MACHINE LEARNING TOOLS #4

Central European University
2024

Photo by Markus Spiske on Unsplash

@divenyijanos


Hot Dog or Not Hot Dog?
@divenyijanos
created by Leonardo.AI in March 2024


Okay. Let's start
with a hot dog.

Cases for Interpretability

- **Apple** issued a credit card offering smaller lines of credit to women 📖
- **Amazon** withdrawn an algorithm used in hiring due to gender bias 📖
- **Google** got criticized for a racist autocomplete 📖
- both **IBM** and **Microsoft** ran facial recognition algorithms that turned out to be better at recognizing men and white people 📖

Categorization of Interpretability

- **intrinsic** vs **post-hoc**
- **feature** vs **model**
- **model-specific** vs **model-agnostic**
- **global** vs **local**

Variable Importance

Model-based	Model-agnostic
for tree-based methods:	permutation-based:
total decrease of impurity due to the splits on that variable averaged over all trees	increase in error due to permutation of that variable

Variable Importance

Pros	Cons

Pros

- model agnostic
- easy to interpret

Cons

- does not reveal the direction between features and outcomes
- does not explain individual predictions
- does not tell how the prediction would change if a feature were changed

Partial-dependence profile

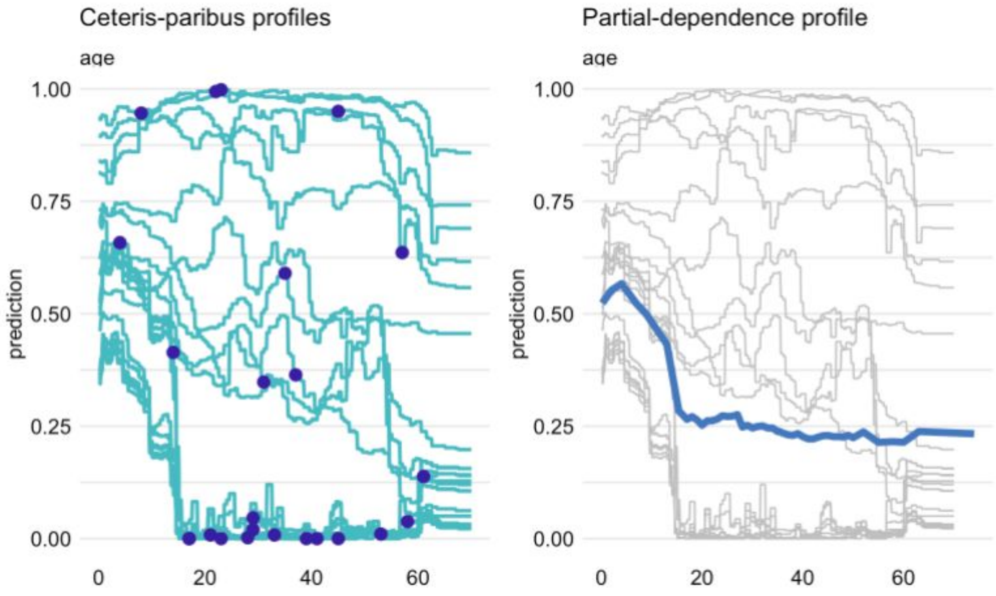


Figure 17.1: Ceteris-paribus (CP) and partial-dependence (PD) profiles for the random forest model for 25 randomly selected observations from the Titanic dataset. Left-hand-side plot: CP profiles for *age*; blue dots indicate the age and the corresponding prediction for the selected observations. Right-hand-side plot: CP profiles (grey lines) and the corresponding PD profile (blue line).

Biecek & Burzykowski:
Explanatory Model Analysis (CRC: 2021)

Partial-dependence profile

Pros	Cons

Pros

- model agnostic
- easy to interpret

Cons

- computationally expensive
- sensitive to correlated features (extrapolate to unlikely regions)

