94-879 FUNDAMENTALS OF OPERATIONALIZING AI

FALL 2023

FORECASTING COVID-19 MORTALITY RATE USING VACCINATION DATA

Prepared For

Dr. Anand S. Rao
Distinguished Service Professor of Applied Data Science and AI
Carnegie Mellon University

Prepared By

Abdul Rehman (marehman@andrew.cmu.edu)
Atharva Joshi (atharvaj@andrew.cmu.edu)
Kateryna Shapovalenko (kshapova@andrew.cmu.edu)
Quentin Auster (qja@andrew.cmu.edu)
Shubham Shastri (sshastri@andrew.cmu.edu)
Heinz College, Carnegie Mellon University

December 12, 2023

94-879 Fundamentals of Operationalizing AI - Fall 2023
Final Report
Andrew IDs: marehman, atharvaj, kshapova, qja, sshastri

Abstract

The ability to accurately predict adverse COVID-related outcomes, in particular COVID-related deaths, is helpful to policy-makers and frontline workers, because it allows for more efficient allocations of scarce resources. In this paper, we predict national COVID-related deaths in the U.S. for the periods April 2022 through June 2022 and July 2022 through September 22, respectively, using machine learning forecasting techniques that rely on bagging and boosting. We introduced a novel set of features, termed "dose administration percentage," which allows for the consistent use of ever-changing vaccination information over time, across the introductions of new vaccines to fight new variants of the virus. From these analyses, we not only predicted COVID deaths, but also drew a connection between the vaccination rates of various age groups and COVID deaths, showing that vaccination rates for 65+ and 12+ populations are of particular importance nationally. Additionally, we evaluated our models for model drift and found that degradation is a definite concern, suggesting that continuous monitoring is absolutely necessary. We monitored the results of our experiments using ML FLow, and have also looked into using docker to package and deploy our work.

Keywords: COVID-19, mortality rate, Random Forest Regressor, LightGBM, time-series forecasting, machine learning

Introduction

COVID-19 started out as an outbreak in medical facilities in Wuhan (Qi et al., 2020). Due to its contagious nature, it swept across the province of Hubei (Xu et. al., 2020). The WHO declared a public health emergency in January 2020 and designated the outbreak as a Public Health Emergency of International Concern (PHEIC) (WHO, 2020). Given resource constraints for treating patients with COVID-19, policy makers, epidemiologists, frontline workers, and critical services providers valued accurate forecasts of patient outcomes in order to allocate resources to areas most in need.

Across the pandemic, researchers have used numerous time-series, machine learning, and deep learning techniques for forecasting, including autoregressive models, boosted models (e.g., XGBoost or LightGBM), bagging models (e.g., Random Forest Regressors), K-Nearest Neighbour (KNN) algorithms (Pourhomayoun et. al, 2021); bi-directional Gated Recurrent Unit (GRU) & bi-directional long short term memory (LSTM) (Ayoobi et. al, 2021); even Generative Adversarial Networks (GANs) (Alizadehsani et al, 2021).

In this paper, we use vaccination rates across age categories, collected at the county level, to forecast COVID-related deaths and examine the linkage between vaccinations and death rates. We primarily rely on time-series machine learning models, such as auto-regressive bagging and boosting models. In addition to forecasting national COVID-related deaths using vaccination data, we also investigated the potential for model drift to occur as the patterns of our data change with time. To do so, we perform modeling in stages: first we forecast the period April 2022 through June 2022 using data collected prior to April 2022 ("Phase 1"); second, we forecast the period July 2022 through September 2022 using all data collected prior to July 2022 ("Phase 2"). The purpose of forecasting in Phase 2 is to see if there were differences in the results and model approaches that yielded good results from our approach in Phase 1. We found that model drift is a significant source of performance degradation, starting from anywhere between 1 month and 5 months, meaning that continuous monitoring is absolutely necessary for these types of high-stakes predictions.

Materials and Methods

Datasets

We relied primarily on two data sources: (i) Johns Hopkins Center for Systems Science & Engineering Data ("JHU Data" or "vaccination data") and (ii) Center for Disease Control and Prevention Data ("CDC Data" or "death data"). The JHU data contained daily county-level information on vaccinations, both raw and as a percentage of population for age groups 5+, 12+, 18+, and 65+. It also contained categorical information regarding whether the county was a “Metro” area and the county’s Social Vulnerability Index (“SVI”). The CDC data contained daily cumulative death counts by county.

Data Cleaning

In the vaccination data, we performed the following cleaning steps:

- We dropped rows for which the county FIPS code was unknown (“UNK”) were dropped.
- We dropped rows for which the county FIPS code was null. Of these, only Kansas City had non-zero populations in the data, but this population should be included in Jackson County, MO (FIPS code 29095).

In the death data, we performed the following cleaning steps:

- We reshaped data from wide to long.
- We converted the FIPS column to be a 5-digit string column.

To merge the two datasets, we did the following:

- We performed a left join using FIPS code and date, with the vaccination data as the left dataset and the death data as the joining dataset.
- We dropped unmatched rows (the vast majority of these were due to non-overlapping time periods between the two datasets).

Exploratory Data Analysis

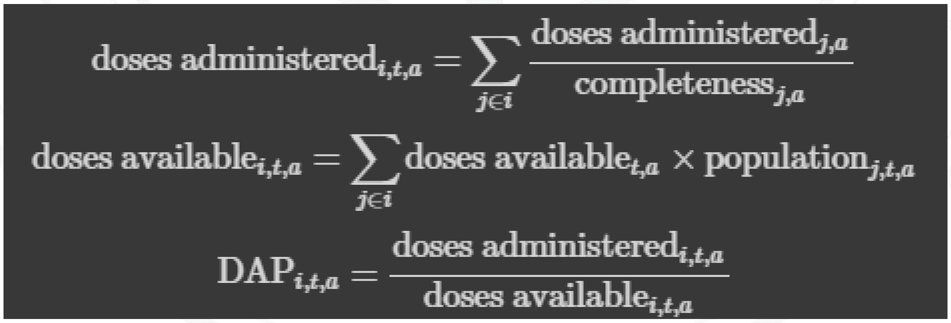
We felt that due to the temporal and spatial aspects of the data, using a medium that was able to display both dimensions simultaneously would provide an effective method for analyzing the spread of covid across states and time periods. From a spatial perspective, variations in the transmission rate among different states resulted in diverse COVID-19 spread patterns. Characteristics of state populations such as age distributions, population density, and healthcare infrastructure to name a few, affected the disparity in the spread of the disease. Temporally speaking, policy methods, disease variants, and vaccination rates were significant in terms of events that shaped the trajectory of the spread.

It was decided that the best method to capture both dimensions of the data was to use a choropleth that captured select metrics across states in conjunction with a slider that could be adjusted to show the statistics for a given date. The data was aggregated to a state level for the visualizations to maintain interpretability and keep the visual unclustered. The progression of deaths, percentage of the population that had received the first dose of a vaccine, and the percentage of the population that had received a complete series of a vaccine were the three statistics plotted using the choropleth - slider method (Figures 1 - 3).

An additional visualization (Figure 4) demonstrates the absolute number of vaccinations per three different age groups: 0 - 12, 12 - 65, and 65 plus.

Preprocessing and Feature Engineering

In order for our models to be robust across time, we recognized that it would be crucial to engineer features that would be available (and useful) across developments to COVID variants and doses. For instance, had we chosen to drop features with counts of null values above a certain threshold, this would effectively remove vaccination data for newer boosters and therefore throw away relevant information. To deal with this issue, we constructed a new feature which we term Dose Administration Percentage ("DAP"), which represents the total number of doses administered to a population over the total number of doses available at the time (assuming each person can only get one of each dose type, e.g., Shot #1, Shot #1, Booster #1, Booster #2, etc.). We calculated DAP for each of the four age groups, 5+, 12+, 18+, and 65+ using the raw vaccination information, number of doses available for each age group at a given time, and census population information provided in the data. We note that we did have to impute some values (with the county median) for some counties and days. Specifically, over counties *i*, states *j*, and time periods *t*, we calculate DAP as follows:



$$\text{doses administered}_{i,t,a} = \sum_{j\in i} \frac{\text{doses administered}_{j,a}}{\text{completeness}_{j,a}}$$

$$\text{doses available}_{i,t,a} = \sum_{j\in i} \text{doses available}_{t,a} \times \text{population}_{j,t,a}$$

$$\text{DAP}_{i,t,a} = \frac{\text{doses administered}_{i,t,a}}{\text{doses available}_{i,t,a}}$$

The additional benefit of this approach is that it allows for flexibility in aggregating at the state or county level. This was another important consideration—for the sake of quick iteration in the modeling phase, we wanted to have the ability to train models quickly, necessitating models to be built at the state level. In order to do this, we had to convert the two categorical variables (Metro Area and SVI) to continuous variables that could be aggregated at either the state or county level. To do this, we constructed variables such as “Percentage of State/County in Metro Areas” or “Percentage of State/County in SVI=A” in order to capture the number of people exposed to each unique category of the categorical variables.

