# Impact Factors on COVID-19 Related Deaths

Jacob Esswein and Tarik Muzughi

April 2020

# Contents

# 1 Introduction

In November of 2019, the first case of COVID-19, nicknamed corona virus, occurred in Wuhan, China, and since proven that it one of the most infectious diseases we've ever seen, with one point four million reported cases, and eighty-seven thousand deaths as of the writing of this document(1). With such a impact disease, it is important to pay attention to those who may be most susceptible to it.

# 2 Data Set

This data was obtained from and accumulated by Bill Petti(2) and contains statistics such as positive and negative tested cases, the number of people on ventilators, number of people recovered, population size and their relation to deaths related to COVID-19

## 2.1 Data Set Description

In this data set there are 1605 rows and 8 columns, meaning we have 1605 locations where this data was pulled, and seven fields of input data, and output field. The input fields are the following:

- Positive Tested Cases
- Negative Tested Cases
- Number of People on Ventilators
- Number of People Recovered
- Population of Location
- Tests Needed Per Population

## 2.2 Input Data Visualization

Here are the histogram plots of the frequency of each input variable:

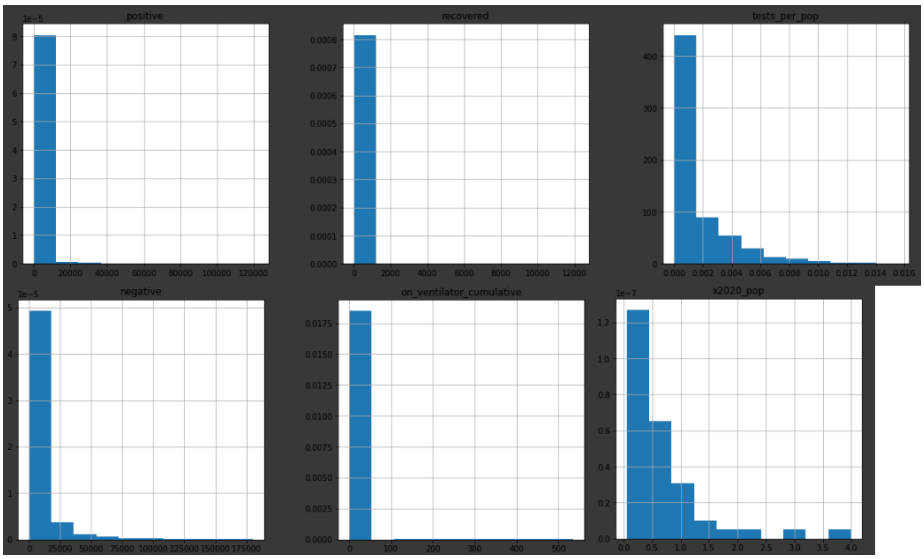


Figure 1: Input Data Distribution Histograms

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| positive | 1604.0 | 1.502624e+03 | 7.261589e+03 | 0.0 | 1.100000e+01 | 1.060000e+02 | 6.432500e+02 | 1.220310e+05 |
| negative | 1604.0 | 7.118349e+03 | 1.613627e+04 | 0.0 | 9.100000e+01 | 1.113500e+03 | 6.579250e+03 | 1.802490e+05 |
| on_ventilator_cumulative | 1604.0 | 3.205736e+00 | 3.195941e+01 | 0.0 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 5.350000e+02 |
| recovered | 1604.0 | 4.563155e+01 | 5.559402e+02 | 0.0 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 1.218700e+04 |
| x2020_pop | 1604.0 | 6.648308e+06 | 7.500173e+06 | 567025.0 | 1.826156e+06 | 4.645184e+06 | 7.797095e+06 | 3.993749e+07 |
| tests_per_pop | 1604.0 | 1.478895e-03 | 2.231617e-03 | 0.0 | 2.547463e-05 | 3.590630e-04 | 2.258463e-03 | 1.554901e-02 |
| deaths | 1604.0 | 3.278055e+01 | 1.923532e+02 | 0.0 | 0.000000e+00 | 1.000000e+00 | 1.100000e+01 | 4.159000e+03 |

