Project Guidance

The aim of the project is for you to take the dataset your prepared in the first coursework and perform an appropriate statistical analysis project. You should aim to build an well optimised statistical model or predictor. The purpose of the report is to explain your data modelling and the choices and to explain your results and conclusions.

Code Guidance

We would expect the code you submit to handle the following concerns

1. Loading the data in to R
2. Preprocessing the data for the planned modelling. This may include things such as rescaling variables, creating dummy or aggregate variables
3. The statistical modelling, model optimisation and benchmarking
4. All code that produces supporting material for the report such as tables of numbers and figures

Importantly your code should run in a base install of R Studio. If you have chosen to use libraries or packages that are not installed by default you should make sure to include the appropriate install.packages() commands at the top of the code. You may make use of the modules that were introduced during the course, though these should be installed as needed. The exception is the Keras package. If you are working with neural networks you may assume that Keras is available for R and that your code does not need to handle installing this. You should **not** make use of GPU acceleration with Keras

We require that you submit a dataset with which we can test your code for and with which your code will run to completion. We ask that this dataset be **no more than 100Megabytes** and that your code with the submitted dataset have a runtime of **no more than 15 minutes.** You may have to submit only a sample subset of your dataset but your code must run correctly to completion on this sample. Your report should be based on the full dataset you used and your personal work can have any run time you deem appropriate to complete the most appropriate model and model optimisation. If they data in your report and submitted data are different, you must explain in your report how the dataset you used for modelling differs from the sample you submitted. We accept that a sampled dataset may not produce the same results and figures as the data used in your report.

Code should be clearly written to a consistent style. For instance (https://www.r-bloggers.com/2018/09/r-code-best-practices/)

Report Guidance

The point of the report is to explain your analysis and the conclusions you have drawn from them

We would expect the following concerns to be addressed in your report’s

1. Abstract. A brief summary of your modelling, about 200 words. You should use this to highlight your main result or conclusion.

2. Introduction. Use this section to introduce the context of your analysis. Prior academic work with your dataset and modelling techniques should be explained here. Algorithms that you use should be clearly summarised here. You should also use this section to clearly explain the hypothesis of your study. Appropriate references should be included as per a research paper.
3. Methods & Materials. In this section you should clearly explain your work such that a motivated and experienced R programmer can replicate what you have done. Explain you dataset and its preparation also explain all the steps in your modelling and analysis. You should explain any important metrics that you have gathered or measured in your results, this is especially critical for any metrics you formulate yourself.
4. Results. In this section you should present your results, the outcomes of any data experiments and optimisation studies you have conducted. This section should be supported with appropriate tables and figures
5. Discussion and conclusion. This section is to allow you to reflect on the work. What can you conclude from your results. What synthesis of new knowledge can be made from the information and facts presented in your results? Take time to comment on the benefits and drawbacks of your approach. You should also discuss what improvements to your modelling could be made and what future work or analysis you would do next.
6. Bibliography. You should cite appropriate papers, texts and material in your report. These should be formatted to a standard Reference Style and the references gathered in this section. You should use Harvard referencing style commonly used in academic work.