trend.

In the partial autocorrelation plot (Figure 3.4), we can see many spikes pierce the thresholds, but the breakthroughs are not so seasonally regular. Seasonal root tests (Figure 3.5), as usual, reject the claim that the trend has seasonality.

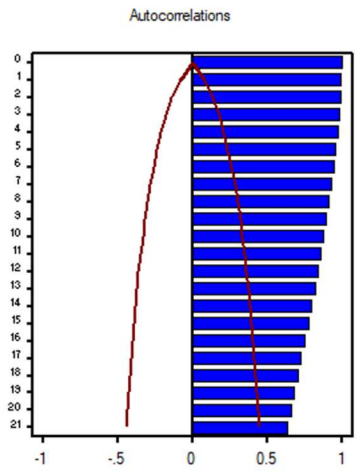


Figure 3.3

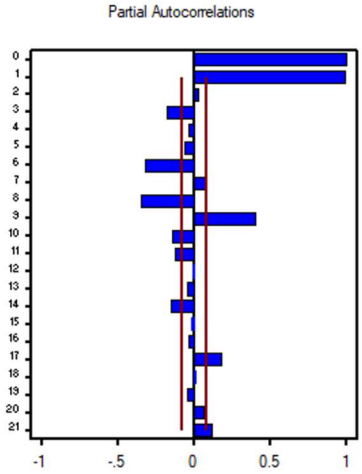


Figure 3.4

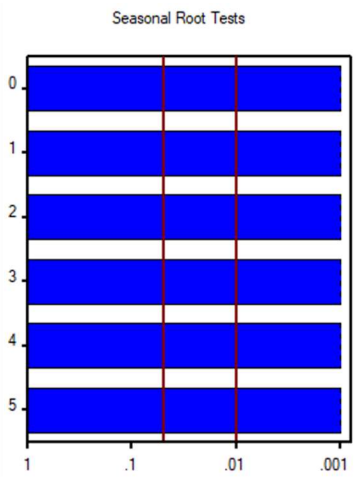


Figure 3.5

For the trend diagnostic, we used unit root tests to test the hypothesis that there existed an underlying trend within the data. The results in Figure 3.6 fail to reject the null hypothesis that the time series itself has a trend. Also, when applying the first difference to the time series, the p values of the unit root tests become so significant that the null hypothesis can't be reasonable. The shift of the p values of the unit root tests from the left boundary to the right suggests there's a trend within the death records.

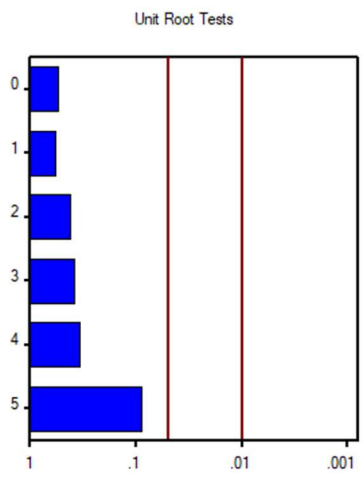


Figure 3.6

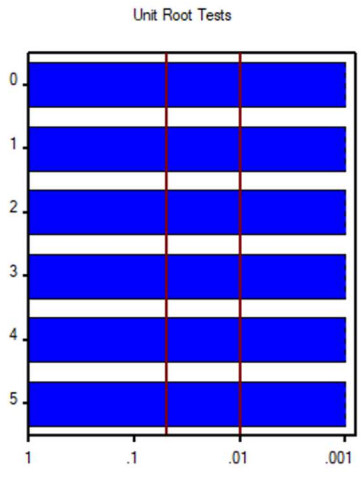


Figure 3.7

The autocorrelation plot and the unit root test suggested that the time series was not likely to be stationary. We had not enough evidence for the claim that seasonality existed using only the autocorrelation plot. Still, we were concerned about the waves of the death cases in the Pandemic and

believed that seasonality should be taken care of in the modeling process. All diagnostics (strong autocorrelation and insignificant unit root test results) suggested a trend in the data that would violate the assumptions of a valid forecasting model.

3. The Benchmarking Model

To eliminate non-stationary factors, SAS provides sophisticated models to address the problems automatically. We used a forecast horizon of 120 days for the model training purposes and a hold-on sample of 80 days. Before we added our key features to our model, we generated an automatically applied forecasting model in SAS, the performance of which could be used as a benchmark for our manually applied forecasting models. The best model SAS could offer in this case is a model of Log-Linear Trend with Autoregressive Errors, in which the trend and seasonality underlying the dataset were addressed by autoregressive, seasonal regressive, and log-linear trend. We wrote down the model performance in terms of its RMSE (2.41044) and set the goal that our best model should not perform worse than this automatically applied model.

4. Model Training

To consider the effects of the top features on the prediction, we added all 19 variables as regressors using the Fit Custom Model settings. Our team also added the parameters of the trend curve and seasonal dummies to our model so that the effects of trend and seasonality could be skimmed off. In light of the benchmark model, we used Logarithmic Trend to deal with the underlying trend. The lagging effects of the single variables were also on our list of considerations. To determine the best lagging period for each shifted variable, we added variables one by one from the top feature pool to the Logarithmic Trend and seasonal dummies until the pool was exhausted. The lagging period for each shifted variable was the optimal period that minimized the predicting error of the model after the variable was assigned. Please note that all the dynamic regressors are not eligible for transformations, so we will not consider transformative regressors in our model. Finally, we migrated the regressors and dynamic regressors to our benchmark model for better model comparison. Using the model forecasting visualization tool in SAS, we summarized the predicting performance of the models in our "basket" at this point as follows.

