

Portfolio Question 10.3.P - Interpretable or Black Box?

Daniel Lawson University of Bristol

Portfolio 10.3.P (v4.0.0)

Portfolio Question 10.3.P - Interpretable or Black Box?

- ▶ Should high-stakes decisions about people be restricted to interpretable models?
 - ▶ Contrast **interpretable models** with **post-hoc explanations** of black boxes, and consider what each means for detecting and challenging unfairness.

Why explanations matter

- ▶ Algorithms help make decisions about **credit**, **hiring**, **benefits**, **medical treatment** and **bail**.
- ▶ An explanation is needed to:
 - ▶ Let the affected person **challenge** a decision, or know what to change;
 - ▶ Let the organisation **audit** the model for bias and proxy discrimination;
 - ▶ Let regulators and courts check that the **law** is being followed.
- ▶ Recall from Part 1: GDPR gives rights around solely automated decisions, often described as a “right to an explanation”. How much it actually requires is still debated¹.
- ▶ **Loomis v. Wisconsin (2016)**: a defendant was sentenced partly using the proprietary COMPAS risk score. He argued he could not challenge a model he could not see. The Wisconsin Supreme Court allowed its use, with warnings attached.

¹Wachter, Mittelstadt and Floridi. “Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation.” International Data Privacy Law 7.2 (2017).

Two routes to an explanation

- ▶ From the lecture, there are two broad approaches:
 - ▶ **Interpretable models**: decision trees, rule lists, sparse linear models and scoring systems. The model *is* the explanation.
 - ▶ **Post-hoc explanations** of a black box: counterfactuals (“you would have been accepted if. . .”), feature attributions (e.g. LIME, SHAP), and surrogate models fitted to mimic the black box.
- ▶ Post-hoc explanations are attractive because they let us keep the most accurate model, whatever it is.
- ▶ Counterfactual explanations were proposed as a way to satisfy GDPR **without opening the black box**².

²Wachter, Mittelstadt and Russell. “Counterfactual explanations without opening the black box: automated decisions and the GDPR.” *Harvard Journal of Law and Technology* 31 (2018).

The case for interpretable models

- ▶ Cynthia Rudin argues that we should **stop explaining black boxes** for high-stakes decisions, and use interpretable models instead³.
- ▶ Her main claims:
 - ▶ The accuracy-interpretability trade-off is often a **myth**. On tabular data with meaningful features, a well-built interpretable model is usually about as accurate.
 - ▶ A post-hoc explanation is a **second model**. If it were always faithful, we would not need the first one; if it is not, it can mislead.
 - ▶ Black boxes make errors in the data (typos, missing values, leakage) much harder to spot.
- ▶ Example: a short **rule list** (CORELS) predicted re-arrest about as well as COMPAS, using only age, sex and number of prior offences:
 - ▶ *If age 18-20 and male, predict yes; else if age 21-23 and 2-3 priors, yes; else if more than 3 priors, yes; else no.*

³Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1 (2019): 206-215.

Problems with post-hoc explanations

- ▶ **Unfaithful**: an explanation can agree with the black box on average while being wrong about *why* it made a particular decision.
- ▶ **Gameable**: it is possible to build a classifier that discriminates on a protected attribute, yet produces innocent-looking LIME and SHAP explanations⁴.
 - ▶ An organisation could therefore pass an audit based on explanations alone.
- ▶ **Not unique**: many different counterfactuals can explain the same decision. Which one is shown, and who chooses?
- ▶ Recall from the lecture: a meaningful counterfactual must update **all correlated attributes** (postcode, social network, etc.), not just the protected one.

⁴Slack et al. "Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods." AAAI/ACM Conference on AI, Ethics, and Society (2020).

The case for black boxes

- ▶ For **images, text and audio**, there are often no meaningful hand-made features, and deep models clearly outperform interpretable ones.
 - ▶ Should a less accurate cancer screening model be used because it is easier to explain? Errors cost lives too.
- ▶ **Interpretable does not mean fair**: a transparent rule using postcode can still discriminate by proxy. Transparency makes this visible, but does not fix it.
- ▶ **Gaming by users**: if people know the exact rule, they may change their behaviour to satisfy it without changing the underlying risk.
- ▶ **Commercial secrecy**: companies argue that revealing a model exposes their intellectual property. Is that a legitimate reason, or simply a cost of making decisions about people?
- ▶ Building good interpretable models takes **time and expertise**. Black boxes are often cheaper.

Discussion

- ▶ Should there be a **legal requirement** for interpretable models in some domains (e.g. criminal justice, credit), and if so, which?
- ▶ Is it enough for a model to be interpretable **to the regulator**, rather than to the person affected?
- ▶ How much accuracy would **you** give up for interpretability?
Does it depend on who bears the errors?
- ▶ Would you rather be judged by a transparent rule that is slightly wrong, or an opaque model that is slightly more often right?