# Covid-19 Deaths Forecast

DATA 200: Graduate Project

Catalina Villouta, Haoran Liao, Vincent Lieng

December 13, 2021

***Abstract*** **– COVID-19 pandemic has created the largest public health crisis in decades. Since the outbreak, there has been tremendous interests in attempting to forecast the confirmed cases and the death tolls, and to predict the course of the pandemic, so as to better inform public health policies. In this project, we make use of publicly available data repository on U.S. COVID-19 related statistics in the year of 2020 and 2021 to build a model that forecasts the death tolls in each state in the next week. We engineered and selected time-lagged features, including how the virus spreads geographically between states, and experimented on several models. In particular, we built a Ridge regression model that achieves a 94% cross-validation R squared with informative interpretations on the various features contributing to the forecast. We hope that our model can be used in assisting the prediction of the course of the pandemic.**

## 1 Introduction

COVID-19 has being a serious public health issue that has impacted the world for the past two years. According to Worldometer,[5] as of December 2021 there are more than 270 million total cases in the world and more than 5 million deaths. In the US alone there are more than 50 million total cases and more than 800 thousand deaths. For those who recovered from the virus, many still suffer from difficulty in concentrating, memory problems, persistent loss of smell and or taste, shortness of breath, fatigue, and more.[2]

The outbreak has drastically changed people's lives and relationships. According to a survey done by Pew Research Center[1] on March 2021, 28% of Americans have degraded physical and mental health, some due to the loss of loved ones. 32% of Americans mentioned the pandemic limited activities they can do in their free time, and 41% of Americans experience isolation and losing connection with family and friends.

Our research question is: what are the factors that are ultimately significant and useful for forecasting COVID-19 deaths in the short term? For answering this question we decided to build a model capable of forecasting total deaths caused by COVID-19 in a week in any particular US state. By constructing this model we could study what are the most prominent factors to predict deaths and how the virus spreads geographically. The insights obtained in this project together with the model itself could be used for short-term public health emergency planning and decision-making.

We followed many careful steps in order to build this model successfully. We used a VAR model to find geographic relationships between states beyond just proximity, which were useful for engineering new promising features. We included sex and age related information to enrich our data. We experimented with an array of regression models: linear regression, Ridge regression, LASSO regression, Elastic net, Random Forest regression, and Gradient Boosting regression. We applied different elegant criteria for feature selection, and measured each model's performance using cross-validation. Finally, we provided our preferred models results and helped interpret the final model and its features.

## 2 Data Description

### 2.1 Sources and data cleaning

We utilized two data sets in our project. The first data set is the USA daily state reports and Provisional COVID-19 Deaths by Sex and Age.

The USA daily state reports are provided by the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.[4] The data set provides 20 features and has a total of 20550 entries. It is an aggregation of multiple data sources, from organizations such as World Health Organization (WHO), European Centre for Disease Prevention and Control (ECDC), DXY.cn, US CDC, BNO News, WorldoMeters, 1Point3Acres, COVID Tracking Project, Los Angeles Times, The Mercury Times, and data from US state department or at county/city level. Since the way each data source collect their data may vary, there might be some inconsistencies and differences in the data from each source. For our project, we included the state name, time each entry was last updated, latitude, longitude, number of confirmed COVID-19 cases, number of COVID-19 deaths, and number of active cases, which is the number of confirmed cases that have not been resolved (total cases - total recovered - total deaths).

During data cleaning, we removed FIPS (Federal Information Processing Standards) code that uniquely identifies counties within the USA, since there was only one per state in almost all cases, making it uninformative. We also removed the cruises, i.e. Diamond Princess and Grand Princess, since they are not US states. We removed a state called "Recovered" which had mostly NaN or 0 values, and thus we concluded it was a typo. We converted the time each entry was last updated to datetime object for easier manipulation. We removed entries that have negative active dates because the number of total cases should not be smaller than the number of recovered cases and the number of deaths. Furthermore, we remove rows with missing dates. Some number of deaths in the data set were negative. For deaths between -5 and 0 we mark them with 0, and for deaths less than -5 we replace the value with NaN. After the preliminary data cleaning, we have 19732 samples for this data set, with the following columns: "Province_State", "Last_Update", "Lat", "Long", "Confirmed", "Deaths", and "Active".

The Provisional COVID-19 Deaths by Sex and Age[3] is provided by National Center for Health Statistics (NCHS). The data is collected by death certificates submitted to NCHS. Because the number of death is only recorded as the number received as of the date of analysis, it may not match the actual number. This causes incomplete data because of the time difference between the actual death and the making of death certificates. This lag may range from 1 week to 8 weeks or more. The data set contains 16 different features and a total of 41310 rows. Features we chose to pay attention to include state, year, month, sex, age group, number of deaths by COVID-19, and number of total deaths. During data cleaning, we removed "United States" and "New York City" from the states column, since they are not states. For the age column we removed "All Ages", and for the sex column we removed "All Sexes" to avoid double counting. There were also many different age groups. For our case, we labeled people from groups 65 - 74 and 75 - 84 as old. Finally, we grouped the data into weeks, instead of months as we had initially intended for, so we had more data points.

## 2.2 Data exploration

To have a general idea of the deaths over time, we show the average of deaths between states in 2020 and into 2021 in Figure 1. We see a peak around the end of 2020. The red line splits our training and testing in time. We also highlight in the plot the start of the testing data, which was chosen to be 2021-02-08 in order to have enough data to compute a robust out-of-sample performance metric. Note that the model will start forecasting weekly deaths for each state, right after the peak of the biggest wave in the training data. We note as well that there is some significant deviations between states based on the size of the confidence interval on the plot. We plot the heatmap of the pairwise correlation of deaths between states, we see that states such as New Jersey, Virginia and Washington D.C. generally have low correlations with other states in the death toll. This is manifested in some light colored strips (low positive correlation) in Figure 2. The most important message related to this graphic is that there seem to be clusters of states that tend to behave similarly. This simple observation made us believe there might be hidden relationships between states that could be useful for forecasting deaths. We will explore this in great depth later in this report.

