Explainability in AI

Machine Learning & Deep Learning

Jawad ALAOUI

Université Gustave Eiffel

NORMA

NORMA

About NORMA
Excellence in AI and Data Solutions

At NORMA, we excel in technology and customer service. Since day one, our team of data and AI specialists has been dedicated to building custom Machine Learning models and multi-agent AI platforms. Our precise engineering approach allows us to tackle complex challenges with innovative and effective solutions.

We've also developed an advanced tool that gives clients real-time performance monitoring of their agents and the ability to test their models instantly. This transparency ensures continuous optimization of deployed systems, guaranteeing top performance and quick adaptation to market changes.

NORMA

NORMA

Jawad ALAOUI

CEO, AI Expert, Tech Leader, Educator

CORPORATE EXPERIENCE

- 15+ years in software development, data analysis and ML.
- Extensive banking and finance experience, including leadership roles at Société Générale, Crédit Agricole, and BPI France.

ENTREPRENEURSHIP

- Co-founder & CEO of NORMA: Leading a tech community that accelerates startups through AI solutions.
- CTO of Indie Plaza: Assists Game Dev in their financing journey

TEACHING

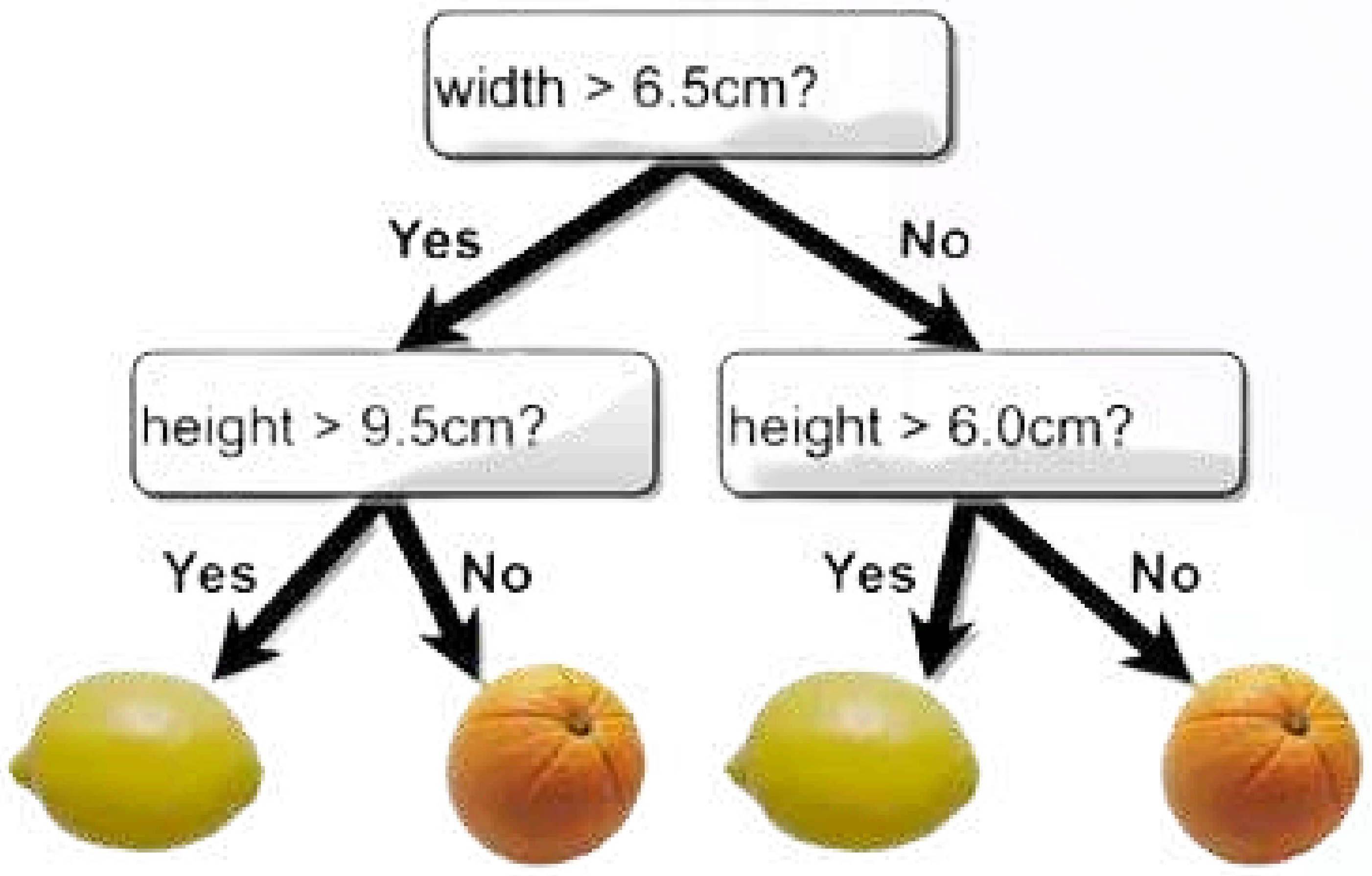
- Part-time Associate Professor at Université Gustave Eiffel: Teaching advanced deep learning and statistical models.
- External Lecturer at UniLaSalle, Centrale Casablanca & Nantes, and Institut Supérieur du Numérique Nouakchott.

