Assignment II - Classification

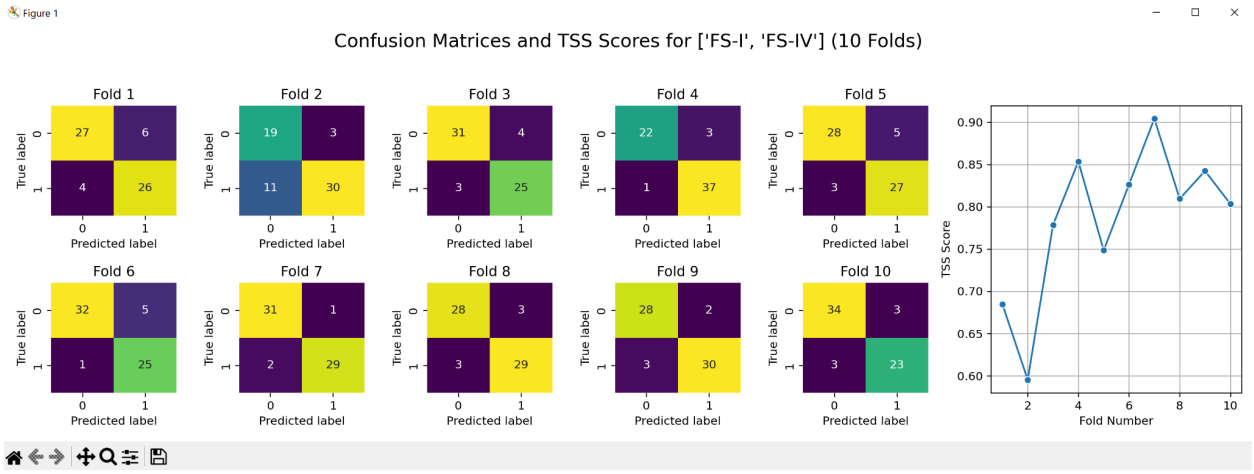
Oct 10, 2024

John Wu

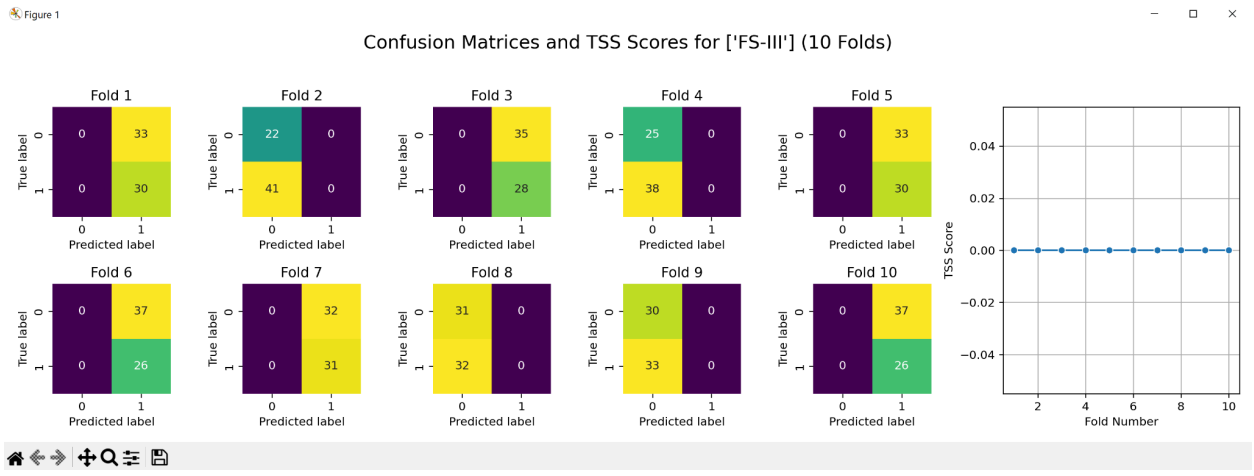
ID# WU103

Which feature set combination worked best and which feature set was worst?

