Interpretability of machine learning

Ignacy Makowski

May 11, 2025

Contents

1 Introduction

1.1 Definition of interpretability

We define interpretable machine learning as the extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model. [1]

1.2 Definition of black-box

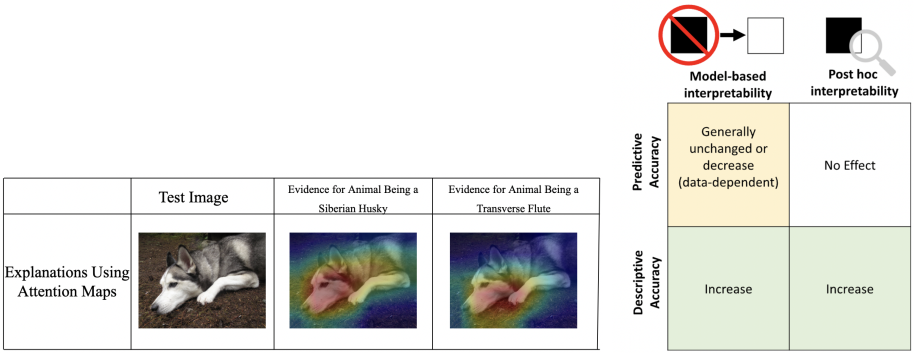
It is any undisclosed or so complicated (*humans can handle at most 7±2 cognitive entities at once*) that it is not possible to understand the function. [2]

1.3 Interpretability vs explainability

The interpretability of a model can be built-in - Interpretable ML or added post factum, usually by means of an additional model - Explainable ML [2]. However, the latter carries the following risks:

- another model requiring explanation is created
- ”explanations” true and false may be identical
- the error from both models accumulates
- the explanation is not inherently linked to the model calculation

Figure 1: Similar explanations [2], accuracy impact [1]



	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

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

1.4 Why is interpretability important? [2]

- Critical for high-stakes decisions (healthcare, criminal justice).
- Auditing predictions for issues like fairness and regulatory compliance (*right to an explanation*).
- Understanding underlying properties (biology).
- Additional sanity check.

2 The Rashomon set argument

2.1 Definition

Consider that the data permit a large set of reasonably accurate predictive models to exist. Because this set of accurate models is large, it often contains at least one model that is interpretable. This model is thus both interpretable and accurate. [2]

3 Evaluation

3.1 Predictive, descriptive, relevant (PDR) framework [1]

- Predictive accuracy - traditional accuracy known, for instance, from supervised machine learning.
- Descriptive - accuracy relevant to post hoc analysis evaluating the proposed explanation. The relevance of this metric is a point of disagreement between the cited articles [2]. It is explained in the following way: *We define descriptive accuracy, in the context ofinterpretation, as the degree to which an interpretation methodobjectively captures the relationships learned by machine-learning models.*[1]
- Relevancy - *We define an interpretation to be relevant if it provides insight for a particular audience into a chosen domain problem.* [1]

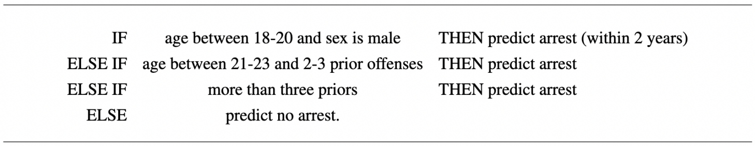
4 Examples

4.1 Rule lists

4.1.1 A

The algorithm works just as well as the closed-source model based on 130 factors used by the US Department of Justice. [2]

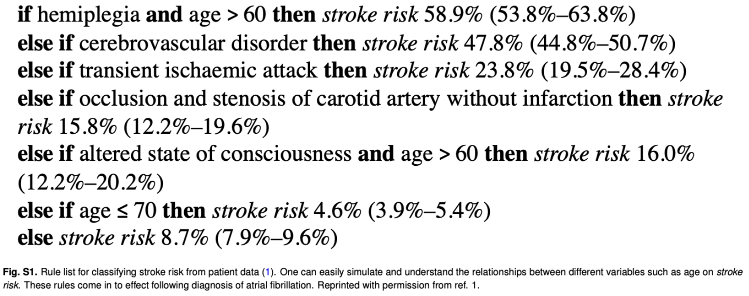
Figure 2: Certifiably Optimal Rule Lists (CORELS) algorithm [2]



IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

4.1.2 B

Figure 3: Another example where an algorithm based on a list of rules has achieved comparable accuracy to traditional ML methods [1]

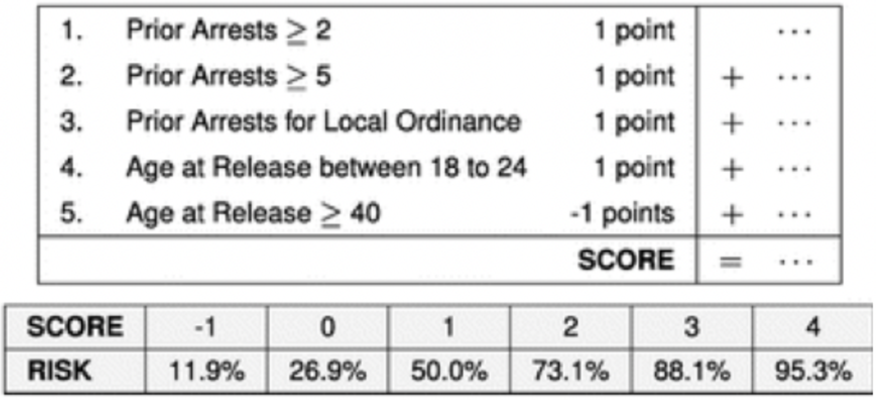


if hemiplegia **and** age > 60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age > 60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age ≤ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)

Fig. S1. Rule list for classifying stroke risk from patient data (1). One can easily simulate and understand the relationships between different variables such as age on *stroke risk*. These rules come in to effect following diagnosis of atrial fibrillation. Reprinted with permission from ref. 1.

4.2 Scoring system

Figure 4: *Scoring system for risk of recidivism from. This model was not created by a human; the selection of numbers and features come from the RiskSLIM machine learning algorithm* [2]



1.	Prior Arrests ≥ 2	1 point		$\cdots$
2.	Prior Arrests ≥ 5	1 point	+	$\cdots$
3.	Prior Arrests for Local Ordinance	1 point	+	$\cdots$
4.	Age at Release between 18 to 24	1 point	+	$\cdots$
5.	Age at Release ≥ 40	-1 points	+	$\cdots$
		SCORE	=	$\cdots$

SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

References

[1] W James Murdoch, Chandan Singh, Katherine Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

[2] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019.