Performance Metrics

Elena Bellodi
elena.bellodi@unife.it

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Regione Emilia-Romagna

FERRARIAE UNIVERSITAS · 13 91 · EX LABORE FRUCTUS

Università degli Studi di Ferrara

Outline

- Metrics for Hard Prediction
- Metrics for Ranking Prediction
- Metrics for Regression

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for Ranking Prediction

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

Receiver Operating Characteristics (ROC) curve

- Plots the TP Rate (Recall) against the FP Rate to graphically illustrate the performance of a binary classifier **in terms of tradeoffs between benefits (true positives) and costs (false positives)**
- A **discrete classifier** is one that outputs only a class label. Each discrete classifier produces an *(fp rate, tp rate)* pair corresponding to a single point in ROC space.

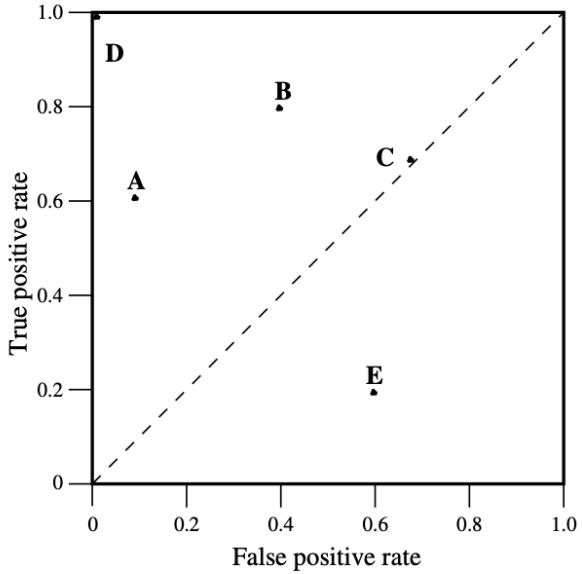


Fig. 2. A basic ROC graph showing five discrete classifiers.

The point (0, 1) represents perfect classification.

D's performance is perfect.

C's performance is random (see next).

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

Receiver Operating Characteristics (ROC) curve

- one point in ROC space **is better than another** if it is to the **northwest** (tp rate is higher, fp rate is lower, or both) of the first.
- Any classifier that appears in the **lower right triangle performs worse** than random guessing.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

Receiver Operating Characteristics (ROC) curve

Ranking/scoring classifiers

- Some classifiers, such as a Naive Bayes classifier or a neural network return a numeric value, **a score or a probability**, which represents the degree to which an instance is a member of class 0 or 1
- These values can be strict **probabilities**, in which case they adhere to standard theorems of probability; or they can be general, **uncalibrated scores**, in which case the only property that holds is that a higher score indicates a higher probability

- **Such a classifier can be used with a threshold to produce a discrete (binary) classifier**: if the classifier output is above the threshold, the classifier produces a Yes, else a No

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

Receiver Operating Characteristics (ROC) curve

- **Each threshold value produces a different point in ROC space**
- Conceptually, we may imagine varying a threshold from -inf to +inf and tracing a curve through ROC space
- FP Rate and TP Rate both have values in the range [0, 1]
- The curve always starts at the point (0,0) produced by the threshold of +inf
- As the threshold is further reduced, the curve climbs up and to the right, ending up at (1, 1)

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

ROC Curve toy example

- test set of 20 instances: 10 positive and 10 negative (actual class)
- the instances are sorted by their scores, the outputs of the scoring classifier
- Thresholds t = [+inf (or 1), 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1]

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

ROC Curve toy example

test set

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

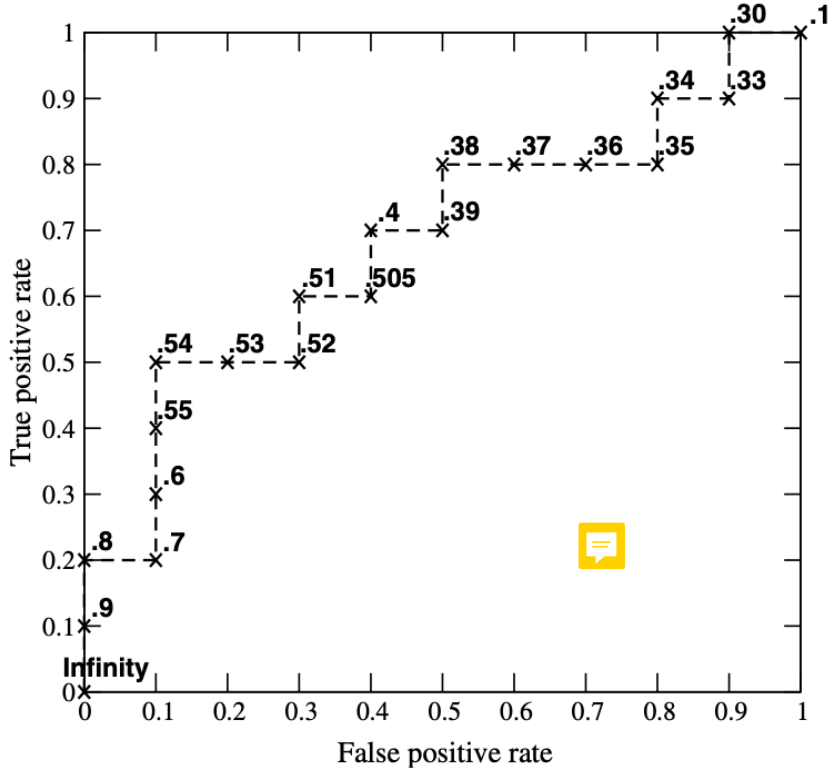


Fig. 3. The ROC "curve" created by thresholding a test set. The table shows 20 data and the score assigned to each by a scoring classifier. The graph shows the corresponding ROC curve with each point labeled by the threshold that produces it.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

ROC Curve toy example

test set

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

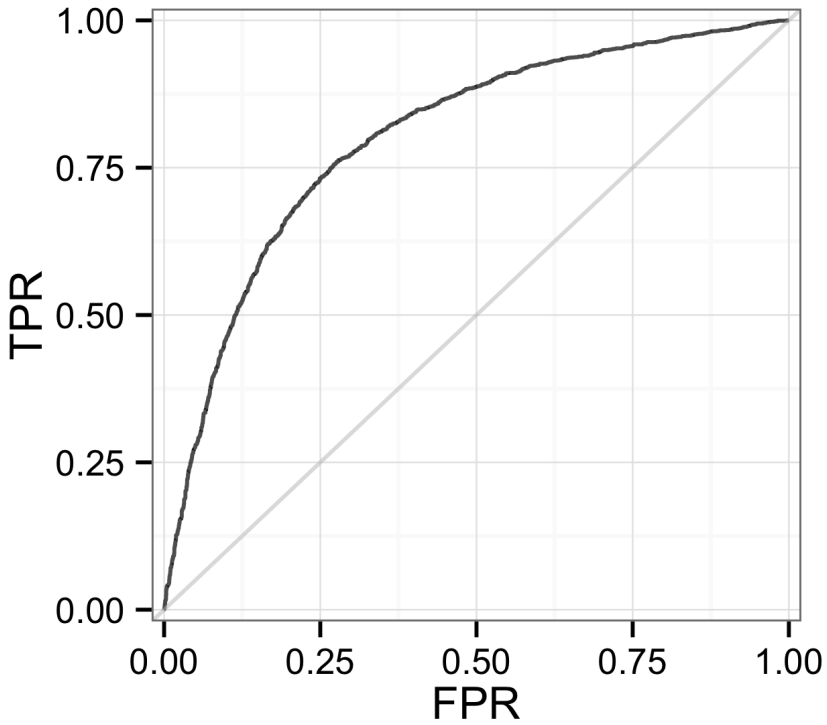
1. if t = 0.9 and score_1 = 0.9 → #inst1 is classified + as its score is >= t
   - fpr = 0, tpr = 1/(1+9)=0.1 → **point (0, 0.1)**
2. if t = 0.8 and score_2 = 0.8 → #inst2 is classified + as its score is >= t
   - fpr = 0, tpr = 2/(2+8)=0.2 → **point (0, 0.2)**
3. if t = 0.7 and score_3 = 0.7 → #inst3 is classified + as its score is >= t, but it's a **n** example
   - fpr = 1/10 , tpr = the value before → **point (0.1, 0.2)**

- ...

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

ROC curve: a real example



threshold values such as [0, 0.02, 0.04, …., 1]

Y	$p < \theta$	$p \geq \theta$
0	True Neg.	False Pos.
1	False Neg.	True Pos.

Observed	0	0	1	1	...
Predicted	0.1	0.6	0.4	0.8	...

