|  |  |
| --- | --- |
| Assignment 4: Essay – Hugo de Heer | Q2 – 2022/23 |

Machine learning for socio-technical systems

*Power, Responsibility, and Fairness*



Content

**Q1:** What did you learn about ML power, responsibility, and fairness? What is responsible ML, and What is “fairness”? (Recommended max 200)

First of all, machine learning and AI are powerful tools that can bring unprecedented opportunities to many different organizations, such as, governments and businesses. With this power also comes a lot of responsibility. Namely there are multiple important subjects that we have to consider because of its direct impact on people’s lives when we are applying machine learning in a socio-technical setting. Responsible AI appends from naively applying AI to reach a certain goal by also looking at the models’ risks, fairness, reliability, safety, privacy, accountability, explainability and transparency. In this way, it aims to address the challenges around AI from both an ethical and legal point of view.

With fairness we mean treatment or behavior without favoritism or discrimination. We can further categorize this in individual fairness and group fairness. In equality (individual fairness) every individual is given the same resource or opportunity to reach a certain outcome. Equity (group fairness) goes a step further and recognizes that people have different circumstances so it allocates the exact resources needed for an equal outcome of everyone. In machine learning, decisions may be considered unfair if there is a presence of biases or if they are based on sensitive variables, such as gender, ethnicity and many more.

**Q2:** What is “Operational Fairness”, and how does it differ from the ethical stand(s) or legal stand(s) on fairness? What are the frameworks and categories of fairness you saw in class (lecture and lab) or others you know of but did not see in class? *(Recommended max 200)*

Operational fairness aims to address the presence of biases in automated decision processes on ML models where fairness from an ethical standpoint is concerned with processes that are morally right. There are multiple techniques to improve the fairness in the ML pipeline, it firstly starts with identifying the protected variables such as gender and age, and the proxies which are features that have high correlation with the protected variables and can still contain sensitive information. A possible technique to address fairness is by simply removing all sensitive attributes and proxies. Another approach is assuring that ML predictions are uncorrelated with the sensitive attributes. One can also assert equal metrics across groups so that there is no bias towards a particular sensitive attribute. Different stakeholders are interested in using different metrics so it is important to select match the interest of the stakeholders with the current fairness metrics. A framework that addresses biases is the aif360 framework in Python that contains methods for removing biases in the pre-processing, in-processing and post-processing stages. Furthermore, it also contains a lot of fairness metrics to evaluate the fairness of the model.

**Q3:** Select a case study: one of the machine learning systems you developed in the first part of the course. *(Recommended max 400)*

## Case 3

I have chosen the case from the third assignment: In this case, ML is used to predict which primary schools perform well and which do not. We don’t measure the school performance directly, but we can predict the expected share of school leavers with an advice for higher education. Then we can compare this with the observed share; the difference may then be explained by the school performance. If the expected share of a school is much higher than the observed share, the school might not perform well.

**Dataset**

The urban and school characteristics are embedded in the used dataset. It contains data of 6293 schools concatenated with neighborhood data in which the school is located. Data on the school advices are collected by the ministry of education to monitor school outflow.

**Stakeholders**

The main stakeholder of interest is the ministry of the Netherlands to monitor the performance of the schools. We could use this model to capture the performance in terms of difference between expected share and true share and list the most underperforming schools which would be of most interest for the ministry for in depth assessment. In this way, the ministry could improve the overall school performance by tackling the underperforming schools.

Society is also a stakeholder since they want fairness in the prediction process of the SHARE\_HIGH variable. There are sensitive variables in the data such as the gender and ethnicity share in a neighborhood of a school which can lead to unfairness in predicting the SHARE\_HIGH variable. There might be an underlying negative correlation between these protected variables and SHARE\_HIGH which would make the model unfair.

**Fairness and mitigating bias**

To keep all stakeholders satisfied we aim to have a fair, well performing model. For this we would need to mitigate bias without losing accuracy. There could be biases introduced in multiple stages in the ML pipeline. In the preprocessing stage, we could take a look if there are any historical and representation biases. We would want that our dataset (school + urban data) is a reflection of the current reality of the world. We could then further investigate what the protected attributes and proxies are and determine the correlation between those and our predicted variable. Options for mitigating bias then could be by removing the sensitive features, or by performing Disparate Impact Remover using the AIF360 framework to increase fairness or by obtaining equal metrics across groups.

**Limitations of fairness approaches**

The main limitation is that there is a conflict of interest of multiple stakeholders, optimizing one fairness metric in interest of society can negatively impact the overall accuracy of the model which our main stakeholder (the ministry) would not favor. This also holds the other way around.