Data Analysis and Visualization

CentraleDigitalLab@Nice

Deborah Dore - PhD - ddore@i3s.unice.fr
MARIANNE, I3S, CNRS, INRIA, Université Côte d'Azur

Imbalanced Datasets
Overview

An **Imbalanced dataset** is a dataset where a class significantly outnumbers the other. In real-life example, this imbalance represents most of the cases.

Example:

You are a bank employee responsible for detecting the validity of credit card transactions. To do so, you have a training set of previously observed transactions, each of which was either:

A. *Normal*

B. *Fraudulent*

Most transactions are normal and it is not unlikely that fraudulent is just 0.1% of the total transactions!

Imbalanced Datasets

Challenges for ML models

Machine Learning Models **may favor the majority class**, leading to poor generalization for minority classes.

Challenges:

- Skewed model performance
- Misleading accuracy metrics
- Difficulty in detecting minority class

Imbalanced Datasets

Metrics for Evaluating Imbalanced Datasets

Basics:

True Positives = Classified as Positive, actually Positive

False Positives = Classified as Positive, actually Negative

True Negatives = Classified as Negative, actually Negative

False Negatives = Classified as Negative, actually Positive

		ACTUAL	
		positive	negative
PREDICTED	positive	**True Positive (TP)**	**False Positive (FP)**
	negative	**False Negative (FN)**	**True Negative (TN)**

Imbalanced Datasets

Metrics for Evaluating Imbalanced Datasets

Why traditional metrics fail:

- **Accuracy** can be misleading (e.g., 99% accuracy by predicting only the majority class).
- $$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} = \frac{\text{Correct predictions}}{\text{Total observations}}$$

Imbalanced Datasets

Metrics for Evaluating Imbalanced Datasets

Why traditional metrics fail:

- **Accuracy** can be misleading (e.g., 99% accuracy by predicting only the majority class).

Better metrics:

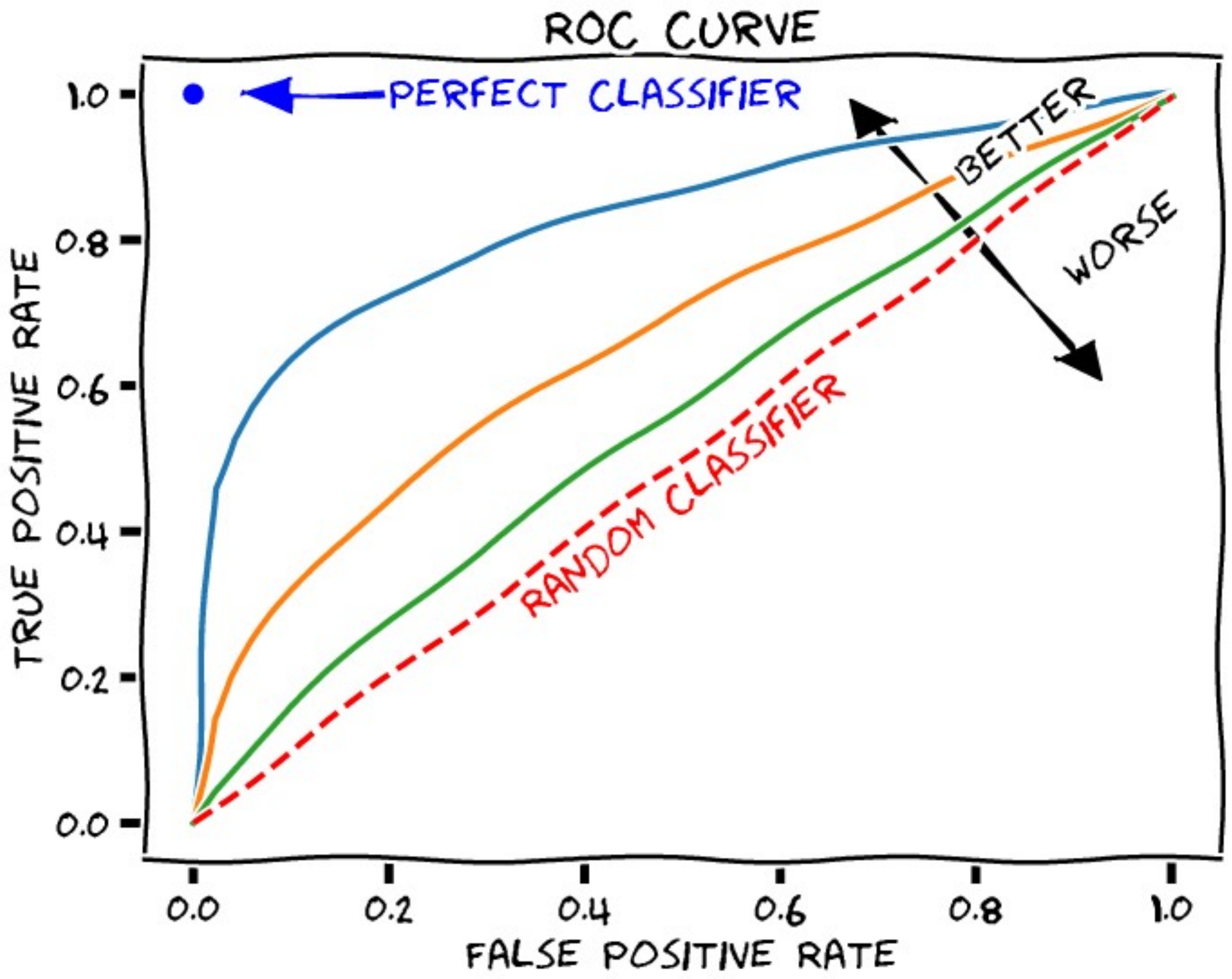
- **Precision** = $\frac{TP}{TP + FP}$
- **Recall** = $\frac{TP}{TP + FN}$
- **ROC-AUC =** Area Under the Receiver Operating Characteristics Curve
- **PR-AUC =** Area Under the Precision Recall Curve

