Models Attempted

To further analyze and locate anomalous time points that can hint/suggest Covid-19 mutations, Principal Component Outlier Detection, Local Outlier Factor, LSTM Autoencoder, Vector AutoRegression and a univariate LSTM was used to find the anomalous time points. The methods that were tested can be used for unsupervised learning problems as our dataset contains no target labels.

The main concept behind PCA (Principal Component Analysis) is that an anomalous time point should exhibit a higher/ab-normal reconstruction error than the nominal time points. The reconstruction error is calculated between the rows with the PCA transformed data and the rows with the original normalized data. So the data is reduced to a lower number of dimensions. The reconstruction error is then computed by taking the sum of squares of the differences between the values in the original rows and the reconstructed rows. Our reconstruction error threshold was set at 200. Time points with a higher error than the threshold is considered an anomalous point and time points with errors below the threshold are considered nominal data points. However, the potential drawbacks we observed in this algorithm were that this method only works with strictly numerical data and since PCA uses matrices in memory it doesn't scale well to large datasets.

The next algorithm that was tested was Local Outlier Factor. Local Outlier Factor (LOF) is an anomaly detection algorithm where the local density deviation of a given point with respect to its neighbors is computed. Samples with a much lower density score are considered as outlier data points. The first parameter considered is the k value which represents the numbers of neighboring data points to consider. This value is usually higher than the minimum number of samples that any cluster contains. But it should be lower than the maximum number of nearby samples, so if there are 30 points close to each other, the k value should be smaller. Using the k value, the "reachability distance" is calculated to understand the maximum of the distance between two points. So if two points are k neighbors, the reachability distance is the k-distance of the data points. Lastly, the local reachability density (lrd) is calculated to inform the distance needed to travel from a data point to the next cluster of points. If the lrd is lower, this means the point is sparse and less dense meaning the point is likely to be an outlier. Finally, the LOF is the average ratio between the lrd scores of the neighbors and the lrd score of a. If LOF is greater than 1, this means the lrd of a is small which indicates that point is an outlier. The potential drawbacks to this method is that the LOF ratio is difficult to interpret because the ratio can be interpreted differently for a wide range of datasets/problems. Choosing the correct k value can also pose problems because if the k value is too large, some anomalous points can be missed.

Vector AutoRegression (VAR) and a univariate Long Short Term Memory (LSTM) model was also tested on our dataset. In VAR, each time series is modeled as a function which uses the past values as model inputs. The model predictors are the lagged past values used to predict the

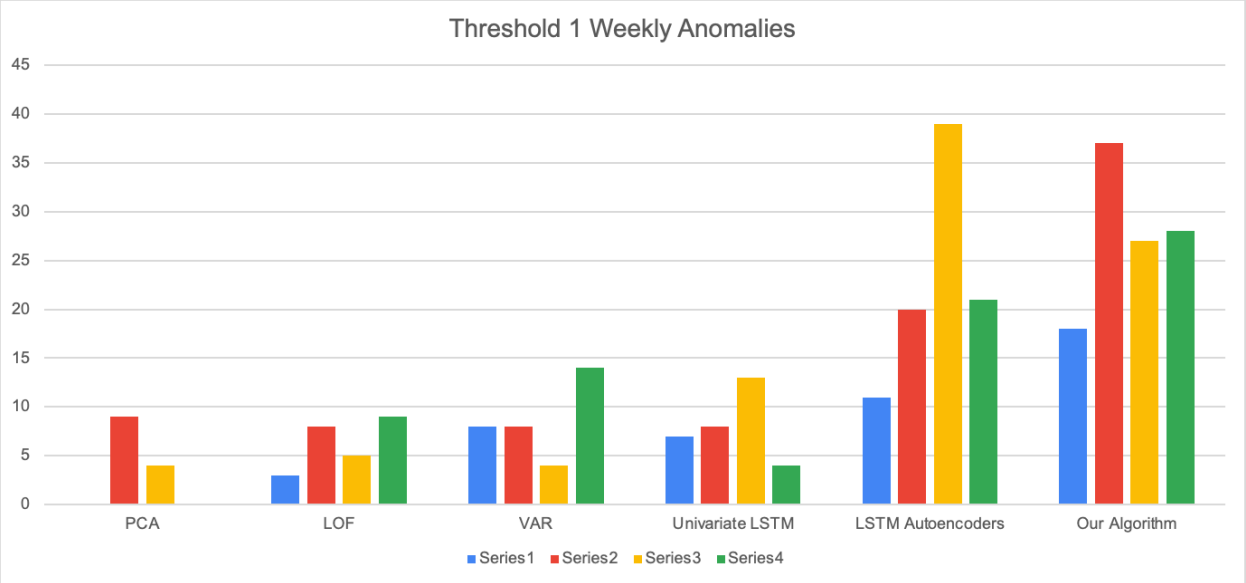
current time series values. Each variable is modeled as a linear combination of the past variables of itself and the other time series values. Before running the VAR model, a causation and stationary test must be conducted. One of the assumptions of VAR is that each time series should be correlated and another assumption is that there shouldn't be any seasonality in the data, which means there shouldn't be data trends based on the time of year. Time points with a higher sum of square error than the error threshold are considered anomalous. The error threshold can be calculated using mean(error) + std(error) *1 or 2. A similar method is used for univariate LSTM models where past data points are considered to predict a current data point. LSTMS are useful because they give the ability for the model to remember long term/short term values with a cell state and hidden state. The error threshold is also the exact same as the VAR threshold. But since this is a univariate technique, over half of the time series must be anomalous for a certain time point for a point to be considered a global anomaly. The drawbacks of these algorithms are VAR doesn't perform well with lowly correlated features (our covid symptoms aren't correlated) and the univariate LSTM rarely detects outliers because majority of the time series columns must classify the specific point as an anomaly to be considered anomalous. This is difficult to accomplish because each time series column will produce different anomalous time points especially if there are many testing points.

