Classification II

1- Evaluating Classification Models

2- Strategies for Imbalanced Data

1- Evaluating Classification Models

- Evaluating a classification model is as important as building it. We are creating models to perform on new, previously unseen data. Hence, a thorough and versatile evaluation is required to create a robust model.

I- Accuracy
II- Confusion Matrix
III- Prediction and recall
IV- Specificity and sensivity
V- ROC and AOC

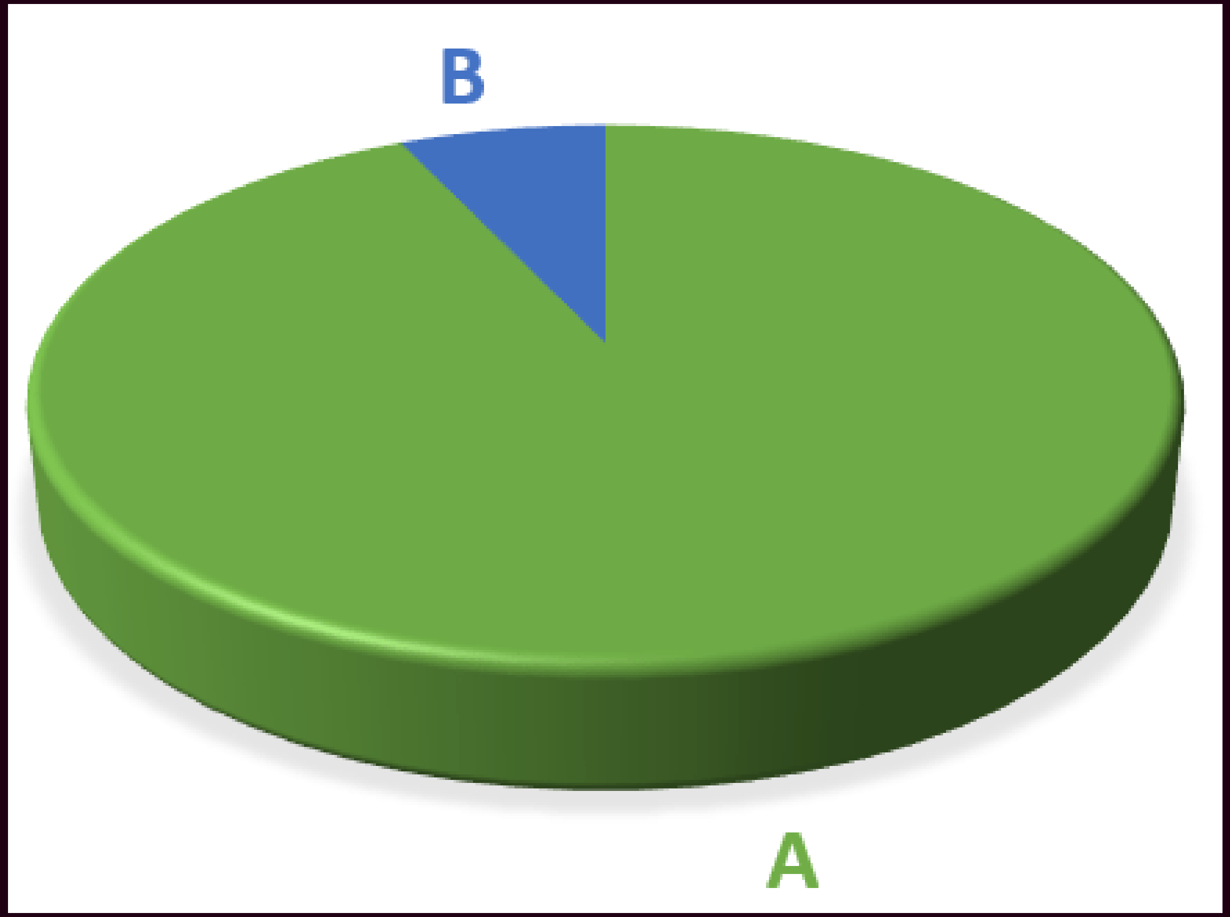
1-1 Accuracy

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ all\ predictions}$$

- Classification accuracy shows how many of the predictions are correct.

1-1 Accuracy

- In some cases, it represents how good a model is but there are some cases in which accuracy is simply not enough. For example, 93% means that we correctly predicted 93 out of 100 samples. It seems acceptable without knowing the details of the task.
- Assume we are creating a model to perform binary classification on a dataset with an unbalanced class distribution. 93% of data points are in class A and 7% in class B.