Threshold	TPR	FPR
0.99998	0.00000	0.00000
0.99998	0.00048	0.00000
0.99974	0.00048	0.00034
0.99953	0.00097	0.00034
0.99946	0.00145	0.00034
0.99942	0.00194	0.00034

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

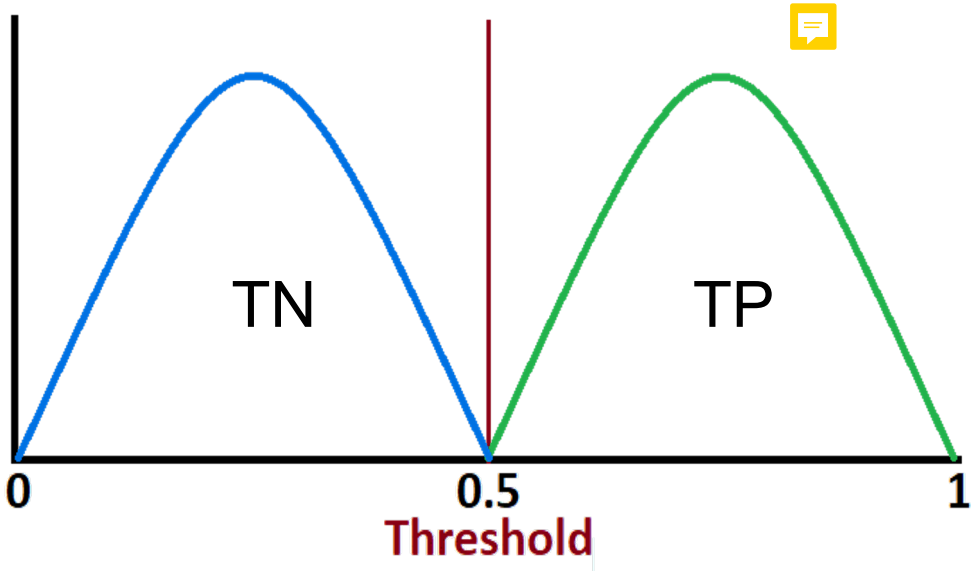
Area Under the ROC Curve (AUC ROC)

- An ROC curve is a two-dimensional depiction of classifier performance. **To compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance**
  - The area under the curve (often referred to as simply the AUC) is equal to **the probability that a classifier will rank a randomly chosen positive instance higher** than a randomly chosen negative one
  - It tells how much the model is capable of distinguishing between classes

- ROC AUC varies between 0 and 1 — with an uninformative classifier yielding 0.5

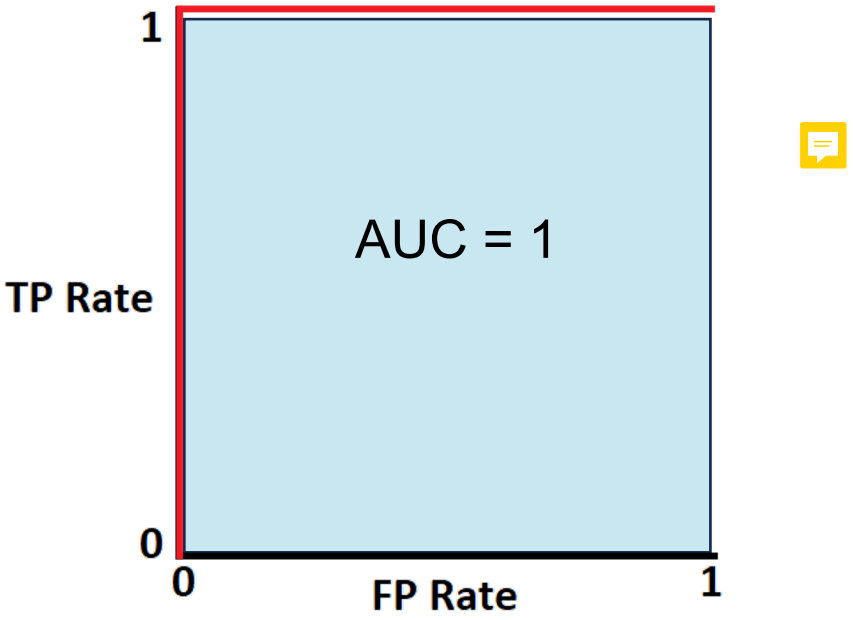
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction



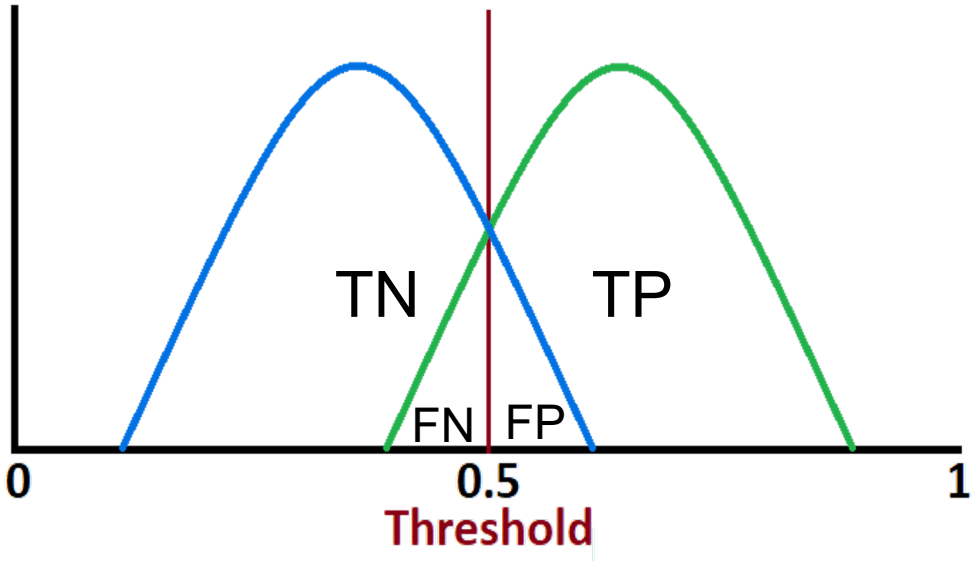
Let the green distribution curve be of the positive class and blue distribution curve be of the negative class.

Best situation: the model can distiguish examples without errors.



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

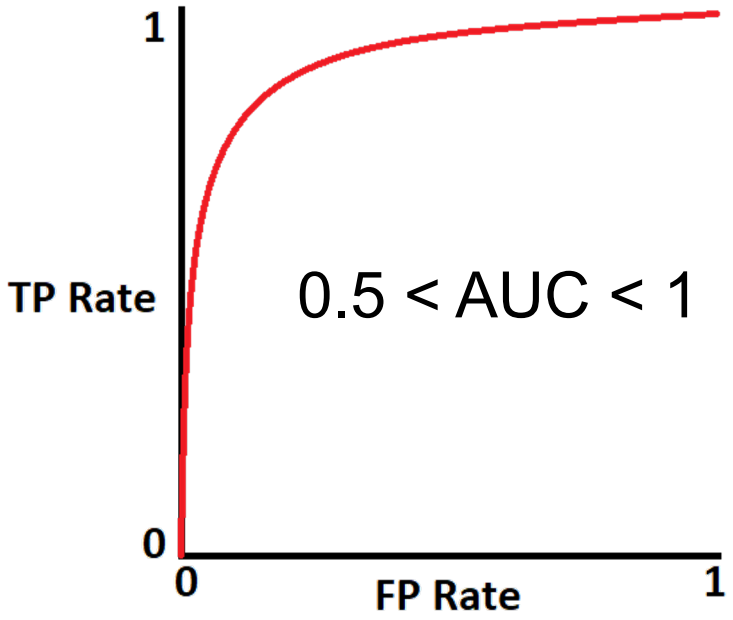
Metrics for ranking prediction



The distributions overlap. The model can distinguish between the two classes with, e.g., 80% chance (the value of AUC).

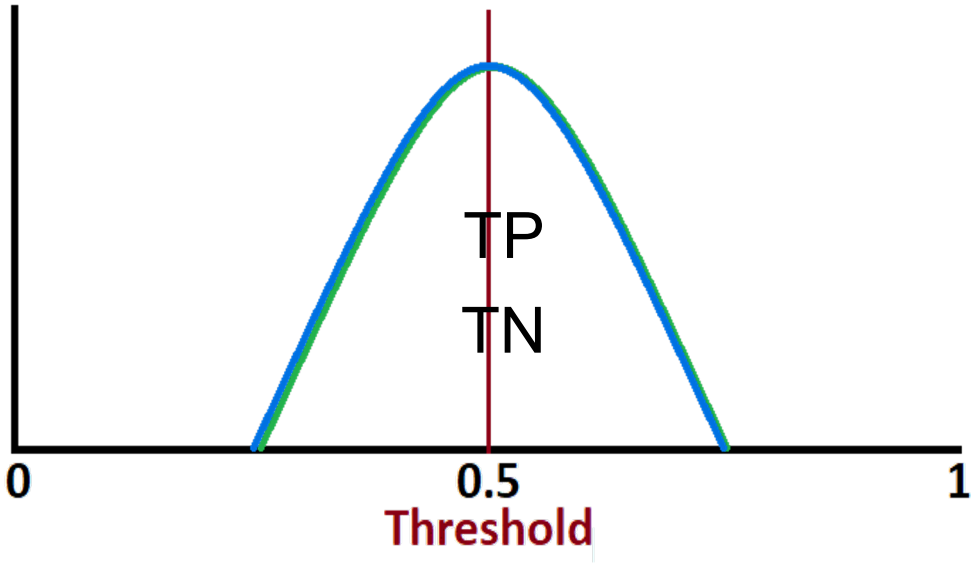
There are two errors that can be maximized or minimized by moving the threshold.