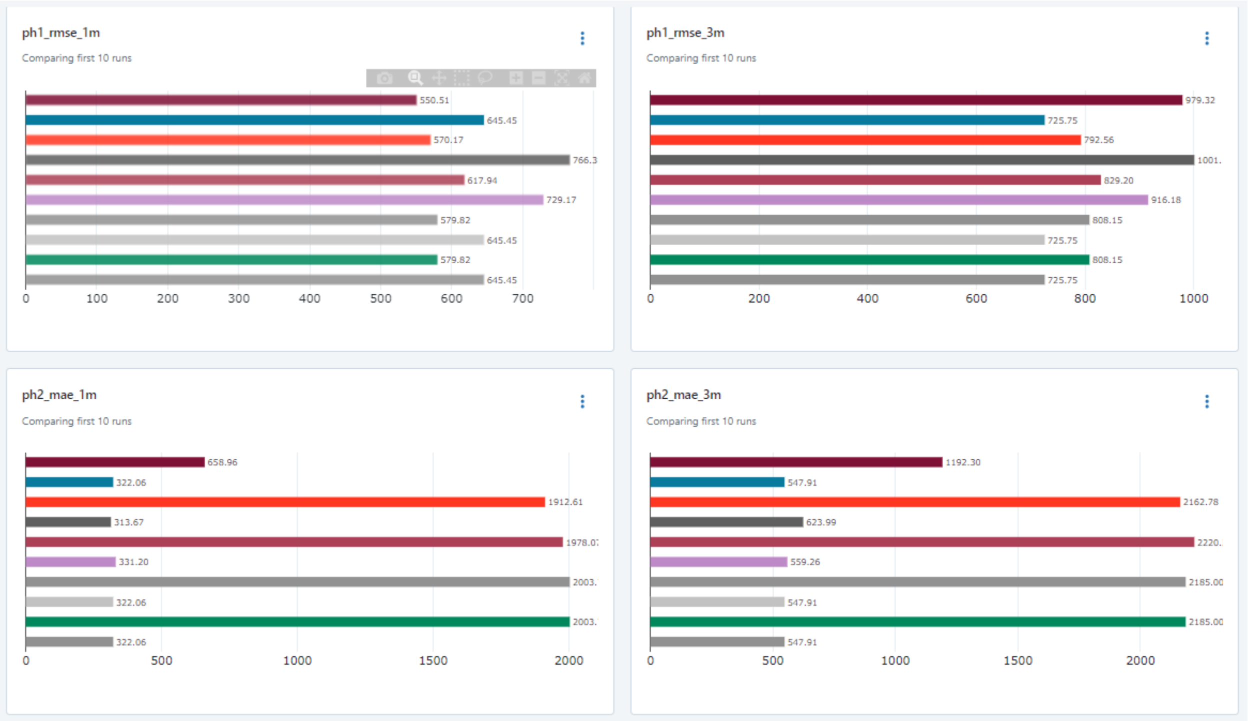
We also converted dates to both month and day indicators to capture seasonal and daily trends.

Finally, we converted cumulative death rates to daily death rates to serve as our target variable.

Modeling and Evaluation

As described above, we perform modeling in stages: first we forecast national daily COVID-related deaths for the period April 2022 through June 2022 (“Phase 1 Test Period”) using data collected prior to April 2022 (“Phase 1 Train Period”) (in union, “Phase 1”); second, we forecast national daily COVID-related deaths for the period July 2022 through September 2022 (“Phase 2 Test Period”) using all data collected prior to July 2022 (“Phase 2 Train Period”) (in union, “Phase 2”).

We used two primary classes of models: bagged architectures (Autoregressive Random Forests) and boosted architectures (Autoregressive LightGBM models). For Random Forests, we experimented with hyperparameters for model complexity, such as number of trees, as well as lag terms. For the LightGBM models, we experimented with hyperparameters such as lag time periods, as well as hyperparameters related to data preprocessing, such as the use of daily death percentage as a predictive feature itself, and the use of county level rather than state-level data. As a default, we use state-level predictions, train a model for each state, and aggregate prediction at the national level. We calculated R2, RMSE, and MAE as evaluation metrics, calculated for the first one month, and then the entire three months of both the Phase 1 and Phase 2 Test Periods.. For the LightGBM models, we logged and monitored results across experiments via ML Flow (see, e.g., below).



