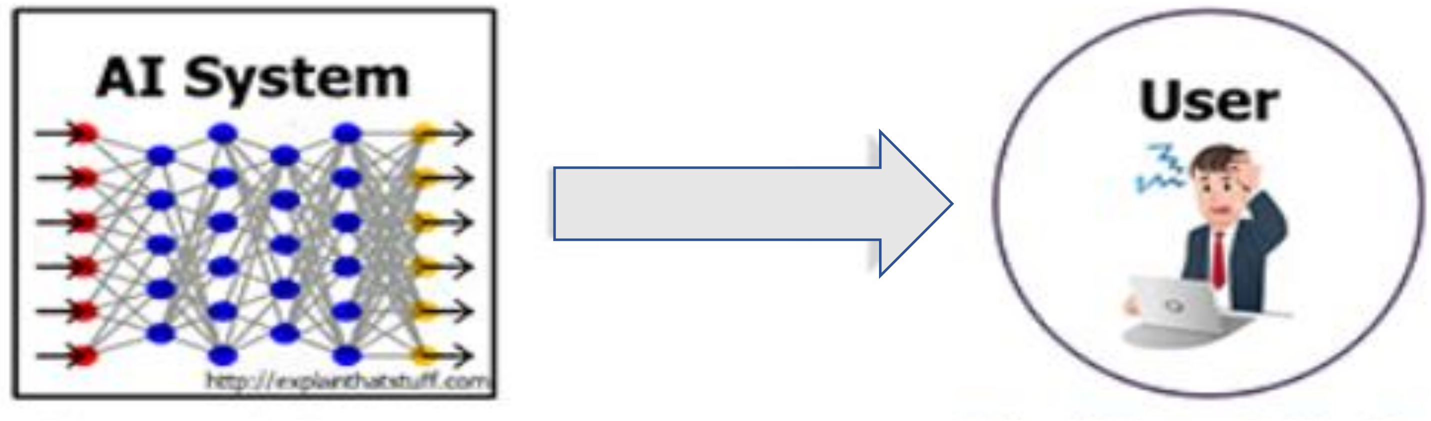
Counterfactual
Local
Explanations
viA
Regression

