MCIS6273 Data Mining (Prof. Maull) / Fall 2021 / HW3a

This assignment is worth up to 10 POINTS to your grade total if you complete it on time.

Points Possible	Due Date	Time Commitment (estimated)
10	Sunday, November 14 @ Midnight	*up to* 3 hours

- **GRADING:** Grading will be aligned with the completeness of the objectives.

- **INDEPENDENT WORK:** Copying, cheating, plagiarism and academic dishonesty *are not tolerated* by University or course policy. Please see the syllabus for the full departmental and University statement on the academic code of honor.

OBJECTIVES

- Learn about the Human-in-The-Loop (HITL) concepts through a candid and practical podcast discussion

- Learn about the ethics of AI through a paper on practical and ethical considerations of Food Recommender Systems

WHAT TO TURN IN

You are being encouraged to turn the assignment in using the provided Jupyter Notebook. To do so, make a directory in your Lab environment called `homework/hw3a`. Put all of your files in that directory.

Then zip that directory, rename it with your name as the first part of the filename (e.g. `maull_hw2_files.tar.gz`), then download it to your local machine, then upload the `.tar.gz` to Blackboard.

If you do not know how to do this, please ask, or visit one of the many tutorials out there on the basics of using in Linux.

If you choose not to use the provided notebook, you will still need to turn in a `.ipynb` Jupyter Notebook and corresponding files according to the instructions in this homework.

ASSIGNMENT TASKS

(50%) Learn about the Human-in-The-Loop (HITL) concepts through a candid and practical podcast discussion

We have talked about the important role humans play not only in the preparation and interpretation of models, but also in the entire pipeline of the data science process.

To drive that point home, we will listen to a 22 minute podcast presented by the *Banana Data Podcast* title: "The Importance of Human in the Loop AI With Christina Hsiao". The podcast is a short, but interesting discussion of HITL concepts in practice.

You can listen to the podcast here:

- https://banana-data.buzzsprout.com/300035/8803284-the-importance-of-human-in-the-loop-ai-with-christina-hsiao

Or you can get the MP3 directly from here:

- [Direct MP3 download from BuzzSprout.com](#)

§ Listen to the entire podcast and answer the questions below:

1. What is the definition of ML Ops according to Ms Hsiao?
   Including in your answer where the intersection of ML Ops.

2. How does the analogy of baking versus cooking related to the conversation on ML Ops?

3. Provide information about the example given where when the "cook" or subject matter experts had to come in and manually validate model output. Include a discussion of cost benefit analysis against false negatives.

4. What was one of the problems Christina identified in oil and gas domain expertise? How is it being addressed?

5. Christina argues against "humanization" in MLOps. What does she suggest instead?

6. According to Christina, what is the goal of MLOps in the Data Science landscape?

7. Write two or three sentences about your reaction to this podcast.

(50%) Learn about the ethics of AI through a paper on practical and ethical considerations of Food Recommender Systems

We have tangentially talked about the ideas of data privacy and the concerns around how, as data scientists, it is important to apply a lens of how data might be used and what privacy concerns need to be considered when working with data, especially personal, human data.

You will read a recent and interesting paper about the ethics of recommendation systems for food recommendation. You may already have experience with smartphone applications which make all kinds of recommendations to you, from which pair of socks to buy on Amazon, what route to take on a drive to shop at your favorite electronics store, to what music to listen to. There are serious issues to consider around these recommendations, not the least of which are *transparency* in how recommendations were made, *what data* were used to make those recommendations, and if there are *biases in the recommendations* which present an unfair advantage to the winning recommendation over alternatives. Consider the recent concern around Amazon bringing out their "house" branded products which compete directly with other products on the same platform (see the 2019 Wired article *"Amazon Doesn't Favor Its Own Brands—Except When It Does"* by Glad Edelman for a more detailed exploration of the problem https://www.wired.com/story/amazon-gating-private-labels-antitrust/).

These issues are complex and nuanced, and this paper by Karpati, et al, 2020 is an interesting exploration into the realities of the AI-driven world we are rapidly accepting, choosing or being forced into. Whichever the case, as a data scientist, you should be aware of the issues involved, as they may very well come up in your work.

Please read:

Daniel Karpati, Amro Najjar, and Diego Agustin Ambrossio. 2020. Ethics of Food Recommender Applications. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 313–319. DOI:https://doi.org/10.1145/3375627.3375874

You can find the pre-print PDF (does not require access to ACM Digital Library) here:

- https://arxiv.org/pdf/2002.05679v1.pdf

Or if you have access to ACM/DL you can get it here:

- https://dl.acm.org/doi/10.1145/3375627.3375874

§ Read the entire paper and answer the questions below:

1. Why is *conceptual vagueness* a problem in the ethical guidelines reviewed by the authors?
2. What are the eleven ethical principles the authors identify as important concepts in any ethical framework?
3. What is IoF? Please detail as many components of IoF as identified in the paper.
4. How do the authors say Food Recommender Systems (F-RS) are different from other recommender systems, and why do the authors think that difference is important? Do you agree with their critique?
5. In terms of autonomy and personal identity, the authors mention the concept of "informed consent" when entering a business relationship. Explain why you agree or disagree with their statement that "[a customer relationship] should entail the awareness about how the other party is compensated for the services they provide". You should be as descriptive as possible when stating your position. Use examples from the paper if you wish.
6. Why do you think defining the non-ethically motivated parts of an algorithm (section 3.5.2) are important in theory, but difficult to actually implement.

7. In section 3.7 "Polarisation and Social Manipulabilty", the authors suggest that "Attracting as many users as possible behind an ethical agenda seems easier if the stress is put on a catchy buzzword" . . . and ". . . we need stakeholders who truly commit to their values and promote their intrepretation and implementation". How might transparency improve the gap that emerges between what companies say about their product and what is true in what they are actually doing? Could you conceive of a way *blockchain technology* might help?
8. List the product assessment criticisms the authors lodge against the Yuka application. Why do they feel the inclusion of "organic" as a criterion for recommendation is problematic?
9. Why are the crowdsourced databases that Yuka rely on a problem for the authors?
10. *Explainabilty* and *Fairness* are two issues that the authors criticize Yuka heavily for. What do you think the Yuka designers could do to address the authors' concerns?
11. Write a three to five sentence reaction to this paper. Include in your reaction whether you thought the authors were objective? Did they clearly communicate the goals of the paper? Were they successful in accomplishing those goals? Do you agree or disagree with one or more points in the paper? How would you suggest improvements to the paper if you were a reviewer or reader?