

# Competition description

Oliver Hinder

The files in the course content repo:

- i. Assignment/competition/competition-data.csv
- ii. Assignment/competition/competition-test-x-values.csv

are the training set, and test set respectively. The columns of this dataset are `X1, ..., X21, outcome`. The goal is to predict `outcome` using `X1, ..., X21`. You should use `competition-data.csv` to build your model. You should submit a csv file called `competition-test-outcome.csv` containing a single column of predictions for the outcome variable for the `competition-test-x-values.csv` dataset. Only the instructor has access to the corresponding outcome values which will be used to evaluate your performance.

You should create a git repo with your code that produces `competition-test-outcome.csv` and add me as a collaborator. If there are multiple files then there should be a readme giving a brief summary. Please note that I will look at the git history to ensure that all team members made meaningful contributions to the repo.

Hint: if your team are novice git users then it may be best to work on separate files in the repo to avoid conflicts.

The assignment will be scored on two criteria:

1. Quality of the code in the git (3 points).
  - a. Is the code readable?
  - b. Were sensible models tried?
  - c. Does the revision history contain descriptive changes?
2. Quality of the predictions on the **test set** in terms of root mean squared error (RMSE). The scores will be calculated as follows
  - i.  $RSME < 12 = 3$  points
  - ii.  $RSME < 8 = 4$  points
  - iii.  $RSME < 5 = 5$  points
  - iv.  $RSME < 3 = 6$  points
  - v.  $RMSE < 2.8 = 6.5$  points
  - vi.  $RSME < 2.6 = 7$  points

Note that you will not know your exact score for this question until the assignments are graded. However, you should be able to use the validation set to get a good estimate.