Local Interpretable Model-agnostic Explanation

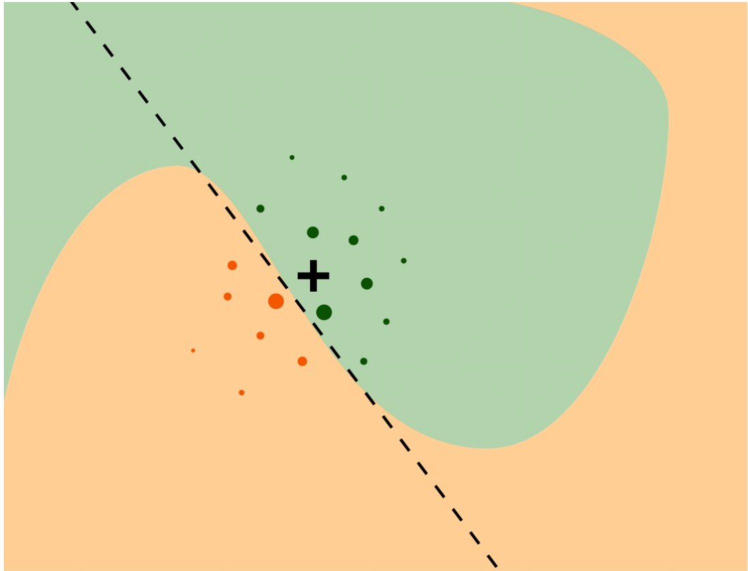
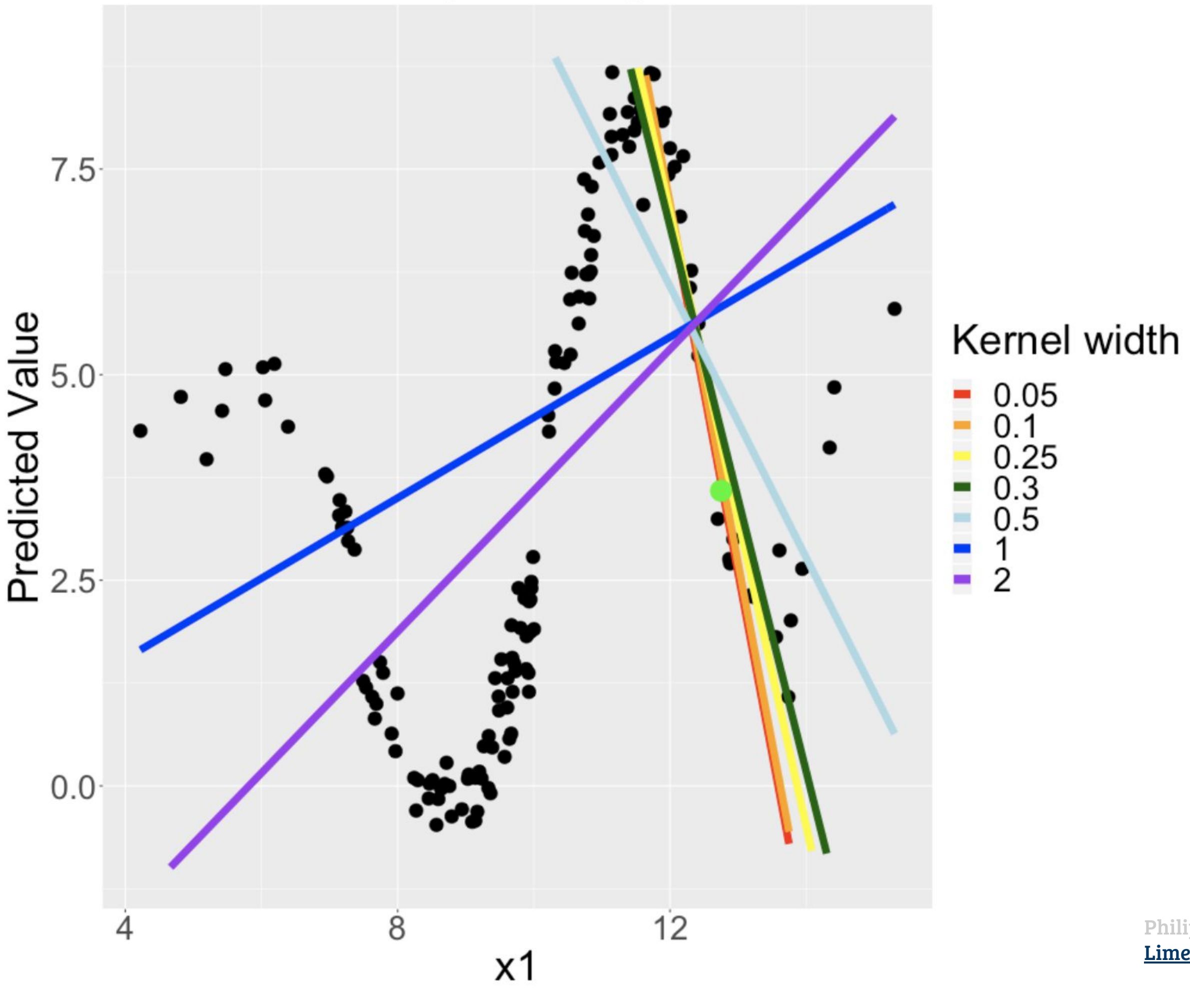


