DATS 6103 Data Mining Project – 2023 Spring

Goal

The goal of this project is to use Python to obtain, pre-process, and clean dataset found online, and use it to present an analysis that includes EDA, various model building tasks that are applicable to the dataset. The skills that all data scientists treasure the most would be all on display. They include technical analysis, critical thinking, team work, communication and visualization skills.

There will be a presentation, 15-20 mins for each team, in-person. The instructor, TA, and your fellow classmates might give you feedback and comments on your presentation. Based on all your work as well as others comment, your team will submit a final paper explaining your analysis, your results, as if you were submitting the formal report to your supervisor, or an article to be submitted to peer-reviewed journals. The paper should be about 10-15 pages long (Not counting any tables, charts and graphs), and less than 5000 words in any case. You can have as many charts and graphs as you like, but it would a good idea to prioritize those you'd like to include in the main body of the report vs. those you can put in the appendix.

We require datasets to have at least 4000 observations (i.e., 4000 rows of data), and eight or more relevant variables/features (i.e., 8+ columns of useful data).

Steps for Completion

This project has these components (100 points total, which is 25% of your course grade):

1. Topic proposal (4 points, team grade*)
2. Presentation slides (12 points, team grade or individually graded, team decision, let me know when you submit the topic proposal)
3. Presentation (24 points, your in-person presentation, individually graded)
4. Git usage (12 points, individually graded, from the git repo history log online)
5. Codes (24 points, team grade*, your technical and coding skills)
6. Final Write Up, any format (24 points, team grade*, your communication skills)

* Even though these items are graded as a team, I reserve the right to make adjustments if there is strong evidence of unequal contributions among team members. I hope I never need to use this special clause.

Key Due dates

1. Topic Proposal: Tuesday, March 28 (to the team folder on OneNote class notebook).
2. Powerpoint or Google Slides for your team presentation: April 23 (to the team folder on OneNote class notebook). You can still change them afterwards. You will receive feedback. This submission is not "graded", but the presentation file used the actual presentation will be graded.
3. Team Presentation (target 15-20 minutes): Apr 25 for Section 11, Apr 26 for Section 10.
4. Discussion Forum (as a separate Homework/Discussion grade) - Every student needs to review at least two other teams' presentations and give them feedback: end of Thursday, Apr 27.

5. Python file (.py file) and Final summary report (separate items) - With the TA/instructor/fellow classmates' comments, modify your codes and/or write up if needed and submit. Deadline: April 28. The final summary report can be Word doc, pdf or html created by QMD. You can also choose to use other word processor to write the summary report.

Part 1: Topic proposal

Your topic proposal (one per team) is a 150-200 word description of

1. the research topic your team comes up with
2. the SMART question(s) of your research (see below)
3. the source of your data set(s) and
4. the link to your team's GitHub repo
5. the modeling methods you propose to use (you can change afterwards)