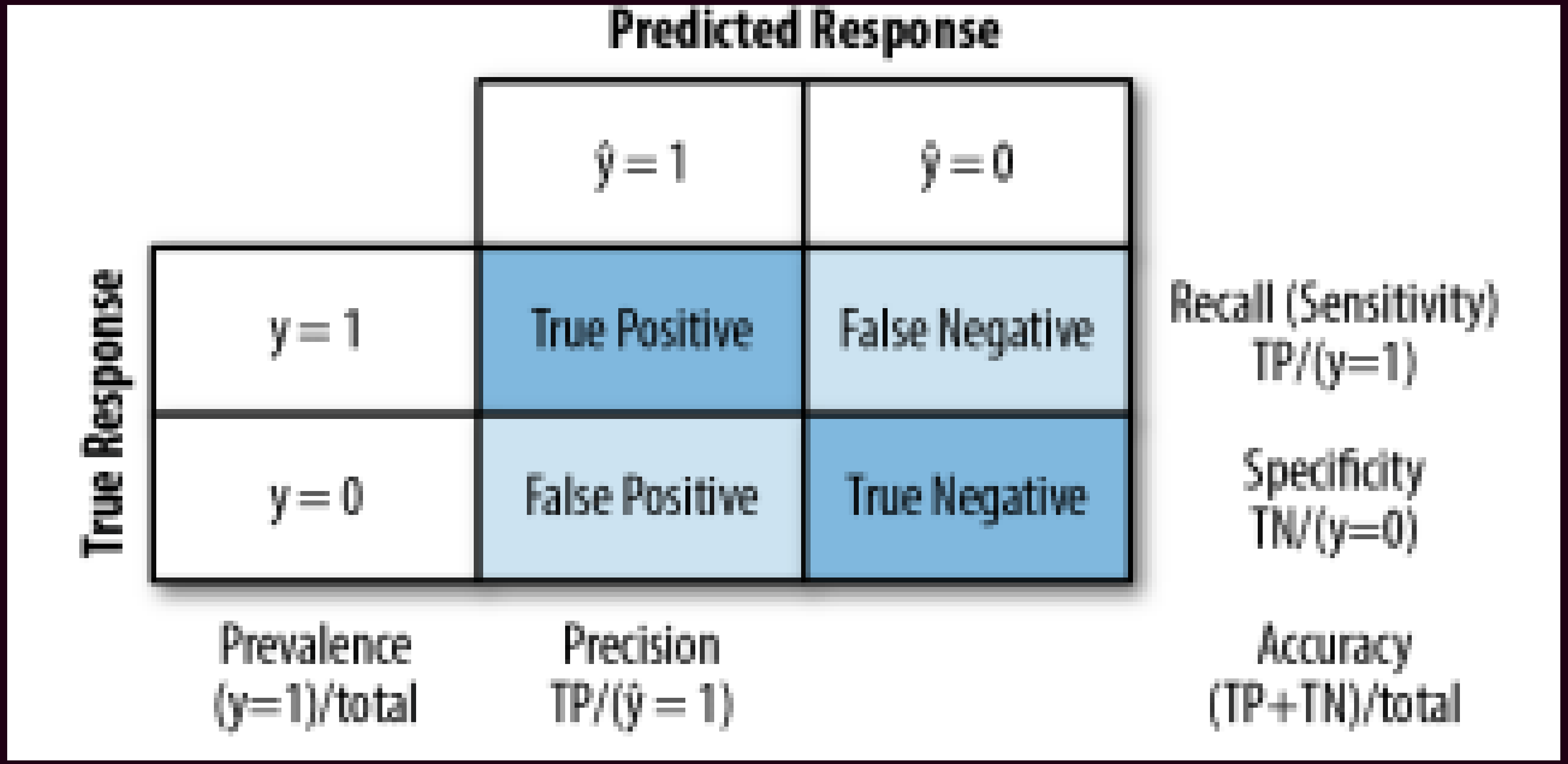


- We have a model that only predicts class A. It is hard to even call it a “model” because it predicts class A without any calculation. However, since 93% of the samples are in class A, the accuracy of our model is 93%.
- What if it is crucial to detect class B correctly and we cannot afford to misclassify any class B samples (i.e. cancer prediction)? In these cases, we need other metrics to evaluate our model.

1-2 Confusion Matrix

- A confusion matrix is a technique for summarizing the performance of a classification algorithm.
- Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset.
- Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making.

--> Following the usual conventions, Y = 1 corresponds to the event of interest , and Y = 0 corresponds to a negative (or usual) event .



