COVID-19 Modeling in US States Using LSTM Neural Networks

By Josh McGuckin, Neil Patel, Daniel Gallegaur, Robert Saporito

Background & Problem:

Since the first recorded case of the COVID-19 disease was reported in Wuhan, China on December 31st, 2019, the infection has spread rapidly throughout the global population. On March 11th, the number of cases had reached 118,000 and had spread to over 100 countries worldwide, at which point the WHO declared this disease to be a global pandemic. As a result of this novel virus' rapid spreading ability, deadly nature, and unpredictability, health experts and community leaders around the world called for strict social distancing guidelines and quarantine measures to attempt to halt the spread of the disease and limit its damage. However, by April 1st, the number of total cases reached to 823,626, and as of the day of this writing (6/03/2020) that number ballooned to 6,295,328, resulting in 380,580 deaths and 2,742,306 recoveries (Dandekar & Barbastathis, 2020; Pal et al., 2020).

To gain deeper insight into the trajectory of this disease researchers have been looking closely at infection and recovery data. These numbers can help public health specialists understand how communities need to continue responding to limit the spread of the virus and the impact this virus is having on communities. It is also important to quantify the effect of distancing and quarantine measures so that better predictions can be made to help put an end to this pandemic. Previously, researchers have used epidemiological models, such as SIR and SEIR to understand the spread of SARS/MERS infections (Dandekar & Barbastathis, 2020). These focus on deriving infection and recover rates for susceptible populations. However, these can be thought of as blind reproduction models because they do not consider real-time data. Thus, they cannot accurately model how diseases are actually spreading through populations, and only tell us the possible trajectory based on underlying infection characteristics. Novel machine learning and neural network techniques have been proposed to tackle this challenge.

In a recent paper Dandekar & Barbastathis (2020) set out to determine how quarantine measures can affect the reproduction number of the SARS-Cov-2 virus (R_t) – a measure of how fast a virus is spreading through a population. The goal was to improve upon previously developed epidemiology models using a shallow neural network to study real-time infection rate data from the US, South Korea, Italy, and Wuhan, China. The authors present two such models, SEIR (developed and used 2006-2019), which is divided into a set of related time-dependent differential equations: S = susceptible; E = exposed; I = Infected; R = recovered, and SIR, which leaves out the exposed variable – assuming direct transition from susceptible to infected. Without taking into account quarantine measures, these models found that the recession in infections in Wuhan (by late February) could not have been accounted for by normal infection rate statistics. This signaled that exponential growth would have been likely to continue if no interventions were put in place. Next, a neural network model was implemented to produce an accurate model of the actual infection spread, by quantifying the effect of quarantine controls on reproduction number. A nonlinear function approximator neural network was used to derive a quarantine strength term (Q), creating the novel SIRT model ($T = Q \cdot I$; total quarantined population). The network was built as a 2-layer feed-forward network, using 10 hidden units and ReLu activation functions. The model described in this current paper proposes a similar assessment of underlying infection data using current infection counts and other epidemiological data in conjunction with mobility data. In our study, we propose a prediction model using various state-wide features using an LSTM deep neural network model.

From the SIRT model, results showed that slower and more relaxed policies in the US, along with its more complicated and diverse geography, led to the US having the lowest quarantine strength. However, the growth rate was estimated to be decreasing and approaching $R_t = 1$. Forecasts showed that continued measures would produce a negative growth rate by April 20th, with a peak of 600,000

infected before stagnation. A mixed Q(t) model showed that stronger control measures, modeled after the other regions, could have arrested the virus even faster. Alternatively, if quarantine measures were to be relaxed over the same period of 17 days (from April 1st – April 17th), stagnation would have been prevented and the case count could have topped 1 million. One of the main weaknesses of this model was the heterogeneity of the different populations being studied, which make it difficult to correlate the various results. One area for further development of this model could be to account for state-to-state infection rate variance in the United States as a result of population density and asymmetric quarantine policies. This would allow for a more accurate and comprehensive model on the effectiveness of the US quarantine as a whole because it is heavily influenced by the quarantine protocols of individual states. Such insight would allow state legislators to make informed decisions on quarantine policies to accelerate the infection plateau for the country. It would also provide insight as to how infection rates for the entire country will change in response to key states opening their borders prematurely.

In another paper by Pal et al. (2020), a neural network model is also presented to predict the growth of COVID-19 based on temporal data and an LSTM model. The objective was to rank risk categorizations of 170 countries based on key features that affect the growth of the SARS-Cov-2 virus. Additionally, the team was able to develop and test multiple models to understand which structures resulted in the best performance. The group presents how AI and machine learning algorithms provide for novel solutions to problems that involve prediction or classification. The temporal neural network model of a recurrent neural network (RNN) was chosen for several reasons. One of the key reasons for this was the fact that RNNs have feature memory which allowed Pal and colleagues to compose a model from a large collection of time-dependent epidemiological data. By extracting underlying features from the large temporal dataset, the team aimed to find key features that affect COVID-19 infection growth for certain countries. They implemented fuzzy logic at the end of their LSTM (Long Short-Term Memory) model to predict short term risk for specific countries.

From the LSMT model, in consideration of ozone, humidity, dew, and temperature data, alongside infection data, a 78% accuracy was achieved for the number of infections 10 days into the future (post training). However, some weaknesses with this model include the small sample size of data (> 90 days), a short time window of prediction (10 days), which would be difficult to effectively act on in our societies, and the isolation of country populations that were modelled, which fails to take into account inter-population spread. This LSTM network that the authors developed shows great potential to be expanded into other areas, such as future pandemics, alternate datasets, and economic effects and indicators. In the future, consideration could be given to flight, traveler, and business data, allowing the model to respond more directly to quarantine and social isolation measures and the impact that highly connected countries have on each other.

In summary, a large portion of the globe has been forced to go into quarantine to stop the spread of COVID-19, the disease caused by the SARS-CoV-2 virus. In order to assess the effectiveness of mitigation efforts and need for continued preventative measures, we must build robust epidemiological models that can accurately analyze and forecast infection and recovery rates. Doing so will help to determine the best path towards recovery from this pandemic. To achieve this goal, we attempt to build a time-dependent LSTM network in the vein of epidemiological modelling, using related infection and demographic data, to understand how different conditions are having an affect on the spread of COVID-19 in various state-based populations in the US.

COVID-19 Modeling in US States Using LSTM Neural Networks

By Josh McGuckin, Neil Patel, Daniel Gallegaur, Robert Saporito

Datasets & Methods:

The dataset used to train the model consists of many features which may impact the pandemic spread. Many of the features of the data required multiple steps of data preprocessing and involved careful dimensionality reshaping. The two major databases used were the Johns Hopkins Dataset and the joint-effort Google and Apple Mobility dataset. Both databases are updated on a regular basis keeping accurate measurements during the lifetime of the pandemic outbreak.The databases were sampled to find the ideal range of the dates which would include the most amount of quantifiable data without any breaks or gaps. Not having breaks in each time series dataset is vital for the forecasting to train properly. Combining two databases containing different fields, required aligning of dates to properly obtain a combinational database which could then be reformed into time series data. The total data consists of most importantly, number of positive cases, number of negative cases, population hospitalized due to the virus, mobility trend patterns in transportation, and in some models the average temperatures and humidity levels. All the data obtained was on a daily basis and the times series forecasting also occurs on the scale of days. The mobility data consists of multiple fields of percent deviations from the mean with a certain baseline (0%). This data needed to be converted to integer type and increased to have only positive numbers being input into the model. This was done by adding a value of 100 to each mobility datapoint and rounding the result to the nearest integer. Examples of mobility data consist of mobility driving, mobility in parks, mobility in workplaces, mobility in transit, etc. The John Hopkins data provided a reputable data source that maps the numbers of the cases in the United States on a state by state basis. Out of this multitude of data, three states data was extracted (NY, NJ and TX) each having a different range of dates, due to discrepancies in the data. Thus each state model was trained on a different dataset ranging from different dates. Because NY and NJ are the two most heavily infected US states and TX is currently one of the least infected states, all three states were modeled to compare prediction accuracy between both extreme and mildly COVID-19 afflicted states.

