My Name (myNetID)

IE598 MLF F19

Module 6 Homework (Cross validation)

Using the ccdefault dataset, with 90% for training and 10% for test (stratified sampling) and the decision tree model that you did in Module 2:

**Part 1: Random test train splits**

Run in-sample and out-of-sample accuracy scores for 10 different samples by changing random\_state from 1 to 10 in sequence.

The individual scores are shown in the following table:

|  |  |  |
| --- | --- | --- |
| Seed | In Sample Accuracy | Out of Sample Accuracy |
| 1 | 0.82303 | 0.82367 |
| 2 | 0.82319 | 0.82467 |
| 3 | 0.82196 | 0.83667 |
| 4 | 0.82456 | 0.81067 |
| 5 | 0.82270 | 0.83000 |
| 6 | 0.82411 | 0.81200 |
| 7 | 0.82381 | 0.81900 |
| 8 | 0.82230 | 0.82967 |
| 9 | 0.82356 | 0.81733 |
| 10 | 0.82200 | 0.83267 |
| Mean | 0.82312 | 0.82364 |
| Standard Deviation | 0.00085 | 0.00833 |

**Part 2: Cross validation**

The score of each fold is shown below

|  |  |
| --- | --- |
| Fold Number | CV scores |
| 1 | 0.80997 |
| 2 | 0.80333 |
| 3 | 0.81733 |
| 4 | 0.81100 |
| 5 | 0.81900 |
| 6 | 0.82800 |
| 7 | 0.83333 |
| 8 | 0.83170 |
| 9 | 0.82767 |
| 10 | 0.82367 |
| Mean | 0.82050 |
| Standard Deviation | 0.00957 |

There is no explicit train/test split for cv. The out of sample accuracy is 0.82050, as the mean number shows.

**Part 3: Conclusions**

My original prediction was the cv method may generate a better result, since the label of Y values is highy unbalanced. But the actual calculated data suggests the results are similar. Two methods generates similar mean around 0.82 and similar standard deviation.

If we don’t specify the random seeds and run the split for multiple times, I assume the train/test split method will have a less accuracy result, since its splitting is more random and a bad splitting may mislead the training model.

**Part 4: Appendix**

Link to github repo