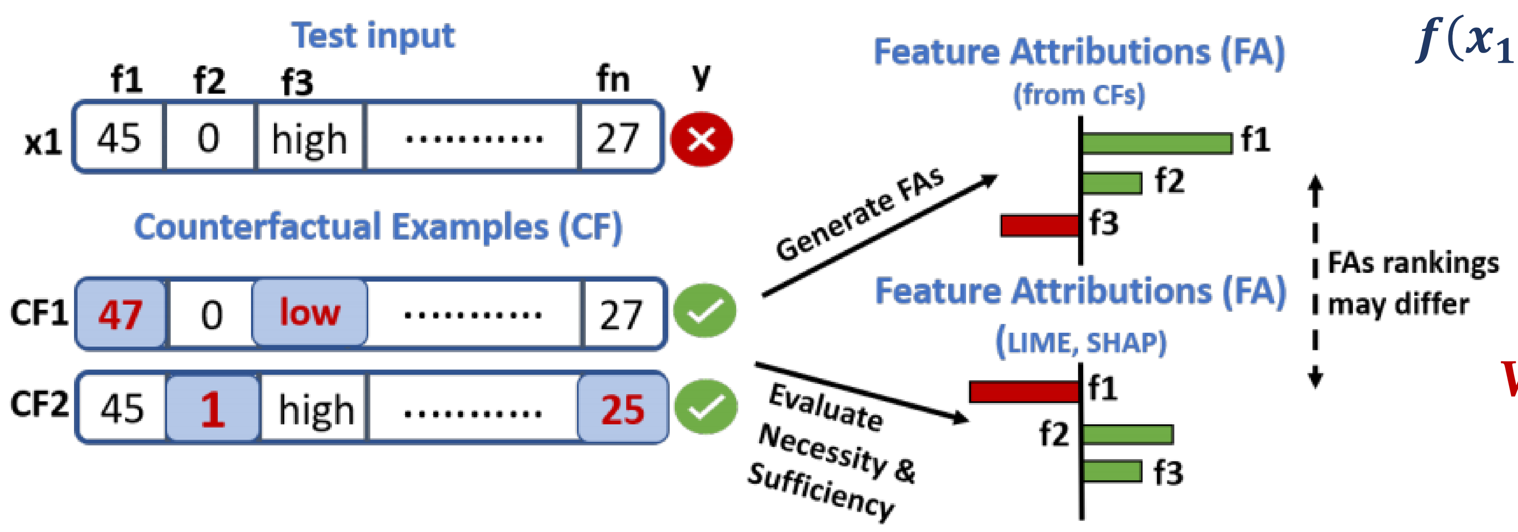
Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End

Local Explanation Methods DISAGREE with Each Other

Feature Attributions and Counterfactuals often disagree even for simple linear models



$f(x_1, x_2) = I(0.45x_1 + 0.1x_2 \geq 0.5),\ x_1, x_2 \in [0, 1]$

	x_1	x_2
LIME	0.34	0.07
SHAP	0.69	0.28
WachterCF	0.98	0.97
DiCE	1.00	0.98

Importance Scores

- Propose an unifying framework based on Actual Causality to interpret these two approaches
- Evaluate attribution-based methods on the necessity and sufficiency of their top-ranked features

Actual Causality and Sufficiency → Ideal Model Explanations

(1) **Existence:** There exists a context $\boldsymbol{u} \in U$ such that $\boldsymbol{x}_j = a$ and $f(\boldsymbol{x}_{-j} = b, \boldsymbol{x}_j = a) = y^*$.

(2) **Necessity:** For each context $\boldsymbol{u} \in U$ where $\boldsymbol{x}_j = a$ and $f(\boldsymbol{x}_{-j} = b, \boldsymbol{x}_j = a) = y^*$, some feature subset $\boldsymbol{x}_{sub} \subseteq \boldsymbol{x}_j$ is an actual cause under $(M, \boldsymbol{u})$

(3) **Minimality:** $\boldsymbol{x}_j$ is minimal, namely, there is no strict subset $\boldsymbol{x}_s \subset \boldsymbol{x}_j$ such that $\boldsymbol{x}_s = a_s$ satisfies conditions 1-2 above, where $a_s \subset a$.

Stronger Necessity condition (But-for):

Changing the value of $\boldsymbol{x}_j$ alone changes the prediction of the model *(that is when all other features are kept the same)*

(4) **Sufficiency:** For all contexts $\boldsymbol{u}' \in U$, $\boldsymbol{x}_j \leftarrow a \Rightarrow y = y^*$.

