Assignment 6 (term project – individual work): Data Analytics (Fall 2024) 30% (25% written + 5% oral presentation)

Written part and presentation slides (both) Due: Tuesday, 10th December 2024 by 08:00 pm ET by email (eleisa2@rpi.edu). (12/10/2024 by 08:00 pm ET on LMS).

Submission method: written part and poster via email. Please use the following file naming for electronic submission: DataAnalytics_A6_YOURFIRSTNAME_YOURLASTNAME.xxx, etc.

Level: _____ Please mention the Level (4000 or 6000) that you are registered for this course and mention it at the top of the written part of the assignment.

Please use font style New Times Roman and font size 12 for text in the written part of the Assignment 6.

Late submission policy: **This assignment is due at the end of term. If you are more than 2 days late, it is likely that you will not have your grade for this assignment included in your final grade before they need to be submitted.**

Note: Your work for this assignment should be the result of your own individual work. Take care to avoid plagiarism ("copying"), and include references to all web resources, texts, and class presentations. You may discuss the project with other students, but do not take written notes during these discussions, and do not share your written assignment or presentation or poster.

General assignment: Your term projects should fall within the scope of a data analytics problem of the type you have worked with in class/ labs, or know of yourself – the bigger the data the better. This means that the work must go beyond just making lots of figures. You should develop the project to indicate you are thinking of and exploring the relationships and distributions within your data to lead to optimized predictive models. Start with a hypothesis, claim, or questions. Think of one or more ways to construct models[1], find or collect the necessary data, and do preliminary analysis, detailed modeling, validation, summary (interpretation) and (if any) resulting decisions.

Note: You do not have to come up with a positive result, i.e. disproving the hypothesis is just as good. Use the section numbering below for your written submission for this assignment.

Guidance: Topics, scope and general nature – please use the feedback from Assignment 4 (project proposals) and further consult the instructor and your classmates.

1. Abstract and Introduction (2%) Describe your motivation, initial hypothesis/ idea that you wanted to investigate, and if applicable any prior work, interest in the topic (like an intro for a paper, with references), Minimum 1/2 page.

2. Data Description and preliminary analysis (3%)

[1]NOTE: 6000-level students must use at least two different datasets during the analysis.

Describe how you determined which dataset(s) you used in this project, the criteria, source, data and information-types in detail, associated documentation and any other supporting materials. Detail any preliminary analysis you did. (e.g. initial scatter/box/line plots) Minimum 1 page text (+ graphics if applicable).

3. Analysis (5%) Explore the statistical aspects of your datasets. Perform any transformations, interpolations, smoothing, cleaning, etc. required on the data, to begin to explore your hypothesis/ questions. Analyze the distributions; provide summaries of the relevant statistics and plots of any fits you made. Discuss and specify or estimate possible sources of error, uncertainty or bias in the data you used (or did not use). Minimum 2 pages text + graphics.

4. Model Development and Application of model(s) (12%)

[1]NOTE: 4000-level students must develop at least two different types of models and 6000-level students must develop at least four different types of models, not just change the number of variables for a given model type.

What types of models you used to describe the data (regression, classification, clustering, etc.), patterns/ trends you found, visual approaches that helped you choose models, and or variables (type/ number) in the model, other parameter choices or settings for the models (e.g. distance metrics, kernels, etc.). Apply the models to assess model performance (i.e. predict). Discuss the confidence in your results including any statistic measures. Discuss how you validated your models and performed any optimization (give details). Minimum 6 pages text + graphics.

5. Conclusions and Discussion (3%) Describe your conclusions; interpret the results, predictions you made, the models and their characteristics, and a give summary of what changed as you went through the project (data, analysis, model choices, etc.), what you would do next, or do differently in a subsequent exploration. Minimum 1 page text (+ graphics if applicable).

References – websites, papers, packages, data refs, etc. You need to properly cite your references in the body of the document and should include those properly cited references at the end. You can choose any citation format (Chicago, MLA, etc…) but be consistent when you cite your references (There is no specific citation format, just be consistent).

Include your R scripts in a zipped folder with the written part of the assignment and send them by email or upload to GitHub.

6. Oral Presentation (5%). Plan for a ~6 minute presentation; slides must cover the following:

a). Title (with your name) and Abstract

b). Problem area – what you wanted to explore/ solve/ predict and why, and what you wanted to predict?

c). The data – where it came from, why it was applicable and the preliminary assessments you made.

d). How you conducted your analysis: distribution, pattern/ relationship and model construction. What techniques did you use/ not use and why? What worked? What did not work?). How did you apply the model? How did you optimize, account for uncertainties?

f). What did you predict and what decisions (prescriptions) were possible. What was the outcome, conclusions?

Graphical Representations

Provide graphical representations related to each of questions 2, 3, and 4, at least. Ensure all figures are numbered, legible, fully explained in the text and annotated with a caption. The final written document should be a minimum of 10 pages (but can be more). All graphics should be within your written assignment unless they are very large. Large graphics files – a link to an online location is acceptable (e.g. Box, Dropbox, Google Drive).

[1] **NOTE: 4000-level students must develop at least two different types of models and 6000-level students must develop at least four different types of models, not just change the number of variables for a given model type.**