In order to train the model properly to obtain the best accuracy, a large dataset was chosen and the data was segmented into a window size. Multiple scaling and dimsonality shaping methods were used to fit the data to the model. In some very small cases, the lack of data for the early days within the dataset was replaced by values of zero, as shaping with complete data is essential for the model to train properly. Before feeding into the models, the data was downscaled to a range of 0 to 1, for faster computational performance. One final important feature to note about the data was that the order of the data was always conserved throughout the data preprocessing and model training. This is essential for the temporal correlation between the data to conserved for forecast based on history,

Implementation:

After scaling of the datasets between 0 to 1 a window size was chosen that would determine the number of preceding rows, which represented individual days in the time series dataset, that would be input into the model for each predicted output row of parameter values. This was implemented by a function which removed the most recent time series data, the number of rows designated by the window size, from the dataset array and reshaped this new input data array (X) into the appropriate 3D format for model fitting (# rows, window size, # parameters). The same function also generated the expected output (Y) from the 3D input data array (X) by

assigning this expected output (Y) equal to the following row in the original dataset outside the window, as depicted below.

Y[window, :] = dataset[window+1, :]

This would ensure that every expected row output (Y) from the model was based off of a window of preceding rows in the original dataset. This allowed for accurate prediction of all parameter values for future time points based off of previous time series data. After generating the scaled input (X) and expected output (Y) arrays these were both separated to form the training and testing datasets. The training dataset was designated as the first 80% of rows from both scaled arrays (X, Y), while the remaining 20% of rows (i.e. the more recent time series data) were designated as the testing dataset which would be used for prediction evaluation.

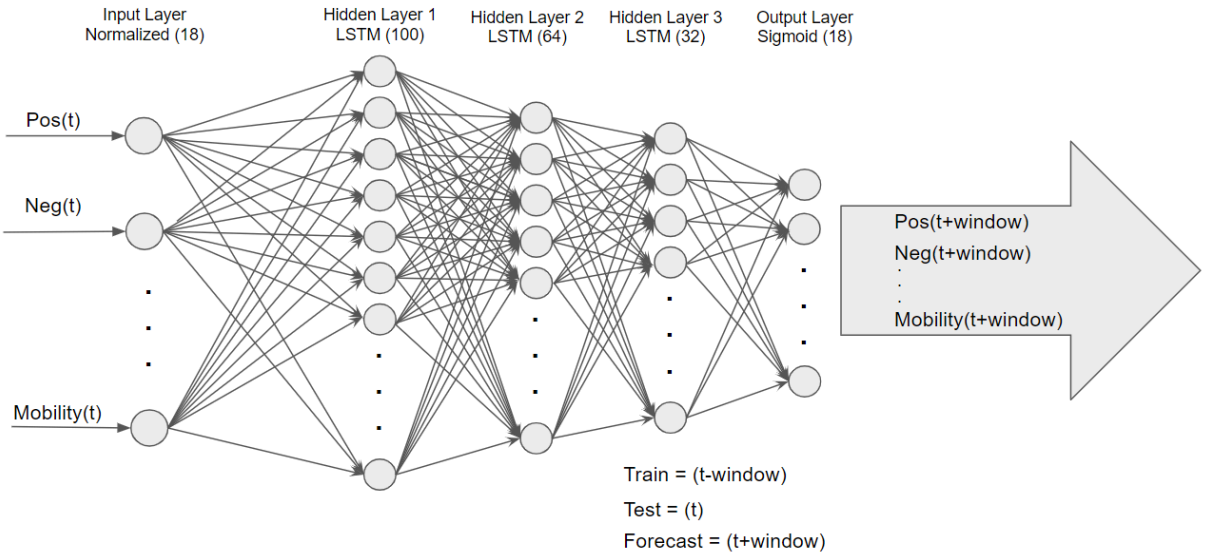


Figure 1. Architecture of LSTM COVID-19 State Models

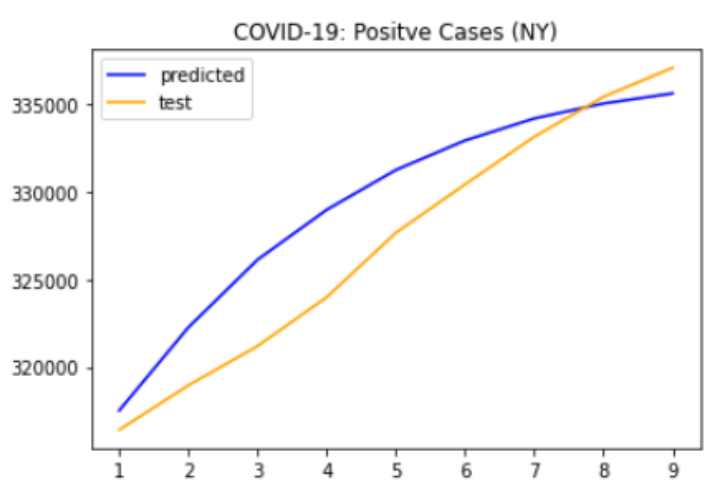
After this data preprocessing procedure a 4-layer LSTM recurrent neural network (RNN) was built using keras (architecture shown in 1) and trained with the processed data arrays. This procedure was followed for all 3 state datasets and resulted in state-specific LSTM models for COVID-19 forecasting (NY, NJ, TX).

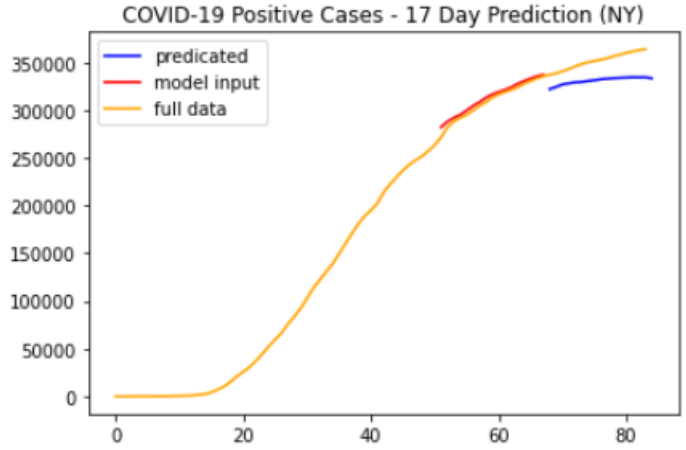
Once trained, these state-specific models were used to make predictions of key parameter trends (positive cases, recoveries, deaths, hospitalizations, etc.) for future timepoints. This was achieved by a for-loop which fed the last window of each state's preprocessed dataset as the input to that state's trained model, adding the resulting prediction (1 row per iteration) as the most recent time point in the original state dataset, removing the earliest time point (row) from the dataset, and re-iterating through the for loop. This row shifting method was found to be necessary as a result of the dimensionality reshaping that needed to be performed for model predictions. The number of iterations through this for-loop was the desired timespan for forecasting. This forecasting timespan was set to the window size of each state model so that it could be compared against similar timespan predictions on the train and test arrays. This would

allow for more accurate evaluation of these forecasts by comparing the prediction accuracy of these models to unseen test and train data with similar timespans.

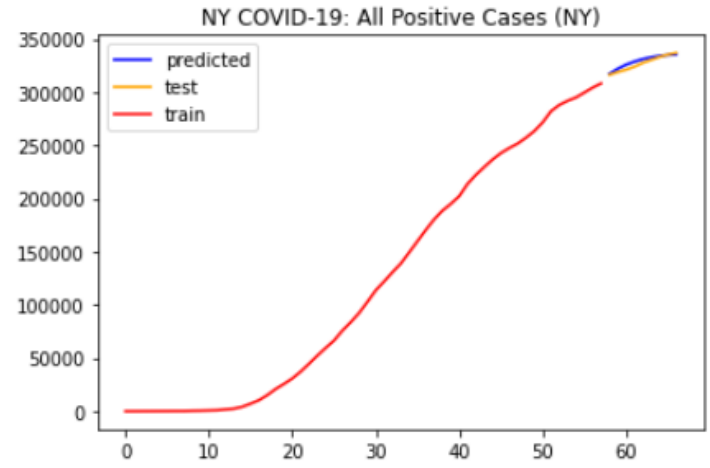
Results:

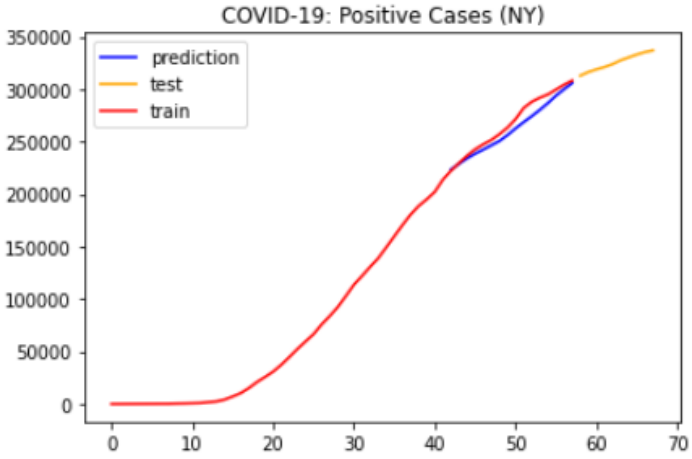
The results of the future and evaluated predictions are shown below in Figures 2-7. The top right of each figure shows forecasted trends for the parameters (positive and recovered cases shown in these figures) described over the window size for each state's model. This was done for accurate comparison to evaluated testing and training predictions that use these window sizes. The results of the evaluation show good accuracy for each model with the least accurate being for recovered cases (20% error) and the most being for positive cases (0.076% error) both in the NJ state model.






```
Test Percentage Error: 5.624
Test RMS Error: 20542.765
```






```
Test Percentage Error: 1.461
Test RMS Error: 5303.632
```

```
Test Percentage Error: 2.299
Test RMS Error: 7423.679
```

Figure 2: Positive cases of COVID-19 as predicted by the NY state model. Figures 2a (top left) and 2b (bottom left) shows the prediction accuracy of the model with a window size of 17 days, compared to the actual number of reported positive cases from the test array. Figure 2c (top right) shows the forecasted trend of positive cases as output by the NY state model over the

window size which is the ultimate goal. Figure 2d (bottom right) shows the accuracy of the model prediction when fit to data that it was already trained to from the train array.

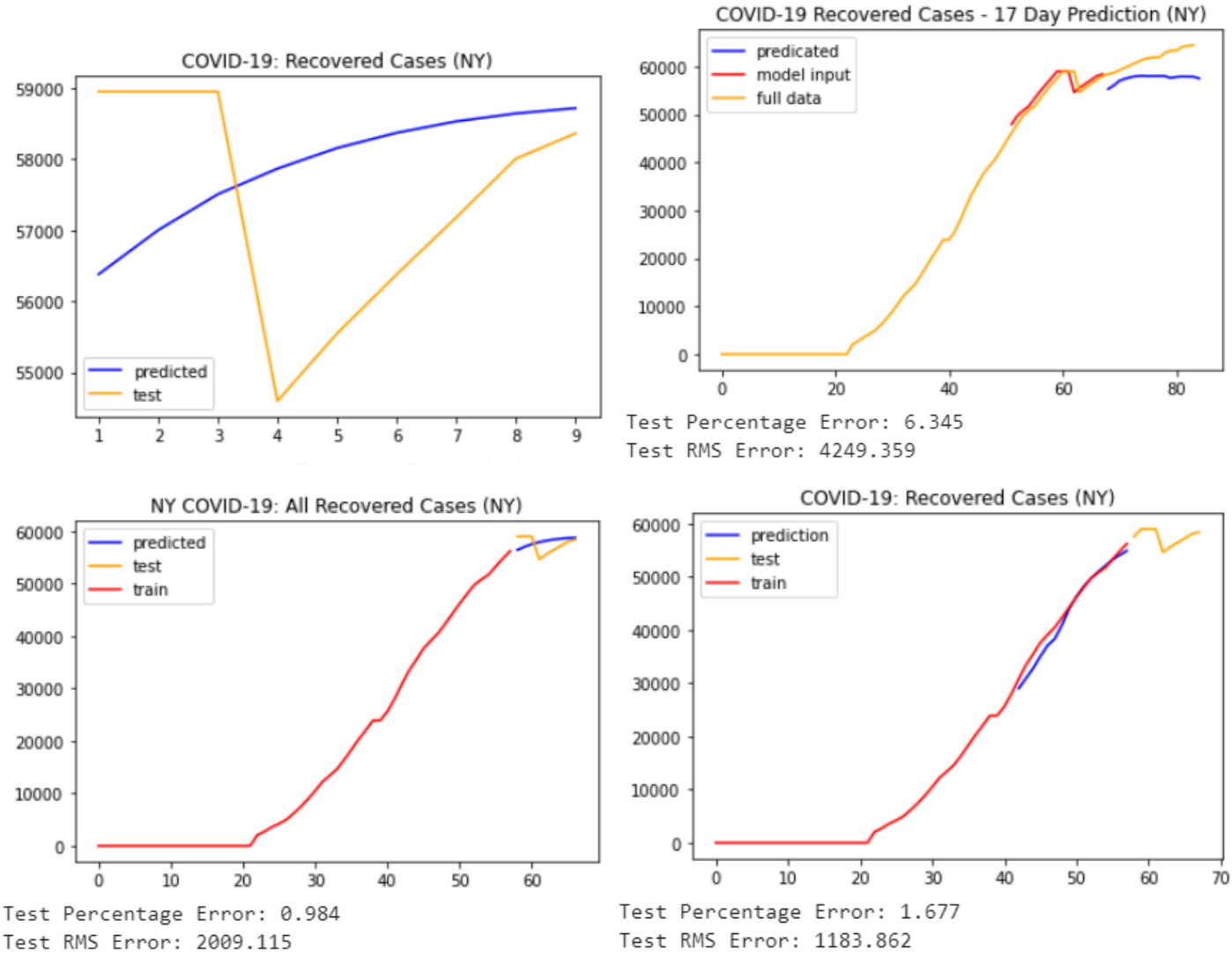


Figure 3: Recovered cases of COVID-19 as predicted by the NY state model. Figures 3a (top left) and 3b (bottom left) shows the prediction accuracy of the model with a window size of 17 days, compared to the actual number of reported recovered cases from the test array. Figure 3c (top right) shows the forecasted trend of recovered cases as output by the NY state model over the window size which is the ultimate goal. Figure 3d (bottom right) shows the accuracy of the model prediction when fit to data that it was already trained to from the train array.

