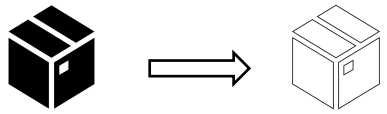
The fidelity of global surrogates in explainable Machine Learning

Carel Schwartzenberg, Tom van Engers and Yuan Li

Part 1: Fidelity metrics

- What metric should be used to measure the fidelity of a global surrogate to the original model?

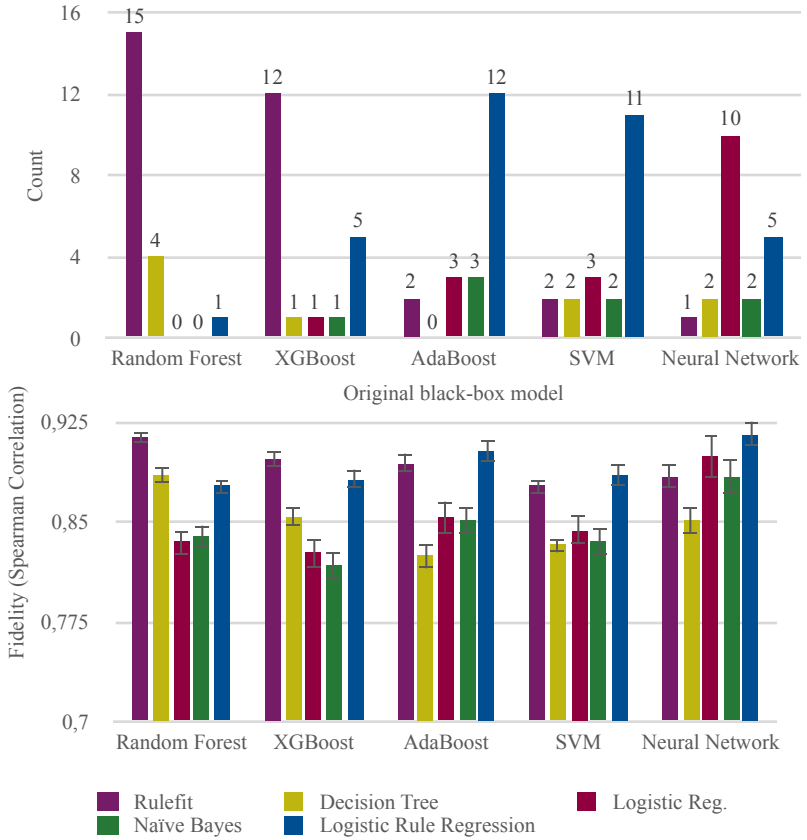


Black-box model **Surrogate white-box model**

Metric	MSE	MSE	MAE	R^2	AUC	SpCorr
Accuracy (Acc.)	1.00	0.677	0.664	0.677	0.761	0.654
Mean Squared Error (MSE)	0.677	1.00	0.941	1.00	0.759	0.827
Mean Absolute Error (MAE)	0.664	0.941	1.00	0.941	0.743	0.796
Coefficient of Determination (R^2)	0.677	1.00	0.941	1.00	0.759	0.827
Area Under the ROC Curve (AUC)	0.761	0.759	0.743	0.759	1.99	0.764
Spearman Correlation (SpCor)	0.654	0.827	0.796	0.827	0.764	1.00
Permutation Importance (PMI)	0.648	0.675	0.666	0.675	0.713	**0.720**
Accumulated Local Effects (ALE)	0.659	0.711	0.723	0.711	0.738	**0.788**

Part 2: Fidelity of global surrogates

- Which global surrogate models have the highest fidelity?
- Do certain classes of surrogate models have higher fidelity to certain classes of black-boxes?



