Interpretable Machine Learning

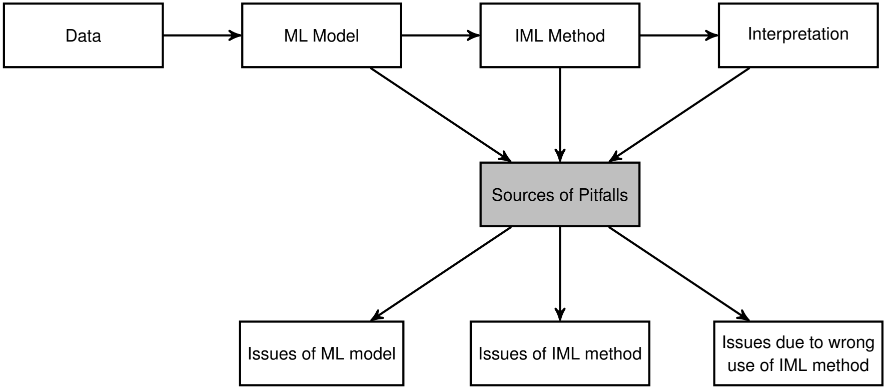
Pitfalls and Best Practices

Learning goals

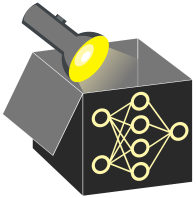
- General pitfalls of interpretation methods
- Practices to avoid pitfalls

SOURCES OF PITFALLS

▸ Molnar et. al (2021)

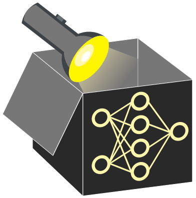


ISSUES OF ML MODEL ▸ Molnar et. al (2021)

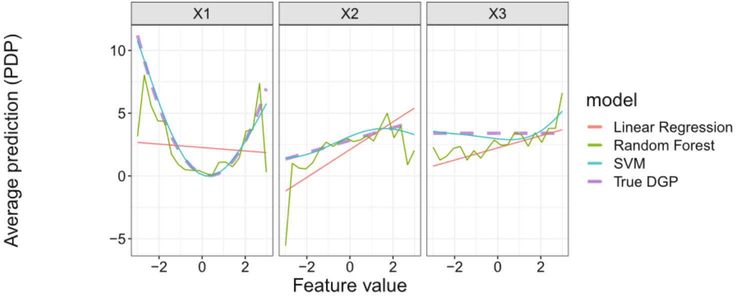


- **Proper training and evaluation**: To gain insights into DGP, deployed model should generalize well to unseen data (garbage in, garbage out)

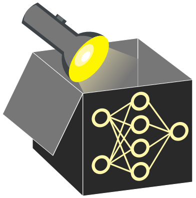
ISSUES OF ML MODEL ▸ Molnar et. al (2021)



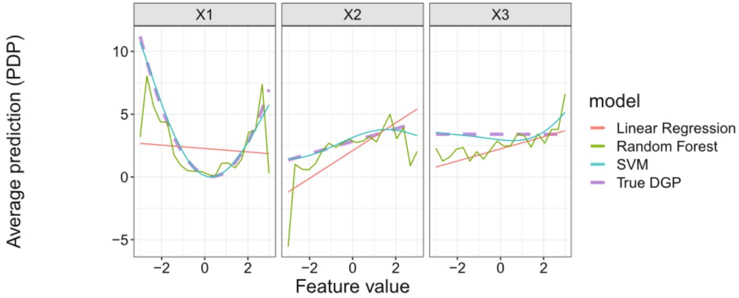
- **Proper training and evaluation**: To gain insights into DGP, deployed model should generalize well to unseen data (garbage in, garbage out)
  Example: $X_1, X_2, X_3 \sim \mathit{Unif}(-3, 3)$ with $Y = X_1^2 + X_2 - 5X_1X_2 + \epsilon$, $\epsilon \sim \mathcal{N}(0, 5)$
  Figure: PDP of DGP (true effect), linear regression model (underfitted), random forest (overfitted), and SVM with radial basis kernel (good fit).



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

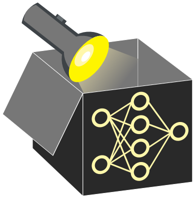


- **Proper training and evaluation**: To gain insights into DGP, deployed model should generalize well to unseen data (garbage in, garbage out)
  Example: $X_1, X_2, X_3 \sim \mathit{Unif}(-3, 3)$ with $Y = X_1^2 + X_2 - 5X_1X_2 + \epsilon$, $\epsilon \sim \mathcal{N}(0, 5)$
  Figure: PDP of DGP (true effect), linear regression model (underfitted), random forest (overfitted), and SVM with radial basis kernel (good fit).