Imbalanced Datasets

Metrics for Evaluating Imbalanced Datasets

ROC-AUC is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability.

In other words, it tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.



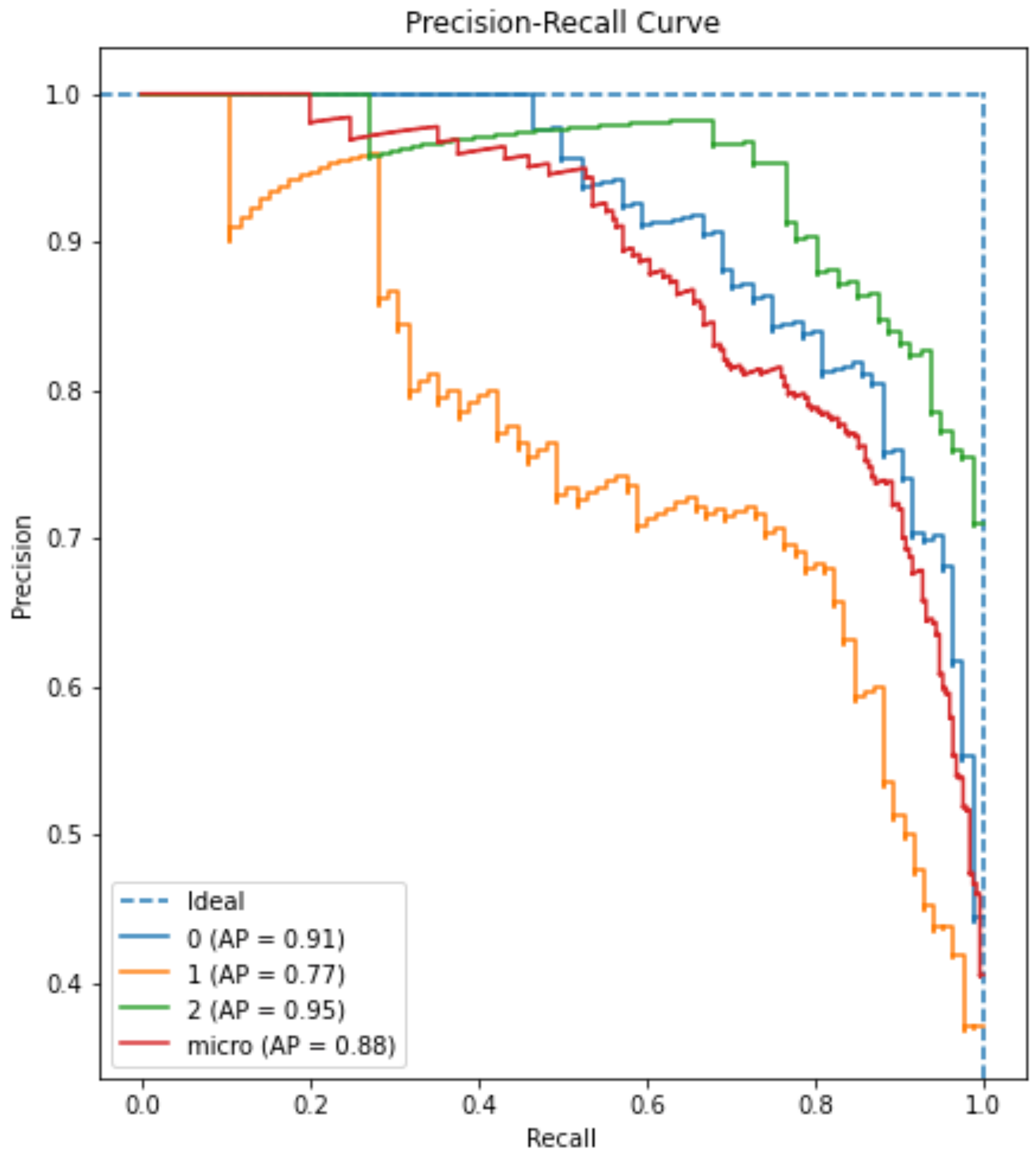
https://commons.wikimedia.org/wiki/File:Roc-draft-xkcd-style.svg

