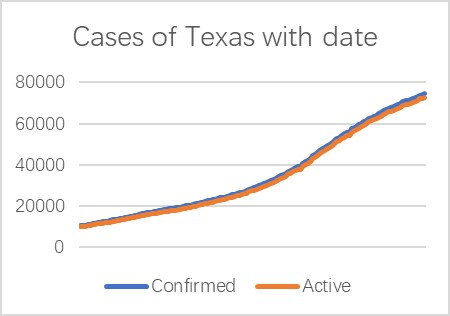
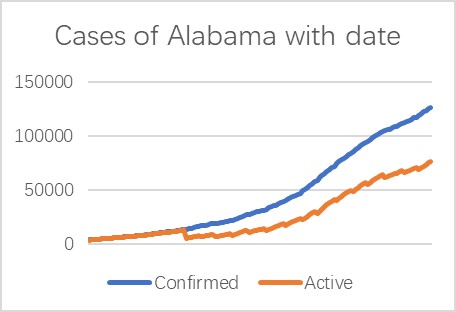
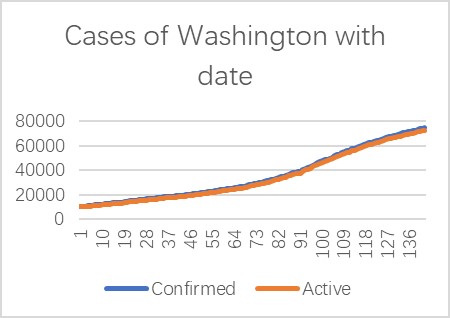
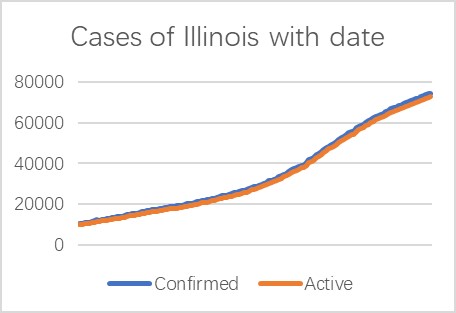
Midterm Report Raw

Data preprocessing:

Before processing the data, we checked how the data looked, and we found a lot of blank data. We filled all these blanks with 0, because we could not train the model with missing values. Additionally, if an attribute has too many missing values, the model will automatically assign a small weight to it to decrease its significance, so it turns out to be fine by replacing missing values with 0.

Also, we had a glimpse on the relationship between cases of each state with related attributes (especially the date) to get a sense of which model may be appropriate for the dataset.





In addition to a crude lookup on the trends, we also noticed that some attributes are dependent on each other. For example, the “hospitalization rate” is simply “People\_Hospitalized” divided by “Confirmed\_cases”. Therefore, we only need to put two of “hospitalization rate”, “people\_hospitalized”, and “confirmed\_cases” into our model. It is also true for other “rates”. Therefore, we omitted some features when training the model.

Smoothing?

Besides smoothing, we also project each feature into a polynomial degree (in our model it’s 3) to enable us to find a polynomial relationship between cases and pertinent attributes. It will be introduced in detail in the following parts.

=============================================================================(Designed and Tested Model)

*LSTM*

Our LSTM models are trained with 3 features(date, confirmed, and deaths number), and those data are separated by state. We tried different numbers and sizes of hidden layers, and temporarily decided to have two hidden LSTM layers each with 100 and 50 units, with a dense layer for generating output, and we set the time seq to 50 when preparing the data.

*SIR*

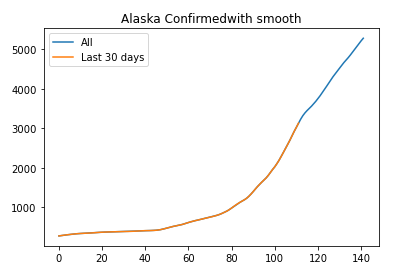
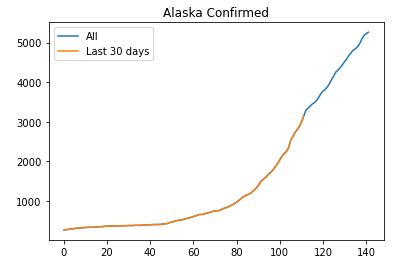
In addition, we explored the SIR model. Under the assumption that the population remains unchanged, the SIR model divides the population into three categories: susceptible,infected and removed.This model is a set of differential equations that conveys how the numbers of the three groups change over time.The main problem with this model is we need to solve the parameters, like beta and gamma that represent infectious rate and recovery rate, directly from data. An existing way is to select a window of time and calculate betas and gammas for each day of that period. Using the mean values of all the betas or gammas as the parameter for prediction. However, it’s hard to find a proper window for getting the optimal parameters and also hard to ensure the sensitivity and stability in prediction.

*Exponential Moving Average (EMA)*

We also examined classic methods of time-series modeling, the auto-regressive and moving-average classes of functions. We chose the exponential moving average (EMA) as it has an exponentiated weight put towards the latest data observed/fed-in to the model within a given window size determined by its weight parameter (alpha). This helps with smoothing out the data observed and use mostly the latest information which is more helpful in determining the upcoming data while taking a kind of weighted average with respect to the previous days. Due to the simplicity and intuitiveness of the model, we first implemented this model as a baseline model to test our performances, as it only uses the predictor variables themselves for predicting their latter changes. We thus tried fitting the Daily Cases (Confirmed) and Deaths by setting different values of alpha parameter that controls the amount of days considered and observed a fair performance for data that were more linear with less drastic extreme changes. For other data, it has more delay in learning the correct changes and often overshoot/undershoot for new predictions given a recently observed sudden surge/drop.