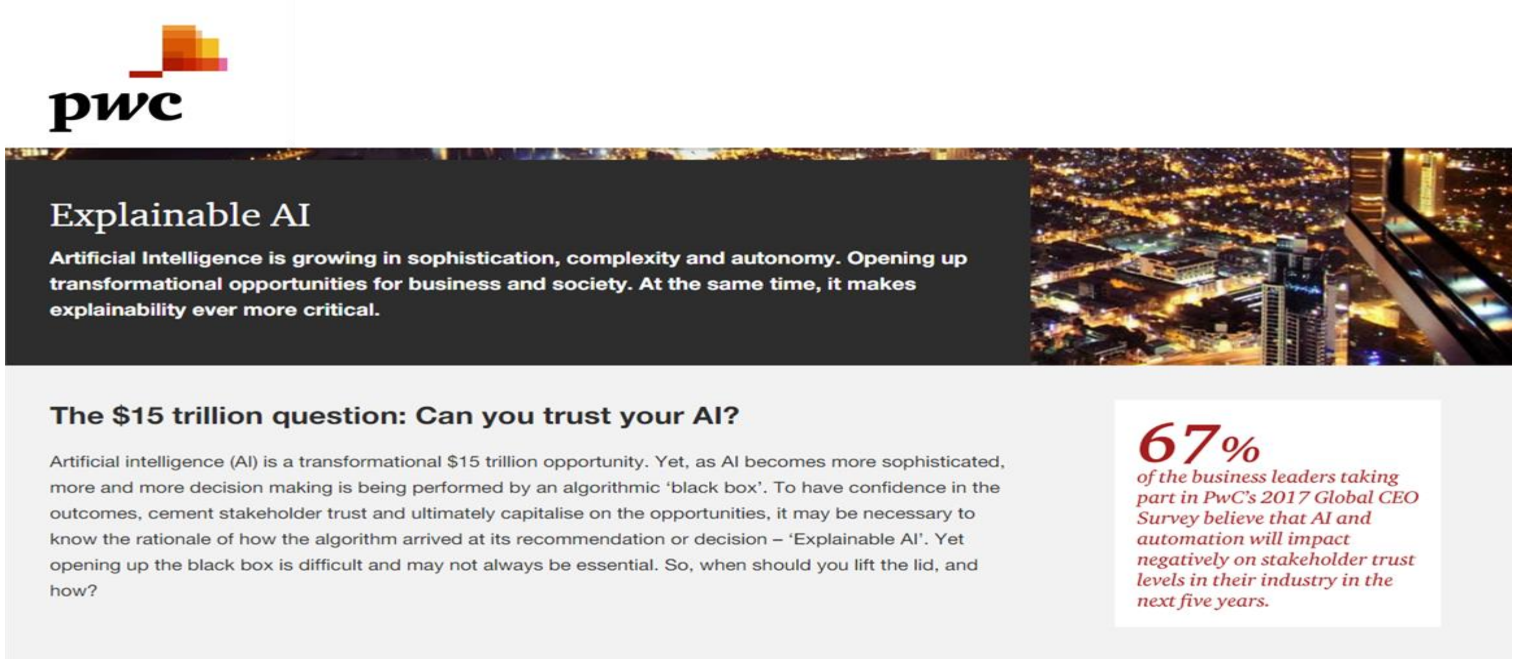
Adam White
Artur d’Avila Garcez

3rd April 2020

CITY
UNIVERSITY OF LONDON
EST 1894
125 YEARS

The Need For Explainable AI




pwc

Explainable AI

Artificial Intelligence is growing in sophistication, complexity and autonomy. Opening up transformational opportunities for business and society. At the same time, it makes explainability ever more critical.

The $15 trillion question: Can you trust your AI?

Artificial intelligence (AI) is a transformational $15 trillion opportunity. Yet, as AI becomes more sophisticated, more and more decision making is being performed by an algorithmic 'black box'. To have confidence in the outcomes, cement stakeholder trust and ultimately capitalise on the opportunities, it may be necessary to know the rationale of how the algorithm arrived at its recommendation or decision – 'Explainable AI'. Yet opening up the black box is difficult and may not always be essential. So, when should you lift the lid, and how?

67%
of the business leaders taking part in PwC's 2017 Global CEO Survey believe that AI and automation will impact negatively on stakeholder trust levels in their industry in the next five years.

State of the Art Method 1: *b*-counterfactuals

b-counterfactuals explain a single prediction by identifying ‘close possible worlds’ in which an individual receives the prediction they desired.

‘Mr Jones would have received his loan, if his annual salary had been $35,000 instead of the $32,000 he currently earns.’

(Wachter et al. 2017)

State of the Art Method 1: *b*-counterfactuals

b-counterfactuals explain a single prediction by identifying ‘close possible worlds’ in which an individual receives the prediction they desired.

‘Mr Jones would have received his loan, if his annual salary had been $35,000 instead of the $32,000 he currently earns.’

(Wachter et al. 2017)

Suppose Mr Jones was assigned a probability of 0.75 for defaulting on a loan.

A satisfactory explanation also needs to explain:

I. why Mr Jones was assigned a score of 0.75. This would include identifying the contribution that each feature made to the score.

II. how the features interact with each other.

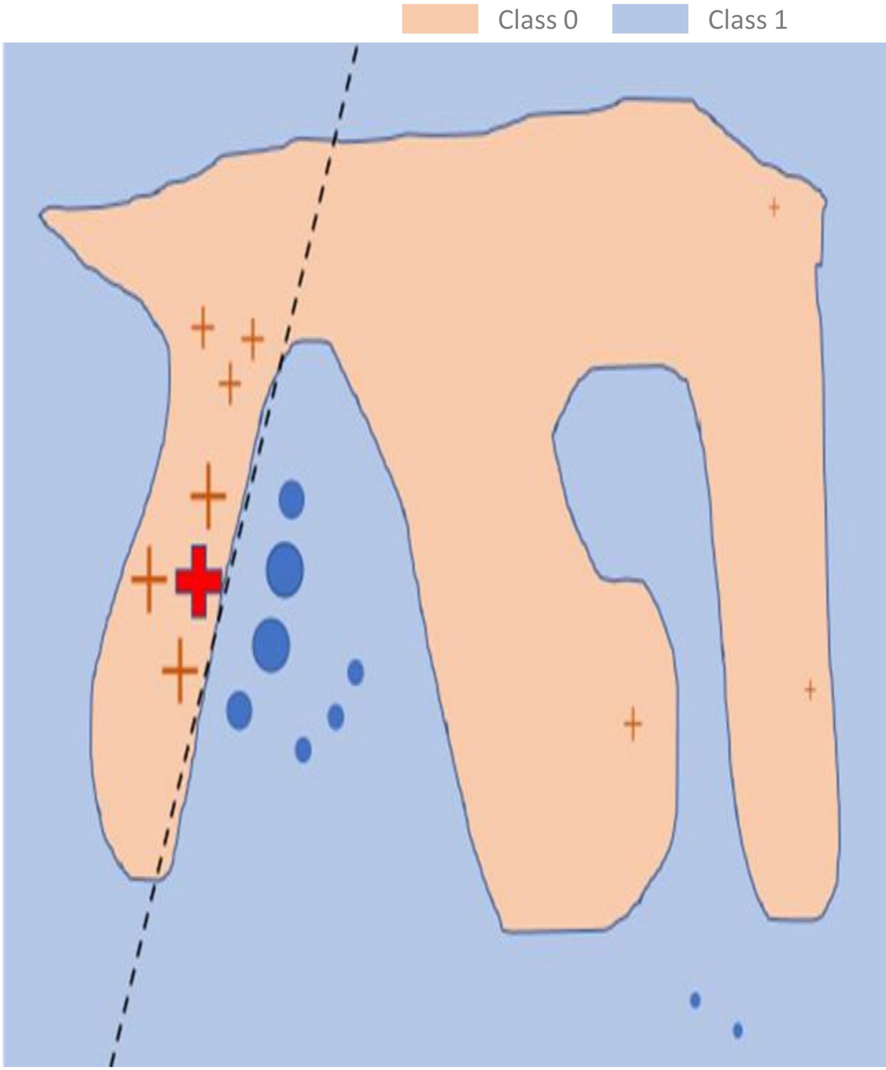
State of the Art Method 2: LIME

Local **I**nterpretable **M**odel-Agnostic **E**xplanations

The LIME algorithm:
Consider a model *m* whose prediction *y* for observation *x* is to be explained.

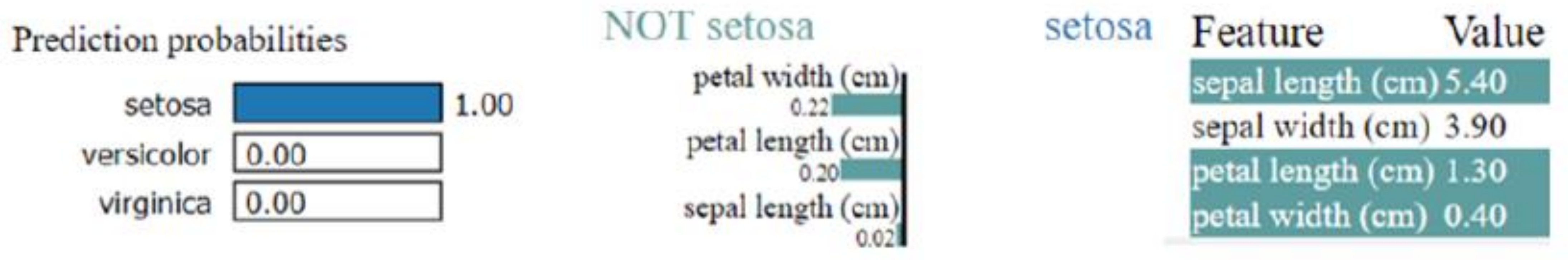
(1) generates a dataset of synthetic observations;
(2) labels the synthetic data by passing it to the model m;
(3) weights the synthetic observations based on Euclidean distance from x
(4) produces a locally weighted linear regression.

The regression coefficients are meant to explain prediction *y*.



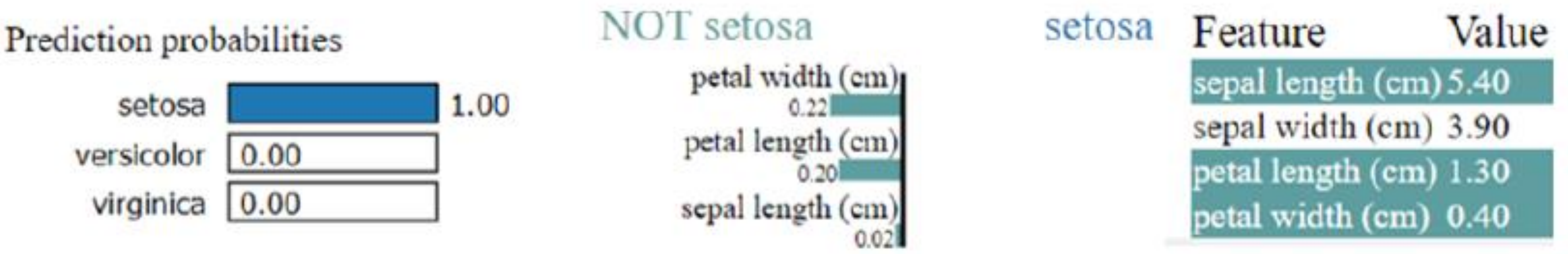
Riberio et al. (2016)