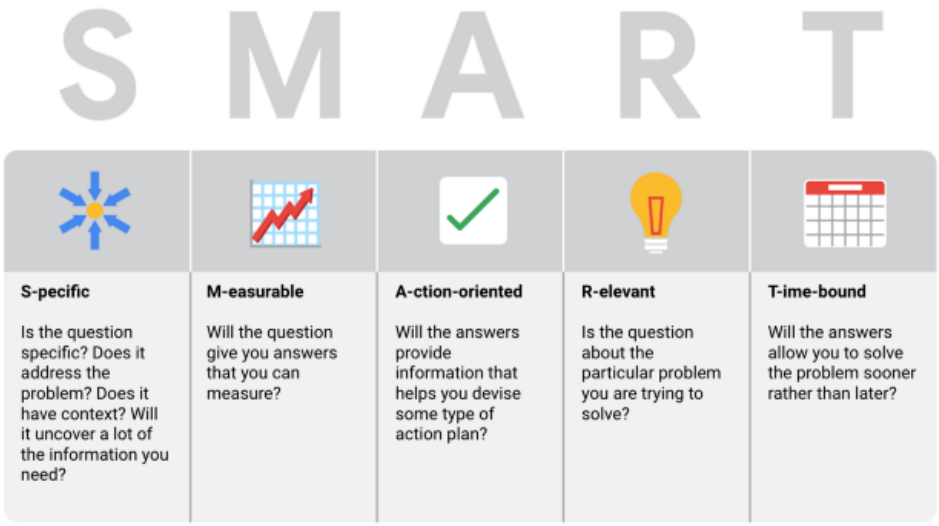

Figure 1. Illustration of SMART Questions
Source: *Highly effective questions are SMART questions* by EM Kautsar, Medium, Mar 30, 2021).

Part 2: Presentation Slides

You can use Powerpoint, Google Slides, or any other appropriate products. Typical slides should not be too "wordy". It's not the purpose of slides to be read like an article.

Even though this following point is not universally agreed on, I personally do not feel complete sentences are needed on slides. Just bullet/lists of the main points you want to deliver. Use charts, graphics, animations to capture the audience's attention. Keep in mind some of your audience might be watching your presentation on a low-res screen, or a mobile device, or in general could be real-time with low band-width. Cramping a lot of words on to the slides would be unwise.

Using succinct bullet points is one way to thoughtfully ensure readability and accessibility for others. Another step to take is using appropriate contrast (e.g., very dark blue background with white or off-white text) and San Serif fonts. To fully check accessibility of a PowerPoint presentation, you can use the Tools > Accessibility Checker option.

The presentation slide deck can be graded either as teamwork or individually. Your team can decide and indicate your choice with the topic proposal. The default choice is team graded and feedback will be provided as a way to improve your work prior to the team presentation.

Part 3: Presentation

Each team will prepare a 15-minute presentation delivered on our presentation day. Each member needs to be a part of the presentation. Although we do not have strict rules on time/duration of appearance for each member, roughly equal time allocation is desired. Any order of appearance and combinations are allowed. Making the best impression on the audience is the goal. The presentation score is awarded individually to each member.

Part 4: Git Usage

While you are working on the project collaborating with your teammates, you all should be using the git source-control-management (SCM) system. Remember to commit often, and fetch/pull regularly, so that the team work goes smoothly.

All this history of work will be available in the repo. Make sure you add "ning-rui" on GitHub as a collaborator to your team's repo (private repo). I need to see everyone is contributing to the project. Definitely NOT just committing and pulling/merging on the last day of the project submission.

This part is graded individually. As long as you are doing your part, updating your codes regularly on the shared repo, you will get full credit for this part.

Parts 5: Codes

Submit a separate python file (.py) that includes your codes. The codes need to be working. Indicates what packages you are importing, so that I can follow and run your codes. Also remember to either include the dataset, or include a link to the dataset either online or on GitHub.

Parts 6: Final Write Up

The final write up can be prepared in any word processing format and style your team chooses.

There is a 5000-word limit (max) for the document. There is no minimum, as long as all the information about the research and the result is conveyed. If your team has to go over the limit, please contact me to get pre-approved.

These are some of the items for the write-up, if applicable. Use your organizational skill to make the logical summary of your research:

1. Overview of your project (if applicable):
    a. Why do your team choose this topic?
    b. What prior research and analysis have been done on this topic?
    c. Information about the dataset(s) you used.
    d. Any unusual EDA results that is worth mentioning?
2. Your SMART questions, and how did they come up?
3. Your models or machine learning algorithms. Score or evaluate your models.
4. Interpret your results.
5. What predictions can you make from your models? Examples?
6. Draw conclusions. How do these answer the SMART questions?
7. References (APA style, see for example GW Himmelfarb: APA Citation Style, 7th Edition)

Grading Criteria

Your final paper will be graded according to the given rubric document.