**Base Model**

For the training period, 7th March 2020 – 31st December 2020 (300 days), the binary vector of Covid-19 case anomalies, L, and the binary vector of search query anomalies , were calculated for each state. The two vectors, for each state, were then compared and matched to find the best lag-threshold combination that maximised the resulting F-score. The comparison algorithm uses unique matching whereby a Covid-19 case anomaly can only match to one symptom rate anomaly.

Using the best lag-threshold combination for each state, the number of forecasted Covid-19 case anomalies and symptom rate anomalies for the 300-day training period were plotted side by side.

Chart, bar chart

Description automatically generated

Figure 1 AK - MO

Chart, bar chart

Description automatically generated

Figure 2 MS - WY

With the exception of AZ, CA, GA, IL, IN, NC, NM, SC, SD, TX and WV, all states had a higher or equal number of symptom rate anomalies than Covid-19 case anomalies.

The same was then done for the 36-day testing period (1st Jan 2021 to 5th Feb 2021) using the best lag-threshold combination for each state.

Chart, bar chart

Description automatically generated

Figure Outbreak Count in Testing Data

In the testing results, only OR and TN have a higher number Covid-19 case anomalies than symptom rate anomalies.

**State CT**

The training data for state CT has 67 symptom rate anomalies and 18 Covid-19 case anomalies. This is 272% more symptom rate anomalies than Covid-19 case anomalies. The best TH was 0.1 and best lag was 3 days.

For a lag of 3 days and threshold of 0.1, the following results were calculated:

Text

Description automatically generated

Figure State CT’s Anomalies (TH = 0.1, lag=3)

Chart

Description automatically generated with medium confidence

Figure State CT's daily delta values compared with symptom rate anomalies (TH =0.1, lag = 3)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **0** | **TP**  15 | | **FN**  3 | **TN**  230 | | Precision = 0.224  Recall = 0.833  F-score = 0.352 |

**Why is the best threshold not higher?**

Given a lag of 28 days, when the threshold is set to 1.1, the number of symptom rate anomalies is 12. This is a more appropriate symptom anomaly forecast than 67 (TH=0.1, lag=3) however the F-score is lower (best F-score=0.352). Although the number of FP has decreased by 83%, this is offset by the decrease in number of TP (80%) and increase in FN (533%).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **FP**  9 | **TP**  3 | | **FN**  19 | **TN**  269 | | Precision = 0.25  Recall = 0.136  F-score = 0.1765 |

A picture containing text

Description automatically generated

Figure State CT’s Anomalies (TH = 1.1, lag = 28)

Chart, line chart

Description automatically generated

Figure State CT's delta values compared with symptom rate anomalies (TH =1.1, lag = 28)

The following graph shows the F-Scores of state CT across a 0–30 day lag when different thresholds are applied. A general downward trend in F-scores (as TH increases) can be seen.

Text

Description automatically generated with medium confidence

TH = 0.3

TH = 0.5

TH = 0.7

TH = 0.9

TH = 1.1

Figure shows F scores (across 0-30 days lags) when TH is increased (0.3 – 1.1)

**State VT**

The training data for state VT has 64 symptom rate anomalies and 17 Covid-19 case anomalies. This is 276% more symptom rate anomalies than Covid-19 case anomalies. The best TH was 0.2 and best lag was 0 days.

For a best lag of 0 days and best threshold of 0.2, the following results were calculated.

Text, application

Description automatically generated

Figure State VT’s Anomalies (TH=0.2, lag=0)

A picture containing text

Description automatically generated

Figure State VT's delta values and Symptom Rate Anomalies (TH=0.2, lag=0)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **FP**  47 | **TP**  17 | | **FN**  0 | **TN**  236 | | Precision = 0.266  Recall = 1.0  F-score = 0.420 |

**State VT: TH = 1.1 and lag = 28 days:**

Text

Description automatically generated

Figure

A picture containing chart

Description automatically generated

Figure

Icon

Description automatically generated

TH = 0.2

TH = 0.4

TH = 0.6

TH = 0.8

TH = 1.2

TH = 1.4

TH = 1.6

Figure F scores (across 0-30 days lags) when TH is increased (0.2 – 1.6)

**The comparison to find the best lag-threshold combination is repeated with the following change:**

**The threshold can take a value from 0-2.0 with 0.01 increments instead of 0.1 increments**. This was expected to allow the model to give a more accurate definition of “symptom rate anomaly” in each state.

However, as seen in the Figure 12, the symptom rate anomalies are still over-forecasted.

Chart, bar chart, line chart

Description automatically generated

Figure

More importantly, as seen in the Figure 13, this change does not produce a lower symptom rate anomaly forecast (compared to the original implementation). Only 11 states see a marginal improvement in symptom rate anomaly forecasts.

Chart, bar chart, line chart, histogram

Description automatically generated

Figure

**Thoughts**

There is a lot of noise in the symptom rate data causing false positive anomalies. This could be due to search queries that are made out of concern or curiosity instead of signalling a positive Covid-19 case diagnosis. At first, I thought that the best TH had to be a larger value so that the number of symptom rate anomalies would be lowered to a more appropriate level. However when the TH was increased, the number of true positive counts decrease while false negatives increase causing the overall f-score to be lower.

**Next Steps:** Moving average and exponential smoothing will be applied to the symptom rate data to minimise noise and see if results can be improved.