COVID-19 Modeling in US States Using LSTM Neural Networks

By Josh McGuckin, Neil Patel, Daniel Gallegaur, Robert Saporito

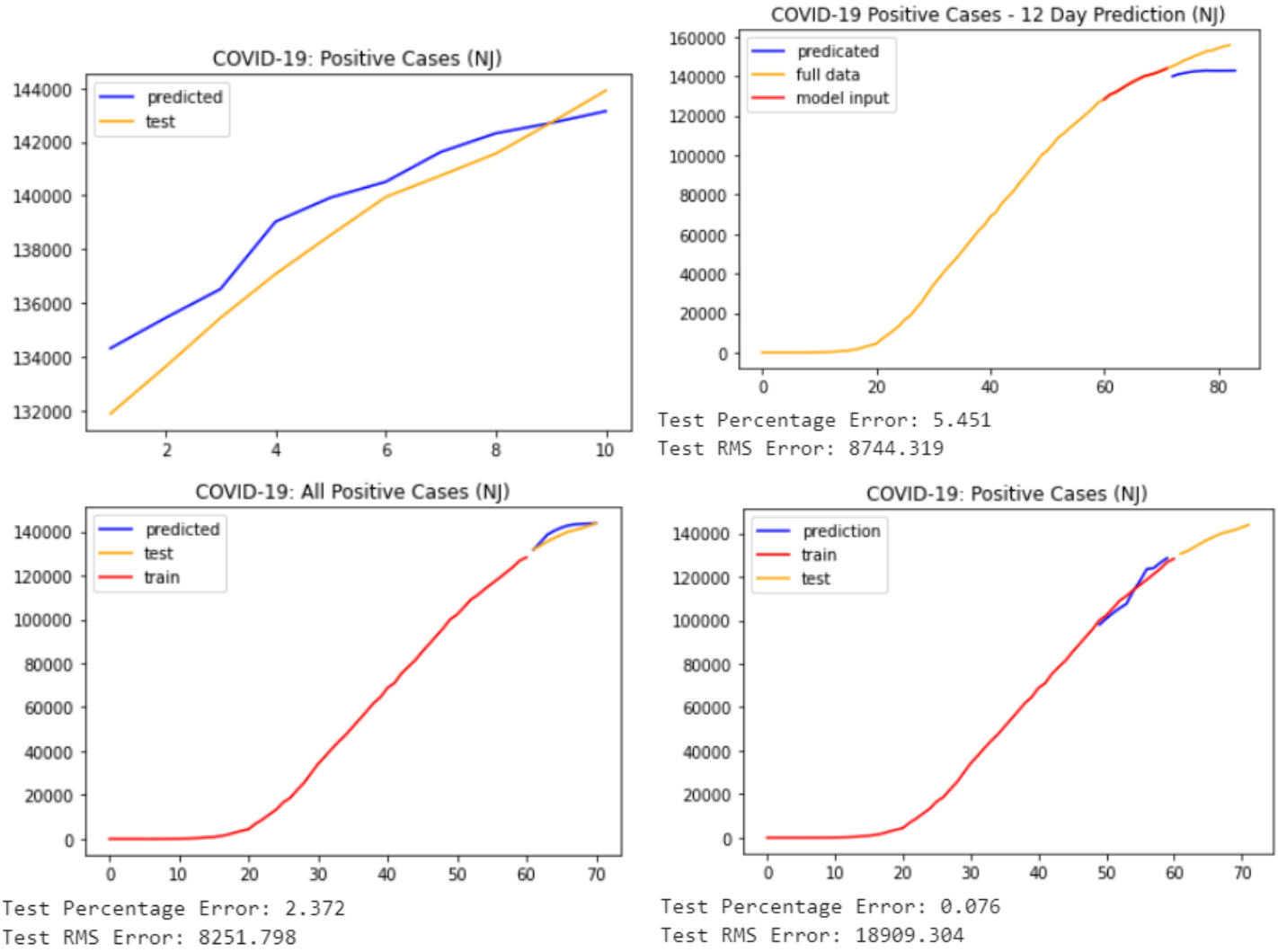
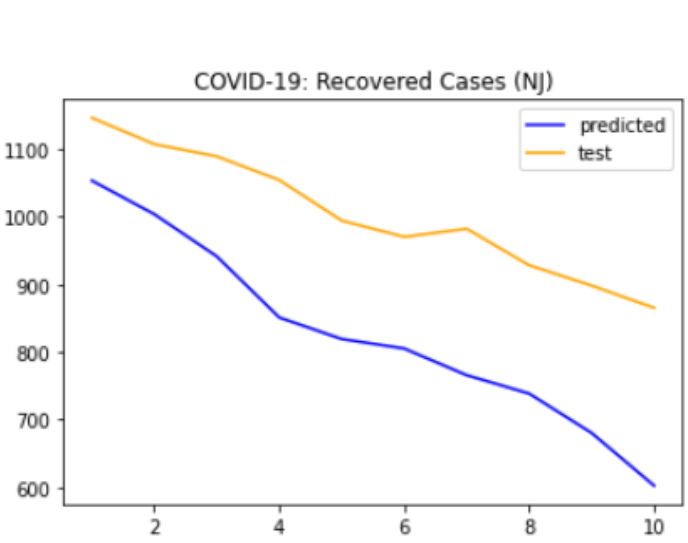
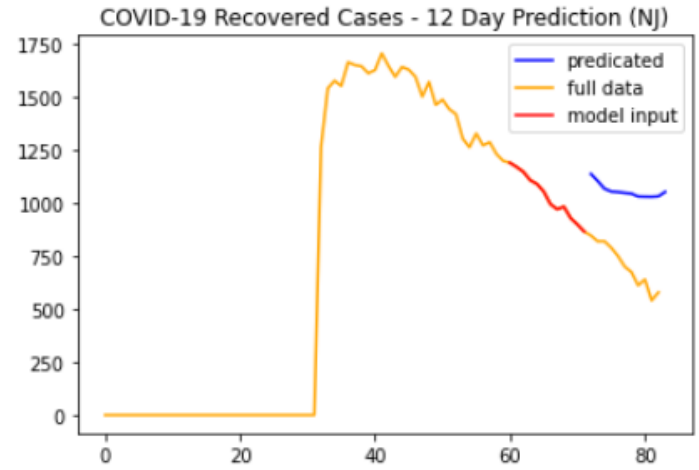


Figure 4: Positive cases of COVID-19 as predicted by the NJ state model. Figures 4a (top left) and 4b (bottom left) shows the prediction accuracy of the model with a window size of 12 days, compared to the actual number of reported positive cases from the test array. Figure 4c (top right) shows the forecasted trend of positive cases as output by the NJ state model over the window size which is the ultimate goal. Figure 4d (bottom right) shows the accuracy of the model prediction when fit to data that it was already trained to from the train array.

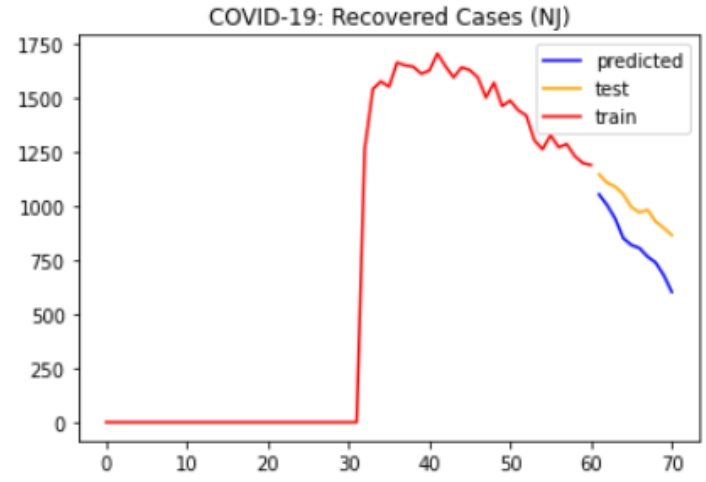
COVID-19 Modeling in US States Using LSTM Neural Networks

By Josh McGuckin, Neil Patel, Daniel Gallegaur, Robert Saporito

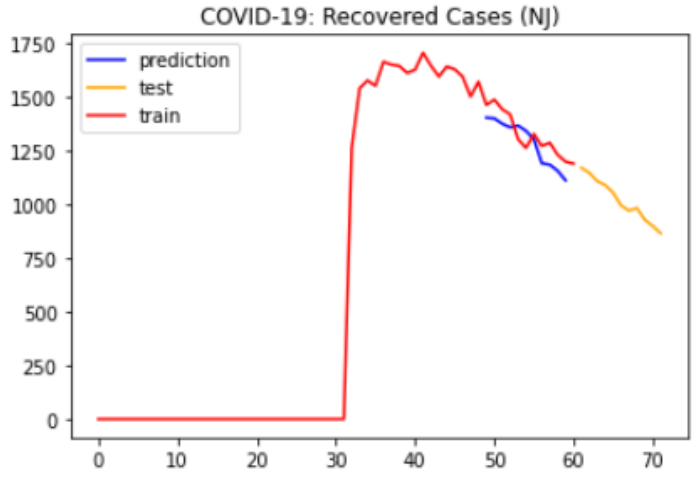




Test Percentage Error: 46.853
Test RMS Error: 347.701

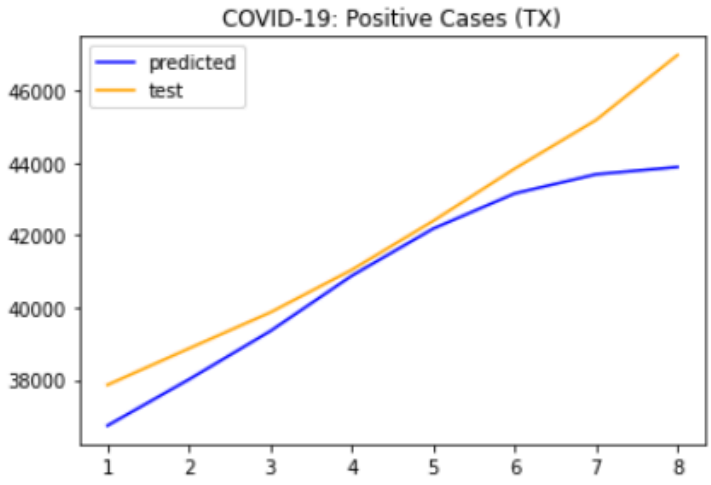


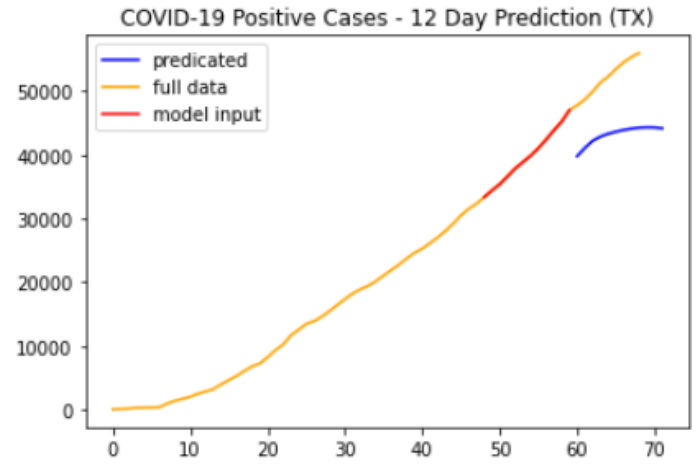
Test Percentage Error: 20.107
Test RMS Error: 302.849



