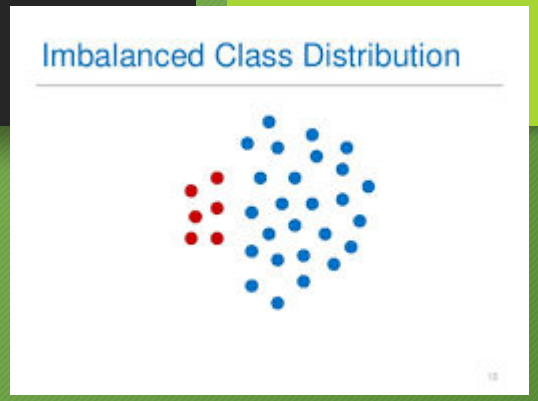
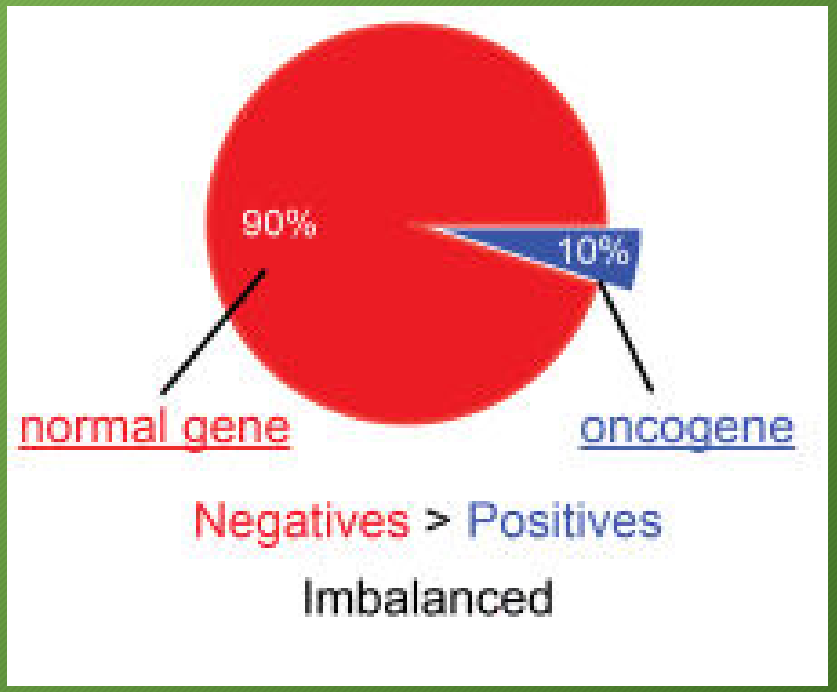
PERFORMANCE OF MODEL WITH BALANCED AND UNBALANCED DATASET-CHALLENGES THAT ARE FACED AND METHODS TO RESOLVE THE ISSUES

Presented by:
Priyansh Jain
A023119820019

UNBALANCED DATASET

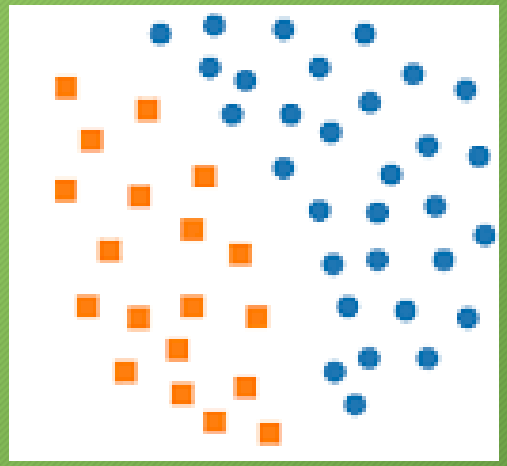
- Imbalanced data typically refers to classification tasks where the classes are not represented equally.
- For example, you may have a binary classification problem with 100 instances out of which 80 instances are labeled with Class-1, and the remaining 20 instances are marked with Class-2.
- This is essentially an example of an imbalanced dataset, and the ratio of Class-1 to Class-2 instances is 4:1.