		Predicted Response		
		$\hat{y} = 1$	$\hat{y} = 0$	
True Response	$y = 1$	True Positive	False Negative	Recall (Sensitivity) TP/(y=1)
	$y = 0$	False Positive	True Negative	Specificity TN/(y=0)
	Prevalence (y=1)/total	Precision $TP/(\hat{y} = 1)$		Accuracy (TP+TN)/total

1-3 Precision and Recall

Precision and recall metrics take the classification accuracy one step further and allow us to get a more specific understanding of model evaluation. Which one to prefer depends on the task and what we aim to achieve.

Precision : ***measures how good our model is when the prediction is positive.***
--> The focus of precision is positive predictions. It indicates how many positive predictions are true.

$$Precision = \frac{TP}{TP + FP}$$

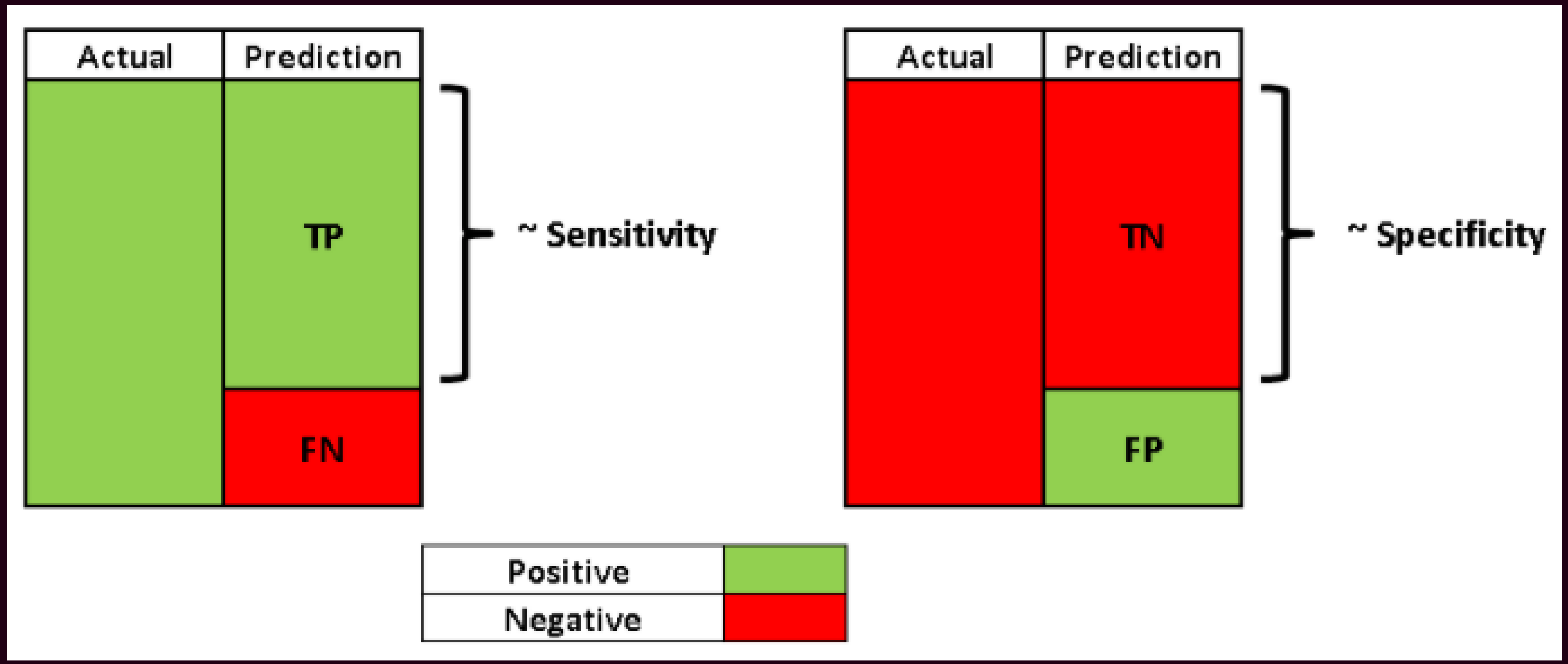
Precision : Recall measures how good our model is at correctly predicting positive classes.
--> The focus of recall is actual positive classes. It indicates how many of the positive classes the model is able to predict correctly.

$$Recall = \frac{TP}{TP + FN}$$

1-4 Specificity and Sensitivity

- **Sensitivity**, also known as the true positive rate (TPR), is the same as recall. Hence, it measures the proportion of positive class that is correctly predicted as positive.