LIME does not measure its fidelity



Feature	Value
sepal length (cm)	5.40
sepal width (cm)	3.90
petal length (cm)	1.30
petal width (cm)	0.40

LIME does not measure its fidelity

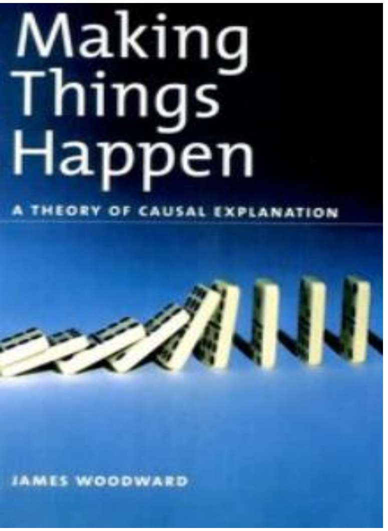


Feature	Value
sepal length (cm)	5.40
sepal width (cm)	3.90
petal length (cm)	1.30
petal width (cm)	0.40

Regression scores:

setosa	0.54
versicolor	0.47
virginica	-0.02

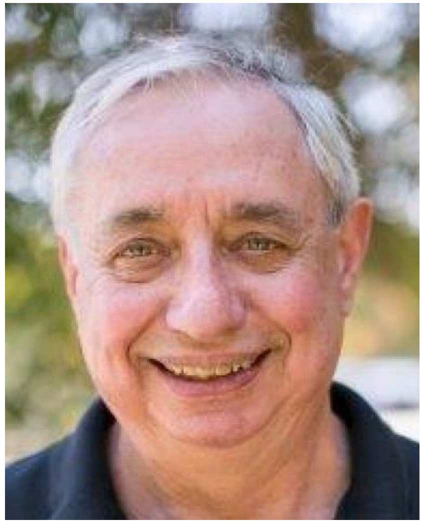
Key Requirements for Satisfactory Explanation



A satisfactory explanation consists in showing patterns of counterfactual dependence.

It needs to include an equation relating a prediction *Y* to its features *X*

(Woodward 2003)



In addition:

Measure its fidelity. It must *know when it does not know*

Counterfactual Local Explanations viA Regression

Provides local explanations:

- with *b*-counterfactuals and the corresponding regression equation;
- using the values of *b*-counterfactuals to significantly improve the fidelity of its regressions;
- that report their fidelity

The *CLEAR* Method:

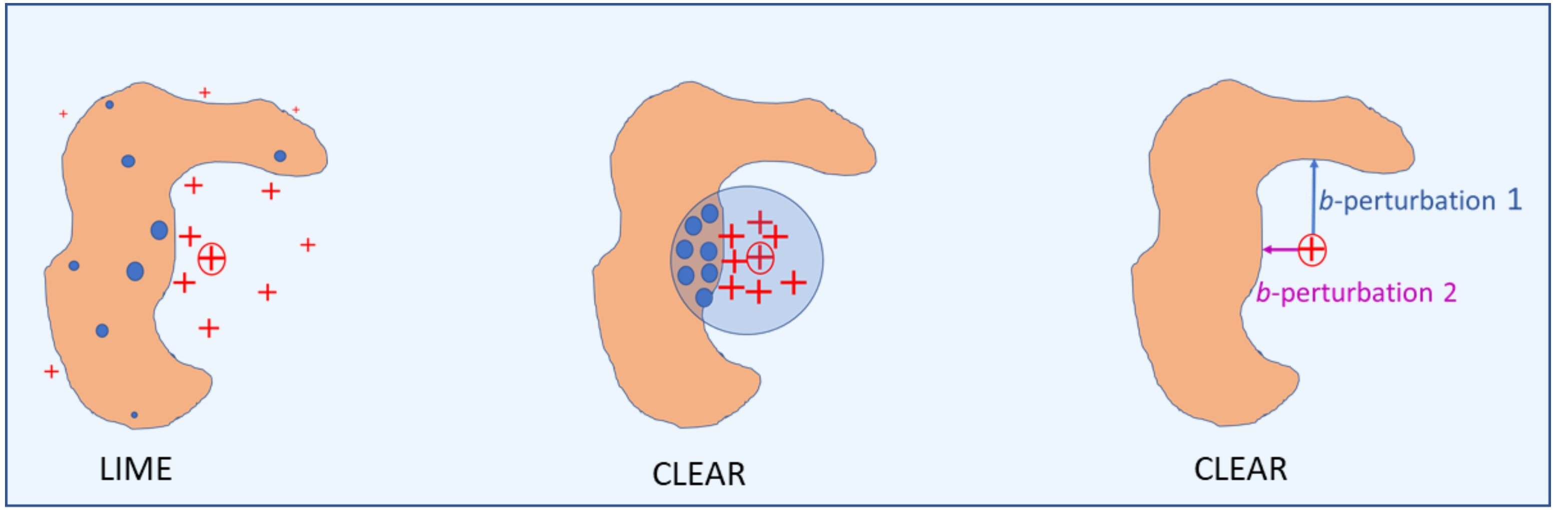
***b*-perturbation**: the minimum change to a feature *f* needed to change the AI system's prediction to a desired class (all other features being kept constant)