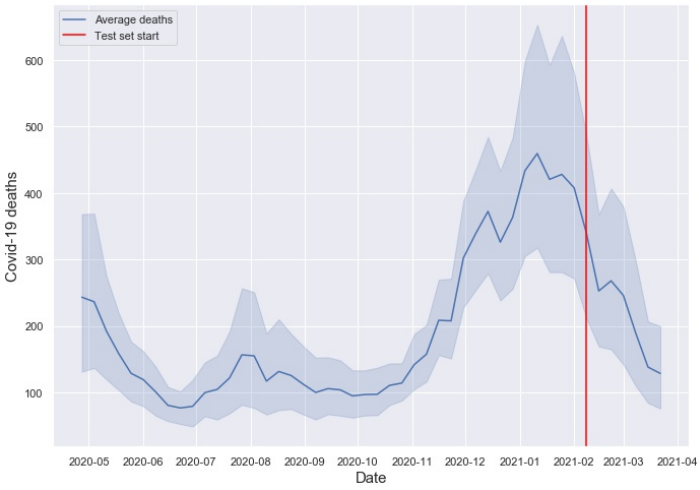


Figure 1: Weekly Covid-19 deaths, average between states

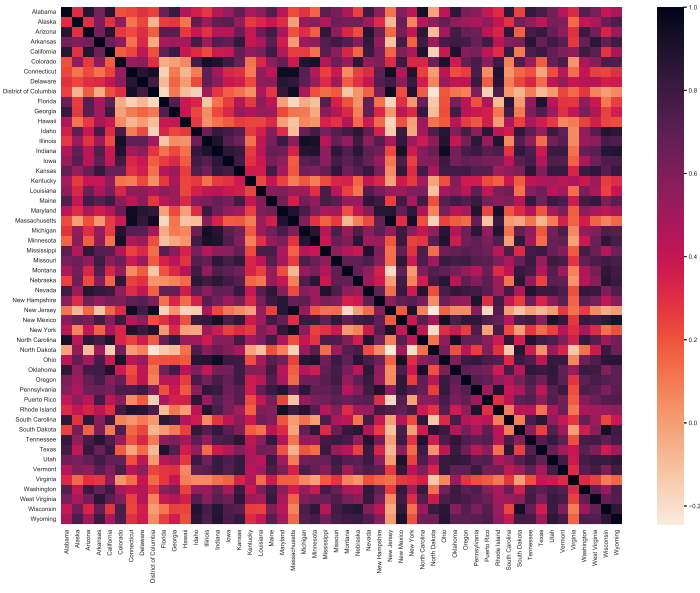

Figure 2: Correlation of deaths between states

## 2.3 Data preprocessing

We computed the difference between deaths and confirmed cases in the source because the given values were cumulative. The difference between a date's value and

the previous date's value would yield the daily cases or deaths. We inputted negative values as mentioned in the Sources and data cleaning subsection. We then grouped the data weekly by state and obtained the sum. We finally dropped all state-weeks which did not have 7 non NaN values.

For the deaths data in New Jersey, we observed an anomalous spike on week 26, as shown in Fig. 3. This spike is possibly due to the adjustment in the death tolls in the early period of the pandemic and we decided to correct it. We replaced it with the average between the deaths on week 25 and 27.

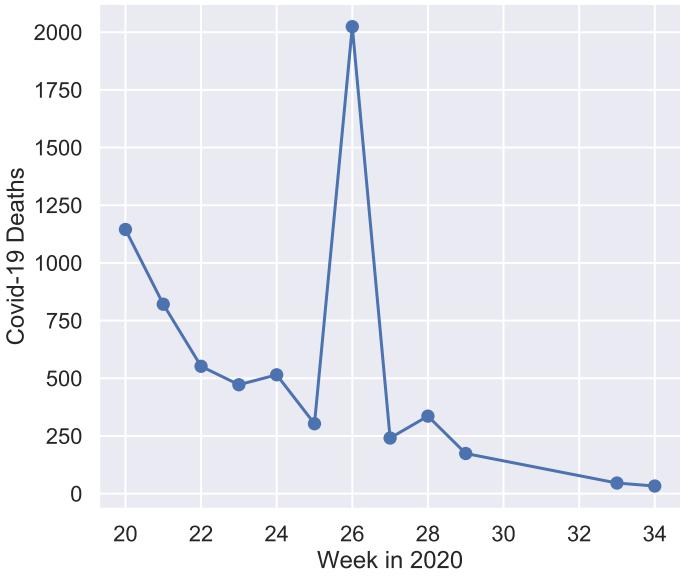


Figure 3: Deaths in New Jersey in period of anomaly

We defined an internal week index for handling the temporal dimension of data more easily. We split the data on week index 41 which starts on 2021-02-08. The training set then consists of all data between week 5 (starting 2020-05-25) and week 41, totaling 36 weeks of data, and the testing set consists of all data after that till week 47, totaling 8 weeks of data. All in all, our training set had 1683 data points and our testing set had 305.

## 2.4 Feature Engineering

### 2.4.1 Same-state features

In order to make forecasts, we make use of the correlations between the past and the future. One way to incorporate the past is to use the statistics from a few weeks ago as our regressors and regress the future week's statistics on these regressors. We therefore chose the cases/deaths four weeks ago, three weeks ago, two weeks ago, and one week ago, respectively denoted as confirmed/deaths_w1, confirmed/deaths_w2, confirmed/deaths_w3, confirmed/deaths_w4 as some of our features. We also included the total deaths and deaths caused by COVID-19 in the past month (m1) as well as the month before the past month (m2), on male, female and old, denoted as total/covid_deaths_male_m1/m2, total/-covid_deaths_female_m1/m2, total/covid_deaths_old_m1/m2. Monthly data for female, male, and old for the past month was mapped carefully to make sure that the information was indeed available before the week being forecasted at that row.

As shown in Fig. 4, looking at the statistics with a 1-month lag, there exist strong correlations among the deaths of male, female and the old , since their pair-plots show very clear linear trends. These features, on the other hand, show relatively weaker correlation with our target – the deaths in the next week. Nevertheless, there are clear linear trends in those features' pair-plots with the target.

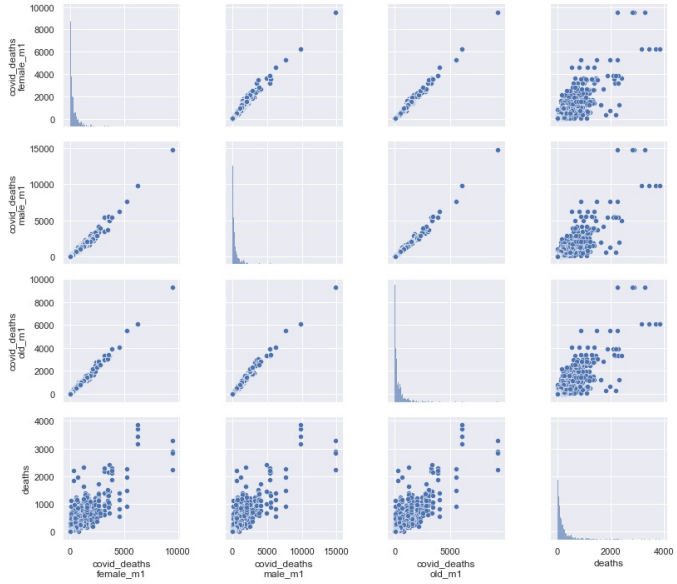


Figure 4: Interactions of deaths and 1-month lagged features

Shown in Fig. 14 are the pair-plots of the deaths of female in the past month (m1) and the month before the past month (m2). We can see a weak correlation between the statistics of the two lags. However, the correlations between these lagged statistics and the death target in the next week are even weaker. We thus concluded that the doubtful forecasting benefit of this m2 features -given the low correlation with the target- is far surpassed by the detriment of loosing one extra month of data in order to include features with a two month lag.