Ideal Model Explanations → Partial Model Explanations

- However, for most realistic ML models, an ideal explanation is impractical.
  - It is rare to find such clean explanations of a ML model's output
  - Example: there is no sufficient feature for $f(x_1, x_2, x_3) = I(0.4x_1 + 0.1x_2 + 0.1x_3 \geq 0.5)$

- (α, β) goodness of an explanation to capture the *extend* to which a feature is necessary or sufficient to "cause" the model's original output

$$\alpha = \Pr(\boldsymbol{x}_j \text{ is a cause of } y^* | \boldsymbol{x}_j = a, y = y^*) \qquad \beta = \Pr(y = y^* | \boldsymbol{x}_j \leftarrow a)$$

Interpretation Using A Unifying Framework

Counterfactual explanation (α_{CF})

- Optimizes Necessity
- Perturbed feature subset $\boldsymbol{x}_j$ is a but-for cause of the original output
- α_{CF} summarizes the outcomes of all such perturbations and ranks any feature subset for their necessity

$$\alpha_{CF} = \Pr((\boldsymbol{x}_j \leftarrow a' \Rightarrow y \neq y^*) | \boldsymbol{x}_j = a, y = y^*)$$

Attribution-based explanations (β)

- Optimizes Sufficiency
- Importance of $\boldsymbol{x}_j$ can be interpreted as its sufficiency
- The fraction of all contexts where $\boldsymbol{x}_j \leftarrow a$ leads to $y = y*$ is given by

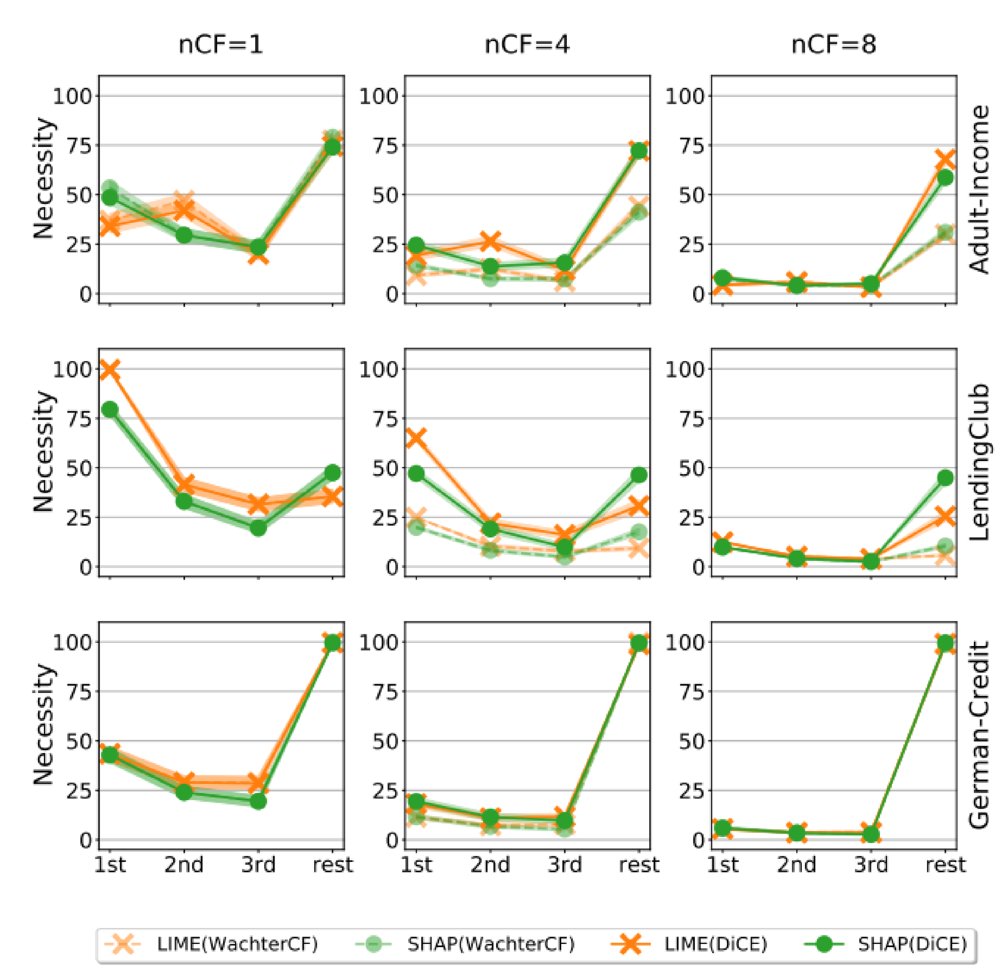
$$\beta = \Pr(y = y^* | \boldsymbol{x}_j \leftarrow a)$$

