Explainability

Milestone 5: Trade-offs and Interactions with other verticals in Trustworthy AI

Roseline Polle

Postgraduate, UCL

roseline.polle.19@ucl.ac.uk

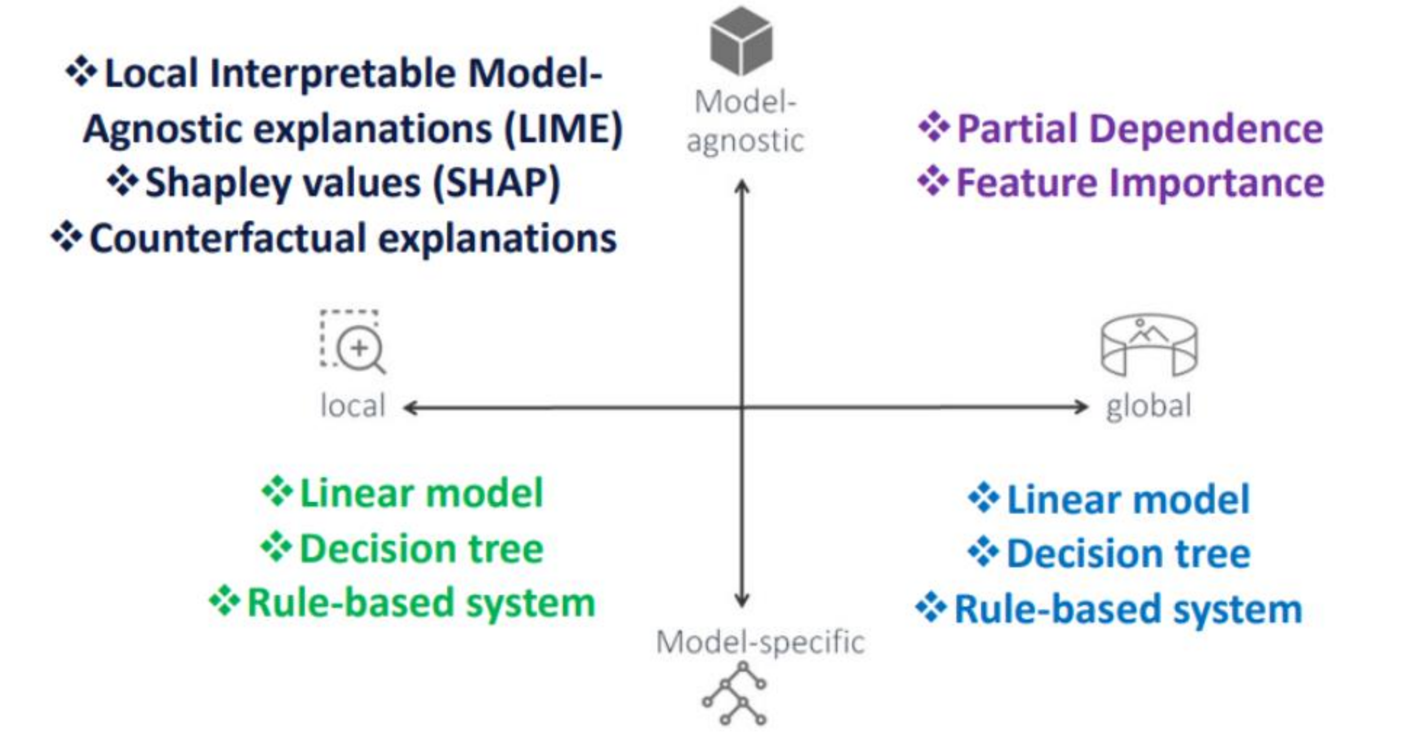
1. Intro on Explainability

Interpretability/Explainability

- **Interpretability**: "the degree to which a human can understand the cause of a decision“ [Millller, 2018]

- **Explainability**: the degree to which the inner mechanics of an algorithm are understood by a human

Types of Explainability



[Koshiyama et al.,2021]