JAWAD

The Scope of Explainability in AI module

- Introduction and Motivations
- Taxonomy of Explainability Methods
- Ante Hoc Explainability Methods
- Post-hoc Explainability Methods

Definition of Explainability in AI



Explainability (also called interpretability) is the degree to which a machine learning model's workings and outputs can be understood by humans. In essence, an explainable AI system can articulate why and how a particular decision was made in a way that "makes sense" to a human

Simple models (like linear models or decision trees) tend to be inherently explainable, whereas complex models (e.g. deep neural networks) often behave as black boxes that defy easy interpretation.

Why it's important?

Trust and Transparency

The ability to explain AI decisions is crucial for building trust. When users and stakeholders understand how a model arrives at its predictions, they are more likely to trust and adopt its recommendations. Conversely, a lack of transparency can erode confidence – users feel uneasy "in the dark" about an algorithm's reasoning.

In high-stakes domains (healthcare, finance, etc.), stakeholders demand clear justifications for model decisions before relying on them, as explainability provides reassurance that the AI is making sound and justifiable choices.

Why it's important?

Ethics and Fairness

Explainability is also essential for identifying and mitigating bias in AI systems. Black-box models can inadvertently learn discriminatory patterns from data, leading to unfair or unethical outcomes. Interpretable AI allows developers to detect biased decision rules (e.g. a model overly relying on a sensitive attribute like race, gender or name)

By examining explanations, we can ensure decisions align with ethical standards and correct any unjust or harmful behavior in the model, thus promoting fairness and accountability in AI-driven decisions.

Why it's important?

Regulatory Compliance

In many sectors, laws and regulations now require AI decisions to be explainable. For example, financial and data protection regulations mandate transparency in automated decisions – the EU's GDPR grants individuals a "right to an explanation" for algorithmic decisions that affect them

Similarly, laws like the U.S. Equal Credit Opportunity Act demand that lenders provide specific reasons for loan denials. Explainable AI facilitates compliance by providing auditable justifications for each outcome, helping organizations meet legal standards and avoid liability.

Real-World Implications

Biased AI Model Case Studies

Lack of explainability has led to serious real-world issues, underscoring why it matters.

Reuters

World Business Markets Sustainability Legal Breakingviews Technology Investig

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

October 11, 2018 2:50 AM GMT+2 · Updated 6 years ago

Aa

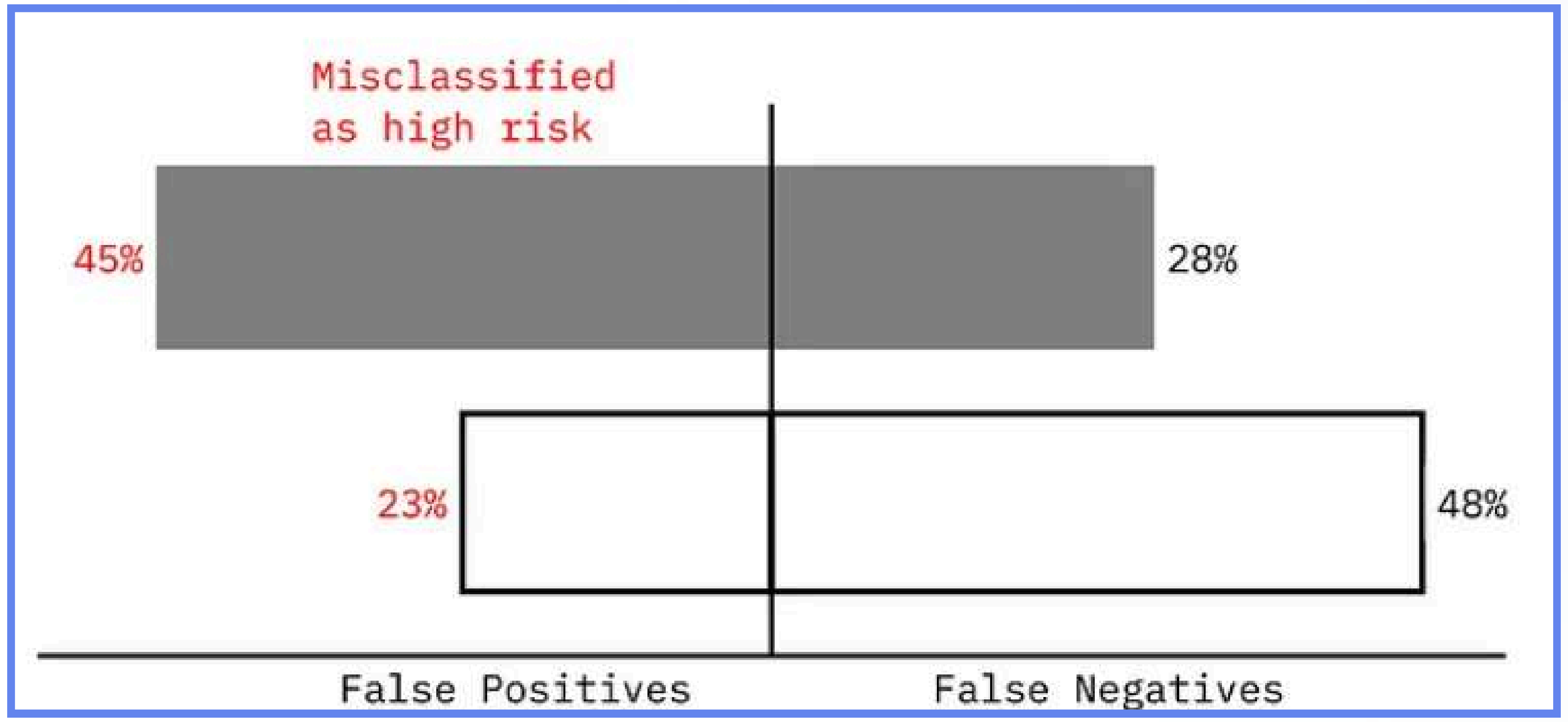
In the case of the Amazon project, there were a few ways this happened. For example, the tool disadvantaged candidates who went to certain women's colleges presumably not attended by many existing Amazon engineers. It similarly downgraded resumes that included the word "women's" — as in "women's rugby team." And it privileged resumes with the kinds of verbs that men tend to use, like "executed" and "captured."

an AI hiring tool developed by Amazon began systematically discriminating against female candidates, downgrading resumes that included the word "women's" – a bias learned from patterns in past hiring data

Real-World Implications

Biased AI Model Case Studies

A criminal risk scoring algorithm (COMPAS) was found to disproportionately predict higher recidivism risk for Black defendants compared to white defendants with similar profiles.



True positives & True negatives are cases of the algorithm's predictions being correct. ProPublica found that COMPAS was correct only ~61% of the time.[1] This accuracy rate was similar for both groups.

False positives predict an incorrectly high probability of recidivism. False negatives predict an incorrectly low probability of recidivism. When COMPAS was wrong, it was wrong in different ways for different groups.

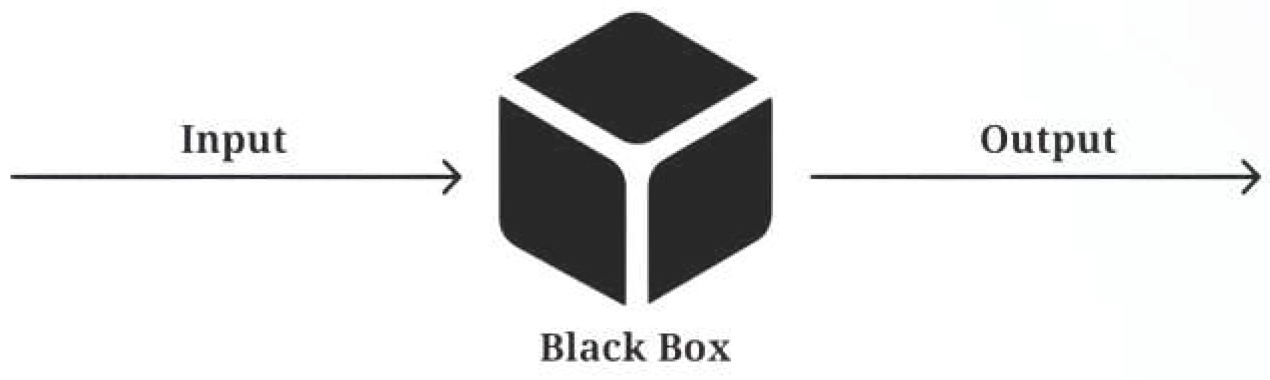
Black defendants were more often predicted to reoffend but they didn't. White defendants were more often predicted to not reoffend but they did.

Such cases illustrate that without explainability, AI can "launder" and perpetuate historical biases, with grave consequences for impacted groups.

Challenges in Black-Box AI Models

Biased AI Model Case Studies

Modern AI models (especially deep learning networks) often operate as black boxes: they have millions of parameters and complex non-linear interactions that are not directly interpretable by humans. This opacity means even the engineers building a model may struggle to understand its internal logic or predict why it makes a given prediction.



The absence of insight limits our ability to debug errors and trust AI models, especially in safety-critical applications. Without explainability, we risk deploying unpredictable models with faulty reasoning.
This drives the need for methods that enhance transparency in AI's decision-making process.

Taxonomy of Explainability Methods

Intrinsic Explainability: Intrinsic (built-in) explainability refers to models that are interpretable by design.

- These models have a transparent structure that humans can directly follow, so no external post-processing is needed to understand their decisions.
- Examples include decision trees (where one can trace the path of decisions), linear/logistic regression (weights indicate feature influence), or rule-based classifiers.

Interpretability Through Model Constraints

- Achieved by limiting complexity or using human-understandable representations
- Trade-off: Less complex models can be easier to interpret but may have lower accuracy

Model-Specific Nature

- Interpretable structure depends on the model type
- Example: A small decision tree is intrinsically interpretable, while a neural network typically is not

Taxonomy of Explainability Methods

Post-hoc Explainability: Post-hoc methods provide explanations after a model has been trained, without altering the model itself.

