Project Overview

PPOL670 – Introduction to Data Science

Spring 2021

Contents

Overview

The following provides an overview of the data science project that you will be responsible for completing by the end of the semester. The project is an opportunity to apply the skills and tools that you've learned throughout the course on an area of substantive interest to you. The aim of this project is to build a model that predicts an outcome of interest (broadly defined).

The project is composed of three distinct parts: a proposal, a presentation, and a report. The proposal should outline the general plan for the project and will serve as an opportunity for the professor and teaching assistant to provide guidance on its feasibility. The presentation is an opportunity to present your work in mid-stream to receive verbal feedback from the professor, TA, and classmates. These comments will hopefully help you as you move forward with the final report. The report is the written analysis of the project in its entirety. The report will be due on May 18 @ 9pm (PPOL670's designated finals slot).

The proposal, presentation materials, and the report should be generated using RMarkdown and should follow all reproducibility practices discussed in class.

Project Proposal

Due	Proportion of Grade	Length
April 13	5%	750-1000 words

The project proposal asks that you sketch out a general project proposal (750-1000 words). The proposal should offer the following information:

1. A high-level statement of the problem you intend to address or the analysis you aim to generate;
    - What is the research question/problem?
    - Have others studied this question/problem before? If so, what are some conclusions they've drawn. What is still unknown?
2. What is the outcome variable that you're looking to explore/predict?
    - Describe the variable and how it relates conceptually to the research question/problem.
    - Provide summary statistics of the outcome variable.
        - Five number summary, plot the distribution of the variable, and/or any other valuable summaries of the variable (e.g. plot the spatial distribution of the variable)
    - Is there missingness in the variable? If so, explore what observations are missing data. Is there something systematic about that missingness (e.g. "a majority of authoritarian countries are missing in the data")
    - Note: *The outcome variable cannot be pulled from the World Bank, United Nations, International Monetary Fund, or Our World in Data without explicit permission from the instructor.*
3. What predictor variables do you plan on using to model the outcome?
    - You should have a minimum of 10 predictor variables in your model.
    - Outline where you plan to get these data.
    - *No more than 20% of your predictor variables* should be from the World Bank, United Nations, International Monetary Fund, or Our World in Data.
    - You do not need to describe these variables at this point in time. Rather I am looking that for a clear plan to get these data. If you have these, you can offer some brief descriptive statistics.
4. A definition for what "success" means with respect to your project.
    - In your words, what would a successful project look like? How will you know that you solved the problem or accomplished your goal?
    - Four weeks isn't a long time to complete a project like this. Thinking serious about what a "finished" or "successful" project might look like. This will help you set realistic goals/expectations.

Please be detailed but *succinct* as possible when writing. *Any material that exceeds 1000 words will not be considered when grading/reading.* There is no advantage/incentive to exceeding the word limit. Be sure to properly cite any referenced materials and/or packages

(it is okay if your work cited runs over the word limit. Your work cited will not be considered in the word count).

Project Presentation

Due	Proportion of Grade	Length
May 4	10%	7 minutes in length

Part 1: Video Presentation

Please prepare and record a 7 minute presentation that walks us through the progress you've made on your project to date. The presentation is an opportunity to summarize your project and talk through your (preliminary) results. Moreover, it will provide an opportunity for both your peers and the Professor/TA/classmates to provide constructive feedback, which you can then incorporate into your final paper.

When preparing your recording, please prepare slides *using* `R Markdown`. **Students should not "live code" or show output from their computer.** This is meant to be a polished presentation as if you were giving it in-person.

The slides should generally adhere to the following format. You should plan on having up to 5-10 slides in total. The layout of the presentation should take on the following form.

1. (1-3 slides) Problem statement and Background
2. (1-3 slides) Methods you explored or considered using.
3. (1-3 slides) The methods/tools you used, and the rationale for their use.
4. (2-4 slides) Results (however preliminary).
    - Show main visuals, analyses/tables, and/or any products built (interactive graphics, websites, etc.)
5. (1-2 slides) Lessons learned thus far and/or plans to mitigate challenges.

Students must submit both their slides *and* the `.rmd` file used to render the slides along with their video recording as a `.zip` file to CANVAS by the end of the scheduled class time. There will be no in-person and/or virtual class meeting this day.

> *Note that it is vital that all students submit their video on time so that others will have sufficient time to provide feedback.*

Part 2: Feedback

Each student will be randomly assigned the names of 5 peers in their class. The names will be circulated on **May 4**. Each student will be required to watch the recordings of their assigned classmates and provide substantive feedback by **Sunday May 9 11:59PM**. All comments/Feedback should be written on a shared **Google Document**, which will be circulated on **May 4** via the class Slack channel. The recorded presentations will be stored in a share folder on CANVAS. All enrolled students will have access to this folder.

Project Report

Due	Proportion of Grade	Length
May 18	30%	3000 words

The report is a complete description of the project's analysis and results. The report should be 3000 words in length and cover the below bullet points. Note that your work cited can exceed the 3000 words limit. Below I've outlined points that one should aim to discuss in each section. Note that paper should read as a cohesive report, so do not respond to these bullet points verbatim.

- **Introduction**
  - What is the aim of the project?
    - Summarize the problem
    - State your goals
  - What do you do in this report?
    - offer a roadmap of the project
- **Problem Statement and Background**
  - Give a clear and complete statement of the problem and/or aim of your analysis.
  - Include a brief summary of any related work that has tried to tackle a project similar to yours (i.e. a *light* literature review)
- **Data**
  - What is the unit of observation?
  - What is the outcome variable?
    - How is it measured?
    - Where does it come from?
    - Please describe how the outcome variable is distributed using a table and/or graph.

  - What are your predictor variables?
    * How are they measured?
    * Where do they come from?
    * Please describe how the predictor variable are distributed using a table and/or graph.
  - Outline any potential issues with the data:
    * Missingness
    * Lack of variation and/or availibility
    * Any potential sources of bias
  - How do you overcome/mitigate these issues in your analysis?
- **Analysis**
  - Describe the methods/tools you explored in your project.
  - Outline in detail our entire analysis.
    * Justify the tools/methods that you used.
    * Assume the reader is smart but doesn't know `R`/Machine Learning well. That is, be crystal clear about what you're doing and why.
  - Note that this section should walk us through what you're doing and how you're planning on doing it. There should be no results presented in this section.
- **Results**
  - Give a detailed summary of your results. Present your results clearly and concisely.
  - Please use visualizations and tables whenever possible.
  - Be sure to:
    * Discuss the performance of your predictive model.
    * Use interpretable machine learning to talk about which variables were important in the prediction task and how they relate to the outcome (i.e. PDP/ICE/Surrogate Models)
- **Discussion**
  - What conclusions should we pull from your analysis?
  - What are the limitations (i.e. what can't we say given your findings)?
  - How would you expand the analysis if given more time?
  - Speak on the "success" of your project (as defined in your proposal).
    * Did you achieve what you set out to do? If not why?

The reports must be submitted as a hardcopy (i.e. the `.rmd` notebook must be rendered as a `.html`) to CANVAS by 9PM on May 18th. ***Note that given the page constraints, no `R` code should be visible in the rendered document.***