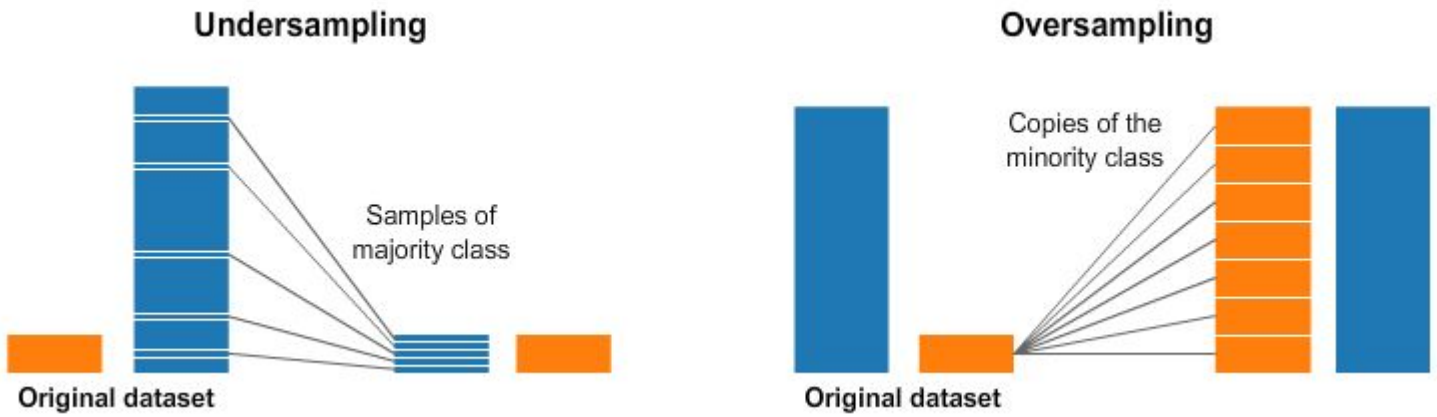
1. We can handle the imbalanced dataset cases to minimize the Type II errors by balancing the class representations

2. To balance the classes we can –
   a. Decrease the frequency of the majority class
   b. Increase the frequency of the minority class   OR



3. Decreasing the frequency of majority class is done using random under sampling. For e.g.

   a. Total observations – 1000

   b. Fraud - 020

   c. Non-fraud - 980

   d. Event rate of interest - 2%

   e. Take 10% of non-fraud cases randomly - 98

   f. Club with the fraud cases – 118 sample size

   g. Modified event rate - 20 / 118 = 17%

4. Random oversampling is used to increase the frequency of minority class. This is done by replicating them in order to increase their representation. For e.g.
    a. Total observations – 1000
    b. Fraud - 020
    c. Non-fraud - 980
    d. Event rate of interest - 2%
    e. Replicate a % of fraud cases n times e.g. 10 cases 20 times
    f. Sample size changes from 1000 to 1200
    g. Modified event rate - 220/1200 = 18%

5. The simplest implementation of over-sampling is to duplicate random records from the minority class, which can cause overfitting.

6. In under-sampling, the simplest technique involves removing random records from the majority class, which can cause loss of information.

Imblearn Techniques

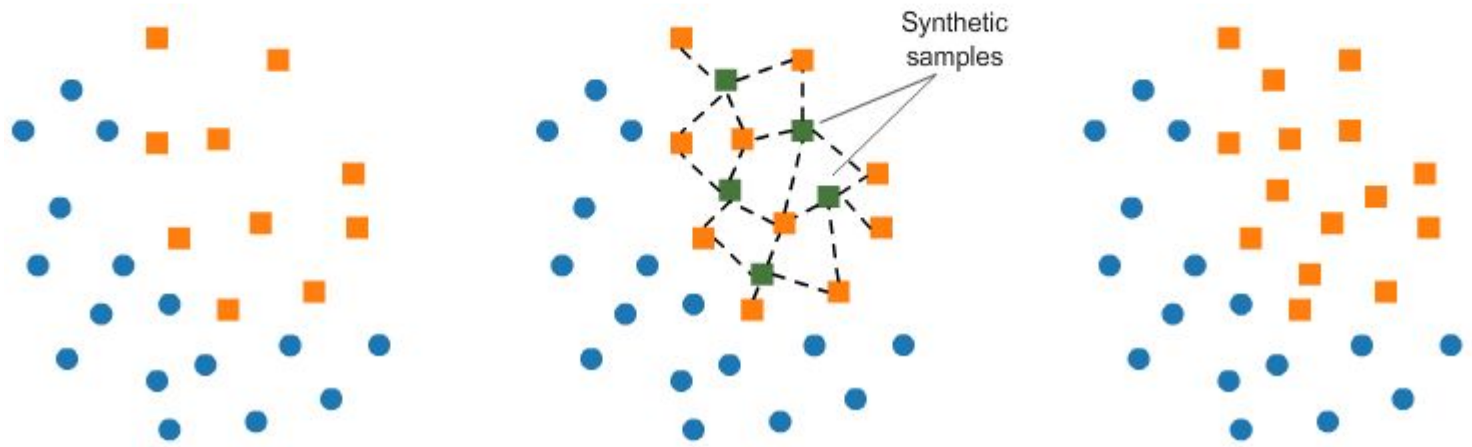
1. Python imbalanced-learn module – provides more sophisticated resampling techniques