Mr Jones would have received his loan if his salary had been $35,000 instead of $32,000

b-counterfactual = $35,000   *b-perturbation = $3000.*

The *CLEAR* Method:

1. Determine $\boldsymbol{x}$'s actual *b*-perturbations.
2. Generate synthetic observations that are then labelled by the AI system.
3. Create a balanced neighbourhood data set (including b-counterfactuals).
4. Perform a step-wise regression – including 2nd degree terms and interaction terms.



The *CLEAR* Method:

1. Determine $\boldsymbol{x}$ ‘s actual *b*-perturbations.
2. Generate synthetic observations that are then labelled by the AI system.
3. Create a balanced neighbourhood data set (including b-counterfactuals).
4. Perform a step-wise regression – including 2nd degree terms and interaction terms.
5. Estimate the *b*-perturbations

MLP on toy diabetes dataset $\boldsymbol{x}$ = {Glucose: 0.537, BloodPressure: 3.1} $P_{\mathrm{MLP}}(\boldsymbol{x} \epsilon \; class1)$= 0.69

CLEAR generates a logistic regression equation (step 4):

$(1 + e^{-\boldsymbol{w}^T\boldsymbol{x}})^{-1}$ = 0.69

& $\boldsymbol{w}^T\boldsymbol{x}$ = - 0.8 + 1.73 Glucose + 0.25 BloodPressure + 0.31 Glucose^2

For $\boldsymbol{x}$ to be on the decision boundary $\boldsymbol{w}^T\boldsymbol{x}$ = 0.

The estimated *b*-perturbation is obtained by substituting:

$\boldsymbol{w}^T\boldsymbol{x}$ = 0 and the value of BloodPressure in $\boldsymbol{x}$

0= 0.31 Glucose^2 + 1.73 Glucose -04

Glucose = 0.025

b-perturbation = 0.025 – 0.537 = -0.512

The *CLEAR* Method:

1. Determine $\boldsymbol{x}$ ‘s actual *b*-perturbations.
2. Generate synthetic observations that are then labelled by the AI system.
3. Create a balanced neighbourhood data set (including b-counterfactuals).
4. Perform a step-wise regression – including 2nd degree terms and interaction terms.
5. Estimate the *b*-perturbations
6. Measure the fidelity of the regression coefficients.

fidelity error = | estimated *b*-perturbation - *b*-perturbation | (step 5 - step 1)

estimated *b*-perturbation (step 5)	= -0.512
actual *b*-perturbation (step 1)	= -0.557
fidelity error	= 0.045

