Forecasting COVID-19 Infections Using Mobility and Socio-Economic Data

An LSTM Neural Network Approach

Endric Daues and Sameh Hameedi

Department of Applied Mathematics
Columbia University in the City of New York

Introduction

- It is estimated that deaths from COVID-19 now exceed 1 million, with infections exceeding 40 million globally
- In the United States, over 8 million infections have been reported, with deaths in excess of 200,000
- Social distancing has been put forward as a mechanism to combat the spread of the virus

Introduction

- The need for social distancing is a topic of ongoing study, though a number of initial investigations, such the work of Courtemanche *et al.* [CGLPA], have found it to be an effective measure

- Regardless of whether social distancing protocols have been impemented, it has been noted that COVID-19 deaths scale with a number of demographic factors

- As found in the work of Kim *et al.* [KMC], death rates among minorties and lower-income groups in the United States are markedly higher

Overview

- In this work, we undertake the following:

    1. We analyze mobility and demographic data for several counties in the Unitied States

    2. We investigate importance of these factors by implementing a Random Forest classifier

    3. We construct a framework of prediciting the infection rates on a Week to Week basis using a Long Short Term Memory (LSTM) neural network

Feature Engineering – Mobility Data

- We utilize the SafeGraph Social Distancing Metrics dataset
- This dataset illustrates the movement patterns of individuals in various counties throughout the United States, using the signals of electronic devices as a proxy
- In particular, we focus on the following metrics:
    1. **Mean Home Dwell Time**
    2. **Completely Home Device Count**
    3. **Full Time Work Behaviour Devices**
    4. **Part Time Work Behaviour Devices**
- In counties where social distancing protocols and lockdowns were implemented, we would expect metrics 1-2 and 3-4 to increase and decrease respectively

Feature Engineering – Mobility Data

- Due to the size of this dataset, we restrict our analysis to the months of June, July and August 2020
- In addition, we focus our analysis at the state level – we subsample 10% of the counties of a given state and average the resulting metrics for these counties
- In doing so, we circumvent the size issues that accompany working at the level of granularity associated with individual counties

Feature Engineering – Demographic Data

- We utilize Open Census and NYT Case/Death datasets
- In particular, we construct the following metrics:
    1. **Infected Population**
    2. **Population Density**
    3. **Proportion of Minorities in Population**
    4. **Proportion of Population Earning < $30,000 Annually**
- Again, we subsample counties and average these metrics to construct a state-level picture

Feature Engineering – Challenges

- The sheer size of the SafeGraph dataset presented a number of challenges
    - In particular, we were unable to analyze a subset of counties over the entire duration of the dataset
    - Each daily dataset includes all counties, and selecting a few requires downloading the entire set
- Preparing the time series data for the LSTM model mandated the shifting of weekly infection data for individual counties, which required expensive computations
- Another major challenge is presented by the fact that mobility data under a strict lockdown is strongly correlated with the level of infections in a region – i.e. a strict lockdown is a response to high infection rates