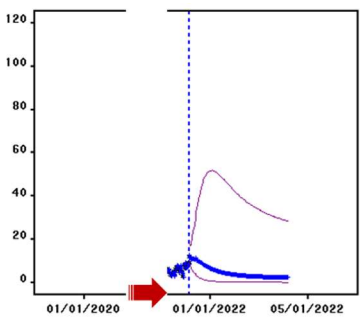


Figure 3.8 Benchmark

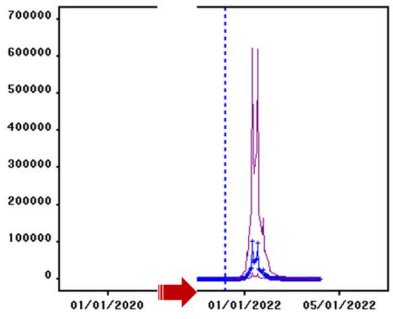


Figure 3.9 Benchmark – All Var

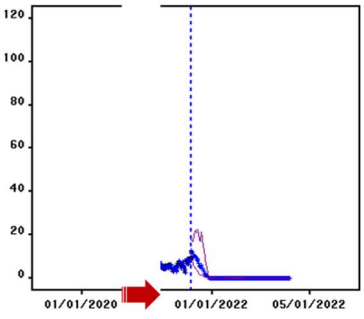


Figure 3.10 Benchmark – All LaggVar

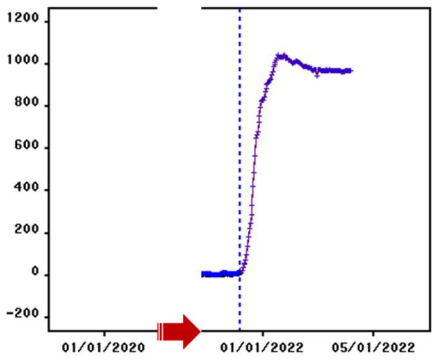


Figure 3.11 Model – All Var

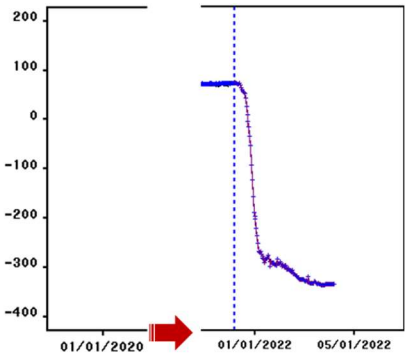


Figure 3.12 Model – All LaggVar

Given all variables as predicting factors, both the benchmark and manually applied models predicted a horrible level of covid death cases in the 120-day validation period (Figure 3.9 & Figure 3.11). The predictions told a totally different story for the models with shifted variables: death cases would drastically fall (Figure 3.10 & Figure 3.12). The parameter estimate of the fall was so significant that the model with shifted variables (Figure 3.12) crashed and was trashed to the bottom of the negatives (around -350).

Compared with the benchmark model in Figure 3.8, the predictions of the model set were extremely volatile and raised the alarm that inappropriate variables were included in the model. We walked through the parameters of estimates of the model set and investigated what could go wrong. We figured out that one of the variables, Omicron, had an extremely high estimate and standard deviation that would force the predictions to go up and down in a drastic manner.

MODEL NAME	OMICRON-RELATED VAR	ESTIMATE	P-VALUE	STANDARD DEVIATION
BENCHMARK	NA	NA	NA	NA
BENCHMARK – ALL VAR	Omicron	3.94851	0.6722	9.2813
BENCHMARK – ALL LAGGVAR	OMICRON[Lag(6)] Lag6	-7.86409	0.4157	9.5894
MODEL – ALL VAR	Omicron	988.46879	<.0001	119.7642
MODEL- ALL LAGGVAR	OMICRON[Lag(6)] Lag6	-620.18041	<.0001	81.3323

Table 3.1

Table 3.1 shows that the benchmarking models (Benchmark – All Var & Benchmark – All LaggVar) do better in suppressing the anomaly of Omicron-related variables, but they fail to take care of other

variables. In Figure 3.9, the Omicron-related variable may violate variable relationships within the model, which fails by predicting unrealistic death cases. Adding a lagging period seemed to address the problem of using an Omicron-related variable in Figure 3.10. Still, the predicting performance showed that all other variables might not be significant, as the predicted values just followed a very similar trend to the benchmark model containing no features. The extremely unstable Omicron-related variable estimates in Table 3.1 explained why the models in Figure 3.11 and Figure 3.12 were out of control. In the case of manually applied models, even though the Omicron-related variables had significant p values, the model's high standard deviation made the parameter hostile to the forecasting.

5. The “Culprit” – the Omicron variant

However, why should Omicron, the new covid variant responsible for the surges in confirmed cases in recent days, be blamed for the failure of our current model set? Even though the variant is not new in our daily life, it's the "alien" for our training dataset. When the start date of the prediction, DEC 08, 2021, was set, the model had no reason to believe that Omicron would dominate other covid variants soon. In section 2.1, Omicron started changing the rule of the game in early December of 2021 and became the main variant in the period of validation (Figure 2.11). Such a drastic change in the viral landscape made our model confused and provided a risky estimate for the parameter. Omicron, in this case, became the "virus" that infected our models and made them "retarded”.

To eliminate the noises from Omicron, our team excluded the variable from each of our model sets. In this way, new models (M4, M5, M8, M9) were added to our model "basket" (with model labeled from M1 to M9), which was summarized in the following table.

MODEL CATEGORY	MODEL NAME	LABEL
AUTOMATIC	Benchmark	M1
	Benchmark – All Var	M2
	Benchmark – All LaggVar	M3
	Benchmark – All Var excpt Omic	M4
	Benchmark – All LaggVar excpt Omic	M5
MANUAL	Model – All Var	M6
	Model – All LaggVar	M7
	Model – All Var expt Omic	M8
	Model – All LaggVar expt Omic	M9

Table 3.2

6. The First Model Comparison

So far, we have a "basket" of models that consider variables that will impact our prediction. We exported the predicted values of each model and calculated the RMSE of each model for the period of validation. The best model stood out with the lowest error in the validation dataset. The performances of controlling error of the validated models are listed as follows.

LABEL	CATEGORY	TRAINING RMSE	VALIDATION RMSE	RANK*
M1	Automatic	2.41	22.65	2
M2		2.28	18,872.88	9
M3		2.35	25.76	3
M4		1.93	403.61	6
M5		2.36	34.69	5
M6	Manual	5.43	790.85	8
M7		4.76	462.87	7
M8		5.31	30.17	4
M9**		5.33	18.60	1

*The ranking is based on the validation RMSEs of the model set. Higher ranking means lower RMSE.

**The best model, in current model set, is M9 (Model – All LaggVar expt Omic).

Table 3.3

Table 3.2 shows that M9 turns out to be the best model. We reviewed the autocorrelation plot of M9's prediction errors to see if all non-stationarities were removed. There remain many unregular significant spikes in the partial autocorrelation plot (Figure 3.13), meaning the prediction errors are not stationary yet.