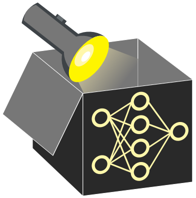
- **Avoid unnecessary complexity**: Prefer simple interpretable models and use them as baseline, move to more complex models if performance not sufficient

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

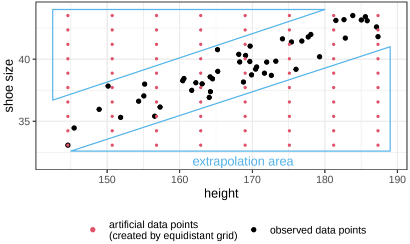


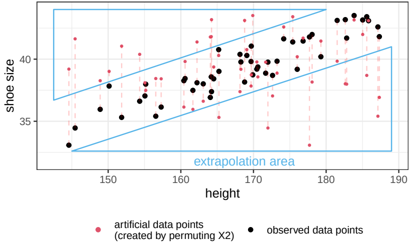
- **Consider dependencies**: Some interpretation methods have issues in case of dependent features
  $\rightsquigarrow$ Check presence of dependencies and use suitable interpretation methods

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

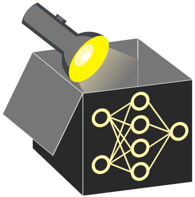


- **Consider dependencies**: Some interpretation methods have issues in case of dependent features
  ⇝ Check presence of dependencies and use suitable interpretation methods
  Example: Explanations may rely on unreliable pred. where model extrapolated

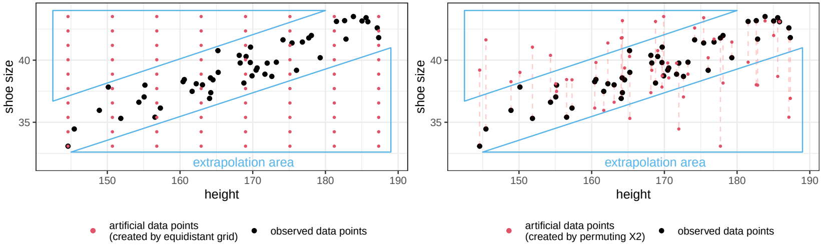




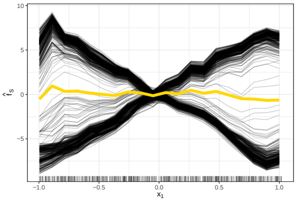
ISSUES OF IML METHOD ▸ Molnar et. al (2021)



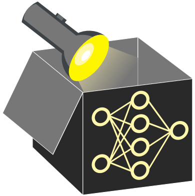
- **Consider dependencies**: Some interpretation methods have issues in case of dependent features
  ⇝ Check presence of dependencies and use suitable interpretation methods
  Example: Explanations may rely on unreliable pred. where model extrapolated



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

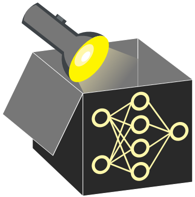


INTERPRETATIONS WITH DEPENDENT FEATURES

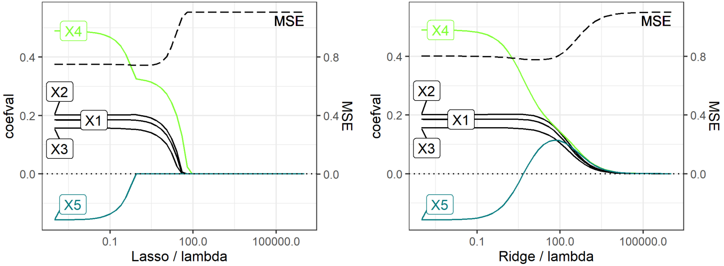


- Highly correlated features contain similar information
  - $\rightsquigarrow$ Model might pick only 1 feat. (regularization), even if it is causally irrelevant
  - $\rightsquigarrow$ Produced explanations can be misleading (true to model, but not to data)
  - $\rightsquigarrow$ E.g., different interpretable models produce different results

INTERPRETATIONS WITH DEPENDENT FEATURES

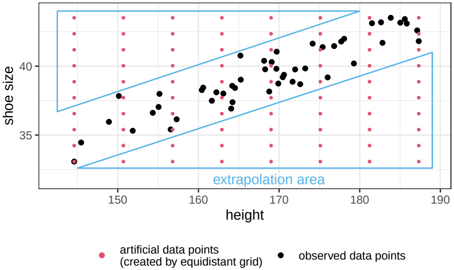


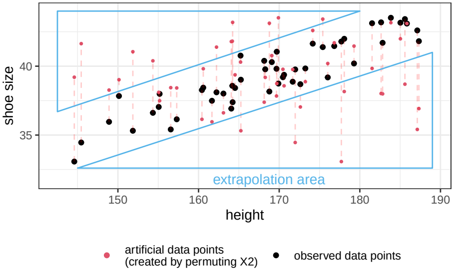
- Highly correlated features contain similar information
  - ⇝ Model might pick only 1 feat. (regularization), even if it is causally irrelevant
  - ⇝ Produced explanations can be misleading (true to model, but not to data)
  - ⇝ E.g., different interpretable models produce different results
- **Example:** Simulate 100 obs. from DGP $Y = 0.2(X_1 + \cdots + X_5) + \epsilon, \epsilon \sim N(0, 1)$