- Treats the model as a black box and infers explanations by examining inputs and outputs
- The goal is to approximate or interpret what the model is doing in human-understandable terms, even if the model is complex.

Approximate Nature of Explanations

- Does not require the original model to be simple
- Explanations are approximations, requiring careful interpretation

Independant from the explained model

- Applied after model training
- Examples: Surrogate models, Feature importance analysis, Visualization tools (graphs, heatmaps, etc.)

Taxonomy of Explainability Methods

Global vs. Local Explanations: Explainability methods can be characterized by the scope of the explanation.

Approximate Nature of Explanations

A global explanation provides a high-level understanding of what factors generally drive the model's predictions. For example, global methods might yield a set of decision rules for the entire model or a ranking of feature importances for the model as a whole.

Independant from the explained model

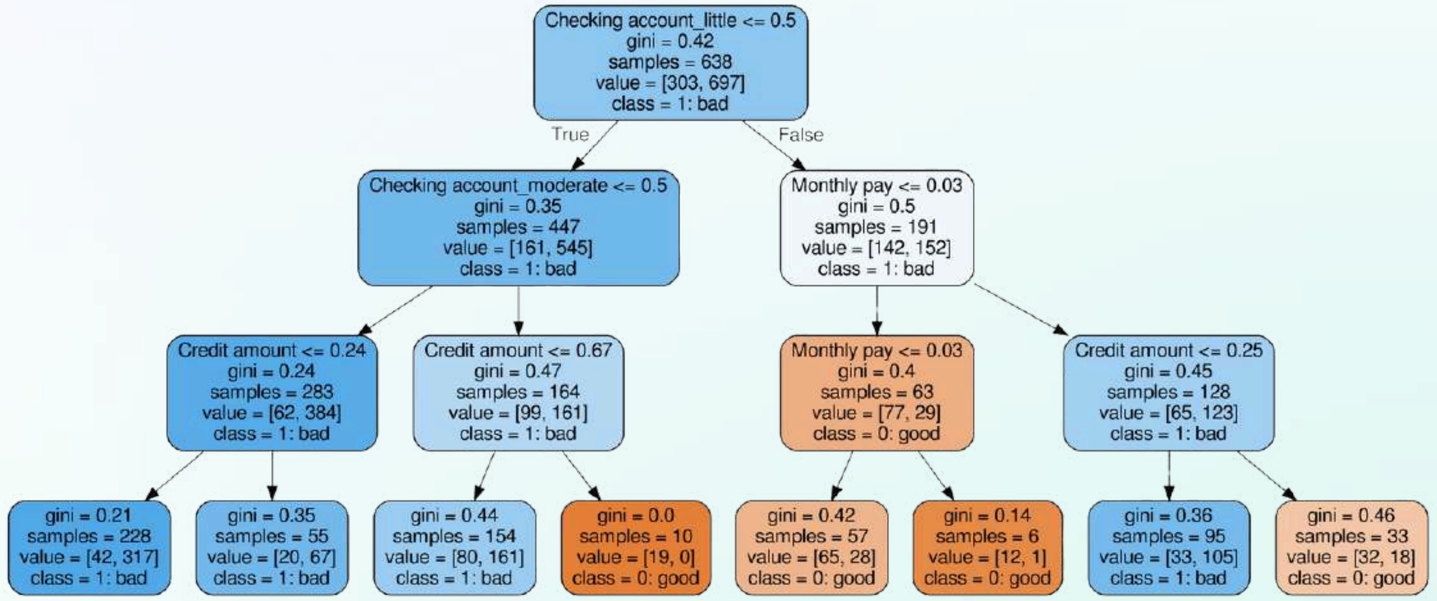
Local predictions explain the model's decision for specific inputs by highlighting influential features. Local methods address "Why did the model do X for this case?" while global methods focus on "How does the model generally make decisions?"

Practitioners often use both to understand general patterns and investigate specific cases, particularly outliers or errors.

Ante Hoc Explainability Methods

Decision Trees

A decision tree is a flowchart-like structure where each internal node represents a test on a feature, each branch represents an outcome of that test, and each leaf node (terminal node) represents a class label (for classification) or a predicted value (for regression).



Interpretability:

- Easy to understand and interpret after a brief explanation
- Can be displayed graphically for non-experts to interpret
- Each path from root to leaf represents a decision rule.

Advantages:

- Handles both numerical and categorical data.
- Requires little data preparation; no need for data normalization or dummy variables.
- Reflects the importance of attributes; features on top are the most informative.