- **Specificity** is similar to sensitivity but focused on negative class. It measures the proportion of negative class that is correctly predicted as negative.



1-5 ROC and AUC

- ROC (receiver operating characteristic) :

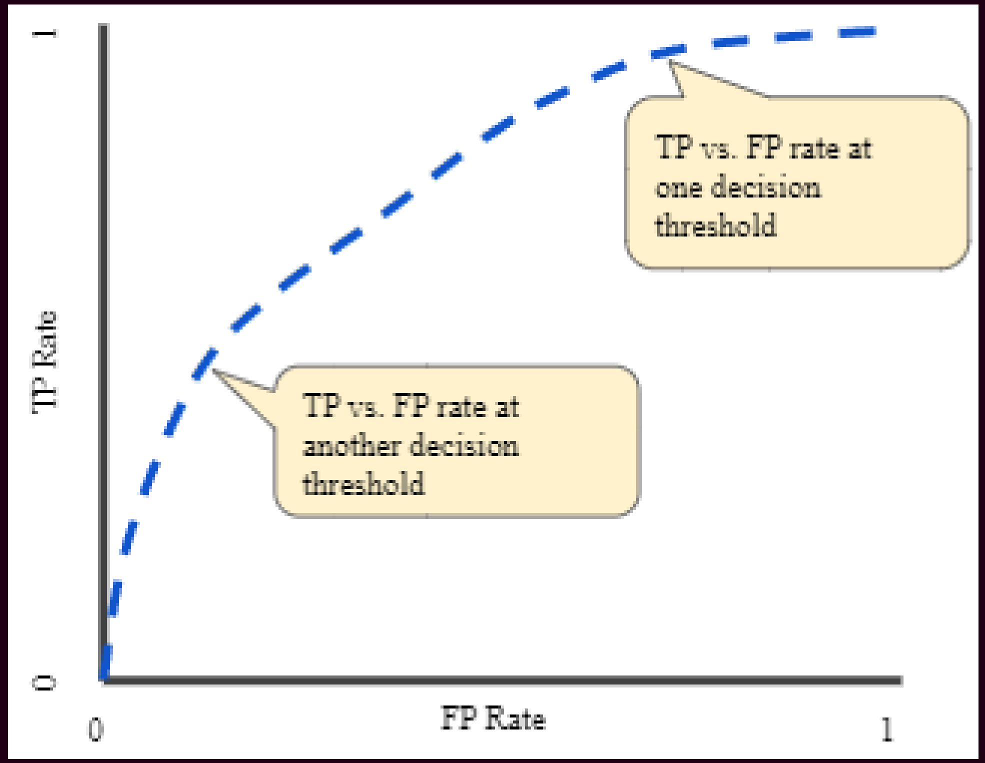
--> An ROC curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

  - True Positive Rate
  - False Positive Rate

- True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows: $TPR = \frac{TP}{TP + FN}$

- False Positive Rate (FPR) is defined as follows: $FPR = \frac{FP}{FP + TN}$

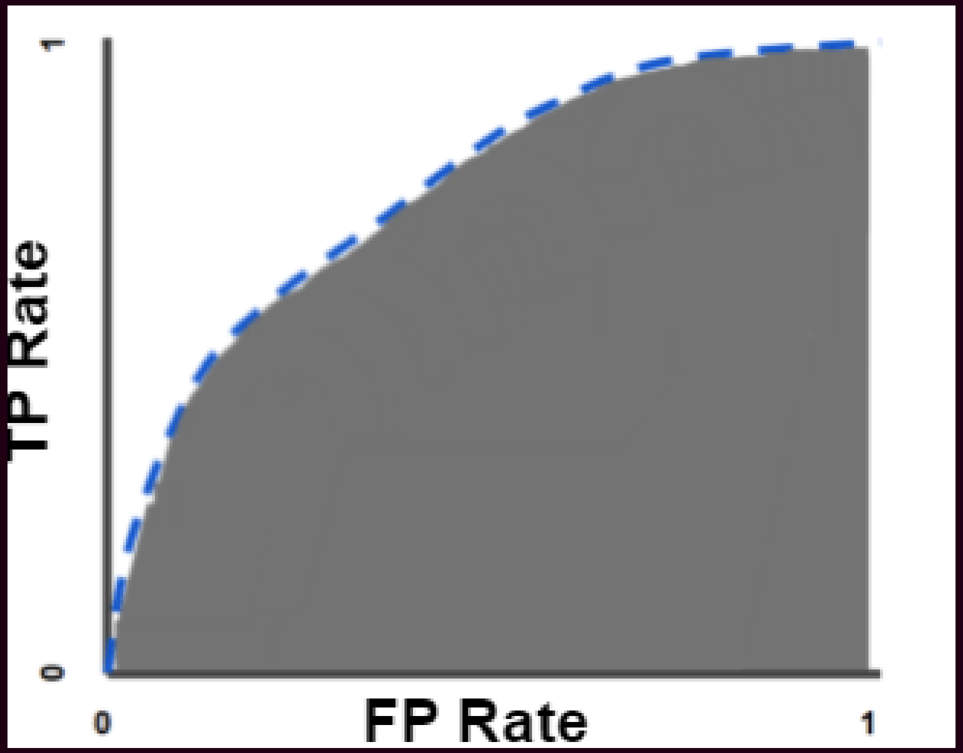
- An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.



