UNIVERSITÀ DI PISA

Peeking inside the Black Box

Visualizing statistical learning with plots of Individual Conditional Expectation (ICE)

P. Argento, A. Montinaro, M. Poiani

September 11, 2024

Statistics for Data Science 23/24
MSc Data Science and Business Informatics
Università di Pisa

Outline

1. ICE Toolbox and Simulations

2. Experiments on Real Data

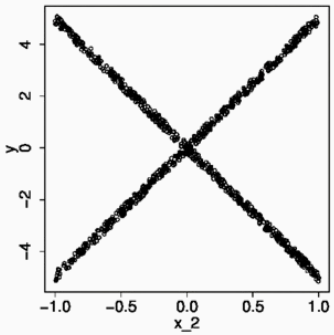
Friedman's Partial Dependence Plot

- Let $S \subset \{1, \ldots, p\}$ and let C be the complement set of S, the **Partial Dependence Function** of f on $\mathbf{x_S}$ is given by

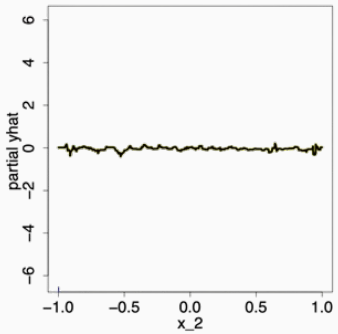
$$f_s = E_{x_C}[f(x_s, x_c)] \approx \hat{f}_S = \frac{1}{N} \sum_{i=1}^{N} \hat{f}(\mathbf{x_S}, \mathbf{x_{C_i}})$$

  where $\mathbf{x_{C_i}}$ represents the i-th observation of the features x_C .

- **Partial Dependence Plot**: is a model-agnostic visualization tool that plots the change in the average value of $\hat{f}$ as a function of x_S, which is the subset of predictors.



(a) Scatterplot of Y versus X_2



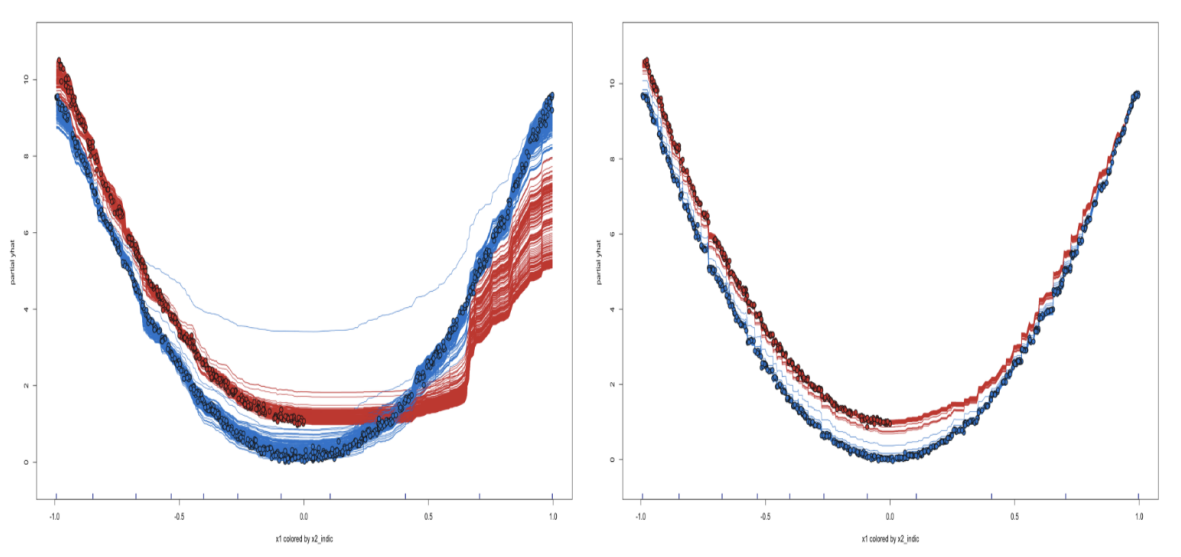
(b) PDP

PDP incorrectly suggests that there is no meaningful relationship between X2 and the predicted Y.

