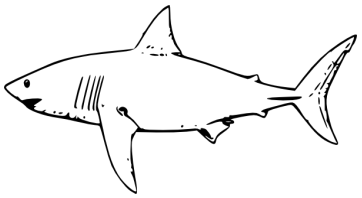

Machine Learning Assignment, Summer 2022

Computing and Maths, University of Stirling

Sydney City Council (SCC) is an Australian government agency responsible for the operation of one of the city's beaches. There have been a number of recent shark attacks at Australian beaches and SCC are keen to minimise risk to the public. They want to be able to predict whether a shark will appear at their beach. With this information, they will choose to close the beach or keep it open. While they do care about possible attacks, they also care about lost revenue to the local economy when the beach is closed. The data tells us if a shark appeared under certain conditions at a beach - under the column "shark".

Your assignment is to use machine learning techniques to produce a system that would be able to tell SCC if a shark appearance is likely.

You can use Orange, Python, R, or any machine learning package of your choice. The data for the assignment is in a file, *shark.csv,* provided for you. A data dictionary, *datadictionary-shark.txt*, is provided describing the columns in this file.

You can post general questions about the assignment in the Teams "Assessment Questions" channel.

Please note that this is an individual assignment, not group work. See the section on academic misconduct at the end of this brief for more details.

<u>Requirement</u>
You should submit a report describing the modelling process you followed and your results. You should try to frame the problem in the form of a CRISP-DM framework to better facilitate the discussion. Refer to the relevant CRISP-DM stages at each stage of your report. You do not need to submit code or data. The report is worth 100 marks in total and must cover the following (with weightings per section as shown):

Business Understanding [10%]
Describe the task you were given: is it clustering, classification or regression?; describe the data you received and the requirements of the finished system, including why machine learning is suitable for this task. Define any terminology that you will use in the report (for example, model, variable, task, etc.). Comment on any issues around ethics or trust, such as model transparency, that may be relevant to the framing of the problem.

Data Understanding [10%]

List the variables that you found in the file provided by the company. For each one, say whether it should be treated as categorical or numeric; nominal, ordinal, continuous or discrete; and whether or not it is likely to be of use in building the solution. Explain your decisions: if you rule out any variables at this stage, you can justify your choice using summary statistics, or a histogram plot of its distribution.

Data Preparation [10%]
Describe what you did with the data prior to the modelling process. Show histograms of the data before and after any pre-processing that you carried out. (you do not need to give histograms of all variables, just the ones that need some cleaning) If you corrected any mis-typed or corrupted entries in the data, report what you changed, such as any rules you used, or examples of specific data points that were cleaned.

Modelling [30%]
You must use three different techniques and build models with each: these should include one tree-based model, one based on logistic regression, and one based on neural networks. Try to make each model perform as well as it can: if you varied the hyperparameters of a model, show which hyperparameters you varied and how this impacted on the results. Describe how you split the data for training, validation and testing purposes. Be methodical and record each result. This stage is a little like scientific research – you are carrying out experiments in your search for the best solution. Once you have a solution, show how you verified its robustness. For the three different techniques report on their comparative ability to make predictions for this problem, but only select a single model for the final test.

Don't try to find a perfect or extremely accurate model - one does not exist! We are interested in the procedure you followed and the justification you give for choosing particular model types/parameters/features, rather than a specific accuracy score.

Evaluation [10%]
Analyse and describe the level of accuracy the model achieves and the errors your model makes. Show a confusion matrix for your model. Are there any areas of the data where it performs worse than in others, and are there any types of error that the company would want to avoid more than others? Show a lift curve or a ROC curve for the model's predictive capability, and explain what this tells you. Comment on whether the modelling approach chosen and the results can be trusted by the company.

Overall Approach and Creativity [30%]
You should adopt a structured approach to the whole process and clearly identify the five CRISP-DM stages excluding deployment in your report. Additional marks are available within this category for going beyond the basic approaches mentioned in the sections above and those covered within the course; for example, you might consider feature construction or adding some explanation to the predictions the model makes. Those showing (per the common marking scheme) originality and exceptional analytical, problem-solving and/or creative skills will be awarded the highest marks.

Submission
You do not need to submit the models you built, just the report. The deadline for this assignment is 4pm on Friday 19th August 2022. Please submit your report via Canvas at https://canvas.stir.ac.uk/courses/12173/assignments/90875. You should save it as a

doc or pdf file bearing your university portal username (3 letters + 5 digits, e.g., xyz00001.pdf).

Just write what you need to provide the required information clearly and concisely. You can assume that the client has a good technical understanding of data mining and statistics, so do not avoid technical terms in your report, but where you use them please explain what they mean in plain language too. To maximise your mark, closely follow the instructions you've been given.

Marking Scheme

Submissions will be marked in alignment with the University Postgraduate Common Marking Scheme, which can be found here: https://bit.ly/3sztOIt

Word Count Limit

For reports that exceed the 3500 maximum word count by:

- Less than 5% (less than 175 extra words), no marks will be reduced.
- From 5% - 10% (176 to 350 extra words), the mark will be reduced by five percentage points (5 marks).
- More than 10% (351 or more extra words), the mark will be reduced by ten percentage points (10 marks) for every 350 extra words.
- Word count penalties will not reduce the final mark below the passing threshold (50 marks) so if your mark was 52 and you have a 5% (5 marks) penalty, your final mark would still be 50.

Note on Avoiding Academic Misconduct:

Work which is submitted for assessment must be your own work. All students should note that the University has a formal policy on Academic Integrity and Academic Misconduct (including plagiarism) which can be found at https://bit.ly/37fYxPw

Plagiarism: We know that assignment solutions by previous students can sometimes be found posted on GitHub or other public repositories. Do not be tempted to include any such code or writing in your submission. Using work that is not your own will be treated as "poor academic practice" or "plagiarism" and will be penalized.

To avoid the risk of your own work being plagiarised by others, do not share copies of your solution, and keep your work secure both during and after the assignment period.

Collusion: This is an individual assignment: working together with other students is not permitted. If students submit the same, or very similar work, this will be treated as "collusion" and all students involved will be penalized.

Contract cheating: Asking or paying someone else to do assignment work for you (contract cheating) is considered gross academic misconduct, and will result in termination of your studies with no award.

Late submissions: This assignment is subject to the usual grade penalties for a late submission. You can email questions about these to alexander.brownlee@stir.ac.uk.