DATA 2020 Rubric final report

Description

The goal of the DATA2020 project is to help you understand the concept of a real data application. It will also help in your job applications (academic or industry jobs alike). In fact, there will be a poster presentation to an audience of undergraduates, graduate students and faculties. The best poster will be chosen, and the group will receive an award "DATA2020, best poster award" (consisting of 75 dollars to spend on amazon). Make it look professional! You can add it to your CV, also providing the GitHub link to your CV so potential employers can check your R coding and writing style when you apply for jobs.

Each group will present its poster.

Students will be asked to analyze a dataset chosen by them. The students will have to be organized into groups of 4-5 and communicate both the dataset chosen and the group composition to the instructor asap.

Each group's draft poster presentation files will have to be submitted in canvas at 4 pm 5 days before (on Thursday). The poster presentations from each group will be on May 2nd or May 9th in the DSI lobby (presence is required for this date). If you are presenting the poster on May 2nd please send the poster to nathanael@brown.edu and maria_cardone@brown.edu not later than Sunday April 30 at 8PM, so they can print the poster and organize the DSI space. If you are presenting the poster on May 9th please send the poster to nathanael@brown.edu and maria_cardone@brown.edu not later than Sunday May 7th at 8PM, so they can print the poster and organize the DSI space. Please include in the e-mail the course (DATA 2020) and the size of the poster (36 inches by 48 inches). Poster received after the deadline will not be printed, and your score for the final project will be 0.

Please show the draft of your posters to one of the TAs before the presentations. Note that if you ask the TAs for feedback at the last minute, they might not be able to get back to you. They are not required to reply to emails during the weekend or outside the regular 9-5 hours. So please plan and give them at least one workday to reply!

Formatting requirements: Your poster should include the title, name of students in the group, affiliation, and (optional) a link to your Github repository. The structure should follow a simple abstract outline. Please include no more than 100 words per section. We will deduct 1 point for each additional ten words in total. Your poster must not be wordy. Too much text can be off-putting for the audience.

- Poster size should be 36 inches by 48 inches (portrait).
- Your text should be concise and clear, with no lots of words and more figures and plots (please see some poster examples in the folder final project/poster on canvas).
- No code should be included in the poster, and the audience should be able to understand your methods and results without looking at your code.
- Using R code is required.
- All figures and tables should be very clear; by looking at them, the audience should understand what you are plotting. It would be best if you had the title of a caption.
- All axes should be labeled, and the visualization type should fit the data you plot.
- : At the end of the poster, add a reference section to cite publications, data sources, and any previous work you have used for your poster.
- The poster will be graded on a scale of 100 points.

Poster sections

Please include the following sections in your poster and do not deviate from this structure.

Introduction

This should report the formulation of the question in terms of a statistical objective. In your introduction, make sure to motivate your problem, describe your dataset, and the previous work. 20 points

EDA

This should be largely focused on exploratory data analysis of variables you decide to consider in the methods. If you decide to remove some variables from your analysis, explain why you did that. Explore the distributions and the symmetry, and explain why you will use the methods in the next section. Describe the most important/interesting aspects of the preprocessing and the EDA briefly. 20 points

Methods

In this section, describe and justify the methods used for analysis. Try at least two different methods we have explored in class. Describe the variables you choose with appropriate methods, theoretical implications, and assumptions. What metric do you use to evaluate your models' performance and why? Measure uncertainties and/or prediction (e.g., bootstrap, cross-validation, etc.). In general, explain what considerations went into each step of the methods. 25 points

Results

This section includes visual and numerical summaries of key components of the analysis, providing an interpretation of results in context. For example, the best model, accuracy, prediction, hypothesis testing and confidence intervals. (in classification: what is the baseline accuracy, etc.; in regression: what's theMSE/RMSE or R^2). Which model was the most predictive? Summarize the performance of the models in a table or using a figure. Calculate at least three different global feature importances and discuss your findings. Discuss the results of model interpretations in the context of the problem. Which features are the most and least important? Did you find something that's unexpected/surprising/interesting? 25 points

References

List the publications, data sources, and any previous work you found. 5 points

Contribution of each member of the group.

Include each member's contribution to the poster. 5 points