The Class Imbalance Problem

Rachit Chhaya
DA-IICT

Outline

- Class Imbalance Problem
- Why is it a problem?
- How to Handle Class Imbalance?

Takeaways

After the session participants can:

- Understand the Class Imbalance Problem
- Understand the Implications of having class imbalance
- Understand some techniques used to handle class imbalance
- Apply some of these techniques using python (after hands on)

The Class Imbalance Problem

- Number of observations having one class label is much less than than the number of observations with other class label.
- Examples: electricity pilferage, fraudulent transactions in banks, identification of rare diseases, etc.

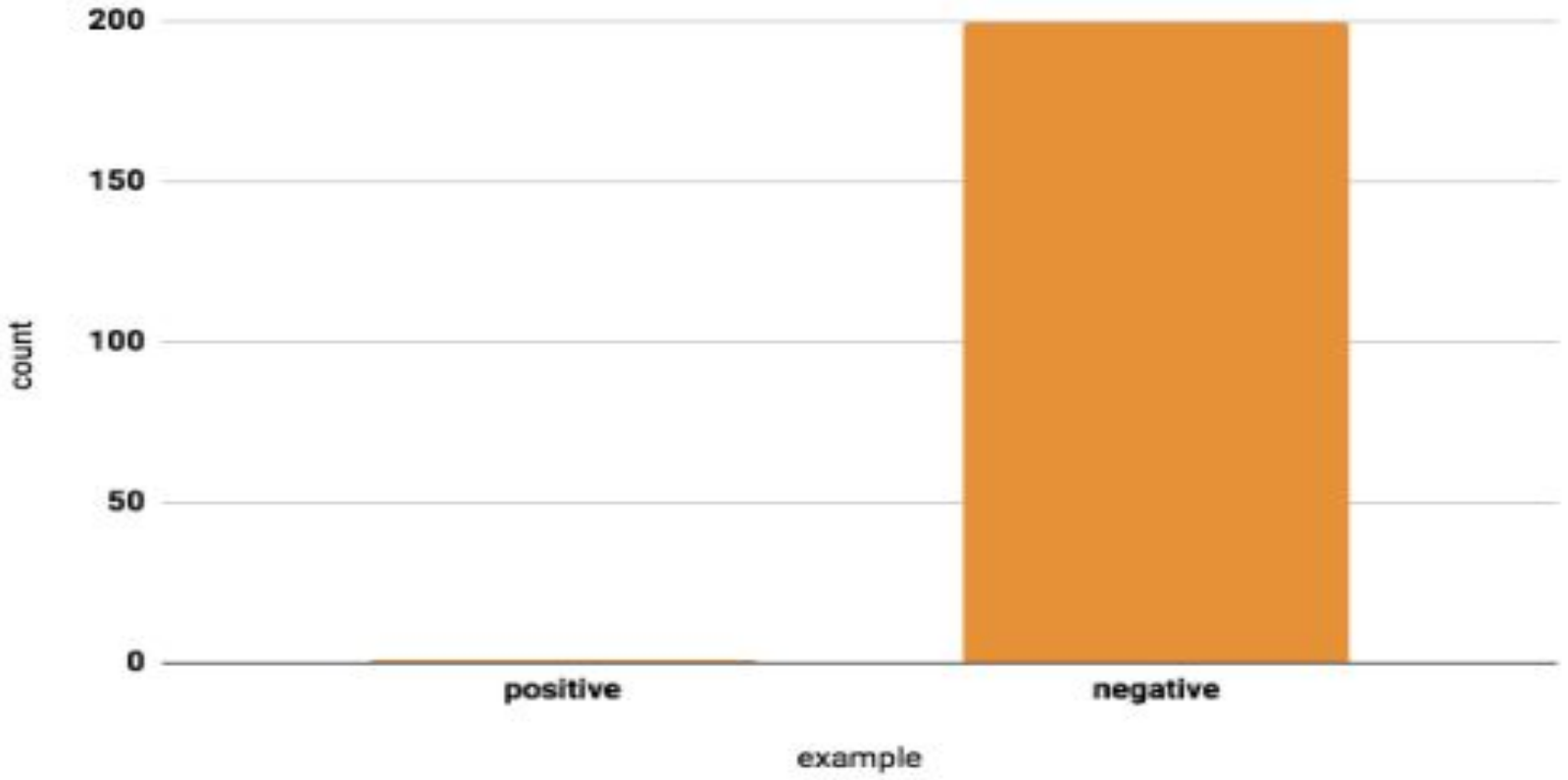

Image From : [Imbalanced Data | Machine Learning | Google Developers]

The Issue

- Conventional classifiers designed to optimize accuracy
- Biases performance towards the majority class
- More pronounced the imbalance, more pronounced is the issue
- Accuracy may not be a good performance measure for classifier when data is imbalanced

What to do??