Extrapolation detection

ICE Plots to understand if and **how the black-box has used extrapolation**

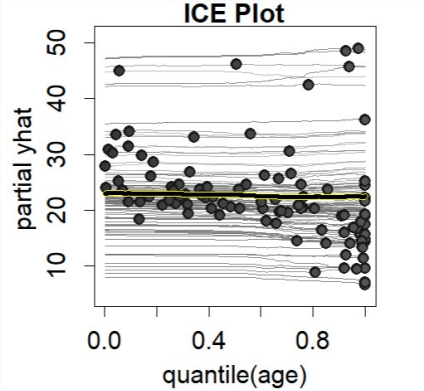
- each curve in the ICE plot includes the points representing the fitted value to find regions without observations
- danger of curse of dimensionality
- choosing the correct model



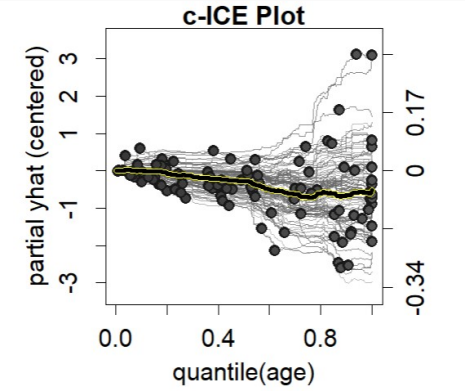
Comparison of ICE plot of RF and GBM on the same simple data synthetically generated

Individual Conditional Expectation Plots

- **ICE plots** break down the output of classical PDPs. **Each curve** is the predicted response $\hat{f}$ for **each observation**, conditioned on other observed variables x_C, showing how the prediction varies with x_S for that specific instance.
- When prediction curves have a wide range of intercepts, the clarity of the plot may be compromised. The alternative is the c-ICE. By pinching prediction lines at a chosen point, it removes **level effects** and centers all curves at a common baseline.



RF ICE plot for BHD for predictor age.

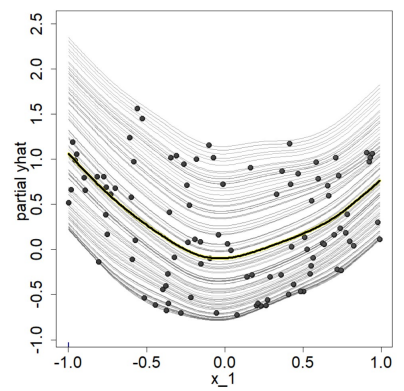


RF C-ICE plot for BHD for predictor age.

Additivity Assessment (I)

ICE plots can be used as a **diagnostic tool** to evaluate whether a fitted model $\hat{f}$ is additive.

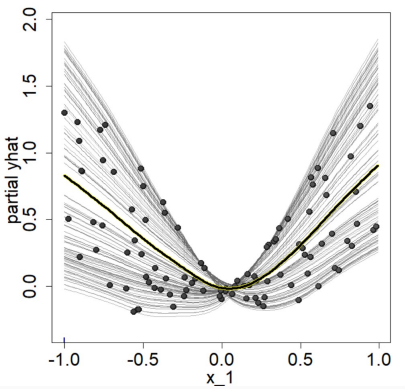
Additive Model ICE Plot



$$Y = X_1^2 + X_2 + E$$

$$X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathrm{U}(-1, 1),\ E \sim \mathcal{N}(0, 1).$$

Non-additive Model ICE Plot

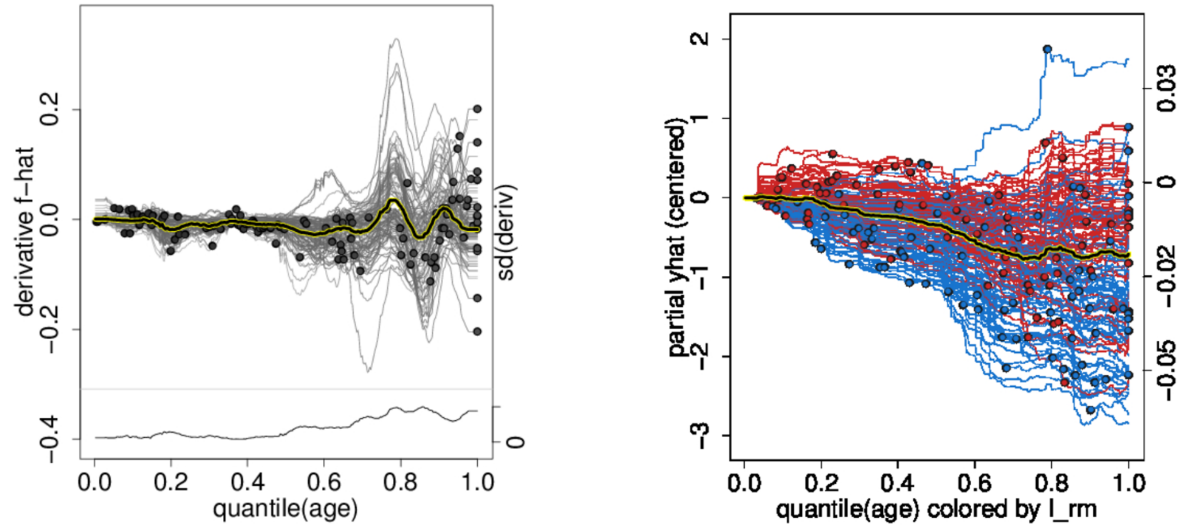


$$Y_{\text{nonadd}} = X_1^2 + X_2 X_1 + E$$

$$X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathrm{U}(-1, 1),\ E \sim \mathcal{N}(0, 1).$$

Derivative ICE Plot and other features

The d-ICE plot extends the ICE plot by focusing on the **rate of change** of predictions with respect to a predictor, enabling the detection of interaction effects in a model. d-ICE plots include also the **standard deviation of the partial derivatives** at each value of x_S, which summarizes the extent of interaction effects. **Color** allows overloading of ICE, c-ICE and d-ICE plots with information regarding a second predictor of interest.



ICEbox main functions

```
ice = function(object, X, y, predictor, predictfcn,
    verbose = TRUE, frac_to_build = 1, indices_to_
    build = NULL, num_grid_pts, logodds = FALSE,
    probit = FALSE, ...) {}
```

- `ice`: constructs ICE plots by creating grid points for the specified predictor and computes corresponding PDP;
- `dice`: computes derivative of ICE curves of the input ICE object and returns a d-ICE object containing also standard deviation of the derivatives at each grid point;
- `plot.ice` and `plot.dice`: plot ICE and d-ICE curves allowing customization of color and fraction to plot.