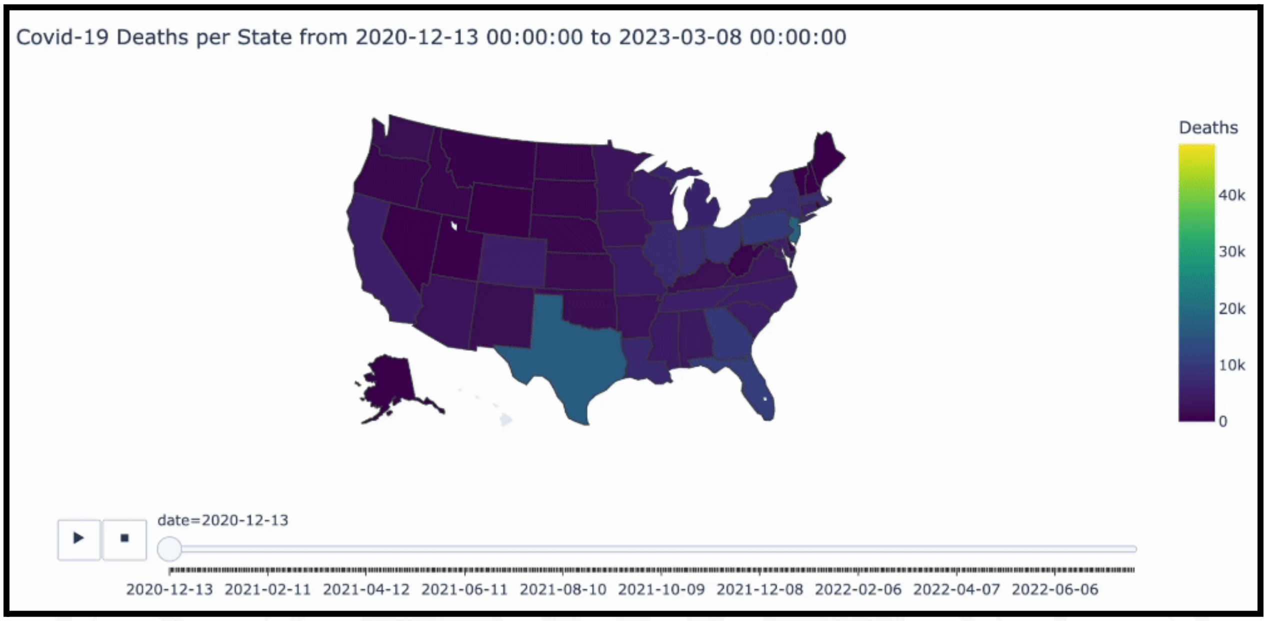
Results and Discussion

Exploratory Data Analysis

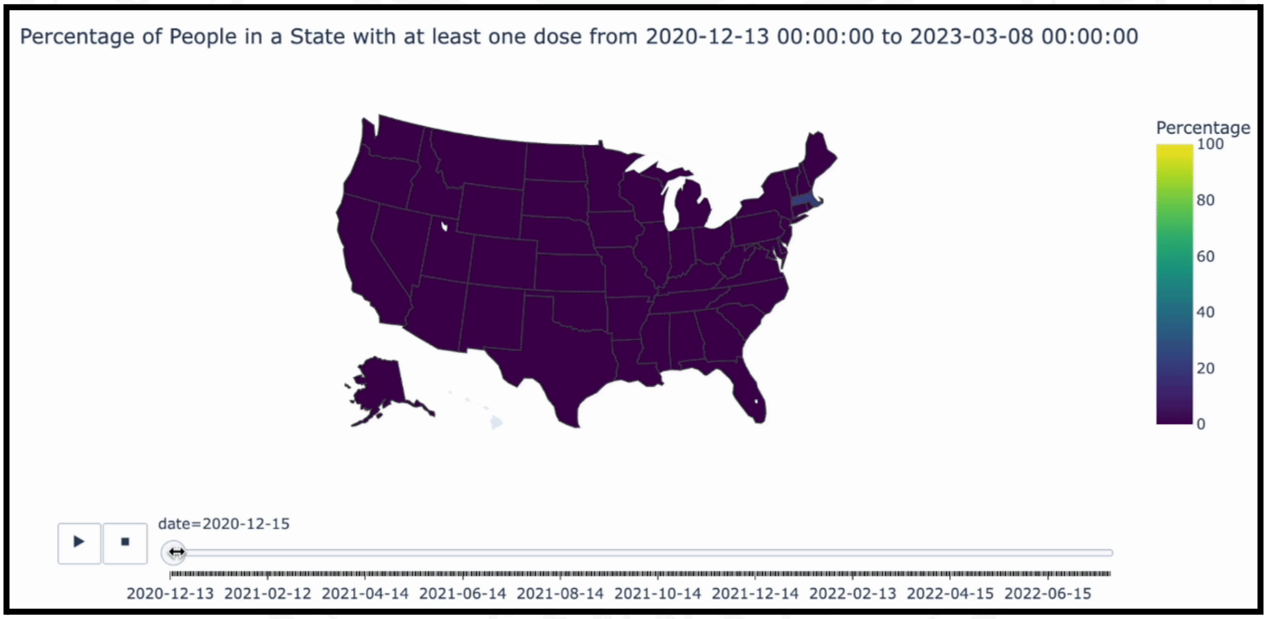
We present the results of our exploratory data analysis in Visualization Appendix, Figures 1-4. Links to animated time-series figures are provided in the appendix.

Figures 1-3, show that the state of Texas experienced the highest number of COVID-19 deaths, and saw a rate of vaccine adoption that was significantly slower than the rest of the country. As expected, higher population areas such as the coasts, experienced higher absolute deaths.

