CLASSIFICATION METRICS XIII

Receiver Operating Characteristics (ROC) Curve

- It is one of the most important evaluation metrics for checking the performance of the classification model. Intuitively, it tells how much model is capable of distinguishing between classes.
- ROC is a probability curve which is plotted against True Positive rate and False Positive rate.
- Here, TPR is takes the X-axis whereas, Y-axis is FPR. TPR is same as recall of positive class. In other words, it is sensitivity, refer Equation 15. On the other hand FPR is measured using Equation 16.

$$TPR = \frac{TP}{TP + FN} \qquad (15)$$

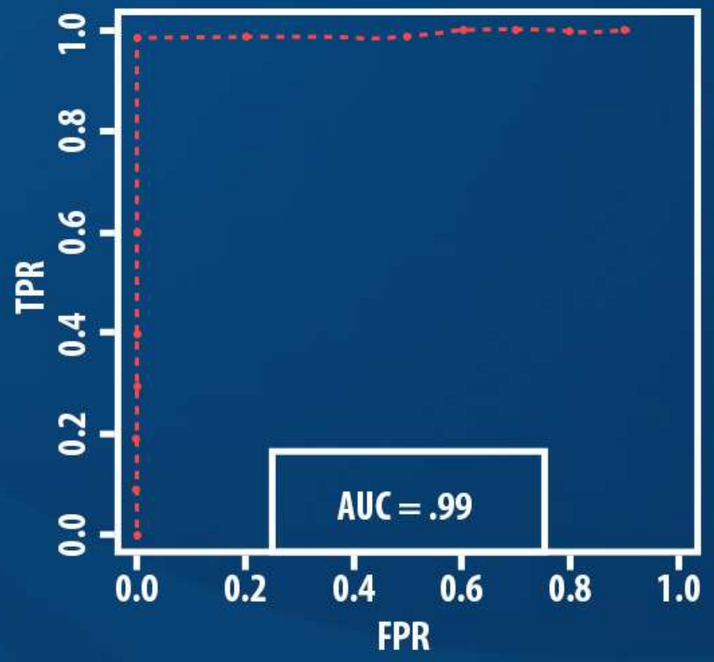
$$FPR = \frac{FP}{TN + FP} \qquad (16)$$

- FPR is also calculated by subtracting 1 from specificity.

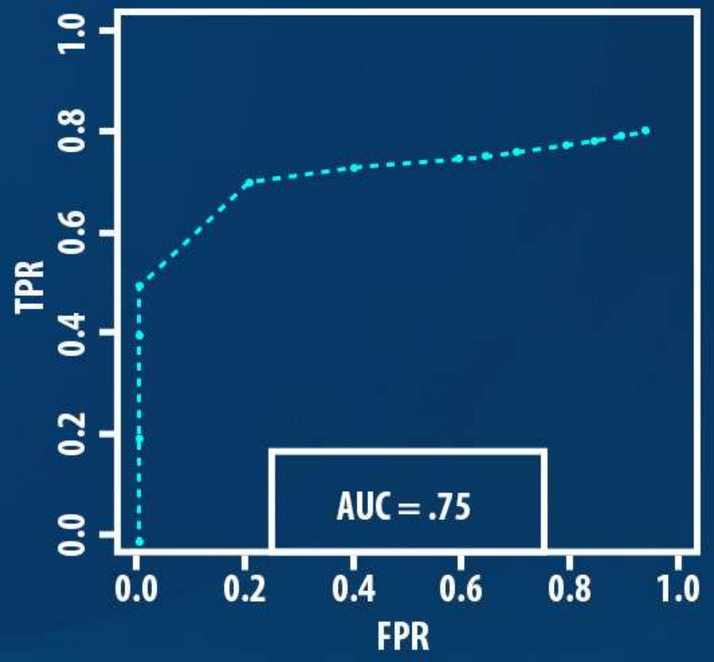
CLASSIFICATION METRICS XIV

ROC interpretation

- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- AUC (area under curve) uses ROC to find the area under the curve.