Figure 9.1: The idea behind the LIME approximation with a local glass-box model. The coloured areas correspond to decision regions for a complex binary classification model. The black cross indicates the instance (observation) of interest. Dots correspond to artificial data around the instance of interest. The dashed line represents a simple linear model fitted to the artificial data. The simple model "explains" local behavior of the black-box model around the instance of interest.

approximates a black-box model by a sparse glass-box model

Biecek & Burzykowski:
Explanatory Model Analysis (CRC: 2021)



Philipp Kopper:
[Lime and Neighborhood]

@divenyijanos

Local Interpretable Model-agnostic Explanation

Pros	Cons

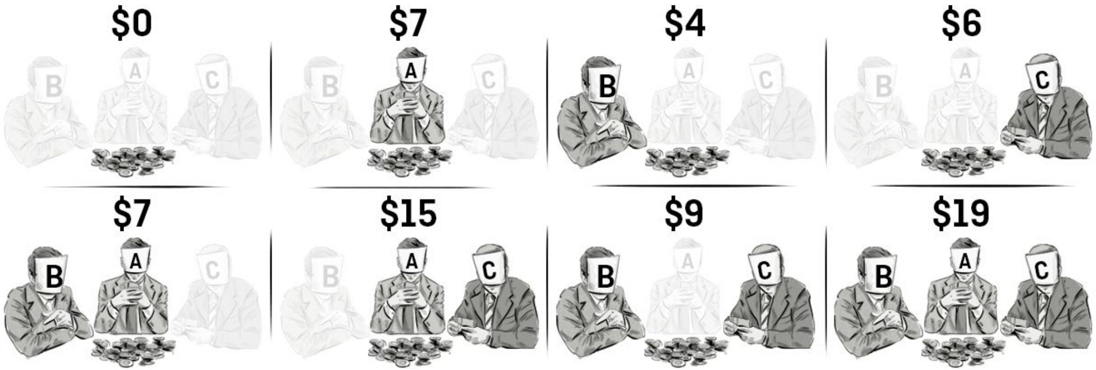
Pros

- model agnostic
- easy to interpret
- sparse even with many features

Cons

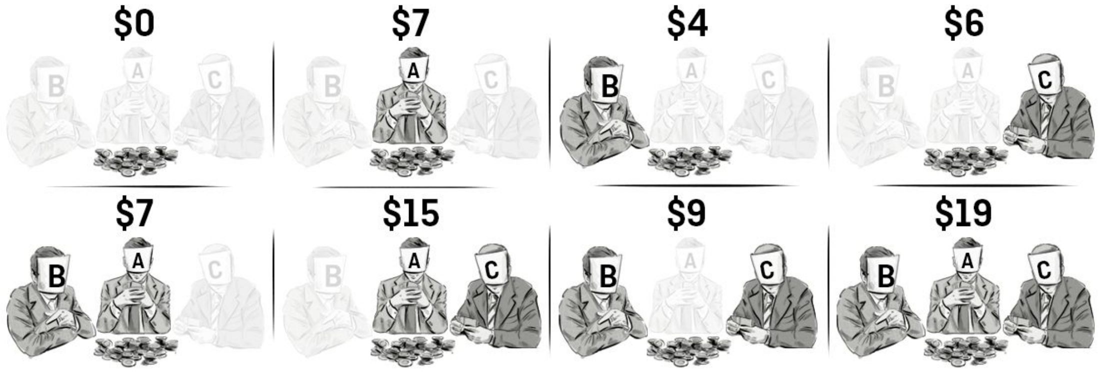
- approximates the black-box model not the data itself
- defining "local neighborhood" might be tricky – especially in high-dim settings