Imbalanced Datasets

Metrics for Evaluating Imbalanced Datasets

PR-AUC represents the area under the Precision-Recall curve, which plots Precision (the proportion of true positives among all positive predictions) against Recall (the proportion of true positives identified correctly) at various threshold settings.

In simpler terms, the PR AUC quantifies how well a model can distinguish between classes, considering both its ability to not mark a negative sample as positive (Precision) and its ability to find all the positive samples (Recall). A higher PR AUC value signifies a better performing model.



https://towardsai.net/p/l/precision-recall-curve

Imbalanced Datasets

Improving Imbalanced Datasets

- Oversampling
- Undersampling
- Hybrid Approaches
- Other Approaches

Imbalanced Datasets

Improving Imbalanced Datasets: Oversampling

Oversampling consists in increasing the size of the minority class by duplicating or synthesizing samples.

Advantages: balances the dataset without discarding majority-class samples.

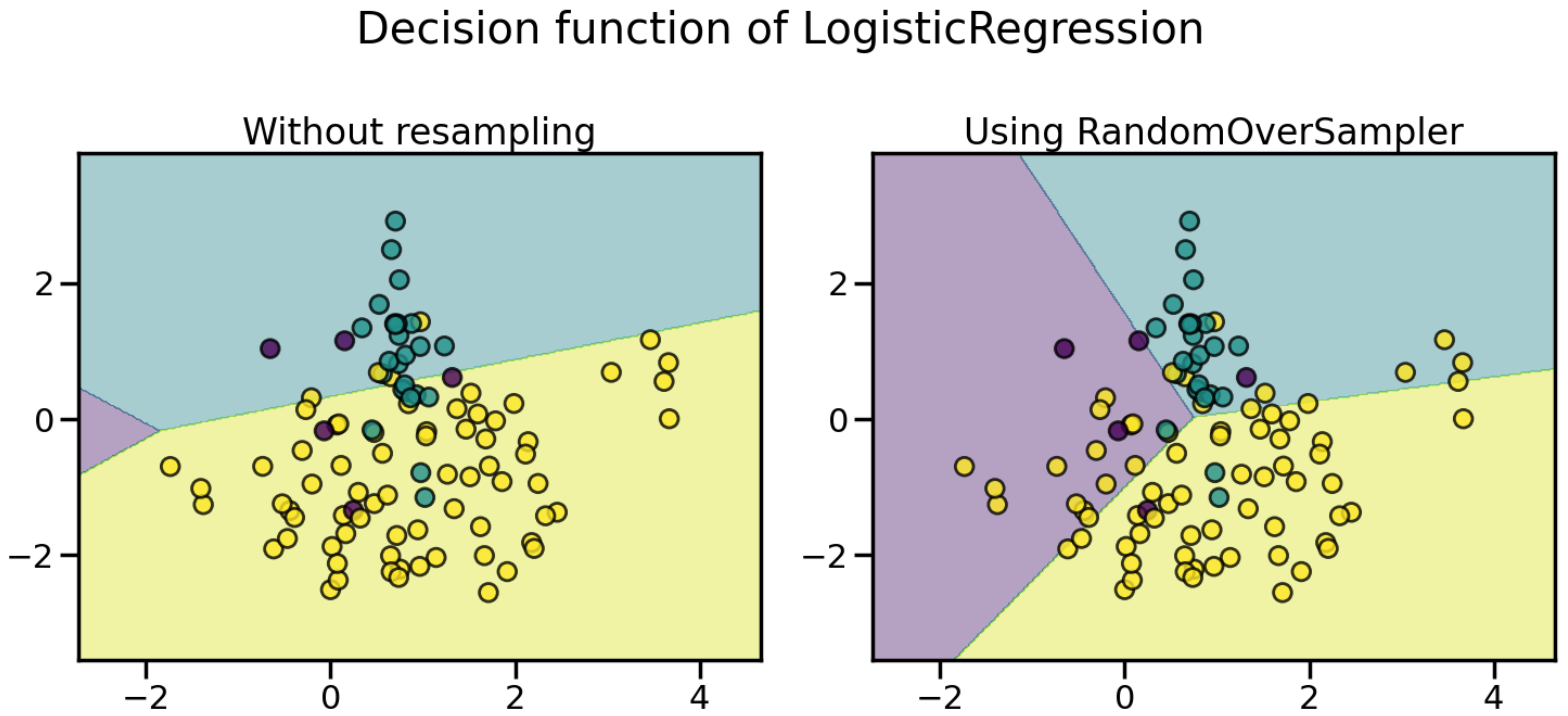
Techniques:

- Random Oversampler
- SMOTE (Synthetic Minority Oversampling Technique)

Imbalanced Datasets

Oversampling → Random Oversampler

Random Oversampler randomly duplicates entities.



https://imbalanced-learn.org/stable/auto_examples/over-sampling/plot_comparison_over_sampling.html#sphx-glr-auto-examples-over-sampling-plot-comparison-over-sampling-py

Imbalanced Datasets

Oversampling → SMOTE

SMOTE (Synthetic Minority Oversampling Technique) creates synthetic samples by interpolating between existing minority-class samples.

Steps:

1. Select k nearest neighbors for a minority-class sample.
2. Generate synthetic points along the line segments joining the sample and its neighbors.