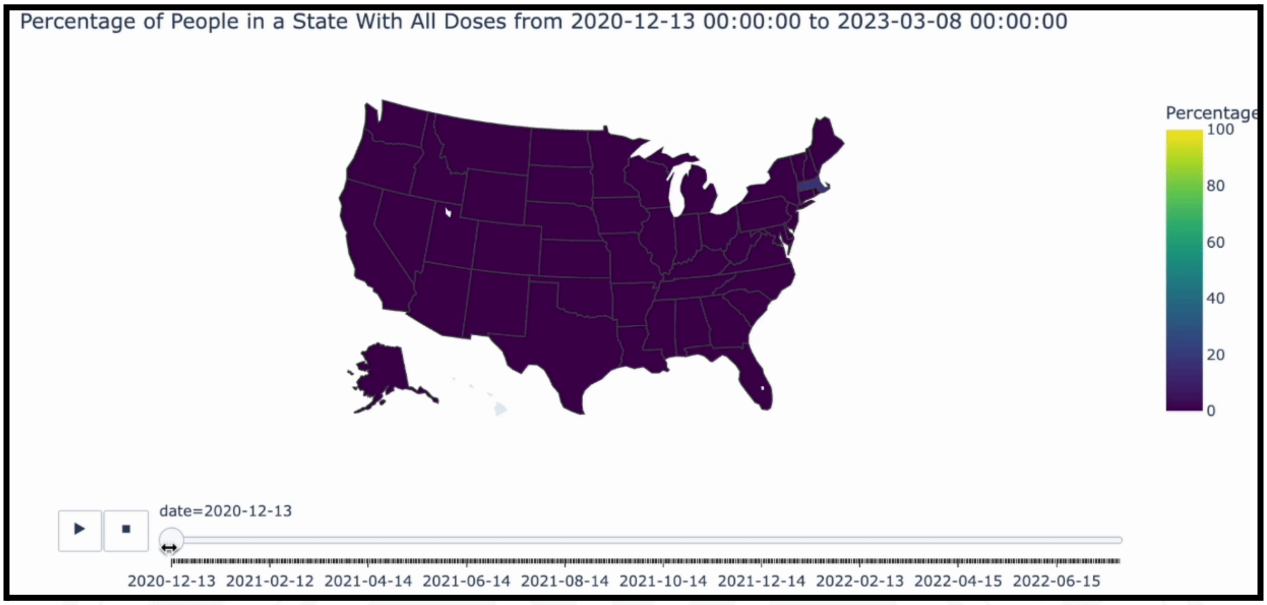
Figure 4 illustrates that the 65 plus population were the key initial targets of the vaccine, as they were deemed to be the most vulnerable to dying from the disease.



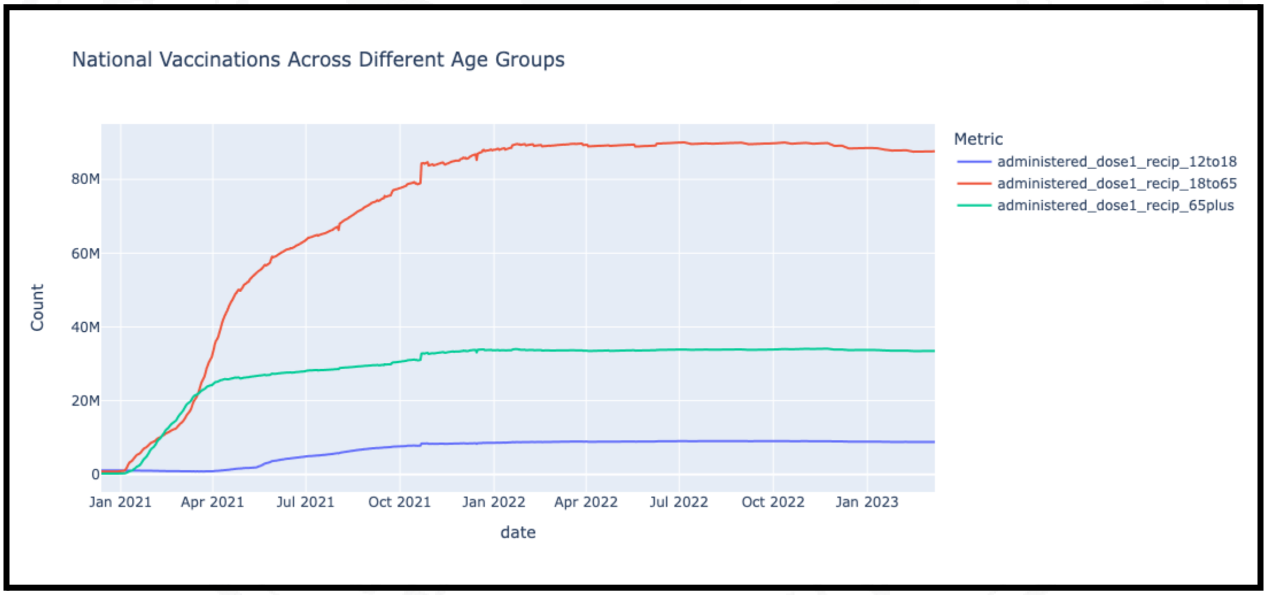
(Fig 1. Geographical & temporal distribution of COVID-19 related deaths, Data : JHU CSSE)



(Fig 2. Geographical & temporal distribution of all age groups with at least 1 dose of vaccination)