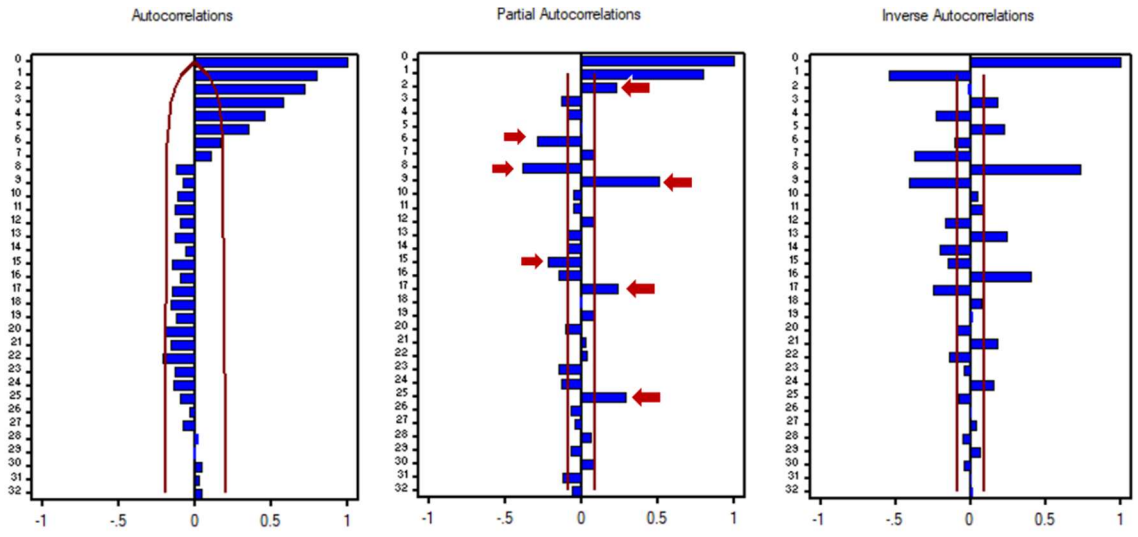


Figure 3.13

To address this problem, we enabled the ARIMA options in model M9 and tried different numbers in the autoregressive p to determine the best p that minimizes RMSE. When p=4, the model M9 performed the best, and the performance was improved significantly. We reviewed the autocorrelation plot once again, and the unregular spikes were removed in the partial autocorrelation plot (Figure 3.14) except for one outlier at lag 8. The autocorrelation of the prediction errors, in this case, looked so small that the value was insignificant at lag 1.

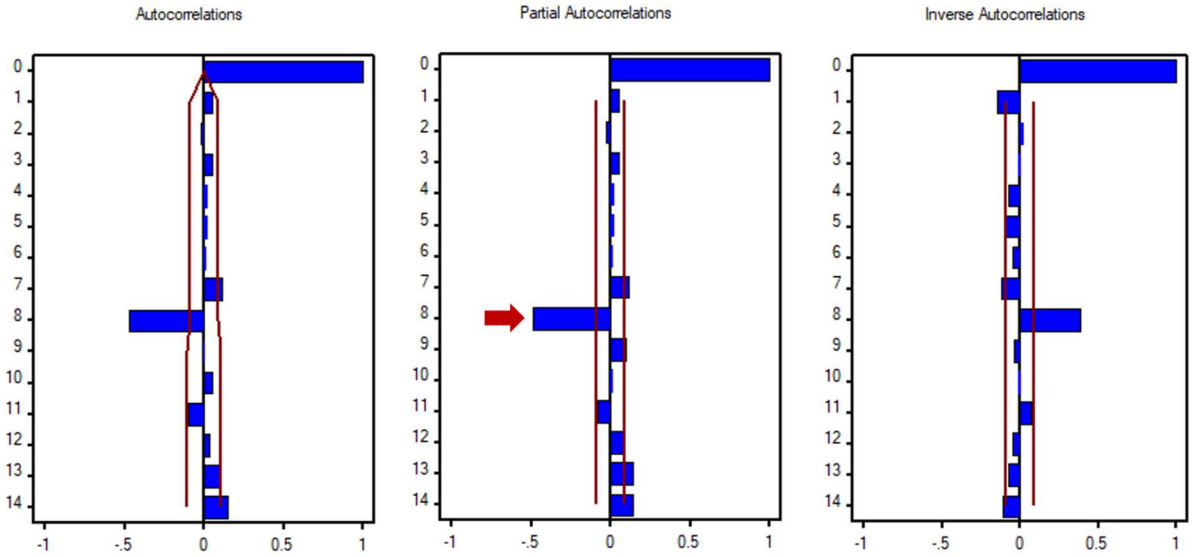


Figure 3.14

7. The Second Model Comparison

We want to call our improved best model M9 & ARIMA(4,0,0) and compare it with M9 in terms of their RMSEs. Table 3.4 shows that training and validation RMSEs are improved when ARIMA(4,0,0) is added. Most importantly, our goal of lowering training RMSE to a level lower than that of the benchmark (2.41) has been achieved.

LABEL	TRAINING RMSE	VALIDATION RMSE
M9	5.33	18.60
M9 & ARIMA(4,0,0)	2.24	16.50

Table 3.4

Looking at the residuals of the M9 & ARIMA(4,0,0) (Appendix II.2) and referring to the prediction period in the prediction plot (Appendix II.1), we can see that the prediction errors at the wave after AUG 2021 are unexpectedly high despite the much lower death cases than the wave started in early 2021. This may result from the occurrences of more predictors at the later pandemic wave (for example, vaccination of young children and the emergence of new covid variant Omicron.)

Our Model Interpretation

After two rounds of model comparison, we filtered out eight models in our model basket and improved the best model, M9, by adding autoregressive parameters. M9 & ARIMA(4,0,0) is our final best model. This section will interpret the model performance and coefficients of our best model.

1. M9 & ARIMA(4,0,0) Model Parameter Estimates

Appendix I lists our key feature parameter estimates and their p values. In section 3, figure 3.2 visualizes the correlations among our key features, and it seems all of them are not independent of each other. The dependency among our top feature pool will mean that model interpretation using coefficients is rather tricky. For this section, we will employ various exploratory data analyses to uncover relationships between key variables so that the effects of those features in our model can be better understood.

People 80 years or older who confirmed covid-19 was one of the two variables with statistical significance when it came to death cases prediction. The model parameter estimates (Appendix I) shows a considerable positive estimate (around 0.2) and a significant p-value for this group of people. For the features related to confirmed cases, data collected from people 80 years or older had the highest parameter estimate. Even though it's difficult to tell the exact coefficient of the variable to the predicted variable, we can use the variable CASES_PER_DAY[Lag(11)] Lag11 as a benchmark since cases per day are related to the widespread infection in the state.