Finally, we plotted the pairwise correlations in a heat map on all the features we have in Fig. 7. We see that the deaths_w1 to deaths_w4 features are correlated in different degrees. We will test the usefulness of these features and we might end up using only _w1. For confirmed_w1 to _w4 features, they are also correlated. At these point, we think it may be good to include _w1, and _w3 or _w4. To help better visualizing the correlations among these features and the target, we additionally plot Fig. 5, where we can see decent linear correlations

between the _w1, _w2 and the target.


Figure 5: Interactions of deaths, lagged deaths, and lagged confirmed cases

### 2.4.2 Inter-state features

We think that geographic locations can play a role in the forecast of deaths. In particular, there may be correlations between the deaths in a state in the near future and the statistics in the states that are close by. We extract the "influential" states to any given state using two methods – using the vector autoregression (VAR) model and using the proximity method, denoted with suffices _influential_var and _influential_prox, respectively.

The VAR model consists of 52 separate regressions wrapped in one convenient algorithm. Each regression will use the deaths of one particular state and regress it against the deaths of all states separately with a lag. We are using daily data here because in this case there would be linear dependence if using weekly as there would be 36 rows (weeks in the training data) and 52 features (number of states), whereas when using daily training data we have 277 rows for each regression and the same 52 features. We say a state X is influential for another state Y if the coefficient of the state X's lagged deaths is significant in forecasting the deaths of Y (we are using a significance of 5% for the p-values). Afterward, we filter only positive coefficients, because we do not believe in a causal relationship where a state's deaths decrease because the deaths in another state increased one day before. To this point we were able to determine a list of influential states for each state. A sample of states with its influential states can be seen in Table 1.

For the proximity model, we just found the 3 states that are closest, by utilizing the latitude and longitude of any state available in the data sources mentioned. It is quite interesting to compare the results of the VAR model, which are based on predictive power, to the results based on proximity in Table 1 and Table 2. We can see that VAR-based influential states are often related to proximity, but it does uncover some other relationships that might be related to air travel. For instance, based on the proximity method, the closest states to New York (NY) are New Jersey (NJ), Connecticut (CT) and Pennsylvania (PA). However, the VAR results indicate that Massachusetts (MA) is better than NJ or PA in predicting deaths one day ahead. For Ohio (OH) the VAR model chose one neighboring state, Kentucky (KY), but also chose Michigan (MI) and West Virginia (WV) which are indeed also next to it, but not by the proximity criteria. Texas (TX) is a very interesting case as none of the proximity-based states were chosen as influential under the VAR criteria. California (CA) and Florida (FL) were chosen as VAR-based influential states, which fits the story that southern states' COVID-19 cases were at some point in the pandemic in a *wave* together. Finally, CA is the state that is influenced by the most states when compared to the states on the table. None of the states in the proximity-based list are in the VAR-based list, which might indicate that CA has more exchanges with a wide range of sometimes distant states.

To obtain the final VAR-based feature, we scale deaths_w1 (or other feature) of all the influential states (var or prox, separately) using the deaths pre-COVID. The reason we do this is that we want the feature to be of the right order of magnitude, and not obtain a lower estimate for states that have less populated states as influential states. To be precise, let $X_1, \ldots, X_k$ be the VAR or proximity based states list of a state Y. We computed a ratio for each of the states $X_i$ on the list defined as the average deaths pre-covid of state Y divided by the deaths pre-covid of state $X_i$. Then, when multiplying a particular feature of an influential state (e.g. deaths_w1) by the the ratio, then the resulting feature will be of a similar order of magnitude as in state Y. After scaling the feature for all neighboring states, we compute the simple average, and that is our final feature for state Y. The same scaling and averaging methodology was used for proximity-based states. We initially computed this feature for deaths_w1 using for both VAR and proximity based states, and called the features deaths_w1_influential_var and deaths_w1_influential_prox, respectively. Then, after noticing that the influential_var feature was useful we came back and added a VAR-based states feature using confirmed_w2, which is naturally called confirmed_w2_influential_var.

| State | Influential States |
|---|---|
| CA | CA, DE, KY, LA, NV, TN, VT |
| NY | CT, MA, NY |
| OH | IN, KY, OH |
| TX | CA, FL, ID, TX |

Table 1: VAR-based influential states sample

| State | Neighboring States |
|---|---|
| CA | NV, AZ, OR |
| NY | NJ, CT, PA |
| OH | WV, KY, MI |
| TX | OK, LA, AR |

Table 2: Proximity-based states sample

In Fig. 6, we plot the pairwise correlations among deaths_w1_influential_var, the deaths_w1_influential_prox, confirmed_w2_influential_var, and the target (deaths). In particular, we observe relatively high linear correlations between the influential states features and the target, indicating potential usefulness of the influential state death and confirmed features.

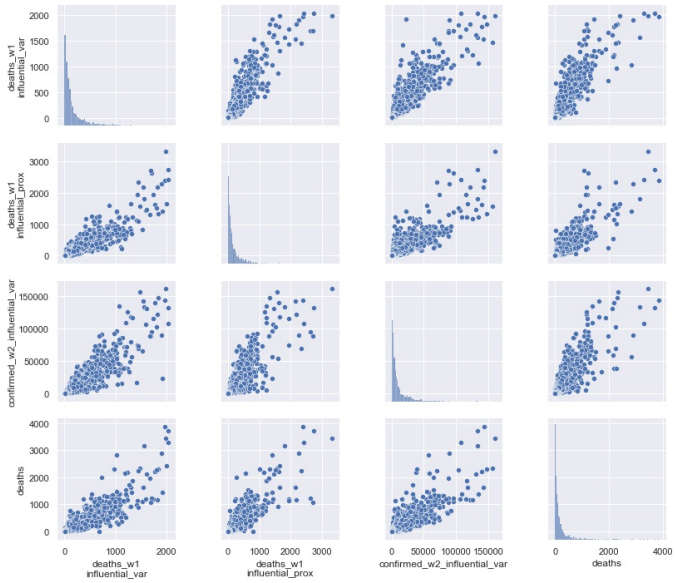


Figure 6: Interactions of deaths and influential states based features

The last neighboring states features are not as highly correlated with the rest and they seem to have a high correlation with the target, so they are very promising features to select. We acknowledge that there is high correlation between our features which could bring colinearity issues when running linear regression for instance. Given the relatively large amount of data as compared with the number of features, we tend to believe that features with a correlation of around 0.9 and less could be used together, but we will be careful going forward. We note that some of the features here are completely dependent, e.g., knowing the deaths of the past four weeks, as well as the deaths of males in the past month, will determine the deaths of females in the past month. We discard the entire *m2* statistics, and the COVID-19 and total deaths of *female* and *male* in *m1* based on the correlations in favor of keeping only the *old* features.

