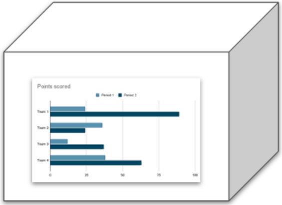
# Interpretable Machine Learning

# Inherently Interpretable Models - Motivation



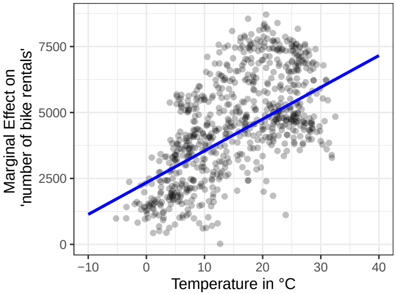

**Learning goals**

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

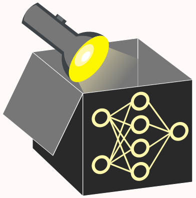
# MOTIVATION

- 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 boosting / component-wise boosting
  - ...

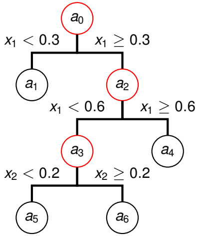


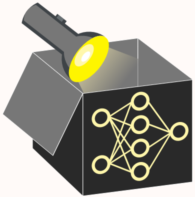

$\rightsquigarrow$ LM provides straightforward interpretation



# ADVANTAGES

- Inherently interpretable - some additional model-agnostic interpretation methods not required
  $\rightsquigarrow$ Eliminates a source of error

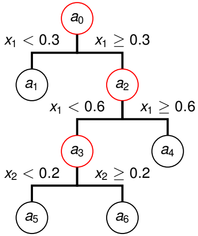


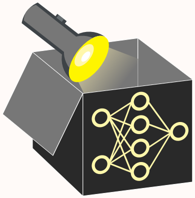



# ADVANTAGES

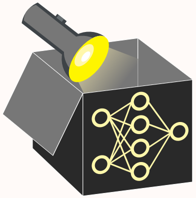
- Inherently interpretable - some additional model-agnostic interpretation methods not required
  $\rightsquigarrow$ Eliminates a source of error

- Interpretable models often simple
  $\rightsquigarrow$ training time is fairly small



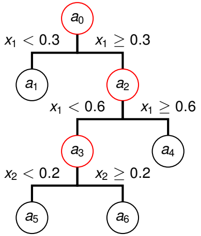
# ADVANTAGES

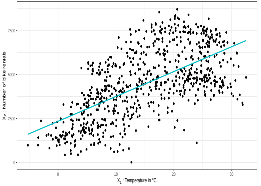


- Inherently interpretable - some additional model-agnostic interpretation methods not required
  $\rightsquigarrow$ Eliminates a source of error

- Interpretable models often simple
  $\rightsquigarrow$ training time is fairly small

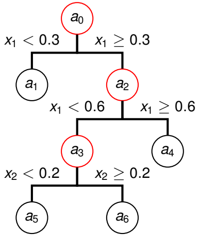
- Some interpretable models fulfill the monotonicity constraint
  $\rightsquigarrow$ Larger feature values always lead to higher (or smaller) outcomes (e.g., GLMs)

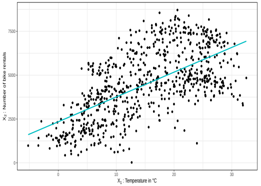


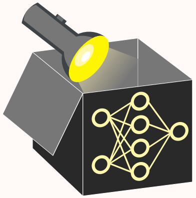



# ADVANTAGES

- Inherently interpretable - some additional model-agnostic interpretation methods not required
  $\rightsquigarrow$ Eliminates a source of error
- Interpretable models often simple
  $\rightsquigarrow$ training time is fairly small
- Some interpretable models fulfill the monotonicity constraint
  $\rightsquigarrow$ Larger feature values always lead to higher (or smaller) outcomes (e.g., GLMs)
- Many people are familiar with traditional interpretable models
  $\rightsquigarrow$ Increases trust, facilitates communication of results

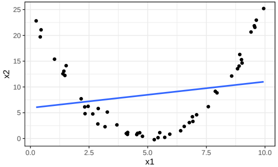


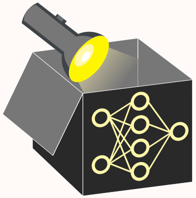



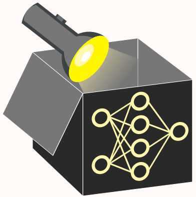
# DISADVANTAGES

- Often certain assumptions about data / model structure required
  $\rightsquigarrow$ If assumptions are wrong, models may perform bad

