Read the first blog post. Based on the descriptions there, which "data scientist rock star" do you think you are most/least like?

I think I am most like Eqaan Librium for whom 'Equilibrium and work-life balance' is important. Yes, accuracy and precision does matter to me, along with being able to explain my work to my stakeholders and the decision of which aspect matters most would differ case by case.

What is the most important thing you learned about XGBoost from these blog posts that either we didn't cover in class or just didn't "click" in class?

I think I did hear this in class at some point but I must have missed out on how XGBoost is an ensemble method that uses a combination of multiple base models to produce one optimal predictive model. I think a combination takes the best aspects from each model and produces a better result. I do know that we touched upon the word gradient but I understood it better from the article of how XGBoost uses a gradient boosting framework meaning that decision trees are grown in a sequential manner as an iterative learning process.

What is the most important thing you learned about doing data science from these blog posts? How does it align with, complement, or run counter to what you have been learning in your data science courses at Calvin?

I've always thought more about missing data so I liked the simple overview that the blog post provided about what to do with missing data.

- Don't always throw away data with missing values because sometimes it could be valuable information because it might have a specific reason for being missing
- In handling missing data, make indicator values, 1 for missing, 0 if not
- Impute all missings in a column with the same value OR
- Replace the missings with a value per group OR
- Predict the missing value based on other variables
- Your choice of what to do with the missing values depends on what model you are building

Look through some of the python code (in the blog posts or at Gitlab) and comment on at least one thing that is interesting because it uses some python thing you didn't know about or would not have thought to do that way.

I liked how the first step of preparing data in the DATA TO PREDICT WHICH EMPLOYEES ARE LIKELY TO LEAVE was changing the target variable to an integer. I may not have thought of that right away but maybe after working with the data multiple times.

The other thing I liked was using indicators for relevant_experience. I think it makes the process much smoother to have dummy fields for uniformity and convenience in analyzing the data.

Now read the last post. Does this change your answer to or the way you think about question 1?

I agree that no data scientist is the same but I think that there may be overlaps in the approaches that a data scientist takes, with one approach being more prominent in a model than another model. And of course, this decision is made by the data scientist and their judgment of which approach takes lead is what makes them different. I think that Eqaan captures this diversity of decisions and I think it is important to always find that right balance depending on the goals of a model.

Just because I'm curious: Which of the blog posts did you read? (Just list off the numbers.)

1, 2, 3, 5, 7, 9

Thinking back on the book reports we have been hearing, give one example of something you heard that got you thinking.

Thinking about Adham's presentation, I remember him mentioning how comfortable humans have become at being stagnant with how fast technology excels and that really made me stop and think how the world is changing so quickly. We have new technologies everywhere. When I first heard about ChatGpt and how popular it was becoming among the student population, I was completely against using it because I did not see why as students, we needed more technology to assist our work. And slowly I realized how important it could be in workplaces but I still was against using it. However, as time went by, I realized that as a computer/data scientist, if I was not familiar with new technologies, I'd be old fashioned and I'd fall behind the rest of the world, which is why I thought it was important for me to make myself aware of technologies like ChatGpt because it will end up being an important technology in the future. So going back to Adham's statement, I think sometimes it is not even the comfort of humans, but it seems like we do not have a choice anymore except to keep up with the technology, no matter how fast it is becoming.

If you were going to read one of the books that someone else reported on, which one would it be? Why?

I would like to read Algorithms of Oppression, presented by Richmond. The book that I read, Data Feminism, talked about oppression and I think reading a book that further explores these different kinds of oppression, especially in the Computer/Data Science field would make me well aware of the bias around me.

Think back a bit about our course as a whole. What do you think you will remember most about it 3 years from now?

I will remember the ethics lecture given by Professor DeYoung because it really opened my eyes to the problems we deal with ethics in our world. I loved how she integrated Christian faith in her version of what ethics should be and I liked how she mentioned that ethics are also embedded in our day to day actions and change begins from the small steps we take.

This course covered quite a variety of things (ethics, visualization, R/python, git/github, creating R package, XGBoost, etc.).

Which topic did you find most challenging? I found XGBoost challenging. I remember trying to create a decision tree with my team and how hard it was to figure out the shape of that along with understanding the algorithms that go inside of XGBoost.

Which topic did you find most rewarding? I loved learning how R/python because I worked with these two languages independently and combining them was a new learn for me. Creating the R package was a great assignment because I learnt about R/python, github, vignettes, documentation, etc.

What topic(s) do you wish we had spent more time on? I do wish we spent more time on XGBoost because I felt like we only started it in the second half of the semester and we didn't really get the chance to work on a hands on assignment using XGBoost.

If you were to assign yourself a grade in this course, what grade would it be? Why? Take into account how much you learned, the quality of the work you did, how much you contributed to your team, etc.

Considering everything that I learnt in this class, I believe I deserve an A. I know there were classes that I missed but I always made sure to ask my classmates about the content of the lecture that day and went back to look at those topics myself either through the class website or ChatGPT. When it came to in class and other assignments, I did do them, individually or as a team. Unfortunately, the last weeks of the semester were rough for me but once I was able to get back to school, I stayed in constant communication with my team about working on the R package. I attended meetings whenever I could (online or in person), I helped debug some github issues (during meetings, outside of meetings and sometimes over email) and I contributed to writing the content of the package. I loved working with Quarto for my book report presentation. It was a new platform for me, but I was quickly able to work my way through it and I was able to execute my presentation without any problems. Overall, I learnt and relearned things in this class that I know will help me in my career field moving forward.