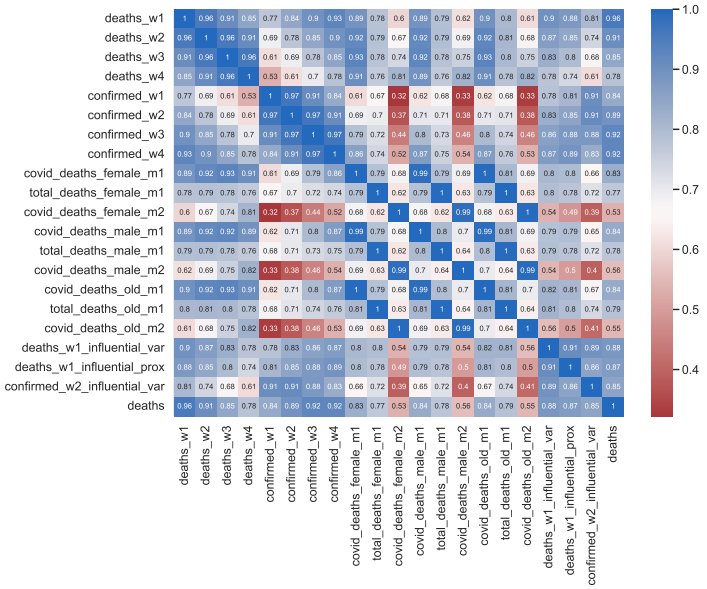


Figure 7: Correlation of deaths and all features

# 3 Methodologies

## 3.1 Performance assessment

Model selection was based on training data results alone. We left the test set to compute the performance only after making our model decision. We computed two metrics, R squared and RMSE, using two methodologies we call train and cross-validation. We refer to train metrics as those obtained by using the whole training set to fit the model and also the whole training set to compute the metric. This metrics will be subject to overfitting. Cross-validation metrics intend to compute a better representation of out-of-sample performance using the training set only. For this we used 5-fold cross-validation. Even though we compute train metrics, our decision to choose a model was based on cross-validation metrics only.

We chose R squared because it is a widely used and easy to understand metric. Values close to 1 mean that 100% of variability was captured by the model, while a value of 0 mean that the performance of the model is equivalent to using the mean as a forecasting model. We also chose RMSE as it can be thought of a weighted average of errors (in deaths) that are weighted more heavily to larger errors. This is because the metric first squares the error, before averaging and taking square root, which exacerbates notable misses. We thought RMSE was a good addition to our evaluation tool kit because RMSE is easy to interpret and also because the units are the same as the target, deaths in this case.

## 3.2 Feature selection

We considered three main methodologies for selecting the features of our model: LASSO, linear regression p-values, and cross-validated metrics.

LASSO uses L1 regularization which has the nice property of quickly eliminating features that do not contribute highly enough to decrease MSE. There is a regularization parameter (alpha in sklearn) that controls the trade-off between MSE and the L1 norm of coefficients. Our methodology here was based on plotting the coefficient values for different values of alpha. Since we want LASSO to eliminate only clearly useless features, we used a small alpha in LASSO in order not to zero out too many features. LASSO is incredibly powerful for feature selection, specially because its execution time is extremely low. It is a great tool for doing a initial selection of features, specially if the number of features considered is relatively large.
In linear regression, we check whether these LASSO features are significant by looking at the t-stat and related p-values of the fitted coefficients for these features. High p-values indicate that we cannot reject that the coefficient accompanying the variable is equal to zero, i.e. the feature is not statistically significant. Removing a feature with a high p-value of the t-test could improve cross-validation metrics, but it is not guaranteed. This methodology also allows to start from a model with more features than those selected by LASSO and then use the p-value as a guide to prune the model into one that generalizes better out-of-sample data.
Ultimately, it is cross-validation performance the driver for decisions. LASSO and linear regression p-values are used as a guide to trying promising models, but the final model selection is based on cross-validated R squared and RMSE.

### 3.3 Models

We propose several models for the death forecast task. The first to consider is ordinary linear regression (OLS). This is a natural choice since we observe clear linear correlations between the lagged death statistics and the target, which was discussed in detail in Sec. 2.4. The second model to consider is Ridge regression. Ridge regression adds a $L_2$-norm regularization term to the OLS, so as to penalize the size of the model coefficients, thus possibly reducing the number of features used. This also allows us to determine which features are useful or not. Some "co-linear" features may be reduced when using Ridge regression. The third model to experiment on is the LASSO regression, which adds an $L_1$-norm to the OLS. This regularization is more aggressive in setting some coefficients to zero than Ridge regression, which is expected to help in the feature selection. The fourth model to try is the Elastic Net. It is a regularized regression method that linearly combines the $L_1$ and $L_2$ penalties of the LASSO and Ridge regression. We will also try Random Forest regression which is an ensemble learning method for regression that combines bootstrapping of the dataset and the bagging of many decision trees trained on the bootstrapped dataset. The last model to consider is the Gradient Boosted regression. This model creates decision trees in sequence where each new tree classifies and improves on the residuals of the previous tree.

### 3.4 Hyperparameter tuning

For both LASSO and Ridge regressions, we have the hyperparameter alpha – the regularization coefficient. We tune the hyperparameter by either doing a linear grid search and choose the alpha that gives the best 5-fold cross-validation RMSE. We also use grid search to choose the hyperparameters of Elastic Net, Random Forest regression, and Gradient Boosted regression.
For Elastic Net we consider alpha and l1_ratio in the grid, the second being the trade-off between L1 and L2 regularization. With l1_ratio of 1 it is in essence running a LASSO regression while a value of 0 is in essence running a Ridge regression (with some slightly different scaling).
For Random Forest regression we included the number of estimators (trees) and the maximum depth (of each tree) into the hyperparameter grid, two choices that greatly affect the dynamics of the model. Similarly, for Gradient Bossted regression we searched over the number of estimators using cross-validate grid search.

## 4 Results

### 4.1 Benchmark

Our benchmark is very simple yet effective. It predicts next week's number of Covid-19 deaths based only on this week's COVID-19 deaths. The training R squared was 93.67% and the cross-validated R squared was 91.32%. We will use the latter as our best reference using the training data alone. The cross-validated RMSE was 85.58 deaths, which looks reasonably small since we are talking about state wide deaths. The model basically forecasts for week $w$ in state $s$ almost the same amount of deaths that occurred in week $w-1$ in state $s$. We can see that our benchmark is highly effective in forecasting next weeks death for any state, and we believe it is a hard benchmark to beat. Fig. 15 shows the Fitted vs Actual deaths on the training data.

### 4.2 Model results and selection

We used LASSO for initial feature selection as detailed in the Methodology section. Fig. 8 shows the values of the coefficients depending on the level of alpha, the regularization coefficient (the higher the stronger the regularization). We can see how LASSO effectively zeroes out coefficients as alpha increases. We decided to choose alpha equals to 5e-4 depicted in the plot as a black line. With this threshold, the selected features are confirmed_w1, confirmed_w2, confirmed_w3,

total_deaths_old_m1, deaths_w1_influential_var, and confirmed_w2_influential_var. Notice that our VAR-based influential states features are included using this methodology, while the proximity-based feature was discarded.

