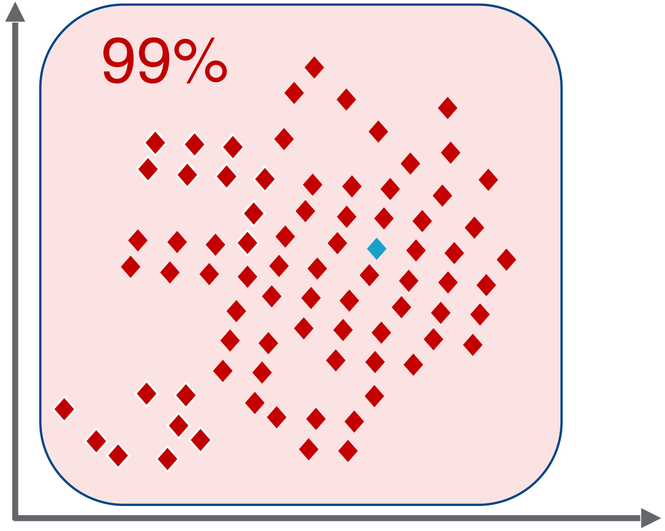
Imbalanced Classes

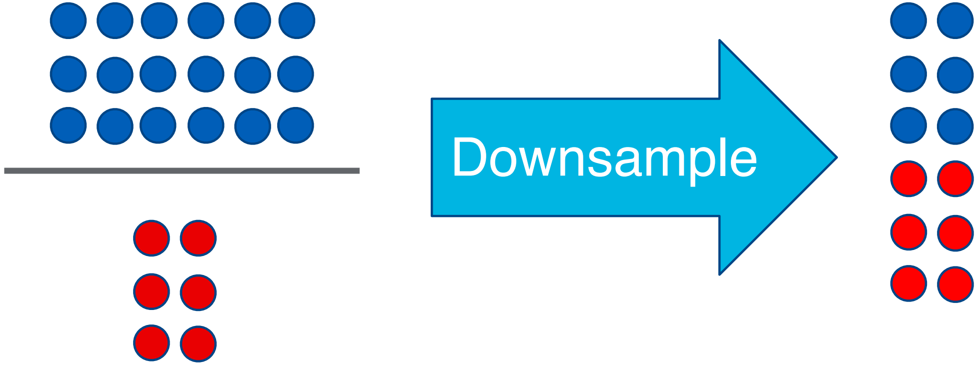
Imbalanced Classes

Classifiers are usually built to optimize accuracy and hence will often perform poorly on imbalanced classes.



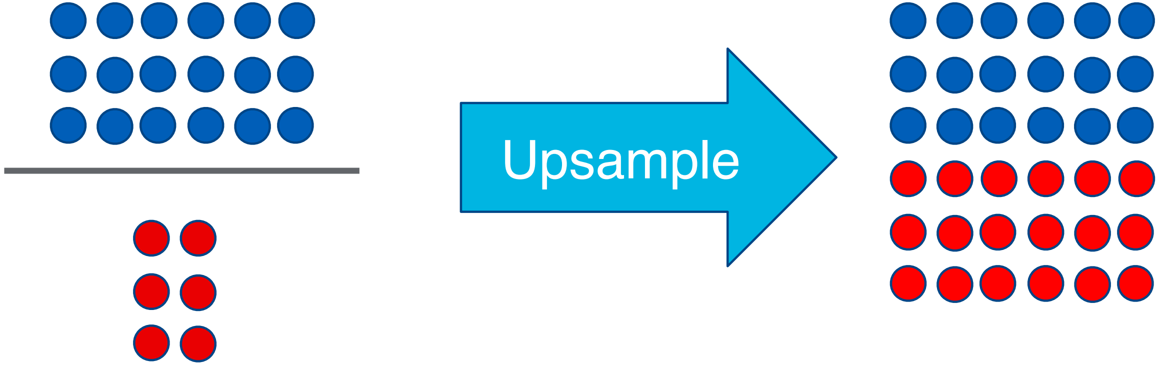
Imbalanced Classes

For imbalanced datasets, we can balance the size of the classes by either downsampling the larger class or upsampling the small one.