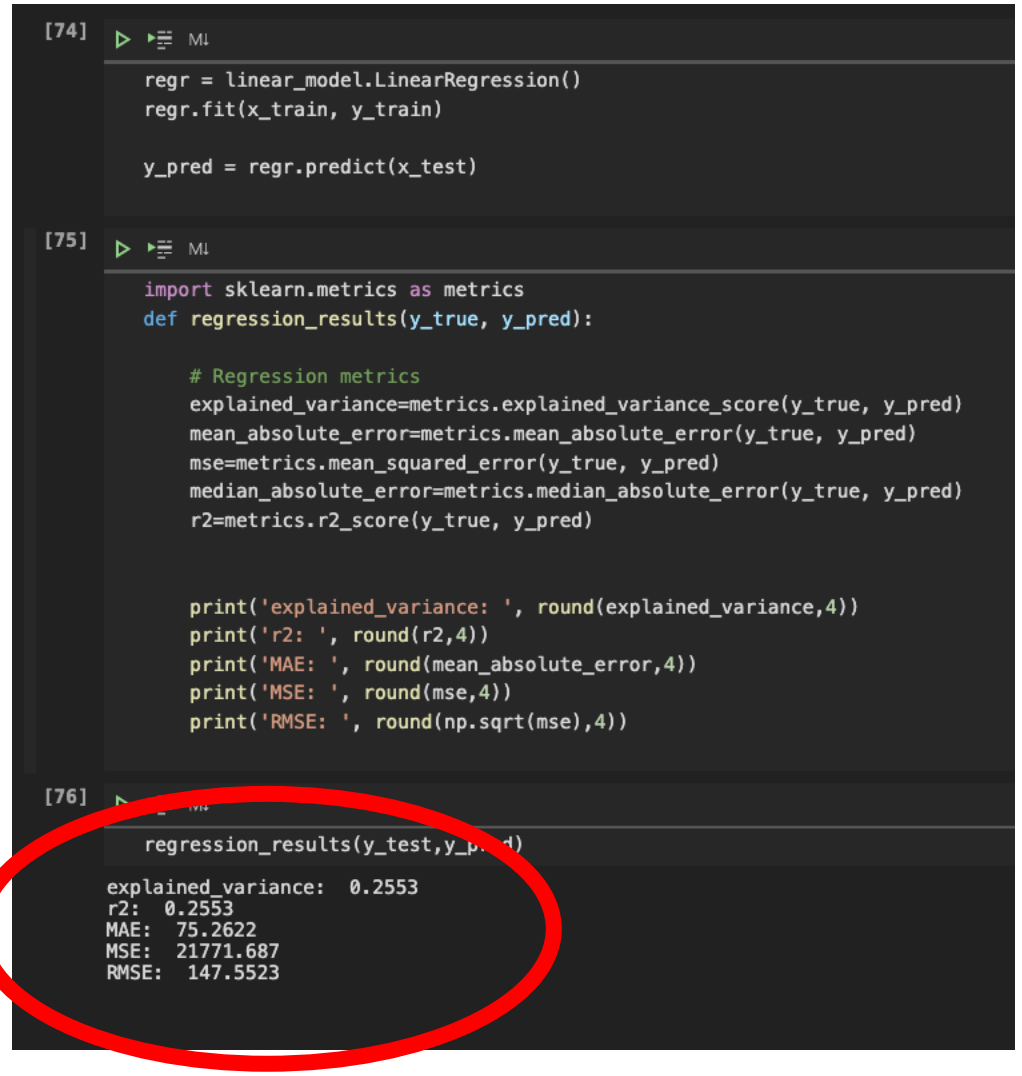
	week	state	county	median_home_dwell_time	median_percentage_time_home	median_non_home_dwell_time	mean_home_dwell_time	mean_non_home_dwell_time	completely_home_device_count	part_time_work_behavior_devices	full_time_work_behavior_devices	delivery_behavior_devices	low_inc	white	pop_density	pop	cases	shifted
0	23	01	015	696.200368	74.552058	165.942800	674.568985	323.131466	0.226135	0.089999	0.039909	0.043576	0.437855	0.739534	0.000285	115883	0.000000	16.714286
324	24	01	015	692.449821	74.869153	167.732934	673.959665	314.478093	0.233734	0.093172	0.042873	0.045135	0.437855	0.739534	0.000285	115883	16.714286	15.571429
648	25	01	015	715.306842	75.963155	186.968884	687.154278	327.722535	0.234762	0.099637	0.046523	0.049140	0.437855	0.739534	0.000285	115883	15.571429	21.142857
972	26	01	015	729.716672	76.098765	182.090292	699.797506	326.955218	0.235278	0.098060	0.048702	0.047329	0.437855	0.739534	0.000285	115883	21.142857	76.142857
1296	27	01	015	705.665130	76.616434	174.289989	683.902553	332.021528	0.243765	0.096922	0.044363	0.054539	0.437855	0.739534	0.000285	115883	76.142857	140.428571

Fig 1: The feature dataframe constructed from SafeGraph, Open Census and NYT datasets

Results

Linear Regression

- As a basline, we select a simple linear regression
- As we see, the regression only explains 25% of the variance, suggestion nonlinearity
- We then move to an ensemble learning method to capture these nonlinearties



Random Forest

- The random forrest classifier is able to explain 72% of the variance and presents significantly lower errors
- This provides evidence that we can use a random forest to approximate the non-linearity, as well as provide feature importances
- We see that population density, the proportion of minorities as well as mean home dwell time are the leading predictors

```
[78]
from sklearn.ensemble import RandomForestRegressor

regr = RandomForestRegressor(n_estimators=500,max_depth=10,bootstrap=True,max_features='sqrt', random_state=0)
regr.fit(x_train, y_train)
y_pred = regr.predict(x_test).reshape(len(x_test))
```

```
[79]
# feature importances
for i in np.argsort(regr.feature_importances_)[::-1]:
    print(features[i])
```

```
pop_density
white
mean_home_dwell_time
low_inc
median_percentage_time_home
full_time_work_behavior_devices
median_home_dwell_time
mean_non_home_dwell_time
median_non_home_dwell_time
delivery_behavior_devices
completely_home_device_count
part_time_work_behavior_devices
```

```
[80]
regression_results(y_test,y_pred)
```