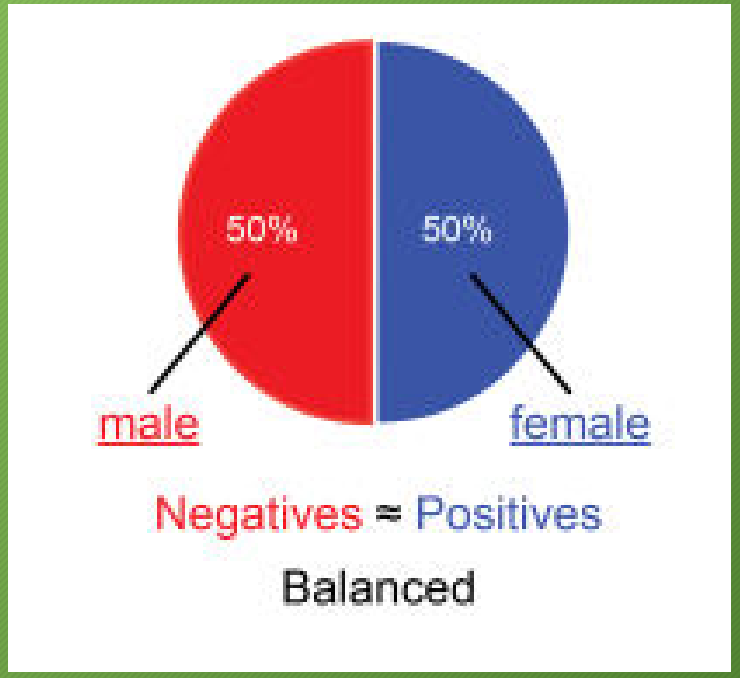




BALANCED DATASET

If in our data we have positive values which are approximately same as the negative values, then we can say that the dataset is balanced.





Why are Imbalanced Datasets a Serious Problem to Tackle?

- When the dataset has underrepresented data, the class distribution starts skew.

- Due to the inherent complex characteristics of the dataset, learning from such data requires new understandings, new approaches, new principles, and new tools to transform data. And moreover, this cannot anyway guarantee an efficient solution to your business problem. In worst cases, it might turn to complete wastes with zero residues to reuse.

Techniques to Convert Imbalanced Dataset into Balanced Dataset

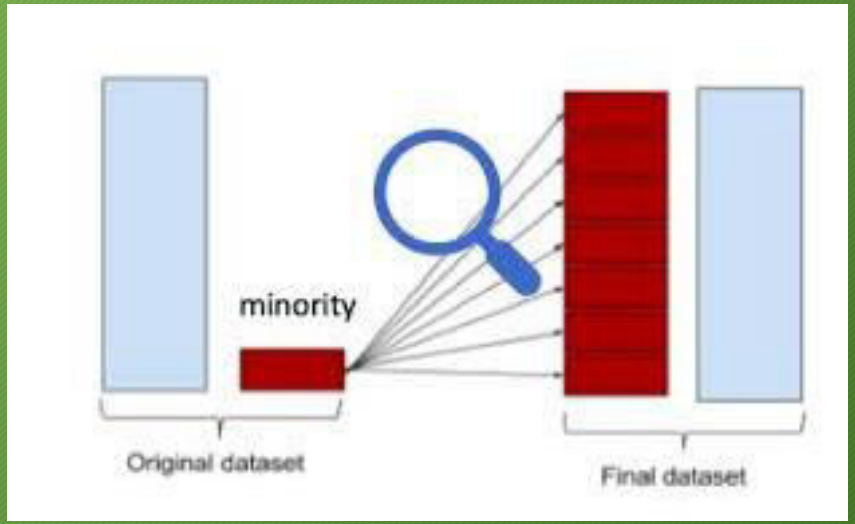
- Use the right evaluation metrics
- Over-Sampling or Up-Sampling
- Under-Sampling or Down Sampling
- Feature Selection
- Cost Sensitive Learning Technique
- Ensemble Learning Techniques

USE THE RIGHT EVALUATION METRICS

- **Confusion Matrix:** a table showing correct predictions and types of incorrect predictions.
- **Precision:** the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier's exactness. Low precision indicates a high number of false positives.
- **Recall:** the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier's completeness. Low recall indicates a high number of false negatives.
- **F1-Score:** the weighted average of precision and recall.

OVER-SAMPLING

This technique is used to modify the unequal data classes to create balanced datasets. When the quantity of data is insufficient, the oversampling method tries to balance by incrementing the size of rare samples.



UNDER-SAMPLING

- This technique balances the unbalanced dataset by reducing the size of the class which is in abundance. There are various methods for classification problems such as cluster centroids and Tomek links. The cluster centroid methods replace the cluster of samples by the cluster centroid of a K-means algorithm.

