COMP47750
Machine Learning with Python

Assignment 2

Pipelines

Feature Selection

1. Rectifying Bias - cross validation

In the first assignment the impact of the strategies to reduce bias was assessed using hold-out testing. The results from this assessment were unstable because of the size of the training data - different train-test splits produce different results. Repeat this assessment using pipelines and cross validation.

Requirements

1. From your first assignment submission, show the hold-out evaluation of the impact of your bias reduction strategy (just one strategy). This will serve as a base-line.
2. Use a Pipeline to repeat this evaluation using cross validation.
3. Comment on the differences between the two evaluations. You may choose to run both evaluations multiple times to show the differences.

Note: Because `imblearn SMOTE` is a slightly odd transformer (changes the size of the training set) it cannot be used with the `sklearn` pipeline. You need to use the `imblearn Pipeline` - it works in exactly the same way.

2. Feature Selection

The objective of this exercise is to assess the impact of feature selection on training and test datasets.

Two datasets accompany this assignment, `heart-train.csv` and `heart-test.csv`. The idea is to identify a good feature subset using the training dataset and test this subset on the test data. In preparing your submission, you should focus on explaining discrepancies between train and test performance rather than maximising performance.

Requirements

4. Use Gradient Boosting as your classifier <link>.
5. As a baseline, you should report performance (accuracy) on the train and test data using all features. The results on the training data can be based on cross validation. The results on the test data can be hold-out, i.e. train on train data and test on test data.

6. Using a feature subset selection method of your choice identify a feature subset that you expect to generalise well for this task. Test the performance of this feature subset on the test data.
7. At no stage should the test data be used in classifier training or in feature selection.
8. In discussing your results, you should comment on the stability and consistency of the results.
9. Your discussion should comment on the insights into the data that can be derived from the overall analysis.

Submission: This is an individual (not group) project. Submission is through the Brightspace page. Your submission should comprise your notebook. Clear all outputs in the notebook before saving for submission. You should use markdown cells in the notebook to report your findings and conclusions.