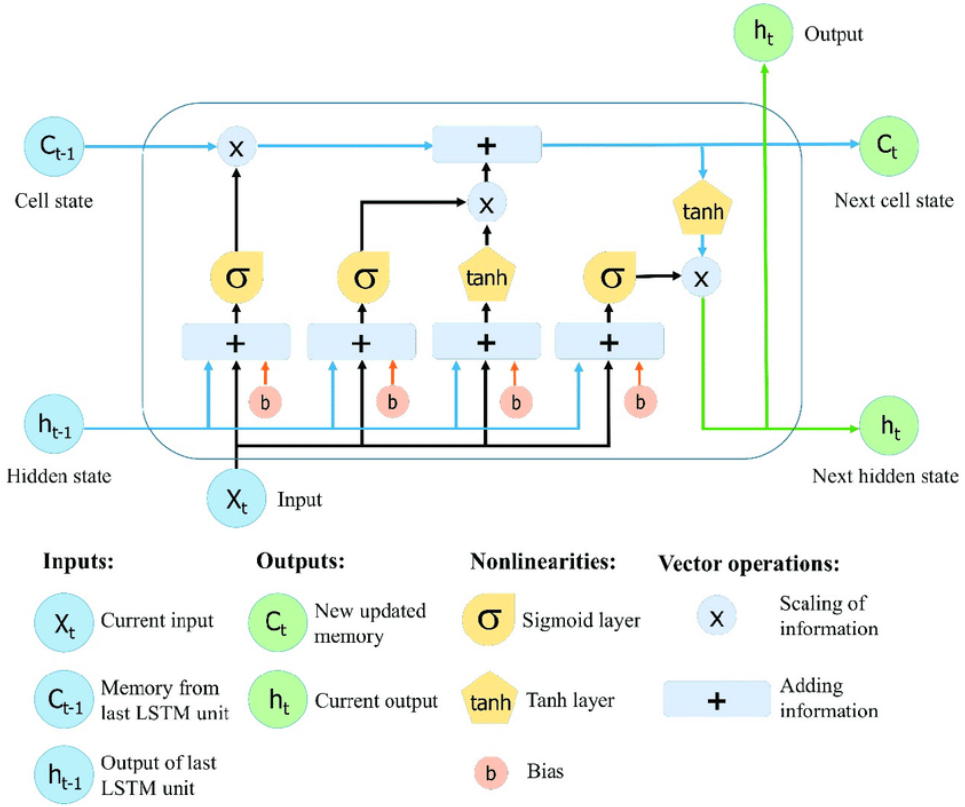
```
explained_variance:  0.72
r2:  0.7196
MAE:  43.3897
MSE:  8196.481
RMSE:  90.5344
```


```
[79]
# feature importances
for i in np.argsort(regr.feature_importances_)[::-1]:
    print(features[i])
```

```
pop_density
white
mean_home_dwell_time
low_inc
median_percentage_time_home
full_time_work_behavior_devices
median_home_dwell_time
mean_non_home_dwell_time
median_non_home_dwell_time
delivery_behavior_devices
completely_home_device_count
part_time_work_behavior_devices
```

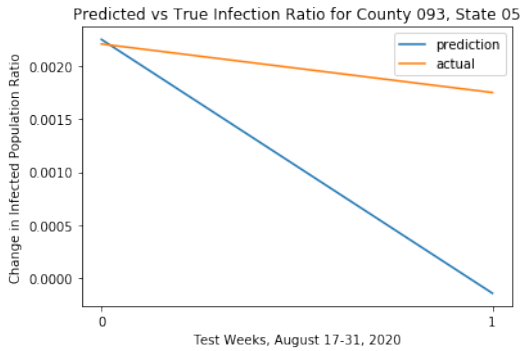
LSTM Neural Network

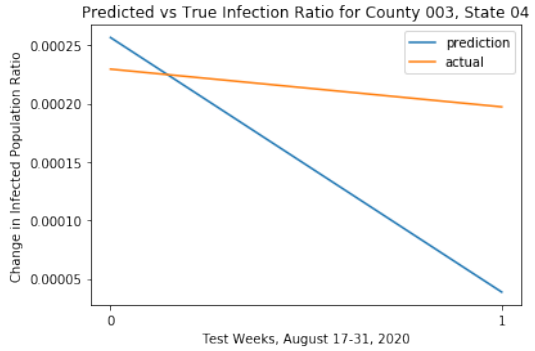
- While the non-linearity can be approximated by a random forest classifier, it ignores the sequential nature of the time series, and the week to week rollover effect of social mobility changes in one week and the changes in infection rates in the next
- To approach this, we use a long short term memory neural network, the aims to learn the rollover effect of social mobility changes by introducing memory nodes in its node structure.
- LSTM netorks have found use in time series approximation in similar instances.

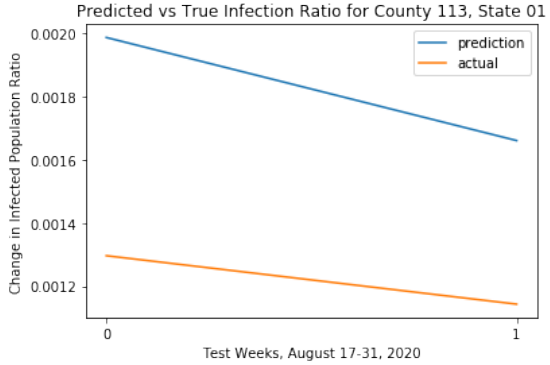


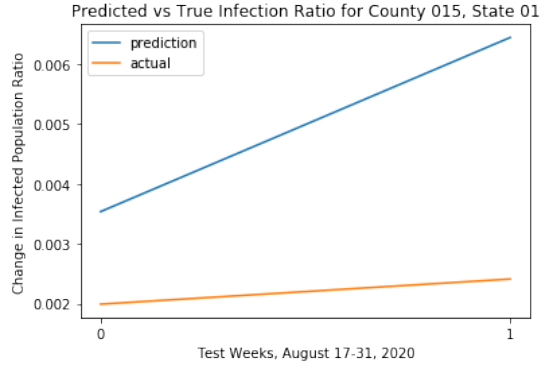
LSTM Neural Network

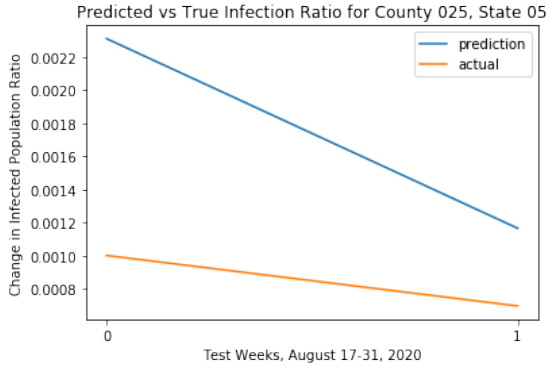
- We separate by county and state, and train the neural network on 10 weeks of data and let it predict the number of cases in a county the following two weeks
- The graphs show the two predicted weeks after the training phase, and indicate that the LSTM is able to capture upcoming trends in the infection rate in a county

