
H2O.ai

Irrelevant Data

- Not all features are related to the target.

- Once identified, remove from the analysis. Do not rely on algorithms to remove irrelevant features. Have doubts? Simulate random numeric and categorical features and find how many of them appear to be important.

Real Data

Solution

Not all data have value

- Noise can be mistaken as signal by machine learning algorithms.

Missing Data

Types of Missing Data

- Unavailable: Valid for the observation, but not available in the data set.
- Removed: Observation quality threshold may have not been reached, and data removed.
- Not applicable: measurement does not apply to the particular observation (e.g. number of tires on a boat observation)

What to Do

- Ignore entire observation.
- Create an binary variable for each predictor to indicate whether the data was missing or not.
- Segment model based on data availability.
- Estimate missing values (Generalized Low Rank Models)
- Use alternative algorithm: decision trees accept missing values; linear models typically do not.

Irrelevant Data

Real Data	• Not all features are related to the target.
Not all data have value	• Noise can be mistaken as signal by machine learning algorithms.
Solution	• Once identified, remove from the analysis. Do not rely on algorithms to remove irrelevant features. Have doubts? Simulate random numeric and categorical features and find how many of them appear to be important.