TRAINING: Data Science and AI for Medicine Training School 2026
Day 2: Explainability in Machine Learning

SPEAKER: Matthias Täschner

GEFÖRDERT VOM
Bundesministerium für Forschung, Technologie und Raumfahrt

SACHSEN
Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

Using materials from Robert Haase (DSC ScaDS.AI / Leipzig University)

Explainability in Machine Learning
Why Does Explainability Matter?

Clinical Trust

Physicians need to understand WHY a model recommends a diagnosis to trust and adopt it.

Regulatory Need

EU AI Act and FDA guidelines increasingly require explainability for medical AI systems.

Scientific Insight

Understanding model decisions can reveal new biomarkers and biological mechanisms.

Explainability in Machine Learning
Where does explainability fit in?

Quick recap: the ML pipeline

Explainability is relevant at multiple stages:

- it helps you select the right features,
- understand your trained model,
- and communicate results.

It bridges the gap between a prediction and a scientific explanation.

Explainability in Machine Learning
White-Box vs. Black-Box Models

White-Box (Interpretable)

Linear Regression

$f(x) = w_1x_1 + w_2x_2$

If $w_1 >> w_2$, then x_1 dominates the prediction. You can directly read the model.

Decision Trees

Follow if/then rules to reach a prediction. Each branch is human-readable.

Black-Box (Opaque)

Deep Neural Networks

Millions of parameters in hidden layers. No simple way to trace a single prediction.

Ensemble Methods (large)

Hundreds of trees with deep branching. Individual logic is lost in the ensemble.

Need explanation / interpretation methods!

Explainability in Machine Learning
Explainability vs. Interpretability

Explainability

A logically consistent description of an algorithm with **complete transparency**.

Focus: How does the algorithm work?

Examples:
Reading model weights, inspecting decision tree branches, understanding the math.

Interpretability

Visualization of intermediate results and their **influence on outcomes**.

Focus: What drives the model's decisions?

Examples:
SHAP plots, feature importance charts, Grad-CAM heatmaps.

Explainability in Machine Learning
Feature Importance

The simplest explainability tool:

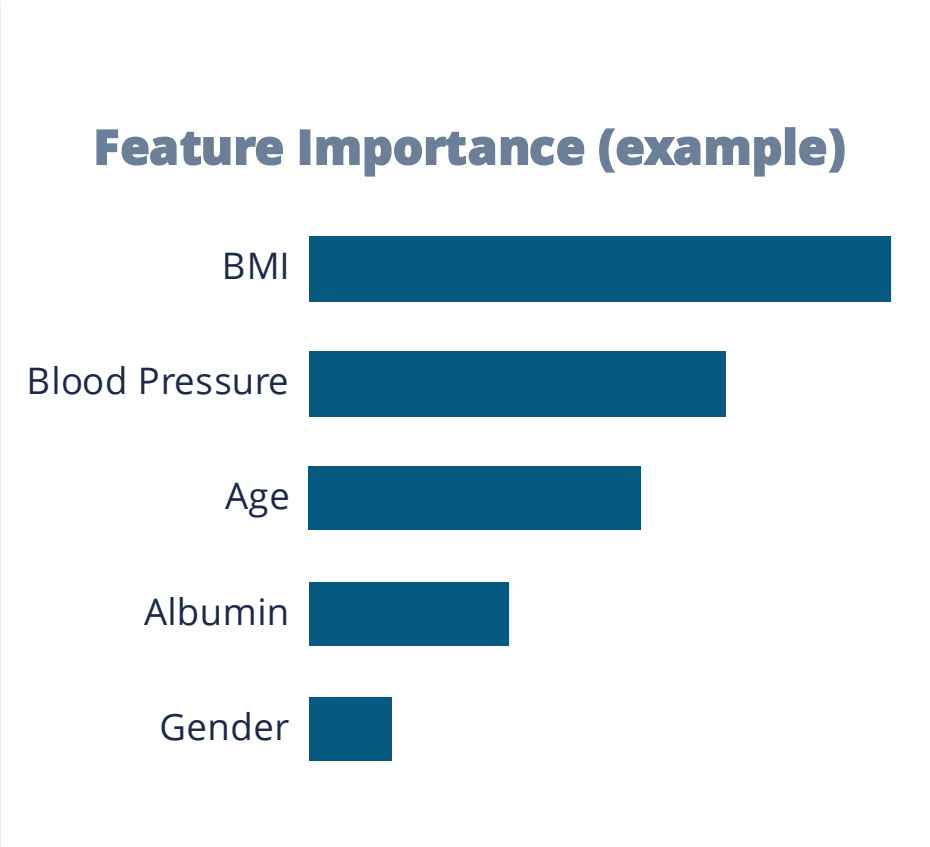
Which features contributed most to the prediction?

Random Forest provides built-in feature importances based on how much each feature reduces prediction error across all trees.

In scikit-learn: *model.feature_importances_*

More robust alternative:

Permutation importance - shuffles one feature at a time and measures how much performance drops



Explainability in Machine Learning
Explanation Methods

Model-Agnostic

Work with ANY model, treat it as a black box.

SHAP
Shapley Additive Explanations. Based on game theory. Shows contribution of each feature per prediction.

LIME
Local Interpretable Model-agnostic Explanations. Approximates model locally with simple model.

Model-Specific

Exploit internal structure of a particular model type.

Grad-CAM
For CNNs. Highlights which image regions activate the classification. Uses gradient information.