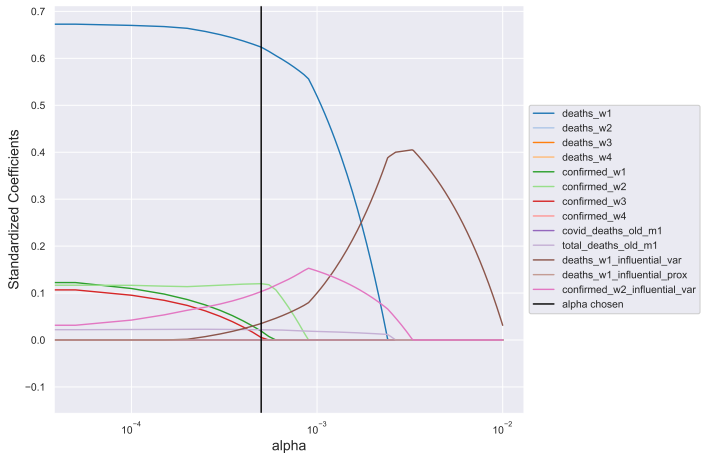


Figure 8: Lasso coefficients as a function of alpha

We fitted a linear regression using only LASSO-based features and their was a considerable improvement of the model over the benchmark. Using cross-validation, we obtained an R squared of 94.02% and a RMSE of 71.25 deaths.

In the linear regression, we checked whether these LASSO features were significant by looking at the t-stat and related p-values of the fitted coefficients for these features. There are two features with a p-value much larger than 0.05, one is the confirmed_w2 with a p-value of 0.78 and the other is total_deaths_old_m1 with a p-value 0.70 as seen in Table 3. Therefore, these two variables are not significant. All the other LASSO features have a t-stat larger than 2 in the linear regression.

| Features | Coef | Std Err | t-stat | p-value |
|---|---|---|---|---|
| deaths_w1 | 0.911 | 0.02 | 59.3 | 0.00 |
| confirmed_w1 | 0.081 | 0.03 | 3.0 | 0.00 |
| confirmed_w2 | 0.012 | 0.04 | 0.3 | 0.78 |
| confirmed_w3 | -0.072 | 0.03 | -2.4 | 0.02 |
| total_deaths_old_m1 | 0.003 | 0.01 | 0.4 | 0.70 |
| deaths_w1_inf_var | -0.093 | 0.01 | -7.1 | 0.00 |
| confirmed_w2_inf_var | 0.171 | 0.01 | 13.5 | 0.00 |

Table 3: Regression coefficient results for LASSO-based features

This gave us enough encouragement to try selecting a set of features based on p-values instead of L1 regularization. We ran a linear regression using a bigger feature set than the one selected by LASSO, where we considered deaths of previous weeks, confirmed cases of previous weeks, covid_deaths_old_m1, total_deaths_old_m1, deaths_w1_influential_var, and confirmed_w2_influential_var. We dropped insignificant variables one-by-one: first total_deaths_old_m1, and then confirmed_w2. This gave us a model where all the variables are significant under linear regression model's assumptions. The regression coefficient results can be seen in Table 4. However, the cross-validated R squared and RMSE where inferior, with 93.89% and 71.87 deaths respectively.

| Features | Coef | Std Err | t-stat | p-value |
|---|---|---|---|---|
| deaths_w1 | 0.84 | 0.03 | 31.1 | 0.00 |
| deaths_w2 | 0.13 | 0.03 | 3.9 | 0.00 |
| deaths_w3 | -0.07 | 0.04 | -1.9 | 0.06 |
| deaths_w4 | -0.07 | 0.02 | -3.2 | 0.00 |
| confirmed_w1 | 0.10 | 0.02 | 4.8 | 0.00 |
| confirmed_w3 | 0.14 | 0.04 | 3.4 | 0.00 |
| confirmed_w4 | -0.22 | 0.03 | -6.3 | 0.00 |
| covid_deaths_old_m1 | 0.09 | 0.02 | 4.0 | 0.00 |
| deaths_w1_infl_var | -0.04 | 0.01 | -3.3 | 0.00 |
| confirmed_w2_inf_var | 0.11 | 0.01 | 8.2 | 0.00 |

Table 4: Regression coefficient results for significant variables

Our next attempt was to add regularization to the model, in particular L2 regularization with Ridge regression. Using only LASSO-based features, we found an optimal cross-validated alpha (regularization parameter) of 0.05 using grid search. The fitted coefficients are shown in Table 5. As anticipated from the linear regression results of the previous model we see a prominent role of deaths_w1 and the confirmed features. It is important to note that since we are not standardizing our features, confirmed features will tend to be 50-100 times larger, since SARS-COV-2 kills around 1-2% of people that contract COVID-19. Taking that scaling note into account, we see that the importance of deaths and confirmed variables is comparable, albeit confirmed has a slightly lesser role. VAR-based features have a lesser role than confirmed using the previous criteria, but are statistically significant nevertheless.

| Feature | Coef |
|---|---|
| deaths_w1 | 0.66 |
| confirmed_w1 | 0.13 |
| confirmed_w2 | 0.12 |
| confirmed_w3 | 0.14 |
| total_deaths_old_m1 | 0.03 |
| deaths_w1_influential_var | -0.01 |
| confirmed_w2_influential_var | 0.03 |

Table 5: Coefficients Ridge regression with LASSO features

The performance of this model was marginally better in R squared than the linear regression model using LASSO-based features, with cross-validated R squared of 94.03% and RMSE of 71.45 deaths. We also tried adding all features that were not discarded at the EDA, to see if Ridge regression might be able to handle spurious relationships on its own better. The optimal cross-validation-based alpha was 0.025. It got a cross-validated R squared of 93.98% and RMSE of 70.57. So an inferior R squared and a better RMSE. Up to this point we were using R squared as our first criteria, and so our incumbent was Ridge with LASSO-based features.

Fig. 16 shows the Fitted vs Actual deaths on the training data. Improvements over the same benchmark chart in Fig. 15 can be spotted looking close enough.

We also tried LASSO and Elastic net directly using all reasonable features, but the grid search yielded an optimal alpha of 0 in both cases, i.e. the models are in essence an OLS model using all reasonable features. The result was a cross-validated R squared of 93.97% and a RMSE of 70.52 deaths. The reason for Elastic net not finding our preferred Ridge model is that it is using MSE (or equivalently RMSE) as the selection criteria when doing cross-validation. LASSO and Elastic net documentations warn users that the algorithms may have instability issues when using alpha equal 0. So if we were to choose this model based on RMSE, then it would be better to go with a linear regression with the same features. Since we were using R squared as our preferred metric, our preferred model was still the Ridge model with LASSO-based features.

Our last trial was to change the model paradigm more aggressively away from minimizing MSE with or without regularization. We tried Random Forest (RF) regression and Gradient Boosting (GB) regression. The optimal hyperparameters for RF regression were a maximum depth of 10 and 200 estimators. An attempt of using weak (high bias) estimators in big numbers to try to reduce both final bias and final variance. Despite our careful hyperparameter search, the cross-validated R squared of Random Forest regression was 90.65% and the cross-validated RMSE was 104.34; much worse than any other model considered, including the benchmark. Results were similar for GB regression.

