# Data Science Toolbox Assessed Coursework 1: Supervised Prediction

**Deadline: Wednesday Noon, Week 12**

## Group Project description

Comparing model performance is an essential part of data science. You will **choose an application domain** that your group will work with during Assessment 0-1.

Your task for this project is to:

- Consider a **binary classification problem**;
- In which each group member will create a **model submission** that can be evaluated on **left-out test data**;
- together agree and test a **performance metric**;
- compare your models according to that performance metric.

Classification is an open-ended problem and you are invited to interpret this aim in a way that you find interesting, tractable and productive. In particular, it can be valuable to deploy traditional mathematical models to understand how they relate to real-world problems.

Remember that the goal of this is **not** to win, but to learn about the appropriateness of the model. An additional goal is to consider the appropriateness of the performance metric. You should create a test and validation dataset, but you may choose how to do this. You may also choose to limit the model to certain covariates.

Half of the effort should be devoted to exploring appropriate performance measures. Think about the circumstances by which your chosen performance metric will lead to real-world generalisability, and how it might compromise this for the purpose of standardization. Demonstrate this with data and/or simulation; for example, if you believe that you can predict **new** types of data, you could demonstrate this by leaving out some types of data and observing your performance. Examine in what sense your group's best method is truly best.

### Appropriate Methods

Your models may include off-the-shelf methodology, including but are not limited to:

- Standard regression techniques such as:
  - transforming the data;
  - creating additional covariates by additional modelling;
  - using variable selection;
  - using penalisation;

  - using non-linear link functions;
- Using more sophisticated forms of regression or classification that you have researched yourself;
- Using models from your personal experience, such as:
  - Time-series models,
  - Point-process models,
  - Multivariate models,
- Any of the advanced topics that are referenced in the Data Science Toolbox lectures.

**Advice on assessment**

You will be assessed on:

a) whether the model implementation is appropriate, that is:
   - you can be awarded credit for additional implementation if an off-the-shelf implementation falls short,
   - you can be awarded credit for exploring multiple implementations,
   - you can be awarded credit for examining the mathematical details of choices.

b) how well explored the model is.
   - you can be awarded credit for simulation work.
   - you can be awarded credit for sensitivity analysis.
   - you can be awarded credit for plotting or otherwise describing various inputs, outputs, or parameters.

c) the correctness of the methods used to achieve their stated goals.

d) the robustness of the results in supporting the conclusions.

e) whether, after the fact, you can explain the limitations of the approach taken.

You do not need to excel in all areas in order to get a high mark. Instead, you need to perform robustly in all areas and additionally demonstrate insight somewhere to score highly.

You will not be penalised if:

- you choose a model and after testing, the implementation proves deficient,
- your chosen model performs poorly, provided that you have implemented it correctly and understood its limitations,
- the implementation makes it impossible to work with the full dataset, and you have to create a smaller artificial comparison dataset. (you may still be penalised if the model could have been simply fit a different way.)

## Topics

Here are suggested example datasets. You are free to find and analyse other datasets on discussion with the course tutor.

- **Cyber Security**: predicting **normal** vs **non-normal** traffic in the KDD99 **(small, 10%) dataset** called `kddcup.data_10_percent.gz` from the week 3 workshop. You are welcome to use additional data from the KDD99 experiment, but this is not required.
- **Network data** based on the Enron email dataset. For example, you could look to predict the next email an account will send, or model the network at a more global level, etc.
- **Epidemiology data**
- Or another topic agreed between the lecturer and your group.

## Individual reflection description

- Discuss the rationale behind the model and/or data extensions that you put into your own model;
- Discuss the performance metric that you chose;
- Discuss the mathematics behind your method;
- Reflect on how you might change your own model, were the intention to "win";
- Reflect on how this could work in a competition setting.

## Coursework guidance

**This section is the same for every coursework.**

### Submission

**Every group member must submit** something to Blackboard, by the deadline. You must submit an **Individual Reflection** and a **Group Report**.