FEATURE SELECTION

In order to tackle the imbalance problem, we calculate the one-sided metric such as correlation coefficient (CC) and odds ratios (OR) or two-sided metric evaluation such as information gain (IG) and chi-square (CHI) on both the positive class and negative class. Based on the scores, we then identify the significant features from each class and take the union of these features to obtain the final set of features. Then, we use this data to classify the problem.

COST-SENSITIVE LEARNING TECHNIQUE

The Cost-Sensitive Learning (CSL) takes the misclassification costs into consideration by minimizing the total cost. The goal of this technique is mainly to pursue a high accuracy of classifying examples into a set of known classes. It is playing as one of the important roles in the machine learning algorithms including the real-world data mining applications.

Cost-Sensitive Learning Framework

- Define the cost of misclassifying a majority to a minority as $C(Min, Maj)$
- Typically $C(Maj, Min) > C(Min, Maj)$
- Minimize the overall cost - usually the *Bayes conditional risk* - on the training data set

$$R(i|x) = \sum_j P(j|x)C(i,j)$$

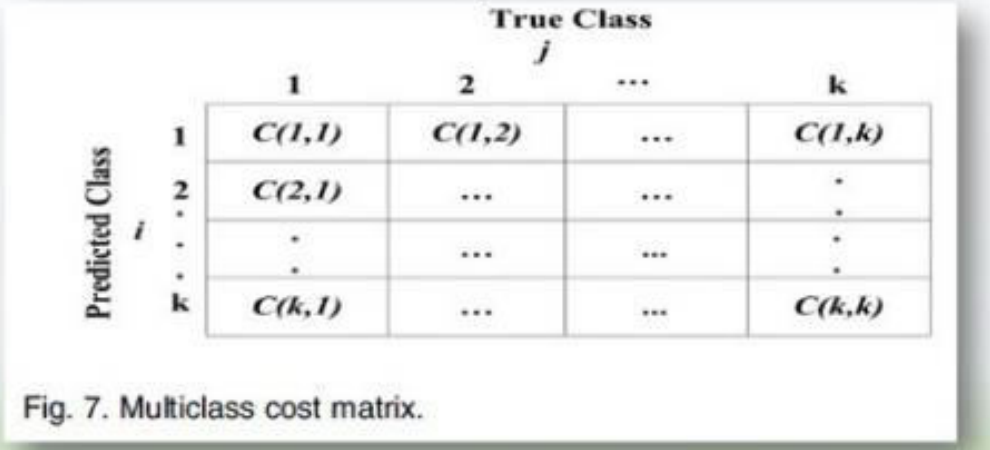
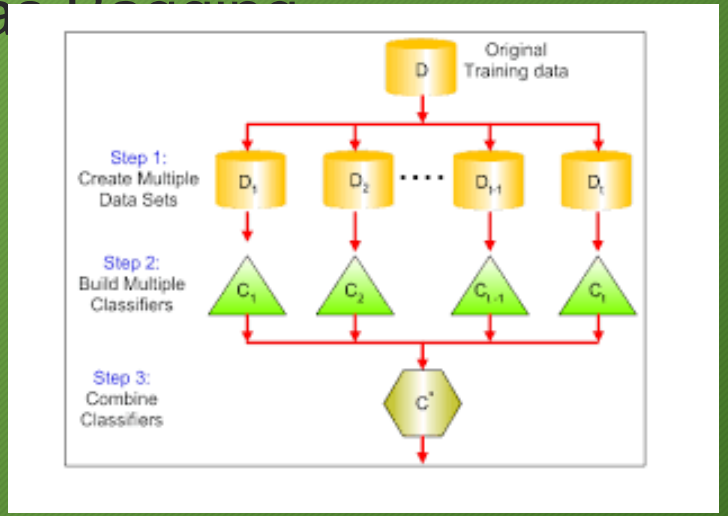


Fig. 7. Multiclass cost matrix.

Advantages:
This technique avoids pre-selection of parameters and auto-adjusts decision hyperplane

ENSEMBLE LEARNING TECHNIQUE

The ensemble-based method is another technique which is used to deal with imbalanced data sets, and the ensemble technique is combined the result or performance of several classifiers to improve the performance of single classifier. This method modifies the generalization ability of individual classifiers by assembling various classifiers. It mainly combines the outputs of multiple base learners. There are various approaches in ensemble learning such as Bagging, Boosting, etc.




thank
you