Group Project for SOWK3136

Hao Luo

Submission deadline: Dec 5, 2021

The goal of COMP3136 is to equip you with the knowledge and skills needed for applying basic statistical learning/machine learning models to real-world problems. The group project is intended to lead you to the direction of real-world exploration.

The project should be completed in small groups of 3-4 persons. You need to do a oral presentation (10%) on Nov 22 and submit your final project report (30%) by Dec 5, 2021.

Your first task is to discuss with your group partners and choose an application that interests all of you. After deciding the topic, the task is to conduct exploratory data analysis and apply appropriate statistical learning techniques to build predictive models. Explore how we can improve the model and discuss the advantages and disadvantages of candidate models.

Some tips:

- Once you have identified a topic, make sure to search for existing literature on relevant topics using Google Scholar or other databases. Since this is an introductory level course, it is OK to "reinvent the wheel". However, it might be more fun to try something new.
- You need to identify one or several datasets that is/are suitable for your topic of interest. There are a lot of very good datasets that are publicly available. Here I list some of the famous ones.
  1. Kaggle
  2. Wisconsin Longitudinal Study
  3. American Time Use Survey (ATUS)
  4. Programme for International Student Assessment
  5. ...you tell me :)

And of course...you can crawl the web and create your own data.

We will not breakdown the 30% that the group project is worth. Instead, the presentation and project report will be evaluated holistically based on the originality, significance, and rigour of the work.

Instruction

Both your oral presentation and final report should include the following information:

- Motivation:

  What problem are you trying to solve? Why do you choose this topic? What will be the potential impact of your project? It is also helpful to discuss the current practice to provide the readers a bigger picture. What do you want to improve?

- Methods:

  - *Data source.* If you are using secondary data, describe the sampling procedure (unless it is data from population); if you are collecting data yourself, explain your data collection procedure.
  - *Variables of interest.* What variables are you interested in?
  - *Analytical procedure.* What statistical learning techniques have your tried and why? What outcome statistics are you looking for? If you are comparing models, describe your evaluation criteria here.

- Results:

  - *Descriptive statistics.* Describe the basic characteristics of your sample: sample size; missing values; the mean, standard deviation, or frequency distribution of variables of interest. Some basic graphs such as scatter plots and box plots might be helpful.
  - *Outcome from your modeling procedure.* Present values of your outcome statistics, such as training MSE, test MSE, error rates, and AUC, etc. If you are comparing results from different models, consider using figures.

- Discussion:

  Summarise your main findings. Discuss the implications of your project. Last but not the least, discuss the merits and limitations of your project.

1 Oral Presentation

Please choose one representative to upload the PPT to Moodle by Nov 21.

2 Final Report

Final project report can be at most 15 pages long including appendices and figures. Please see the presentation and language requirement below:

- Use Time New Roman font and standard font size (12 cpi).
- Set the line spacing option to 1.15.
- To improve the layout of your proposal and make it easier to read, you can divide it into sections and sub-sections, each with a relevant heading. Use line spaces to separate the sections from one another, and bold, capitals or italics to highlight the headings.

3 Individual Report

Apart from the group report, you are required to hand in a brief report (less than 150 words) specifying the individual contributions in the whole procedure, including your own and your partners' contribution. Note that the purpose of individual report is just to identify extreme cases.