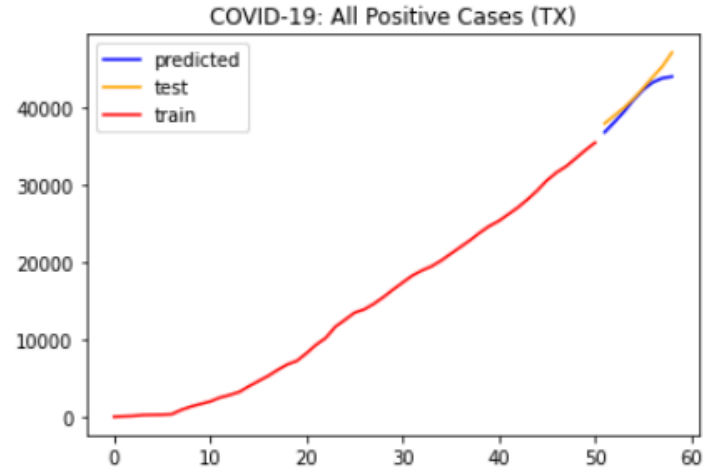
Test Percentage Error: 3.413
Test RMS Error: 201.229

Figure 5: Recovered cases of COVID-19 as predicted by the NJ state model. Figures 5a (top left) and 5b (bottom left) shows the prediction accuracy of the model with a window size of 12 days, compared to the actual number of reported recovered cases from the test array. Figure 5c (top right) shows the forecasted trend of recovered cases as output by the NJ state model over the window size which is the ultimate goal. Figure 5d (bottom right) shows the accuracy of the model prediction when fit to data that it was already trained to from the train array.

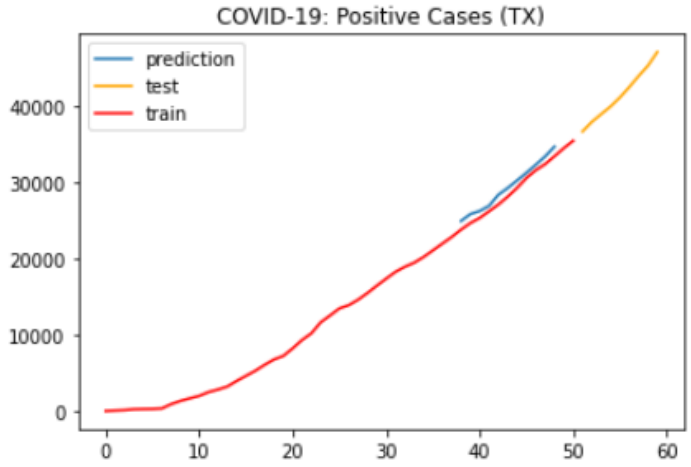




Test Percentage Error: 8.814
Test RMS Error: 4661.730



Test Percentage Error: 0.680
Test RMS Error: 5315.538



Test Percentage Error: 7.143
Test RMS Error: 6449.003

Figure 6: Positive cases of COVID-19 as predicted by the TX state model. Figures 6a (top left) and 6b (bottom left) shows the prediction accuracy of the model with a window size of 12 days, compared to the actual number of reported positive cases from the test array. Figure 6c (top right) shows the forecasted trend of positive cases as output by the TX state model over the window size which is the ultimate goal. Figure 6d (bottom right) shows the accuracy of the model prediction when fit to data that it was already trained to from the train array.

COVID-19 Modeling in US States Using LSTM Neural Networks

By Josh McGuckin, Neil Patel, Daniel Gallegaur, Robert Saporito

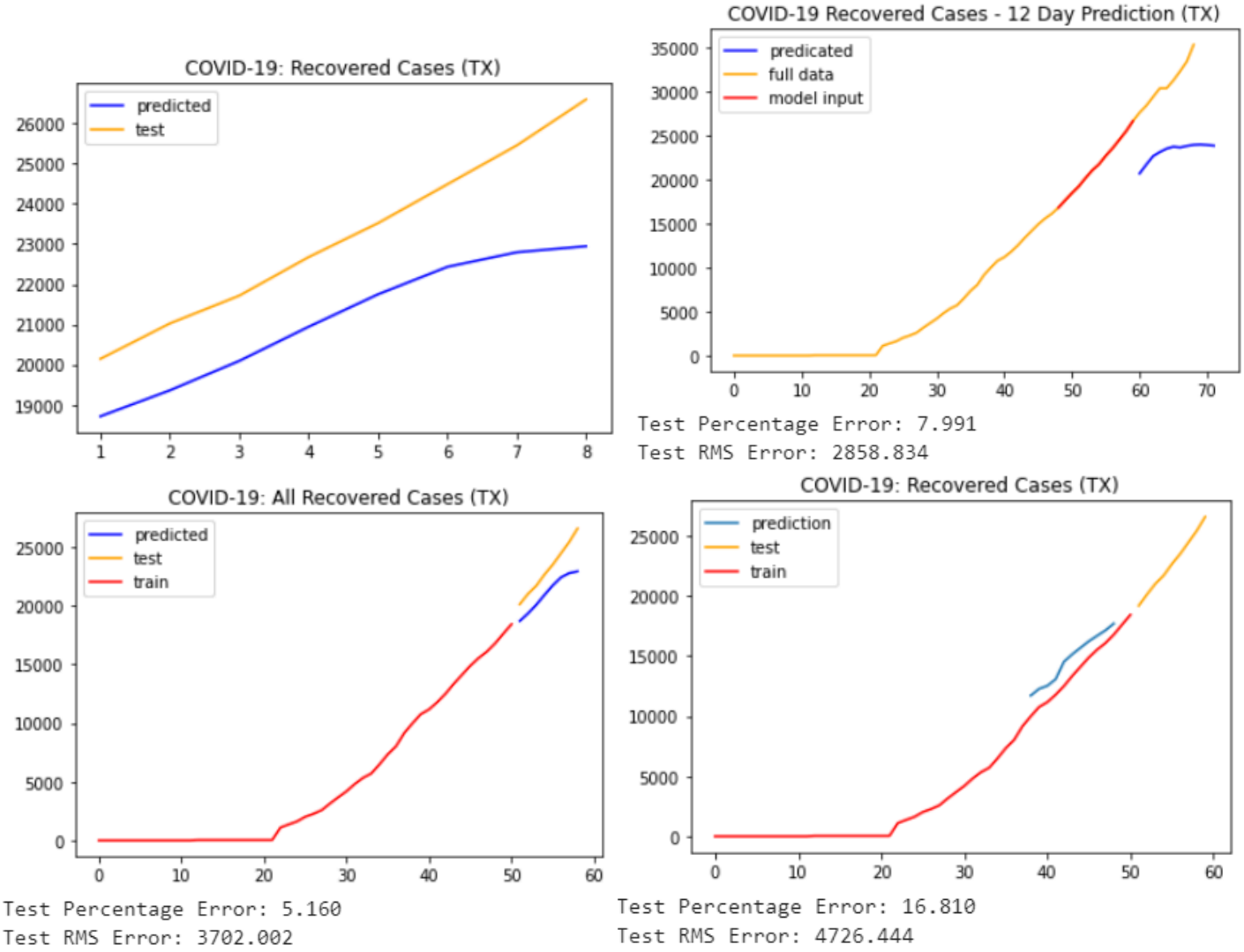


Figure 7: Recovered cases of COVID-19 as predicted by the TX state model. Figures 7a (top left) and 7b (bottom left) shows the prediction accuracy of the model with a window size of 12 days, compared to the actual number of reported recovered cases from the test array. Figure 7c (top right) shows the forecasted trend of recovered cases as output by the TX state model over the window size which is the ultimate goal. Figure 7d (bottom right) shows the accuracy of the model prediction when fit to data that it was already trained to from the train array.