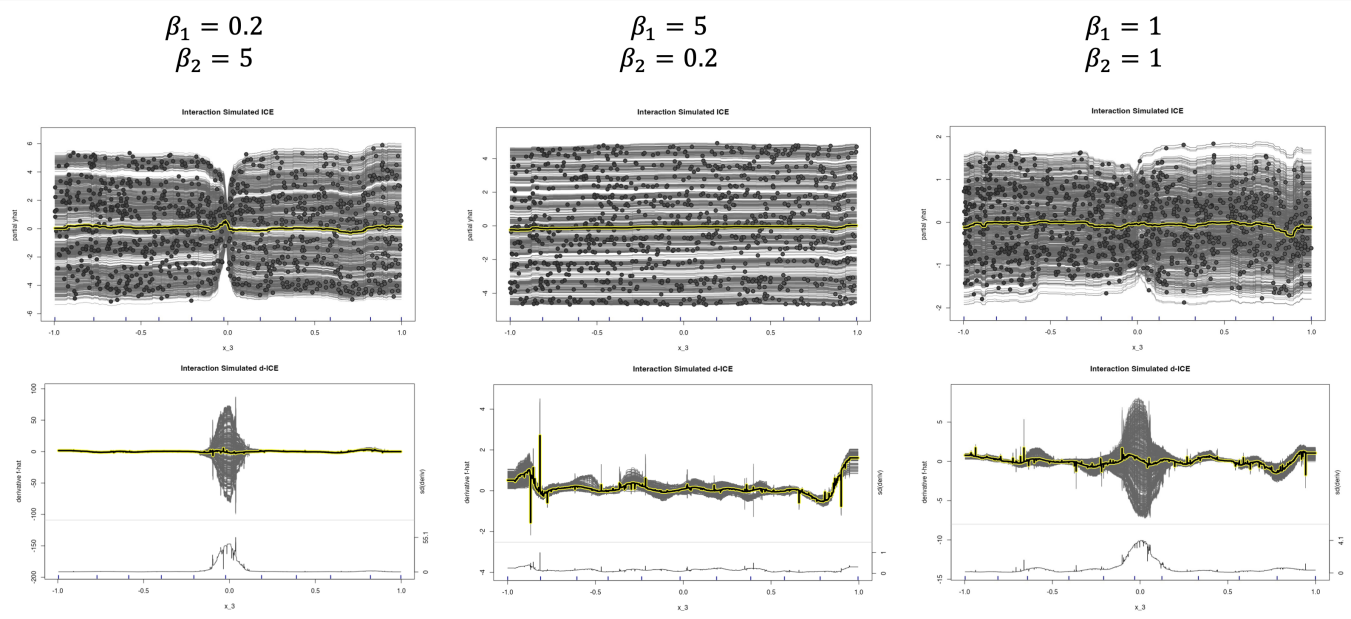
Finding interactions

In the presence of **interaction effects**, the averaging procedure in the PDP can obscure any heterogeneity in the prediction. *Example model:*

$$Y = \beta_1 X_1 - \beta_2 X_2 + 2\beta_2 X_2 \mathbb{1}_{X_3 \geq 0} + \varepsilon$$

$$\varepsilon \overset{iid}{\sim} \mathcal{N}(0,1) \qquad X_1, X_2, X_3 \overset{iid}{\sim} U(-1,1)$$

ROI (Regions of Interaction) are regions of the plot where the fitted model's interactions are concentrated.



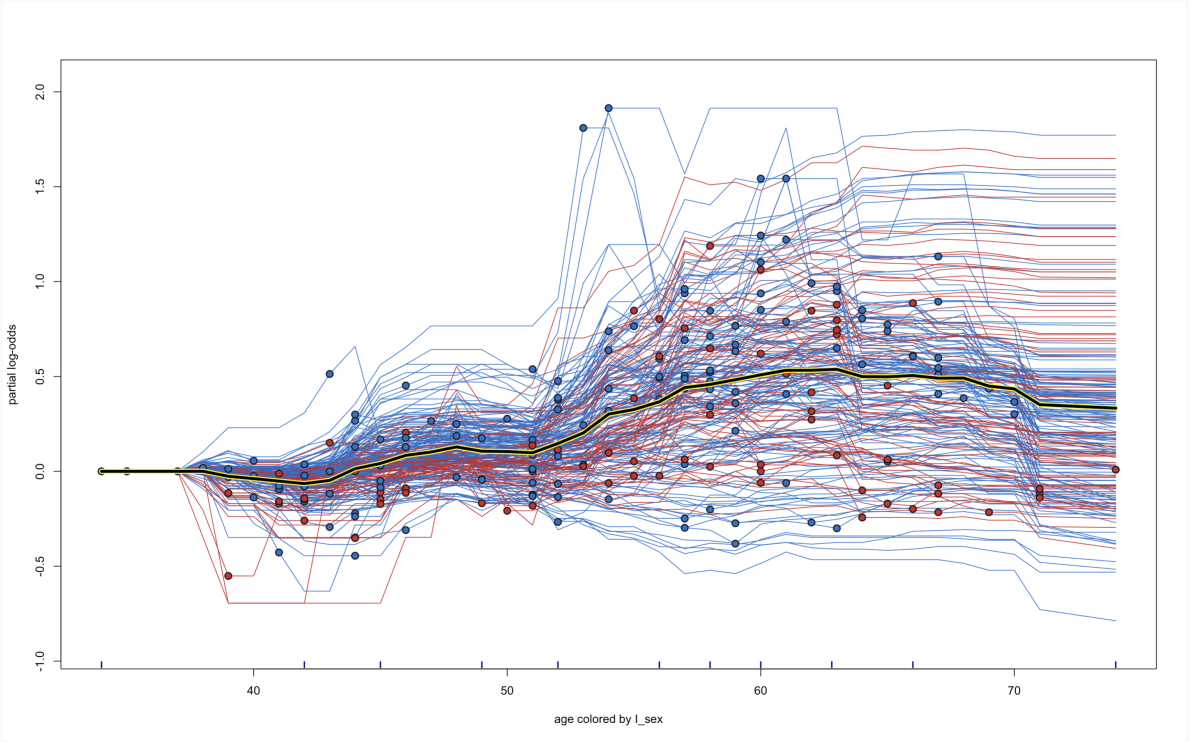
Outline

1. ICE Toolbox and Simulations

2. Experiments on Real Data

Heart Disease Dataset (I)