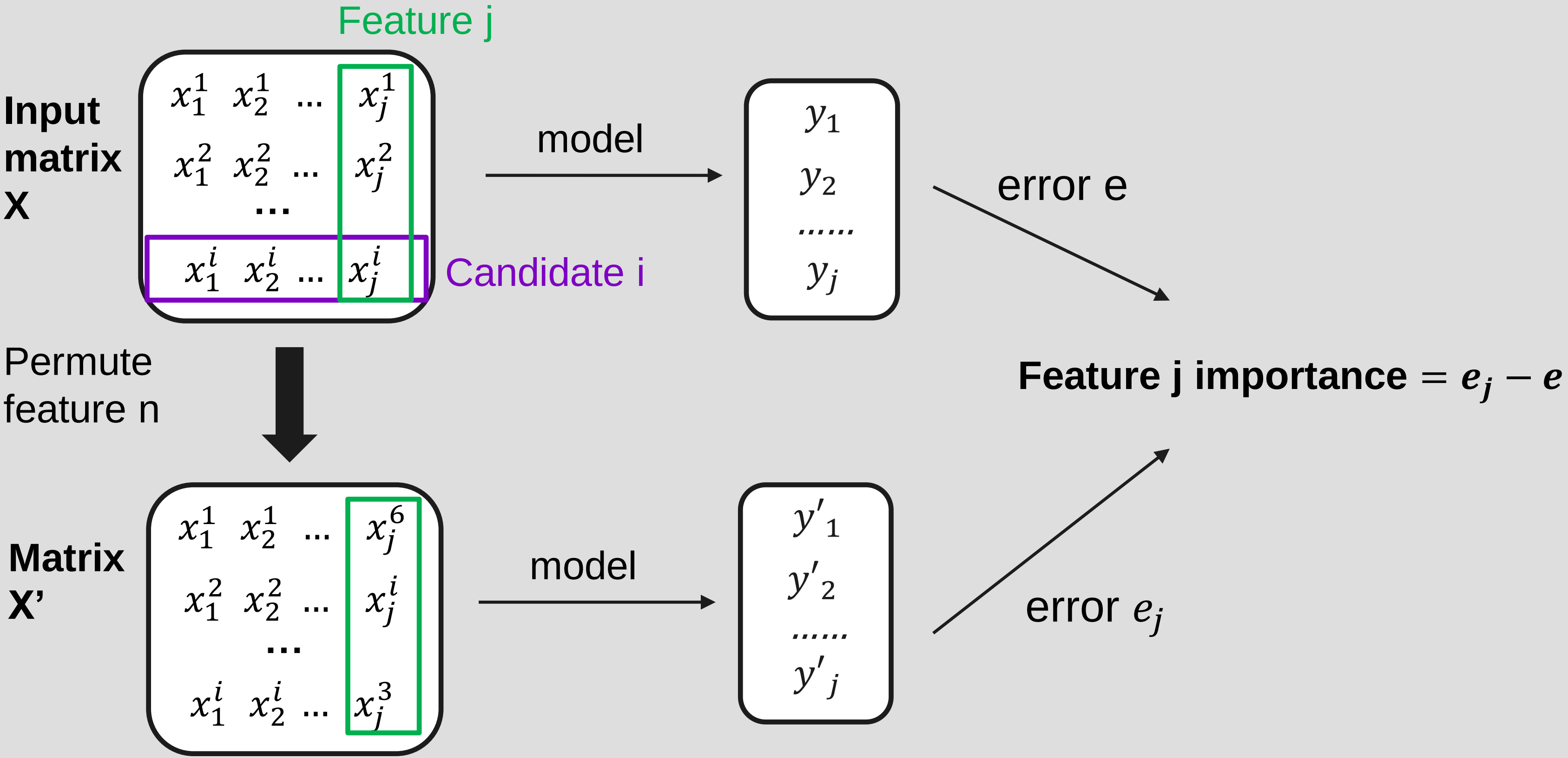
Feature importance

- Looks to assign a score to each feature relative to its importance in the prediction.