Performance Evaluation:

Unlike classification models, in which the model either predicts the correct class or does not, a forecasting model is unlikely to predict the exact expected value. This means that a forecasting model requires a method of evaluation that takes into account how closely the predicted data reflects reality. There are many options for this, but for this model the root mean squared forecast (RMSE) and percentage errors were used. The RMSE metric was chosen because the normal mean squared error makes the forecasting errors positive and gives larger weights to errors that are further away from the actual value, while also providing absolute error representation. Taking the square roots gives a metric in which the units are the same as those of

the data that is being predicted. In the COVID-19 dataset that was used in this model, the values of parameters of interest are frequently in the thousands. This results in a situation where the RMSE can be in the thousands, but the model is still making predictions that very closely resemble the expected data. This is what prompted the use of the mean percentage error which shows the relative difference between the predicted values and the actual values from the validation data.

During validation it was observed that the accuracy of the model depended very heavily on the quality of the data. In the rapidly unfolding quarantine situation, the standards and quality in data acquisition varied from state to state and in some cases from one parameter to another. In ideal cases, such as in the case of predictions of deaths in New Jersey, the model performed with an error of less than 1% compared to test arrays. In the worst cases, such as the predictions of hospitalizations in New Jersey, the error exceeded 25%. The poorer performance is likely due to a lack of hospitalization data, which is represented as zeros, in the earlier days of the dataset followed by a very steep slope once data was accurately recorded. These sudden changes are not reflective of the true underlying pattern and therefore distort the function approximation. It is also possible that certain features are inherently harder to predict or rely on some other feature that was not accounted for in this experiment. For example, hospitalization predictions were generally less accurate than predictions of positive cases, negative cases, and deaths in all 3 states.

Future Experiments:

As it is, the proposed model would be useful for predicting trends for future developments in the continued spread of COVID-19 which could be of use to public health agencies and policy makers. Additionally the methods developed in this paper such as correlating epidemiological data with climate and population mobility could be used to produce accurate models of other pandemics that may emerge in the future. These methods should be improved upon before such application, specifically the data preprocessing which could improve model accuracy when early time points of key parameters are sparse. Model accuracy might also be improved by collecting other data which might be related to how certain features evolve, such as the age of those hospitalized, number of COVID-19 tests conducted per day, or air travel in and out of the state.

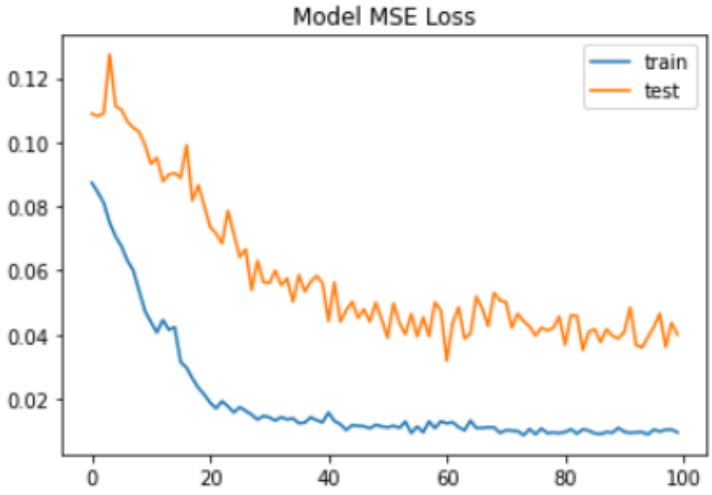
In the immediate future predictions could be made about the remaining US states, and the existing predictions could be extended using more recent data. This may be particularly important as states begin to lift quarantine measures and individuals begin gathering again. This allows for the possibility of using these models for the investigation of a second wave of infections occurring in the relatively near future. Such investigations could be modeled by a controlled increase of mobility in each state model to simulate the lifting of quarantine-induced restrictions and observing the predicted change in positive cases. Advanced warning of a second wave would be helpful in preparing mitigation methods. Additionally comparing the predictions between states would be helpful in identifying which state's quarantine methods have been the most effective thus far.

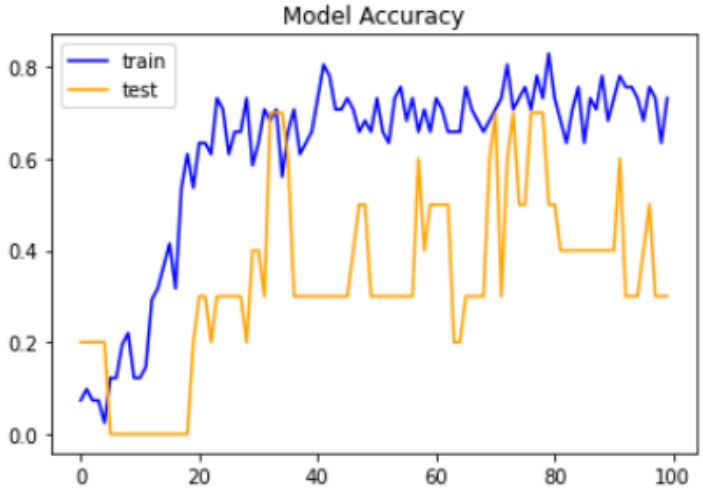
COVID-19 Modeling in US States Using LSTM Neural Networks

By Josh McGuckin, Neil Patel, Daniel Gallegaur, Robert Saporito

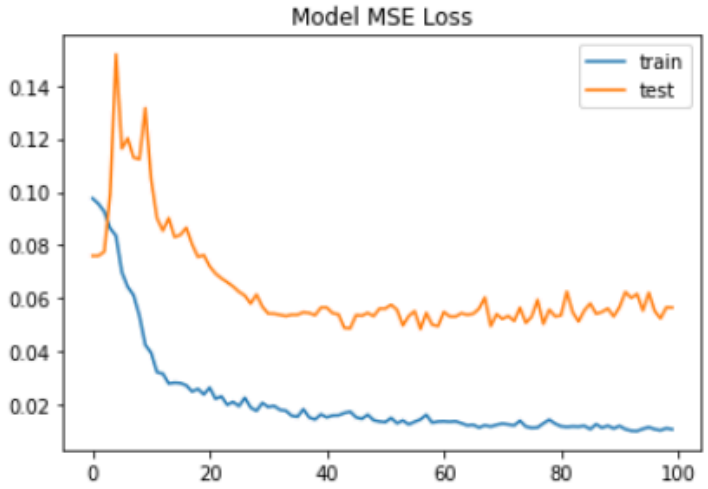
Appendix A: Model Loss and Accuracy

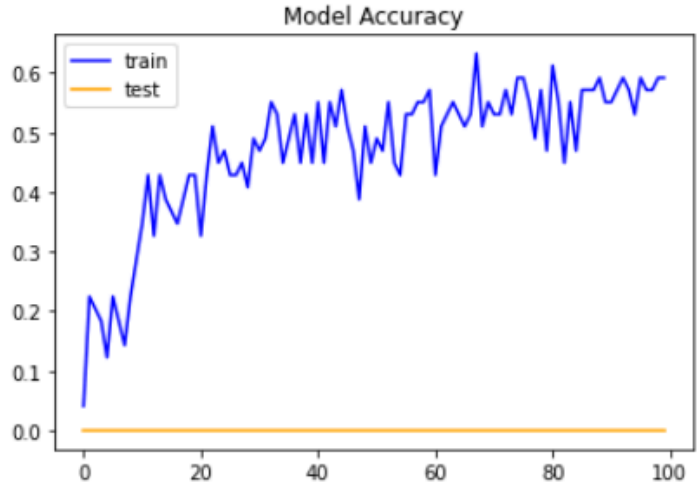
NY State Model:



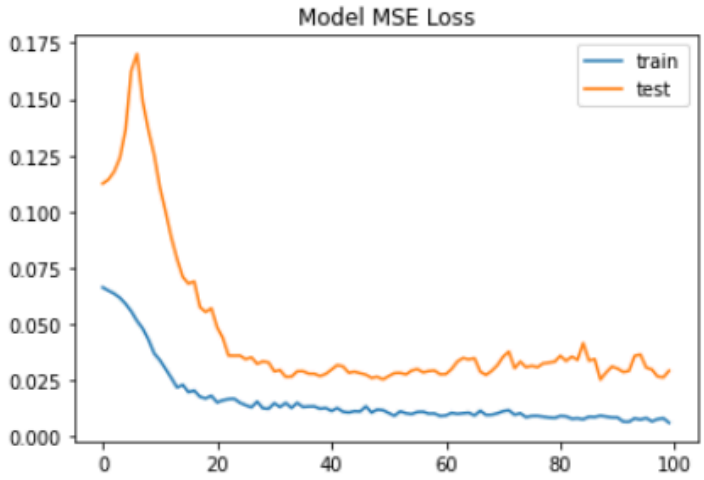


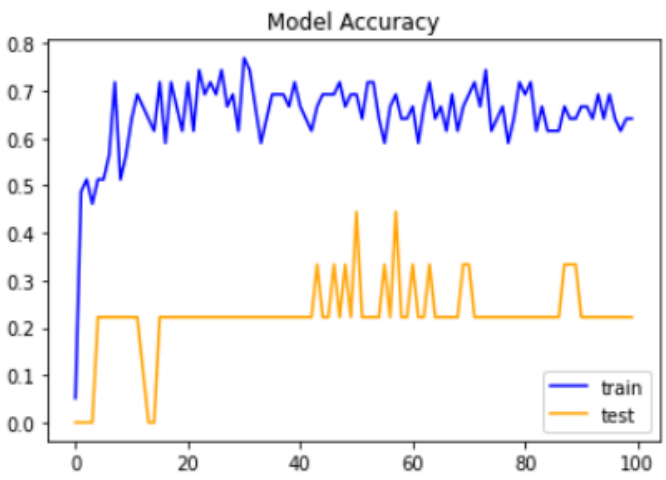
NJ State Model:





TX State Model:

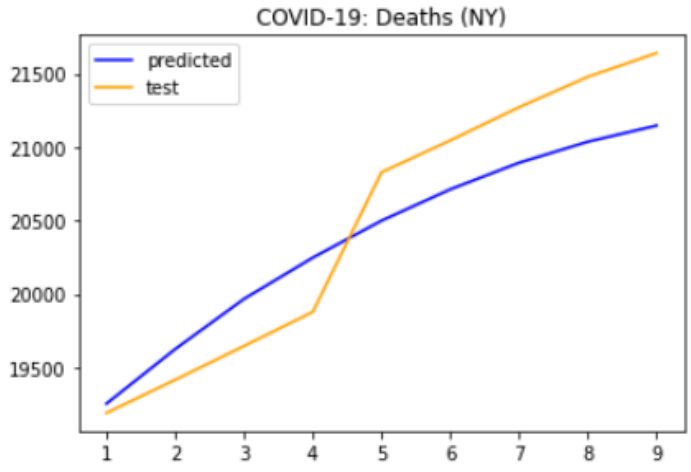


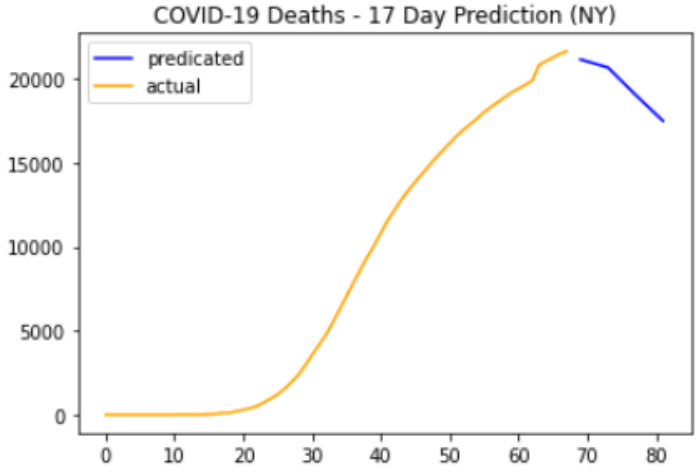


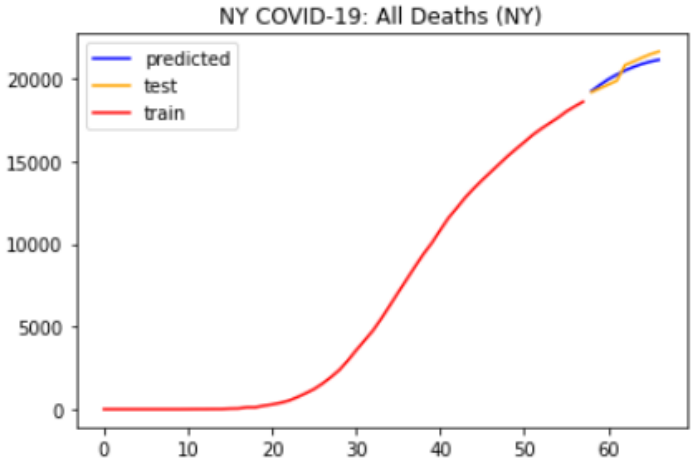
COVID-19 Modeling in US States Using LSTM Neural Networks

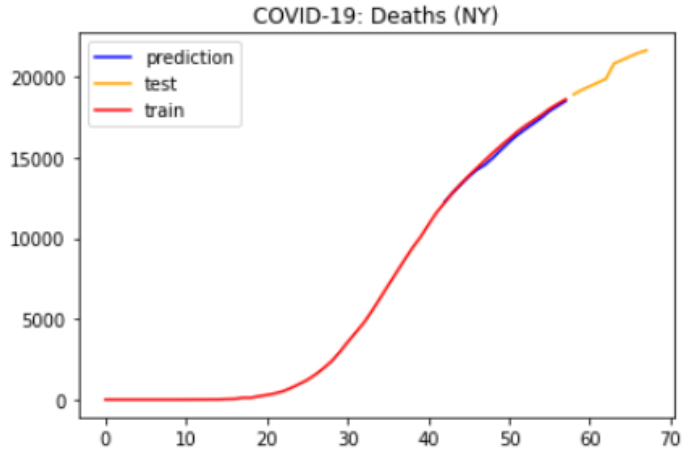
By Josh McGuckin, Neil Patel, Daniel Gallegaur, Robert Saporito

Appendix B: State Model Parameter Evaluations



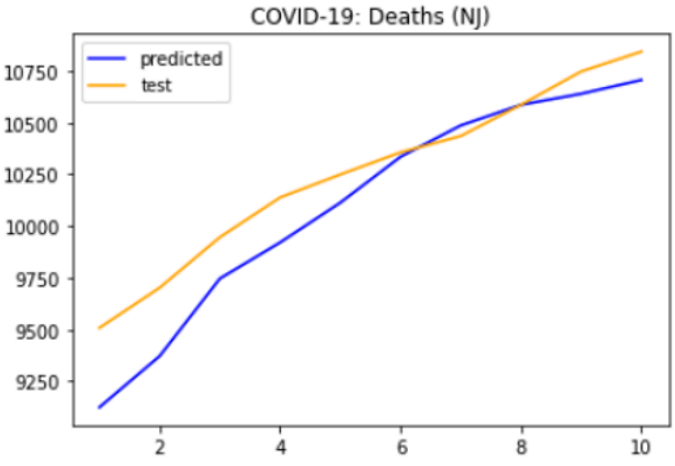


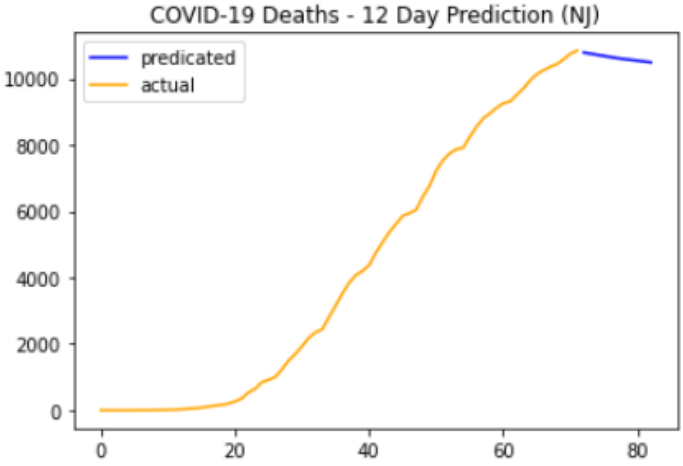





```
Test Percentage Error: 0.945
Test RMS Error: 416.397
```