It is recommended to upload the Reflection directly to blackboard, and in the notes add a link to your Report repository.

You should submit a Report Repository containing:

1. **README.md**: An explanation of:
    - **Project Group**: List who was in the group.
        - It can be helpful to describe briefly what their contribution was, here or in the reading order. Any **Equity** variation should be clearly noted.
    - **Reading order**: the order that your files should be read in, which should explain:
        - Preparation: how to install any packages or software etc that should be installed.
        - Your report content reading order (if there are multiple files), with any additional info about the file you feel appropriate (e.g. if only some of the team were an author; the purpose of the file such as "data downloading", etc)

2. **report/**: a FOLDER, containing all of the files that will be read as part of assessing your project.
    - Label these alphanumerically as "<number>-<name>.<file ending>" for the reading order.
3. **Documentation folders**:
    - Each member of the group should be using the repository to work on the project. Make **one folder** per group member, name and merge your content into the report as possible. This will act as documentation that you have contributed to the project.

There is an Example for you to emulate, with the structure:

- README.md
- report/
    - 01-Data.Rmd
    - 02-R_analysis.Rmd
    - 03-Python_Analysis.ipynb
    - 04-Wrapup.Rmd
- RachelR/test.Rmd
- PeterP/work.ipynb

**Assessment**

- 75% of your mark will be for the group project itself. All students in a project should submit the same project; only one project will be run. The individual marks may be moderated away from the group project mark.
- 25% of your mark will be for an individual reflection, which should be written by you. It should be approx 500-800 words (not strict) which should be individually written.

## Report

All coursework for this unit is based on group work in teams of around 3. Your team will address a single data science challenge. You will have choice about the topic, within the remit of the project description. It is always the intention that you each learn from, and teach, your teammates any skills you can bring to bear on your chosen problem. Your team will submit a single project report, which is a script that can be run to a) obtain data, b) analyse data, and c) produce any figures and tables that you feel are illuminating.

Your project script would **typically** take the form of an **Rstudio markdown** project or a **Jupyter Notebook**. It should be annotated with factual statements describing what you have done and why in basic terms. Unless otherwise stated, you may choose the programming language but we recommend sticking with python or R since all students are expected to become familiar with these. The results of computations including plots should be displayed and labelled (e.g. with numbers) and if you have not used a seamless method then you must provide a zip file containing both a script, and a pdf or similar document that

also contains the output of your script. Your script is expected to run, and if at any stage some manual step is required (for example, to wait for a bluecrystal job submission to finish, or data must be downloaded) this should be carefully noted. You may lose marks if your script needs debugging.

There is no word, page or other limit. Credit will be awarded for making your arguments thoroughly but without repetition or meandering off-topic. Only include material that you feel makes a contribution to the overall project scope. If some research led to a dead end, work it into the results.

Remember to **reference** websites and other resources for content and ideas, in addition to the usual academic referencing. This will assist you in your future projects.

### Report Assessment Criteria

Your project will be assessed against the following criteria:

- Fit and Success.
  - *Choice of project is very important for learning. This section includes finding good datasets and matching them to questions; making progress on hard problems; fulfilling the learning outcomes.*
- Innovation.
  - *Thinking outside of the box and finding resources that are not presented in the course. Innovation can come in the form of data, methods, and mathematical ideas brought in from elsewhere.*
- Citations, Referencing, Literature.
  - *Cite your sources, build up a repertoire of useful content. Link your results to those on analogous problems. Note that many resources are not published papers.*
- Structure and Description.
  - *Your project should be well introduced, and easy to read and undestand. Make good figures and explain them. Structure your project well, stick to the point and note what your results mean.*

### Equity

Your team should try to agree an **equity** or proportional contribution to the group project, accounting for both practical (implementation) and conceptual (theory, methods choice, etc) contributions. If you cannot agree, you should approach the tutor to try to agree equity before submitting divergent opinions. Try to agree any non-even equity before the project gets underway.

Contributions will be taken into account when assigning individual marks from group reports. Small deviations are unlikely to be given divergent grades.