Let the green distribution curve be of the positive class and blue distribution curve be of the negative class.



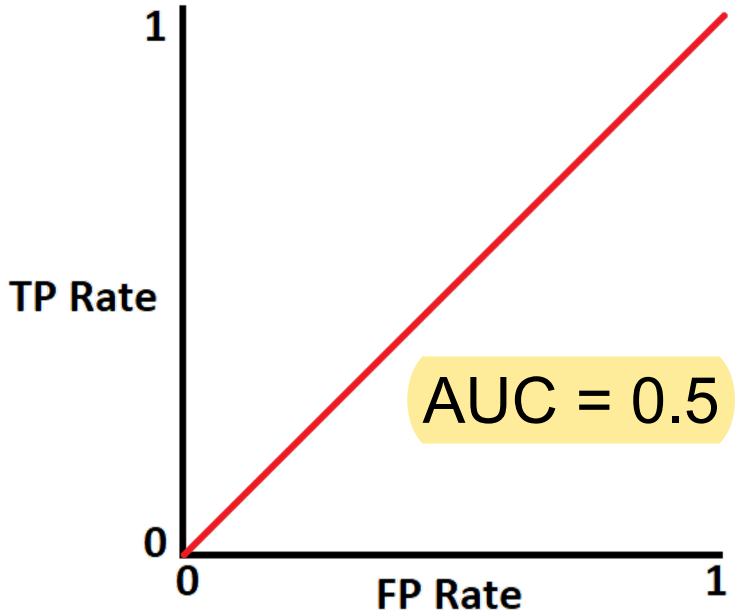
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction



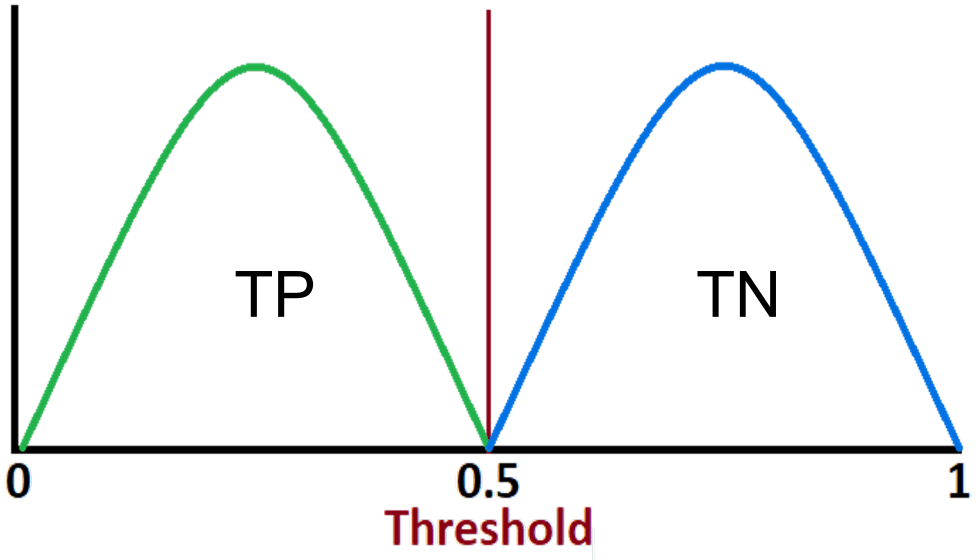
Let the green distribution curve is of the positive class and blue distribution curve is of negative class.

Worst situation: the model is not able to distiguish the examples, it performs a random guess.