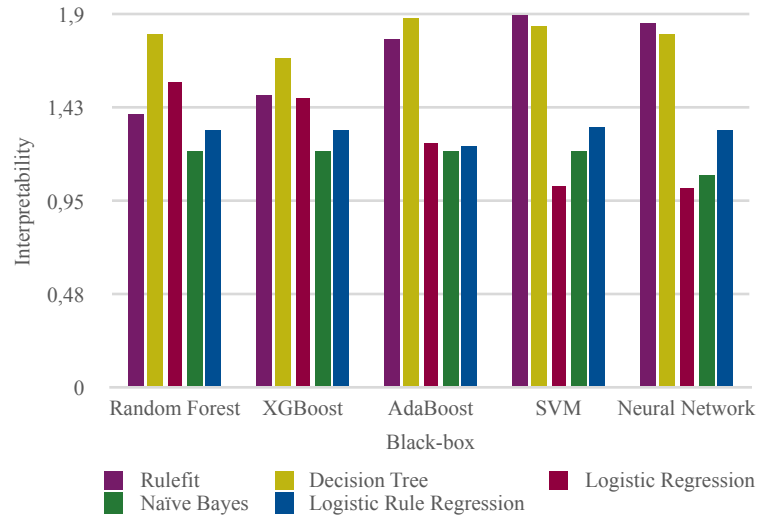
Part 3: Interpretability of global surrogates

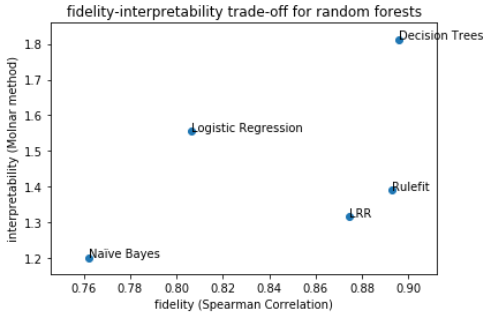
- What does the fidelity-interpretability curve look like for global surrogates?
- Which global surrogate models have the highest fidelity relative to their interpretability?

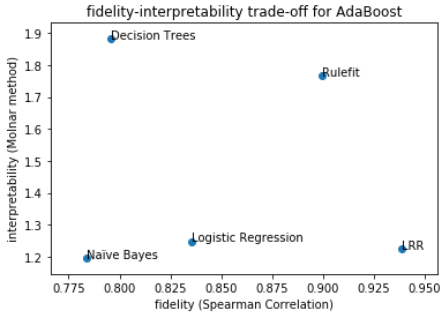
Molnar's interpretability measure

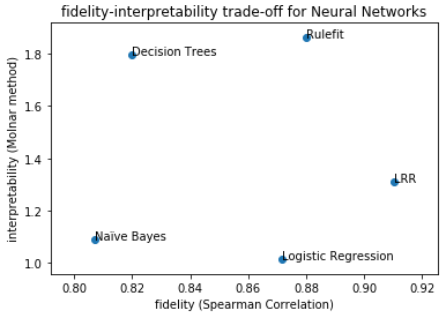
0: low relative interpretability.
3: high relative interpretability

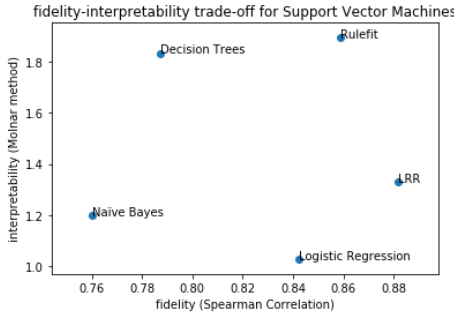
$$Interpretability = 3 - (NF_{scaled} + IAS_{scaled} + MEC_{scaled})$$

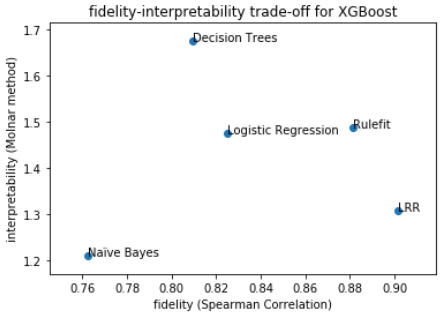












University of Amsterdam
Faculty of Science