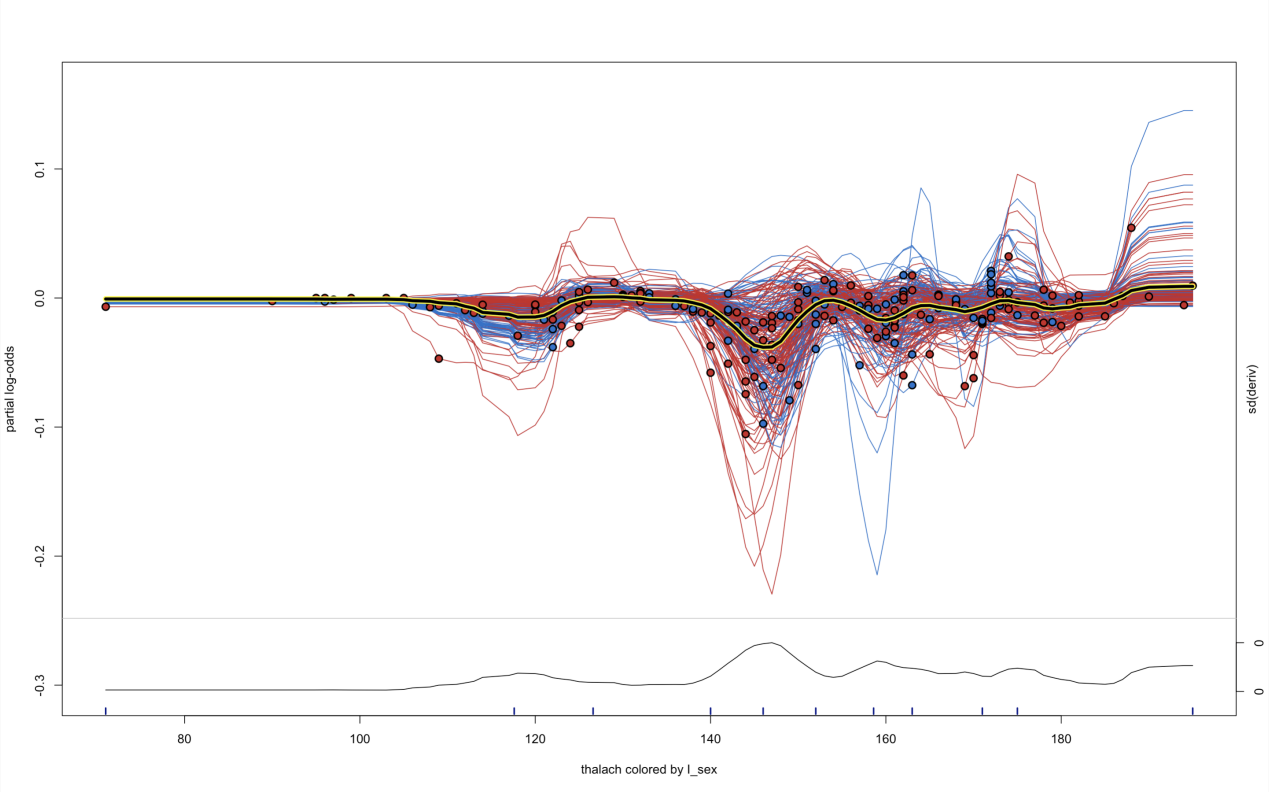
- Mixed variable dataset containing 14 variables of 297 patients for their heart disease diagnosis (binary classification task);
- ICE plots of the predictors with the highest RF importance;
- Importance of the PDP;
- Extrapolation of young females as they age.



c-ICE of age colored by sex

Heart Disease Dataset (II)

- `thalac` is the maximum heart rate achieved, meaning that when this parameter goes below 150 circa the change in prediction is the highest;
- Different peaks for male and female patients.



