
H2O.ai

Extreme Values & Outliers

- Extreme values can exist in response or predictors
- Valid: rare, extreme events
- Invalid: erroneous measurements

- Remove observations that represent outliers.
- Apply a transformation to reduce impact: e.g. log skewed data, create categorical bins, impose cap on low/high values (winsorize).
- Choose a more robust loss function: e.g. MAE vs MSE.
- Ask questions: Understand whether the values are valid or invalid, to make the most appropriate choice.

Types of Outliers

What to Do

What Can Happen

- Extreme values can have a disproportionate effect.
- MSE will focus on handling extreme observations more to reduce squared error.
- Boosting will spend considerable modeling effort fitting these observations.

H2O Extreme Value Handling

```r
h2o_frame["log_x"] <- h2o.log1p(h2o_frame["x"])

h2o_frame["cat_x"] <- h2o.cut(h2o_frame["x"], breaks)

h2o_frame["winz_x"] <- h2o.ifelse(h2o_frame["x"] < low, low, h2o_frame["x"])
h2o_frame["winz_x"] <- h2o.ifelse(h2o_frame["winz_x"] > high, high,
                                  h2o_frame["winz_x"])
```

```python
h2o_frame["log_x"] = h2o_frame["x"].log1p()

h2o_frame["cat_x"] = h2o_frame["x"].cut(breaks)

h2o_frame["winz_x"] = h2o.H2OFrame.ifelse(h2o_frame["x"] < low,
                                          low, h2o_frame["x"])
h2o_frame["winz_x"] = h2o.H2OFrame.ifelse(h2o_frame["winz_x"] > high,
                                          high, h2o_frame["winz_x"])
```

Extreme Values & Outliers

Types of Outliers

- Extreme values can exist in response or predictors
- Valid: rare, extreme events
- Invalid: erroneous measurements

What Can Happen

- Extreme values can have a disproportionate effect.
- MSE will focus on handling extreme observations more to reduce squared error.
- Boosting will spend considerable modeling effort fitting these observations.

What to Do

- Remove observations that represent outliers.
- Apply a transformation to reduce impact: e.g. log skewed data, create categorical bins, impose cap on low/high values (winsorize).
- Choose a more robust loss function: e.g. MAE vs MSE.
- Ask questions: Understand whether the values are valid or invalid, to make the most appropriate choice.