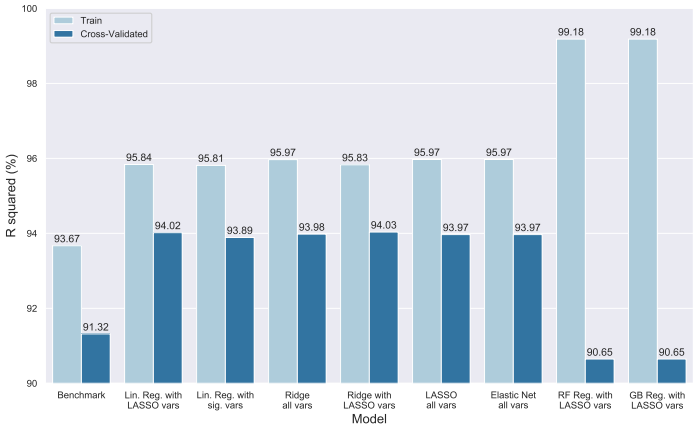


Figure 9: Train and Cross-Validated R squared for different models

Fig. 9 and Fig. 10 show train and cross-validation metrics for R squared and RMSE respectively. We can see that all linear regression based models (Lin. Reg., Ridge, LASSO, Elastic Net) with enough features performed similarly and better than the benchmark, with around 95% train R squared and 94% cross-validated R squared. However, RF regression with LASSO variables and GB regression with LASSO variables both suffer from overfitting as it can be evidenced by the wide gap between train metrics and cross-validated metrics. For instance, train or fitted RMSE for RF is 30.52, while the cross-validated RMSE is 104.35. The fitted RMSE for GB regression is also 30.52, and the cross-validated RMSE is 104.35.

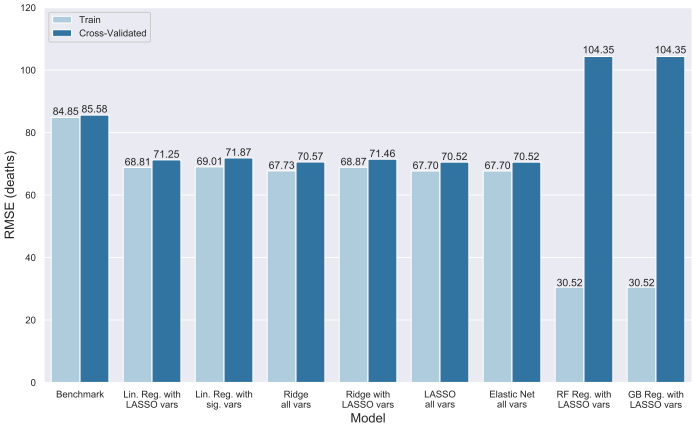


Figure 10: Train and Cross-Validated RMSE for different models

We decided to choose Ridge regression with LASSO variables as our final model since it had the best performance based on cross-validated R squared score of 94.03%. Fig. 17 and Fig. 18 show the same previous plots but on test data now. Again, the differences in these plots are subtle and seems like the improvement in metrics is driven by forecasting better the mass of points rather than the occasional off-pattern weeks.

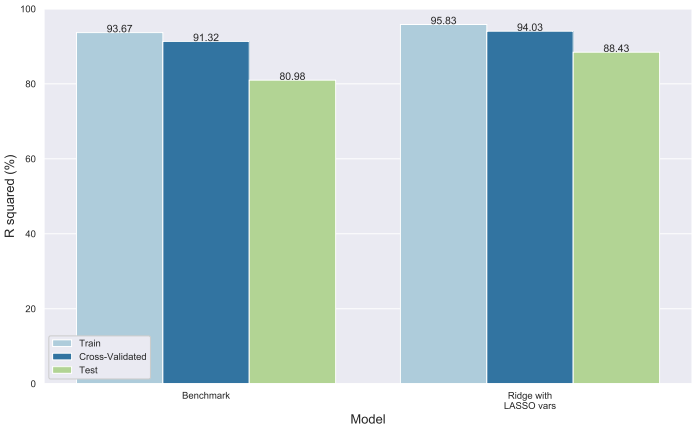


Figure 11: Train, cross-Validated, and test R squared for different models

Finally, Fig. 11 and Fig. 12 show train, cross-validation, and test performance for both the benchmark model and our preferred model: Ridge regression using LASSO-based features. We stress that this performance was genuinely checked only after the model had been selected, in order to make it a true reflection of out-of-sample performance. Test R squared was 88.43% for the preferred model while 80.98% for the benchmark. Test RMSE was 137.63 deaths for the preferred model and 176.60 deaths for the benchmark. Both metrics show a notable

improvement over the baseline. We will further discuss our test results in the discussion, including the reasons why we believe the test performance was significantly worse than what we measured using cross-validation.

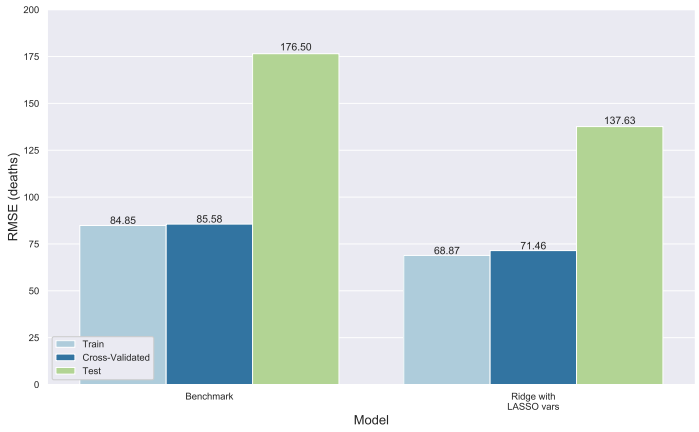


Figure 12: Train, cross-Validated, and test RMSE for different models

### 4.3 Bootstrapping

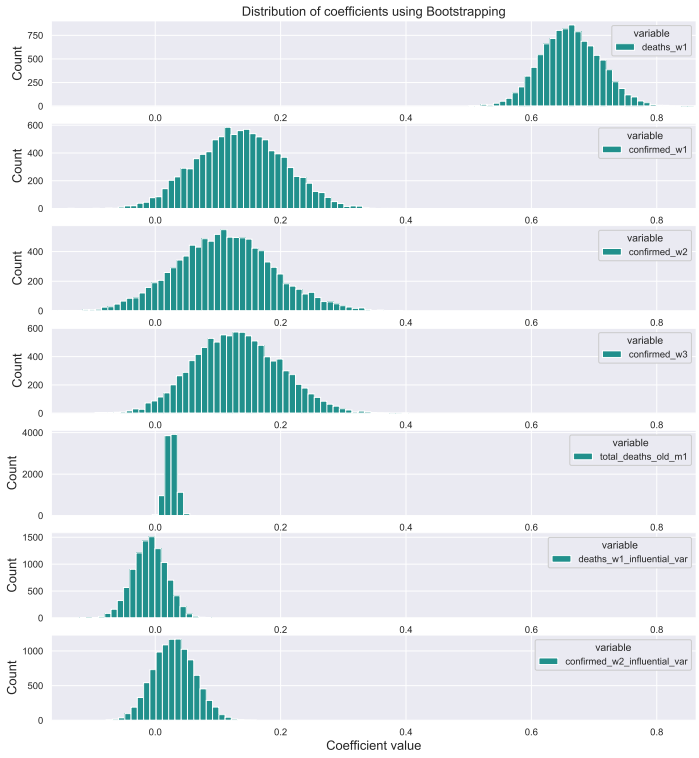


Figure 13: Distribution of coefficients using Bootstrapping

We performed boostrapping to discover the empirical distributions for the coefficients. This distributions are well studied in the case of linear regression, but for Ridge regression, our preferred model, the added regularization term can introduce some distortions to the distributions we already have for the linear regression case available in Table 3.