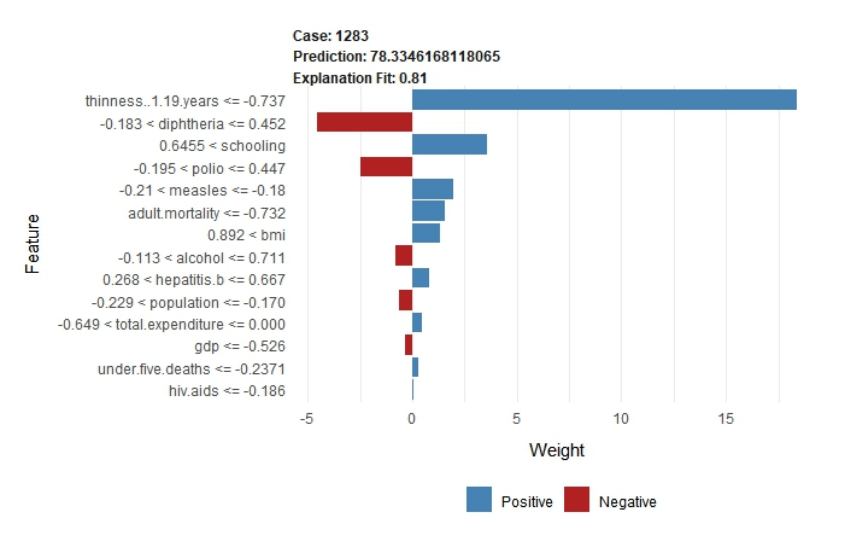
d-ICE of thalac colored by sex

Life Expectancy dataset: LIME and ICE Toolbox (I)

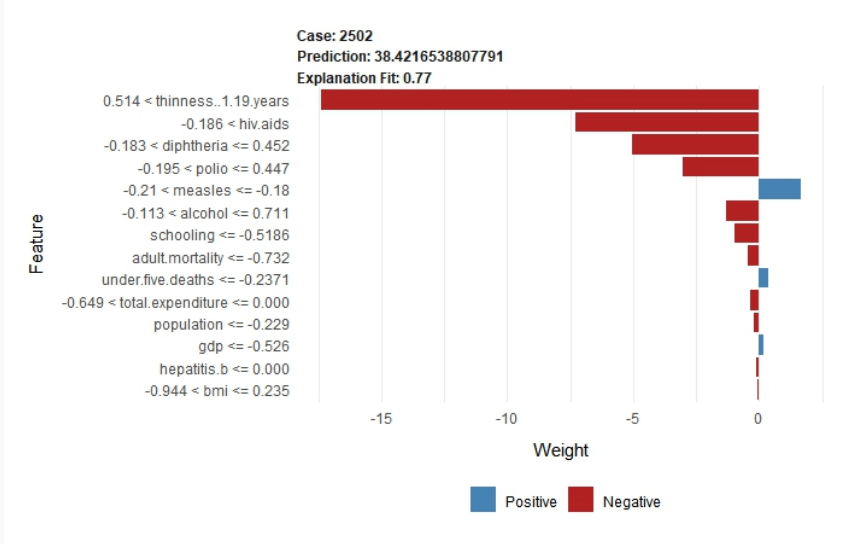
- We combined two explainability methods to deeper investigate the utility of the ICE Toolbox. The model we explained was a **Neural Network** with 4 layers.
- For conducting our tests we utilized a public dataset about **Life Expectancy**, comprising 22 columns and 2938 rows, which includes 21 predictor variables and one target variable. The health data for 193 countries come from the WHO data repository, with corresponding economic data obtained from the United Nations website.

Life Expectancy dataset: LIME and ICE Toolbox (II)

The plots present two distinct scenarios: on the left, the visualization depicts an instance characterized by a low HIV/AIDS level, whereas the plot on the right illustrates a scenario where the HIV/AIDS level is high.