- Simple example in linear regression:

$$\boldsymbol{y} = \underbrace{\boldsymbol{w_1} * \boldsymbol{x_1}}_{\text{Importance of feature 1 in outcome y}} + \boldsymbol{w_2} * \boldsymbol{x_2} + \dots + \boldsymbol{w_n} * \boldsymbol{x_n}$$

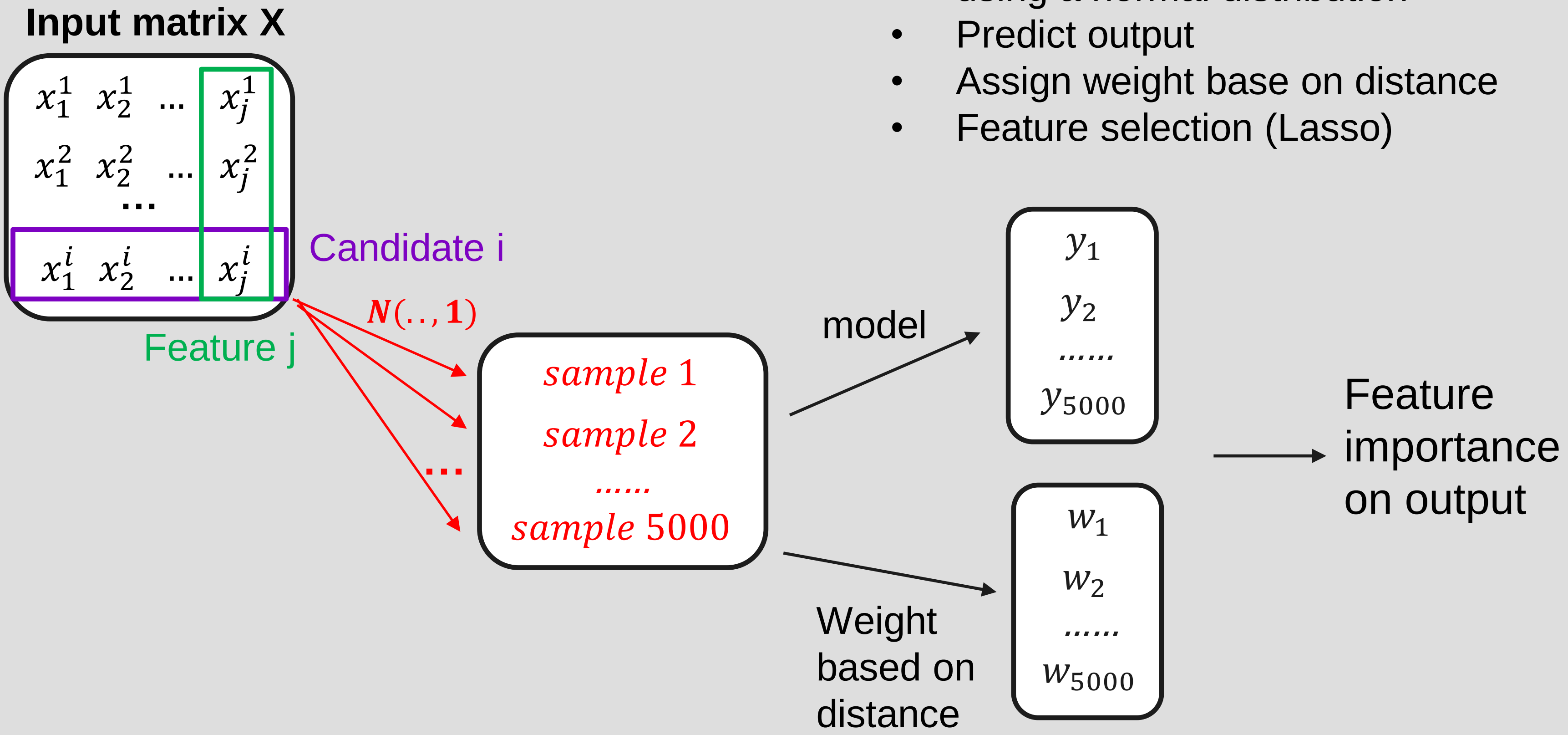
Permutation feature importance



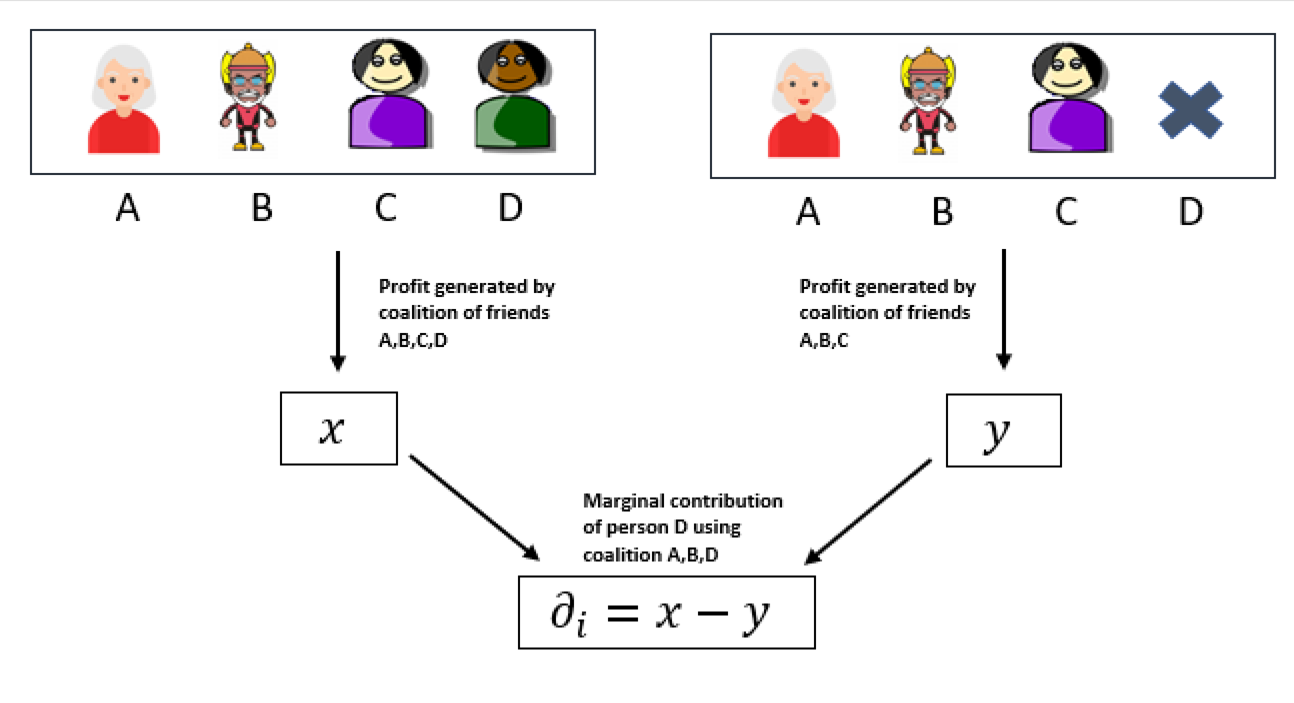
LIME

Explain results for candidate i:

- Sample 5,000 times feature vector i using a normal distribution
- Predict output
- Assign weight base on distance
- Feature selection (Lasso)