Top Features of LIME/SHAP are Neither Necessary Nor Sufficient

We use counterfactual explanations to evaluate feature attribution methods based on Necessity and Sufficiency

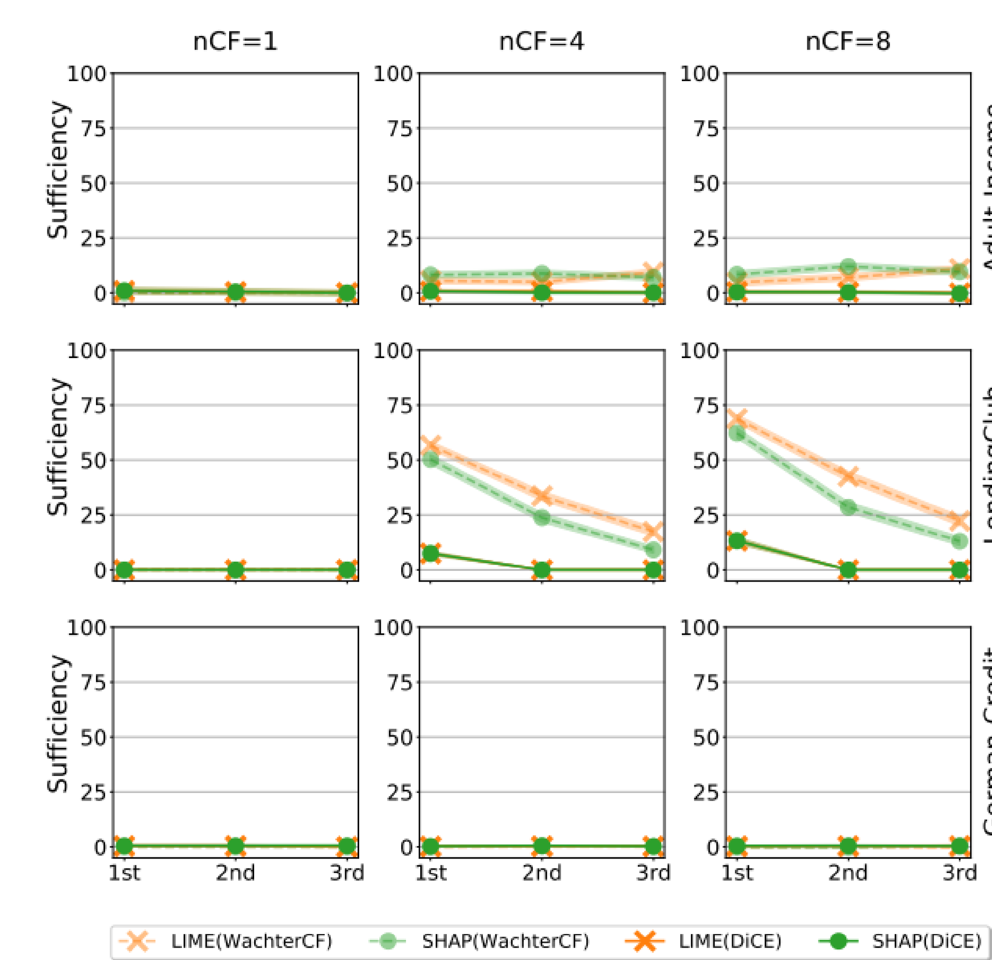
Generate CFs by changing only $\boldsymbol{x}_j$

$$\text{Necessity} = \frac{\sum_{i, \boldsymbol{x}_j \neq a} \mathbb{1}(CF_i)}{\text{nCF} * N}$$



Generate CFs by fixing only $\boldsymbol{x}_j$

$$\text{Sufficiency} = \frac{\sum_i \mathbb{1}(CF_i)}{\text{nCF} * N} - \frac{\sum_{i, \boldsymbol{x}_j \leftarrow a} \mathbb{1}(CF_i)}{\text{nCF} * N}$$



- Highly ranked features may often neither be necessary nor sufficient explanations of a model's predictions – Other features are (sometimes more) meaningful and can potentially provide actionable changes
- Necessity and Sufficiency become weaker for top-ranked features as the number of features in a dataset increases
- Important to consider multiple explanation methods to understand the predictions of a ML model

Ramaravind K. Mothilal (raam.arvind93@gmail.com)
Divyat Mahajan (t-dimaha@microsoft.com)
Chenhao Tan (chenhao@uchicago.edu)
Amit Sharma (amshar@microsoft.com)

Paper: https://arxiv.org/abs/2011.04917