CITY UNIVERSITY OF LONDON EST 1894 125 YEARS

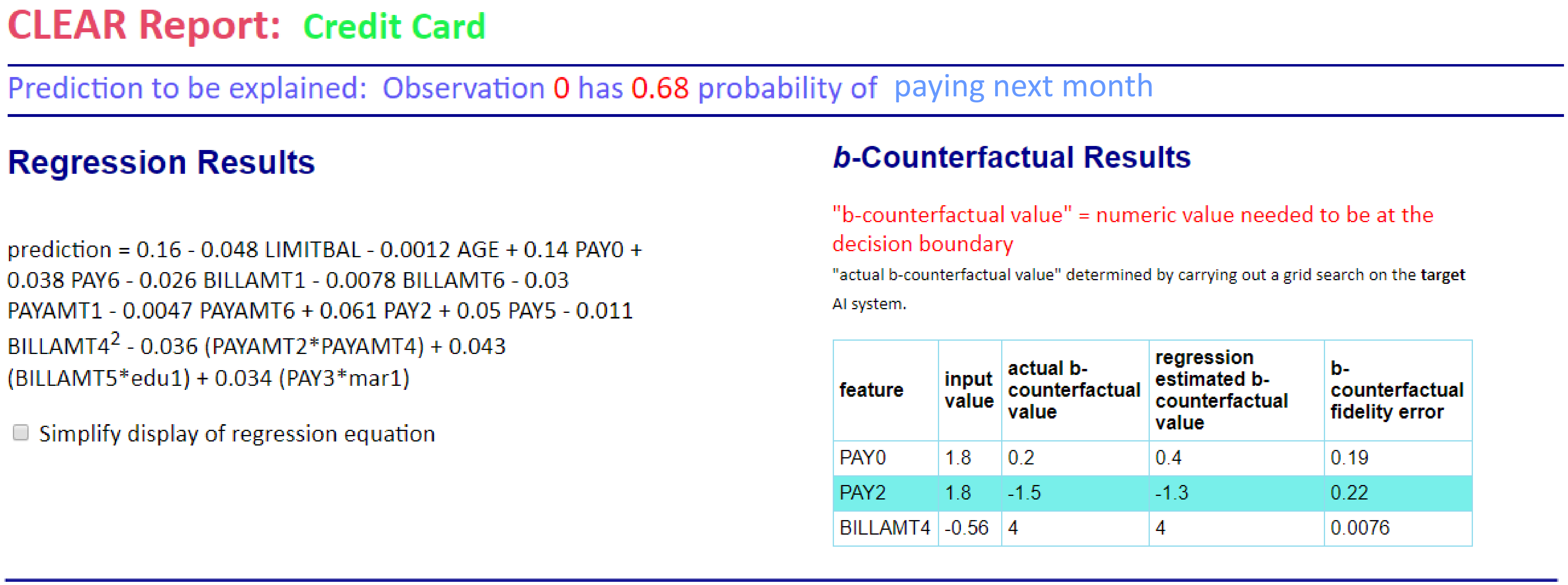
The *CLEAR* Method:

1. Determine $\boldsymbol{x}$ ‘s actual *b*-perturbations.
2. Generate synthetic observations that are then labelled by the AI system.
3. Create a balanced neighbourhood data set (including b-counterfactuals).
4. Perform a step-wise regression – including 2nd degree terms and interaction terms.
5. Estimate the *b*-perturbations
6. Measure the fidelity of the regression coefficients.

   fidelity error = | estimated *b*-perturbation - *b*-perturbation | (step 5 - step 1)

7. Iterate to best explanation.

An Example of *CLEAR*’s Output:



CLEAR Report: Credit Card

Prediction to be explained: Observation 0 has 0.68 probability of paying next month

Regression Results

prediction = 0.16 - 0.048 LIMITBAL - 0.0012 AGE + 0.14 PAY0 + 0.038 PAY6 - 0.026 BILLAMT1 - 0.0078 BILLAMT6 - 0.03 PAYAMT1 - 0.0047 PAYAMT6 + 0.061 PAY2 + 0.05 PAY5 - 0.011 BILLAMT4^2 - 0.036 (PAYAMT2*PAYAMT4) + 0.043 (BILLAMT5*edu1) + 0.034 (PAY3*mar1)