Table 1: Input Feature Statistics

### 2.3 Output Data Visualization

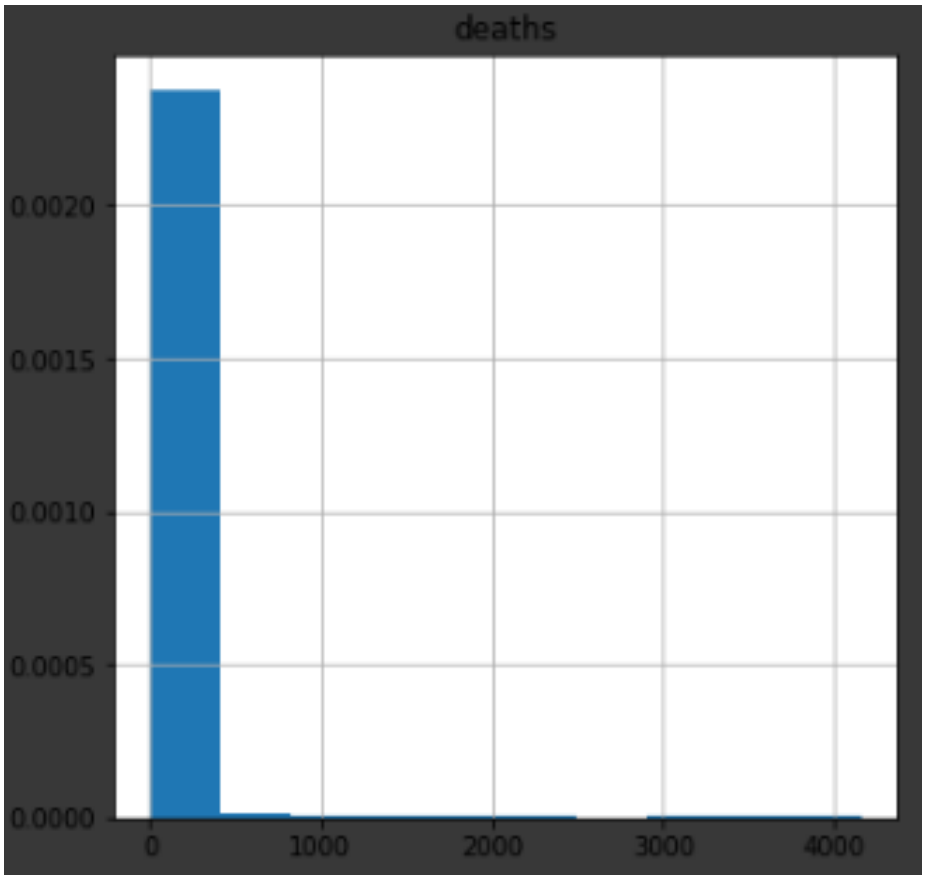


Figure 2: Output Data Distribution Histogram

## 3 Data Processing

### 3.1 Data Normalization

The data in this data set is not distributed uniformly, thus we need to pre-process all of the data with normalization. Our data will be normalized using Z-Score Normalization for accuracy purposes. The formula for Z-Score Normalization is:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Figure 3: Z-Score Normalization Formula

