Explainable Machine Learning in Deployment

Umang Bhatt,¹,² Alice Xiang,² Adrian Weller,¹,³ & Peter Eckersley²
¹University of Cambridge, ²Partnership on AI, ³The Alan Turing Institute

Correspondence to: usb20@cam.ac.uk

Summary
We study how various organizations deploy explainable machine learning tools in practice. Through semi-structured interviews with over thirty individuals from across twenty organizations, we find that feature importance was the most common explainability technique, and Shapley values were the most common type of feature importance explanation. The most common stakeholders were machine learning engineers (or research scientists), followed by domain experts (loan officers and content moderators). We also conclude the following:

- Model explanations are mostly used by machine learning scientists internally for model debugging and are not shown to end users.
- The goals for explainability are not clearly defined within organizations.
- Technical limitations make it hard to reliably deploy explainability tools in real-time.

Explanation Types
We study explainability as one form of algorithmic transparency that scrutinizes how a model behaves via three common explanation types.

Stakeholders
Most organizations deploy explainability atop their existing ML workflow for one of the following stakeholders, who are not necessarily domain experts themselves.

Table: Interviews
We present six example use cases from our interviews. For each use case, we define the domain of use, the model’s purpose, the explainability technique used, the stakeholder consuming the explanation, and how the explanation is evaluated.

Establishing Clear Desiderata
We suggest a framework for organizations to decide what type of explanation to deploy depending on who is consuming the explanation.

1. Identify stakeholders. Who are your desired explanation consumers? Typically this will be those affected by or shown model outputs.
2. Engage with each stakeholder. Ask them some variant of “What would you need the model to explain to you in order to understand, trust, or contest the model prediction?” and “What type of explanation do you want from your model?”
3. Understand the purpose of the explanation. Once the context and helpfulness of the explanation are established, examine what will be done with the explanation [1].
   - Static Consumption: Will the explanation be used as a one-off sanity check or shown as reasoning for a particular prediction?
   - Dynamic Model Updates: How does the stakeholder interact with or provide feedback to the model after viewing the explanation?

Limitations
Our interviews shed light on prohibitive limitations of various explainability techniques. These include:

- Spurious correlations in datasets can lead feature level explanations awry.
- Sample importance is computationally infeasible to deploy at scale.
- Few organizations raise privacy concerns after explanations were shown to be useful for model inversion [2].

References