STAT 451 Group Project

Timeline

The project is worth 50 points, with intermediate deadlines as listed in the schedule:

- Form a group (0 points) of 4-5 students:
  - To choose your own group, go to our Canvas page's People tab, click on the Project tab, and put your group members into "Project *n*" (where *n* is the lowest number from 1 to 30 that is not already in use by another group).
  - Right after this deadline, we will randomly assign students to groups for those who do not choose their own groups.
- Write a one-page proposal (4 points) including a few lines of code to read data, descriptions of the question(s), variable(s), and methods you will use. Turn in a `proposal.html` (or `.pdf`), once per group.

  Write a discussion post to claim your data set by clicking on "Project proposal data set" in the "Discussion" tab of our Canvas page.
- Meet as a whole group with teacher and/or TA (3 points) for 5-10 minutes to discuss your proposal.
- Turn in presentation slides as `presentation.html` (or `.pdf`), once per group.
- Presentation (25 points) to class.
- Attend peers' presentations and give feedback (8 points spread over two or three days).
- Turn in report (10 points) of 750 words or less with supporting graphics as `report.html` (or `.pdf`), once per group.

A description of project requirements follows below.

Data

Find a data set on the internet that interests you. Pick one or several questions about one or several variables about which you are curious. Choose a data set and question different from your peer groups' choices. **Do original work.** Include citations (URLs, etc.) for data, graphs, and methods you get from others.

Here are links to many data sources: `www.stat.wisc.edu/~jgillett/451/project/dataLinks.pdf`

Presentation

Make presentation of 5-8 minutes (in class for in-person courses or via video for online courses) that summarizes your report (below).

- Introduce the topic to a general audience.
- Summarize your method.
- Highlight outcomes of your project.

Here are some procedures:

- Include group members' names on your first slide.
- Each group member of your group should speak for at least one minute. Introduce yourself as you start speaking. (Something like "Hi. I am John Gillett" is sufficient.)
- For in-person courses:
  - Before the first day of presentations, you will turn in a `.pdf` or `.html` file of slides to Canvas. I will have all these files on my laptop, which will be connected to the projector.
  - On each day of presentations, we will randomly select which group should present next.
  - A short period of Q&A will follow your presentation.
  - Outside Q&A and the gaps between presentations, the audience must be silent. (Exchange texts or notes or discretely leave the room if conversation is necessary.)

Here are suggestions:

- Practice, as there is no other way to give a good short talk. It lets you experience awkard moments that call for revision.
- Speak to the audience (not the screen), which is the class (not its teacher).
- Do not put words on slides you don't give your audience time to read.
- Instead of flashing a slide for 1 second, revise the talk.
- There is excellent "Oral Presentation Advice" at

  `http://pages.cs.wisc.edu/~markhill/conference-talk.html`

Report

Write a report of 750 words or less (about three pages of text, possibly plus supporting graphs) in three sections describing your data, variables, question, machine learning analysis, and conclusion.

- Its *introduction* should summarize the data you analyzed, the question you pursued, your machine learning analysis, and your conclusion. It should outline the body of your report. A reader who quits after your introduction should understand your work broadly.
- Its *body* should describe your data, machine learning analysis, and results.

  Here are ideas to consider:

  - Mention your data source and size.
  - Explore your dataset with numerical and graphical summaries.
  - Describe feature engineering like data cleaning, rescaling, and imputation.
  - Try several models, compare their strengths and weaknesses, and explain your modeling decisions.
  - Use cross-validation or other hyperparameter-tuning techniques to justify your selection of hyperparameters.
  - Analyze your chosen model with performance metrics on unseen test data, explaining parameters and hyperparamers where possible. Discuss overfitting.
  - Include informative graphics where possible efficiently to communicate your conclusion.
  - Mention weaknesses of your work.
- Its *conclusion* should revisit your question and conclusion in the light of your report's body. It could suggest future work.
- Include a post-conclusion "Contributions" paragraph briefly describing the contributions of each group member. Here is an example (other "Contributions" designs are ok too):

Member	Proposal	Coding	Presentation	Report
Lucy Van Pelt	1	1	1	1
Charlie Brown	1	1	1	1
Linus Van Pelt	0	0.5	0.4	0
Spike	0	0.7	0	0.3

  Notes:

  - In the chart above, 1 = full contribution, 0.1-0.9 = partial contribution, 0 = no contribution.
  - Linus attended the presentation without preparation.
  - Spike sent a video, but it was unrelated to our presentation slides.

Write your report using any software you like. Turn in a `.pdf` or `.html` file.

(There are two examples of projects from Sebastian Rashka's 2018 STAT 479 course (479 became 451) at the bottom of `https://sebastianraschka.com/blog/2019/student-gallery-1.html`. His course and project requirements differ from mine.)

Grading

Peer feedback on presentations will include voting/ranking that leads to three awards for

- best presentation
- most creative or interesting project
- best visualizations

These are some things we will consider when grading.

- Does the project demonstrate knowledge of the course?
- Is the report no more than 750 words long? (Paste your text into `https://wordcounter.net` to check, as we will stop reading at 750.)
- Is the question engaging?
- Is the analysis correct and persuasive?
- Are the graphical and numeric summaries informative? Are their fonts easily legible?
- Is the writing vigorous? (Strunk and White say, "Vigorous writing is concise. A sentence should contain no unnecessary words, a paragraph no unnecessary sentences, ...")
- Are report authors listed at the top?
- Are numbers rounded? "0.3 vs. 4.1" conveys information faster than "0.337885 vs. 4.078801".
- If you report a classifier's accuracy, put it in the context of P(success) in the data. (If P(success) is near 0 or 1, high accuracy is obtained by merely predicting 0 always or 1 always, respectively.)
- Do you report counts where rates are more useful? e.g. "There are 10 times more car accidents on cloudy days than sunny days" is uninteresting if there are 10 times more cloudy days than sunny days. On the other hand, "There are 123 accidents per 100,000 miles on cloudy days but only 12.3 accidents per 100,000 miles on sunny days" is interesting.