STAT 4010/5010

Statistical Methods and Applications II

Final Project Description

Conduct a thorough analysis of a data set using statistical modeling techniques that we have learned in this class. I expect you to include a minimum of 5 of the major concepts that we have discussed in class this semester. Below is a listing of techniques you may choose from.

- hypothesis testing
- hypothesis testing with randomization
- confidence intervals
- bootstrapping
- regression modeling
- diagnostics of the model – goodness of fit, feature selection, residuals
- t-tests
- F-tests
- Causal Inference
- Bayesian networks
- ANOVA
- Model Selection – AIC, BIC, MSPE, R^2, Ra^2
- Generalized Linear Modeling

You will be expected to show your work and process in a jupyter notebook. The last markdown cell should formally describe your results. You should think of the last markdown cell as writing a 1-2 page "paper". I expect 400-500 words.

Grading Criteria:

The project grade will be determined on the basis of the quality of the statistical analysis, paper, and presentation. The project will be graded out of 100 points. A specific rubric follows:

10 points: Clear layout of work and inclusion of the 1-2 page "paper" at the end of your notebook. Please include formatting and give a brief description of what you are doing above each coding cell (and include comments) so that your process can be followed.

25 points: The written paper that summarizes your results. You will receive a grade of 0-5 in each of the following categories: (1) context and purpose of writing; (2) content development; (3) sources and evidence; (4) explanation of your statistical analyses; and (5) syntax and mechanics. It is quite obvious when chatGPT is used. If chatGPT is suspected, you may receive a 0 on this portion. Please use your own thoughts and words.

65 points: Related to problem solving. You will receive a grade between 0-10 on how well you define the problem/question of interest. You will receive a grade between 0-10 on how well you propose solutions to answer the questions. You will receive a grade between 0-25 on implementing solutions to answer the question of interest. I will be specifically looking for the minimum of 5 topics in the bullets above and breaking this score into 0-5 sub-scores. You will receive a grade between 0-20 on how well you evaluate outcomes, interpret results, etc.

For the report/"paper" at the bottom of your jupyter notebook submission, please use the following outline as your guide:

1. **Introduction/Background**

- Why are you interested in this problem?
- What is the relevant background information for readers to understand your project? Assume that your audience is not an expert in the application field.
- Is there any prior research on your topic that might be helpful for the audience?
- From where did the data come? Is this an experiment or observational study? Who collected the data? Why was the data collected (if you weren't the one doing the collecting)?
- What are the questions of interest that you hope to answer?

2. **Methods/Results (experimental design and data collection)**

- How did you obtain the data? A research lab? Government website? Web scraping?
- Describe your exploratory data analysis methods. What needed to be done to the dataset to make it amenable to analysis?
- What analyses are most appropriate to answer the question of interest?
- Describe the analyses used. Check your assumptions!
- Present relevant graphics and interpret results.
- Explicitly connect your technical (e.g., statistical, mathematical) results to your research questions.

3. Conclusions

- What are your conclusions? What did you learn?
- How would you extend this research? What future research ideas come to mind based on your results and experience with this analysis?