Interpretable Machine Learning

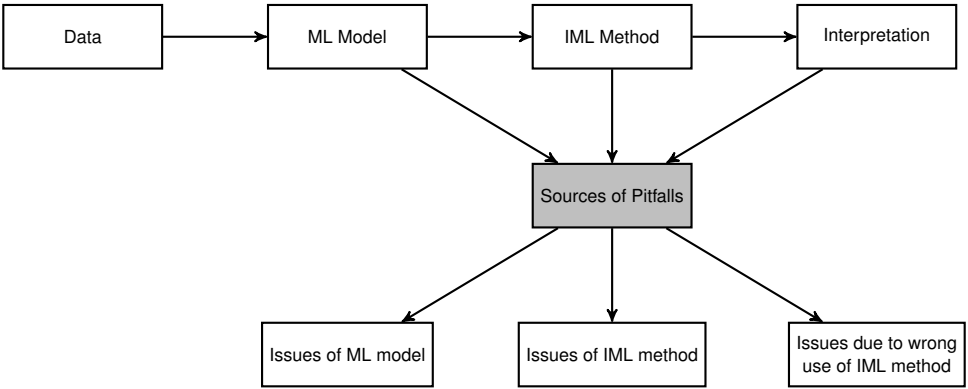
Pitfalls and Best Practices

Learning goals

- General pitfalls of interpretation methods
- Practices to avoid pitfalls

SOURCES OF PITFALLS

▸ Molnar et. al (2021)



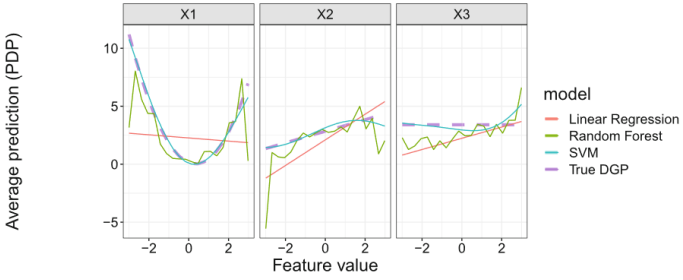
ISSUES OF ML MODEL ▸ Molnar et. al (2021)

- **Proper training and evaluation**: To gain insights into data generating process, deployed model should at least generalize well to unseen data (garbage in, garbage out)

ISSUES OF ML MODEL

▸ Molnar et. al (2021)

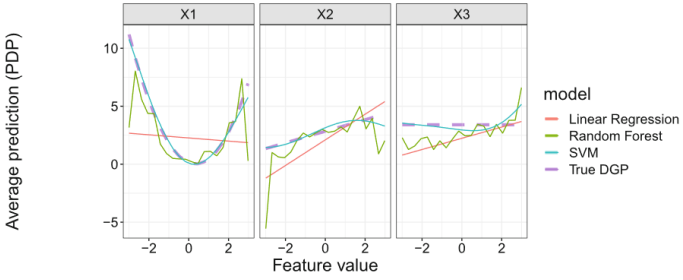
- **Proper training and evaluation**: To gain insights into data generating process, deployed model should at least generalize well to unseen data (garbage in, garbage out)
  Example: Three features are drawn from a uniform distribution, and the target is generated as $Y = X_1^2 + X_2 - 5X_1X_2 + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 5)$. Figure: PDPs for the DGP and for a linear regression model (underfitted), a random forest (overfitted) and a support vector machine with radial basis kernel (good fit).



ISSUES OF ML MODEL

▸ Molnar et. al (2021)

- **Proper training and evaluation**: To gain insights into data generating process, deployed model should at least generalize well to unseen data (garbage in, garbage out)
  Example: Three features are drawn from a uniform distribution, and the target is generated as $Y = X_1^2 + X_2 - 5X_1X_2 + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 5)$. Figure: PDPs for the DGP and for a linear regression model (underfitted), a random forest (overfitted) and a support vector machine with radial basis kernel (good fit).