Linear and Polynomial Models

The best result of our team so far comes from the linear model with L1 and L2 normalization (Lasso and Ridge). By observing the plots of ‘confirmed cases’ and ‘deaths’ of each state, we found that the trend of the curves remain the same in the last month. Moreover, the variation of gradient is small, some of the curves even remain a straight line. Therefore, we propose that only the last few data should be retained, and the Linear model and Polynomial model might work well. Since the size of training data decrease to a very small number, we conducted the smoothing on data and applied Lasso and Ridge to relieve the overfitting issue. As a starting experiment, we apply the same setting (data, smoothing window size, hyperparameter of Lasso and Ridge) to all states. The best results of Linear model and polynomial are close (2.28 for linear. 2.30 for polynomial). So far, the linear model is slightly better than the polynomial model. Intuitively, we believe the polynomial model should fit the data better since the curves are not all straight lines. This happened because we haven’t tuned the parameters state by state. One of the problems with polynomial models with degree greater than 2 is that they are likely to drastically drop down which is also the reason why we smoothed the data. On the other hand, LSTM can give us more stable prediction curves even though its trend may be wrong. In the future work, we are going to make a trade off between them and figure out how to solve their problems.

=============================================================================

Conclusions and Findings

The best result we have achieved is a test MAPE of 2.27630 using a linear regression with lasso and ridge regularization and varying window sizes (approximately less than 10 days for all) hand tuned for each state.

For all of the models we have attempted, restricting a tighter window size would always lead to better performance on the test set, which fit our assumption as recent events (e.g. lockdown measures placed/lifted, increased state testing due to new site openings, etc.) might elevate/lower the daily changes of our target data (number of confirmed cases and number of deaths), in which taking account of past information would only prevent the model from learning and adapting quickly for recent predictions. However, our current worry is that this might be overfitting for the short-term predictions (e.g. predicting the next 10 days) and would affect long-term predictions (predicting 10 days of data in the next month).

Moreover, we observed that the distribution of our data varies quite drastically between each state, not only in their slopes but also second derivatives that marked how their slopes changed and represented a measure of stability of data (the more stable the data, the closer their second derivative to zero). And oftentimes, the drastic changes that occurred in one state would not correspond to that of another while states that had similar trends at first would not often continue to exhibit similar trends, which posed significant challenges for our model to provide excellent predictions across-the-board without careful, extensive hand-tuning for each state.

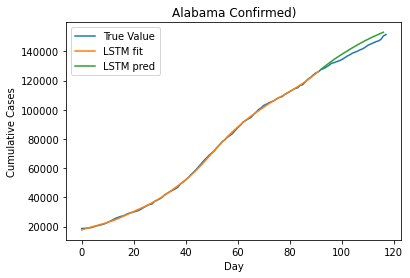
* Mobility (data) more exploration (Jason)

=============================================================================

(Analysis of current models & techniques)

So far, the best MAPE we got is 2.27630. Compared to our previous results, it has been improved a lot. This result was achieved by using LSTM model and fitting linear model to data.The bottleneck in this case is around 2.There could be a chance of getting better results by using polynomial.For getting such better performance, we did data smoothing before training. This method was effective since it reduced the data bias and increased the precision of the data without changing the tendency of data.

LSTM has two major issues. The first one is that it is hard to stabilize due to the lack of features and data categorized by states. As we tested our model independently for each state, we found many states have unstable curves, which are hard to be predicted by our LSTM model. However, we also found that some states have great performance. For example, Alabama and California tend to have very accurate predictions on the confirm and deaths number, as we can see that the prediction curve is well fitted to the true value both for the training set and the validation set. The second issue is that our LSTM seems to simply regenerate the shape of the previous curve when predicting the future trends, which is not reasonable for the curve of confirm and deaths number. It is probably because our LSTM cannot ideally generalize high dimensional features. So, we decided to reconsider the implementation of LSTM model and use or incorporate other models as our current priority.



=============================================================================

Timeline of future plans

Given the success we observed from our linear model, we set out to improve its results by trying out different polynomial transformations and possible combinations of other features to predict an intermediate value (e.g. Death rate) of which we can later transform back to the two predictor variables we are interested in (Deaths and Confirmed). We have also researched and seen promising results achieved using Facebook's Prophet package (which uses both linear and logistic regression under the hood) with flexibility of encoding seasonality and special dates to improve our prediction.

We also think that incorporating the mobility data of each data would help the model to learn by using for instance, the net change of population dynamics (influx - outflux) or a measure of mobility rate for each state after careful processing and transformations. Naturally, the more mobility within a state that has a steep increase of confirmed cases should only face more new confirmed cases and deaths (with delay) and vice versa. We also hope to specify and train for time-varying window sizes of each state at a given time using a more automatic method of selection to improve its learning (e.g. to use more days of previous data to smoothen one sudden change but shrink the window size for faster learning for continuous days of spike).