Interpretability of machine learning

Katsiaryna Dubrouskaya

Many high-performing models, which are used nowadays across diverse fields, operate as "black boxes," meaning that internal decision-making processes are hidden and difficult for humans to understand. Due to lack of transparency there is a growing need in interpretability to trust model predictions, debug unexpected behaviors, etc.

Some basic terminology:

Interpretability itself (ML) – as the extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model.[1]

Post hos analysis – a statistical analysis conducted after a study has been concluded and the data collected.

Predictive accuracy – measures the ability of the predictive model to predict (the proportion of correct predictions on the test(unseen) data).

Descriptive accuracy (interpretation context) – the accuracy of the interpretation method (objective judgement of the relationships learned by machine-learning models).[1]

Relevancy – refers to the levels of “meaningfulness” which interpretation provides to the intended audience

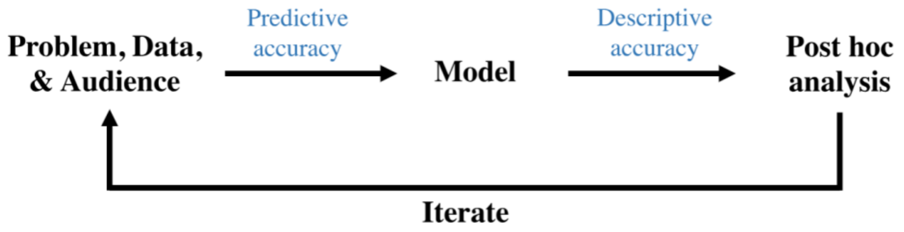


Fig. 1. Overview of different stages (black text) in a data–science life cycle where interpretability is important and two transitions where errors can arise. [1]

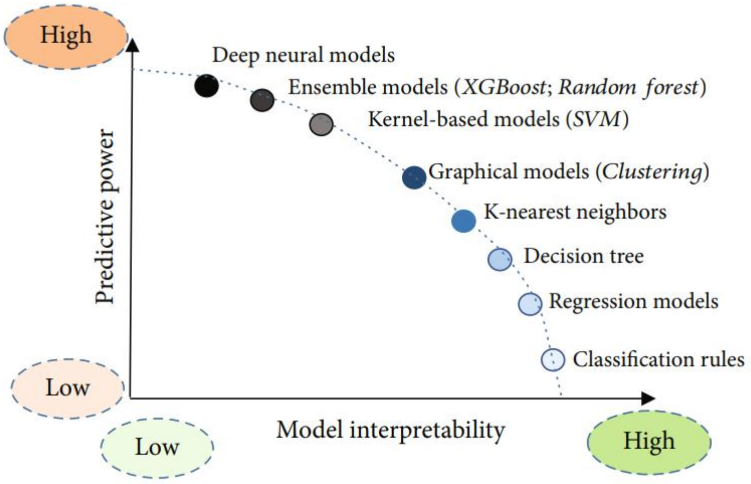
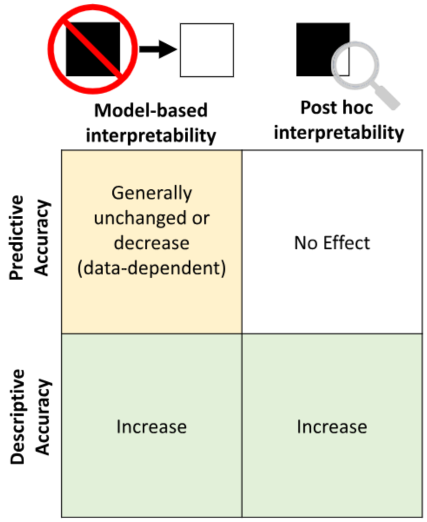


Fig 2: Predictive power vs. interpretability trade-off. [2]

A trade-off between the model interpretability and predictive power is commonly observed, as shown in the figure on the left. As the model gets more advanced, it becomes harder to explain how it works. High interpretability models include traditional regression algorithms (linear models, for example), decision trees, and rule-based learning. On the other hand, low interpretability models include ensemble methods and deep learning where the black-box feature extraction offers poor explainability.[2]



	Model-based interpretability	Post hoc interpretability
Predictive Accuracy	Generally unchanged or decrease (data-dependent)	No Effect
Descriptive Accuracy	Increase	Increase

If the interpretability is focused on methods in modeling or post hos analysis stage, it’s called **model-based interpretability**. (which means simpler models(so human could understood them) which can results to the lower accuracy and usually require retraining).

