Imbalanced data

When is data imbalanced?

45-55 split is not imbalanced however, a 90-10 split is imbalanced. Imbalance lies somewhere between these extremes.

Types of Imbalance

1. Between-class

When one class has large number of observations and other class has very few

2. Within-class

When both classes have abundant observations, but observations for one class are concentrated in one-region only, whereas the other class is spread out over the entire range.

Other issues with imbalanced data – class overlapping

The class more represented in the overlap regions tends to be better classified this is because the algorithm is able to get a more informed picture of the data distribution of the majority class

The class less represented in such regions tends to be better classified by local methods. If we take k-NN as an example, as the value of k increases, it becomes increasingly global and increasingly local. It can be shown that performance for low values of k has better performance on the minority dataset, and lower performance at high values of k.