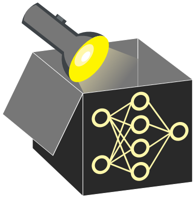


  - $X_1, \ldots, X_4 \sim N(0, 2)$ (uncorrelated)
  - $X_5 = X_4 + \delta, \delta \sim N(0, 0.3) \Rightarrow \rho(X_4, X_5) = 0.98$ (highly correlated)
  - LASSO: Shrinks coef. of X_5 to zero, coef. of X_4 about $1.5\times$ higher
  - Ridge: Similar coef. for X_4 and X_5 for higher lambda

EXTRAPOLATION DUE TO DEPENDENCIES

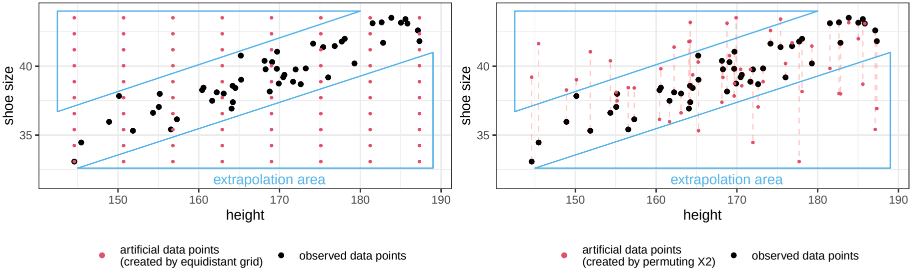


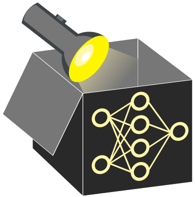




- Many interpretation methods are based on artificially created data points
  - ⇝ Many points lie in low-density regions if features are dependent
  - ⇝ Predictions in such regions have high uncertainty
  - ⇝ Explanations can be biased if they rely on pred. where model extrapolated

EXTRAPOLATION DUE TO DEPENDENCIES

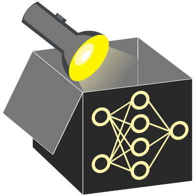




- Many interpretation methods are based on artificially created data points
  - ⇝ Many points lie in low-density regions if features are dependent
  - ⇝ Predictions in such regions have high uncertainty
  - ⇝ Explanations can be biased if they rely on pred. where model extrapolated
- There is no definition of when a model extrapolates and to what degree
  - ⇝ Severity of extrapolation depends on model
  - ⇝ Density of train data may helps identify regions where extrapolation is likely
    But: Density estimation in many dimensions is often infeasible

ISSUE: WRONG USE OF IML METHOD

▸ Molnar et. al (2021)

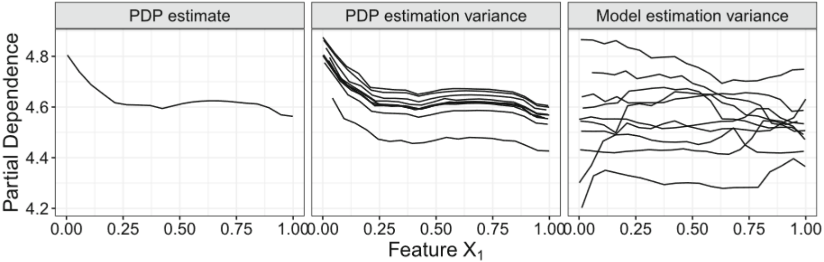


- **Quantify uncertainty**: Interpretation methods are often (statistical) estimators
  $\rightsquigarrow$ Beware of uncertainty, we may need confidence intervals

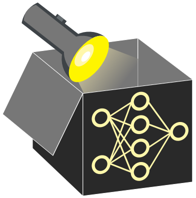
ISSUE: WRONG USE OF IML METHOD

▸ Molnar et. al (2021)

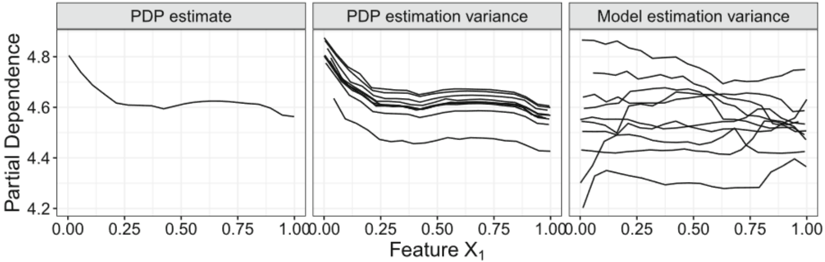
- **Quantify uncertainty**: Interpretation methods are often (statistical) estimators
  $\rightsquigarrow$ Beware of uncertainty, we may need confidence intervals
  Example: Left plot (IML method output) misleading compared to fitted models in right plot



ISSUE: WRONG USE OF IML METHOD ▸ Molnar et. al (2021)



- **Quantify uncertainty**: Interpretation methods are often (statistical) estimators
  ⇝ 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**: Want to understand the model or the nature of DGP?
  ⇝ Goal should guide the choice of interpretation method