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

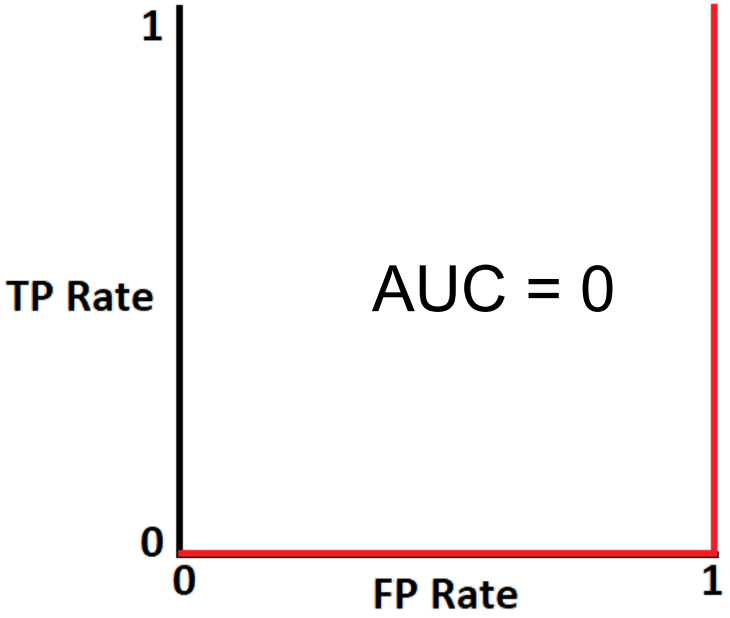
Metrics for ranking prediction



The model assigns the positive class to negative examples and negative class to positive examples.

If the AUC = 0, it can perfectly distiguish the two classes but assigns inverted labels.

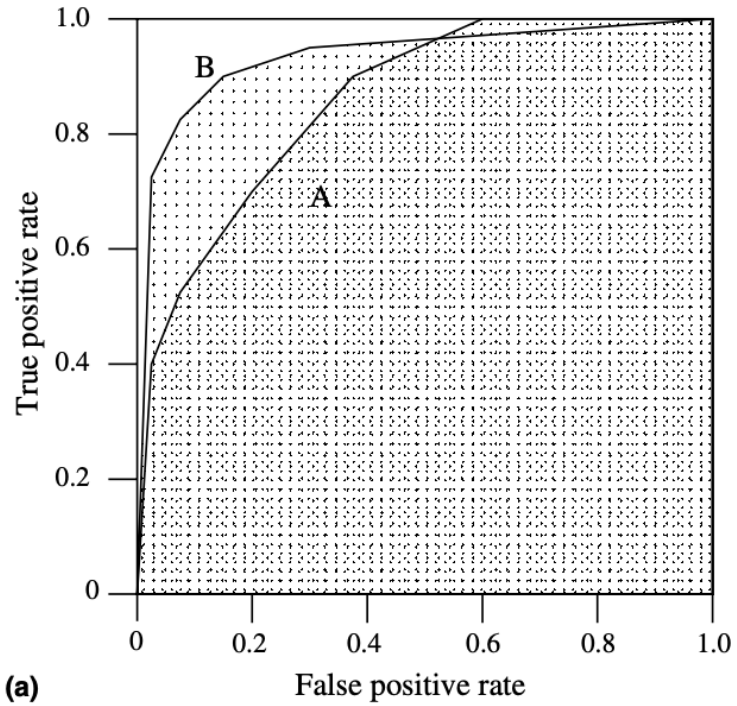
Let the green distribution curve be of the positive class and blue distribution curve be of the negative class.



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

- Comparison of classifiers performance

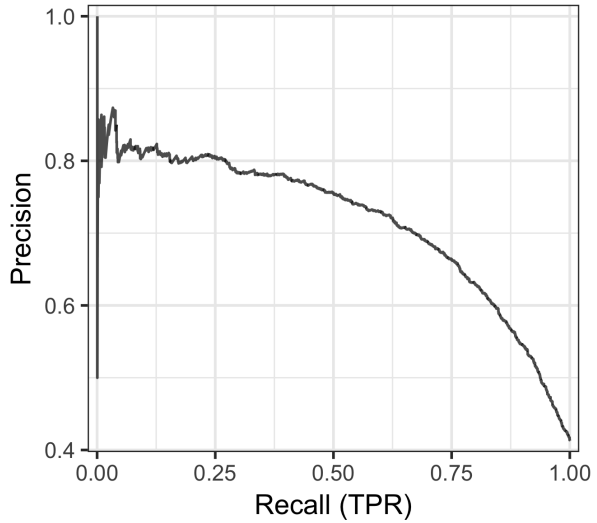


Classifier B has greater area and therefore better average performance.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

- **Precision Recall (PR) curve**
  - Alternative to ROC curves for tasks with a large skew in the class distribution
  - Looking at PR curves can expose differences between algorithms that are not apparent in ROC space
  - In PR space the goal is to be in the **upper-right-hand corner**