### 3.2 Relationship Between Input and Output Data

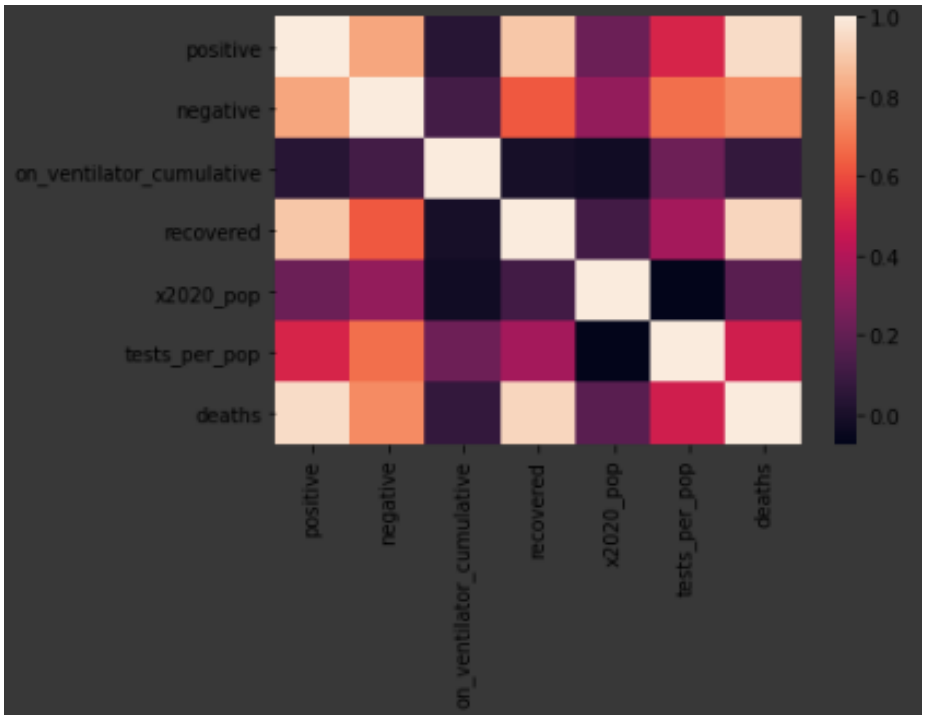


Figure 4: Heat Map of Input and Output Data

From this heat map we can see obvious correlations such as positive tests and correlated, but less expected correlations such as amount recovered and deaths.

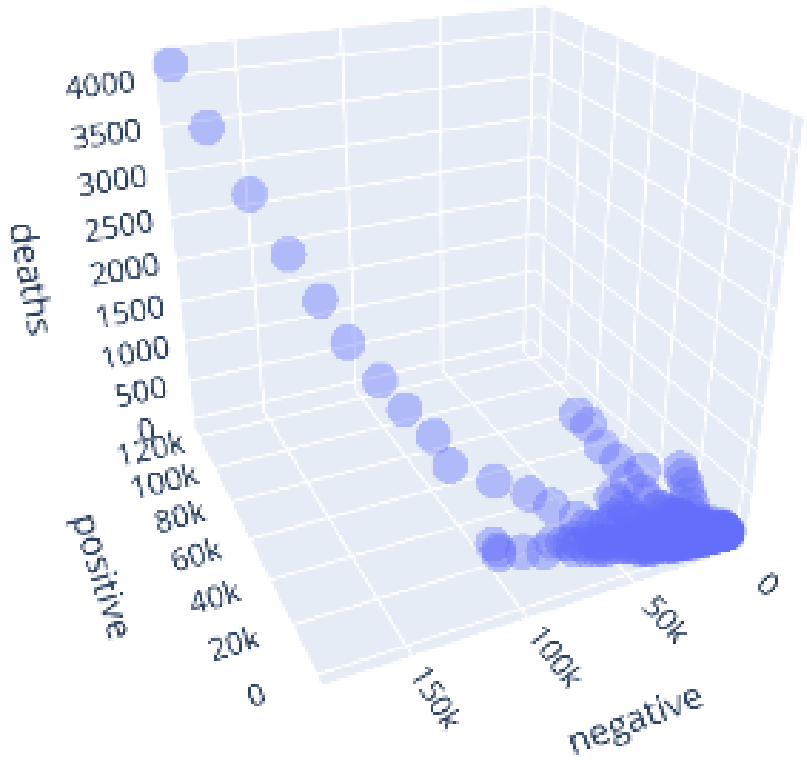


Figure 5: 3-D Graph of Relationship of Positive Tests, Negative Tests, and Deaths

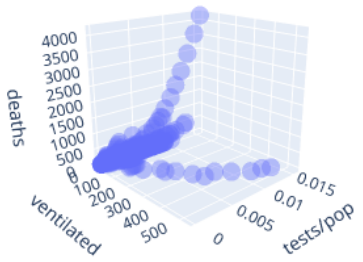


Figure 5: 3-D Graph of Relationship Between Deaths, People with Ventilators, and Tested Population

# 4 Modeling and Neural Networks

## 4.1 Single-Layered v.s. Multi-Layered

While starting the processes of splitting our data and training our neural networks we first began by using a single-layered model, then slowly transitioning this model into a multi-layered model. The results of this are about what you would expect. When using the single-layered model we saw that our neural network's predictions were astronomically off, so much so that any normal person could see that it was predicting incorrect data. One of the biggest factors we believe that could be causing this is the fact that their exists some outliers in our data, specifically, our data only covers one month worth of data, and during the very early stages of COVID's breakout in the United States, thus resulting in areas that didn't have any fatalities related to COVID yet.

As we transition into multi-layered models and increased the layers we saw our neural network start producing even more accurate predictions that we were then able to use.