Imbalanced Datasets

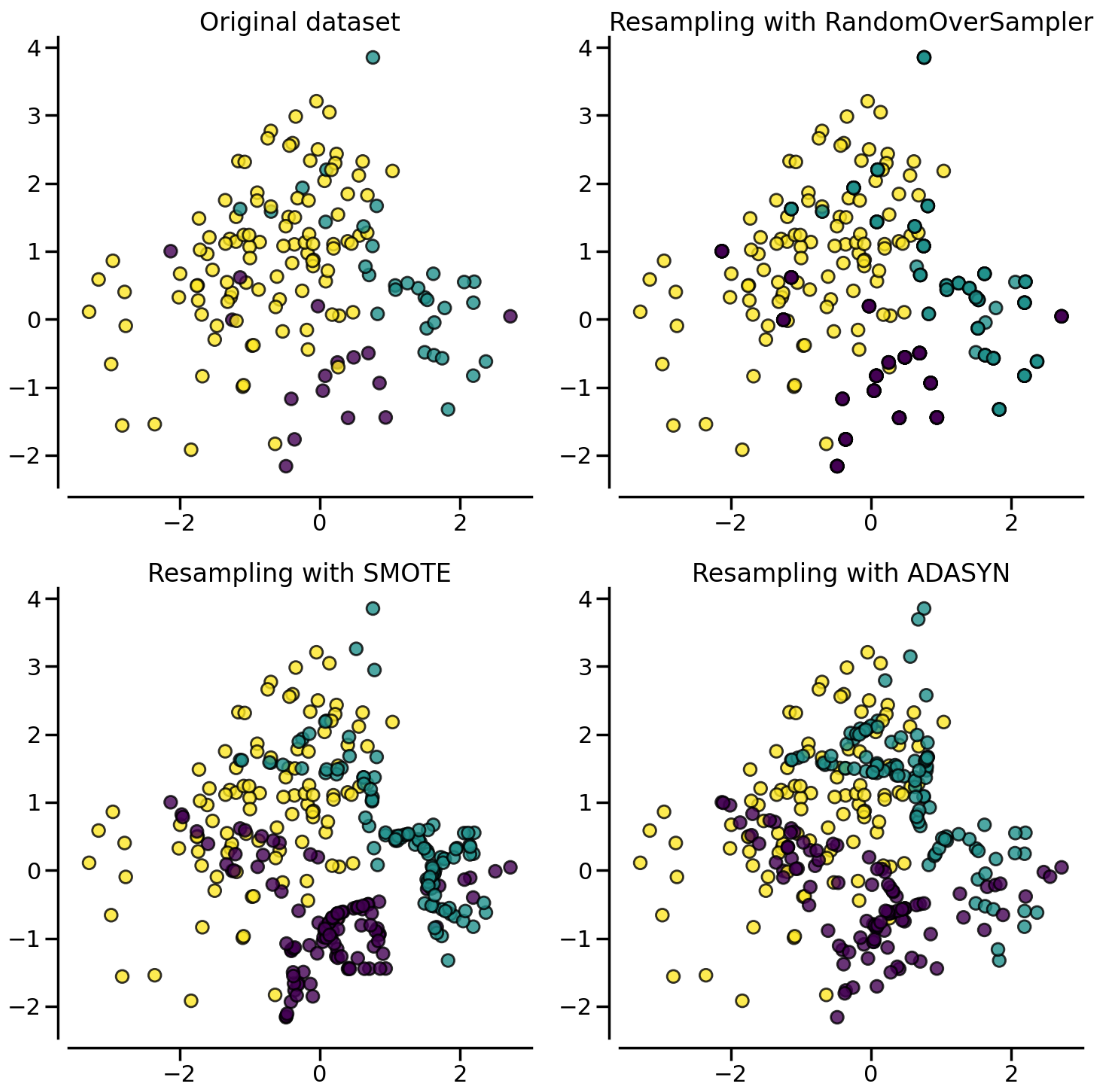
Oversampling → ADASYN

ADASYN (Adaptive Synthetic Sampling) builds on SMOTE but focuses more on difficult-to-learn samples.

Improvement: Weights minority samples based on their difficulty to classify.

Imbalanced Datasets: Oversampling

SMOTE vs ADASYN vs RANDOM OVERSAMPLER



https://imbalanced-learn.org/stable/auto_examples/over-sampling/plot_comparison_over_sampling.html#sphx-glr-auto-examples-over-sampling-plot-comparison-over-sampling-py

Imbalanced Datasets

Improving Imbalanced Datasets: Undersampling

Undersampling consists in reducing the size of the majority class.

Advantages: simplifies the dataset and reduces computational costs

Techniques:

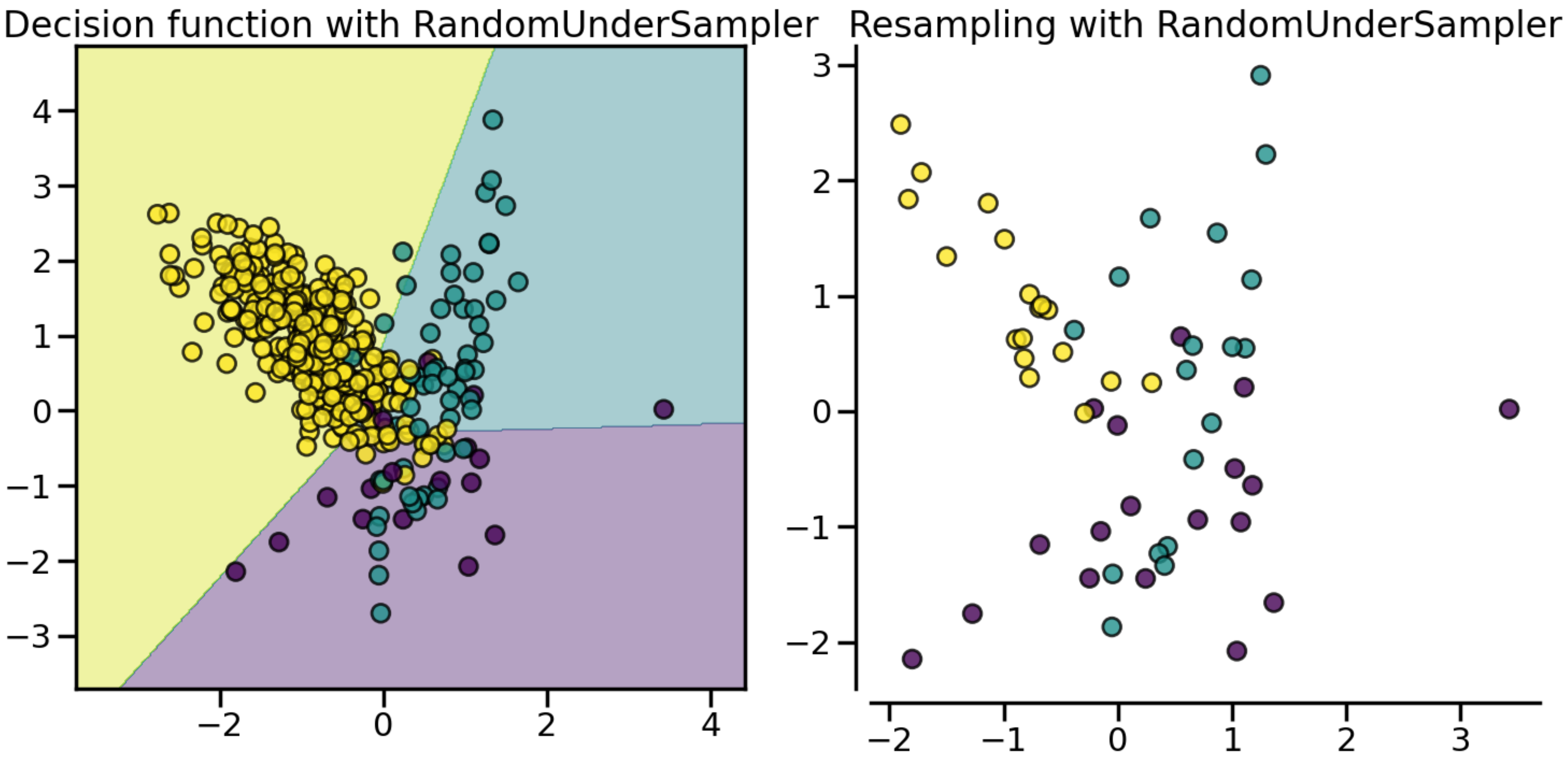
- Random Undersampling
- Cluster Centroids

Challenges: Risk of losing important information

Imbalanced Datasets

Undersampling → Random Undersampling

Random Undersampling randomly deletes majority class entities.

