Statistical/Data Literacy Task

Rough drafts to be discussed on Wednesday, February 14th; "Final" drafts are tentatively due on Friday, February 16th

Purpose:

The goal of this activity is to connect research on statistical/data literacy to practice. We will be reading and learning a lot about statistical and data literacies, so this activity is an opportunity to apply what we are learning to our teaching and/or research practices. You may choose to work on this activity individually or in groups – it is completely up to you! 😊

Directions:

For this activity, please find or develop a task or example focused on an aspect of statistical/data literacy that could be integrated within a course (or research project) of interest. The task/example may be adapted from another source (please cite), or it may be of your own creation.

For the chosen task/example, please explain the ways in which it focuses on an aspect of statistical/data literacy, referencing any readings/sources as appropriate to support your explanations, and describe how you envision the task/example could be integrated within the chosen course (or research project) of interest.

During class on Wednesday, February 14th, 2024, we will get to share our ideas and tasks/examples with one another, with the goal of continuing to workshop them together. "Final" versions of the tasks/examples and the corresponding explanations are not due until Friday, February 16th (or possibly later if needed).

All our work including the below information and the Jupyter Notebook can be viewed in this repository. This task was developed by Emily Bolger and Rachel Roca.

Task:

Play the Game (15 minutes)

Learning Objective: Step into the role of using AI to solve a task, identify the potential consequences when utilizing these tools.

TO DO:

1. Play through "Survival of the Best Fit" while considering the following questions:
    a. Reflect on intention vs. impact. Why was an algorithm used in the first place? Was the intention malicious? What was the impact on the candidates?
    b. What was your strategy when choosing who to hire? How did the game react to your choices?
        i. (Facilitator's note: mention what is built in deterministically- how does this extend to limitations of simulations/models in general?)
    c. Were you surprised when you reached the end of the simulation? Why or why not?
2. If you have time/want to, play through the simulation a few times. What aspects stay the same? What changes? Regardless of the choices that you make.

Look at How the Data was Constructed (10 minutes)

Learning Objective: Dive deep into the data creation process and the model building for "Survival of the Best Fit," including the choices that were made to implement bias.

TO DO:

1. Read through the pieces of the Jupyter Notebook used to create the data and the model for "Survival of the Best Fit" on Google Collab (based on their GitHib repository).
2. Discuss/think about the questions posed in the notebook regarding data creation.
3. What remaining questions do you have? Pose it to your group or class.

Real Life Application and Discussion (~15-20 minutes)

Learning Objective: Understand bias in data and algorithms in the real world, outside of toy simulations.

TO DO:

1. Read the article "Amazon ditched AI recruiting tool that favored men for technical jobs".

2. Consider the following questions:
    a. What connections do you see between “Survival of Best Fit” and Amazon’s real-life recruiting tool?
    b. Why was the algorithm biased in this case? What was biased about the training data?
        i. How did “Survival of the Best Fit” integrate this type of bias when creating their data? How did it affect the performance and results of their model?
        ii. How does the simulated data and the real data represent systemic issues present in our world?
    c. What components of bias did you notice when playing the game, if any?
    d. How would you feel if you knew the company you applied to was using AI to judge candidates? Why or why not? Related, if you were in a hiring role, would you feel comfortable using AI to aid in hiring decisions? Why or why not?

How the Task can be Integrated:

We see two potential applications for where this activity could be integrated into a data science or data science-related course.

Option 1: This activity can be integrated into a terminal quantitative reasoning or mathematics course as is. Therefore, non data science students can be exposed to the need for ethics and the dangers of bias in algorithms in data. As we've seen within the activity and the data literacy readings Wolff, et. al. (2016), these ideas are relevant to everyone. This option of the assignment would be used to teach what Wolff, et. al. (2016) call *Communicators* and/or *Readers*. *Communicators* and *Readers* do not necessarily need complex technical skills, but should be able to interpret and make sense of data seen in their everyday life, as well as communicate these data stories to others. We note our activity relies heavily on discussion, and doesn't particularly require any specialized background knowledge in mathematics or coding. This activity may also be used to not only emphasize algorithmic and data bias, but can also link to concepts such as reading histograms, understanding different types of random distributions, and viewing (machine learning) algorithms as more than an inaccessible black box. We would imagine this would take up a full class period, but hope that the ideas of ethics and interrogating bias would appear as a common thread throughout the whole course. The "Play the Game" section can easily be adapted as a pre-class assignment, to give more time during the class period for discussion and looking at the data creation. We recommend an active learning, flipped classroom approach, where students can discuss in small groups before a full class debrief.

Option 2: This activity can be integrated into an mid-level computational data science course. Thus, this activity would be for the education of what Wolff, et. al. (2016) calls the *Makers* and/or *Scientists*. These archetypes of data literate citizens require more technical skills to contribute data-driven models and solutions for society. The "Play the Game" and the "Real Life Application and Discussion" sections would remain generally unchanged. However, prior to the "Look at How the Data was Constructed" section, we would have the students create their own code to generate the data and model used for the simulation. This would include:

- Identifying components of the simulation that rely on the bias.
- Developing a plan to use code to mimic the identified components.
- Writing the code to create a bias dataset, with justified choices, and applying a classifier of their choice to the data.
- Evaluate whether the choices made integrate bias into the data in a way that is similar to the simulation.

Students can then complete the "Look at How the Data was Constructed" section. Students will compare their implementation with the creators of the simulation. They will explore: How their choices affected the outcome of the model? What choices did they make to integrate bias compared to the creators? What are the affordances and constraints of each method? What would they do differently if they were tasked with generating this data again? Ideally, this would be taught in a flipped classroom style (or semi-flipped classroom style). Thus, the creation of the data and model will be constructed in small groups.

Connecting to Data Literacy:

While the definition of data literacy is actively still being developed (Gummer, 2015; Wolff, et.al, 2016; Schield, 2005; Dichev, 2017), Wolff, et. al. (2016) has converged on the following definition through conducting a literature review of relevant papers in the field:

> "Data literacy is the ability to ask and answer real-world questions from large and small data sets through an inquiry process, with consideration of **ethical use of data**. It is based on core practical and creative skills, with the ability to extend knowledge of specialist data handling skills according to goals. These include the abilities to **select**, clean, analyse, visualise, **critique and interpret data, as well as to communicate stories from data and to use data as part of a design process."**

The aspects of the definition we highlighted in our task are indicated in bold above.

Wolff, et. al. (2016) emphasizes the importance of ethics and understanding the role and impact of data in different contexts as competencies that should be present in all aspects of the data inquiry cycle. Additionally, Utts (2003) identified understanding "Biases in Surveys" as one of their seven topics that educated citizens should know about statistics and probability. As presented by Utts, this included phrasing of survey questions, ordering of survey questions, and development of these questions for varying audiences. As surveys are a common form of data collection, the concerns presented extend to our considerations. Both of these authors highlight both the high level of bias we face in data, as well as the inevitability of interacting with this data on a daily basis.

We (Emily and Rachel) decided to showcase the importance of ethical use of data (particularly human data) in the case of hiring practices. Despite this importance, it is not always emphasized explicitly in classrooms. This aligns with our lived experiences in TAing for various classes in CMSE; even data science students aren't exposed to issues of ethics and bias in their data science classwork. We thus included a more in-depth integration of this task (see Option 2), to mitigate this. While we are unfamiliar with the curriculum of terminal math and quantitative or data literacy courses, we feel that this would be a useful and important task to engage students with these issues.

With this in mind, we (Rachel and Emily) feel that it is important to note that across the literature there is not a cohesive and comprehensive definition of data science literacy. Most include some aspect of collecting, analyzing, visualizing, and interpreting data, but make no claims as to the necessary level of these skills and how they are defined on a fine-grain level. Additionally, courses covering data science and literacy describe the key components they feel are necessary to the course which align with the stated aspects, but few deeply describe the implementation of these skills (Dichev, 2017; Schreiter, et.al, 2024). Since it is unclear at what skill level one becomes 'literate,' we created multiple access points to the material that can be expanded and adapted for various levels.

Related to the presented definitions, our task focuses on giving students exposure to the ethics in data. This includes how bias is embedded into data collection, the existence of this bias in the model is used in, and the implications of using the results from these models. Students think

critically about the choices the creators made to create the simulation and how this represents real-world stereotypes and bias.

Algorithms are increasingly being used in applications such as hiring practices to online advertisement and even incarceration decisions (Weapons of Math Destruction by Cathy O'Neil, Topaz, 2023). Therefore, it is imperative to give students exposure to these topics, since their lives have been and will inevitably continue to be affected by decisions made by an algorithm. We thus present multiple perspectives in which they can interact with algorithms, data, and ethics. The first part of the assignment provides them with the perspective of implementing these algorithms in a workplace to understand the motivations behind using them and the potential consequences. The second part introduces students to the creation of data and the models from a developing perspective, highlighting just how many places bias can be introduced. Finally, students are faced with a news article where they must engage with the implications of being barred from a job due to gender identity. It is with these three parts and perspectives that we hope students leave this task with a more comprehensive understanding of how and why ethics are so important in data and algorithm driven approaches.

Emily and Rachel Notes:

Tie questions, etc towards articles/readings and their definitions of what data science literacy is/should include
https://www.survivalofthebestfit.com/

- Playing the game
- Purpose of using ML (speeding up, etc)
    - What do you notice about your tendencies as you play the game
- Try to break/outsmart the game
    - It's rigged…
    - Some aspects are pre-determined/ built in

https://github.com/survivalofthebestfit/survivalofthebestfit/blob/master/data-engineering/biased-data-gen.ipynb
https://github.com/survivalofthebestfit/survivalofthebestfit/blob/master/game-source/public/game/assets/text/textTemplate.js ,
https://github.com/survivalofthebestfit/survivalofthebestfit/blob/master/game-source/public/game/controllers/machine-learning/modelTraining.js

- Digging to the data and how it was builD
- Built in bias into the data
- What choices were made in the creation of the data that affects the performance of the model? How does this capture aspects of real world bias? Did you notice this when playing the game?

https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

- Amazon hiring debacle, real world example of the simulation above

(https://callingbullshit.org/syllabus.html)

To do:

- Read the Data Science Literacy articles with specific frame of how it relates to task
    - DONE
- Add markdown/comment explanations on data generation
    - DONE
- Come up with discussion questions for simulation activity, Amazon article, and data generation file
    - DONE
- Create a general explanation/overview of the activity hitting these topics:
    - *For the chosen task/example, please explain the ways in which it focuses on an aspect of statistical/data literacy, referencing any readings/sources as appropriate to support your explanations, and describe how you envision the task/example could be integrated within the chosen course (or research project) of interest.*