Imbalanced vs balanced dataset

- Classification is one of the most common machine learning problems. One of the common issues found in datasets that are used for classification is imbalanced classes issue.
- Imbalanced dataset means when we have unequal distribution of classes.
- For example, let's consider a dataset called anomaly detection on AIOPS data(server data). Mostly we have 2 classes like anomaly and not anomaly.
- whenever you see this kind of data we used to have a lot of non-anomolous data and very few anomaly data.
- This leaves us with something like 50:1 ratio between the anomaly and non-anomaly classes.

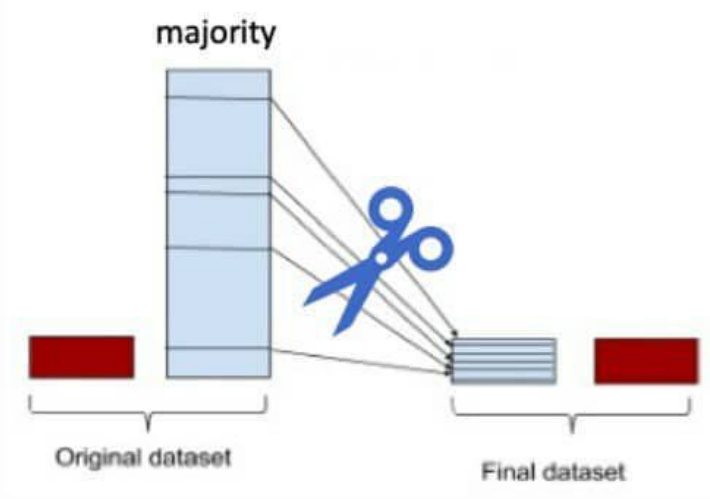
Imbalanced vs balanced dataset

- If we train a binary classification model without fixing this problem, the model will be completely biased and mostly it always predicts the majority class.
- The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class.
- Let's see how to fix this in the next series of posts.

Fixing Imbalanced dataset

Resampling (Undersampling)

- Undersampling is the process where you randomly delete some of the observations from the majority class in order to match the numbers with the minority class.

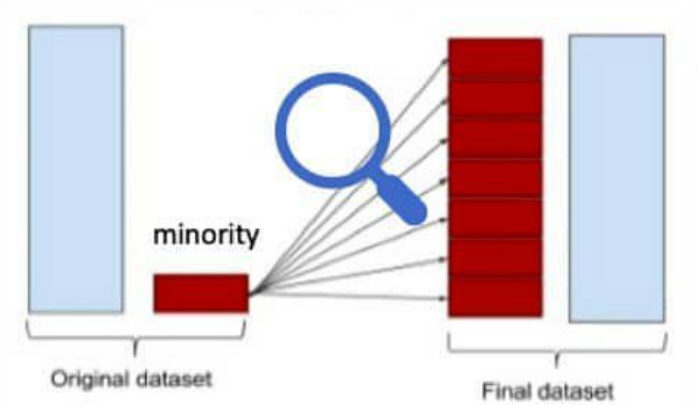


- It can help improve run time and storage problems by reducing the number of training data samples when the training data set is huge.
- It can discard potentially useful information which could be important for building rule classifiers.
- The sample chosen by random undersampling may be a biased sample. And it will not be an accurate representation of the population. Thereby, resulting in inaccurate results with the actual test data set.

Fixing Imbalanced dataset

Resampling (Random Over-Sampling)

- Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample.



- Unlike undersampling, this method leads to no information loss.
- Outperforms undersampling.
- It increases the likelihood of overfitting since it replicates the minority class events.

Fixing Imbalanced dataset

Resampling (SMOTE)

- In the last technique, we using to replicate the same data multiple times which leads to overfitting.
- To overcome this method we use SMOTE (Synthetic Minority Over-sampling Technique).
- Here we take the subset of minority class data and then we create new synthetic similar instances.
- And there will be no loss of useful information.
- While generating synthetic examples SMOTE does not take into consideration neighbouring examples from other classes. This can result in an increase in the overlapping of classes and can introduce additional noise
- SMOTE is not very effective for high dimensional data