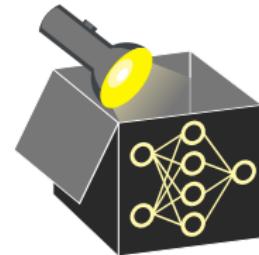
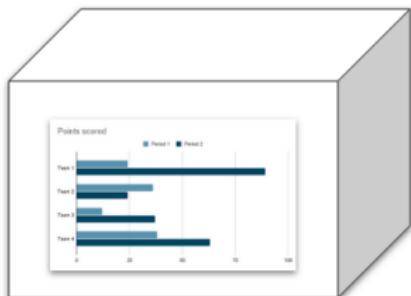


Interpretable Machine Learning



Interpretable Models 1

Inherently Interpretable Models - Motivation

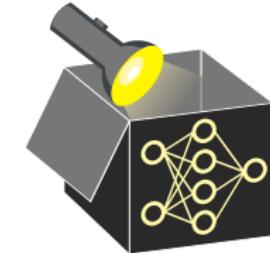
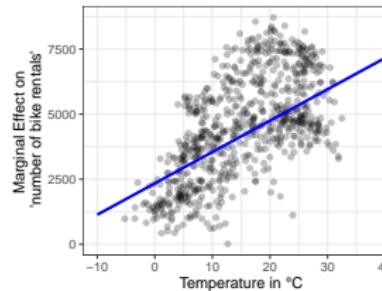


Learning goals

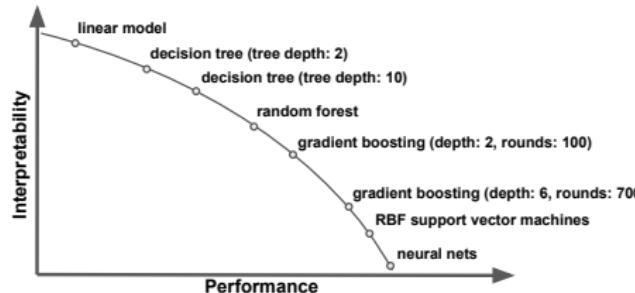
- Why should we use interpretable models?
- Advantages and disadvantages of interpretable models

MOTIVATION

- Achieving interpretability by using interpretable models is the most straightforward approach
- Classes of models deemed interpretable:
 - (Generalized) linear models (LM, GLM)
 - Generalized additive models (GAM)
 - Decision trees
 - Rule-based learning
 - Model-based / component-wise boosting
 - ...
- Often there is a trade-off between interpretability and model performance

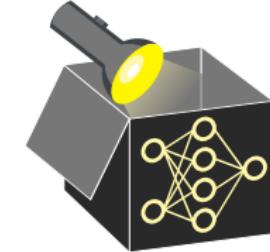
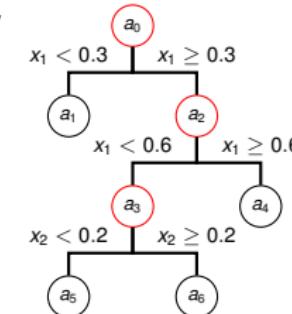


~~ LM provides straightforward interpretation



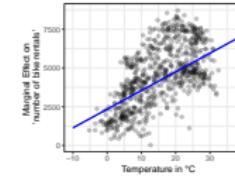
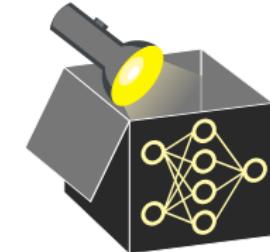
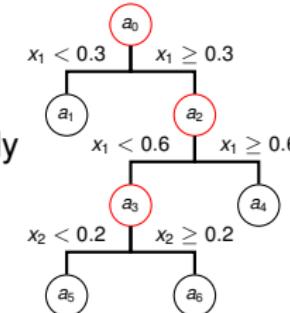
ADVANTAGES

- Interpretable models are transparent by design, making many model-agnostic explanation methods unnecessary
~~ Eliminates an extra source of estimation error



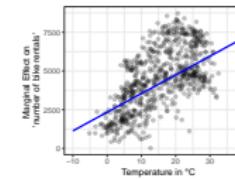
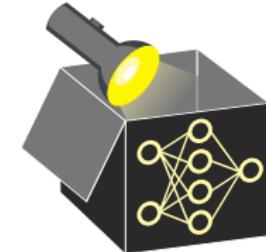
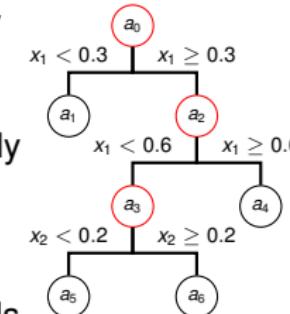
ADVANTAGES

- Interpretable models are transparent by design, making many model-agnostic explanation methods unnecessary
~~ Eliminates an extra source of estimation error
- They often have few hyperparameters and are structurally simple (e.g., linear, additive, sparse, monotonic)
~~ Easy to train, fast to tune, straightforward to explain



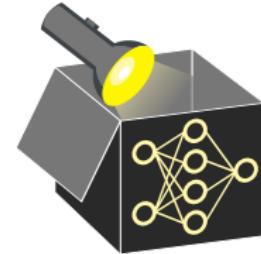
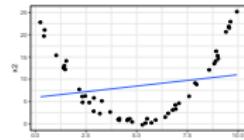
ADVANTAGES

