Audit project: Salary predictions

Assessing the responsibility and ethical aspects of a ML model

Algorithmic Auditing

- A process to open up the black-box of a ML system
- Enhance documentation, traceability and transparency
- Covers goal, context as well as technical aspects
- AI regulation in the EU coming up, partly to facilitate audits
- Critique of the audits:
    - Who audits the auditors?

Salary Raise Prediction

The small start-up company 2Cool2Care have closed their latest funding run and are hiring a bunch of people at the same time. To keep things simple and efficient, they plan to have two levels of compensation (High and Low) and to use ML to automate which of the new employees get which of the salary levels.

You've been tasked with doing a algorithmic auditing on their model. They would like know stuff like the risks involved with their method, how different groups are affected and if the features they use make sense. They'd really like to see numbers on your claims!

Setup

The **goal** is to scrutinize the transparency, fairness, sustainability and data privacy of the model.

- 3 groups, one for each topic/theme
- Choose yourself which group to join!
- Evaluate the model by the questions asked under each topic, or in any way you find suitable or exciting!
- If you have time, make suggestions for improvements
- Present at 11.30!

Climate impact

Try to answer one or several of these questions:

- What is the estimated carbon emission for this model?
- Do you think it’s the training or prediction phase that would produce the highest larger carbon footprint in this case?
- What do you propose to the company take action on to reduce their climate impact?

Fairness

Try to answer one or several of these questions:

- Which fairness metric would be suitable to evaluate this method?
- What is the value on this fairness metric on the model as of today?
- What do you propose to the company take action on to improve this metric?

Transparency

Try to answer one or several of these questions:

- What is the feature importance of this black-box model? Does it seem legit?
- Choose two people from the test set and inspect their predictions. If any of them were to get the low prediction, create a counterfactual explanation to let them know what they need to change to receive the higher salary?

Tips

- To load the model artifact, run:

```
from tensorflow import keras
model = keras.models.load_model('model/salary_model')
```

- Load the data files and look at it in a notebook, Excel or whatever you are comfortable with.