DATS 6101 Introduction to Data Science

Final Project Outline (Spring 2022)

Description and Purpose:

The goal of this Final Project is to apply the various model building techniques we've learned to a real-life data science project. For completeness, your team should summarize here your project's initial stages such as background research and EDA which were performed in Midterm Project. This is especially important if your team has switched topic or dataset between Midterm and Final. Your team will have new SMART question(s) now that are to be answered by building models.

Same as before, we require datasets to have at least 3000 observations (i.e., 3000 rows of data). Each team will choose their own research topic and question. You can use the same dataset as your Mid Term Project, although it is not required.

Your team should submit a topic proposal with SMART question(s) about one week before the presentation is submitted. After the presentation is submitted, you will receive feedback from your fellow classmates, TA, and instructor within the next 2 days. You can then make modifications to your **Final Write Up** before submitting.

Your team will prepare and submit a research paper, no more than 4000 words (in the html, not counting TOC, figure caption, etc). Even though each picture/chart/graph might be worth a thousand words, they do not count towards the word-limit.

Details:

I. **Topic Proposal** (Due: April 13, Wednesday, End of Day). Each team must submit one proposal on Blackboard. In 150-200 words describe a) the research topic, b) the SMART question(s) of your research (you can still change them afterwards), c) the source of your data set(s) and how many (roughly) observations, and d) the link to your team's GitHub repo.

II. Development of a **research driven question (SMART)** focused on a dataset either inside of R or one of your choosing from any online sources (3000+ observations). It can be the same dataset as your Midterm Project. The SMART questions would be different, however.

III. (Due: April 20, Wednesday, End of Day) Provide an **R-markdown file**, knitted into **HTML**, which shows the R-code and brief explanations for the technical work in your project. (**Also** submit your **data file**, or give the online source URL.) It should include:
   - Summary of the dataset (just basic xkablesummary (dataframe) will do here)
   - Model(s) used
   - Model evaluation(s) and comparison

IV. (Due: April 20, Wednesday, Class time) Develop a **15-to-20-minute** presentation for the team that effectively communicates the results of your data science project to be presented during class.

V. (Due: April 27, Wednesday, End of Day) Write a roughly 10-page (No more than 4000 words) summary of the research and EDA process of your project. The summary should be prepared in **R-markdown**, and knitted into **HTML**. You may take some of the work in part II (such as graphs and results) to include here. They can overlap. This summary is to-be presented to your boss, your client, or to-be submitted for publication in journals. Potential area of topics to address in this summary may include:
    - Some basic EDA.
    - How did you select and determine the correct model to answer your question?
    - What predictions can you make with your model? Examples.
    - How reliable are your results?
    - What additional information or analysis might improve your model results or work to control limitations?
    - References (APA style preferred)

Grading:

I. Topic Proposal, 5%
II. R-codes (RMD), Technical analysis, 25%
III. 25% Individual presentation score + 10% team presentation doc (pptx or google slides)
IV. Summary paper, 25%
V. Git usage – 10%, individually graded, based on your git activity in repo history)

Grades for parts I, II, IV, and part of III, are team-based. But I reserve the right to award different grades to team members if there is evidence of unfair contribution within the team. A peer evaluation form will be submitted individually by all students after the completion of the project.