
H2O.ai

Target Class Imbalance

- You can **balance** the training set using sampling.

- H2O has a `balance_classes` argument that can be used to do this properly & automatically.
- You can manually **up-sample** (or **down-sample**) your minority (or majority) class(es) set either by duplicating (or sub-sampling) rows, or by using row weights.

Artificial Balance

Solutions

Potential Pitfalls

- Don’t balance the test set! The test set should represent the true data distribution.
- The same goes for a hold-out validation set and cross-validation sets.
- Cross-validation will probably require custom coding.

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.

Target Class Imbalance

Artificial Balance

- You can **balance** the training set using sampling.

Potential Pitfalls

- Don't balance the test set! The test set should represent the true data distribution.
- The same goes for a hold-out validation set and cross-validation sets.
- Cross-validation will probably require custom coding.

Solutions

- H2O has a `balance_classes` argument that can be used to do this properly & automatically.
- You can manually **up-sample** (or **down-sample**) your minority (or majority) class(es) set either by duplicating (or sub-sampling) rows, or by using row weights.