☐ Simplify display of regression equation

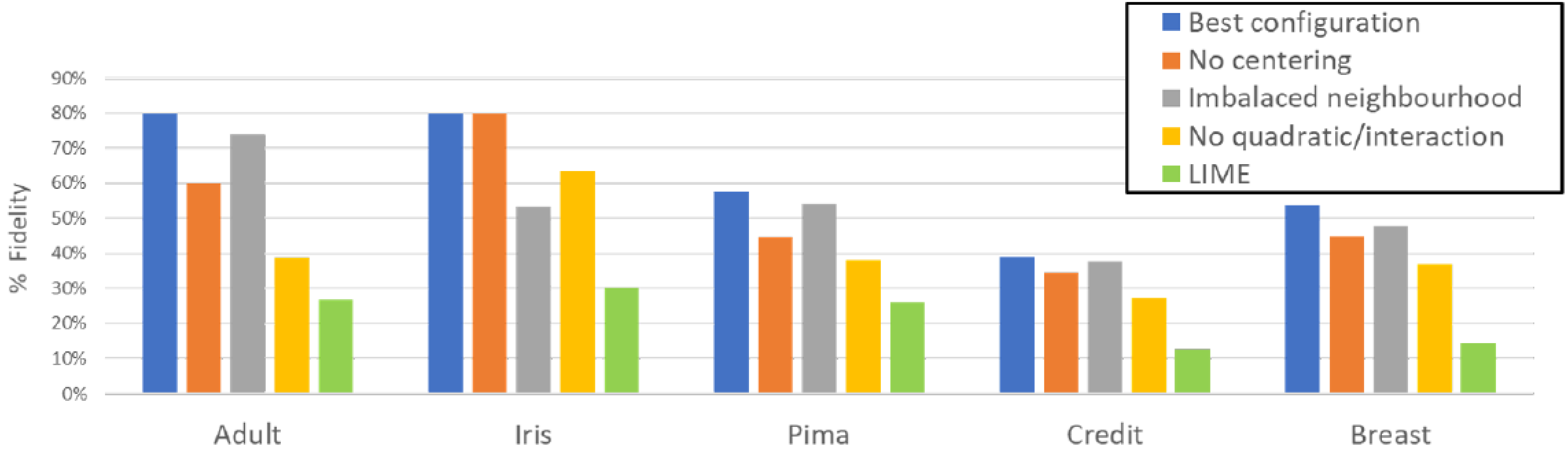
b-Counterfactual Results

"b-counterfactual value" = numeric value needed to be at the decision boundary

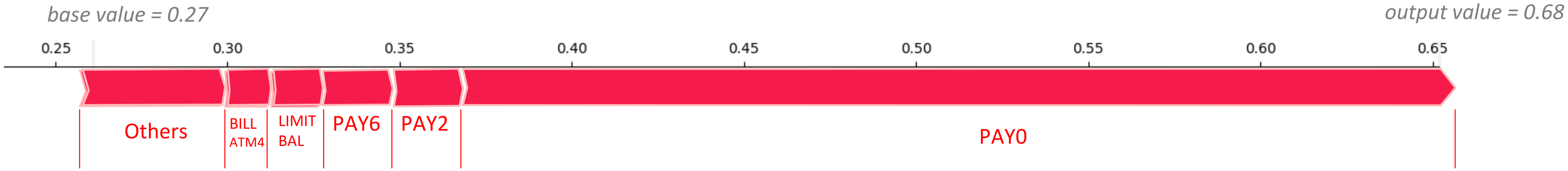
"actual b-counterfactual value" determined by carrying out a grid search on the **target** AI system.

feature	input value	actual b-counterfactual value	regression estimated b-counterfactual value	b-counterfactual fidelity error
PAY0	1.8	0.2	0.4	0.19
PAY2	1.8	-1.5	-1.3	0.22
BILLAMT4	-0.56	4	4	0.0076

CLEAR has much higher fidelity than LIME



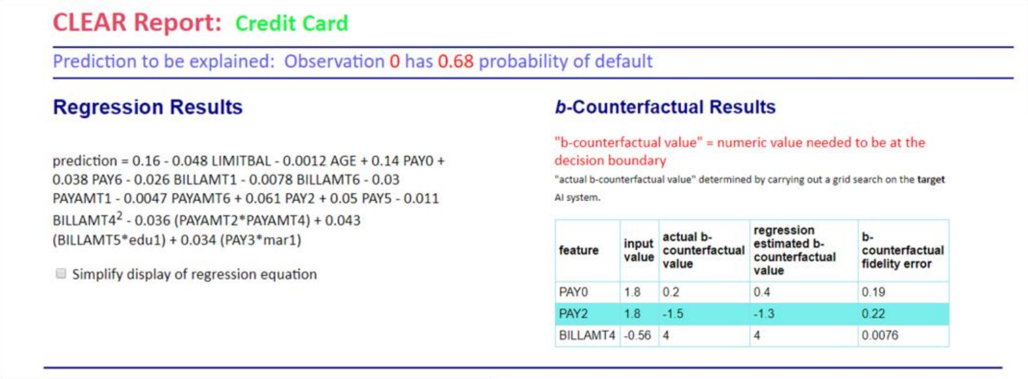
SHAP (SHapley Additive exPlanations)