- Want to use our conventional classifier
- Imbalance will create bias
- Solution : Use some preprocessing/ post-processing of our data to still be able to use our conventional classifiers

Solutions

Random Undersampling

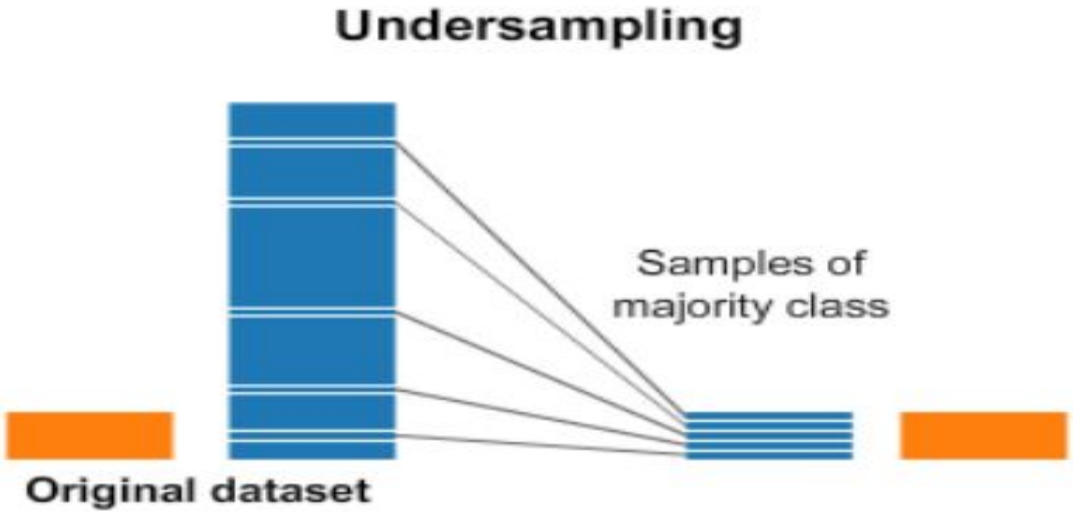

Image Taken From: Important Techniques to Handle Imbalanced Data in Machine Learning… – Towards AI

Advantages:

- Simple Implementation
- Efficient

Disadvantages:

- Throws away lots of data
- Is not representative of test data

Random Oversampling

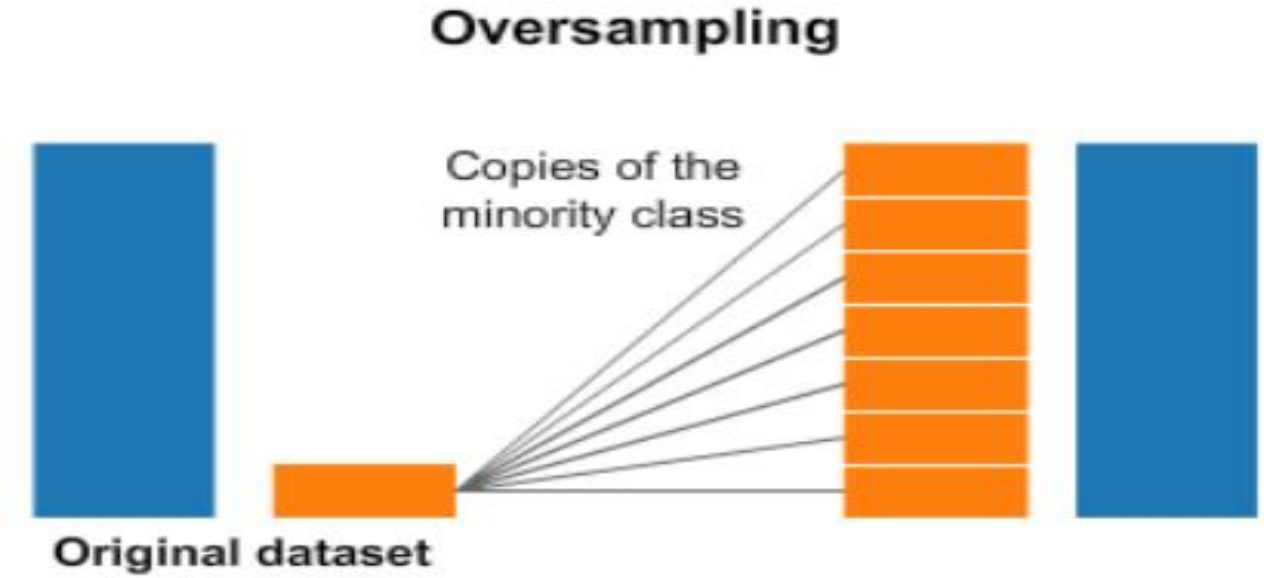

Image Taken From: [Important Techniques to Handle Imbalanced Data in Machine Learning… – Towards AI]

Advantages:

- No data loss
- Empirically seen to perform better than undersampling

Disadvantage:

- Prone to overfitting

SMOTE: Synthetic Minority Over-sampling TEchnique

- Avoid overfitting due to exact replicas of minority class samples
- Subset of Minority class is taken
- New synthetic data samples similar to the subset are created and added

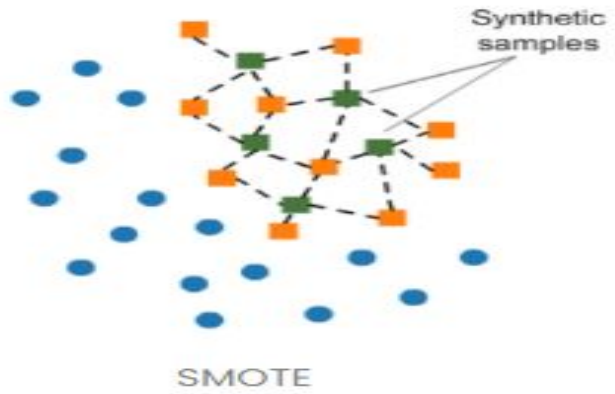

Image Taken From : [Important Techniques to Handle Imbalanced Data in Machine Learning… – Towards AI]

Advantages:

- No loss of useful information
- Reduced Overfitting

Disadvantages:

- May introduce noise
- Empirically seen to be less effective when data is high dimensional

Other Techniques:

- Penalization based models
- Class Weight Based Models
- Try to solve for a different evaluation measure
- Ensemble based models