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

- **Precision Recall (PR) curve**

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

$$FP\ Rate = \frac{FP}{FP + TN}$$

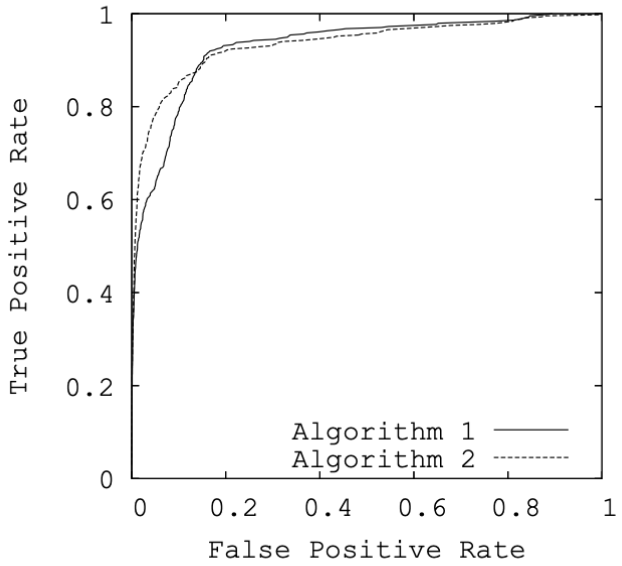
- If the number of negative examples greatly exceeds the number of positive examples:
  - a large change in the number of FP can lead to a small change in the FPR used in ROC
  - instead, precision captures the effect of the large number of negative examples on the algorithm's performance

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

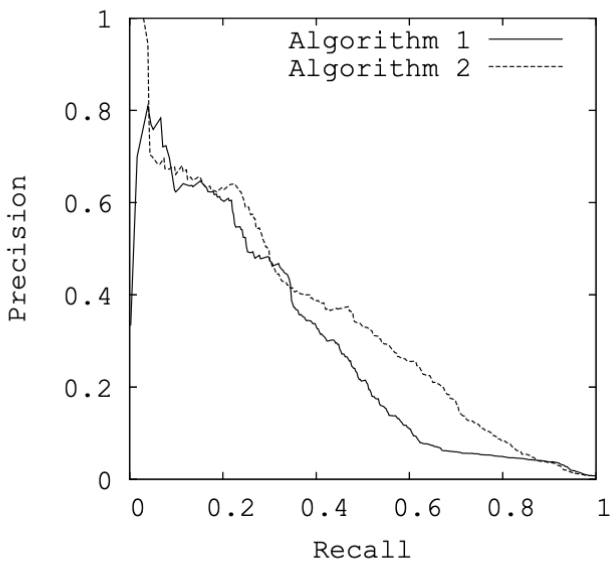
Metrics for ranking prediction

- **Precision Recall (PR) curve**

The 2 ROC curves appear to be fairly close to optimal



(a) Comparison in ROC space



(b) Comparison in PR space

In PR space the goal is to be in the upper-right-hand corner→ there is room for improvement

Figure 1. The difference between comparing algorithms in ROC vs PR space

Metrics for ranking prediction

Multi-class classifiers

- With n classes the confusion matrix becomes an n x n matrix containing the n correct classifications (the major diagonal entries) and n^2 –n possible errors (the off-diagonal entries)
- For instance, in case of ROC curves, instead of managing trade-offs between TP and FP, we have n benefits and n^2 –n errors. With only three classes, the surface becomes a 9-3 = 6-dimensional polytope

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

Multi-class classifiers

- Solutions
  1. **Micro-average:** Plot one PR/ROC curve for each class by summing *globally* the respective TP, FP, and FN values across all classes and average the curves.
  2. **Macro-average:** Plot one PR/ROC curve for each class considering each class *separately* (does not take into account class imbalance) and average the results.
  3. **No average:** Consider the curves without combining them.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

- Micro average: simulates a binary classification. What is correct vs what is not correct.
  - TP: labels correctly predicted → **TP = 4**
  - FP: labels predicted when they shouldn't → **FP = 5**
  - FN: same as FP → **FN = 5**