2. For example, we can cluster the records of the majority class, and do the under-sampling by removing records from each cluster, thus seeking to preserve information.

3. In over-sampling, instead of creating exact copies of the minority class records, we can introduce small variations into those copies, creating more diverse synthetic samples.

Imblearn Techniques

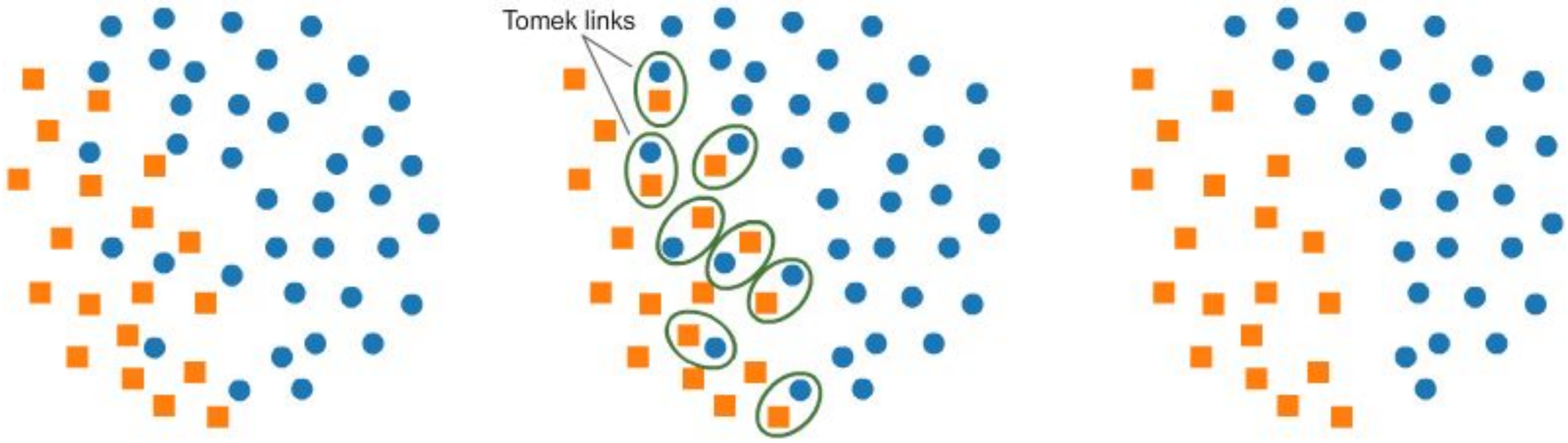
4. SMOTE (Synthetic Minority Oversampling TEchnique)
   a. consists of synthesizing elements for the minority class, based on those that already exist.
   b. It works randomly picking a point from the minority class and computing the k-nearest neighbors for this point.
   c. Synthetic points are added between the chosen point and its neighbors.



Imblearn Techniques

5. Tomek links T-Link
    a. Tomek links are pairs of very close instances, but of opposite classes.

    b. Removing the instances of the majority class of each pair increases the space between the two classes, facilitating the classification process.



Imblearn Techniques

6. Cluster centroid based under sampling -

   a. Method that under samples the majority class by replacing a cluster of majority samples by the cluster centroid of a KMeans algorithm.

   b. This algorithm keeps N majority samples by fitting the KMeans algorithm with N cluster to the majority class and using the coordinates of the N cluster centroids as the new majority samples.