On the other hand, if interpretation methods take a trained model as input and extract information about what relationships the model has learned, it’s called **post hos interpretability**. (no model modifications; is helpful on complex and high-dimensional data)

Fig. 3. Impact of interpretability methods on descriptive and predictive accuracies.[1]

Into the model-based interpretability we can highlight important approaches, such as:

- Sparsity – limitation of the number of non-zero parameters in a model.
- Simulatability – ability of a model that humans can understand and reason about the entire decision-making process.
- Modularity – model is modular if you can divide decision process into meaningful parts which can be interpreted independently.
- Domain-Based Feature Engineering – usage of expert knowledge in construction informative and meaningful input features.
- Model-Based Feature Engineering – usage of unsupervised learning and dimensionality reduction techniques to construct features from data by uncovering underlying structures or creating lower-dimensional representations. [1]

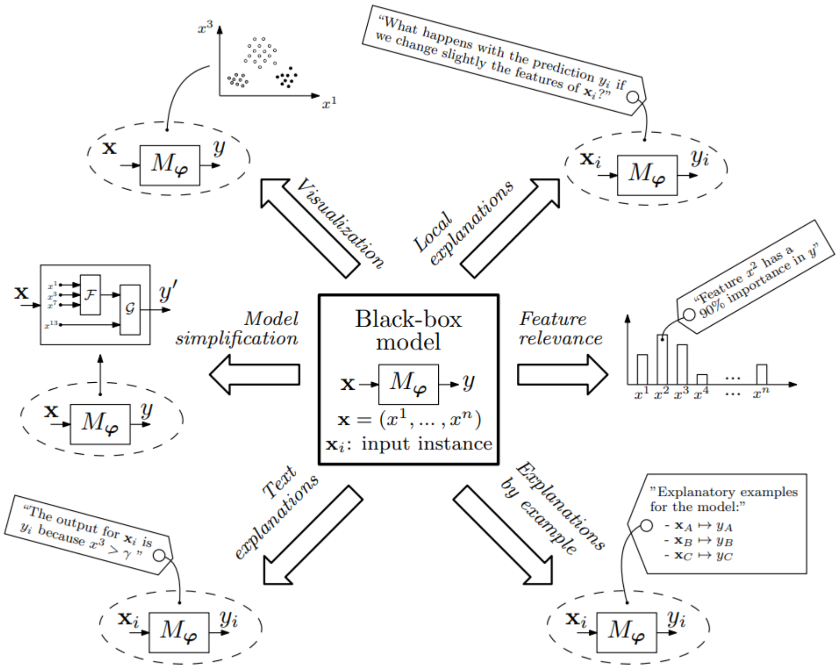

Black-box model
$\mathbf{x} \to M_\varphi \to y$
$\mathbf{x} = (x^1, \dots, x^n)$
$\mathbf{x}_i$: input instance
Visualization
Local explanations
Model simplification
Feature relevance
Text explanations
Explanations by example
“What happens with the prediction y_i if we change slightly the features of $\mathbf{x}_i$?”
“Feature x^2 has a 90% importance in y”
“The output for $\mathbf{x}_i$ is y_i because $x^3 > \gamma$”
”Explanatory examples for the model:”
- $\mathbf{x}_A \mapsto y_A$
- $\mathbf{x}_B \mapsto y_B$
- $\mathbf{x}_C \mapsto y_C$
x^1
x^3
y'
$\mathcal{F}$
$\mathcal{G}$
x^1 x^2 x^3 x^4 … x^n

Fig. 4. Conceptual diagram showing the different post-hoc explainability approaches available for a ML model Mϕ. [4]

Post-hoc explainability uses a variety of techniques to improve the interpretability of models that are not easily interpretable by design, such as:

text explanations, visual explanations, local explanations, explanations by example, explanations by simplification and feature relevance explanations techniques.

• *Explanations by simplification* are a group of methods in which a whole new system is built using the trained model that needs to be explained. This new, simpler model usually tries to be as similar to the old one as possible while also being less complicated and maintaining the same performance score. This is one of the approaches of "explainable AI" (XAI).
XAI is criticized for creating explanations that may not be entirely faithful to the original black-box model's computations and can sometimes use misleading terminology or lack meaningful detail about the decision-making process.[3]

Literature:

1. Murdoch, W. James, et al. "Definitions, methods, and applications in interpretable machine learning." *Proceedings of the National Academy of Sciences* 116.44 (2019): 22071-22080.
2. Kumar, Akshi, Shubham Dikshit, and Victor Hugo C. Albuquerque. "Explainable artificial intelligence for sarcasm detection in dialogues." *Wireless Communications and Mobile Computing* 2021.1 (2021): 2939334.
3. Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature machine intelligence* 1.5 (2019): 206-215.
4. Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information fusion 58 (2020): 82-115.