Ante Hoc Explainability Methods

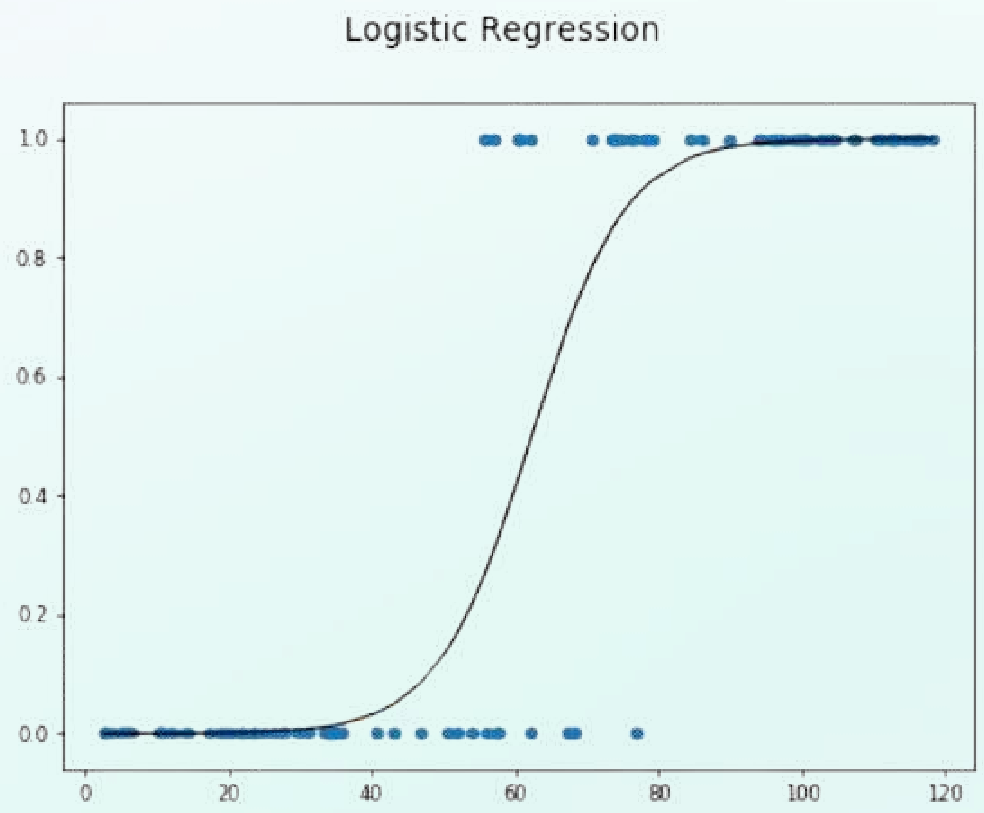
Generalized Additive Models (GLM)

The generalized linear model (GLM) is a generalization of ordinary linear regression, defined by the formula:

$$E[Y] = g^{-1}(Xa)$$

Interpretability:

- Coefficient significance is typically assessed using t-tests or z-tests, where a low p-value indicates a statistically significant predictor.
- Each coefficient represents the change in the transformed mean response per one unit change in the predictor.
- Example: In logistic regression (a GLM with a logit link), a coefficient of 0.5 implies that a one-unit increase in the predictor multiplies the odds by exp(0.5) ≈ 1.65 (i.e., a 65% increase in odds).



Advantages:

- Flexible framework accommodating various types of response variables.
- Maintains interpretability through model coefficients.

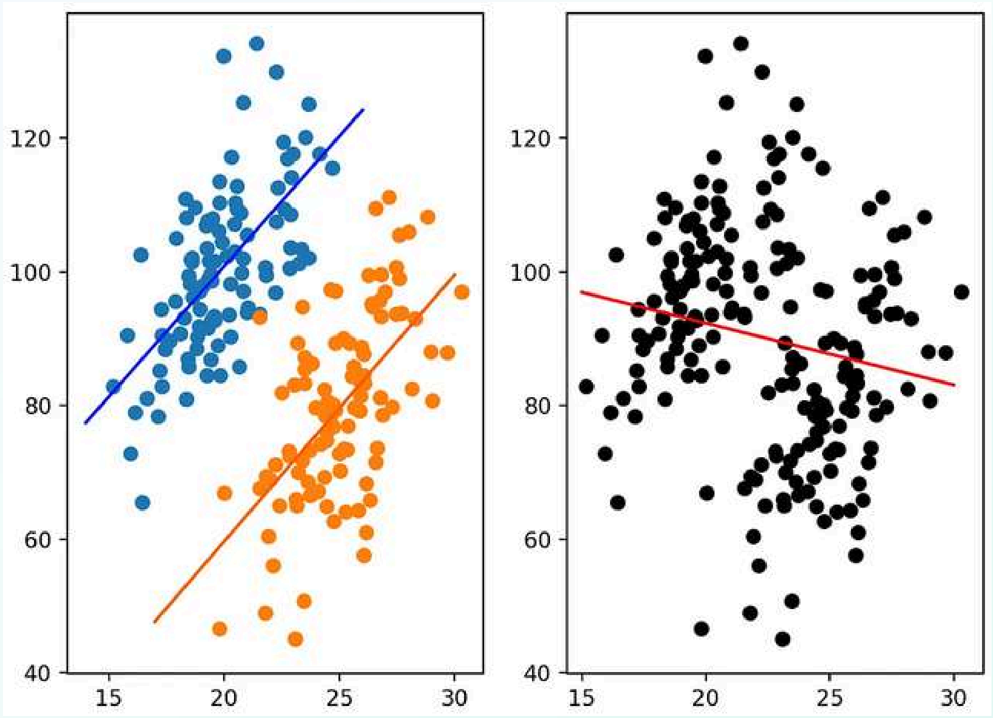
Ante Hoc Explainability Methods

Simpson's paradox & confounding variables

Simpson's Paradox refers to a phenomenon in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