## 4.2 Linear Activation v.s. Sigmoid Activation

While building our neural network and learning curves we tested results between using linear activation and sigmoid activation. In our resulting learning curve from using linear activation we saw that our MAE was very volatile between

epochs, meaning that it peaked and dipped a lot when changing the epochs.

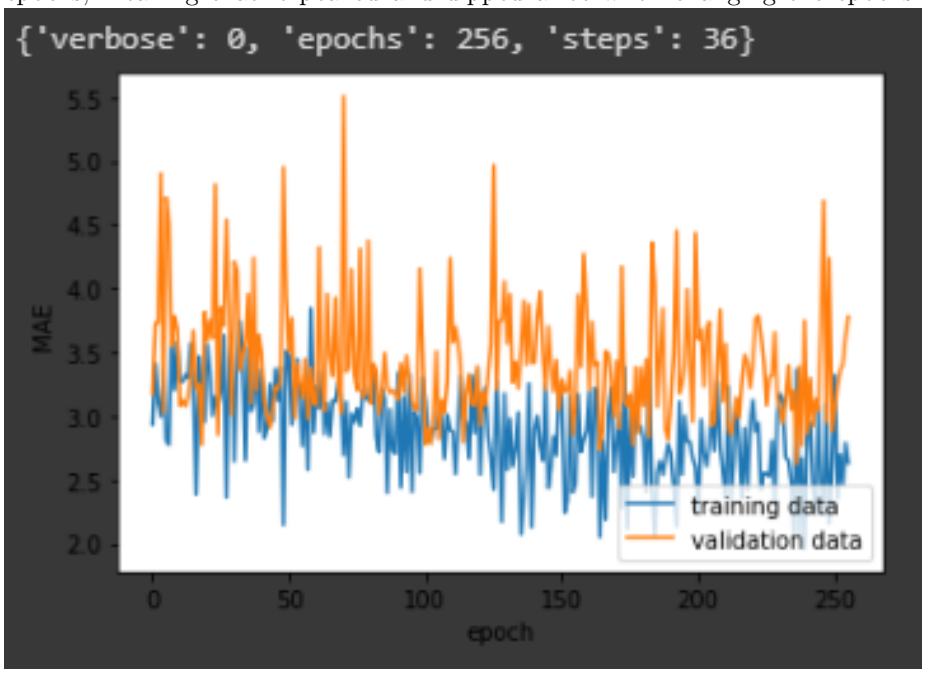


Figure 6: Learning Curve when using Linear Activation (Epoch v.s. MAE)

When we used sigmoid activation we saw much different results in our learning curve. While the linear activation resulted in a volatile learning curve, in the sigmoid activation we saw a very consistent graph, only occasionally peaking

compared to the linear activation.

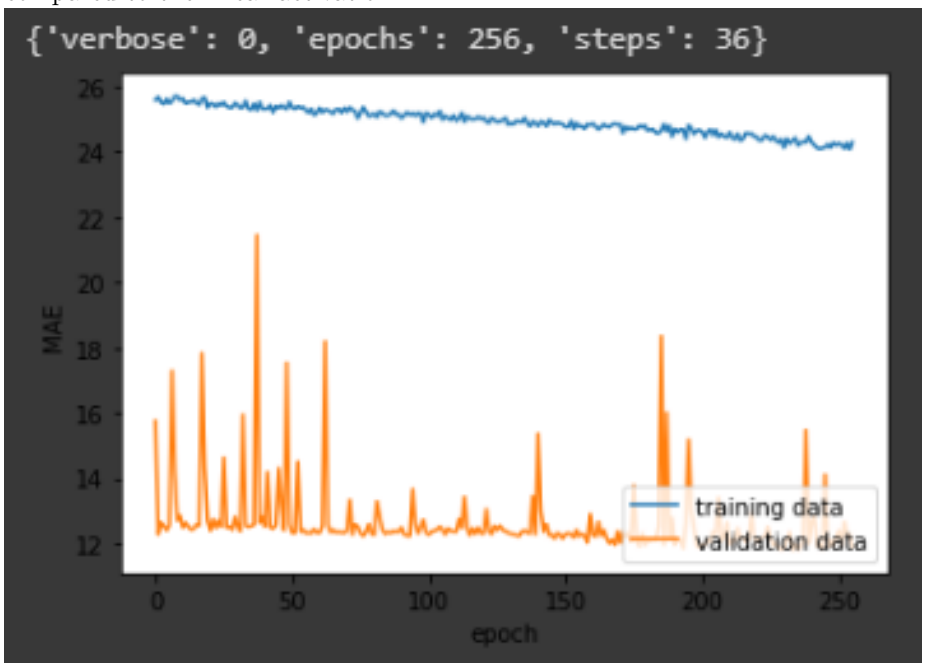


Figure 7: Learning Curve when using Linear Activation (Epoch v.s. MAE)

