Transparency and Interpretability

Preethi Narayan

Apr 29, 2022

Problem 1: Online Job Ads

Part A

Give three distinct reasons why gender disparities might arise in the operations of such a system.

Feedback loop - This system is trained on biased data (from historical bias as mentioned below) and this bias will be continually reproduced as the system will learn from its own recommendations that carry the preexisting bias, and then continue to make more recommendations holding that bias. Because this job search website collects interaction data from its own users, more bias will be created because, resulting from the fact that women are not shown high paying jobs due to the existing bias of the system, they will continue clicking on lower paying jobs. The system will then learn from this behavior and assume that women are not interested in the higher paying jobs, and then continue to show women lower paying jobs, therefore creating a feedback loop for this bias.

Historical bias - This hypothetical job search website uses historical data, and therefore the system would have historical bias. In the past, marginalized groups like women and people of color have been discriminated against, and if this past data is used in the training data for the new machine learning algorithm, then historical bias will make its way into the new website's predictions.

Employer bias - As mentioned earlier, the service receives data from employers, detailing which users were invited to job interviews, and which were hired. This data is used in the machine learning algorithm used by the system, and as a result the decision that these employers have previously made will be used to influence future recommendations made by the algorithm. This would cause the service to take on biases that were exhibited by employers during their hiring processes.

Part B

Suppose that the job search service decides to increase the number of times it presents job openings in STEM to women. To do so, the service observes that STEM job experience (in years) is positively associated with the likelihood that a user clicks on an advertised STEM job opening: the more years of experience, the more likely a user is to click. Consider the following intervention:

> *Pre-process the training dataset, replacing the value of the "job experience" feature for women with the best (highest) possible value for the feature in the dataset.*

Part B(i)

Under what conditions will this intervention increase the number of times job openings in STEM are shown to women?

In the case that the job openings in STEM being shown to people is highly correlated with the years of experience that people have in the field, this would result in an increase of the quantity of job offerings that are shown to women because the best value of the feature is what is used in the dataset. In summation, if the value "job experience" is used for determining who should be shown STEM jobs, then women will not be shown as many STEM jobs because the system is rather prioritizing people with more experience than prioritizing women when it comes to choosing who is shown the STEM jobs. If this is adjusted so that the best possible value is women instead, then women will be prioritized and shown more STEM jobs.

Part B(ii)

Under what conditions will this intervention fail to increase the number of times job openings in STEM are shown to women?

In the case that the job openings in STEM being shown to people is highly correlated to gender in the model, then this change would not increase the number of times job openings in STEM are shown to women. The model would learn that there are less women in STEM jobs, and would correlate that women should be shown less STEM jobs than men. In this case, the intervention would fail to increase the number of times job openings in STEM are shown to women.

Problem 2

In this part of the assignment, you will watch a lecture from the AI Ethics: Global Perspectives course and write a memo (500 words maximum) reflecting on issues raised in the lecture.

This lecture covered the concepts of how companies decide their marketing tactics that they use online by analyzing mountains of consumer data. Companies use this data to create targeted advertisements, prices based on different consumers, and other tactics related to privacy. The stakeholders that could be affected by the data science issues discussed in this lecture include the companies that are using the data that is collected from the consumers, and the consumers whose data is being collected. The general public is also affected by the data science issues discussed in this lecture because the precedents set moving forward will affect everyone in society. The companies that are using the data that is collected from consumers are the ones that highly benefit from the data science issues discussed in this lecture because they are able to profit off of the data collected from consumers. With the data, they are able to find new customers that are more likely to generate revenue for them, and they are able to create advertisements that are also able to make it even more likely that these consumers provide the companies with more revenue. The companies are also able to tailor their price points for different products and services based on the information that they have about the consumers that they are targeting. Though this disproportionately benefits the companies, some consumers prefer targeted advertisements rather than random ones, and as a result, this might also mean that it benefits the consumers on a small scale as well.

Though consumers may receive a small benefit from targeted advertisements, in general they are usually the ones that experience the most negative effects as a result of companies amassing and using data from consumers. In regards to the tactics used by companies that relate to privacy, differential pricing, and targeted advertising, consumers are the ones that are the most hurt by them. Consumers' privacy is breached because companies often collect the data of consumers without their knowledge or consent. Targeted advertising without informing consumers that their data is being collected is morally and ethically incorrect because it preys on consumers that don't realize their information is being used to coerce them into making purchases or support companies, and they are not given any choice or agency in the matter of their data being used to influence their actions and thoughts.

Differential pricing can negatively affect consumers in many different ways. One way that differential pricing can affect consumers is by companies quoting higher prices to more affluent consumers while quoting lower prices to consumers that are less likely to have as much disposable income, as people coming from lower income communities or students that are in college. Differential pricing can also further drive existing biases into society by giving companies the power to discriminate between different types of consumers, which could result in consumers from marginalized groups being shown unfavorable products or prices in comparison from privileged consumers. One example of this that was discussed in the lecture was how Black consumers were not shown ads for certain rental properties, which occurred as a result of a company exhibiting bias after receiving data on these Black consumers. In this example that was mentioned during the lecture, Facebook/Snapchat owned the data that was being shared, and the data protection was certainly not adequate for the consumers because the consumers did not

know that their data was being collected to such an extent, nor did they know that their data would be used to carry out racial biases. Consumers have little to no information about how their data is being used by the companies that collect/sell it, or how it is being used by the companies that buy it. In regards to one of the specific situations above, consumers have no understanding of how an algorithm might take their race and income into account to show them ads for one property vs another. This black box negatively affects consumers because the lack of information leads to discomfort, but benefits companies as they have no requirement to disclose their methods.

Currently, vendors have very few incentives to increase data protection, transparency or fairness. One incentive may be to avoid backlash from current and potential consumers and prevent loss of income from such backlash. Data protection laws are being written and enacted in many places across the world, and in these places companies could face severe consequences for breaching the privacy of consumers. Once regulations are improved over time, vendors will have more incentive to protect the data of consumers.

Problem 3

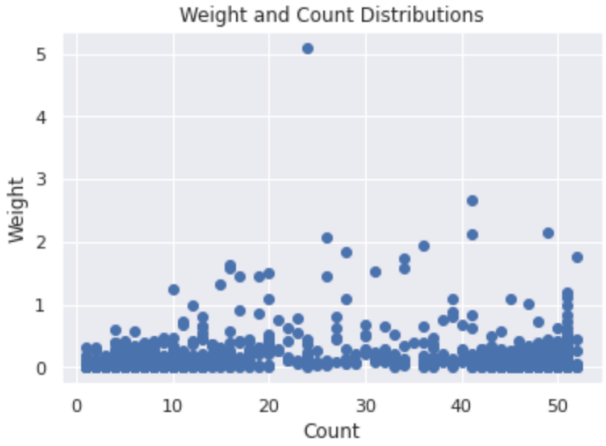
Part A

See Notebook

Part B

See Notebook

Part C



As shown in this plot, there is not a heavy correlation between count and weight in this subset of words that we have studied. Some words have high weight and low counts while other words

have low weights and high counts. There is one outlier in the dataset that has a high count and a high weight (the outlier that has a weight greater than 5). In summation, there is not much of a relationship between count and weight for these words.