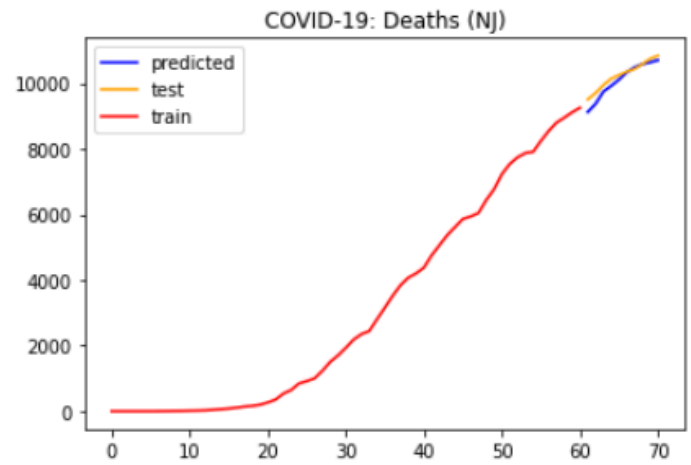
```
Test Percentage Error: 1.871
Test RMS Error: 332.308
```



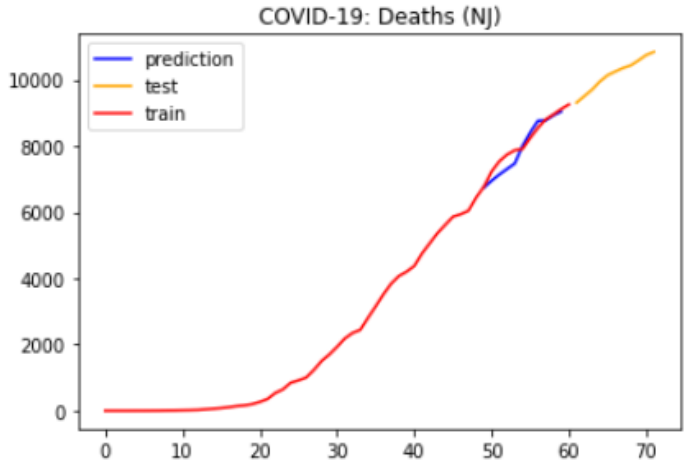


COVID-19 Modeling in US States Using LSTM Neural Networks

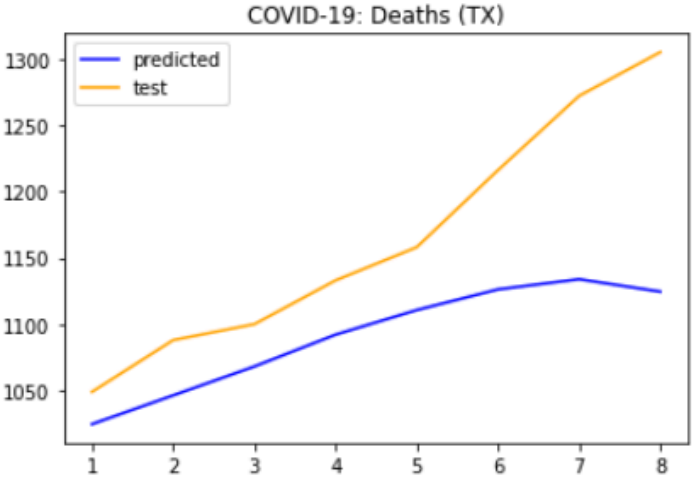
By Josh McGuckin, Neil Patel, Daniel Gallegaur, Robert Saporito

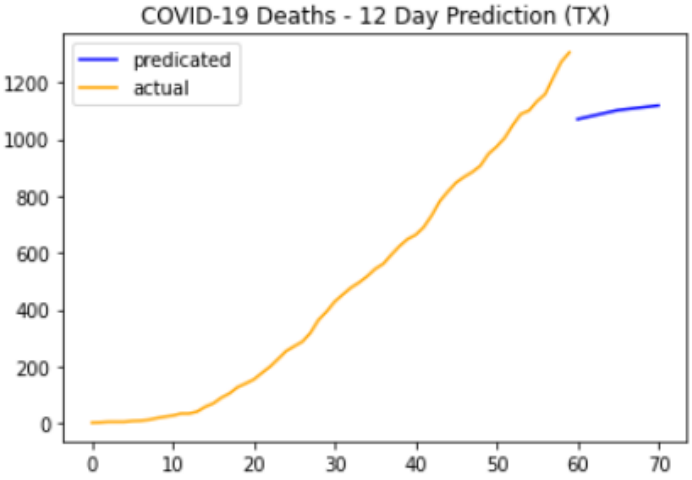


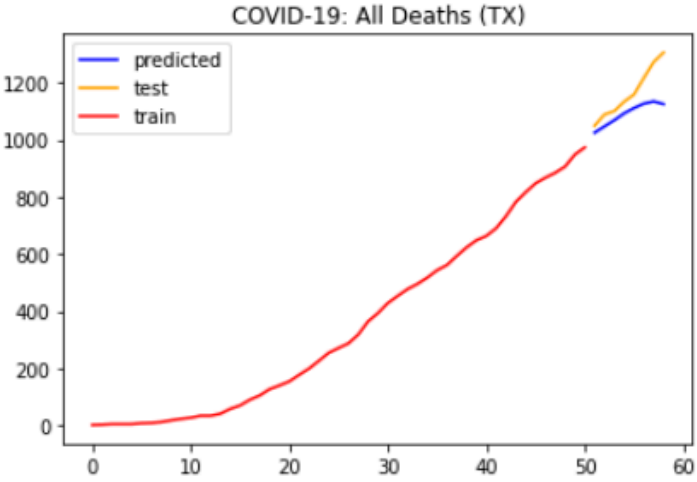
Test Percentage Error: 0.050
Test RMS Error: 955.501



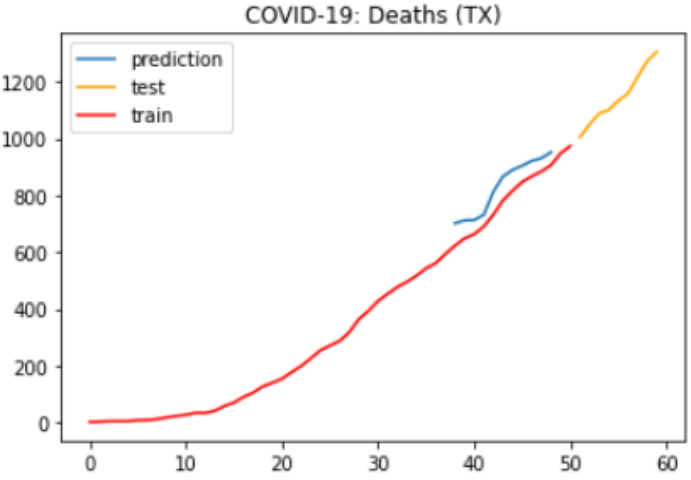
Test Percentage Error: 1.344
Test RMS Error: 1521.462







Test Percentage Error: 3.254
Test RMS Error: 123.994



Test Percentage Error: 12.134
Test RMS Error: 212.662

COVID-19 Modeling in US States Using LSTM Neural Networks

By Josh McGuckin, Neil Patel, Daniel Gallegaur, Robert Saporito

Sources:

Johns Hopkins Dataset;
COVID19Tracking. (2020, June 02). COVID19Tracking/covid-tracking-data. Retrieved June 03, 2020, from https://github.com/COVID19Tracking/covid-tracking-data

Dandekar, R., & Barbastathis, G. (2020). Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning. https://arxiv.org/abs/2004.02752

Elflein, J. (2020, May 29). U.S. COVID-19 cases by state. Retrieved June 03, 2020, from https://www.statista.com/statistics/1102807/coronavirus-covid19-cases-number-us-americans-by-state/

Google Sponsored Open COVID-19 Dataset:
Open-covid-19. (2020, June 02). Open-covid-19/data. Retrieved June 03, 2020, from https://github.com/open-covid-19/data

Pal, R., Sekh, A. A., Kar, S., & Prasad, D. K. (2020). Neural Network Based Country Wise Risk Prediction of COVID-19. https://arxiv.org/abs/2004.00959