Bootstrapping results can be seen in Fig. 13. Feature deaths_w1 has the largest overall coefficient values, with a minimum of 0.50. Thus, it is safe to assume that the true coefficient of such a model would be positive and significant. Feature confirmed_w1, confirmed_w2, confirmed_w3, and total_deaths_old_m1 have also a distribution on the positive side with very little chance of the coefficient being equal to 0.

Finally, confirmed_w2_influential_var is likely positive, but there is a chance that it is not-significant based on this methodology. Feature deaths_w1_influential_var is centered very close to 0, making this feature the most likely to be insignificant in the presence of all other features.

## 5 Discussion

Table 5 shows the coefficient values of our chosen Ridge regression model with LASSO-based features. We see that a main feature of the model is deaths_w1, which has a coefficient of 0.66. Hence, our starting point is 66% of last weeks cases and we add other terms from there. We noticed that confirmed_w1, confirmed_w2, and confirmed_w3, are also prevalent factors when forecasting deaths. This makes perfect biological sense since people who died this week must have contracted the disease previously. Having only this result for understanding the evolution of the disease, we have strong indication that people that contract COVID-19 may die 1-3 weeks later.

We also see that total_deaths_old_m1 contributes to the forecast, although we have to remember from Table 3 that the coefficient ends up being not statistically significant. So we caution on the direct interpretation and its use for planning. We did however found that the variable is likely to be positive and significant using bootstrapping on the Ridge regression model, which is the right one to use (Figure 13). Additionally, the feature does makes sense. States with more total deaths of people with more than 65 years of age, will tend to have more people of that age, and thus tend to have more deaths due to the virus, since the virus tends to be more lethal the older you are.

The last set of features in the model are those based on influential states. We believe that using a VAR model for capturing geographical connections between states was a creative feature which allows for interesting interpretations. We saw in the Feature Engineering subsection that we found for each state a list of influential states based on daily regression results using training data only. Influential states tended to be states that are close, but not always, offering interesting relationships between states in terms of spreading the virus across border enabled by car and air travel mostly. We see in Table 3 that confirmed_w2_influential_var had a positive sign of around 0.03 while deaths_w1_influential_var had a negative value of around -0.01, which tends to net the former one, in a relationship which is not entirely

intuitive.
Negative signs are not always incorrect or even not intuitive. For Table 4 we see that the linear regression chose positive signs for the first weeks of deaths and confirmed and negative signs for the last weeks. This is actually an acceleration signal when they are taken into account. Take confirmed as an example. The values of confirmed_w1, confirmed_w3, and confirmed_w4 are respectively 0.1, 0.14, and -0.22. Now if confirmed cases are constant, it will forecast around 2% of deaths driven by confirmed cases. However, if the confirmed cases in w1 and w3 are higher than in w4 it will predict more, while if confirmed cases are decreasing it will predict less. This could be deemed as trend following, and we see that our Ridge regression model avoids for the most part this kind of strategy.
Finally, we note that test error was much larger than what would have been expected based on cross-validation metrics, as seen in Fig. 11 and Fig. 12. This could be a sign of overfitting, but we believe there are also other rationals behind this result. Fig.1 from the Data Description depicts very clearly the last *wave* available up to the point the data was downloaded. We can see that average deaths between states was much lower for the most part of the training set, while the test set starts from the peak of the deadliest wave seen so far. That is partly -we believe- a reason for the weak test performance as compared with the cross-validation results. Another reason is just a pure miss in deaths, which is also related to the choice of test set. The test set has the biggest drop in cases seen up to that point. The training set had no reduction in cases as dramatic as the one seen in the test set. This makes the model forecast less deaths than actually happened.

## 6 Conclusions

In this project we have carefully built a regression model to forecast the number of weekly deaths that will happen in a state one week in advance. Our motivation was to understand the factors that are significant and useful in predicting deaths.
Our data consisted of daily deaths, and confirmed cases for each state, together with monthly age and sex related data with COVID-19 related deaths and total deaths per group. We engineered a number of features we had reason to believe might be useful for the task. We first grouped daily deaths into weekly deaths which was our target variable. We then mapped the deaths and confirmed cases of previous weeks as features. We also mapped all sex and age related monthly data making sure the data was indeed available before the week we were trying to forecast. Finally, we used a VAR model to determine the ability of each state in forecasting one time lag ahead deaths of another state. This allowed us to build a feature based on so called influential states, which were often close to the target state, but not always, uncovering deeper functional relationships between states.

Our preferred model was a Ridge regression with LASSO-based variables. We call LASSO-based variables to the features that were selected after using LASSO regression to zero out most non-significant variables. Our preferred model had the highest cross-validation R squared of 94% as compared to a cross-validation R squared of 91% for our benchmark model which uses only past week deaths. In terms of cross-validation RMSE, our model achieved 71 deaths as compared with 86 deaths. A substantial improvement if we consider that we are erring 15 deaths less on average for all states a week.

We found that the most important features considered in forecasting COVID-19 deaths were deaths the week before, confirmed cases each week for the last 3 weeks, total deaths of people over 65 years old, and confirmed cases of previous weeks of states found to be influential. All of this features were significant from a statistical and bootstrapping perspective, and all contributed to the improved cross-validation performance.

Our model can be used by health professionals and authorities to plan when there are potentially higher number of death and people in ICU which is highly correlated with deaths. Authorities may need to take drastic measures as adding temporal clinic beds to receive and treat critical patients. An early sign of a sharp increase in deaths can help better plan the capacity of health facilities and the state system which can save lives. The downside of this kind of work is that forecasting deaths can cause high anxiety for many people, which can affect their mental health and ultimately their quality of lives. It might be in the best interest of the general population to not publish this kind of model, even it is valuable information. It is a hard line to draw.

For future work, we could consider using backtesting to mimic how the model would update coefficients weekly as more information from test set becomes available. In the current setting, the models coefficient are fixed after training, and are not updated any more as it then forecasts week by week COVID-19 deaths in the test set (the last 9 weeks of data available). We noticed that COVID-19 deaths decreased sharply right at the beginning of the test set. The pandemic's dynamics change over time, and a backtesting would allow the model to update its parameters as each week of data becomes available, which would be a more realistic out-of-sample test. Regardless of this possible improvement, we can say that COVID-19 waves are very hard to predict! Hence, public health decision makers would benefit from being more conservative.