### 4.3 Architecture

When building out the architecture for our neural network we found that when using the linear regression model, adding more epochs only resulted in less and less accurate data. Specifically we found that increasing the epochs any higher then 250 would cause over fitting. We also found that decreasing epochs below 200 resulted in under fitting, thus us choosing to keep epochs in within that range.

### 4.4 Prediction v.s. True Data

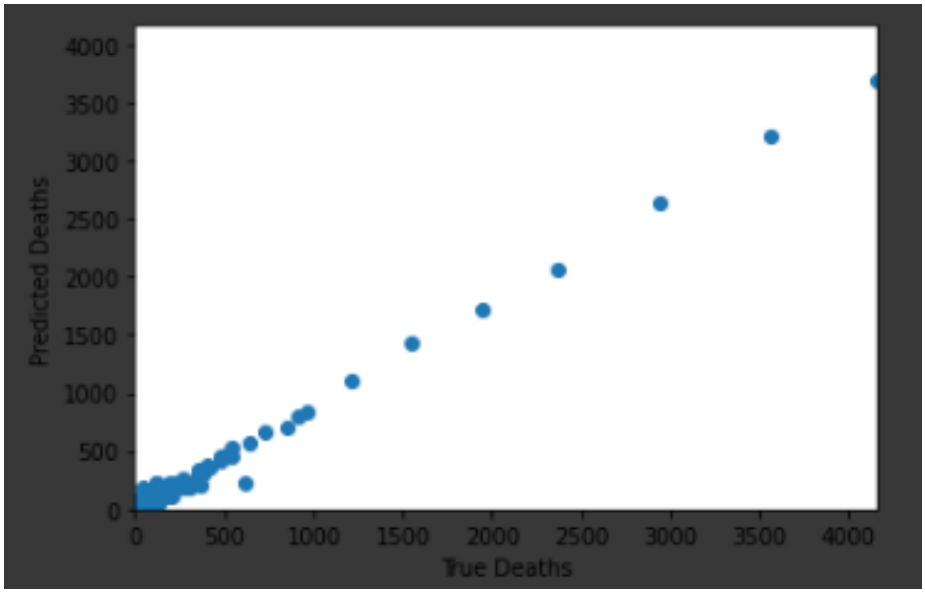


Figure 8: Graph of Deaths Recorded v.s. Deaths Predicted

Based on all work done by our neural network previously stated and the graph above it would appear that our neural network is well trained, in the case of it not being overly trained (as its not a perfect straight, line) as well as not being under trained (a semi-straight lined to formed).

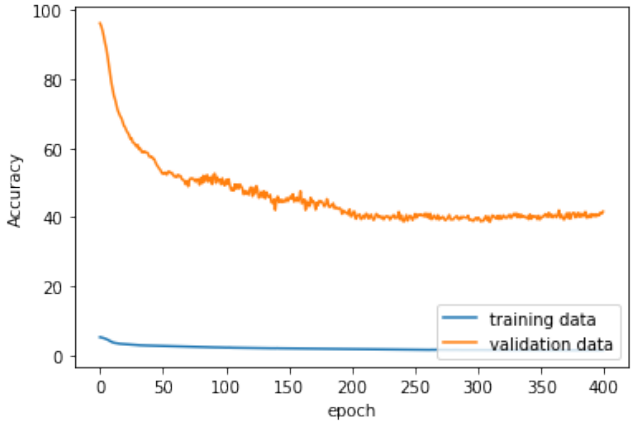


Figure 8.5: Graph of Epochs v.s. Accuracy

# 5 Feature Importance and Reduction

## 5.1 Importance Before Feature Reduction

Firstly through looking at Figure 9, the most interesting thing is that tests per population is the least significant input feature, however in the heat-map from Figure 4, we see that Test Per Population and Deaths are relatively correlated, which is rather contradictory to Figure 9. A possible reason for this is the number of tests that are being handed out is small compared the deaths caused by COVID-19, making test per population to be deflated. A second causation for this discrepancy is the people being tested are only on those being symptomatic and thus causing its importance to be misleading.