SHapley Additive exPlanation



@divenyijanos

SHapley Additive exPlanation



Value of *A*:
~AVG(7, 7-4, 15-6, 19-9)

average the marginal contributions
across each permutations

SHapley Additive exPlanation

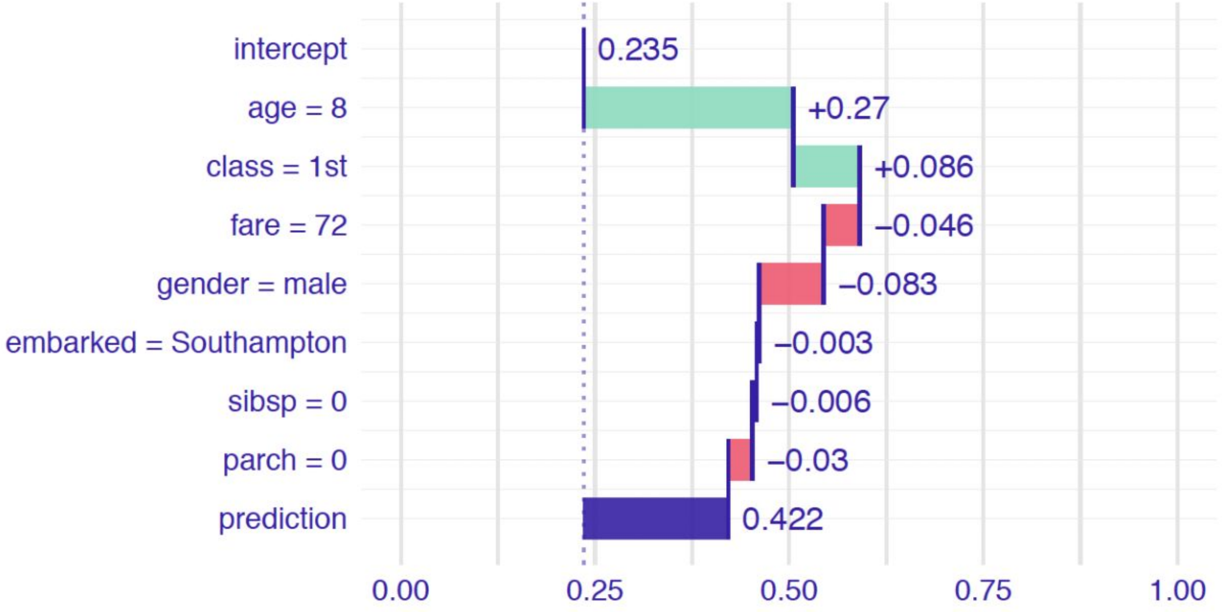


Figure 6.1: Break-down plots show how the contributions attributed to individual explanatory variables change the mean model's prediction to yield the actual prediction for a particular single instance

Biecek & Burzykowski: Explanatory Model Analysis (CRC: 2021)

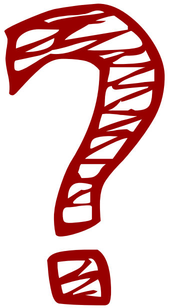
@divenyijanos

SHapley Additive exPlanation

Pros	Cons
• model agnostic • strong formal foundation derived from the cooperative games theory	• additive: if the model is not additive, SHAP values can mislead • time-consuming for large models

Which set to use for calculating these metrics?

Does it help for actionability?



Prediction vs Causation

Double Machine Learning (Chernozhukov et al.)

$$
\begin{aligned}
y_i &= \theta d_i + g_0(x_i) + \zeta_i \\
d_i &= m_0(x_i) + v_i
\end{aligned}
$$

Double Machine Learning (Chernozhukov et al.)

$$y_i = \theta d_i + g_0(x_i) + \zeta_i$$
$$d_i = m_0(x_i) + v_i$$

Double Machine Learning (Chernozhukov et al.)

$$y_i = \theta d_i + g_0(x_i) + \zeta_i$$
$$d_i = m_0(x_i) + v_i$$

- estimate nuisance functions *g0* and *m0* with flexible (ML) models
- orthogonalize *D* to avoid regularization bias
  - estimate D by X $\rightarrow$ calculate residuals *eps_m*
  - estimate Y by X $\rightarrow$ calculate residuals *eps_g*
  - regress *eps_g* on *eps_m* to recover *theta*

Double Machine Learning (Chernozhukov et al.)

$$y_i = \theta d_i + g_0(x_i) + \zeta_i$$
$$d_i = m_0(x_i) + v_i$$

- estimate nuisance functions *g0* and *m0* with flexible (ML) models
- orthogonalize *D* to avoid regularization bias
  - estimate D by X → calculate residuals *eps_m*
  - estimate Y by X → calculate residuals *eps_g*
  - regress *eps_g* on *eps_m* to recover *theta*

👉 simulation exercise
on the Double ML package site

Recommended Materials

Video:

- Florian ?? (Deepfindr): Explainable AI (videos 1-4)

Text:

- Christoph Molnar: Interpretable Machine Learning