
H2O.ai

# 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)

- 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.

# Types of Missing Data

# What to Do

# H2O Simple Missing Data Imputation

```r
h2o.impute(data, column = 0, method = c("mean", "median", "mode"),
           combine_method = c("interpolate", "average", "lo", "hi"),
           by = NULL, groupByFrame = NULL, values = NULL)
```

```python
h2o_frame.impute(column = -1, method = u'mean',
                 combine_method = u'interpolate',
                 by = None, group_by_frame = None, values = None)
```

# 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.