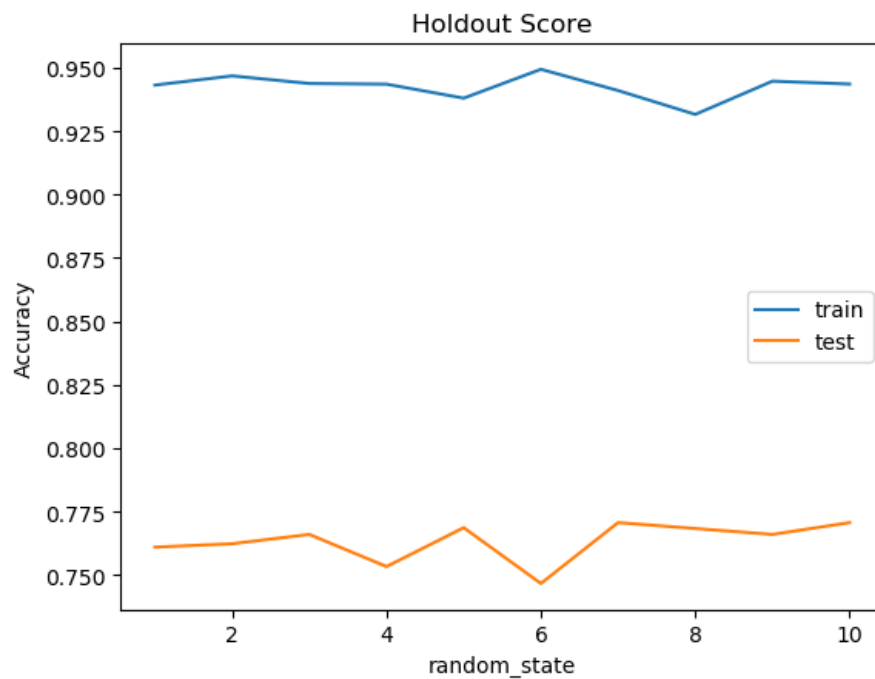


My Name (shixuef2)
 IE598 MLF F19
 Module 6 Homework (Cross validation)

Part 1 Random test train splits

Individual Scores:

Random_s tate	1	2	3	4	5
Train Score	0.94311111 11111111	0.94674074 07407408	0.94374074 07407408	0.94348148 14814815	0.93796296 2962963
Test Score	0.761	0.76233333 33333333	0.766	0.75333333 33333333	0.76866666 66666667
Random_s tate	6	7	8	9	10
Train Score	0.94933333 33333334	0.94096296 2962963	0.93155555 55555556	0.94466666 66666667	0.94351851 85185185
Test Score	0.74666666 66666667	0.77066666 66666667	0.76833333 33333333	0.766	0.77066666 66666667



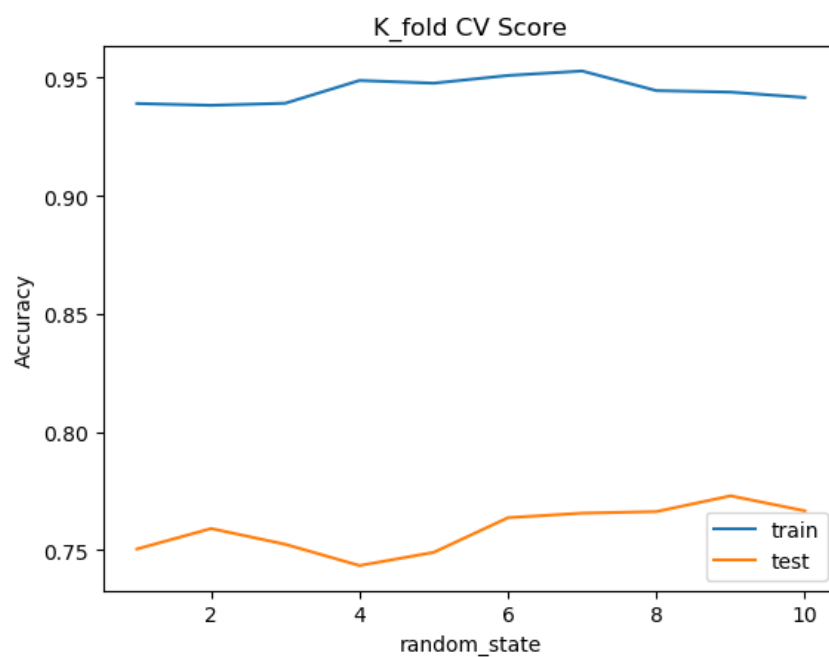
Summary:

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_s tate	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_s tate	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:

The screenshot displays the PyCharm IDE interface. The main editor window shows a Python script named 'HW6.py' with the following code:

```
53 print(scores['train_score'])
54 print(np.mean(scores['train_score']))
55 print(np.std(scores['train_score']))
56 print(scores['test_score'])
57 print(np.mean(scores['test_score']))
58 print(np.std(scores['test_score']))
59 plt.title('K_fold CV Score')
60 plt.plot(range(1,11),scores['train_score'])
61 plt.plot(range(1,11),scores['test_score'])
62 plt.legend(['train', 'test'])
63 plt.xlabel('random_state')
64 plt.ylabel('Accuracy')
65 plt.show()
66
67 print("My name is Shixue Feng")
68 print("My NetID is: shixuef2")
69 print("I hereby certify that I have read the University policy on Academic Integrity and that I am not in violation.")
70
```

The Run window at the bottom shows the output of the script:

```
My name is Shixue Feng
My NetID is: shixuef2
I hereby certify that I have read the University policy on Academic Integrity and that I am not in violation.
Process finished with exit code 0
```

A notification bubble in the bottom right corner states: "You are using Jupyter notebooks. PyCharm Professional Edition supports it." The system tray at the bottom shows the date and time as 16:52 on 2019/10/6.