MISCADA ASML: Classification
Summative Coursework

Lecturer: Dr LJM Aslett

DEADLINE: 2pm, 18th March 2022

The summative coursework for this submodule takes the form of a short report based on applying the techniques you have learned during the course to a real classification data set.

Data set

First, select a data set from among the the approved options listed on the MISCADA ASML Classification summative coursework page on Ultra. If you have another data set you would like to use that is not among those listed, **you must confirm this is ok** with the lecturer before proceeding.

Try to select a data set that you find interesting! The aim is to then explore it and formulate a classification problem to tackle.

There are three parts, two in a written report and the third is reproducible code.

Part 1: Executive Summary (max 1 page, worth 10% of mark)

In this part you should describe in terms that would be *easy for a non-expert to understand*

- some detail about the data set being analysed;
- the real-world objective that your analysis aims to predict/answer;
- a non-technical explanation of how well the finally chosen model performs.

Imagine you are writing this for a friend who is studying a totally different subject! So for the third point, do *not* report technical details such as AUC, true positive rate, etc and instead provide tangible performance numbers and easily interpretable plots. Feel free to be creative here if necessary to define an interesting question (eg assign monetary values to different ground truth/prediction outcomes).

Part 2: Technical Summary (max 4 pages, worth 80% of mark)

Describe *in full technical detail* how you approached the modelling problem and summarise the performance metrics for your final solution. The following list provides some pointers to things you should think about addressing, together with the proportion of marks covering 4 core areas:

10% { **Problem description**, including for example:

- description of the data and the explanation of the objective of the analysis;
- initial data summary;
- simple visualisations of the data;

25% **Model fitting**, including for example:

- any train/test/validate, cross-validation, nested resampling or bootstrap strategies employed;
- the approach taken to fitting, including any design, loss function, early stopping criteria, and algorithm choices;

20% **Model improvements**, including for example:

- what model was finally selected and why;
- summaries of different approaches tried before selecting the final model;
- hyperparameter selection or tuning to improve model fits;
- insights into improvements achieved through different architectures (deep learning), data augmentation approaches, regularisation methods, etc;

25% **Performance report**, including for example:

- details on the performance of the model, including calibration;
- reporting and justification of objective function choices;
- any post-model analysis such as tuning true/false positive rates (e.g would you be more worried about false negatives or false positives in this problem, and how could you address that concern).

- supporting plots for any of the above points.

Do *not* include any code in the written report, see part 3 next.

Part 3: Reproducible code (working web link from report, worth 10% of mark)

The report should contain a link to the code used to generate the results in the report (Github or zip only). It is crucially important that this should:

- successfully run on any R 4.1.2 or later installation (so include `install.package` commands for any non-base packages);
- all plots and values included in the report should be reproducible from the submitted code (fix random number seeds to help with this);
- include a master file named `report.R` which will be run.

Some part of your code will be run for marking, but *no attempt* made to fix anything that does not work, so make sure it runs unaided. Further guidance will be in a lecture video on Ultra.

Submission details

The final report should be a **maximum of 5 A4 pages total in PDF** (any work beyond 5 pages will not be marked). Submit a PDF file with a link to the code used to generate the results. **Remember to add your anonymous student number (Z-code) to the submission.**

Regulations

Please remember that plagiarism and collusion are taken very seriously by Durham University. All work you submit should be your own. See official policy on the MISCADA DUO pages.