Sampling

Deal with an imbalanced dataset

When your dataset does not represent all classes of data equally, the model might **overfit to the class that's represented more in your dataset and become oblivious to the existence of the minority class**. It might even give you a good accuracy but fail miserably in real life.

A widely adopted technique for dealing with highly unbalanced datasets is called **resampling**. **Resampling is done after the data is split into training, test and validation sets.** Resampling is done only on the training set or the performance measures could get skewed. Resampling can be of two types: **Over-sampling and Under-sampling**.

Over-sampling and Under-sampling

Under sampling involves removing samples from the majority class and over-sampling involves adding more examples from the minority class. The simplest implementation of **over-sampling** is to **duplicate random records** from the minority class, **which can cause overfitting**. In **under-sampling**, the simplest technique involves removing random records from the majority class, **which can cause loss of information**.

There are two oversampling functions: **Random Oversampling** and **Synthetic Minority Oversampling Technique (SMOTE). Random Oversampling repeats the existing samples randomly** and **SMOTE creates new samples through simulation** based on the distribution of the data that belongs to a class. SMOTE is similar to interpolation.