1-5 ROC and AUC

- AUC (Area Under the ROC Curve) :

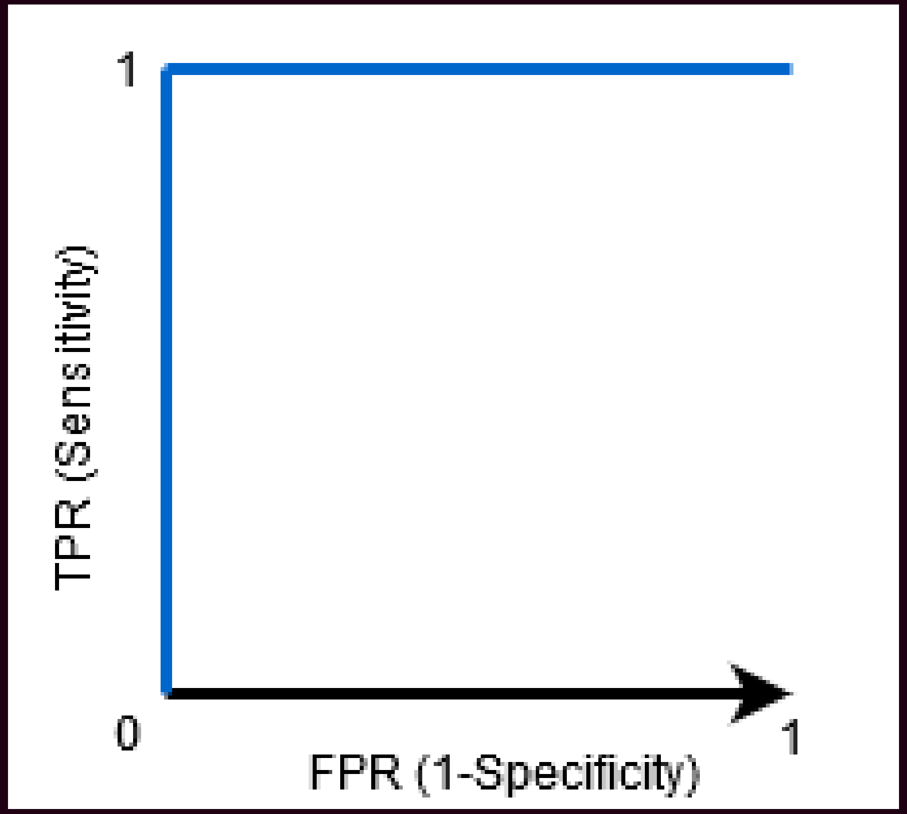
--> AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1).



