Imbalanced Data:

Understanding and Addressing Challenges

Praveen Singh

AMS Meeting, New Orleans, January 12, 2025

Agenda

- Introduction
- Imbalanced Problem Domains
- Challenges
- Techniques to Handle Imbalanced Data
- Case Study
- Summary

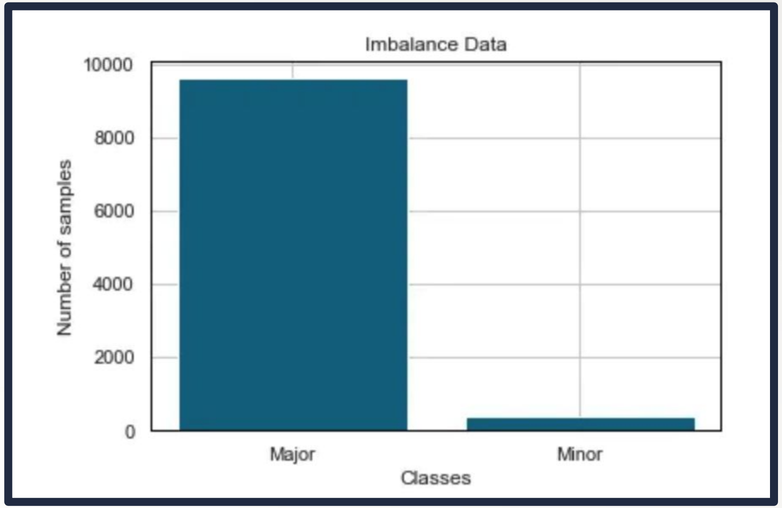
Introduction

Imbalanced Data

Major Class: with more instances
Minor Class: with fewer instances

- Occurs when the number of instances for different classes are significantly out of proportion.

- The minority classes with fewer instances usually contain the <u>essential</u> information.



Imbalanced Data

- ❖ Most ML algorithms work best when the number of instances of each classes are roughly equal. When the number of instances of one class far exceeds the other, problems arise.

- ❖ Many typical classifiers may generate unsatisfactory results due to concentration on global accuracy while ignoring the identification performance for minority classes.

Imbalance Domains

Imbalanced Problem Domains

There are cases where we expect data to be imbalanced:

Fraud Detection

Credit card transactions, each of which are either:

- ❖ Normal
- ❖ Fraudulent, account for less than 0.1% of the total transactions!

Medical Diagnosis

"Imagine you have 10,000 lung X-Ray images and only 100 of them are diagnosed with Pneumonia." These images are classified into:

- ❖ Healthy
- ❖ Not Healthy

Spam Detection

Another typical example of imbalanced data is encountered in e-mail classification problem where emails are classified into:

- ❖ Ham (relevant)
- ❖ Spam

Rare Events

Imbalance is also common in Weather Data:

- ❖ Severe storm events vs. normal weather conditions
- ❖ Tornado vs. non-tornado days

Imbalance in Weather Data may lead to poor predictions.

Challenges

Challenges with Imbalanced Data

- **Bias toward majority class:** Models tend to predict the majority class more frequently.
- **Poor generalization:** Imbalanced data makes it difficult for models to learn from the minority population.
- **Inaccurate evaluation:** Standard metrics like accuracy can be misleading due to the majority class.
- Increased complexity and training time
- Rare event detection

Techniques to Handle Imbalanced Data

Imbalanced Data Techniques

Resampling Methods

-Undersampling
-Oversampling

Algorithmic Approaches

-Adjusting Class Weights:
Decision Tree, SVM
-Cost-sensitive Learning:
Cost-sensitive loss function