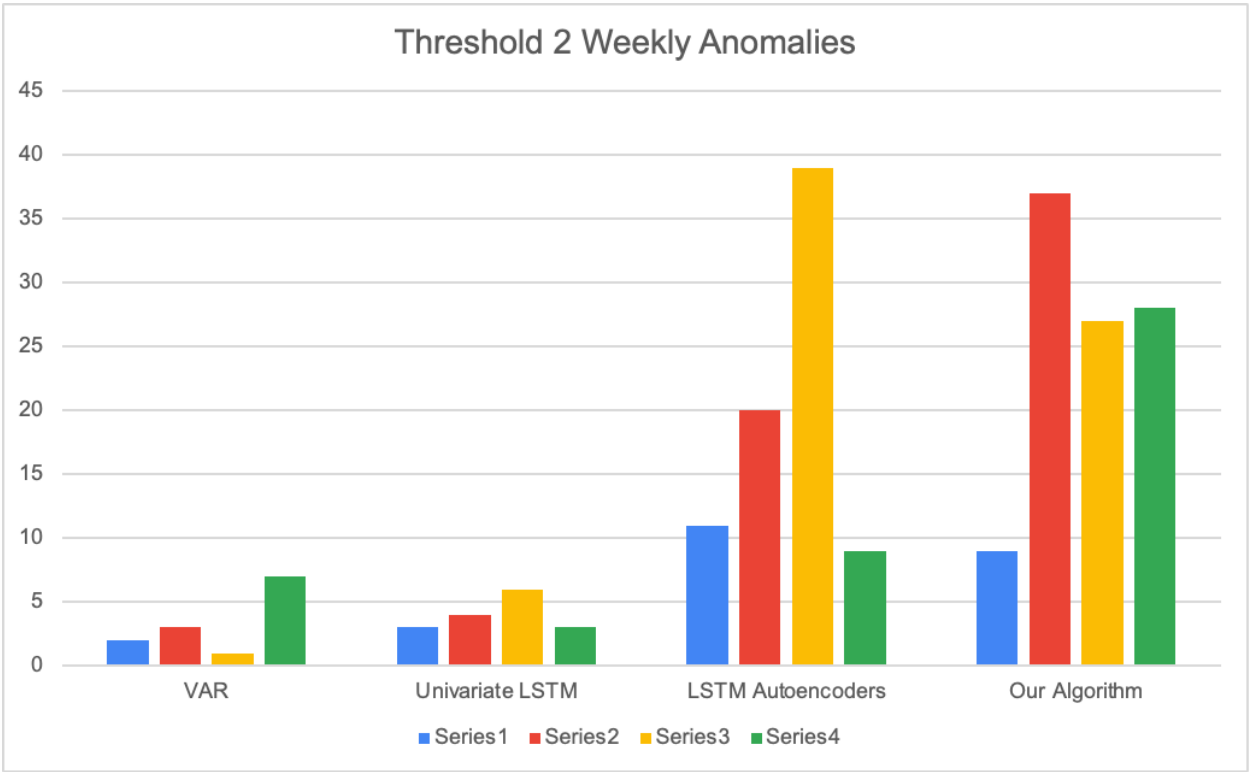
The last method that was tested was the LSTM autoencoder. This algorithm follows a similar structure to PCA because reconstruction errors are used to compute anomalies, but this model seems to be more accurate because the model learns the data representation through a more extensive training process. LSTM models utilize a cell state to learn the long term relationships of the data, if data is worth remembering it will pass through a sigmoid function and return a number closer to 1. But if it is worth forgetting, the sigmoid function will return a number closer to zero. The LSTM autoencoder takes LSTMs to another by using their layers as encoders/decoders. Typical autoencoders use encoder functions to decompress the data into a lower dimension and then use decoding functions to learn/recompute the decompressed data into the data's original dimension. So in an LSTM autoencoder, the first layer uses n neurons, the second layer uses n/2 neurons to decompress the input data, the third layer than repeats the vector to pass it on to the fourth layer and then the last layer recompresses the data into the original data dimension. In our results we observed PCA and the LSTM Autoencoder were the strongest models as it was able to learn the representation of the multivariate data as a whole. But the LSTM autoencoder uses more complex activation functions such as RELU compared to PCA which uses a linear function, which at the end had a greater impact to detect more anomaly time points.

