Overall project structure

The course project ties together multiple topics covered in this course. In this project, students will propose and implement their own predictive modeling analysis to address a selected task. The course instructor provide a list of data (Project_Datasets.xlsx) from which students may select their course project. Optionally, students may propose their own project idea with instructor approval.

- Individual effort or teams up to 3 students
- Submissions are through Canvas, no email submission will be accepted
- Semester-long project
  - 1-2 weeks to pick a question and find data
  - 3-4 week to perform analysis
  - 3-4 week to choose models and implement them
  - 1-2 week to write report

Important Dates:

- If you decide to use another dataset, please send it to me for approval by September 5th
- Checkpoint 1: Due 10/03 11:59 PM
- Checkpoint 2: Due 11/07 11:59 PM
- Final Report: Due 12/12 11:59 PM

Checkpoint 1 - 20 points

Submit a 2–3 page summary (Microsoft Word or PDF format) of your EDA.

Your summary should include the following information:

- Description of the problem background and objective. Indicate the analysis paradigm you will use to frame the problem; that is, are you framing the problems as regression, classification, clustering, or some other form of statistical analysis? Indicate your rationale.
- Summary of the data set that, at a minimum, answers the following questions: How was the dataset obtained? What is the unit of analysis? How many observations in total are in the data set? How many unique observations are in the data set? What time period is covered?
- Summary of any data cleaning steps you have performed. For example, are there any observations / time periods / groups / etc. you have excluded? Detail missing data (e.g, the number of incomplete samples, percent of samples for which a given feature is missing). Specify the approach you used or will use to address missing data (e.g., complete case analysis, missing data imputation).

- Description of the outcome variable. Provide a visualization of the outcome variable if appropriate. Provide distributions of the outcome variable as appropriate.
- Description of key predictors/features with appropriate visualization techniques that compare predictors to the response. You should investigate all predictors in your data as part of your project. For this assignment, pick one or two predictors that you think are going to be most important in explaining the outcome. Your selection of predictors can either be guided by your domain knowledge or be the result of your EDA on all predictors.

Checkpoint 2 - 20 points

The second element is a summary of the baseline model selection and results. Each project team will submit a summary document in Microsoft Word or PDF format. The summary document should include the following:

- Justification of your model choice based on how your response is measured and any observations you may have made in your EDA.
- Report the model's test error rate using one of the techniques discussed in lecture. Justify your choice.
- Visualizations (e.g., plots) as needed to explain model performance.
- Based on the estimated test error rate, discuss how well the model performs.
- Use the model to make predictions for at least three test cases of interest.

Final presentation & report - 60 points

Each team will complete a final report. For the final report, **your team should implement an alternative analysis or model that is anticipated to outperform your baseline analysis.** The final project written report should be a detailed description of the final project that integrates all your findings and includes the following sections:

• **Introduction:** This section should be an approximately one (1) page description of the task the project addresses.

• **Materials and Methods:** This section should describe the methods used including those needed for data preprocessing, feature selection, machine learning and/or statistical analysis, and performance evaluation.

• **Results:** This section should include tables, figures, and text descriptions necessary to detail the analysis results.

• **Discussion and Conclusions** This section should provide an interpretation of the analysis results, conclusions drawn from the analysis, review any challenges that were encountered, discuss limitations of the analysis, and a proposal of how the analysis could be improved further or applied more broadly to other tasks.

• **Code** Include your executable code files as an attachment to the final report

References should be cited as appropriate. The final written report should be submitted in PDF format using the following naming convention: *Lastname_Firstname_Report.pdf.* The project source code should be submitted as an archive file (.zip or .tar) file.