Ensemble Methods

-Boosting Techniques:
Gradient Boosting
-Bragging Techniques:
Random Forest

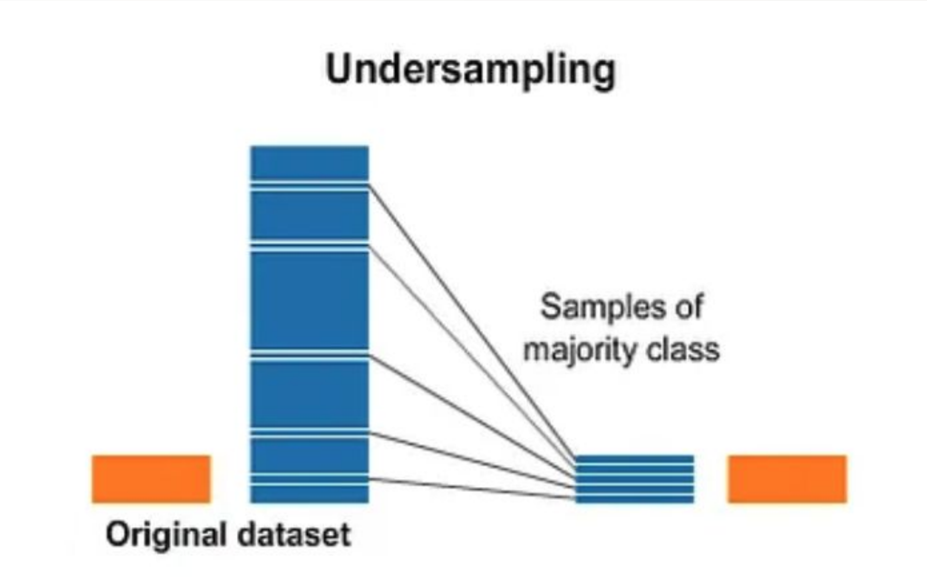
Evaluation Metrics Adjustment

- **Accuracy** is often NOT an appropriate evaluation metric for imbalanced data problems.
- **Precision** and **Recall** capture different characteristics of the classifier.
- **Area Under the Curve (AUC)** and **F1** can be used as a single metric to compare algorithm variations.

Undersampling

Undersampling means to get all of the classes to the same amount as the minority class or the one with the least amount of rows.

By removing some of the majority class instances so it has less effect on the machine learning algorithm.



Undersampling

Pros:

- Easy to implement
- Training becomes much more efficient (smaller training set)
- For some domains, can work very well

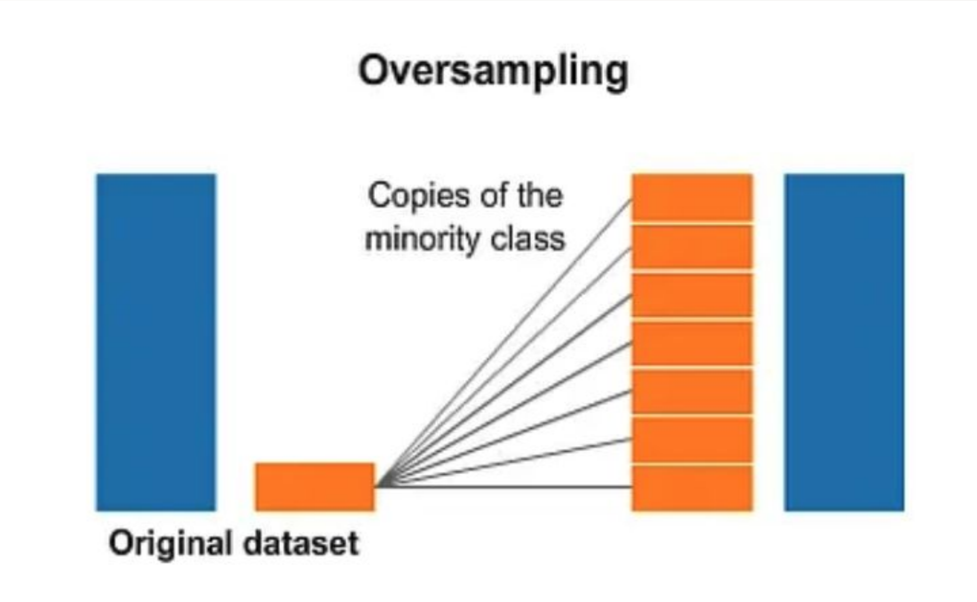
Cons:

- Throwing away a lot of data / information

Oversampling

By adding more of the minority class instances so it has more effect on the machine learning algorithm.

With oversampling, instances can (and do) appear multiple times.

