
H2O.ai

Target Class Imbalance H2O Parameters

- **`balance_classes`**: balance training data class counts via over/under-sampling.
- **`class_sampling_factors`**: desired over/under-sampling ratios per class (in lexicographic order). If not specified, sampling factors will be automatically.
- **`max_after_balance_size`**: maximum relative size of the training data after balancing class counts.
- **`sample_rate_per_class`**: variable row sampling rate per class.

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.

H2O.ai

Target Class Imbalance H2O Parameters

- **`balance_classes`**: balance training data class counts via over/under-sampling.
- **`class_sampling_factors`**: desired over/under-sampling ratios per class (in lexicographic order). If not specified, sampling factors will be automatically.
- **`max_after_balance_size`**: maximum relative size of the training data after balancing class counts.
- **`sample_rate_per_class`**: variable row sampling rate per class.

H2O.ai