Attention Maps
For Transformers. Visualize which tokens/patches the model attends to.

Explainability in Machine Learning
SHAP - The Game Theory Intuition

Think of it like a team game:

Imagine each feature is a player on a team. The team's goal is to make a correct prediction.
SHAP asks: "If this player joins the team, how much does the team's performance improve?"
This is done by testing all possible combinations of players.

1

Try all coalitions

Test every subset of features with and without feature *i*

2

Measure marginal contribution

How much does adding feature *i* improve prediction quality?

3

Average across all orderings

The SHAP value is the weighted average of all marginal contributions

Explainability in Machine Learning
SHAP – Pitfalls: Correlated Features

Correlated features can harm interpretability

The Problem

When two features are highly correlated (e.g., weight and BMI), SHAP splits importance between them.
Each feature may appear less valuable than it truly is, leading to misleading conclusions about what matters.

What To Do

- Check correlations before training
- Consider removing redundant features
- Use domain knowledge to group related features
- Be cautious interpreting importance of correlated features

Explainability in Machine Learning
SHAP – Pitfalls: Other Mistakes

Mistake 1: High feature value means high SHAP value	Not necessarily. It depends on the learned model relationship. A high value can push up, push down, or do either depending on context.
Mistake 2: SHAP proves causality	It does not. SHAP explains the model's behavior, not real-world cause and effect.
Mistake 2: SHAP baseline is zero	Often false. The baseline is the expected model output under the reference data. The SHAP plotting docs refer to this as the expected value / base value.

SHAP tells, for a specific prediction, how much each feature moved the model output away from a baseline, where those contributions are computed as fair averages of marginal effects over many possible feature subsets.

Explainability in Machine Learning
Practical Tips

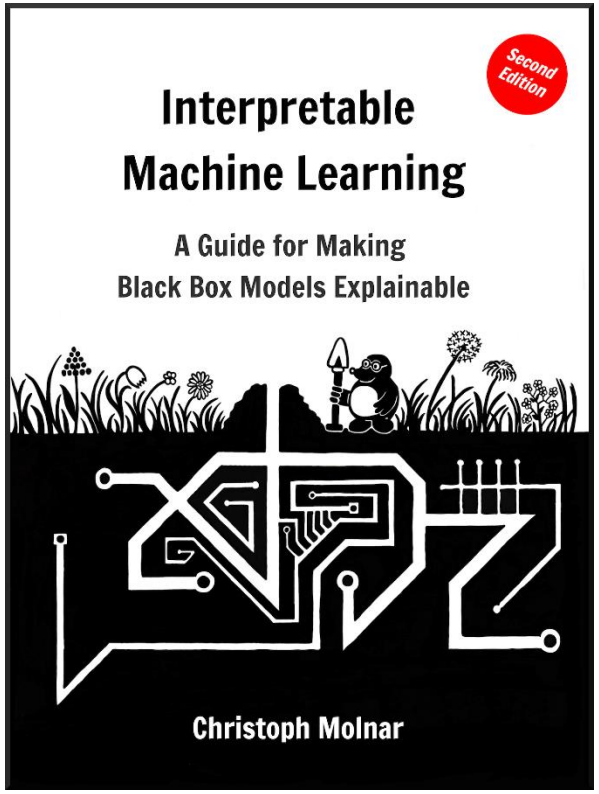
Start simple
Use an interpretable model first (logistic regression, small decision tree).
Move to complex models if needed.

Always plot feature importance
Even a simple bar chart of feature importances can reveal surprising insights about your data.

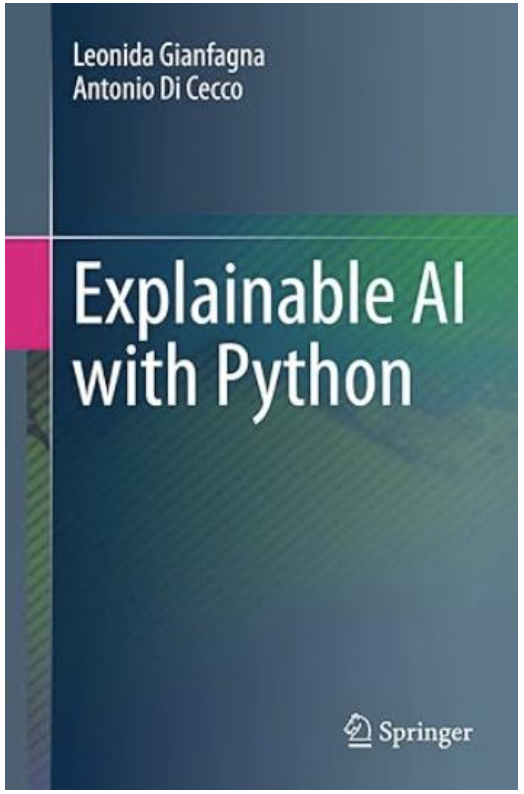
Use SHAP for deeper insight
SHAP summary plots show not just which features matter, but how their values affect predictions.

Validate with domain expertise
If a model relies on features that don't make clinical sense, investigate further.

Explainability in Machine Learning
Read more...



https://christophm.github.io/interpretable-ml-book/



https://www.amazon.de/dp/3030686396

Explainability in Machine Learning

Practice

Explainability with scikit-learn and SHAP

- Navigate to the training materials folder
- Use uv to start JuyperLab
- Open the notebook “day2.3_ml-basics/01_explainability.ipynb”
- We will use the Python packages scikit-learn and shap