(a) ROC curve with AUC value 0.99

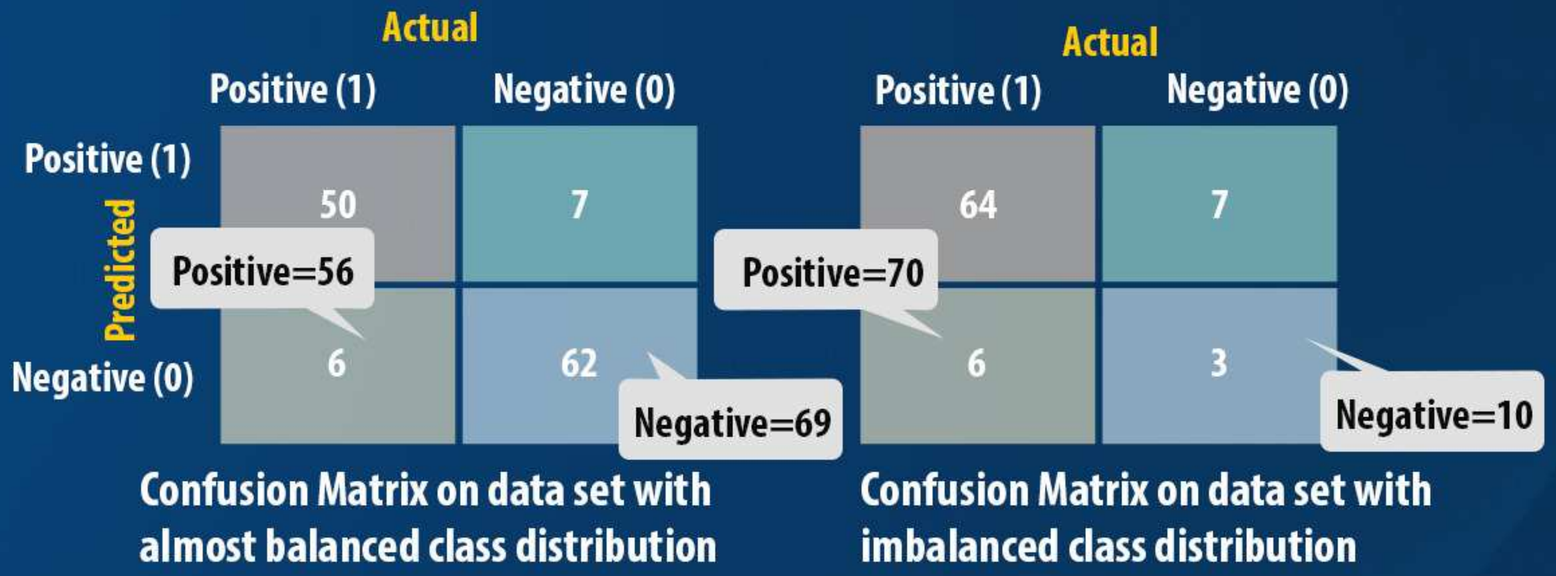


(b) ROC curve with AUC value 0.75

Figure 59: ROC curve comparison. Performance of model in Figure 59a is better than the model with performance in Figure 59b

ISSUE OF CLASS IMBALANCE IN CLASSIFICATION METHODS I

- There is an issue in classification where the number of observations belonging to one class is significantly lower than those belonging to the other classes.
- In result, performance of classification may drop on the class with lower number of observations.



Confusion Matrix on data set with almost balanced class distribution

Confusion Matrix on data set with imbalanced class distribution

Figure 60: Performance of classifier on two different data sets with imbalance distribution of classes

Confusion Matrix 1	Confusion Matrix 2
N = 100	N = 100
Positive class = 55	Positive class = 70
Negative class = 45	Negative class = 30

Table 9: Class distribution for Confusion Matrixes above

ISSUE OF CLASS IMBALANCE IN CLASSIFICATION METHODS I

Approaches to handle Imbalanced Class Distribution

- Dealing with imbalanced datasets entails following strategies:

1. Balancing classes in the data set (data preprocessing) before input to the machine learning algorithm.
2. Improving classification algorithms using Ensemble Techniques.

ISSUE OF CLASS IMBALANCE IN CLASSIFICATION METHODS III

Handling imbalance class distribution using Data preprocessing

1. Random Under-sampling
   - Random Undersampling aims to balance class distribution by randomly eliminating majority class examples. This is done until the majority and minority class instances are balanced out.
2. Random Over-Sampling
   - Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample.

