

Summary/Review

Modeling Unbalanced Classes

Classification algorithms are built to optimize accuracy, which makes it challenging to create a model when there is not a balance across the number of observations of different classes. Common methods to approach balancing the classes are:

- Downsampling or removing observations from the most common class
- Upsampling or duplicating observations from the rarest class or classes
- A mix of downsampling and upsampling

Modeling Approaches for Unbalanced Classes

Specific algorithms to upsample and downsample are:

- Stratified sampling
- Random oversampling
- Synthetic oversampling, the main two approaches being Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic sampling (ADASYN)
- Cluster Centroids implementations like NearMiss, Tomek Links, and Nearest Neighbors