In figure 3.2, variables related to confirmed cases cluster in the lower right corner with labels from c1 to c36. The color of the deep red of these variables means they will affect the death cases in the same direction, with the higher level of the variable leading to the higher death cases. Given the insignificant estimate of the CASES_PER_DAY[Lag(11)] Lag11 (-0.00628), we can tell death cases are much more sensitive to the confirmed cases of the people 80 years or older. This is not the latest news that older people are exposed to much higher risks of losing their lives in the Pandemic. In its guidelines for older people in countering the coronavirus, CDC revealed that people in their 80s, especially those 85 years or older, were the most likely to get very sick in the Covid situation (CDC, 2021). The life-threatening condition of the people of such age group had resulted in 80% of Covid-19 deaths in the US, despite their much lower proportion of the population in the country (KFF, 2020).

Another variable with statistical significance is CASES_AGE0_9_PER_DAY[Lag(9)] Lag9, which has a much lower estimate (0.03236) than CASES_PER_DAY[Lag(11)] Lag11 but a higher estimate than the benchmark. Turning the page back to Figure 2.1 in section 2, we can see that confirmed cases among children of 9 years old or younger were surging at a concerning pace when Omicron struck the state. The Guardian reported that the lifting of mask mandates and the alarmingly low vaccination rate among children had caused 20% of pediatric death when Omicron dominated (The Guardian, 2022). Our model alarms, too.

Young adults (of 20-29 years old), however, have an estimate (-0.00569) of being just in line with the overall population. Even though our EDA in the previous section (Figure 2.4) revealed that people of such age group account for the highest proportion of the confirmed cases, they had much lower risks given the much lower parameter estimate.

Temperature (At. Obs) can be an exciting parameter in predicting Covid death cases. A paper in the National Library of Medicine studies that Covid-19 is one of the infectious diseases that display seasonal patterns in their instance (Mohammad M. Sajadi, MD et al., 2020). The study reveals that areas with significant community transmission of Covid-19 distributed roughly along the 30-50° N" corridor at consistently similar weather patterns consisting of average temperatures of 41-52°F, combined with low specific and absolute humidity. It seems the Covid trend is highly predictable in Connecticut. The latitude of the state is from 40°58′ N to 42°03′ N. Also, Our EDA (Figure 4.1) suggests that most high frequencies of Covid confirmed cases cluster at a temperature range from 34°Fto 60 °F. Figure 4.2 shows that high frequencies of the death cases occur in a range of much lower temperatures. We were not sure about the lagging period before the temperature would have significant effects on Covid death cases, but our model suggested that AT_OBS_[Lag(77)] Lag77 (with a parameter estimate of 0.08587) was the best for the use of temperatures in predicting the target.

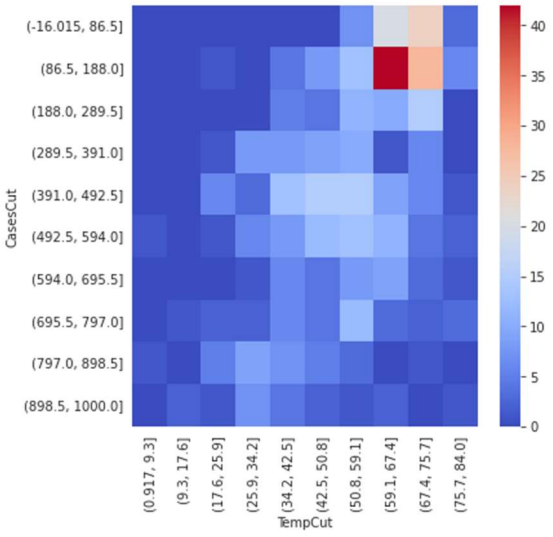


Figure 4.1 Frequencies of temperature vs cases (<1000)

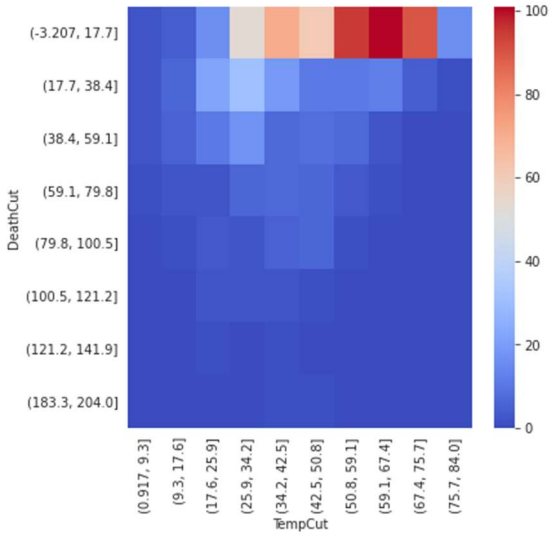


Figure 4.2 Frequencies of temperature vs deaths

2. The Threat of Covid Variants

The question is, why the variant related variables, GEMMA[Lag(34)] Lag34, BETA[Lag(18)] Lag18, ALPHA[Lag(34)] Lag34, and DELTA[Lag(10)] Lag10, are not "tested significantly positive" in the death cases prediction process?

Looking at the "survival plot" of the Covid variant (Figure 2.11), we can see that the realm of Delta demises extremely fast in the Winter while Omicron starts the role on the stage in early DEC 2021

and totally wins the competition at the beginning of 2022. The perfect "X-shape" relationship of Delta and Omicron indicates there exists a formula of the "wildcard" used by the coronavirus in the Winter:

$Delta + Omicron = 1$, given all other variants replaced by the two dominating. (Formula1)

It's not difficult to tell that Omicron is negatively related to Delta by looking at the correlation a6 in Figure 3.2. More interestingly, Omicron is the only variant in "distracting deep blue" in the line of At Obs. in Figure 3.2, meaning, being different to its “predecessor”, the virus is strongly eager to survive the Winter and evolved to counter low temperatures. Correlations from b1 to b45 indicate that Omicron has already been responsible for confirmed cases across all age groups, given relatively strong positive figures.

