Fall 2024

Project Guidelines Stat 441/ Stat 841/ CM 763

Report (Grad students):

One student can upload on crowdmark on behalf of the team.

7 pages maximum. Pdf format. The 7 pages include everything except references. References are not required but could appear on an 8th page. You can submit code as an appendix, but do not refer to it in the main report. Sometimes people try to circumvent the page limit by using tiny margins and font sizes. This is not a good idea because it makes it harder to read the report. Please use 1-inch margins on all sides and a 12-point font.

Please start the report with an abstract (=summary) of your findings like any scientific paper. Most abstracts are between 100-200 words, but feel free to make it longer or shorter as you see fit.

Including appropriate figures and/or tables in the report are strongly encouraged. Please make sure they are readable.

Presentation (Ugrad)

The presentation has a time limit (10 minutes). If you ignore the time limit you have an advantage, and I will have to take off marks. I will not penalize if you are a few seconds over. Beyond that you are taking your chances. Speeding up the video to fit 10 minutes is not permitted (and also risky as it may affect my ability to understand what you are saying).

Generative AI

If you use generative AI for writing or video production you must disclose details. Videos: please send an email to me. Reports: Add another page explaining how you used generative AI and what text passages are affected. Use of generative AI for coding is also allowed and does not require disclosure.

Presentation and Report: What to write/ talk about?

The main goal is to be understandable. I would like to learn about your approach(es) to prediction for the particular data and any insight you may have gained. When something did not work out, I would typically not go into great detail (unless you learned something that is interesting to the reader). Just mention what you tried and that it didn't improve predictions. Please see the rubric for details.

Rubric

The following rubric is generic and not specific to this data set. It is meant as a guide and may change somewhat.

Preprocessing

- Highlights from exploratory data analysis, if appropriate
- Appropriate choices for variables with missing values
- Appropriate choices for categorical and ordinal variables.
- Appropriate choices for any unusual circumstances
- Statements made are correct

Feature engineering

- (The opportunity to create new variables varies widely by dataset, may not be appropriate for all data sets)
- Were any new variables derived?
- Are the new features creative/innovative/sensible?

Modeling

- Use of at least 2 different techniques (gradient boosting and xgboost are both boosting and are NOT considered two different techniques)
- If an algorithm was used that was not discussed in class, there is sufficient explanation to make the reader believe the authors understand the algorithm
- Appropriate tuning
- It is clear how the model is used (what key choices have been made)
- (negative) Algorithm is presented in terms of which software button/software option) to push
- (negative) Reader gets the impression that the algorithms are not really understood
- Statements made are correct.

Stacking

- (stacking may not be appropriate/useful for all data sets)
- Did the authors attempt stacking?
- Is it clear how exactly stacking was used?
- Statements made are correct.

Clarity

- (key) Writing is easy to understand. Ideas are easy to follow.
- Figures: axis labels/any text are readable
- Tables are readable
- Are human-understandable names used? (x5 being an important variable means nothing to me. Income being an important variable does. If brevity demands, you can use x5 in a graph or formula as long as there is an explanation nearby).
- (report only) Figures and tables are referred to in the text.

- (report only) Figures and tables have a caption.
- (report only) Report starts with a well written abstract

Insight

- Did the reader gain any special insight about the data or techniques used? Examples:
    - A particular feature engineered variable was key (ideally quantify in some way)
    - None of the feature engineering variables mattered. If appropriate, a brief reflection about why this is surprising.
    - Stacking was key (ideally, quantify in some way)
    - Something was surprising because …
    - Original variable <readable name> was key. (ideally, quantify in some way)
    - A pre-processing step was key
- (strong negative) A laundry list of things that are not really insightful. It means the authors have difficulties understanding the essential aspects of their results.
- Statements made are correct.

Participation

Participation is assessed as pass/fail.

- (Presentation) All team members present. The presentation time should not be highly uneven (a bit uneven is ok, it is the nature of things)
- All team members contribute to analysis, coding and presentation/report.