Using random seed 1 for all the experiments, 5-fold cross validation. Changed from 3 to 5 folds because even in N = 212, 3-fold means the training set is only twice the testing set, which is not very good and common.

Seems like the time-series data gives consistently worse results, unfortunately.

So we will use the non time-series data for now, even though I think theoretically time-series makes more sense.

For non time series, the yellow highlights all show similarly good results. I decided to use #5, cumulative\_cases, new\_cases\_today, and high\_school\_orless, because this one also has a slightly better F1 score.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | **Average sensitivity** | **Average specificity** | **Average F1 score** |
| **1** | **Cumulative\_cases** | **N = 212** | 0.794 | 0.813 | 0.781 |
| **N = 6458** | 0.580 | 0.851 | 0.654 |
| **2** | **Cumulative\_cases, new\_cases\_today** | **N = 212** | 0.839 | 0.839 | 0.823 |
| **N = 6458** | 0.680 | 0.846 | 0.724 |
| **3** | **Cumulative\_cases, new\_cases\_today,**  **age34\_or\_less** | **N = 212** | 0.834 | 0.839 | 0.824 |
| **N = 6458** | 0.679 | 0.848 | 0.725 |
| **4** | **Cumulative\_cases, new\_cases\_today,**  **non\_white** | **N = 212** | 0.793 | 0.839 | 0.794 |
| **N = 6458** | 0.677 | 0.847 | 0.723 |
| **5** | **Cumulative\_cases, new\_cases\_today,**  **high\_school\_or\_less** | **N = 212** | 0.846 | 0.839 | 0.829 |
| **N = 6458** | 0.677 | 0.847 | 0.722 |
| **6** | **Cumulative\_cases, new\_cases\_today,**  **age34\_or\_less,**  **non\_white** | **N = 212** | 0.765 | 0.840 | 0.782 |
| **N = 6458** | 0.670 | 0.850 | 0.719 |
| **7** | **Cumulative\_cases, new\_cases\_today,**  **age34\_or\_less,**  **high\_school\_or\_less** | **N = 212** | 0.781 | 0.856 | 0.794 |
| **N = 6458** | 0.670 | 0.844 | 0.716 |
| **8** | **Cumulative\_cases, new\_cases\_today,**  **Non white,**  **high\_school\_or\_less** | **N = 212** | 0.810 | 0.839 | 0.806 |
| **N = 6458** | 0.671 | 0.845 | 0.718 |
| **9** | **Cumulative\_cases, new\_cases\_today,**  **age34\_or\_less,**  **high\_school\_or\_less,**  **non\_white** | **N = 212** | 0.691 | 0.865 | 0.739 |
| **N = 6458** | 0.648 | 0.848 | 0.703 |

**Compare if tied performances (highlighted) are statistically different**

Using Wilcoxon ranksum test on the metric from each fold

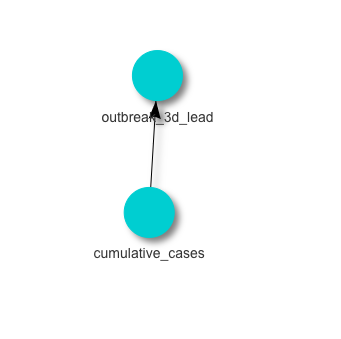
3 vs 5: p = 0.674

2 vs 5: p = 0.674

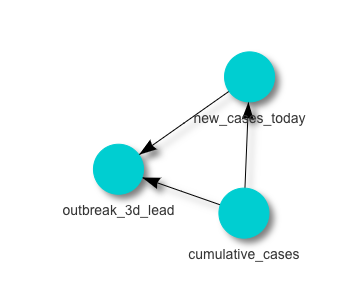
2 vs 3: p = 1

**So they’re not significantly different.**

**Variables: cumulative\_cases**

****

**Variables: cumulative\_cases, new\_cases\_today**

****

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(16, 27)"

[1] "The sensitivity of the fold is 0.556"

[1] "The specificity of the fold is 1"

[1] "The F1 score of the fold is 0.714"

[1] "-----------------"

[1] "Cross validation on fold 2"

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(25, 17)"

[1] "The sensitivity of the fold is 0.882"

[1] "The specificity of the fold is 0.8"

[1] "The F1 score of the fold is 0.811"

[1] "-----------------"

[1] "Cross validation on fold 3"

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(22, 20)"

[1] "The sensitivity of the fold is 0.9"

[1] "The specificity of the fold is 0.864"

[1] "The F1 score of the fold is 0.878"

[1] "-----------------"

[1] "Cross validation on fold 4"

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(21, 21)"

[1] "The sensitivity of the fold is 1"

[1] "The specificity of the fold is 0.714"

[1] "The F1 score of the fold is 0.875"

[1] "-----------------"

[1] "Cross validation on fold 5"

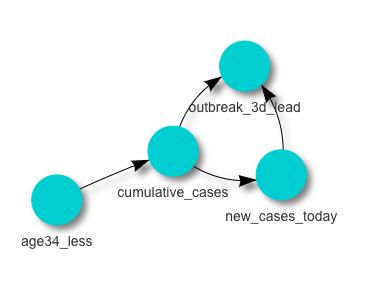
[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set 22:21"

[1] "The sensitivity of the fold is 0.857"

[1] "The specificity of the fold is 0.818"

[1] "The F1 score of the fold is 0.837"

**Variables: cumulative\_cases, new\_cases\_today, age34\_less**



[1] "Cross validation on fold 1"

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(16, 27)"

[1] "The sensitivity of the fold is 0.63"

[1] "The specificity of the fold is 1"