As the LSTM autoencoder was our best performing model, the advantages from the other baseline algorithms were incorporated into the LSTM autoencoder to create an enhanced version which we will be showcased over the coming sections.

Model Anomaly Results (Threshold1, Threshold2)

Date in January 2021	Principal Components Analysis	Local Outlier Factor	LSTM Autoencoders	Vector Autoregression	Univariate LSTM	Our Enhanced LSTM Autoencoders Method
(Threshold 1, Threshold 2)						
1						
2						
3	0	0	2,2	1,0	1,0	3,2
4	0	0	3,3	3,1	2,0	2,1
5	0	2	2,2	0,0	0,0	5,2
6	0	0	1,1	0,0	1,1	1,1
7	0	0	1,1	0,0	3,2	2,1
8	0	1	1,1	1,1	0,0	4,2
9	0	0	0,0	0,0	0,0	0,0
10	0	0	1,1	3,0	0,0	1,0
11	0	0	3,2	1,0	0,0	4,0
12	0	0	1,1	0,0	0,0	2,2
13	6	3	1,1	2,1	2,0	7,3
14	0	0	1,1	3,1	2,2	2,1
15	1	0	1,1	1,0	0,0	2,2
16	0	1	0,0	0,0	2,2	6,1
17	0	0	1,1	0,0	0,0	0,0
18	2	2	12,11	2,1	1,0	14,13
19	0	2	11,8	0,0	1,0	8,6
20	1	1	5,5	1,0	0,0	6,16
21	0	0	2,2	1,1	2,0	4,2
22	0	0	3,3	0,0	1,1	2,1
23	0	1	3,3	0,0	2,2	2,0
24	1	0	1,1	1,0	4,3	2,0
25	2	1	5,1	0,0	0,0	1,1
26	0	1	9,9	1,0	2,0	2,2
27	0	2	7,7	0,0	2,2	9,2
28	0	2	4,4	2,1	2,0	2,1
29	0	1	7,1	2,1	0,0	5,1
30	0	0	2,1	0,0	2,1	1,0
31	0	1	1,0	4,2	0,0	3,1
	0	3	0,0	6,4	0,0	7,2
	0		0,0	0,0	0,0	1,0
Total Mutations between January 1 and January 15	8	7	19,18	15,4	13,7	41,18
Total Mutations Between January 1 and January 26	14	19	81,76	23,7	30,15	93,62





The series represents each week in January which was used as our testing data