# 7 Annex

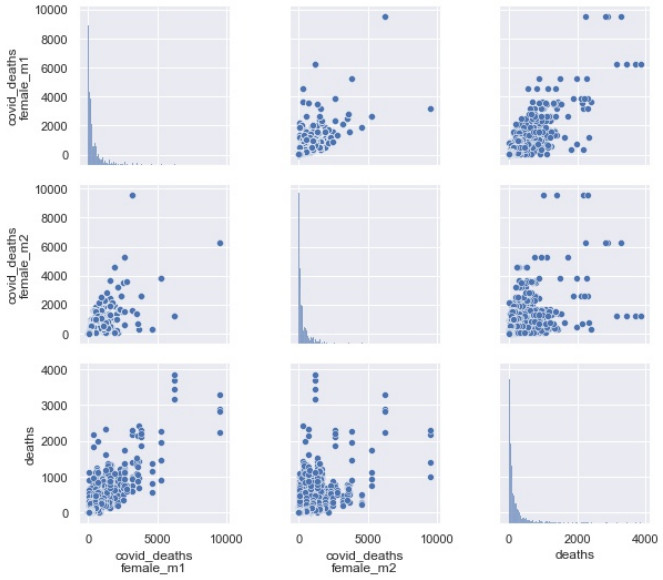


Figure 14: Interactions of deaths and 2-month lagged features

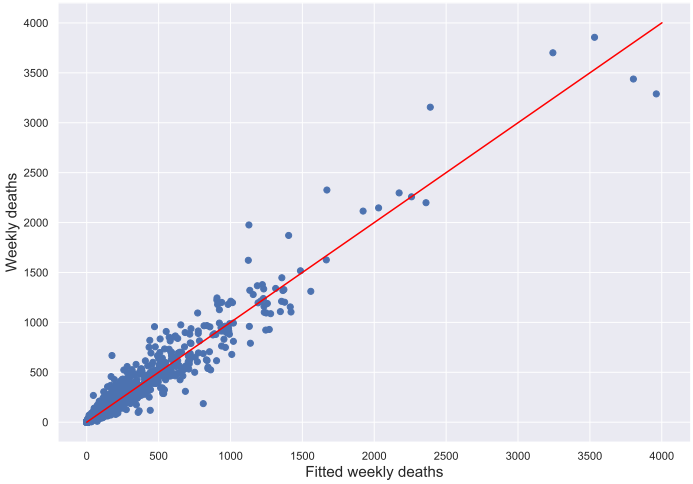


Figure 15: Fitted vs actual using training data with Benchmark model

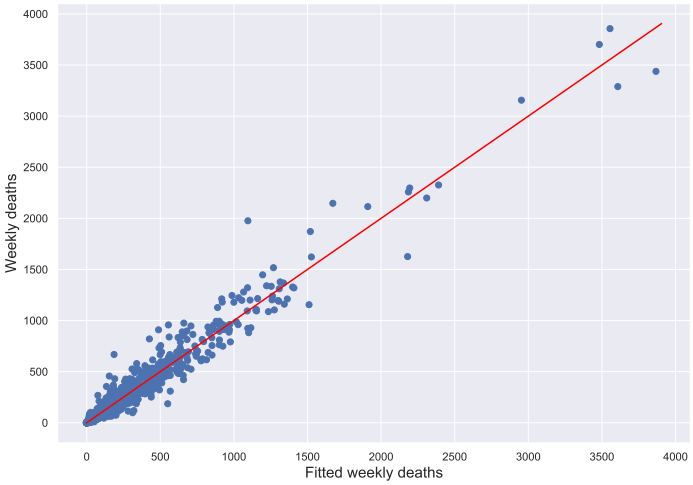


Figure 16: Fitted vs actual using training data with Ridge model

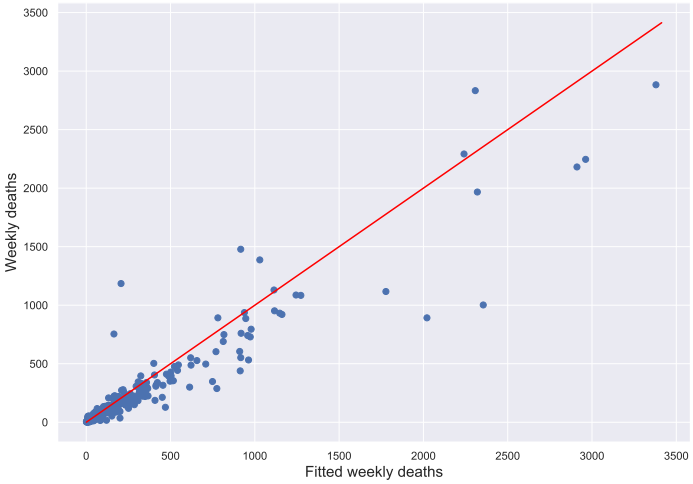


Figure 17: Fitted vs actual using test data with Benchmark model

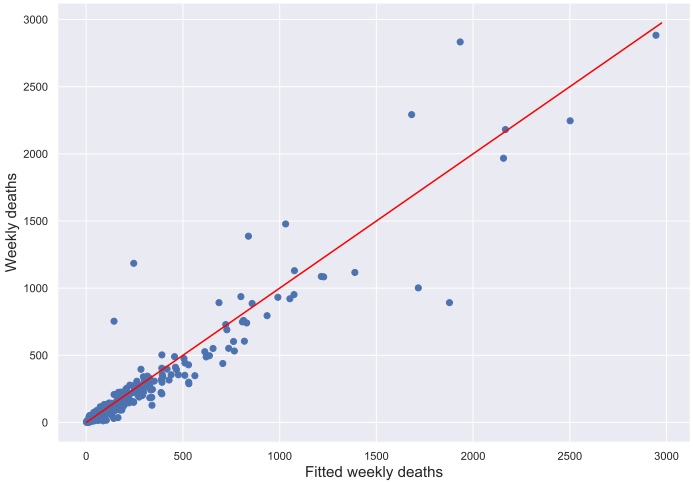


Figure 18: Fitted vs actual using test data with Ridge model

## References


[1] Pew Research Center. *In Their Own Words, Americans Describe the Struggles and Silver Linings of the COVID-19 Pandemic*. Mar. 2021. URL: https://www.pewresearch.org/2021/03/05/in-their-own-words-americans-describe-the-struggles-and-silver-linings-of-the-covid-19-pandemic/.

[2] Centers for Disease Control and Prevention. *Post-COVID Conditions*. Sept. 2021. URL: https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html.

[3] National Center for Health Statistics. *Provisional COVID-19 Deaths by Sex and Age*. Dec. 2021. URL: https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku.

[4] Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*. Dec. 2021. URL: https://github.com/CSSEGISandData/COVID-19.

[5] Worldometer. *COVID-19 CORONAVIRUS PANDEMIC*. Dec. 2021. URL: https://www.worldometers.info/coronavirus/.