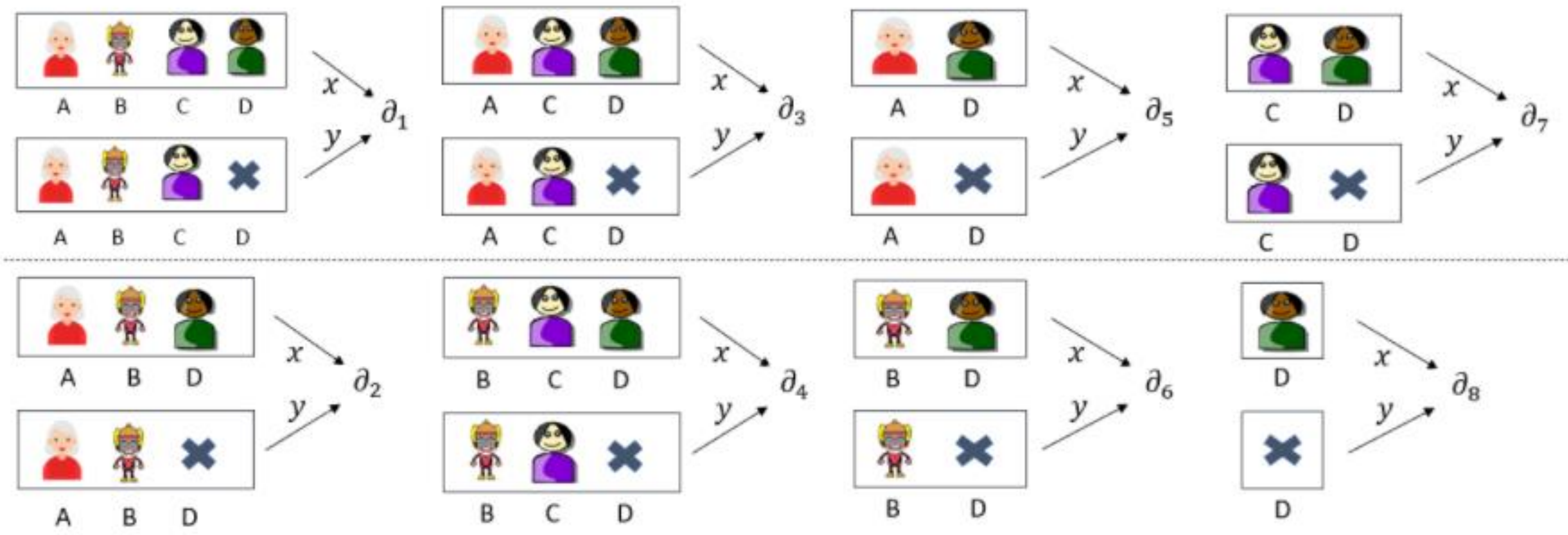


SHAPLEY Values



Adapted from https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3 by Fernando Lopez

SHAPLEY Values



The shapley value for person D is therefore: $\Phi_D = \frac{\delta_1+\delta_2+\delta_3+\delta_4+\delta_5+\delta_6+\delta_7+\delta_8}{8}$

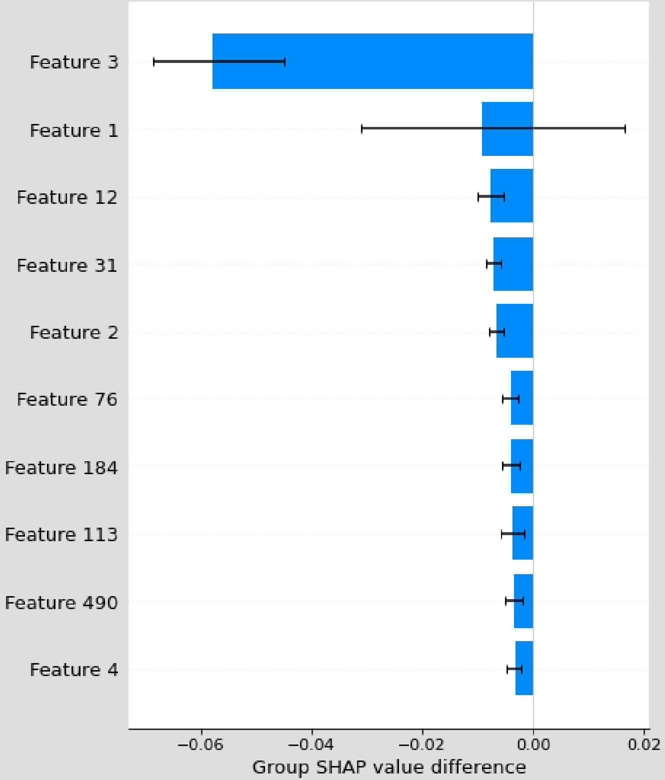
2. Interactions with Fairness

Questions one can answer

- Are the most influential factors reasonable? Are they proxy for a protected characteristics?

- Is the model relying too much on one feature?

- Are they the influential factors the same across different groups ?



Adaptation to fairness

- Instead of explaining output → explain fairness metric
- Example: effect of permutation importance on Disparate Impact metric.
- Answers the question: what are the features most responsible for the observed bias (if any) ?

Further readings

- **Interpretable Machine Learning.** *A Guide for Making Black Box Models Explainable* (https://christophm.github.io/interpretable-ml-book/)