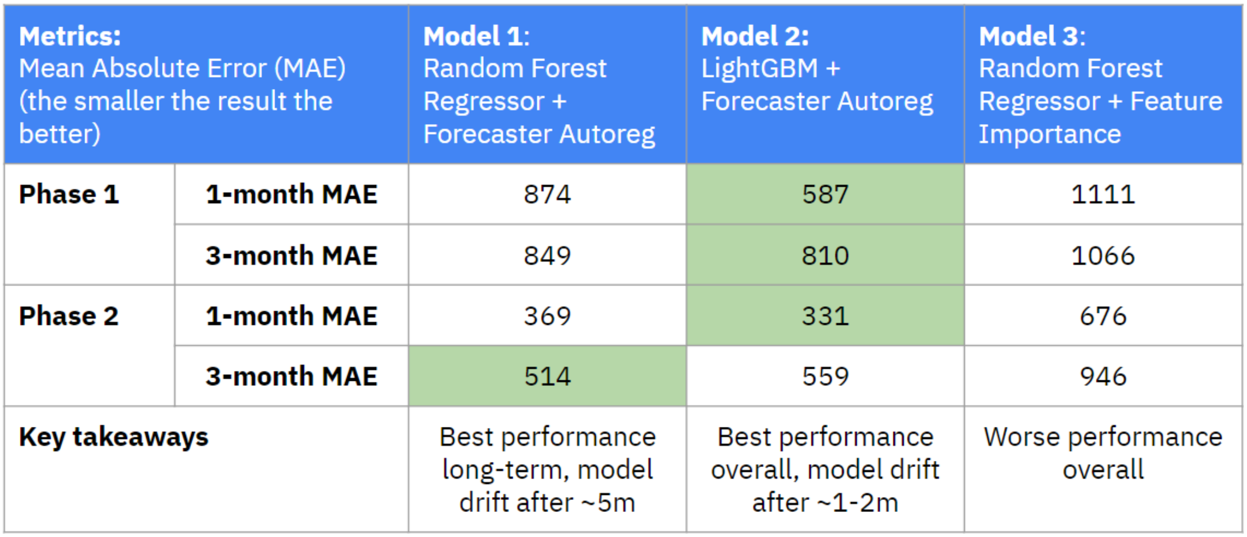
(Fig 3. Geographical & temporal distribution of all age groups with full vaccination status)



(Fig. 4 National Vaccinations across different age groups)

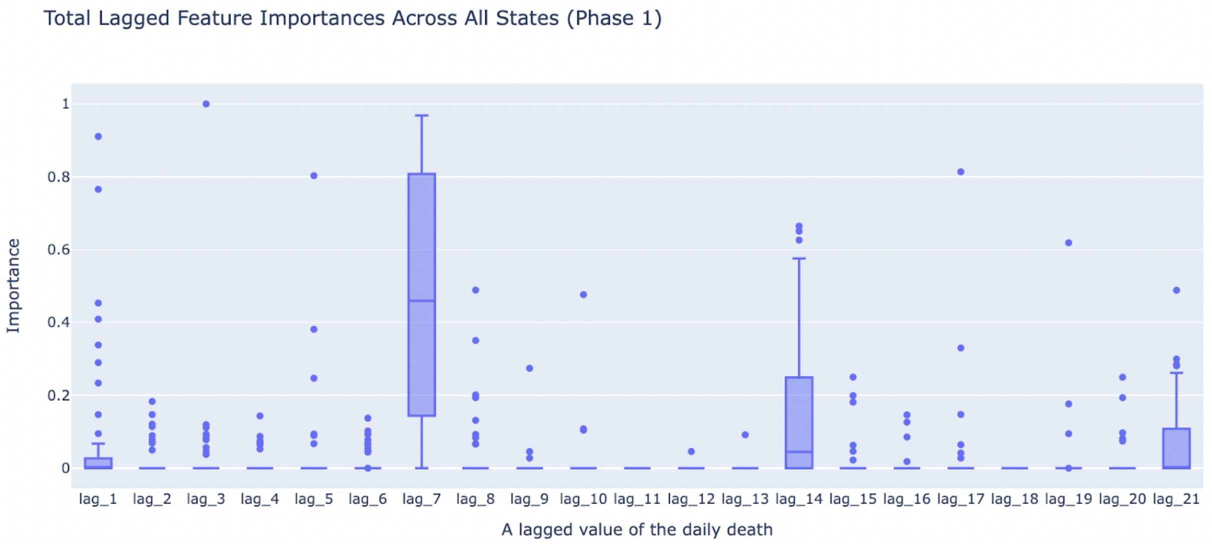
Modeling

Of the models we tried, we found that the Autoregressive Random Forest Regressor performed best for the entirety of the Phase 2 Test Period, while the Autoregressive LightGBM model performed best for the first month, and the entirety, of the Phase 1 Test Period, as well sa the first month of the Phase 2 Test Period. Both of these models outperformed the Random Forest Regressor that did not use autoregression.