Examples:

	Drug A	Drug B
Effectiveness in male (%)	$\frac{60}{200} \times 100 = 30\%$	$\frac{90}{180} \times 100 = 50\%$
Effectiveness in female (%)	$\frac{240}{300} \times 100 = 80\%$	$\frac{45}{50} \times 100 = 90\%$
Combined (%)	$\frac{300}{500} \times 100 = 60\%$	$\frac{135}{230} \times 100 = 58.69\%$



Confounding variables are factors that, while not the primary focus of a study, can significantly impact how we interpret the relationship between the main variables under investigation.
These variables introduce biases or distortions, making it difficult to attribute any observed effects solely to the studied variables.

Post Hoc Explainability Methods

Feature Importance - Permutation Importance

Permutation importance is a model-agnostic method that assesses the contribution of a feature by measuring the change in model performance when its values are randomly shuffled. This breaks the association between the feature and the target, revealing its importance.

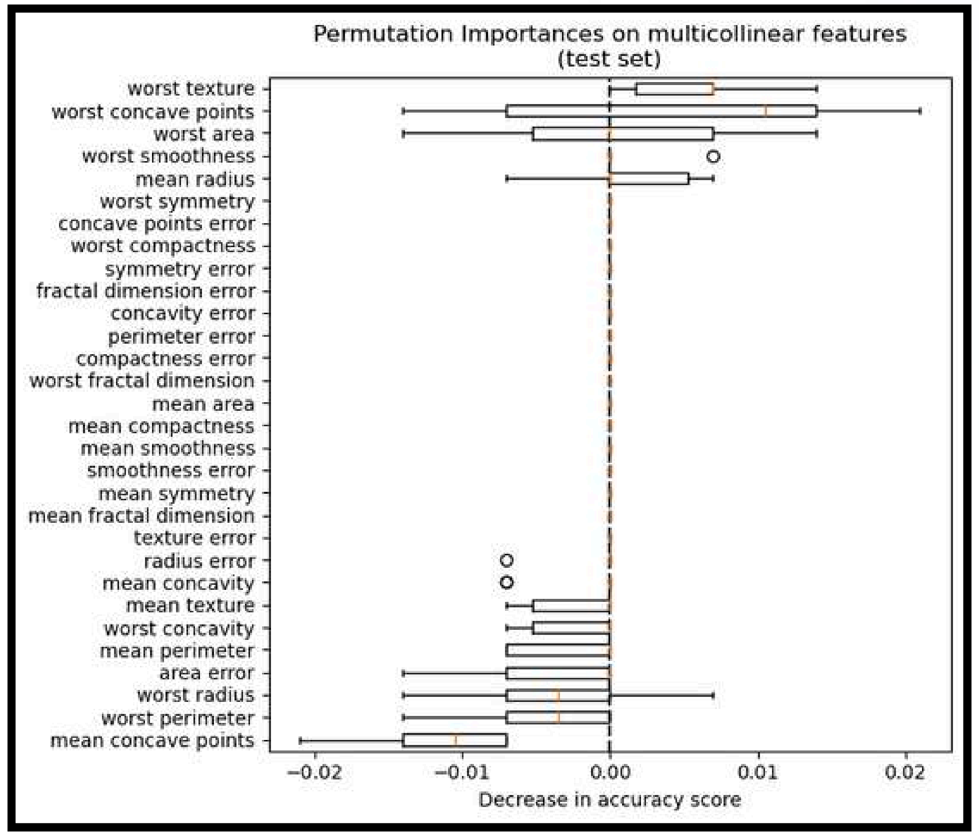
- Inputs: fitted predictive model m, tabular dataset (training or validation) D.
- Compute the reference score s of the model m on data D (for instance the accuracy for a classifier or the R^2 for a regressor).
- For each feature j (column of D):
  - For each repetition k in $1, \ldots, K$:
    - Randomly shuffle column j of dataset D to generate a corrupted version of the data named $\tilde{D}_{k,j}$.
    - Compute the score $s_{k,j}$ of model m on corrupted data $\tilde{D}_{k,j}$.
  - Compute importance i_j for feature f_j defined as:

$$i_j = s - \frac{1}{K} \sum_{k=1}^{K} s_{k,j}$$

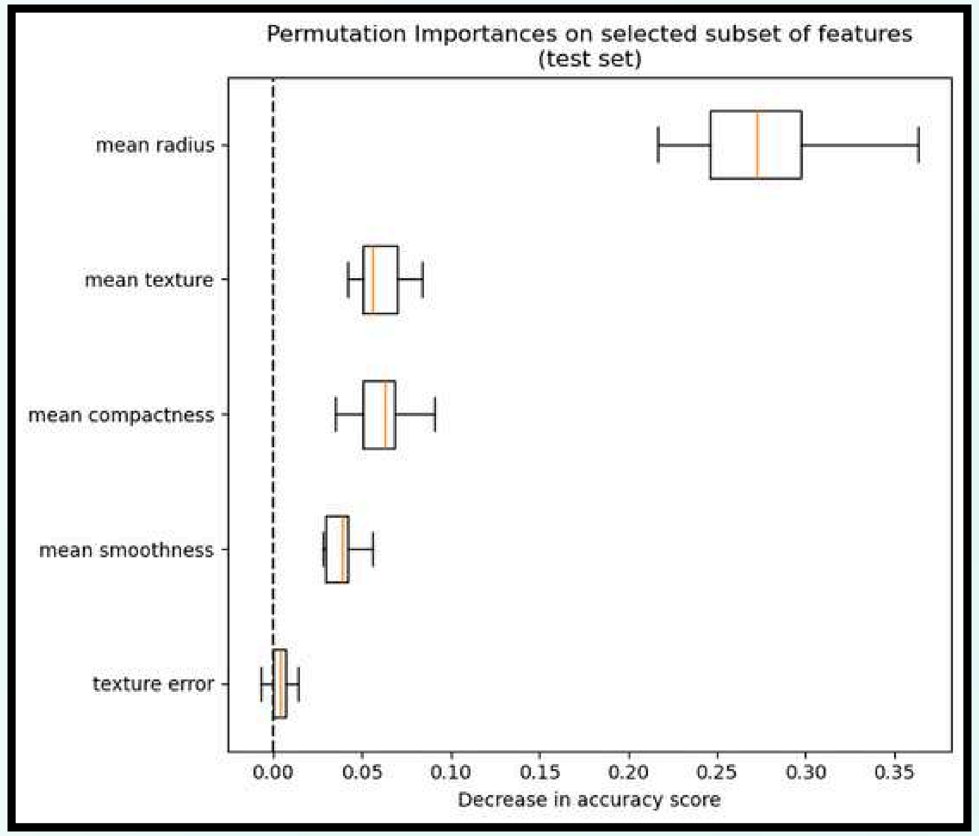
Post Hoc Explainability Methods

Feature Importance - Permutation Importance

- Features deemed unimportant in a poor model may be crucial in a good one. Thus, evaluating a model's predictive power with a held-out set or cross-validation is essential before assessing feature importance. Permutation importance indicates a feature's relevance to a specific model, rather than its intrinsic predictive value.
- When two features are correlated and one of the features is permuted, the model still has access to the latter through its correlated feature. This results in a lower reported importance value for both features, though they might actually be important.



after clustering features
that are correlated



RandomForestClassifier trained using the Breast cancer wisconsin (diagnostic) dataset,

Post Hoc Explainability Methods

Feature Importance - Partial Dependence Plots (PDP)

The partial dependence plot (short PDP or PD plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model. A PDP can show whether the relationship between the target and a feature is linear, monotonic or more complex. For example, when applied to a linear regression model, partial dependence plots always show a linear relationship.

Let XS be the set of input features of interest and let XC be its complement. PDF at point xs:

$$\hat{f}_S(x_S) = \mathbb{E}_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C)\, dP(X_C)$$

$\hat{f}$ the machine learning model

The partial function is estimated by calculating averages in the training data, also known as Monte Carlo method:

$$\hat{f}_S(x_S) \approx \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(x_S, x_C^{(i)}\right).$$

In this formula, $x_C^{(i)}$ are actual feature values from the dataset for the features in which we are not interested, and n is the number of instances in the dataset.

Post Hoc Explainability Methods

Feature Importance - Individual Conditional Expectation (ICE) plots

Similar to a PDP, an individual conditional expectation (ICE) plot shows the dependence between the target function and an input feature of interest. However, unlike a PDP, which shows the average effect of the input feature, an ICE plot visualizes the dependence of the prediction on a feature for each sample separately with one line per sample.