Under-Sampling	Over-Sampling
N =100, Positive class = 70, Negative class = 30	
Take random 50% samples from Positive class	Increase samples in Negative class
Data set = 35(Positive class) + 30(Negative class)	Data set = 70(Positive class) + 60(Negative class)

Table 10: Example illustration of Random under-sampling and Random over-sampling

ISSUE OF CLASS IMBALANCE IN CLASSIFICATION METHODS IV

Characteristics of Random Under-Sampling and Random Over-Sampling

1. Random Under-Sampling can discard potentially useful information which could be important for building rule classifiers.
2. Random Under-sampling can impact classifier's performance.
3. Random Over-Sampling method leads to no information loss.
4. Random Over-Sampling increases the likelihood of overfitting since it replicates the minority class events.

ISSUE OF CLASS IMBALANCE IN CLASSIFICATION METHODS V

Improving classification algorithms using Ensemble Techniques

The main objective of ensemble approach is to improve the performance of single classifiers. The approach involves constructing several two stage classifiers from the original data and then aggregate their predictions.

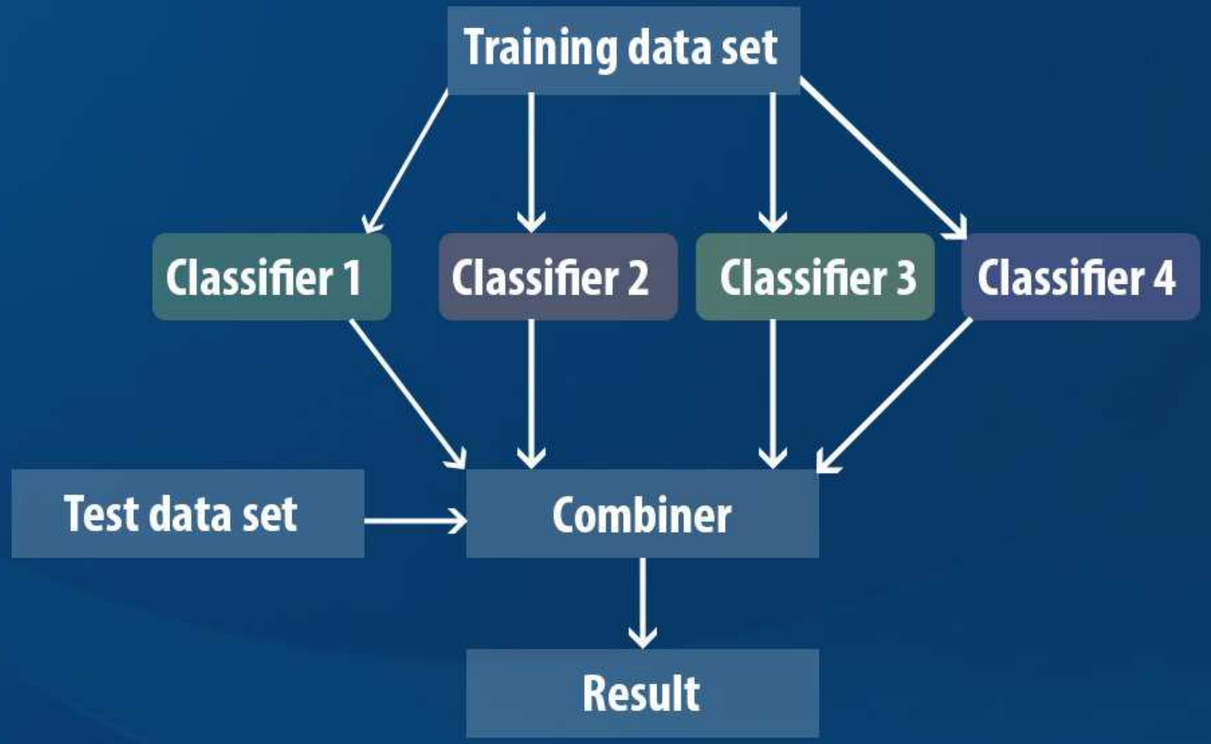


Figure 61: Ensemble Classification