
H2O.ai

Target Class Imbalance

- A dataset is said to be **imbalanced** when categorical responses occur at widely varying rates.
- Standard optimizations by machine learning algorithms may favor majority classes.
- Rule of thumb for binary response: If the minority class makes < 10% of the data, this can cause issues.

- Advertising — Probability that someone clicks on ad is very low… very very low.
- Healthcare & Medicine — Certain diseases or adverse medical conditions are rare.
- Fraud Detection — Insurance or credit fraud is rare.

Imbalanced Response Variable

Common Examples Across Industries

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.

Target Class Imbalance

Imbalanced Response Variable

- A dataset is said to be **imbalanced** when categorical responses occur at widely varying rates.
- Standard optimizations by machine learning algorithms may favor majority classes.
- Rule of thumb for binary response: If the minority class makes < 10% of the data, this can cause issues.

Common Examples Across Industries

- Advertising – Probability that someone clicks on ad is very low… very very low.
- Healthcare & Medicine – Certain diseases or adverse medical conditions are rare.
- Fraud Detection – Insurance or credit fraud is rare.