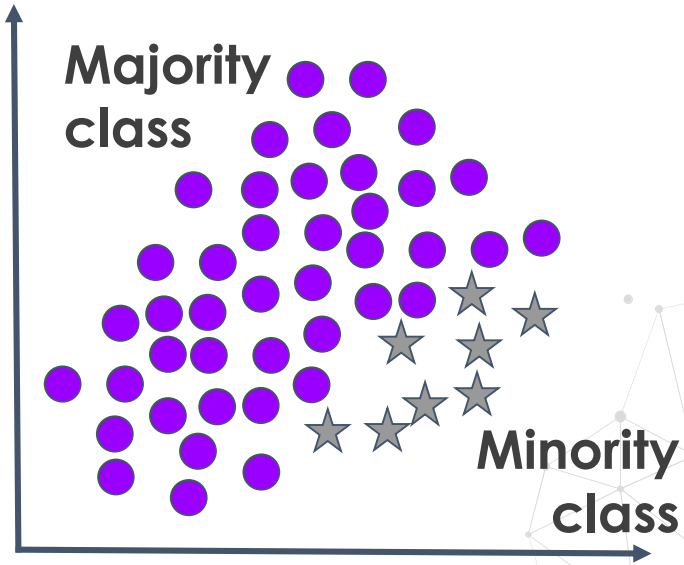
Train In Data

The Nature of Imbalanced Classes

Imbalanced class distribution

- **Class distribution**: the proportion of instances belonging to each class.
- **Imbalance ratio** = $\frac{Xminority}{Xmajority}$

In some cases, a ratio as low as 1:35 can be hard for building a good model, while in other cases, 1:10 is tough to deal with.



The nature of imbalanced classes

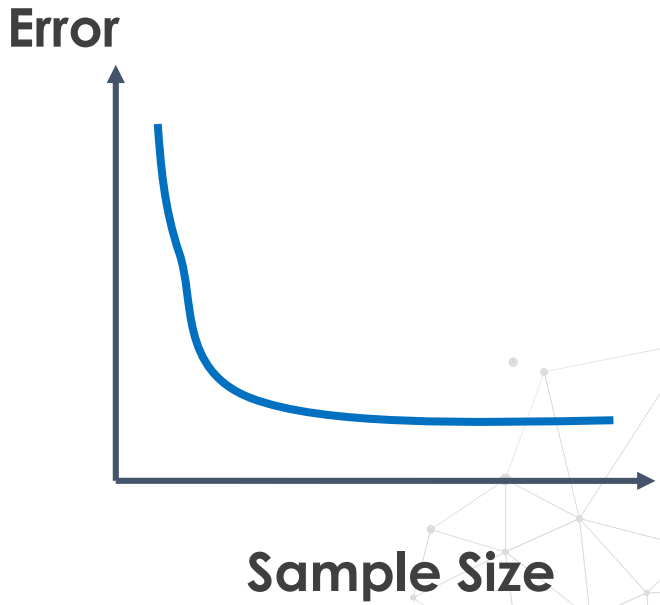
Factors that influence the ability of a classifier to identify rare events

- Small sample size
- Class separability
- Within-class sub-clusters.

Small Sample Size

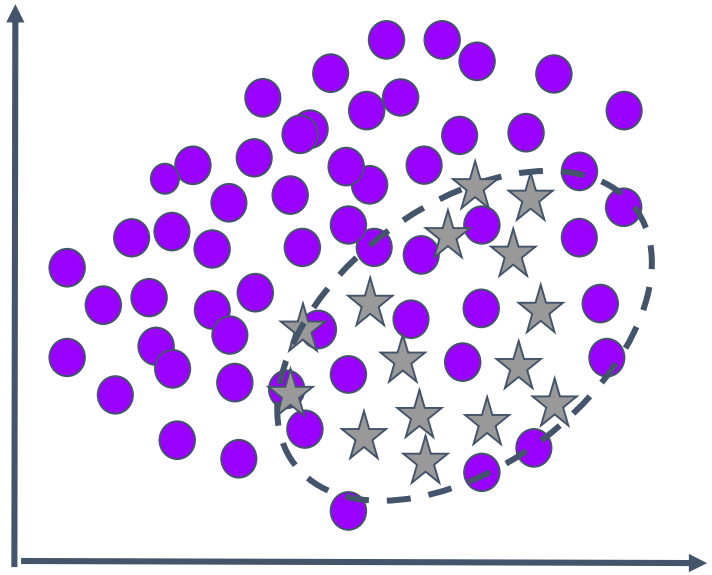
Sample size plays a crucial role in determining the "goodness" of a model.

- If sample size is limited, finding patters inherent to the small class is hard.
- As the data size increases, the error in the prediction decreases.



Imbalanced classes may not be a problem if the data is big enough

Class separability

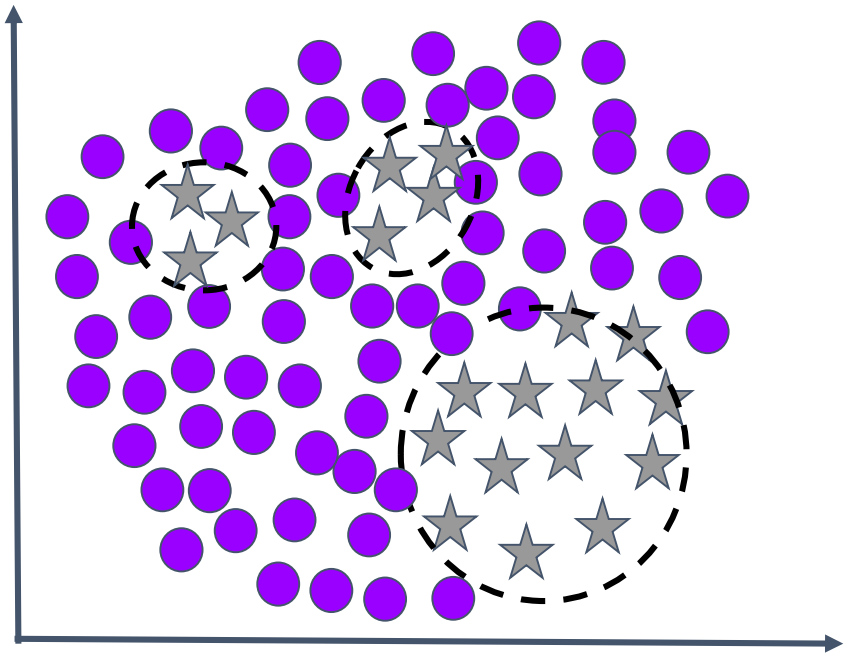


If patterns among classes overlap, it is harder to find rules.

The class imbalance per se may not be a problem, instead the separability makes it harder to finds rules to classify correctly the minority class

Linearly separable domains are not sensitive to any amount of imbalance

Within class sub-clusters



In many classification problems, a single class is composed of various sub-clusters or concepts.

These sub-clusters do not always contain the same number of examples.

This phenomena is referred to as **within-class imbalance**, corresponding to the imbalanced class distribution among subclasses

Within class sub-clusters increases the complexity and makes it harder to find boundaries to separate the classes.

Train In Data

THANK YOU

www.trainindata.com