SHAP
- based on Game Theory concept of Shapley values
- identifies the ‘marginal contribution’ each feature makes to a prediction
- used by consultancies and banks (eg RBS)

But
- makes major assumptions (eg linearity, ignores the complex topography of AI systems)
- often answers the *wrong question* (wrong baseline)
- does not measure fidelity

SHAP explanations often appear to be poor



CLEAR Report: **Credit Card**

Prediction to be explained: Observation 0 has 0.68 probability of default

Regression Results

prediction = 0.16 - 0.048 LIMITBAL - 0.0012 AGE + 0.14 PAY0 + 0.038 PAY6 - 0.026 BILLAMT1 - 0.0078 BILLAMT6 - 0.03 PAYAMT1 - 0.0047 PAYAMT6 + 0.061 PAY2 + 0.05 PAY5 - 0.011 BILLAMT4^2 - 0.036 (PAYAMT2*PAYAMT4) + 0.043 (BILLAMT5*edu1) + 0.034 (PAY3*mar1)

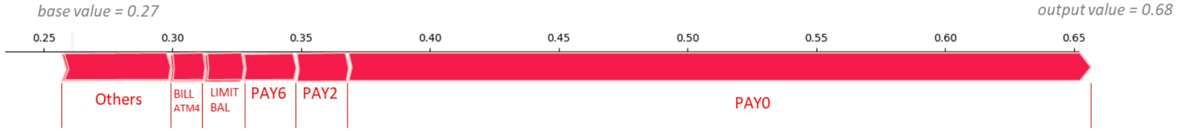
□ Simplify display of regression equation

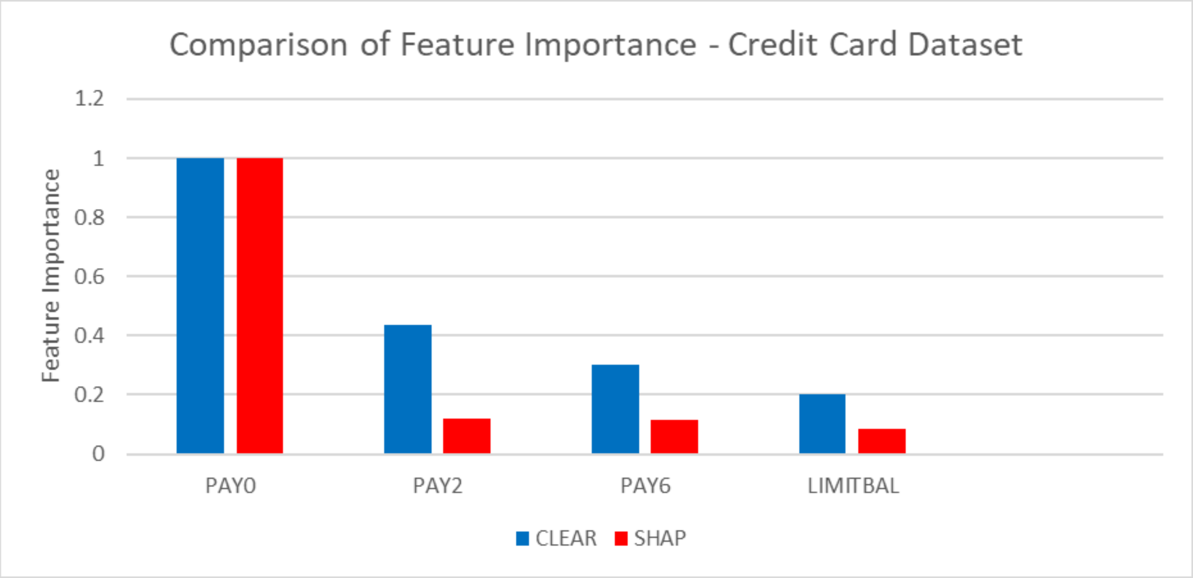
b-Counterfactual Results

"b-counterfactual value" = numeric value needed to be at the decision boundary

"actual b-counterfactual value" determined by carrying out a grid search on the **target** AI system.

feature	input value	actual b-counterfactual value	regression estimated b-counterfactual value	b-counterfactual fidelity error
PAY0	1.8	0.2	0.4	0.19
PAY2	1.8	-1.5	-1.3	0.22
BILLAMT4	-0.56	4	4	0.0076





Next Steps

- apply to time series
- identify multi-feature *b*-counterfactuals
- user trial