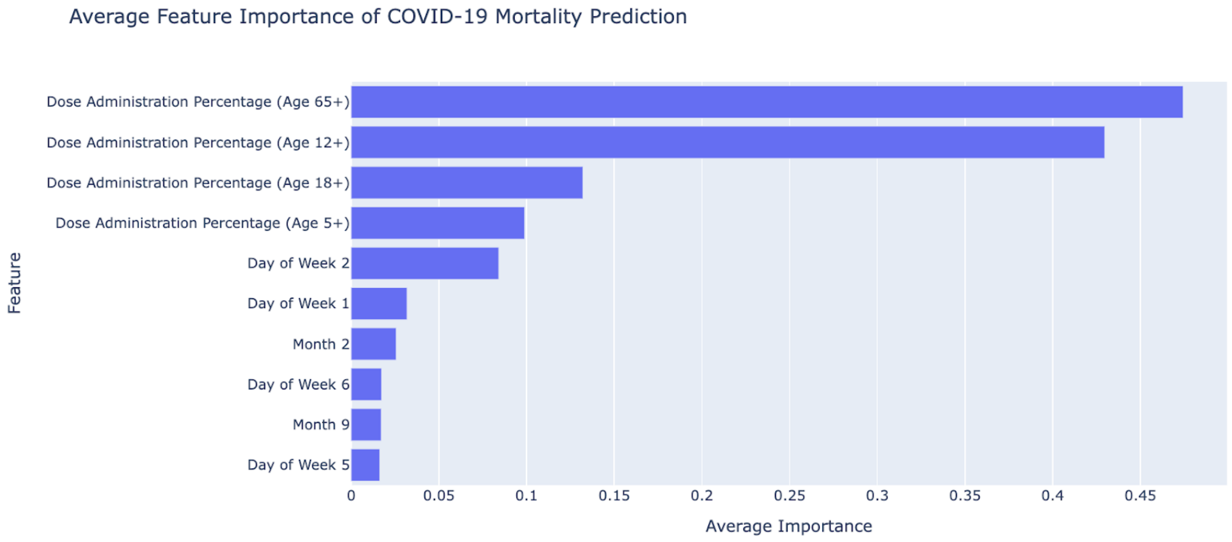


Metrics: Mean Absolute Error (MAE) (the smaller the result the better)		**Model 1:** Random Forest Regressor + Forecaster Autoreg	**Model 2:** LightGBM + Forecaster Autoreg	**Model 3:** Random Forest Regressor + Feature Importance
Phase 1	**1-month MAE**	874	587	1111
	3-month MAE	849	810	1066
Phase 2	**1-month MAE**	369	331	676
	3-month MAE	514	559	946
Key takeaways		Best performance long-term, model drift after ~5m	Best performance overall, model drift after ~1-2m	Worse performance overall

We found interesting patterns in the feature importance of lagged deaths, which coincidentally line up with one, two, and three week lags (Visualizations Appendix, Figure 11, below).