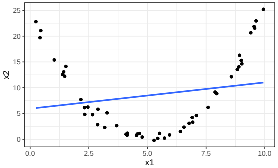


# DISADVANTAGES

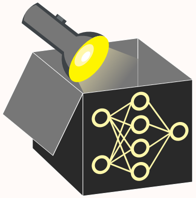


- Often certain assumptions about data / model structure required
  $\rightsquigarrow$ If assumptions are wrong, models may perform bad

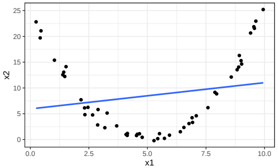



- Interpretable models may also be difficult to interpret, e.g.:
  - Linear model with hundreds of features and interactions
  - Decision trees with huge tree depth

# DISADVANTAGES



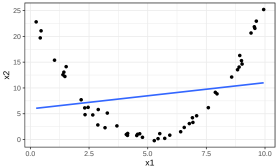
- Often certain assumptions about data / model structure required
  $\rightsquigarrow$ If assumptions are wrong, models may perform bad

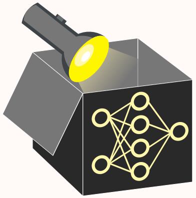



- Interpretable models may also be difficult to interpret, e.g.:
  - Linear model with hundreds of features and interactions
  - Decision trees with huge tree depth
- Often not able to automatically model complex relationships due to limited model flexibility
  e.g., high-order main or interaction effects need to be specified manually in a LM

# DISADVANTAGES

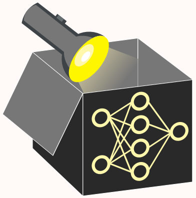
- Often certain assumptions about data / model structure required
  $\rightsquigarrow$ If assumptions are wrong, models may perform bad



- Interpretable models may also be difficult to interpret, e.g.:
  - Linear model with hundreds of features and interactions
  - Decision trees with huge tree depth
- Often not able to automatically model complex relationships due to limited model flexibility
  e.g., high-order main or interaction effects need to be specified manually in a LM
- Inherently interpretable models do not provide all types of explanations
  $\rightsquigarrow$ Methods providing other types of explanations still useful (e.g., counterfactual explanations)

# FURTHER COMMENTS

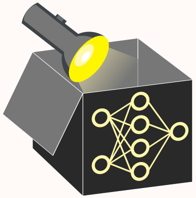


- Some argue that one should always use interpretable models in the first place
  ▸ Rudin 2019
  - . . . and not try to explain uninterpretable models post-hoc
  - Can sometimes work out by spending enough time and energy on data pre-processing

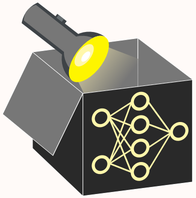
# FURTHER COMMENTS

- Some argue that one should always use interpretable models in the first place
  ▸ Rudin 2019
  - . . . and not try to explain uninterpretable models post-hoc
  - Can sometimes work out by spending enough time and energy on data pre-processing

$\rightsquigarrow$ Drawback: Hard to achieve for data for which end-to-end learning is crucial (e.g., hard to extract good features for image / text data $\rightsquigarrow$ information loss = bad performance)



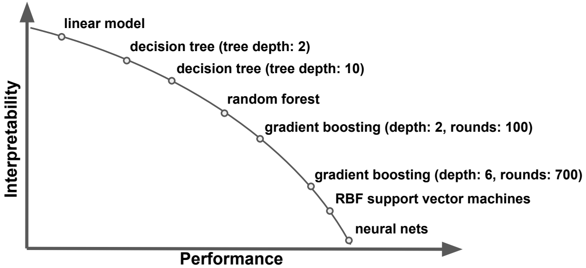
# FURTHER COMMENTS



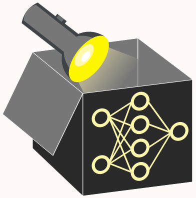
- Some argue that one should always use interpretable models in the first place ▸ Rudin 2019
  - . . . and not try to explain uninterpretable models post-hoc
  - Can sometimes work out by spending enough time and energy on data pre-processing

$\rightsquigarrow$ Drawback: Hard to achieve for data for which end-to-end learning is crucial (e.g., hard to extract good features for image / text data $\rightsquigarrow$ information loss = bad performance)

- Often there is a trade-off between interpretability and model performance



# RECOMMENDATION



- Start with most simple model that makes sense for application at hand
- Gradually increase complexity if performance is insufficient
  $\rightsquigarrow$ will usually lower interpretability and require additional interpretation methods
- Choose the most simple, sufficient model (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 |