Homework 3

Due Date: 11:59pm on Thrs, Dec 14th, 2023

Submission:

You will need to make two submissions**.**

- A PDF file. Please submit a written report under the “Problem Set 3” assignment on Blackboard using the filename “lastname_firstname_report.pdf”

Late Submission:

Late assignments will not be accepted without a valid medical certificate or other documentation of an emergency.

Grading:

Make sure that your solutions to the problems are written clearly and concisely. You will be graded both for answering the questions correctly and for writing up your answers in a readable manner.

Problem 0. Real World Impact [2 points]

Think of one practical use case for your project. It could be something passive, such as analyzing/monitoring trends over time, or something more active, such as a decision-making tool or a data point to be used by decision-makers in positions of authority. Perhaps it's improved patient treatment. Try to be specific. For example, instead of "monitoring COVID cases over time," your practical use case could be "monitoring COVID cases among graduate students at BU over time".

Briefly explain how your project could be used for that use case, then answer the rest of this problem set in the context of that practical deployment. Keep your answers concise; there's no need to add excess details if just a few details are enough to get your point across. You may also want to include some of these answers in your project final report.

Problem 1. Dealing With Data [6 points]

(a) The data you are working with for the project has already been collected and processed. However, in a real world setting, there is often a delay between when the data is collected and when it is available to you for analysis. How does your model deal with that delay? Does it still achieve its purpose even with the delay, or would you need to work around the delay for the model to be effective? If you will need to work around it, how will you do so?

(b) The data you are working with has its own distribution, but in a real world setting you may have outlier data points that fall far beyond the tails of that distribution. How do you detect and handle outliers in such a setting?

(c) The distribution of your data might change over time, and if your model is not calibrated for this change, it might start to perform below expectations. How do you expect to handle this data drift? Briefly explain the process you would follow.

Problem 2. Performance Evaluation [3 points]

How will you evaluate the performance of your project post deployment? If your practical use case involves a model that will be running in real time, your evaluation can include (but should not be limited to) performance metrics that could be used to evaluate the performance of that model. But regardless of whether a model is involved, how will you evaluate the success of your project? Will you need some kind of A/B testing that evaluates your project compared to other existing alternatives, or will you rely on some objective metrics of success? Briefly describe your evaluation scheme.

Problem 3. Social Implications of Model Deployment [4 points]

(a) Who are the stakeholders for your project, and how would they be affected by the deployment of your model?

(b) Would you need to involve the IRB to deploy your project? Briefly explain why or why not your project might need IRB approval.

Problem 4. Paper Submission [1 point]

What academic journal would be a good candidate for your paper? Why? How much work would it take to make your project publication ready?