The feature set combination that worked the best was FS-I and FS-IV for me, it has the confusion matrix values and TSS scores per k-fold as illustrated below with mean TSS = 0.78



On the other hand the worst possible feature set is FS-III, as its data values are almost entirely empty. This also means that it has the least ability to contribute correlations to our classification



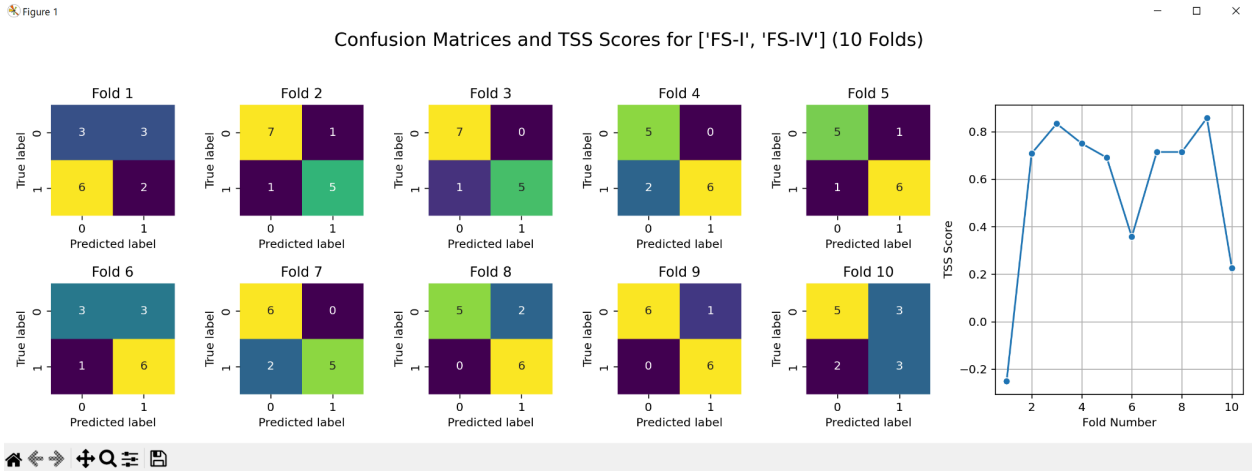
Does adding additional FS-III and FS-IV features improve the TSS score?

Adding FS-IV features to combinations of feature sets tended to improve the TSS score very consistently for almost every other feature set. As such our best combination includes FS-IV. Even when experimenting with different parameters on the SVM any combination with FS-IV leads to the best resulting TSS scores almost every time too. On the other hand, adding FS-III

to any other feature set combination had little to no effectiveness in improving overall TSS score, and from our previous observation about FS-III it would make intuitive sense as to why.

Which dataset led to a better TSS score (2010 or 2020)? What happened?

2010's data set led to the better TSS scores overall and on average when running the same experiment independently using 2010 vs 2020 data. I observed that performing experiments on the 2020 data set had resulted in significantly lower TSS scores and performance overall in comparison to 2010s data set.Taking the best combination of feature sets [FS-I, FS-IV] but using 2020 data also still yielded a lower TSS score than the corresponding best combination of feature sets used from 2010 by an incredible amount too. With 2020's mean TSS = 0.56 for the found best feature combination of [FS-I, FS-IV].



I believe a large reason as to why 2020's dataset performed significantly weaker or poorly is due to the fact the data instances are significantly smaller in 2020's data than in 2010's. Looking at the distribution, despite being evenly distributed, we can see that 2010 has almost 5 times more overall data for classifying positives and negatives compared to 2020. Machine learning models are data hungry, and larger datasets play a critical role in increasing accuracy and allowing models to learn effectively. I would predict that 2020's lackluster data is why its model performed significantly worse than the model trained on 2010's data despite using the best feature set combinations.

Data Set			
2010-15		**2020-24**	
Class	**Distribution**	**Class**	**Distribution**
Positive	315	*Positive*	66
Negative	315	*Negative*	66