MATH97309 Data Science Coursework 6 - Final Report (70%)

Type the title of your project here

Your name here

Submission Deadline: April 26, 2021

Instructions:

- You can download and use Coursework_6.Rmd file as the template for your submission. You can also use LaTeX on your computer or Overleaf to write your coursework.
- Upload the knitted HTML/PDF from the .Rmd file or a PDF from LaTeX or Overleaf to the Coursework 6 Turnitin assignment on Blackboard. The last submission will be considered the final one and previous submissions will be over-written. Avoid last minute uploads.
- Additionally, upload the .Rmd `R` Markdown file (or your code file if you are using LaTeX or Overleaf) to the Coursework 6 assignment (non Turnitin one).
- Your file should contain the code to answer each question in its own code block. Your code should produce output that will be automatically embedded in the output file.
- Writing Code: Use either Google's or Hadley Wickham's style guide for your code. Use functionals instead of loops when possible (refer to Hadley Wickham).
- As this is assessed work you need to work on it INDIVIDUALLY. It must be your own and unaided work. You are not allowed to discuss the assessed coursework with your fellow students or anybody else. All rules regarding academic integrity and plagiarism apply. Violations of this will be treated as an examination offence. In particular, letting somebody else copy your work constitutes an examination offence.

For this coursework you will investigate bias within a black box model. You can choose any algorithmic model that is approriate for your problem and choose one of the following datasets:

- US CDC COVID-19 Case Surveillance Public Use Data (`COVID19.csv`)
- Teacher Evaluations (`evals.csv`)
- US Communities and Crime Data Set (`communities.zip` - Includes `communities.csv` with the data and `communities.names` that contains information about the column names.)
- Predictors of likelihood of sharing disinformation on social media 2019-2020 (`Disinformation.zip` - Read the `ReadMe.txt` and `Techical_Report.pdf`. Additionally, take a look at `sample_code.R`.)

For the dataset of your choice, choose a problem that you are interested in and train an algorithmic model to answer your problem. Then your task is to investigate bias (ethnic, gender, nationality, etc.) within your trained model and the dataset.

You should follow the PPDAC cycle and create a 1 page plan (1% of overall marks for course) due on 18 February consisting of the first three steps:

1. Define a problem that you want to investigate in your dataset.
2. Make a plan in terms of what dataset/variables are you going to use, which black box model(s)? What statistical analyses do you plan to do?
3. Do you expect to see any bias in your model, e.g. will you investigate gender bias or something else? How do you define bias?

For the final report (70%) due on April 26, you will write up your findings in the form of a 10 page research report, using LaTeX or RMarkdown (the 10 page limit is inclusive of figures, references, and appendix!)

It must include the following sections [weighting for marking criteria shown]:

1. Abstract (~200 words) [3 points]
2. Introduction (this should include an improved and expanded version of your problem statement submitted on 18 February. Additionally, it should include motivation for why you chose this problem. Mention your data source and include any outside sources or other information that motivates this project.) [8 points]
3. Literature review [10 points]
4. Data (this should be an improved and expanded version of what you submitted on 18 February. Describe your dataset and the variables that you work with. Include exploratory data analysis, including graphics or analyses that show some initial results on your problem. If you make any transformations, explain why in this section. Make sure your graphs have clear labels, titles and captions.) [10 points]
5. Results: including data visualizations and tables (Train an algorathmic model. Explain your choices for variables. Explore if the trained model is biased.) [15 points]
6. Conclusion: describe your overall results and discuss limitations to your data or analysis. Also include what other things you would have liked to do if you had more time or more data. [10 points]
7. References [2 points]
8. Appendix with minimal working example source code for reproducibility. (Note that if you are using RMarkdown, don't include code in the previous parts of the report. Hide your code and include the code in the Appendix.) [2 points]

For this coursework, practice thinking about how to analyze data. While there are wrong answers, there are many possible right answers. Any data analysis decisions or conclusions that you make should be justified and explained. Your job is to correctly analyze the data, not force the analysis to match a pre-conceived idea.

Note: you will receive written feedback on your submission of a single page on 18 February. You should incorporate this feedback into an improved and expanded version for sections 2 and 4 in the final report, and you will be assessed accordingly.