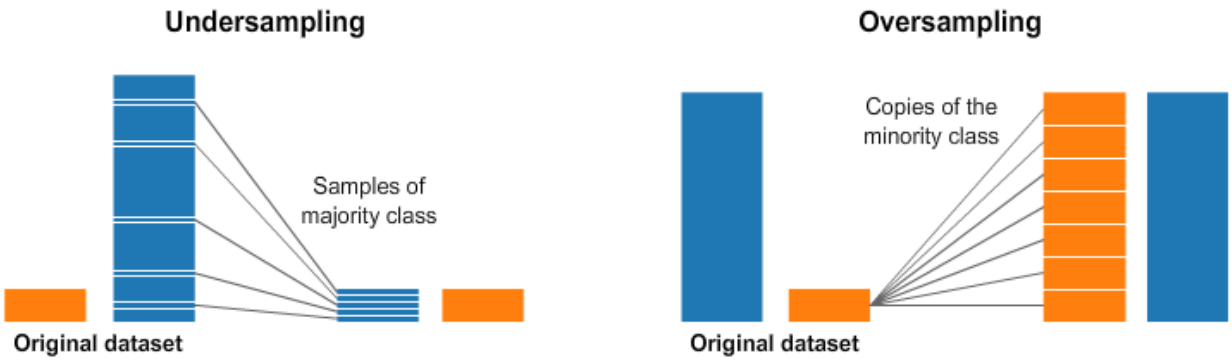


It is common to find datasets with imbalanced class distribution. Building classification model with imbalanced dataset will cause the under-represented class been overlooked or even ignored. Under sampling and oversampling are techniques used to combat the issue of unbalanced classes in a dataset.

**Over sampling: increase sample size of minority class (could introduce noise and redundancy).**
Oversampling is replication the events of minority class. Potential problem could be for this method is overfitting for noisy data, because noisy data will be replicate. To avoid overfitting the procedure of randomized oversampling is proposed with cleaning noisy data.

**Under sampling: decrease sample size of majority class (could remove representative samples).**
In case of undersampling, random samples of majority class (BG) can be taken. Potential problem is that some of useful BG instances may not be chosen for training and classifier will not be optimal. Reduction of majority class without losing performance of classification can be used.