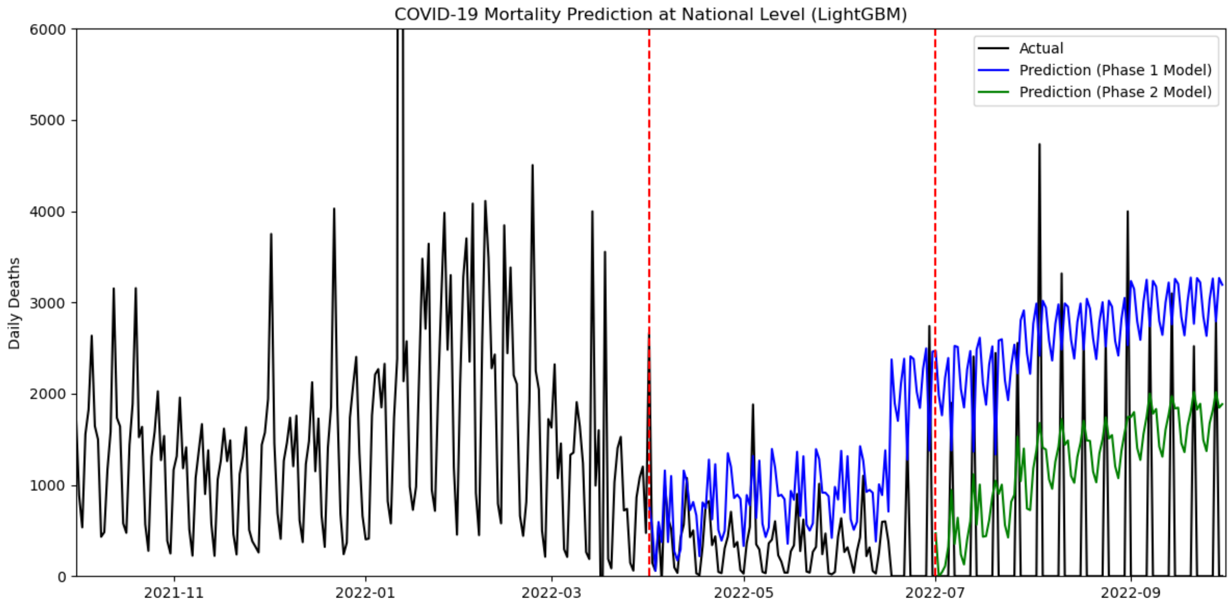


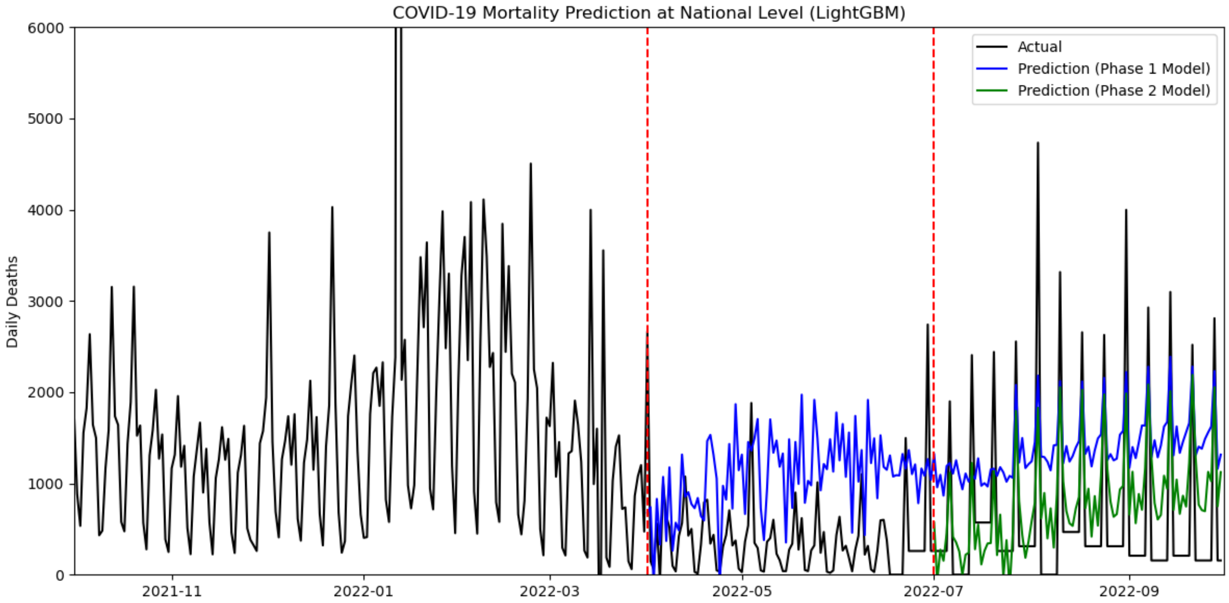
Additionally, as one might expect, we found that dose administration percentages, in particular for 65+ and 12+ had high feature importances (Visualizations Appendix, Figure 12, below).



Model Drift

We saw definite evidence of model drift, in particular for the LightGBM model, which, though more accurate than other models across most of the Test Periods, showed signs of model drift after 1-2 months (Visualizations Appendix, Figures 4-10). As an illustrative example, see, Visualizations Appendix, Figures 4 and 7, below.





Experiments and Deployment

ML Flow

As discussed, we logged experiments for the LightGBM models to ML Flow. The tool proved useful for monitoring the performance of models against each other when using different hyperparameters. However, we did not use it as much for the monitoring of model drift. For this, we had to rely on visualizations, such as those provided above. In order to use ML Flow mode effectively, we would likely want to create a more standardized training and testing pipeline to allow for time-series analysis of performance. Given more time, we might have simulated a weekly re-training and deployment operation in which we could monitor the evolution of model performance on a weekly basis.

Finally, since we trained a model for every single state (or county, depending on the level of data aggregation), we found it difficult to save model artifacts to ML Flow. We leave this as an area for future work.

Docker

Given the dependencies in our code, particularly data science, ML, and forecasting packages, we believe containerization using Docker to be the best packaging and deployment strategy.

Conclusions

The ability to accurately predict adverse COVID-related outcomes, in particular COVID-related deaths, is helpful to policy-makers and frontline workers, because it allows for more efficient allocations of scarce resources. In this paper, we predict national COVID-related deaths in the U.S. for the periods April 2022 through June 2022 and July 2022 through September 22, respectively, using machine learning forecasting techniques that rely on bagging and boosting. We introduced a novel set of features, termed "dose administration percentage," which allows for the consistent use of ever-changing vaccination information over time, across the introductions of new vaccines to fight new variants of the virus. From these analyses, we not only predicted COVID deaths, but also drew a connection between the vaccination rates of various age groups and COVID deaths, showing that vaccination rates for 65+ and 12+ populations are of particular importance nationally. Additionally, we used the process to evaluate the possibility of model drift occurring in our models. We find that model drift and degradation is a definite concern, with even our highest-performing model showing signs of model drift within a 1-2 month period. We monitored the results of our experiments using ML FLow, and have also looked into using docker to package and deploy our work.

References

Q. Li, et al. Early transmission dynamics in wuhan, China, of novel coronavirus–infected pneumonia, New England Journal of Medicine (2020), 10.1056/NEJMoa2001316

B. Xu, B. Gutierrez, S. Mekaru, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. Nature Sci Data, 7 (2020), p. 106, 10.1038/s41597-020-0448-0

World Health Organization. Statement on the second meeting of the International Health Regulations Emergency committee regarding the outbreak of novel coronavirus (2019-nCoV), World Health Organization (WHO). Archived from the original on 31st January 2020

Mohammad Pourhomayoun, Mahdi Shakibi, Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making, Smart Health, Volume 20, 2021, 100178, ISSN 2352-6483, https://doi.org/10.1016/j.smhl.2020.100178. (https://www.sciencedirect.com/science/article/pii/S2352648320300702)

Ayoobi N, Sharifrazi D, Alizadehsani R, Shoeibi A, Gorriz JM, Moosaei H, Khosravi A, Nahavandi S, Gholamzadeh Chofreh A, Goni FA, Klemeš JJ, Mosavi A. Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods. Results Phys. 2021 Aug;27:104495. doi: 10.1016/j.rinp.2021.104495. Epub 2021 Jun 26. PMID: 34221854; PMCID: PMC8233414.