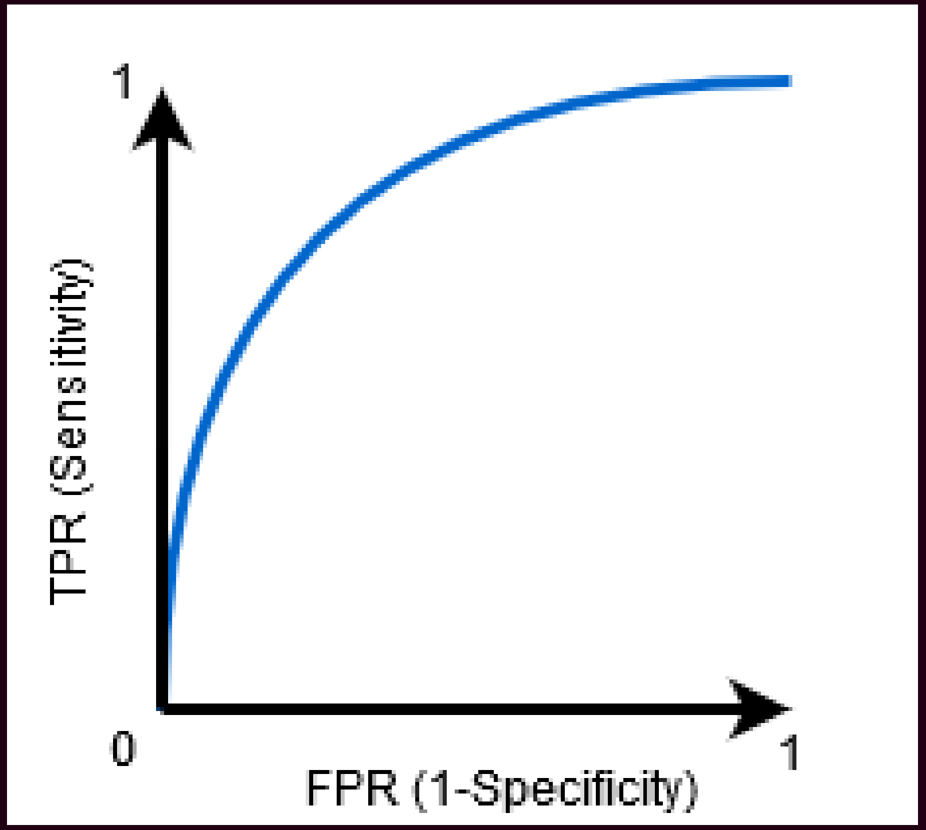
- The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

