**ST3002 Final Project 2019/2020**

*Submission of your work:* This assignment is weighted as 60% of your module mark. Your final program file (.R) and report (.pdf) should be submitted as a single zip file in the Project section of Blackboard **no later than 11:59 pm on Sunday 5th April 2020**. Please name your files as follows:

code\_<NAME\_WITH\_NO\_SPACES>\_<YOURSTUDENTNUMBER>.R

report\_<NAME\_WITH\_NO\_SPACES>\_<YOURSTUDENTNUMBER>.pdf

**Reports which are not submitted in pdf format will not be accepted**. **Please note that** **no late assignments will be accepted**.

**Project description:**

An online news portal has hired you as a consultant. They have carefully collected and compiled data on a number of their articles. This data contains information on 39 potentially important attributes for popularity of articles. It also includes another variable, which records the popularity of each article (shares on social media). The data file has encoded categorical variables as individual and sometimes composite binary “dummy variables”, for a total of 54 columns in the dataset.

You have two tasks.

1. To come up with a way to predict the number of shares an article will receive using all or some subset of the attributes provided. You should demonstrate how well your prediction works by splitting the data into training data (the data used to fit the model) and test data (the data which is used to predict from the model).
2. The editor of the World Section is interested in whether the articles that are published in that section on Thursdays fall into defined groups, and if so, what distinguishes one group from another. You have been asked to investigate this question.

The way in which you approach these tasks is entirely up to you as a consultant, but you should be able to defend your approach.

After you have analysed the data, you should make a short professional report (7 pages max, *including* any figures) for the online news portal outlining for task a) your model and your testing and for task b) your answer to the editor’s questions.

**Files:**

There are two files in the project folder; STU33002ProjectDataset gives the actual data and STU33002ProjectDataDescription gives a description of all of the attributes.