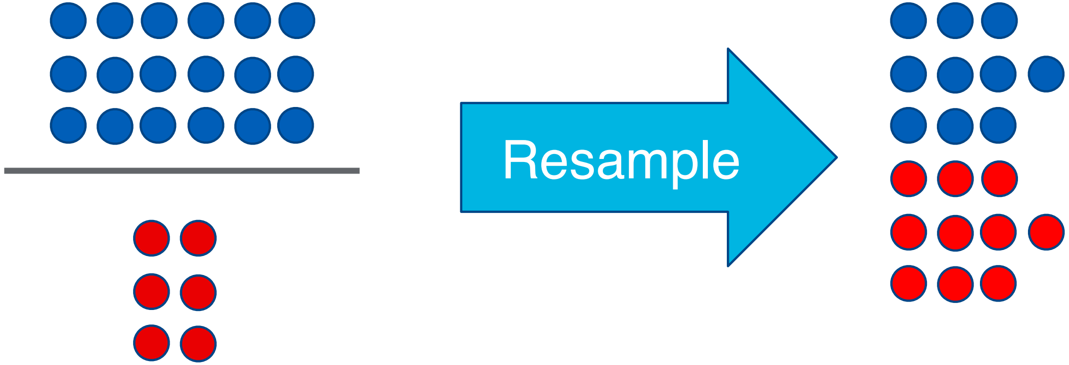
Imbalanced Classes

For imbalanced datasets, we can balance the size of the classes by either downsampling the larger class or upsampling the small one.



Imbalanced Classes

We can also do a mix of the two. In this example 6 < s < 19 for sampling. If we choose s = 10:



Imbalanced Classes

Steps for imbalanced datasets:

- Do a **stratified** test-train split
- Up or down sample the training dataset
- Build models

Imbalanced Classes

With imbalanced classes, the data often isn’t easily separable. We must choose to make sacrifices to one class or the other.

Drunk Driver	Bad Road Condition	Speeding	**Accident**
Y	Y	Y	**N**
Y	Y	Y	**N**
Y	Y	Y	**Y**

Imbalanced Classes

For every minor-class data-point identified as such, we might wrongly label a few major-class points as minor-class.

So as recall goes up, precision will likely go down.

Drunk Driver	Bad Road Condition	Speeding	**Accident**
Y	Y	Y	**N**
Y	Y	Y	**N**
Y	Y	Y	**Y**

Imbalanced Classes

Downsampling adds tremendous importance to the minor class, typically shooting up recall and bringing down precision.

Values like 0.8 recall and 0.15 precision isn’t uncommon.

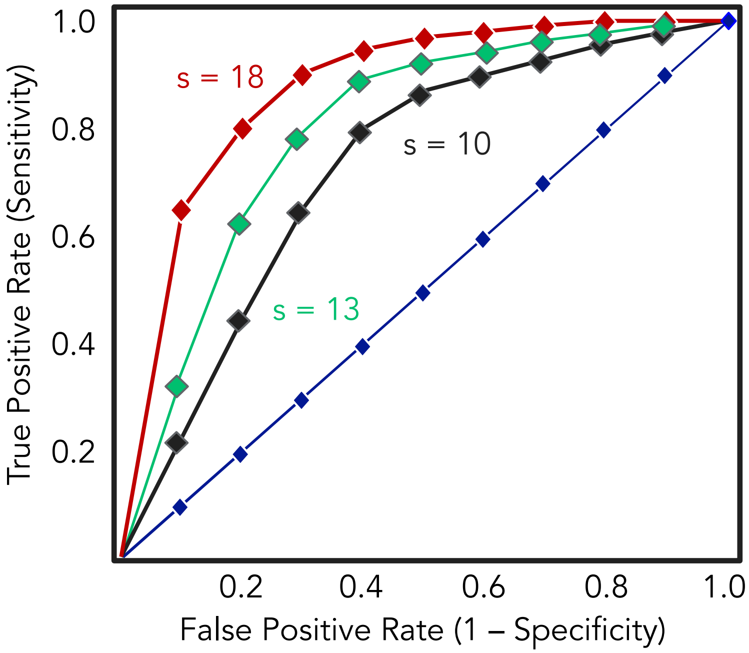
Imbalanced Classes

Upsampling mitigates some of the excessive weight on the minor class. Recall is still typically higher than precision, but the gap is lesser.

Values like 0.7 recall and 0.4 precision isn’t uncommon. And are often considered good results for an imbalanced dataset.

Imbalanced Classes

Cross-validation works for any global model-making choice. Even sampling.



Imbalanced Classes

Every classifier used produces a different model.

Every dataset we use (produced by various sampling, say) produces a different model.

We can choose the best model using any criteria including AUC (area under the curve). Remember each model produces a different ROC curve.

Once a model is chosen. You can walk along the ROC curve and pick any point on it. Each point has different precision/recall values.