Lynxx Pty Ltd
ABN 17 617 542 622
www.lynxx.com.au

Case Study:
Responsible Algorithmic Decisions

Background

As a specialised data and advanced analytics company, our everyday work at Lynxx consists of working with machine learning and decision support models including artificial intelligence or AI[i].

This case study demonstrates our algorithmic decision making auditing tool, which provides confidence in the processes and modelling undertaken by an organisation. Our auditing tool allows us to evaluate an algorithmic approach and to identify potential biases that are inherent in the training data as well as any biases that may be introduced during modelling decisions. The aim of this tool is to promote trust in decision making tools as well as improve the quality of decision models in general. We believe that this trust will translate to increased confidence in governance processes and thus enhance brand trust and reduce reputational risk.

Approach

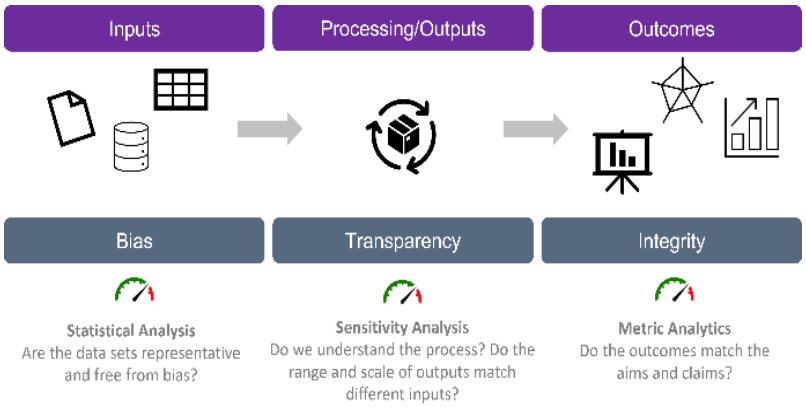
As a professional data science organisation, we were increasingly being asked to build decision making tools (algorithms, data models, machine learning functions, intelligence tools) but seeing that there was a lot of focus on a decision and, in our view, insufficient focus on understanding how inputs, or the models themselves, impacted the outcomes of those models, this gave rise to an increasing level of discomfort around the ethics of algorithmic decision making. Given that a team of professional data scientists were struggling to fully understand how features of models, including at times those that we were building ourselves, were impacting decisions, we recognised that the wider public would struggle with how a model made its decisions. We decided to term this discomfort as relating to the fairness in the model. Not fairness in the subjective sense "it's not fair" but whether the inputs, processes/outputs and outcomes were fairly represented in the model, and vice versa, at a level of granularity that would allow someone to whom the model was being applied, would make sense of what was happening within the model.

> *Algorithmic Fairness: the extent to which the inputs, processes, outputs and outcomes of a model fairly represent the environment in which the model is being applied, and vice versa.*

We decided therefore to define a framework that would allow us to maximise fairness in our own models but soon realised that this approach could be used more widely to audit external models.

Our Algorithmic Fairness Framework

At a high level, our fairness framework identifies the key potential data characteristics at each of the three main stages of modelling (inputs, processing, outcomes), and will provide a way of improving fairness within the model. The precise measure of fairness depends on the stage as follows:



Stage	Inputs	Processing / Outputs	Outcomes
Fairness Measure	Bias	Transparency	Integrity
Description	When the input data contain biases, the model cannot fairly represent the real world.	Does the process do what it says it will do in the model.	Do the metrics match the aims and claims of the process.
Test	We use statistical analysis to test the bias in the input data (or to ensure the biases are disclosed and accounted for in the model).	We use sensitivity analysis to test the outputs based on different inputs to validate the processing claims.	We use matric metric analytics to check that what is being presented is actually measured by the model.

Case Study

To demonstrate our framework, we identified a use case with publicly available data and a high accuracy modelling approach. The data[ii] and model[iii] described a fictional university recruitment open day and were obtained from Kaggle. The data and model were evaluated for fairness through our audit model. The analysis indicated that the input data used would be relatively fair (limited bias in inputs). Analysis of the modelling decisions identified a data leakage affect (in the cleaning of data), which was done to achieve high accuracy, but this data leakage also increased the existing bias to a level that would be considered unfair if processed as raw data (so there was insufficient transparency in the processing). In this particular analysis, the outcome was a significant under-representation of females as suitable candidates – an outcome that was neither intended nor transparent. As part of the framework, we identified potential mitigations that could be used to increase the integrity of the outcomes, and would in our view, increase the acceptance from the wider community. These mitigations included statistical techniques to reduce the processing bias as well as processes to improve decision making while building this type of model.

Learnings

This case study (and several subsequent model audits) has demonstrated the need to take fairness in algorithmic decision making seriously. We can point specifically to the situation where unbiased input data can be biased in the processing steps (in the name of data cleaning) and the subsequent incorrect belief that just because unbiased data went into a model, that the outcomes were fair. We also draw a distinction between known biases (which may be understood and accounted for, or not, in the outcomes) but would also argue that to engender trust in algorithms, complete transparency is needed, so that people having their data processed by algorithms can understand how their data are being used, and to challenge those processes if needed. In a world where more 'intelligence' is being used, we argue that artificial intelligence cannot be assumed to be intelligent.

---

[i] See https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf We are deliberately careful with the concept of "AI" as common language and marketing usage of the term has a very broad definition (stretching in extremis to anything with a computing process attached). Under our definition, we use the Stanford University Human-Centered Artificial Intelligence (HAI) definition where Artificial Intelligence is narrower than "Intelligence". Intelligence is what we would argue common language usage tends to incorrectly label "AI", rather than the more precise Artificial Intelligence processes that have a self-learning mechanism.

[ii] See https://www.kaggle.com/benroshan/factors-affecting-campus-placement

[iii] See https://www.kaggle.com/bariscal/placement-prediction-with-ml-94