The underlying negative correlation between Omicron and Delta (Formula1) will mean that excluding Omicron-related variables in the modeling process will not mean excluding the Omicron effect. The Omicron effect, in this way, exists in our model as (1-Delta) and will have a positive coefficient on the death cases (given DELTA[Lag(10)] Lag10 < 0 (Appendix I)). Despite its insignificant coefficient p variable, the variant will still affect Covid-related deaths since it has a considerable positive correlation with confirmed cases. The rule of "natural selection" makes the comprehension of our variables sustainable, despite the literal disappearance of Omicron. It makes our model follow a very similar path to that of the actual (Figure 4.3 & Figure 4.4).

Figure 4.3 Prediction vs the actual **Figure 4.4 Prediction vs the actual (Zoomed-in)**

Unfortunately, the "inheritance" of the Delta variant will make the model interpret only the effect of the Delta variant itself, not that of the newly emerging one since the model has no intelligence to understand Omicron. Apart from being more resilient in low temperatures, Omicron differs from Delta with its "milder survival strategy." CDC's Morbidity and Mortality Weekly Report revealed that the percentage of adults who died during Omicron (7.1%) was much lower than during the Delta period (12.3%), along with a lower percentage of hospitalized people receiving IMV from DEC 2020 to JAN 2022 (CDC, 2022). This may explain why the predicted trend did not slip as fast as the actual after JAN 23, 2022 (Figure 4.4).

Implications and Recommendations

As the cliché goes, “all models are wrong, but some of them are useful” our model bears the limitations of using historical data for future event prediction. Even though a time series model alleviates the pain of using historical data by considering only data of the validation period for the prediction in that period, the curse of passive data modeling cannot be lifted as the future cannot be known. While Omicron is competing with its only main competitor, Delta, our model becomes the “lucky dog” that still works even when the competition just begins. Therefore, this model holds two very limiting assumptions. First, there exist only two variants in the future, and one of them is the alternative to the other (variant1+variant2=1). Second, we can use our understanding of the predecessor to predict the behaviors of the successor. Such assumptions will cause problems when an event that will change things fundamentally in the future occurs and may result in the failure of the prediction. However, the acceptable overestimation of the death cases of the model after JAN 23, 2022, indicates there were no fundamental changes when Omicron struck Connecticut.

As the model indicated that people over 80 years old are exposed to the highest risk of Covid death among people of all age groups, we should provide more assistance and community support for those people of such groups to help them overcome such difficult times. CDC shares information and guidelines for protecting older people around us. For people who live in the residential communities for older adults, getting vaccinated is one of the keys to protecting friends and family members who live in these communities (CDC, 2021). CDC is concerned about the increased transmissibility of Omicron variant in the nursing homes & long-term care facilities and recommends various measures to control the outbreak of the infection if necessary. For example, empiric use of Transmission-Based Precautions (quarantine) is recommended for residents who are newly admitted to the facility and for residents who have had close contact with someone with Covid-19 infection if they are not up to date with all recommended COVID-19 vaccine doses (CDC, 2022).

As the new variant, Omicron, emerges, and more and more young adults are getting vaccinated, young children less than ten years old should be recommended to get vaccinated. Our model suggests that the rule that young people are less likely to suffer from Covid-19 may be no longer valid, as pediatric deaths become significant indicators for the prediction. CDC reported that children ages 5 through 11 years are most frequently affected by MIS-C, where different body parts become inflamed, including the heart, lungs, kidneys, brain, skin, eyes, etc. Getting children vaccinated helps prevent such conditions associated with Covid-19 among young children (CDC, 2022). Vaccination will also help control the spread of the disease to others by children and teens (CDC, 2022).

Our model also suggests that high temperatures can be a good predictor for the next wave of Covid-associated deaths, with a lagging period of 77 days. It’s good news that Covid-19 has seasonality that can be foreseen and used for the disease countering strategies. Disease controlling institutes should

tackle the infections in Winters and Springs and inform prevention measures like masking mandates to the public. Despite the seeable seasonal patterns of the disease, our EDA reveals that the pattern may become more difficult for the use of prediction as Covid revives at a much faster pace.

Our team also wants to warn the interpretation of our model using variable coefficients, which may lead to misleading decision-making and misunderstanding of the event. Users should be cautious of the hidden variables that will still have significant effects on the target, despite observing insignificant p-values. Covid variants, in our model, affect our prediction with very strong correlations with confirmed cases-related variables. Our model proves that Omicron results in lower death rates than its predecessor. However, the highly infectious feature of the variant should not be ignored, as the variant is still causing more infection instances and pulling up the covid-associated death cases.

Conclusion

Our team has processed a dataset with around 744 records and 34 predicting variables using Python and selected our models' top 19 features. While narrowing our target for proper models and considering the emergence of Omicron, we evaluated our model "basket" in terms of RMSE and "filtered" out models with poor performance in controlling errors. Given the difficulties in dealing with the non-stationarity problems, we paid attention to the stationarity of the predicting models throughout the project. The M9 & ARIMA(4,0,0) model satisfied all our requirements and became our best model for the Covid death cases prediction. At the end of the project, our team concluded that the confirmed cases of older people (80 years older) and young children (0-9 years old) were the most effective "test strips" for the prediction. Other variables like temperatures and conformed cases of young adults (20-29 years old) were strong indicators for the prediction, but those hidden in the variable list and have a strong correlation to the significant variables should not be ignored. For instance, we observed that Omicron worked in the way of lowering the Covid-associated death rate, despite its "recessive trait" in our model. At the bottom of the report, we recommended several measures or actions to protect people whose life was threatened by the disease. Our team also warned against the limiting assumptions of our best model for the best interest of the users who predicted such a future event.

References

CDC (2021, August 4). COVID-19 Risks and Vaccine Information for Older Adults. COVID-19 Recommendations for Older Adults. https://www.cdc.gov/aging/covid19/covid19-older-adults.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fneed-extra-precautions%2Folder-adults.html

CDC (2022, Apr. 6). Why Children and Teens Should Get Vaccinated Against COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/why-vaccinate-children-teens.html

CDC (2022, Feb. 2). Interim Infection Prevention and Control Recommendations to Prevent SARS-CoV-2 Spread in Nursing Homes. Nursing Homes & Long-Term Care Facilities. https://www.cdc.gov/coronavirus/2019-ncov/hcp/long-term-care.html

CDC (2022, January 28). Trends in Disease Severity and Health Care Utilization During the Early Omicron Variant Period Compared with Previous SARS-CoV-2 High Transmission Periods — United States, December 2020–January 2022. Morbidity and Mortality Weekly Report (MMWR). https://www.cdc.gov/mmwr/volumes/71/wr/mm7104e4.htm#:~:text=The%20percentage%20of%20hospitalized%20COVID-19%20patients%20who%20received%20IMV,groups%20(p%3C0.001).