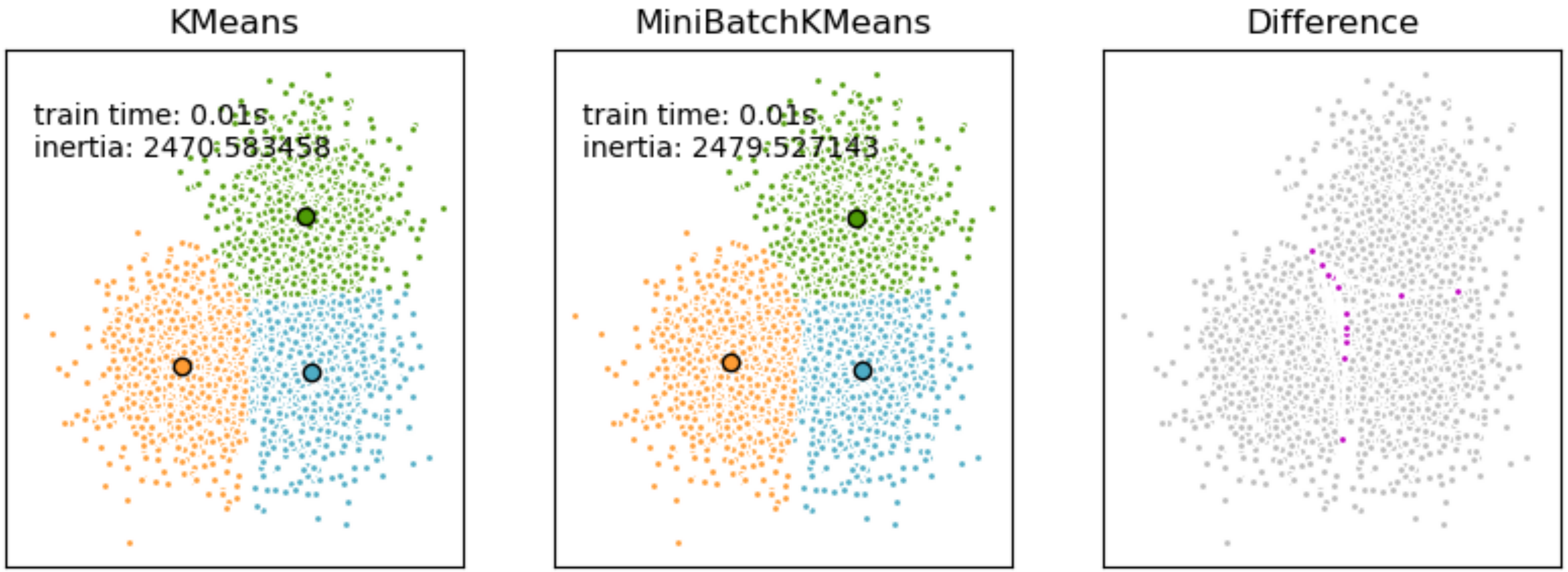


https://imbalanced-learn.org/stable/under_sampling.html

Imbalanced Datasets

Undersampling → Cluster Centroids

Cluster Centroids replaces majority-class samples with centroids of their clusters → Preserves overall data distribution



https://imbalanced-learn.org/stable/under_sampling.html

Data Augmentation

Overview

Data Augmentation is a technique to increase dataset diversity/size by adding new entries created from existing ones.

Example:

$$A = [\text{red},\ 10,\ A]$$

$$B = [\text{blue},\ 5,\ B]$$

<u>New:</u>

$$C = [\text{red},\ 5,\ A]$$

Data Augmentation

Overview

Data Augmentation is a technique to increase dataset diversity/size by adding new entries created from existing ones

Example:

Rotate an image by 15 degrees to create a new sample

Data Augmentation

Overview

Other ideas:

For images: flipping, rotation, scaling, cropping, color adjustments

For text: synonym replacement, back-translation, random insertion or deletion (“The cat sat on the mat.” → “The feline sat on the rug.”)

For tabular data: adding noise to numerical features (es. Gaussian noise), synthetic feature generation

Combining Oversampling and Data Augmentation

Overview

Strategy:

- Use oversampling to balance the dataset and data augmentation to increase diversity

Example:

SMOTE for oversampling, then use data augmentation for increase the minority-class samples

Oversampling & Undersampling

Common Pitfalls

- **Overfitting:** caused by excessive oversampling
- **Loss of information:** caused by excessive undersampling

Models won’t be able to learn if data is inconsistent!

As we say “ Garbage in = Garbage out”!

Useful Links

- https://imbalanced-learn.org/stable/over_sampling.html
- https://imbalanced-learn.org/stable/under_sampling.html
- https://imbalanced-learn.org/stable/combine.html

Notebook
Notebook_Imbalanced_Dataset.ipynb