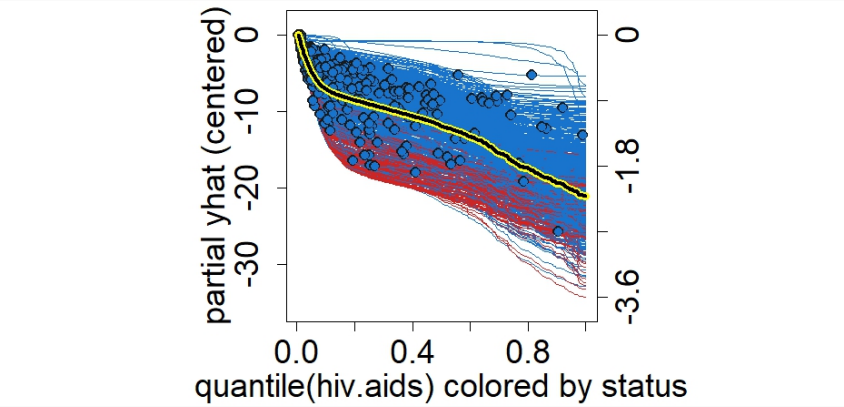
LIME results for instance 1283



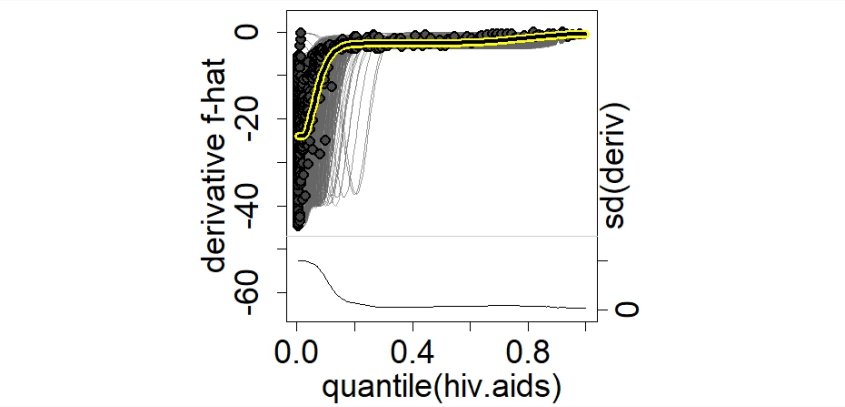
LIME results for instance 2502

Life Expectancy dataset: LIME and ICE Toolbox (III)

ICE toolbox and LIME exhibit a coordinated result. The d-ICE plot demonstrates that an increase in the independent feature has a pronounced negative impact on the prediction when the initial value is near 0. Conversely, when the actual value is already beyond the 0.2 quantile, the effect is less pronounced, and further increases in the feature level exert minimal influence on the prediction.



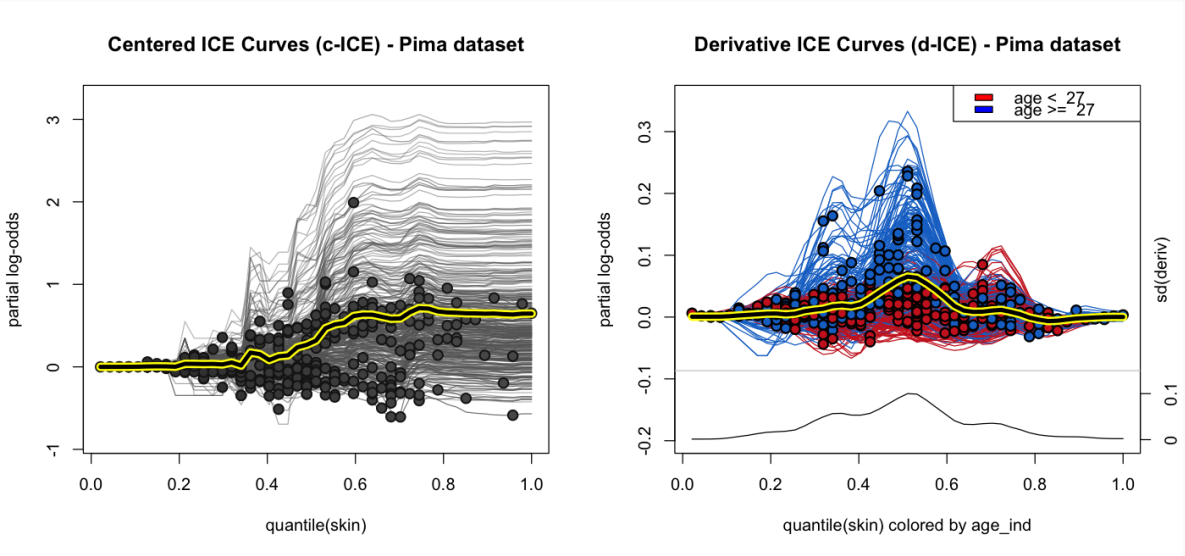
c-ICE for feature hiv.aids. Blue curves concern developing countries, red curves developed ones.