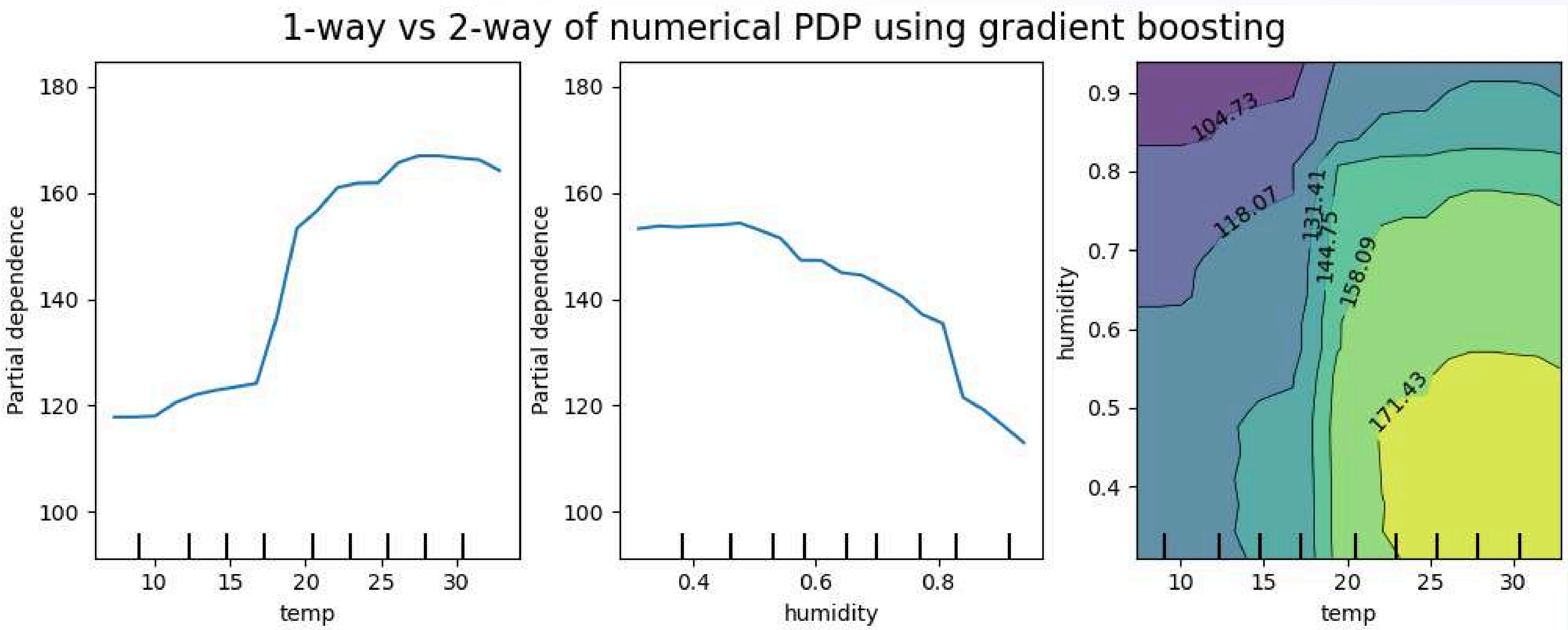
$$\hat{f}_S^{(i)}(x_S) = \hat{f}(x_S, x_C^{(i)})$$

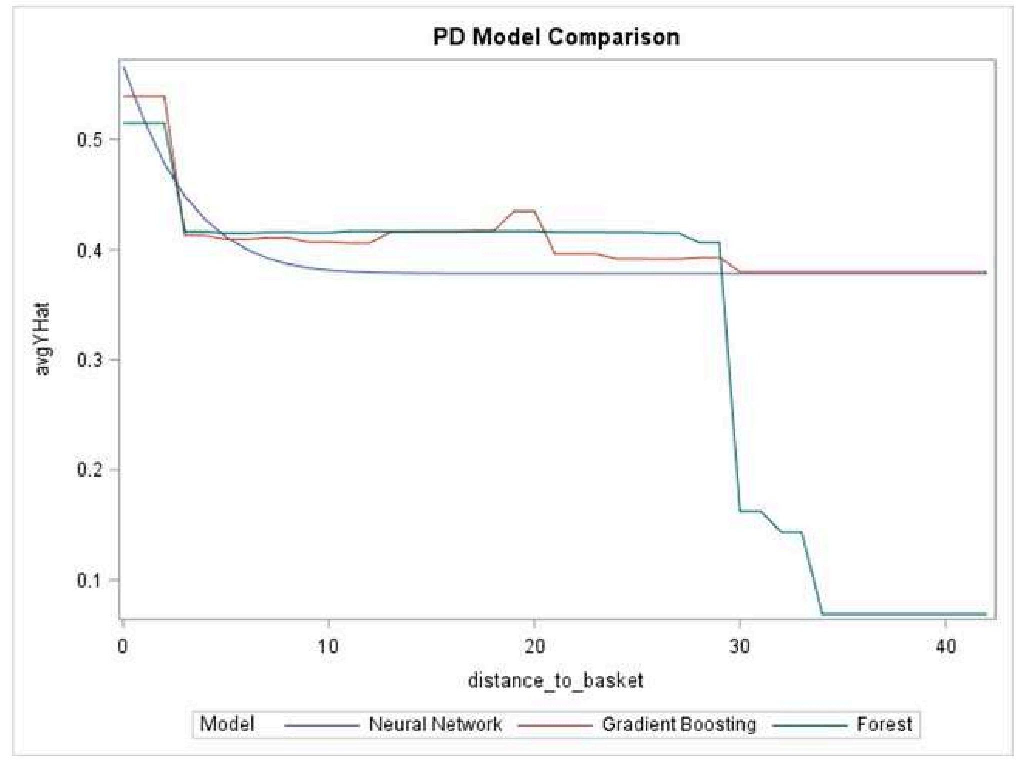
- Each line in an ICE plot represents a single instance, showing how its predicted outcome changes as xs varies.
- Reveals whether the relationship between xS and the prediction is consistent across instances or if subgroups behave differently.
- If there are too many lines in an ICE plot, it can be difficult to see differences between individual samples and interpret the model. Centering the ICE at the first value on the x-axis, produces centered Individual Conditional Expectation (cICE) plots

Independence Assumption: Like PDP, assumes other features remain fixed. this can be misleading if features are correlated.

Post Hoc Explainability Methods

Feature Importance - PDP & ICE





Average prediction by candidate model

Post Hoc Explainability Methods

Feature Importance - PDP & ICE