Christophi, C.A., Sotos-Prieto, M., Lan, FY. et al. Ambient temperature and subsequent COVID-19 mortality in the OECD countries and individual United States. Sci Rep 11, 8710 (2021). https://doi.org/10.1038/s41598-021-87803-w

CT Data (2022, April 28). COVID-19 Vaccinations by Town and Age Group. Health And Human Services. https://data.ct.gov/Health-and-Human-Services/COVID-19-Vaccinations-by-Town-and-Age-Group/gngw-ukpw

KFF (2020, Jul 24). What Share of People Who Have Died of COVID-19 Are 65 and Older – and How Does It Vary By State?. https://www.kff.org/coronavirus-covid-19/issue-brief/what-share-of-people-who-have-died-of-covid-19-are-65-and-older-and-how-does-it-vary-by-state/

Mashinchi, N. (2020, December 29). Predicting number of covid19 deaths using time series analysis (Arima model). Medium. https://towardsdatascience.com/predicting-number-of-covid19-deaths-using-time-series-analysis-arima-model-4ad92c48b3ae

Mohammad M. Sajadi et al. (2020, March 9). Temperature, humidity, and latitude analysis to predict potential spread and seasonality for COVID-19. National Library of Medicine. Version 1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7366819/

The Guardian (2022, March 11). A fifth of all US child Covid deaths occurred during Omicron surge. https://www.theguardian.com/world/2022/mar/11/us-child-covid-deaths-omicron-surge

The Office of Governor Ned Lamont (2021, February 16). Governor Lamont Receives First Dose of COVID-19 Vaccination. Press Releases. https://portal.ct.gov/Office-of-the-Governor/News/Press-Releases/2021/02-2021/Governor-Lamont-Receives-First-Dose-of-COVID-19-Vaccination#:~:text=(HARTFORD%2C%20CT)%20%E2%80%93%20Governor,Trinity%20Health%20of%20New%20England.

Appendix

I. Parameter Estimates of the best model (M9 & ARIMA(4,0,0))

Parameter Estimates

DEATHS_PER_DAY

] + Alpha[Lag(34)] + Delta[Lag(10)] + FullyVax_16_24 + At_Obs_[Lag(77)] + cases_age80_Older_per_day[Lag(21)] + Fu

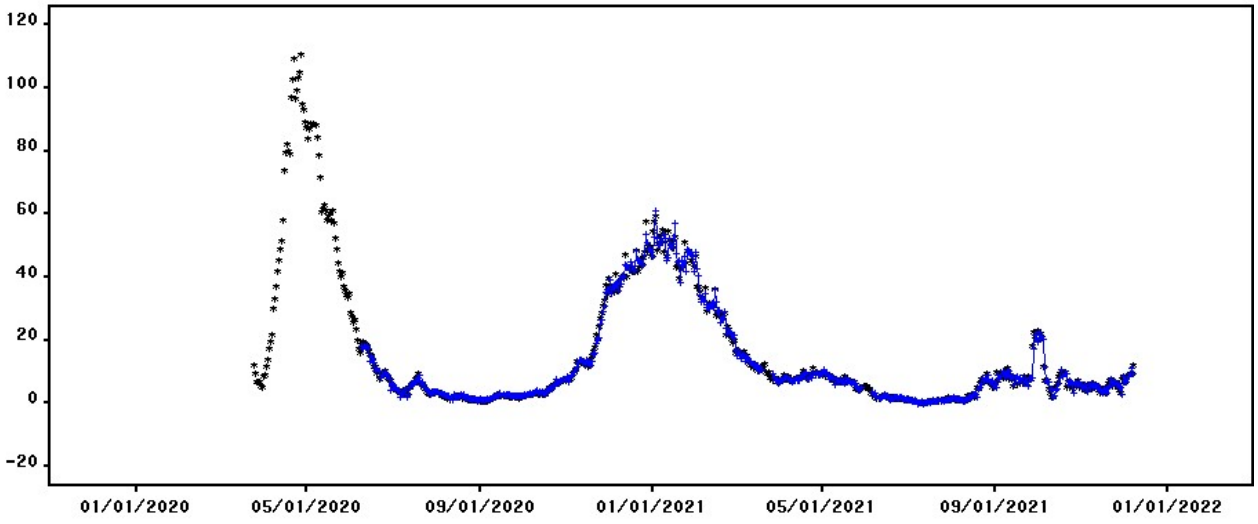
Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	-42.81756	19.9506	-2.1462	0.0362
Autoregressive, Lag 1	0.85615	0.0515	16.6360	<.0001
Autoregressive, Lag 2	0.05414	0.0710	0.7629	0.4487
Autoregressive, Lag 3	0.07512	0.0713	1.0543	0.2963
Autoregressive, Lag 4	-0.04032	0.0536	-0.7516	0.4555
FullyVax_75_	0.05325	0.0595	0.8951	0.3746
GEMMA[Lag(34)] Lag34	-3.08324	10.3920	-0.2967	0.7678
BETA[Lag(18)] Lag18	9.41440	41.1140	0.2290	0.8197
ALPHA[Lag(34)] Lag34	-0.14388	4.0727	-0.0353	0.9719
DELTA[Lag(10)] Lag10	-3.84813	4.6500	-0.8276	0.4114
FullyVax_16_24	0.03944	0.0903	0.4369	0.6639
AT_OBS_[Lag(77)] Lag77	0.08587	0.0458	1.8744	0.0661*
CASES_AGE80_OLDER_PER_DAY[Lag(21)] Lag21	0.21633	0.0126	17.1358	<.0001**
FULLYVAX_45_54[Lag(16)] Lag16	-0.02661	0.0784	-0.3397	0.7354
CASES_AGE70_79_PER_DAY[Lag(10)] Lag10	0.02830	0.0229	1.2346	0.2221
Population_With_At_Least_One_Sho	-0.20348	0.1400	-1.4536	0.1516
CASES_AGE0_9_PER_DAY[Lag(9)] Lag9	0.03236	0.0083	3.9189	0.0002**
CASES_AGE10_19_PER_DAY[Lag(10)] Lag10	-0.0006044	0.0084	-0.0722	0.9427
CASES_AGE20_29_PER_DAY[Lag(6)] Lag6	-0.00569	0.0030	-1.8934	0.0635*
CASES_AGE50_59_PER_DAY[Lag(11)] Lag11	0.0000810	0.0281	0.002881	0.9977
CASES_AGE30_39_PER_DAY[Lag(11)] Lag11	0.00670	0.0277	0.2414	0.8101
CASES_PER_DAY[Lag(11)] Lag11	-0.00628	0.0061	-1.0252	0.3097
CASES_AGE60_69_PER_DAY[Lag(11)] Lag11	0.04798	0.0296	1.6225	0.1103
Logarithmic Trend	8.32236	3.9564	2.1035	0.0399
Model Variance (sigma squared)	2.18224	.	.	.