d-ICE for feature hiv.aids

Pima Indians Diabetes (I)

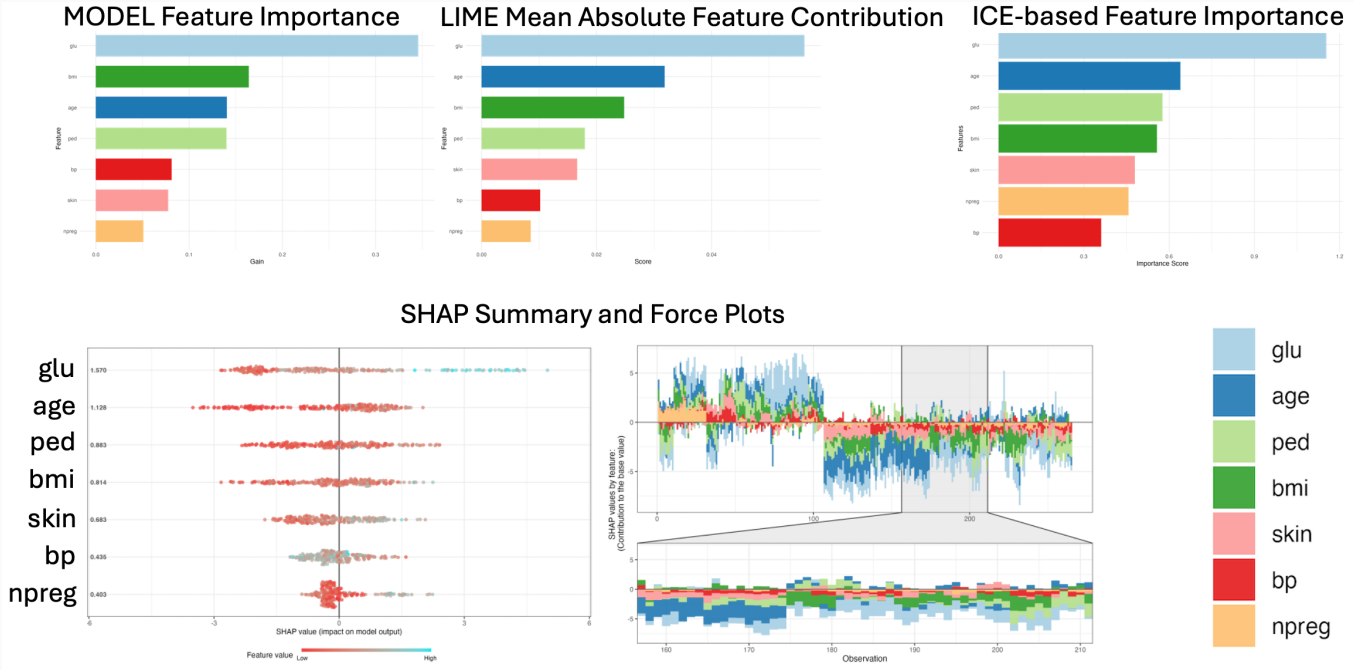
Dataset: 332 Pima Indians women, *binary classification*
Model: Random Forest, misclassification rate 22%



c-ICE shows an increased likelihood of diabetes with high "skin" values, but d-ICE shows high heterogeneity highlighting potential interactions of external factors influencing the development of diabetes.

Pima Indians Diabetes (II)

ICE-based feature importance: identifies dense regions of the feature domain to compute the variability of the ICE curves using the standard deviation. A *feature importance score* is then determined by averaging the variability of the ICE curves in these regions, with higher variability signaling a stronger impact on model predictions.



XGBoost model with 89% accuracy. All methods agree in identifying glucose concentration (`glu`) as a key predictor, followed by age or `bmi`.