

# Data Science Toolbox Assessed Coursework 2

**Deadline: Friday 17:00 Week 8**

## Group Project description

Comparing model performance is an essential part of data science.

Your task for this project is to:

- each create a **model submission**, with the goal of predicting **normal** vs **non-normal** traffic.
- together agree and test a performance metric.
- compare your models according to that performance metric.

Modelling normal versus non-normal traffic is an open-ended problem and you are invited to interpret this aim in a way that you find interesting, tractable and productive. For example, you could look to predict the next event on each edge based on past events on this edge, or you could model the network at a more global level, and many other approaches are possible.

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.

Your models should be off-the-shelf methodology, but should differ by considering extensions of different forms. These can include but are not limited to:

- transforming the data;
- creating additional covariates by additional modelling, for example clustering or creating attributes for factors such as the average of a covariate for a given IP-pair;
- using variable selection;
- using penalisation;
- using non-linear link functions;
- using more sophisticated forms of regression or classification that you have researched yourself.

You should consider the KDD99 (**small, 10%**) **dataset** called kdd-

cup.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.

The models may each be written in a separate language.

### 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.**

#### In brief

You should submit:

1. **Report:** A team report, shared across your team.
2. **Equity:** A proportion of the team report that you each believe you have contributed.
3. **Documentation:** An individual supplement demonstrating the work that you have done that led to the production of the report.
4. **Reflection:** An individual writeup describing the report content in more thought.

#### 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 500-800 words (10% grace is allowable) which should be individually written.

#### Report

As with all coursework for this unit, you will work in your assigned team 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.

Remember to reference where content and ideas come from, in addition to the usual academic referencing. This will assist you in your future projects.

## 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.

### **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. 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 Assessment Writeup**

You are being assessed on your 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.

The individual writeup should:

- Briefly explain 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 about it in their own words.
- Justify the decisions made in the project.
- Explain the results and discuss the conclusions.
- Reflect on the strengths and weaknesses of your approach, and how you might do it differently next time.

### **Recommendations**

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.

If you want 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, Dropbox, or GoogleDrive.
- c) Get together to 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.

### **Learning outcomes**

You are reminded that:

- teamwork is a learning outcome.
- 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.