**For variables with p value > 0.05 but slightly higher than the threshold, we would still consider these are variables significant for the death cases prediction.*

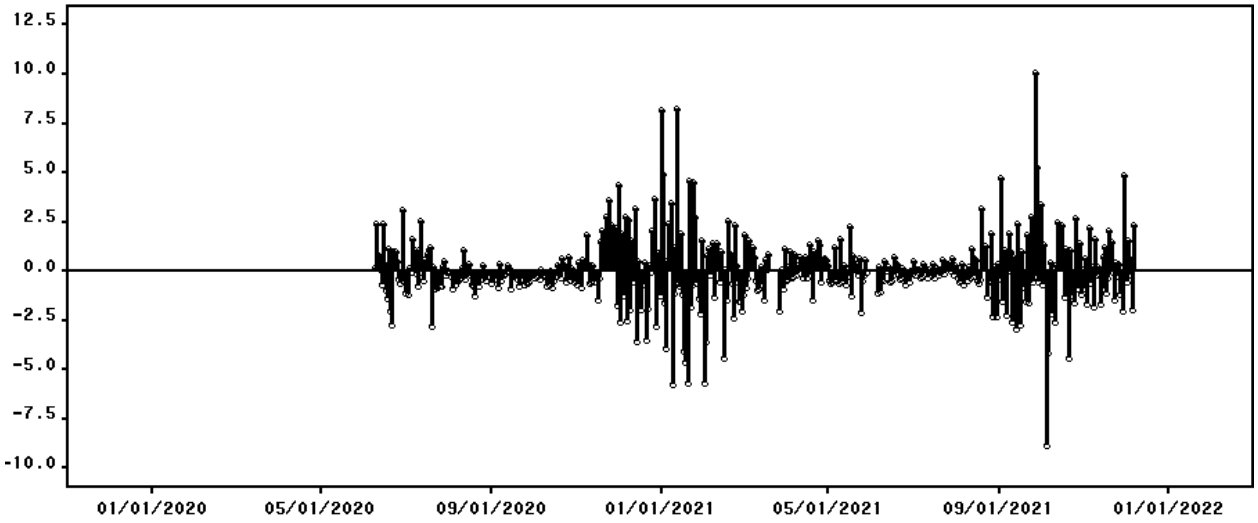
*** For variables with p value < 0.05, we consider these are variables that have statistical significance.*