1-5 ROC and AUC

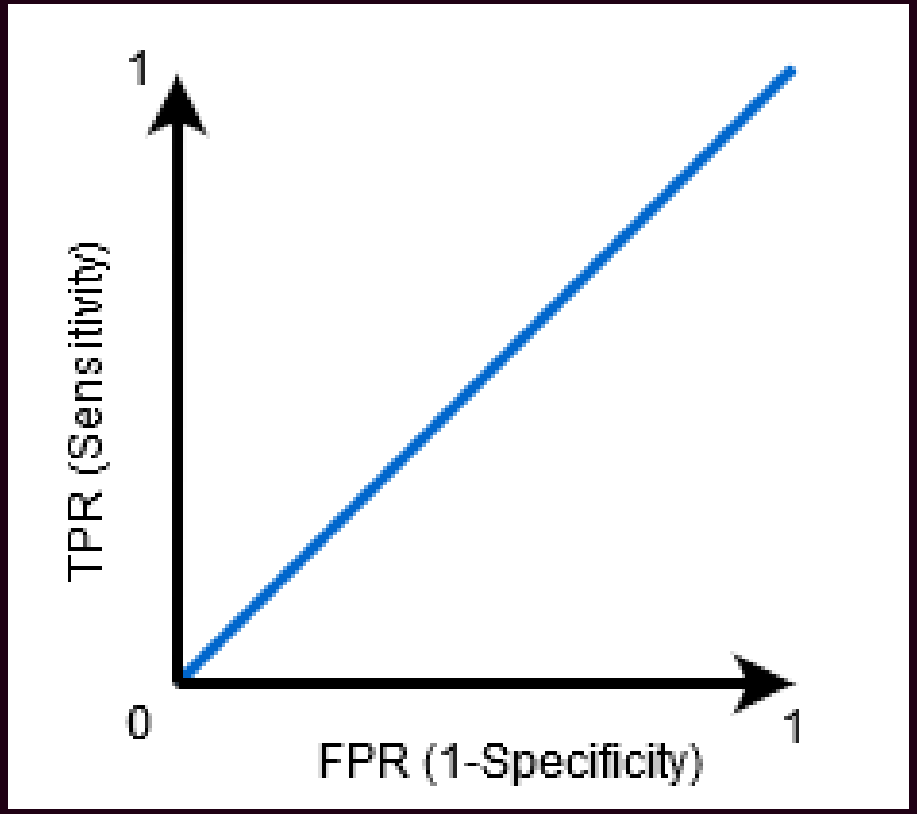
AUC = 1



0.5 < AUC < 1



AUC = 0.5



2- Stategies for imbalanced data

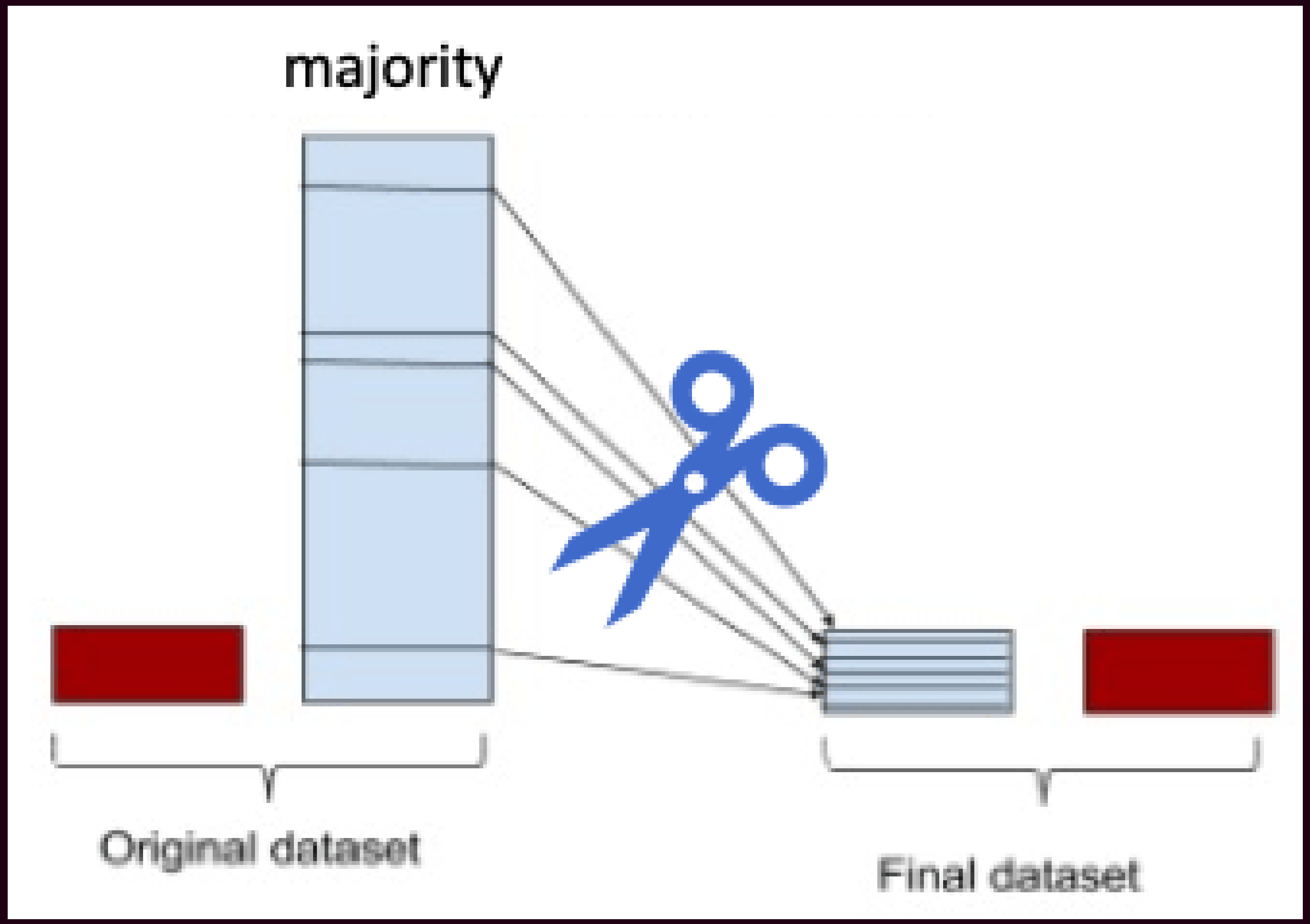
- An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. The distribution can vary from a slight bias to a severe imbalance where there is one example in the minority class for hundreds, thousands, or millions of examples in the majority class or classes.