[1] "The F1 score of the fold is 0.773"

[1] "-----------------"

[1] "Cross validation on fold 2"

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(25, 17)"

[1] "The sensitivity of the fold is 0.882"

[1] "The specificity of the fold is 0.8"

[1] "The F1 score of the fold is 0.811"

[1] "-----------------"

[1] "Cross validation on fold 3"

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(22, 20)"

[1] "The sensitivity of the fold is 0.85"

[1] "The specificity of the fold is 0.864"

[1] "The F1 score of the fold is 0.85"

[1] "-----------------"

[1] "Cross validation on fold 4"

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(21, 21)"

[1] "The sensitivity of the fold is 1"

[1] "The specificity of the fold is 0.714"

[1] "The F1 score of the fold is 0.875"

[1] "-----------------"

[1] "Cross validation on fold 5"

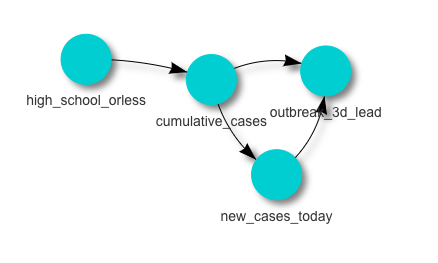
[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set 22:21"

[1] "The sensitivity of the fold is 0.81"

[1] "The specificity of the fold is 0.818"

[1] "The F1 score of the fold is 0.81"

**Variables: cumulative\_cases, new\_cases\_today, high\_school\_less**

****

[1] "Cross validation on fold 1"

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(16, 27)"

[1] "The sensitivity of the fold is 0.593"

[1] "The specificity of the fold is 1"

[1] "The F1 score of the fold is 0.744"

[1] "-----------------"

[1] "Cross validation on fold 2"

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(25, 17)"

[1] "The sensitivity of the fold is 0.882"

[1] "The specificity of the fold is 0.8"

[1] "The F1 score of the fold is 0.811"

[1] "-----------------"

[1] "Cross validation on fold 3"

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(22, 20)"

[1] "The sensitivity of the fold is 0.9"

[1] "The specificity of the fold is 0.864"

[1] "The F1 score of the fold is 0.878"

[1] "-----------------"

[1] "Cross validation on fold 4"

[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set c(21, 21)"

[1] "The sensitivity of the fold is 1"

[1] "The specificity of the fold is 0.714"

[1] "The F1 score of the fold is 0.875"

[1] "-----------------"

[1] "Cross validation on fold 5"

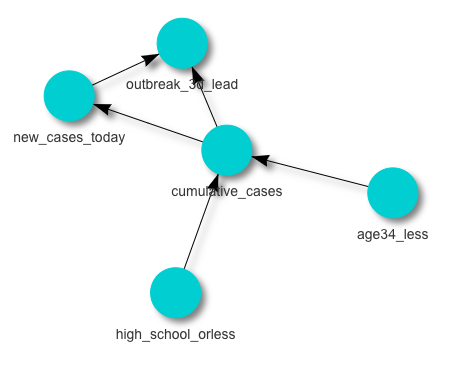
[1] "Target class distribution on testing set 1:2" "Target class distribution on testing set 22:21"

[1] "The sensitivity of the fold is 0.857"

[1] "The specificity of the fold is 0.818"

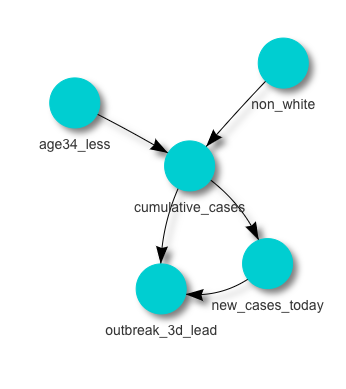
[1] "The F1 score of the fold is 0.837"

**Variables: cumulative\_cases, new\_cases\_today, age34\_less, high\_school\_orless**

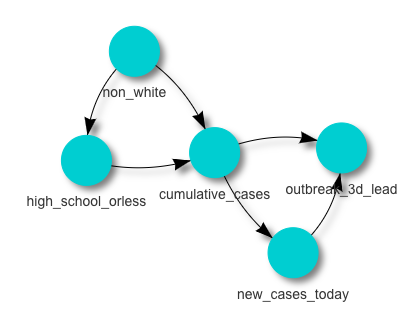
****

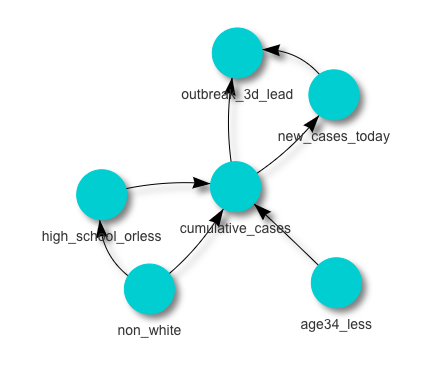
"

**Variables: cumulative\_cases, new\_cases\_today, age34\_less, non\_white**

****

**Variables: cumulative\_cases, new\_cases\_today, high\_school\_orless, non\_white**



**Variables: cumulative\_cases, new\_cases\_today, age34\_less, high\_school\_orless, non\_white**

**Check class imbalance in each fold**

In training and testing sets of each fold, the outbreak YES and NO classes are relatively evenly distributed.

For when N = 6458, (no, yes) in each testing set of the fold: (732, 560), (720, 571), (739, 553), (720, 571), (720, 572)

For when N = 212, (16, 27), (25, 17), (22, 20), (21, 21), (22, 21)