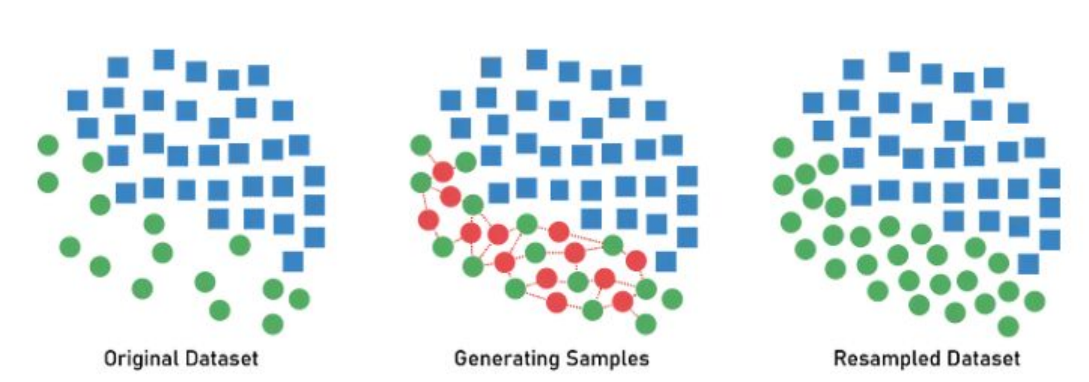


Synthetic Minority Oversampling Technique

- SMOTE is another algorithm to oversample smaller classes.

- The main idea behind SMOTE is that generated instances should be constructed from available observations, but should not be identical.

- SMOTE variants: Adaptive Synthetic Sampling (ADASYN), BorderlineSMOTE, SVM SMOTE.



Oversampling

Pros:

- ❖ Easy to implement
- ❖ Utilize most of the training data
- ❖ Tends to perform well in a broader set of circumstances than subsampling

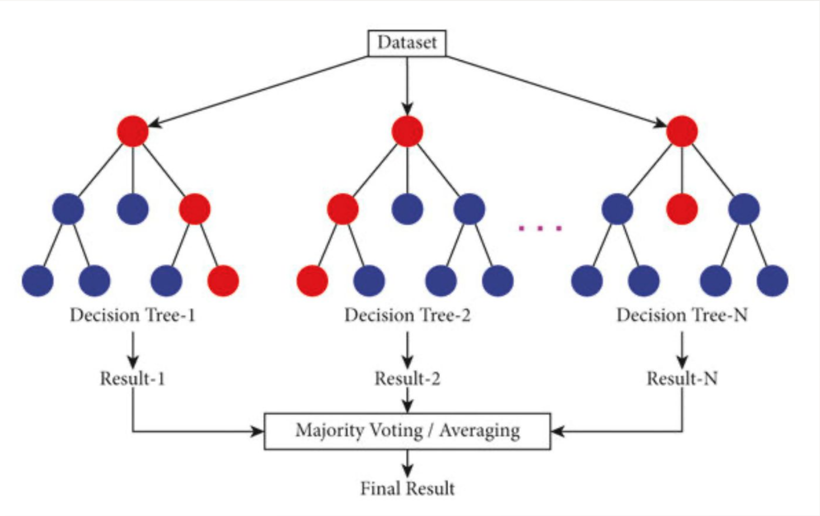
Cons:

- ❖ Computationally expensive

Ensemble Learning

Random Forest

RF is a machine learning method that leverages the power of multiple decision trees and randomness to improve model performance.



Random Forest

Pros:

- Robust to overfitting
- Ability to handle large datasets
- Class imbalance-specific adjustments

Cons:

- Computationally expensive
- Harder to interpret, due to the complexity of having multiple trees

Case Study

Case Study:

- **Dataset Description**
  - A synthetic imbalance dataset
- **Methodology**
  - Apply -
    - Resampling Methods
    - Ensemble Methods - Random Forest
- **Results**
  - Comparison of model performance before and after addressing imbalance.

Tools and Libraries

- **Tools**
  - Anaconda
  - Google Colab
- **Python Libraries**
  - Scikit-learn for machine learning and model evaluation
  - Imbalanced-learn for specific techniques (SMOTE, RandomUnderSampler)
- **Visualization Libraries**
  - Matplotlib and Seaborn for plotting class distributions and evaluation metrics

Summary

Take Home Message

- ❖ Imbalanced data presents a significant challenge in machine learning.
- ❖ Various techniques, like resampling etc can mitigate its impact. The choice of technique depends on the specific problem and it's essential to understand the dataset's characteristics.
- ❖ By using appropriate techniques and metrics, we can build models that are more sensitive to the minority class and, therefore, more useful in real-world applications.

Thank You!