I- Undersampling
II- Oversampling

2-1 Undersampling

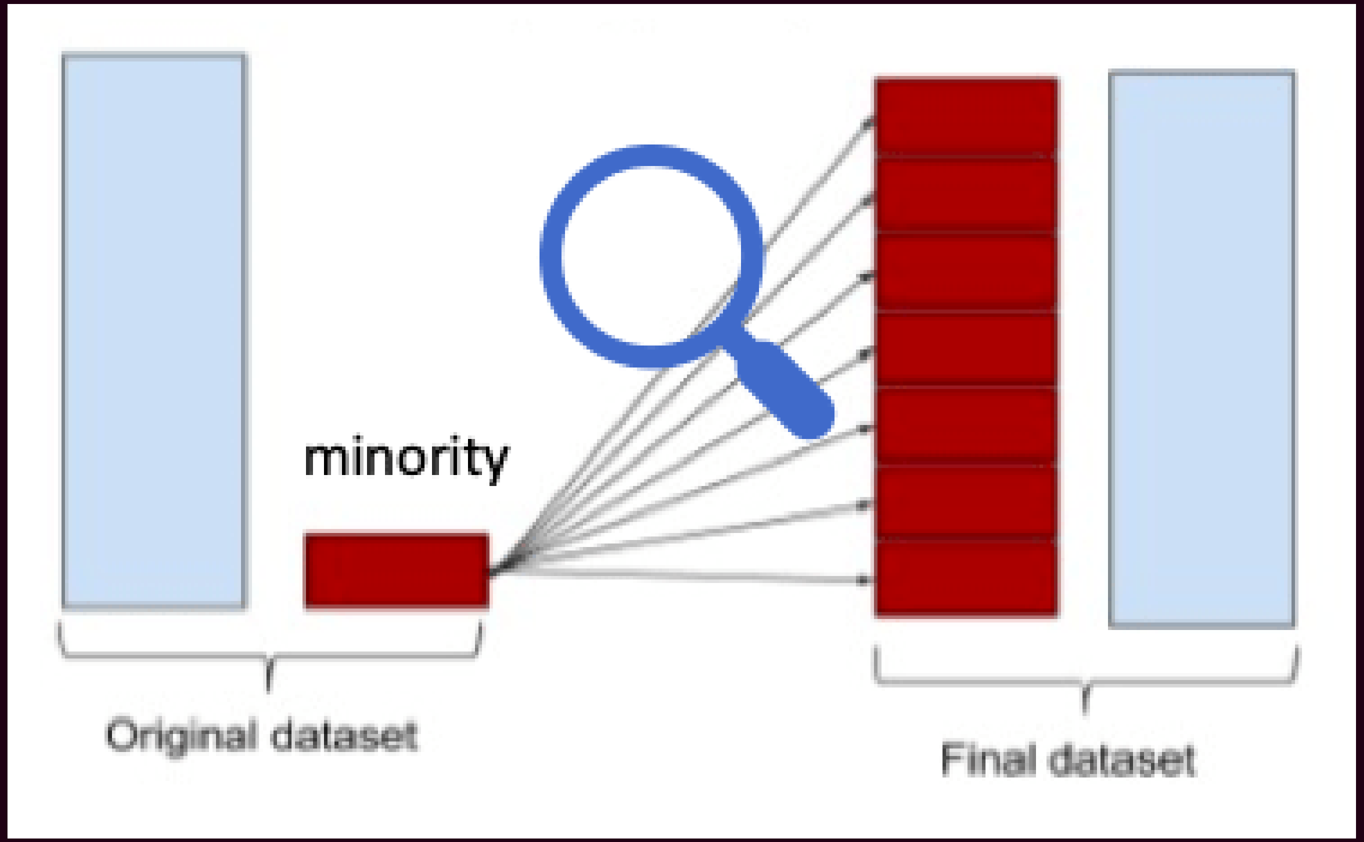
- If you have enough data, as is the case with the loan data, one solution is to under- sample (or downsample) the prevalent class, so the data to be modeled is more bal- anced between 0s and 1s. The basic idea in undersampling is that the data for the dominant class has many redundant records. Dealing with a smaller, more balanced data set yields benefits in model performance, and it makes it easier to prepare the data and to explore and pilot models.

- A huge disadvantage of undersampling is the risk of information loss.



2-1 Oversampling

- If the majority class in a dataset doesn't have the abundance of entries, we can opt to increase the size of the rarity class.
- It's the case that data scientist are most faced with
- The disadvantage with this technique is the increase in variance which may result in overfitting



Take away

- Accuracy (the percent of predicted classifications that are correct) is but a first step in evaluating a model.
- Other metrics (recall, specificity, precision) focus on more specific performance characteristics (e.g., recall measures how good a model is at correctly identifying 1s).
- AUC (area under the ROC curve) is a common metric for the ability of a model to distinguish 1s from 0s.
- Highly imbalanced data (i.e., where the interesting outcomes, the 1s, are rare) are problematic for classification algorithms.
- One strategy for working with imbalanced data is to balance the training data via undersampling the abundant case (or oversampling the rare case).

Thank You
For Your
Attention