4AL3 Assignment 2 – Question 9 report

A) For the 2010-2015 dataset, the feature set [FS-I, FS-IV] proved to be most effective with a TSS of approximately 0.87. However, this set proved to be ineffective for the 2020-2024 dataset, which yielded a TSS of 0.35. The worst performing feature set for the 2010-2015 dataset was FS-III alone, since it is an array of only zeroes it was unable to make correct predictions, and yielded a TSS of 0.0. In another situation where the historical data was populated, this feature set might have been more competitive. Aside from FS-III, FS-II was the worst performing set with a TSS of 0.32.

B) Because FS-III is a blank feature set, it does not improve the TSS score at all, as no new information is provided. Adding FS-IV was very effective improving the TSS score, as it improved the score of every set it was added to.

Set	Score	Score after adding FS-IV
FS-I	0.72	0.87
FS-II	0.32	0.53
FS-III	0.0	0.74
FS-I, FS-II	0.54	0.67
FS-I, FS-III	0.72	0.87
FS-II, FS-III	0.32	0.53
FS-I, FS-II, FS-III	0.54	0.67

C) With the 'best set' found earlier, the 2010 dataset led to a significantly better TSS score. The 2010 dataset produced a score of 0.87, while the 2020 dataset produced a score of 0.35. I think this is due to the drastic difference in size between the 2010 and 2020 datasets. The 2010 dataset contains 630 datapoints, while the 2020 contains 132: almost 4.8 times less datapoints. Because of its small size, the 2020 dataset was significantly more prone to overfitting and more sensitive to biases in the training data. There is also an increased chance that the variance is low because of the small sample size, which could also result in an inaccurate model.

This also means an extremely small test set during cross-validation of the 2010 dataset: with k=10, the average test set size would be just 13 datapoints! This emphasizes even more the above points, where the overfit nature of the model could prove ineffective in evaluating a certain test set.

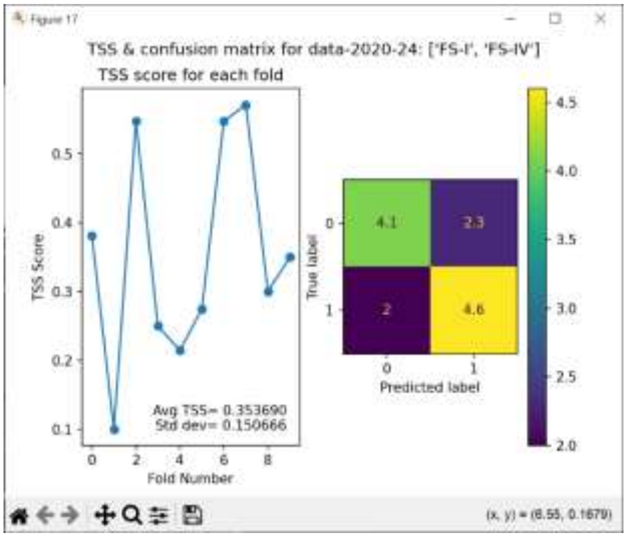
Figure 17
TSS & confusion matrix for data-2020-24: ['FS-I', 'FS-IV']
TSS score for each fold
TSS Score
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
Fold Number
Avg TSS= 0.353690
Std dev= 0.150666
True label
0
1
4.1
2.3
2
4.6
0
1
Predicted label
4.5
4.0
3.5
3.0
2.5
2.0
(x, y) = (8.55, 0.1679)

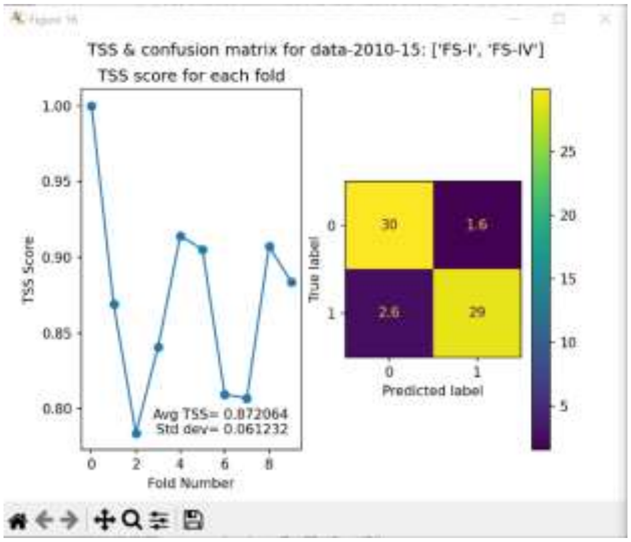
Figure 16
TSS & confusion matrix for data-2010-15: ['FS-I', 'FS-IV']
TSS score for each fold
TSS Score
1.00
0.95
0.90
0.85
0.80
0
2
4
6
8
Fold Number
Avg TSS= 0.872064
Std dev= 0.061232
True label
0
1
30
1.6
2.6
29
0
1
Predicted label
25
20
15
10
5