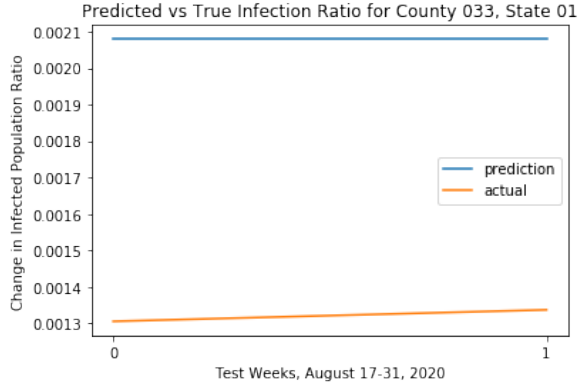






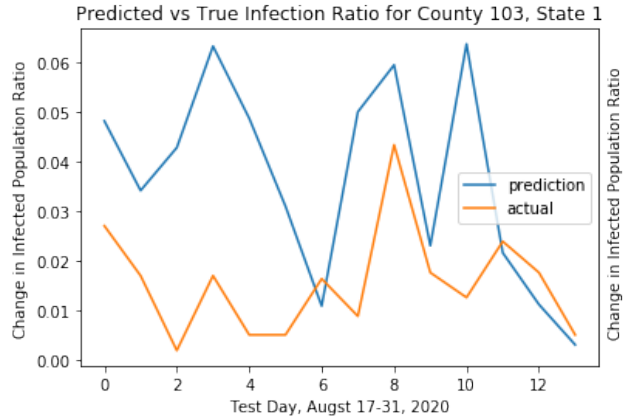


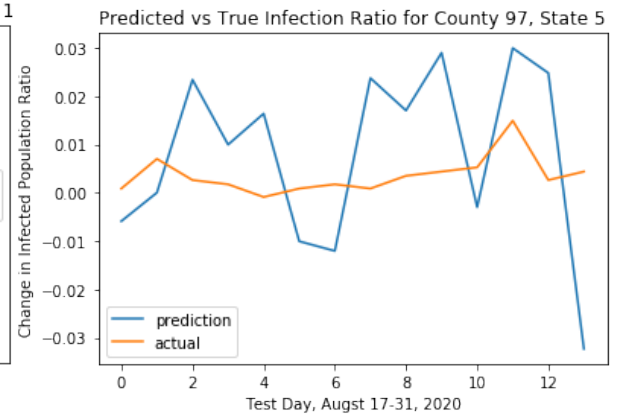


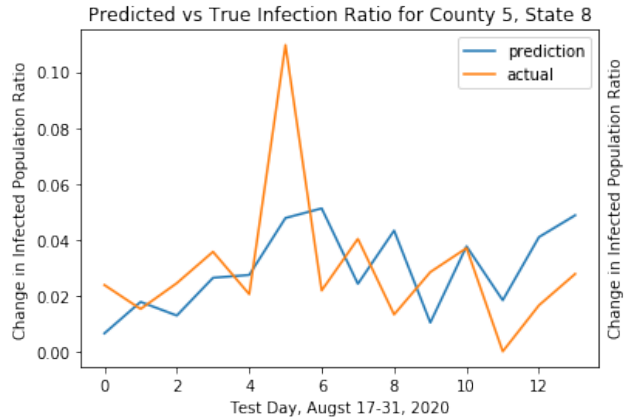


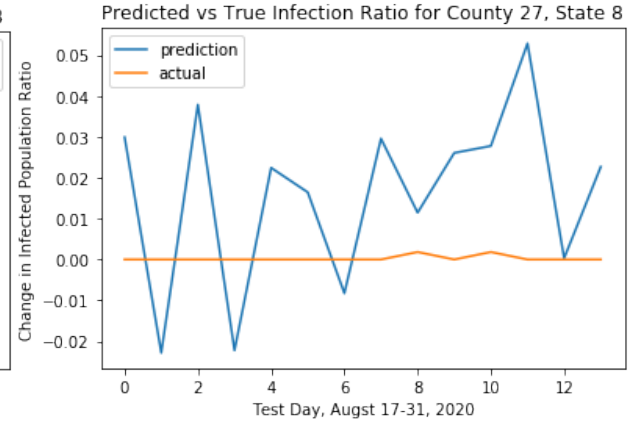
LSTM Neural Network

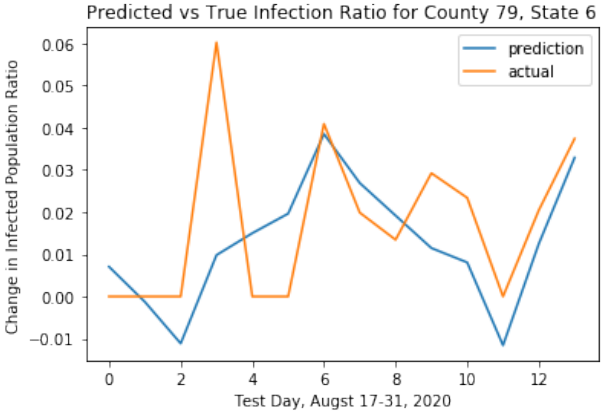
- We further look at the use of an LSTM neural network in predicting the next day new cases.
- While this granularity introduces more variance due to the lack of smoothing from the weekly average, it may find use in instances where day-to-day schedules for healthcare professionals need to be made
- Again, we see significant evidence that trends and even spikes are captured
- Nevertheless, some stationary cases seem to be exaggerated by the predictor, as in county 27, state 8.











LSTM Neural Network

- With more training, the predictive power of the LSTM model may help states evaluate social distancing and lockdown measures

- Moreover, it may provide inisight into **how medical resources can be optimally allocated as the pandemic progresses**

- For states deemed suscetible to increased infection rates by the model, it may be prudent to have medical equipment and personel ready in advance

References

All our code can be found on the following link:

https://github.com/endric-daues/stanford_datathon

[CGLPA] Courtemanche *et al.*, *Strong Social Distancing Measures in the Unites States Reduced the COVID-19 Growth Rate*, Health Affairs, 2020.

[KMC] Kim *et al.*, *COVID-19: Magnifying the Effect of Health Disparities*, Journal of General Internal Medicine, 2020.