Individual grades can be moderated up and down based on equity but are unlikely to be increased as much as they are decreased, and the final decision takes into account documentation.

Additional notes:

- It is expected that all group members understand the group submission.
- It is also the intention that they put in equal effort.
- It is not expected that the final script contains content proportional to equity. There are many good reasons that work does not make the final report.
- If you put in lower effort and agree a lower equity, you may receive a proportionally lower group mark.
- If you put in extra effort and agree a higher equity, you may receive a higher mark but the reward is not linear. It is better to have an equal share of a good project, than a high share of a poor project.
- Mathematical contributions and programming contributions can be considered. All contributions should be documented.
- If some people choose lower equity because they could not contribute fully, make this clear in all reflections. The lower manpower may mitigate a low grade.

**Documentation**

All students are expected to contribute to programming. You should each submit your own scripts, session history or similar, that demonstrate that you made some independent effort, even if these did not make it to the final report. If you cannot demonstrate an amount of effort commensurate with your claimed equity, then your mark may be reduced.

Your documentation is likely to take the form of an Rstudio markdown or Jupyter Notebook. It can be long and contain dead ends. It does not need to be documented, nor be able to run from top-to-bottom. It should be unique to you, though is likely to contain content from others' notebooks. You may refer to it in your individual reflection, but if there is excessive material that should have been shared with the group then you will not receive credit for it. You should not try to boost your individual grade by doing extra work here. It may not be carefully read and you may not receive feedback on it. It should be no additional effort to produce this as it should consist of files that you already have.

## Individual Reflection

The purpose of your reflection is to encourage changes in your practice that improve your understanding of Data Science, as well as improve your ability to work in a team on Data Science projects.

You are being assessed on your progression and understanding of the content of the project. It is better to note deficiencies with what you have done, than to try to post-hoc justify something. It is understood that you are under time pressure and may make a poor irreversible decision for the project performance, but that will not strongly affect your mark if the reason for the failure is clear.

You must write your writeup independently of the other students, though using the shared understanding gained from working with them.

### Reflection Assessment Criteria

Your reflection should always adress the following areas in addition to what is asked in the specific project description.

- Fit.
  - *Introduce the area and explain the overall goals.*
  - *Justify the decisions made in the project.*
- Depth.
  - *Explain the results and discuss the conclusions. This can focus on your contributions but should also include the project as a whole.*
  - *Briefly explain some aspect of the mathematical model(s) that has been used.*
    - *It is expected that your group will discuss this in detail, and that contribution of understanding is included in the project contributions.*
    - *Each student still must write something in their own words.*
  - *Reflect on the strengths and weaknesses of your approach, and how you might do it differently next time.*
  - *Reflect on any aspects of the project that could be improved, paying particular attention to* **group working**, *technological barriers and solutions.*
- Structure and Description.
  - *Write clearly, use referencing if appropriate. This is a reflection, not a report, so use appropriate language.*
- Evidence.
  - *Your documentation and reflection together are assessed to evidence your individual contribution. Explain this and how it fits into the whole.*

### Learning outcomes

You are reminded that:

- Teamwork is a learning outcome.
- Progress, teaching and sharing is more important than individual technical ability or project success.
- The difficulty of these assessments is beyond what would be expected of an average student alone.
- Most groups will contain a mixture of expertise which should be exploited.
- In the event that your entire group is inexperienced at programming, you still need to meet a minimum standard. However, you can still score well if you focus on a mathematically interesting question.

## Working practice

You should work on this project together. This may mean all group members trying different things and coalescing on a final approach. Trying things that fail is still a contribution. Failure can be included in the report if something meaningful was learnt.

To work physically separately, you should:

a) arrange a suitable **discussion forum** for your group such as a WhatsApp group, slack, etc.
b) arrange a suitable **file sharing location** such as github, OneDrive, Dropbox, or GoogleDrive.
c) Collaboratively decide the final content, merging all versions of the analysis.

You should finalise the project content at least 48 hours before the deadline, so that individual writeups can be written.