Prajwal Kumar Chinthoju (pkc3)

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:

EDA:

Correlation heatmap:

Chart

Description automatically generated with medium confidence

We see that pay 0-5 and Bill amt 1-5 are correlated with each other and hence we can try to use PCA to reduce dimensionality of the problem.

PCA analysis gives the below explanatory variance ratio distribution:

Chart, histogram

Description automatically generated

We use the first 4 components in our analysis going further

**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.

Display the individual scores, then calculate the mean and standard deviation on the set of scores. Report in a table format.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **split no.** | **accuracy** | **f1 score** | **recall** | **precision score** |
| **1** | 0.80 | 0.36 | 0.60 | 0.26 |
| **2** | 0.81 | 0.42 | 0.63 | 0.31 |
| **3** | 0.80 | 0.41 | 0.62 | 0.30 |
| **4** | 0.79 | 0.36 | 0.57 | 0.27 |
| **5** | 0.80 | 0.37 | 0.62 | 0.27 |
| **6** | 0.80 | 0.42 | 0.59 | 0.33 |
| **7** | 0.81 | 0.42 | 0.66 | 0.31 |
| **8** | 0.81 | 0.42 | 0.68 | 0.30 |
| **9** | 0.80 | 0.37 | 0.60 | 0.27 |
| **10** | 0.80 | 0.41 | 0.62 | 0.31 |
| **mean** | **0.804** | **0.397** | **0.620** | **0.292** |
| **standard deviation** | **0.006** | **0.024** | **0.031** | **0.023** |

**Part 2: Cross validation**

Now rerun your model using cross\_val\_scores with k-fold CV (k=10).

Report the individual fold accuracy scores, the mean CV score and the standard deviation of the fold scores. Now run the out-of-sample accuracy score. Report in a table format.

|  |  |
| --- | --- |
| **split no.** | **accuracy** |
| **1** | 0.80 |
| **2** | 0.81 |
| **3** | 0.80 |
| **4** | 0.80 |
| **5** | 0.81 |
| **6** | 0.80 |
| **7** | 0.80 |
| **8** | 0.80 |
| **9** | 0.80 |
| **10** | 0.80 |
| **mean** | **0.802** |
| **standard deviation** | **0.004** |

Out of sample accuracy: 0.8046666666666666

**Part 3: Conclusions**

Write a short paragraph summarizing your findings. Which method of measuring accuracy provides the best estimate of how a model will do against unseen data? Which one is more efficient to run?

Using the cv method with holdout is certainly better as in the other case where we manually split data into test and train with 10 random states doesn’t make use of a holdout set. In cases where there is no hold out set used to check the accuracy, one might argue that the entire data is being used to tune the model and we do not actually test the data against an out of sample data set.

From the cv scores- out of sample case, we have 130 false positives in 3000 test samples (<0.5 false positive rate). The other scores for the holdout set are as depicted below:

f1 score: 0.1465863453815261

precision score 0.10993975903614457

recall score 0.21987951807228914

ROC curve for different tresholds of probability determined by decision tree algoirithm:

Chart, line chart

Description automatically generated

**Part 4: Appendix**

Link to github repo: