**Fairness - intermediate**

**Module 4: Technical interpretation of fairness/bias (2 lessons)**

This module will provide an understanding of fairness and bias in the context of technology development

1. Understanding fairness and bias requirements with respect to data available, models to build and end-user requirements
2. Definition of common metrics/quantification methods of bias and fairness and bias mitigation techniques
3. Overview of available open-source tools for fairness/bias assessment

Apply learnings to use cases from TRI’s projects and products. The learner will be provided with Jupyter notebooks showing application of different techniques to assess fairness/bias on datasets and models performances together with multiple choice questionnaires and code snippets to modify.

**Syllabus material**

*a. Understanding fairness and bias requirements with respect to data available, models to build and end-user requirements*

**Why do we care about fairness/bias in tech?**

What is fairness? General definition: impartial and just treatment or behaviour without favouritism or discrimination. How is this relevant to data science? AI and ML technologies have been rapidly rolled out into various systems governing people’s lives. These models form components of decision-making pipelines. It is important to understand impact of these technologies in respect to the consequences of their application. Models inform actions, actions inform outcomes. Examples include policing algorithms (COMPAS), facial recognition (PULSE image depixeliser), HR tools for hiring etc. In general, we want to make not just the models fair, but the outcomes as well. However, from the perspective of data science, we only have influence over the models and the implementation of data, not the actions that are undertaken as a result. Therefore, it is extremely important to communicate with end users about diverse types of bias resulting from data, model selection and model tuning, so that models can be tuned to meet the needs of the end users and impacted groups, and end users can make informed decisions about how to act on the model outputs.

Examples:

Graphical user interface, application

Description automatically generated

Output of the PULSE face depixeliser. The underlying image is of Barrack Obama, but the output from the model, informed by biased training data, outputs a caucasian face.

**What does bias look like from a data science perspective?**

* Certain aspects of a dataset are more heavily weighted and/or represented than others. This can cause low accuracy and skewed outcomes.
* Different performance for diverse groups of people

Some sources and types of bias in data:

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Definition | Relevance | Mitigation |
| Confirmation bias | Tendency to interpret or seek out new evidence that confirms previously held beliefs | Seeking out data relevant to our preconceptions or building models premised on preconceptions. | Record all starting assumptions and priors before beginning technical work and resist temptation to generate hypotheses based on presumptions. |
| Selection bias | Bias in the selection of samples used to train/test a model | Models trained on an unrepresentative subset of the overall data will perform poorly in contexts that are not similar to the training data | Compare sampled data to the final use-case situation to check that the sampled data is truly representative of the contexts the model will be deployed in. Or trial some outcomes of the model in different contexts. I.e., test the performance of a model trained on US data to UK |
| Historical bias | Historical and or social prejudices and beliefs are reflected in data. | Especially relevant for training machine-learning models, as there may be a tendency for a model trained on historical data to try and repeat that same data. | Engagement with groups that will be affected by a model or technology and efforts to include groups that may be mis- or under-represented in historical datasets |
| Survivorship bias | Tendency to focus on characteristics of salient examples rather than whole populations. | This kind of bias arises because of inability to distinguish correlation and causation. Over-indexing based on salient examples leads to under-focus on the properties of the entire population. | Do not over-weight based on true positives or what survives a process, consider the full process and the multiple pathways through it, considering what influences individual trajectories through a system. |
| Reporting bias | Datasets tend to be based around recording noteworthy incidents. For example, the word ‘laugh’ is more common than ‘breathe’ in writing because laughter is more worthy of comment than breathing, but laughter (unfortunately) is not more common than breathing | Skews datasets and the outputs of models based on them to reflect the perspective of the data / data recorders. Related to selection bias. |  |

Diagram

Description automatically generated

Illustration of survivorship bias. Hypothetical damage indicators from planes returning from a battlefield appears to indicate areas where the planes get damaged the most. However, we are only looking at the planes that **return**, not the planes that are shot down/destroyed. That is, we can conclude that a plane receiving damage in these marked areas is potentially still able to return. If you were going to reinforce sections of this plane, where would you focus your efforts?

### **What does fairness look like from a data science perspective?**

The definition of fairness is very context dependent. The key here is to understand: **what does fairness mean to you end user?**

Below are some concepts of fairness to think about

* Group fairness - diverse groups of people should be treated the same on average. This could mean applying the analysis to everyone in the same way.
* Individual fairness - individuals who are similar should be treated similarly. This could mean that a model applied to two people who have similar features, should give similar predictions/results, I.e., fairness through awareness

#### **How to quantify fair model performance?**

* The performance of the model for all groups should be the same
* The performance of the model in the lowest performing group should be as high as possible
* The variation in performance across diverse groups should be as low as possible

### **Case study scenario - suppose we are building an algorithm that predicts whether a given individual is likely to go missing in the next 6 months**

#### **Questions to explore**

* Who are the end users?
* What are they most concerned about in terms of bias? Protected characteristics?
* Are there particular groups that are particularly important in understand bias?
* What is their definition of fairness?

#### **Bias**

There will be some biases in the data. We have identified with the end user the important characteristic/groups to investigate when considering biases. We have identified what biases exist within those groups. Now the questions are, does the model amplify these biases?

Return to previous scenario. Suppose we found that the end user was particularly concerned about gender bias and so that is an important characteristic to explore. First, we look at the results in the test dataset:

* What is the overall performance of the model?
* What is the performance of the model for men? What about for women?
* Are the performances for these two groups significantly different from the overall performance? If so, it suggests a gender bias in the model
* Define significance

Here is another way to explore bias in the test set. What proportion of people in the test dataset that have gone missing are men? Let us assume the answer to this is 70%. What proportion of people in the test dataset that are predicted to go missing are men? If these answer to the second question is significantly higher than 70%, it suggests your model is perpetuating the bias of the data. In other words, the model is heavily using the fact that more men go missing than women, which exaggerates the gender imbalance that already existed within the data.

b. Definition of common metrics/quantification methods of bias and fairness and bias mitigation techniques

There are many ways to detect and quantify bias, and each of those metrics suggests a corresponding method to mitigate the bias by optimising/minimising the metric. However, it is important to note that not all metrics are compatible with one another. Indeed, it is possible to demonstrate that some fairness criteria conflict with one another. This means that the metrics we choose to focus on must be selected with careful consideration to the actions and outcomes our model will inform.

**Case study**: Implementing a model to determine whether to accept candidates for a particular project. We want the model to select candidates based on whether they are a good fit for the project. Let us assume we are hiring from two populations, blue and orange.

**Some examples of bias metrics**

Positive rate – the rate at which a model returns ‘positive’ results (n.b. the ‘positive’ here should not be confused with ‘positive’ as a value judgement, but ‘positive’ in the sense of a test result or detection)

True positive rate – Understood at the ratio of true positives to the total number of positive results.

Unaware model

It might be tempting to create a model that is ‘blind’ to sensitive characteristics. In this case, we could simply drop all data that explicitly identifies which group each candidate belongs to and create a model that is unaware that the subgroups even exist, which seems like it would make discriminating on that basis impossible for the model. However, this approach ignores the possibility of redundant encodings existing. If there are already disparities between the orange and blue subgroups, it may be possible to infer subgroup by looking at the other, non-sensitive characteristics of the candidates. It never possible to create fairness through unawareness, especially for machines, because if there is a pattern to be recognised in the data, the algorithm can and will be sensitive to it.

Statistical or demographic parity

Requires that the positive rate of the model be independent of subgroup. For example the chances that a model returns a positive result should be the same for the two subgroups.

One of the advantages to this formulation of fairness is that it’s independent from the target variable (the candidate’s qualification for the task), making it well suited to achieving fairness in scenarios where that target variable is difficult to assess.

However, this is also a flaw. Since demographic parity is focused only on the outcome of the model it doesn’t necessarily ensure that the two populations are treated fairly, only that the outcomes for both are not disproportionate. We could implement a trained model to select from blue, and then roll a die to select from orange and as long as the chance of being successful is the same for both blue and orange, demographic parity is achieved. This is known as ‘laziness’.

Equalised Odds

Equalised odds requires that a classifier or model has the same error rate for all subgroups / subpopulations. In the hiring process model, this would mean that as long as a candidate is qualified, the outcome of their application should not be dependent on whether they are from the blue or orange group. While this sounds ideal, owing to biases that are present in data, it is difficult to develop and train models that perform the same for all subgroups without compromising on overall model performance.

Equal Opportunity

A relaxed version of equalised odds, it applies the same criteria only to the true positive rate, I.e. the outcome of the model should only be the same for both subgroups when the candidate is qualified. This results in better model performance but may also come at the cost of demographic parity, meaning that there will still be a disparity in the overall success rates of the two groups.

Further, whilst equality of odds and equality of opportunity aim for fair treatment, they are not able to account for biases that exist in the historical data. If there are disparities between the groups that exist in the data, then models aiming for equality of odds and opportunity will replicate these biases. That is, if there is an imbalance between the two classes to begin with, then there will be imbalances reproduced by the model.

Which definition of fairness to use depends on the use-case, with demographic parity aiming to try and correct for biases over a long period of time, and with equalised odds/opportunity ensuring fair treatment during the process, but not addressing bias in the long term (instead perpetuating or even exacerbating it).

c. Overview of available open-source tools for fairness/bias assessment  
SHAP – allows us to assess the importance of different features in a model. This is an important tool when trying to illustrate the relative influence of variables in the model and can be used to identify bias. If the SHAP values are large for any sensitive variables, or variables that are believed to be correlated or proxies for sensitive variables, then this can be a signal that the model may be producing biased outputs which require evaluation.

On top of this, there are

Fairlearn  
Aequitas

Which can both be used to do fairness testing and model selection based on hyperparameters – show some example outputs of this and discuss how to go about selecting a model.

#### **Hands on example (notebook support)**

* Load a dataset
* Plot various histograms, scatter plots, split by the target variable.
* Identify where are the imbalances?
* Train a model on the whole population, evaluate performance
* Check performance for different subgroups and comment on disparities
* Try a mitigation strategy (resampling and possibly fine tuning with aequitas?)
* Check performance differences