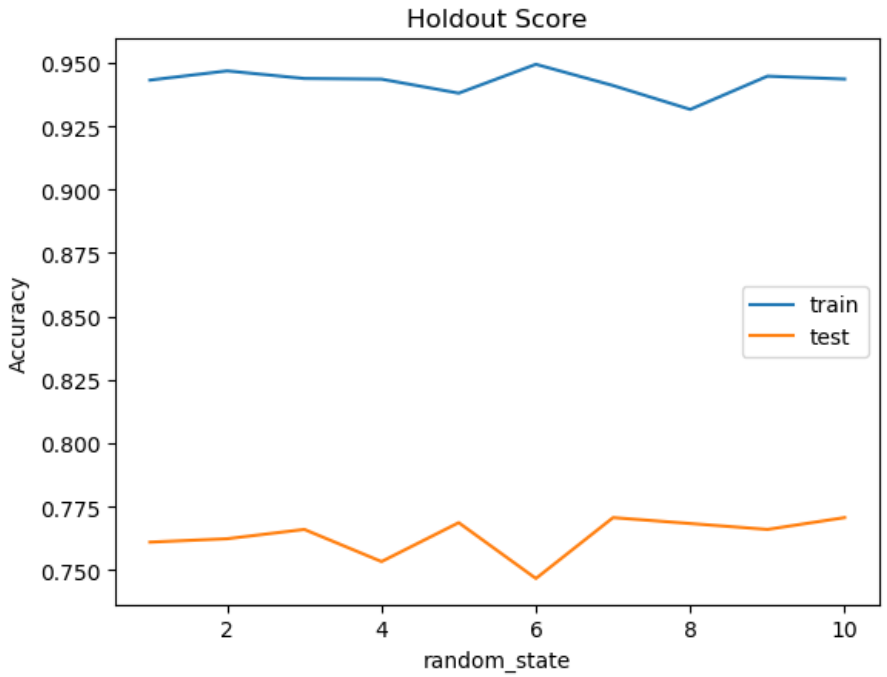
My Name (shixuef2)
IE598 MLF F19
Module 6 Homework (Cross validation)

Part 1 Random test train splits

Individual Scores:

Random_state	**1**	**2**	**3**	**4**	**5**
Train Score	0.9431111111111111	0.9467407407407408	0.9437407407407408	0.9434814814814815	0.9379629629629630
Test Score	0.761	0.7623333333333333	0.766	0.7533333333333333	0.7686666666666667
Random_state	**6**	**7**	**8**	**9**	**10**
Train Score	0.9493333333333334	0.9409629629629630	0.9315555555555556	0.9446666666666667	0.9435185185185185
Test Score	0.7466666666666667	0.7706666666666667	0.7683333333333333	0.766	0.7706666666666667



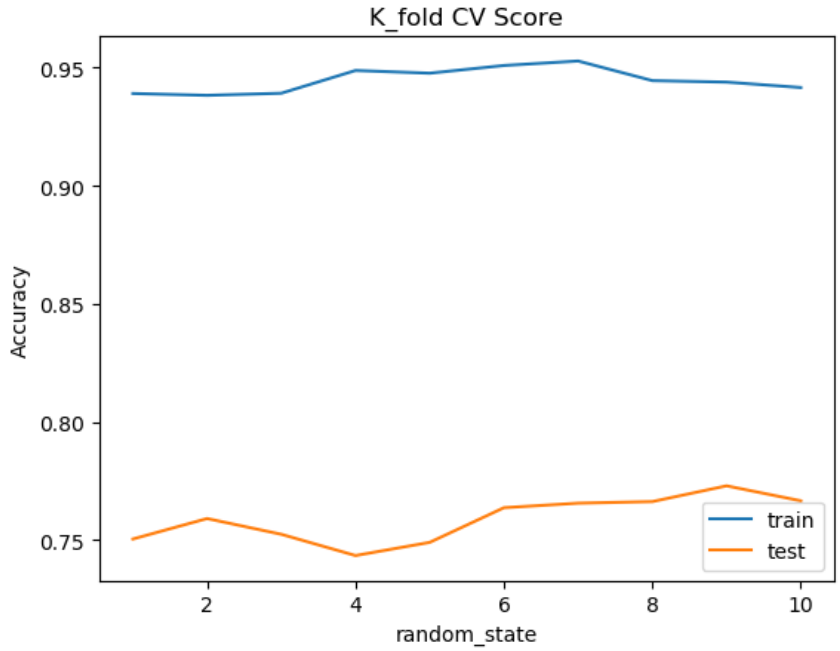
Summary:

Holdout	Mean of Accuracy	Standard Error of Accuracy
Train Set	0.9425074074074073	0.004647729064772772
Test Set	0.7633666666666667	0.007479527614317197
Run Time	6.566829743998824 s	

Part 2 Cross validation

Individual Scores:

Random_state	**1**	**2**	**3**	**4**	**5**
Train Score	0.9389607	0.9381829	0.93907182	0.9487018	0.94755556
Test Score	0.75041653	0.75908031	0.75241586	0.74341886	0.749
Random_state	**6**	**7**	**8**	**9**	**10**
Train Score	0.95081481	0.95270546	0.94440947	0.94377986	0.94148365
Test Score	0.76366667	0.76558853	0.76625542	0.77292431	0.76658886



Summary:

K_fold CV	Mean of Accuracy	Standard Error of Accuracy
Train Set	0.9445666019955137	0.004937950640124042
Test Set	0.7589355341854297	0.009105319718363617
Run Time	6.39438854400214 s	

Part 3 Conclusions

According to the forms above, we can get two conclusions:

1. The average of random test train splits provides the best estimate of how a model will do against unseen data. Since the mean score of test set from part 1 is higher than part 2, and the standard error of test set from part 1 is lower than part 2.
2. Cross validation is more efficient to run according to its less run time.

Part 4 Appendix

Link to my code:
https://github.com/fengzixue96/IE598_F19_HW6/blob/master/IE598_F19_HW6.py

The screenshot: