STAT 4510/7510 - Applied Statistical Models I
Project Description (Spring 2022)

Project Overview

The goal of the course project is to apply the statistical learning techniques (either regression or classification) we have covered in this course to a real data set. This data could be from a public source, or data that you have gathered as a result of your own research.

Some sources of publicly available data

- UCI Machine learning repository: `https://archive.ics.uci.edu/ml/index.php`
- Kaggle: `https://www.kaggle.com/`
- Google dataset search: `https://datasetsearch.research.google.com/`

There should be enough feature variables (5 or more) and data points (a few hundreds or more) to make for an interesting analysis . In the end, you will be assessed on the process you used to try to answer the question(s) you posed and not the result itself. At this point in the course you have only seen basic techniques for regression and classification, but the project will require you to apply three or more unique techniques to your dataset. Students enrolled in STAT 7510 must complete the project independently. Students enrolled in STAT 4510 may choose to work alone or in a group of 2 or 3.

The project for the course is worth 20% of the final grade and will be evaluated by the final written report.

1. Project proposal: Due Friday, March 18
2. Project presentations: in class during the final week (May 3 and 5, tentatively)
3. Final report / R script: Due Tuesday, May 10

Project Proposal

Each student/group should identify a data set to be used for the project. Note that you can use the data set you select to address any research objective you can think of (not just the ones for which the data set was intended). I encourage you to be creative because many of these data sets can be very open-ended.

To be submitted (Friday, March 18)

1. Names of team members.
2. Brief summary of the data: Clearly describe the response variable, feature variables (predictors), and data source.
3. Research objective: What is the goal of the analysis? Is this a regression or classification problem?

Project Presentation

At the last week of the semester, each student/group provides presentations (about 10 minutes) for their preliminary analysis outcome. The project doesn't need to be complete for the presentation, and you can exploit it as an opportunity to take some feedback from your colleagues and the instructor for your final report.

Final Report / R script

Your final report should roughly have the following structure:

- Introduction: Describe the data's background and relevance, your research question(s), etc. Include a description of the data, some exploratory data analysis, and any pre-processing/clean-up you performed prior to completing your analysis.
- Methods: How will the data be analyzed in a way which can address the questions you have proposed?
- Outcome/assessment/interpretation: Evaluate your data analysis results. You may want to compare data analysis outcomes obtained from different techniques you used. Interpret your results in the context of your questions and discuss the implications.
- Conclusion: Summarize everything and suggest possible studies which could be conducted to extend your work in the future.

You should submit the complete R script (.R file) that used for your data analysis separately.

Academic Integrity

Plagiarism is taken very seriously and if a violation of the university's academic integrity policies is discovered, sanctions will be applied. Copying another person's code, output, or other analysis is expressly prohibited.