Black Boxes and Biased Implementations

Nick Eubank

Black Boxes

What does it mean to be a Black Box algorithm?

Black Boxes

What does it mean to be a Black Box algorithm?

- Not public (e.g. proprietary)

Black Boxes

What does it mean to be a Black Box algorithm?

- Not public (e.g. proprietary)
- Not transparent (e.g. neural network)

Black Boxes

What does it mean to be a Black Box algorithm?

- Not public (e.g. proprietary)
- Not transparent (e.g. neural network)
- Both

Black Boxes Hide...

Proprietary:

Black Boxes Hide...

Proprietary:

- Can’t evaluate accuracy issues

Black Boxes Hide...

Proprietary:

- Can't evaluate accuracy issues
- Can't evaluate coding / data errors

Black Boxes Hide...

Proprietary:

- Can't evaluate accuracy issues
- Can't evaluate coding / data errors
- Can't tell if it's using factors we think are unjust
  Has mother been arrested?

Black Boxes Hide...

Non-Transparent Models (SVMs, Neural Networks, etc.):

Even if public, these models have no constant marginal effects!
No $\frac{\partial Y}{\partial X}$!

e.g.:

- High School $\rightarrow$ College:
  Decreases probability of recidivism if 24
- High School $\rightarrow$ College:
  Increases probability of recidivism if 25

Black Boxes

Factor Importance:

- Contributions to predictive power, *not*

Black Boxes

Factor Importance:

- Contributions to predictive power, *not*
- Clear report of how factor impacts outcomes

Black Boxes

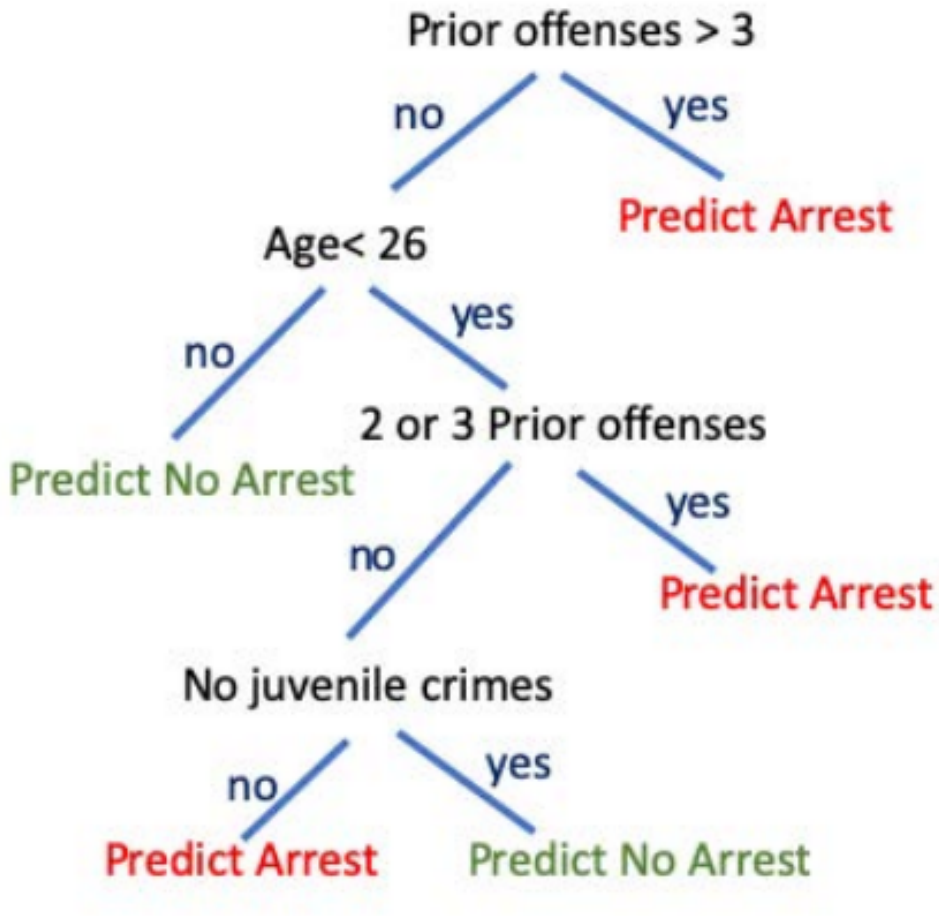
Factor Importance:

- Contributions to predictive power, *not*
- Clear report of how factor impacts outcomes

Can get *averages* of the data you have, which is fine for advertising...
But if errors send people to prison / prevent from getting medical treatment, not ok!

