Explainable Machine Learning in Deployment

Umang Bhatt,[1,2] Alice Xiang,[2] Adrian Weller,[1,3] & Peter Eckersley[2]
[1]University of Cambridge, [2]Partnership on AI, [3]The Alan Turing Institute
Correspondence to: usb20@cam.ac.uk
https://arxiv.org/abs/1909.06342

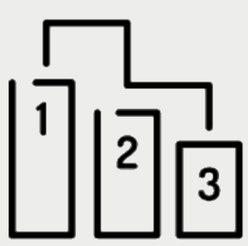
Summary

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.

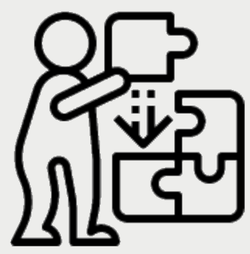
Explanation Types

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



Feature Importance

Sample Importance



Counterfactuals

Stakeholders

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

Executives

Engineers

End Users

Regulators

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.

Domain	Model Purpose	Explainability Technique	Stakeholders	Evaluation Criteria
Finance	Loan Repayment	Feature Importance	Loan Officers	Completeness
Insurance	Risk Assessment	Feature Importance	Risk Analysts	Completeness
Content Moderation	Malicious Reviews	Feature Importance	Content Moderators	Completeness
Finance	Cash Distribution	Feature Importance	ML Engineers	Sensitivity
Facial Recognition	Smile Detection	Feature Importance	ML Engineers	Faithfulness
Healthcare	Medicare access	Counterfactual Explanations	ML Engineers	normalized ℓ_1 norm

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

[1] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[2] Smitha Milli, Ludwig Schmidt, Anca Dragan, and Moritz Hardt. Model reconstruction from model explanations. *In Proceedings of ACM FAT* 2019*, 2019.