- Interpretable models are transparent by design, making many model-agnostic explanation methods unnecessary
~~ Eliminates an extra source of estimation error
- They often have few hyperparameters and are structurally simple (e.g., linear, additive, sparse, monotonic)
~~ Easy to train, fast to tune, straightforward to explain
- Many people are familiar with simple interpretable models
~~ Increases trust, facilitates communication of results



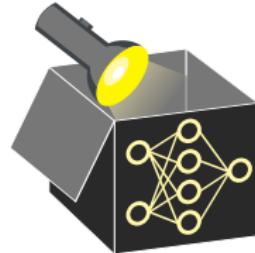
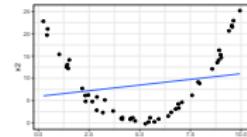
DISADVANTAGES & LIMITATIONS

- Often require assumptions about data / model structure
~~ If assumptions are wrong, models may perform bad



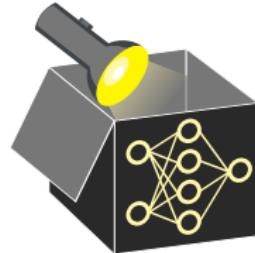
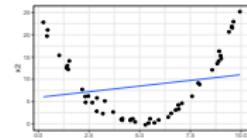
DISADVANTAGES & LIMITATIONS

- Often require assumptions about data / model structure
~~ If assumptions are wrong, models may perform bad
- Interpretable models may also be hard to interpret, e.g.:
 - LM with lots of features and interactions
 - Decision trees with huge tree depth



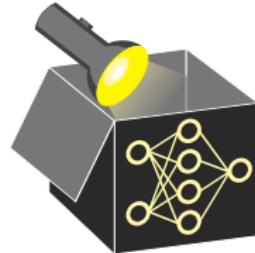
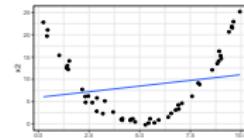
DISADVANTAGES & LIMITATIONS

- Often require assumptions about data / model structure
~~ If assumptions are wrong, models may perform bad
- Interpretable models may also be hard to interpret, e.g.:
 - LM with lots of features and interactions
 - Decision trees with huge tree depth
- Often do not automatically model complex relationships due to limited flexibility
e.g., high-order main or interaction effects need to be specified manually in an LM



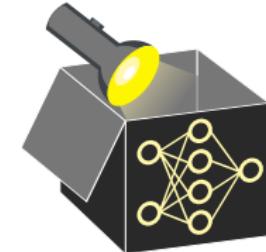
DISADVANTAGES & LIMITATIONS

- Often require assumptions about data / model structure
 - ~~ If assumptions are wrong, models may perform bad
- Interpretable models may also be hard to interpret, e.g.:
 - LM with lots of features and interactions
 - Decision trees with huge tree depth
- Often do not automatically model complex relationships due to limited flexibility
 - e.g., high-order main or interaction effects need to be specified manually in an LM
- Inherently interpretable models do not address all explanation needs
 - ~~ Complementary model-agnostic methods (e.g., counterfactuals) remain valuable for specific tasks



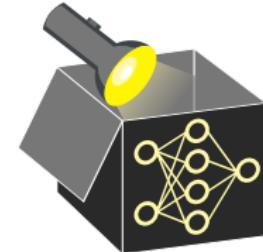
FURTHER COMMENTS

- Some researchers advocate for inherently interpretable models instead of explaining black boxes after training ▶ "Rudin" 2019
 - Built-in interpretation
~~ fewer risks from misleading post-hoc explanations
 - Good performance possible with effort on preprocessing and/or feature engineering
 - But interpretability depends on meaning of created features
~~ E.g., PCA keeps models linear, but yields hard-to-interpret components



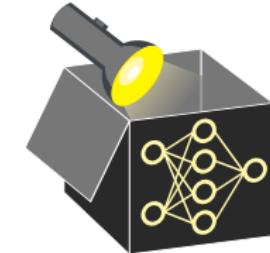
FURTHER COMMENTS

- Some researchers advocate for inherently interpretable models instead of explaining black boxes after training ▶ "Rudin" 2019
 - Built-in interpretation
~~ fewer risks from misleading post-hoc explanations
 - Good performance possible with effort on preprocessing and/or feature engineering
 - But interpretability depends on meaning of created features
~~ E.g., PCA keeps models linear, but yields hard-to-interpret components
- Limitation: Less suited for complex data requiring end-to-end learning
 - Applies to image, text, or sensor data where features must be learned from raw input
 - Manual extraction of interpretable features is difficult
⇒ Information loss and lower performance



RECOMMENDATION

- Begin with the simplest model appropriate for the task
- Increase complexity only if necessary to meet performance requirements
 - ~~ Typically reduces interpretability and requires model-agnostic explanations
- Choose the simplest model with sufficient accuracy ~~ Occam's razor



Bike Data, 4-fold CV

Model	RMSE	R^2
LM	800.15	0.83
Tree	981.83	0.74
Random Forest	653.25	0.88
Boosting (tuned)	638.42	0.89