
H2O.ai

H2O Extreme Value Handling

```
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"])
```

```
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"])
```

Low Frequency Categories

Real Data

- Most real world datasets contain categorical data.

Too Many Categories

- Problems can arise if you have **too many categories**:
  - Computational complexity during estimation
  - Infrequent categories can lead to overfitting

Solutions

- Use knowledge about hierarchical data to collapse categories.
- Use Cross-Validated Mean Target Encoding.
- Use Cross-Validated Weight of Evidence Encoding when modeling binary outcome.
- Use H2O's `categorical_encoding` and `nbins_cat` decision tree tuning arguments

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"])
```