- **Avoid unnecessary complexity**: Prefer simple interpretable models and use them as baseline

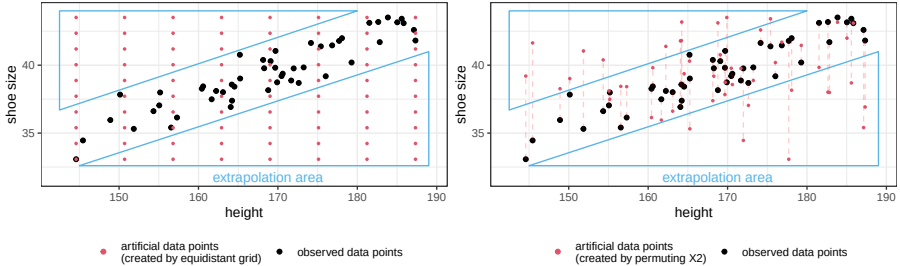
ISSUES OF IML METHOD ▸ Molnar et. al (2021)

- **Consider dependencies**: Some interpretation methods suffer when features are dependent
  $\rightsquigarrow$ Check presence of dependencies and use suitable methods

ISSUES OF IML METHOD

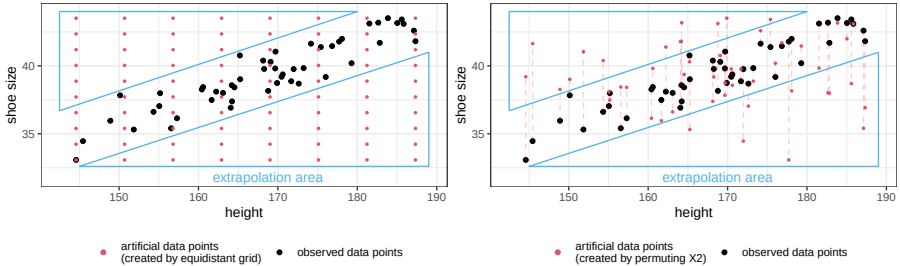
Molnar et. al (2021)

- **Consider dependencies**: Some interpretation methods suffer when features are dependent
  ⇝ Check presence of dependencies and use suitable methods
  Example: Extrapolation



ISSUES OF IML METHOD ▸ Molnar et. al (2021)

- **Consider dependencies**: Some interpretation methods suffer when features are dependent
  ⇝ Check presence of dependencies and use suitable methods
  Example: Extrapolation



- **Beware of simplifications**: Mapping of complex models to low-dim. explanations
  ⇝ Information loss, e.g., some interpretation methods hide interactions (Figure: PDP and ICE Curves)

slides/03_feature-effects/f

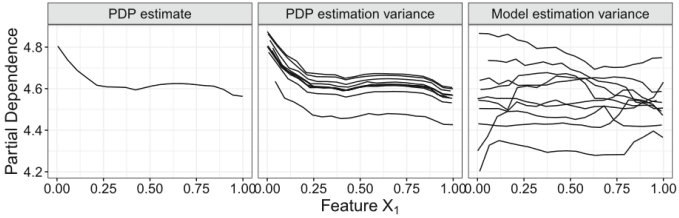
ISSUES DUE TO WRONG USE OF IML METHOD ▸ Molnar et. al (2021)

- **Quantify uncertainty**: Interpretation methods are often (statistical) estimators
  $\rightsquigarrow$ Beware of uncertainty, we may need confidence intervals

ISSUES DUE TO WRONG USE OF IML METHOD

▸ Molnar et. al (2021)

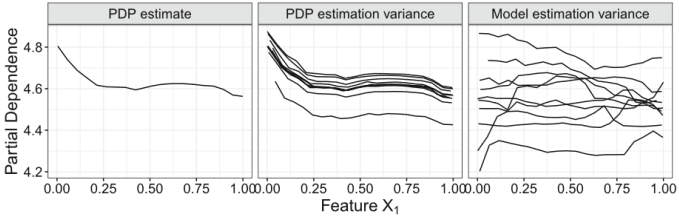
- **Quantify uncertainty**: Interpretation methods are often (statistical) estimators
  $\rightsquigarrow$ Beware of uncertainty, we may need confidence intervals
  Example: Left plot (IML method output) misleading compared to fitted models in right plot



ISSUES DUE TO WRONG USE OF IML METHOD

Molnar et. al (2021)

- **Quantify uncertainty**: Interpretation methods are often (statistical) estimators
  $\rightsquigarrow$ Beware of uncertainty, we may need confidence intervals
  Example: Left plot (IML method output) misleading compared to fitted models in right plot



- **Careful with causality**: Do you want to understand the model or the nature of DGP?
  $\rightsquigarrow$ Your goal should guide the choice of interpretation method