SMM284 – Applied Machine Learning

Group assignment

This assignment covers 100% of your course grade and is due at 5pm on Monday 18th July. Due to school regulations, it will not be possible to extend this deadline, so please ensure you submit on time. The assignment is conducted in groups of up to five and standard school peer assessment practices apply.

For your assignment you are asked to first identify a dataset on Kaggle.com that is relevant to your area of study interest. Check that the dataset has sufficient variables and data points to be useful for analysis. Check the usability score to determine if Kaggle has ranked the dataset as sufficiently robust.

Your assignment then is to:

1. Clean the dataset, providing notation describing the choices made.
2. Apply *your interpretation* of at least one analysis from each of the module's Classes 2-6 to the dataset. There should be a **table** provided in the introduction which maps how the material from each class has been mapped to the analysis provided in your report. If there is a particular reason why the material from one of the Classes of the course cannot be carried out for your dataset, then explain this in the Table. A focus should be on the accuracy obtained from your machine learning code compared to naïve or baseline models.
3. Each of the analyses should consist of a set of code, the output, and an interpretation and implications from your analysis.

The assignment should **only** be submitted in the form of a Python worksheet plus a PDF printout of the Python worksheet with all output displayed. You can use 'markdown' boxes in Jupyter/Colab to type any textual analysis. I would expect about 3,000 words of written analysis, including the introduction describing the dataset, but won't be counting the words. References (please concentrate on practice references, or empirical academic references if suitable) can be provided in a markdown box at the end of the worksheet. In-text hyperlinks are also fine for practice-based references, such as reports.

I will grade your assignment using standard Bayes Business School rubrics, paying particular attention to the extent to which you have built on the codes learned in the class and provided your own interpretation of these, as well as the business-relevance of the analysis.

Honesty: Most databases on Kaggle also provide code where either the uploader, or others, have created their own testing models. Please do look at these. It's a great way of learning. You can use snippets of these codes if you wish and find the code to be particularly well-created, but do not use extensive amounts of the codes. The coding you use should primarily build on the course material, the course textbooks, and your own interpretations of these.