II. Best model viewer in SAS

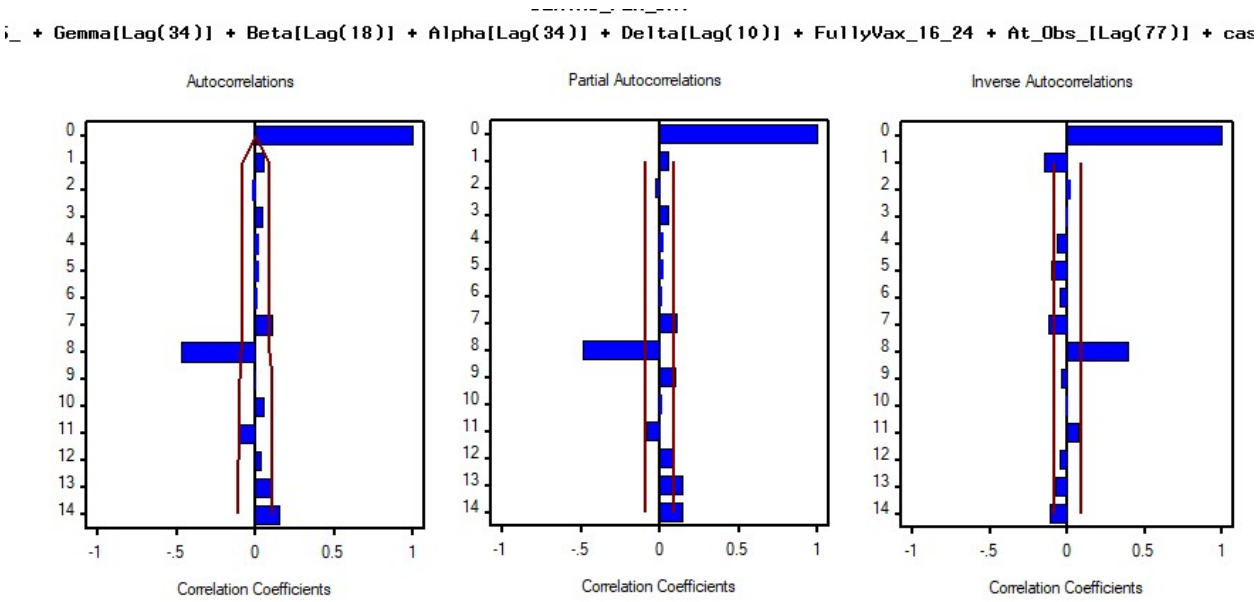
1. Model predictions



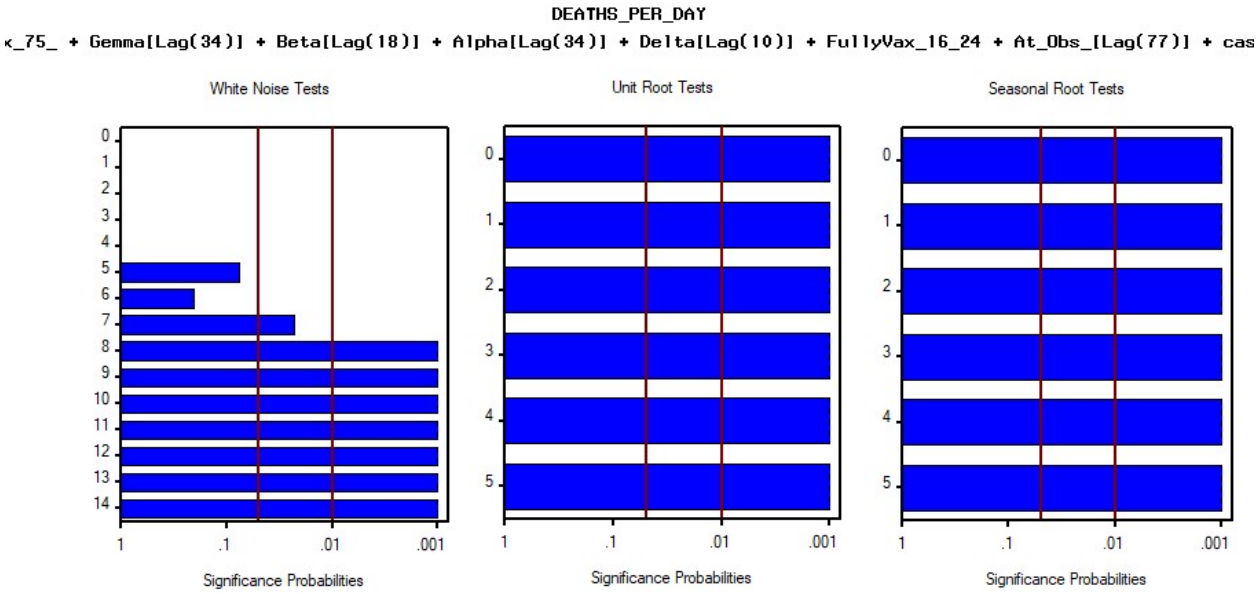
2. Prediction errors



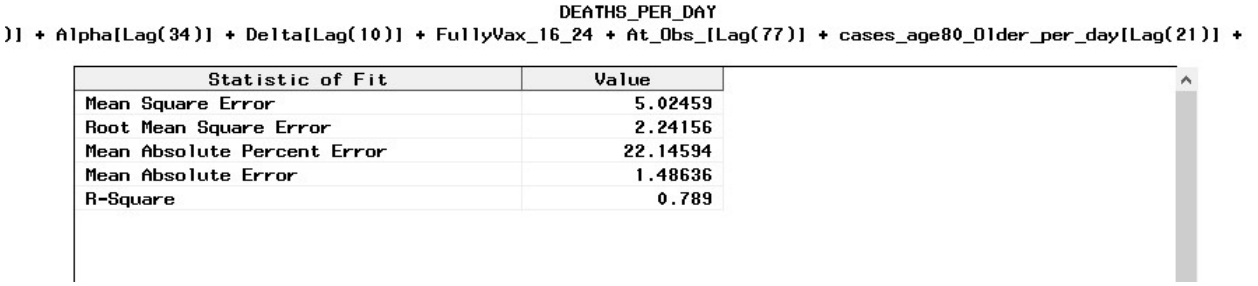
3. Prediction error autocorrelation plots



4. Prediction errors white noise/stationarity test probabilities

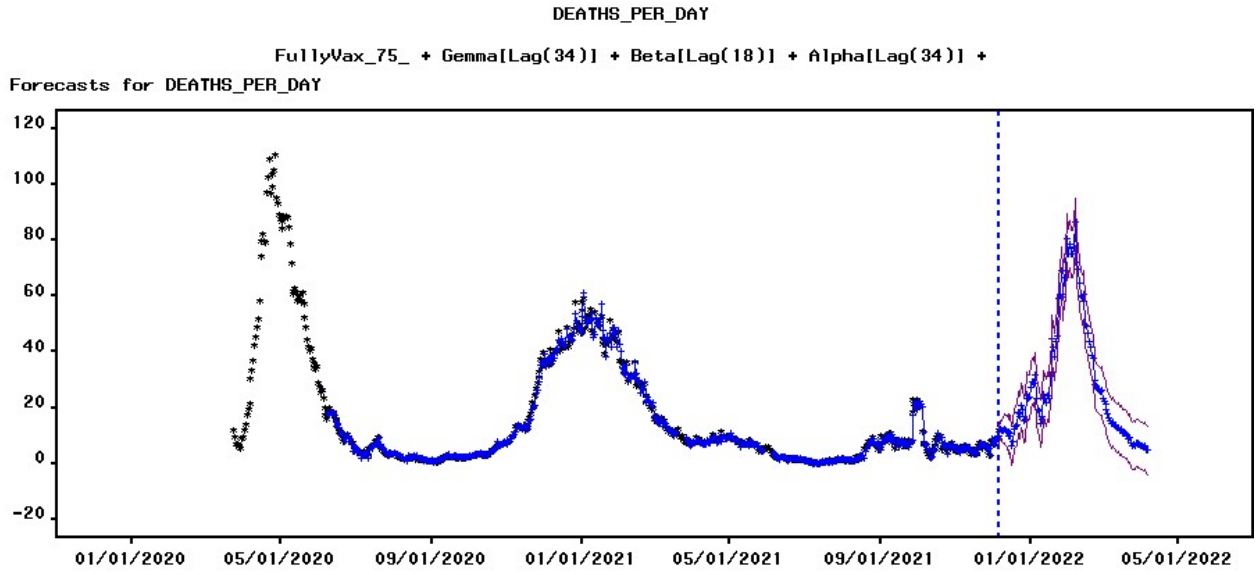


5. Best model statistics of fit



Statistic of Fit	Value
Mean Square Error	5.02459
Root Mean Square Error	2.24156
Mean Absolute Percent Error	22.14594
Mean Absolute Error	1.48636
R-Square	0.789

6. Best model forecasting view in SAS



III. Procedures and Scripts

Pleases refer to our Colab file for more details:

https://colab.research.google.com/drive/1zrQd04KAhAI1kvhbQUGbJzQ96p-f4h9X?usp=sharing

IV. Validation RMSEs in Excel

Please refer to our shared Excel file for more details:

https://docs.google.com/spreadsheets/d/1ABOejtO_SUbPvRAhENVSElxbozbXJEbq/edit?usp=sharing&ouid=114200681388729193508&rtpof=true&sd=true

V. SAS Modeling Memorandum

Please refer to our shared word document for more details:

https://docs.google.com/document/d/1q-daA1ehyhkXVlX6nIb1F0dmNrGDflfj/edit?usp=sharing&ouid=114200681388729193508&rtpof=true&sd=true