



Making models more interpretable



Challenges to interpretability

Challenges related to:



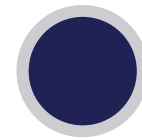
Input data

Correlation
Bias



Model

Complexity
Inscrutability
Performance



Us

Bias



Data and features

Difficult for models to be transparent if the training data is not.

- ✓ Avoid overly complex feature engineering, i.e., use of autoencoders, PCA, polynomial combinations, complex encodings.



Data and features

- ✓ Use features that make sense for a given domain.
- ✓ Avoid features that discriminate based on age, ethnicity, gender, etc.
- ✓ Use monotonic features when possible → it makes it easier to understand.



Data and features

- ✓ Use data that accurately represent the population you are going to serve with your model.

Feature selection

Simpler models are easier to understand.

- ✓ Use less features, remove noisy or non-relevant features.
 - Lasso.
 - Other feature selection methods.



Models

- ✓ Choose intrinsically explainable models whenever possible.
- ✓ Add constraints to limit their complexity, like monotonic constraints, limits to the depth and the feature interactions.
 - ✓ More understandable to humans
 - ✓ Easier to troubleshoot and to use in practice
- ✓ Optimize both the fit (performance metric) and the interpretability of the model.



Black box models

- ✓ Use more than one IML to interpret their results.



THANK YOU

www.trainindata.com