Health and Environmental Data Science

Spring 2019 Course Project

Deliverable Due Dates:

Project Proposal ..... 11:59 PM, March 8th
50% Status Report ..... 11:59 PM, April 11th
Project Final Report ..... 5:00 PM, May 1st (2 hours before final exam period)
Project Presentation ..... 5:00 PM, May 1st (2 hours before final exam period)

Overview:

The best way to learn about using data science is to actually *use it*. The purpose of this course project is to give you an opportunity to explore elements of the data science pipeline from start to finish, from preprocessing, data exploration, model selection, and effective communication of results. Even more importantly, this is an opportunity to explore the techniques on an application that's interesting to you.

The course project will be completed either individually or with a group. I suggest that you work in groups of 2 because it can be very helpful to bounce ideas off one another and share the data preprocessing load. *If you would like to work in a larger group (3 or 4 people),* you need to talk with Dr. Carlson about how you will split up responsibilities and how the project scope encourages larger groups prior to the project proposal. The expectation for the project will scale with the number of students on a team; a team of 2 Is expected to progress further than a single person, including greater data exploration, method exploration, and detailed writing. Regardless, I highly encourage working in groups.

Your course project should address an interesting and meaningful problem. Inspiration can come from your future career goals, a personal pet project, ongoing research. The project's topic is fairly open as long as it pertains to this class in some way, given that (i) the problem is difficult enough to be interesting (i.e. linear regression doesn't give you 100% accuracy) and (ii) the problem is feasible. **The goal of the initial project proposal** is to check that the proposed idea fits these two criteria. Be creative about the problem you want to address; the project should be something that engages you!

I do not expect you to invent new algorithms as part of this project. Largely, I want you to show that can apply techniques from class, and explore additional material related to your application, to a novel data problem.

Note that the project will be evaluated on the approach; not on the metrics of the results. In the real world, conclusions from data science techniques are often **negative**. There is information in a negative result, and you **will not** be penalized for negative results, as long as

the problem was approached in a reasonable way and the methods were approached and evaluated correctly.

Details on Deliverables:

There are 4 different deliverables associated with this project. As mentioned above, the project proposal is designed to check that the proposed project is appropriate. The proposal is 5% of the total grade. The purpose of the proposal is to make sure that groups are on track, and we will make efforts to give feedback quickly.

The progress report is 10% of the total grade, and is designed to make sure that students are progressing on the project, and to catch any issues before the final report and presentation. The final report is 15% of the total grade, and is expected to be a complete and comprehensive report. Each group will give a presentation during the final exam period, which will account for 10% of the final grade.

Project Expectations:

To give a sense of the final expectation of the project, the final report is expected to be complete reflection of the work on the project, and should be modeled on something you could hand to a manager or research advisor. *If you are graduating soon, think about this as an opportunity to have a portfolio-quality item that you could show to a potential employer.* Page limits will be given on individual assignments, and the report should be concise and self-contained with enough information to recreate and reproduce what you have done.

The content of the final report will be highly individualized to your project, but it should contain material to fully address the following categories:

- Problem Description: Background and context for the problem you have selected, and the motivation for considering it for this project. Some questions here to think about are: why is this an interesting problem? Why might others be interested in this problem? What is the *exact* question that you're trying to answer? What metrics will be used to assess your performance on this question?
- Data description: provide a description of the data you used, where it came from, and why it will help you answer the question you're asking from the data. Also, note any peculiarities of the data and how it would influence the analysis procedure that you're choosing.
- Data preprocessing: What preprocessing approaches were chosen, and why? What were their effect on the metrics of performance? What is the sensitivity of the final result to the choices made here?
- Chosen algorithmic pipeline: what approach did you choose? Why did you choose this approach? How sensitive is the final result to the chosen approach?
- Performance results: What are the estimates of the performance metrics from your complete pipeline on the dataset? How does this help answer the question you're trying to ask from the data? Provide curves that address trade-offs in model selection.

- Conclusions: What is your overall assessment of the system you built? What are its strengths? What are its weaknesses? What would you of done differently if you had to redo the project? What answer does the data support to the question you asked?