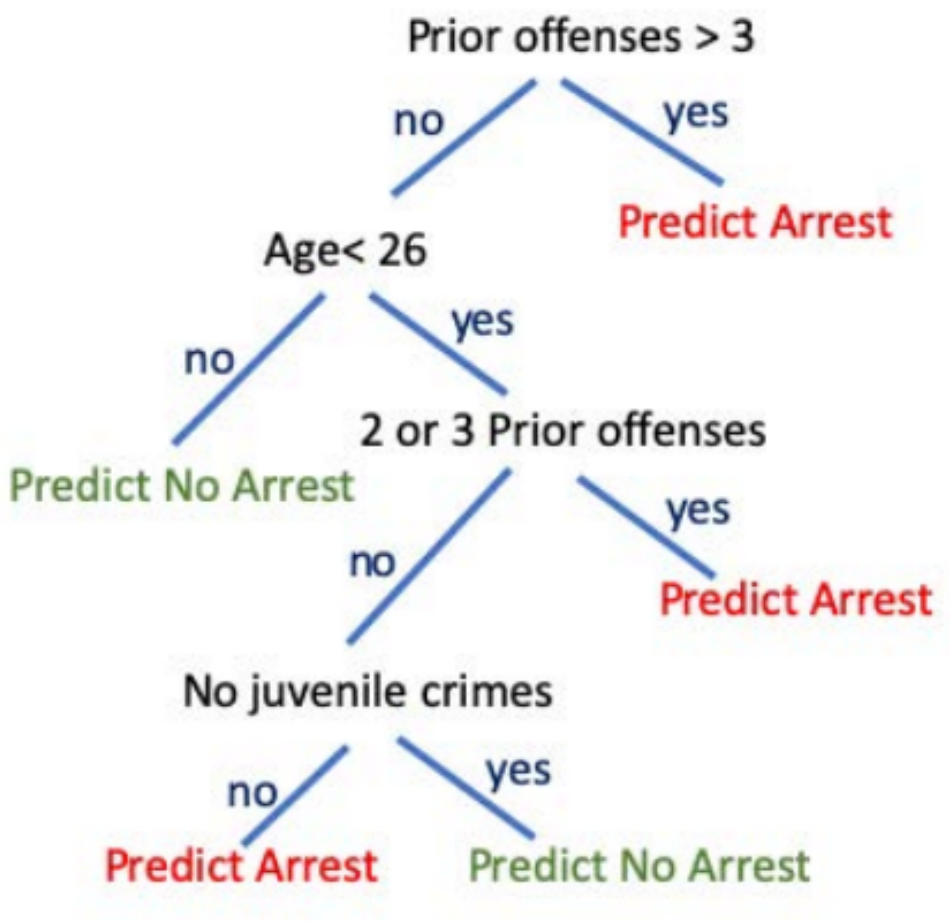
Interpretable Models

Interpretable Models



An interpretable decision tree to predict whether an individual will be arrested in the future. Hu et al. NeurIPS 2019

Interpretable Models



An interpretable decision tree to predict whether an individual will be arrested in the future. Hu et al. NeurIPS 2019

Interpretable Models

Interpretable Models

1. Any cEEG pattern with Frequency **2** Hz	1 point	. . .
2. **E**pileptiform Discharges	1 point	+ . . .
3. Patterns include [**L**PD, LRDA, BIPD]	1 point	+ . . .
4. **P**atterns Superimposed with Fast or Sharp Activity	1 point	+ . . .
5. Prior **S**eizure	1 point	+ . . .
6. **B**rief Rhythmic Discharges	**2** points	+ . . .
	Score	= . . .

Score	0	1	2	3	4	5	6+
Risk	<5%	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

2HELPS2B score for predicting seizures in ICU patients (Struck et al 2017), constructed by the RiskSLIM ML algorithm (Ustun & R 2019).
The factors and point scores were chosen (by an algorithm)

Interpretable Models

1. Any cEEG pattern with Frequency **2** Hz	1 point	. . .
2. **E**pileptiform Discharges	1 point	+ . . .
3. Patterns include [**L**PD, LRDA, BIPD]	1 point	+ . . .
4. **P**atterns Superimposed with Fast or Sharp Activity	1 point	+ . . .
5. Prior **S**eizure	1 point	+ . . .
6. **B**rief Rhythmic Discharges	**2** points	+ . . .
	Score	= . . .

Score	0	1	2	3	4	5	6+
Risk	<5%	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

2HELPS2B score for predicting seizures in ICU patients (Struck et al 2017), constructed by the RiskSLIM ML algorithm (Ustun & R 2019).
The factors and point scores were chosen (by an algorithm)

Don’t preclude bias, but make models easier to audit!