The most significant feature appears to be the amount of people recovered from COVID-19. This makes sense as if a person has recovered from the disease that means they have not died from it, thus having a direct impact on the data.

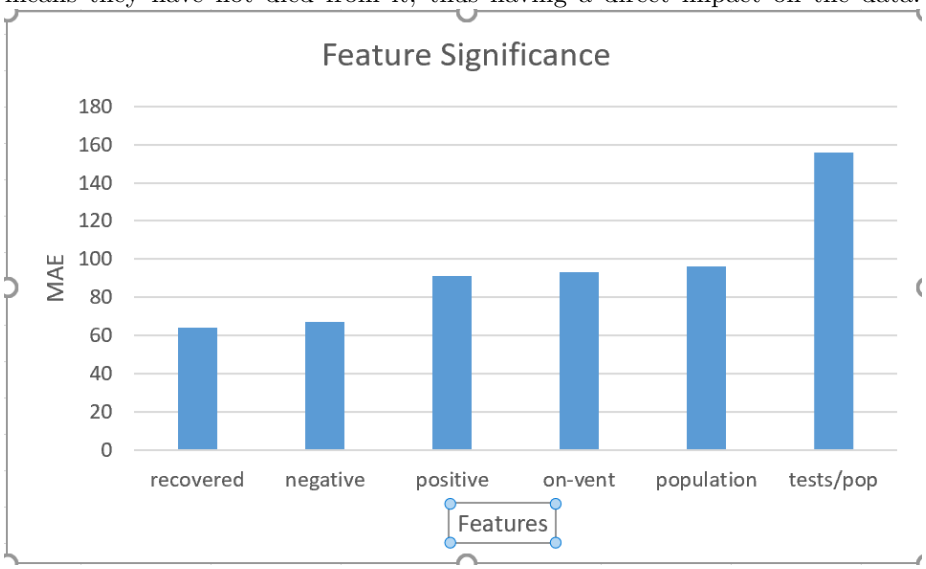


Figure 9: Feature Significance (Input Features v.s. MAE)

## 5.2 Importance After Feature Reduction

Looking at Figure 10 we can see that the MAE improves quite a bit when Test per Population is removed. This is interesting as previously stated, Tests Per Population had the least significance as seen in Figure 9. Furthermore after

removing Population and Tests per Population the data improves as a whole.

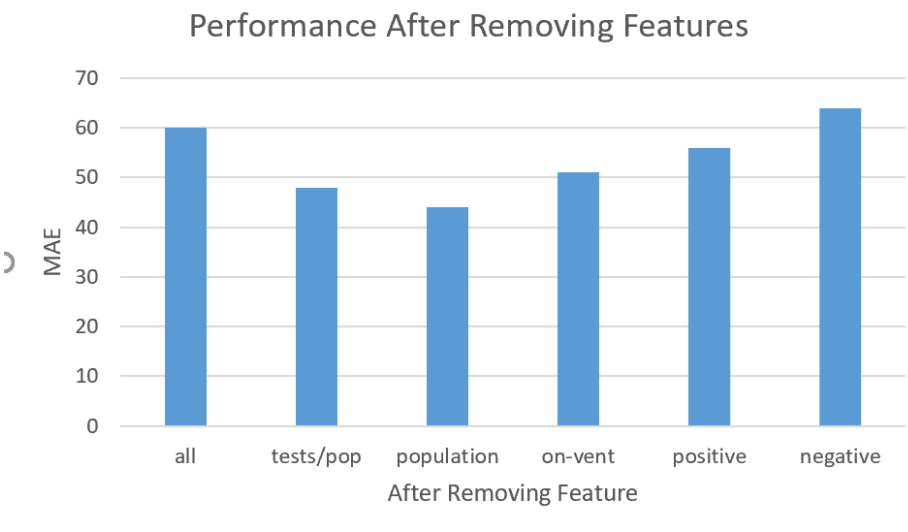


Figure 10: Performance After Removing Features (Input Features v.s. MAE)

# 6 Citations

1. https://www.worldometers.info/coronavirus/
2. https://billpetti.shinyapps.io/covid$_19_{c}ountry_{s}tate_{d}ashboard$