P = 0.444 **R = 0.444**

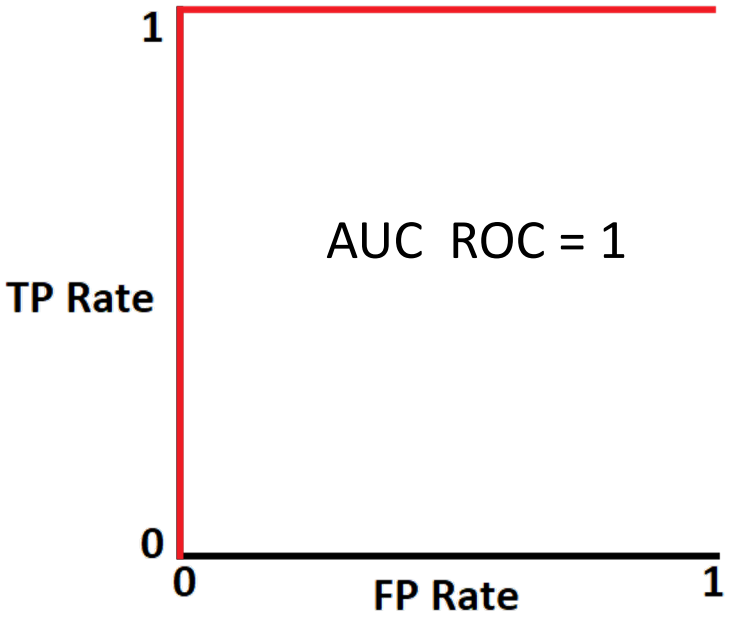
- Macro average
  - Class 1: TP = 0, FN = 2, FP = 2 → P = 0, R = 0
  - Class 2: TP = 3, FN = 1, FP = 2 → P = 3/5, R = 3/4
  - Class 3: TP = 1, FN = 2, FP = 1 → P = 1/2, R = 1/3

P = 0.367 **R = 0.361**

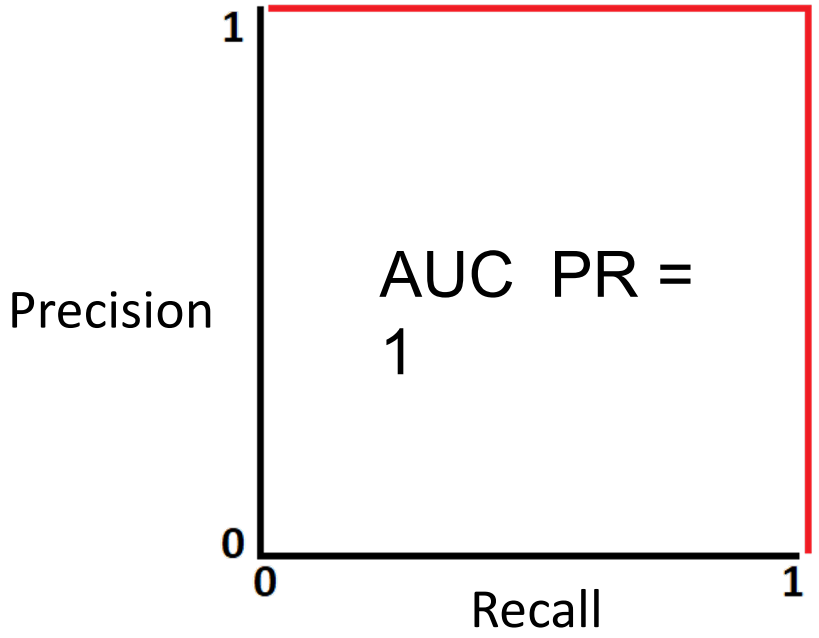
Label	1	2	3	2	3	3	1	2	2
Prediction	2	2	1	2	1	3	2	3	2

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

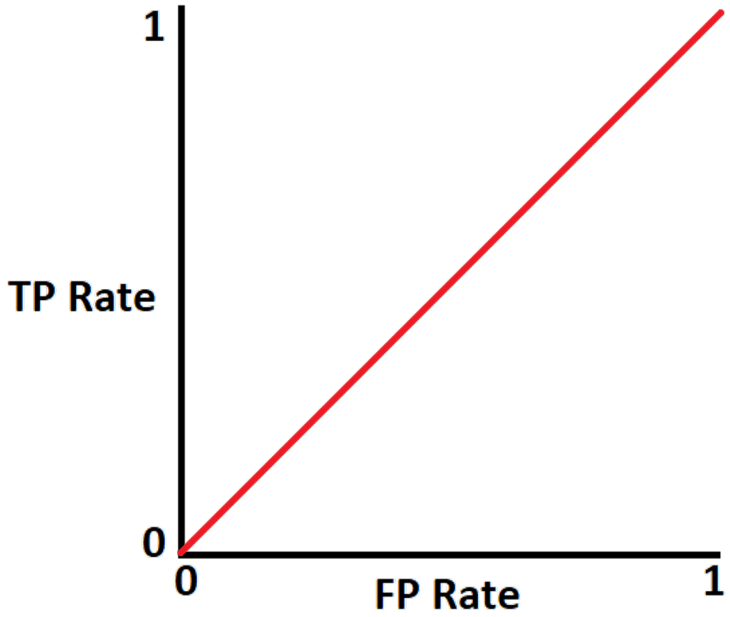


Best situation: the model can distiguish examples without errors.



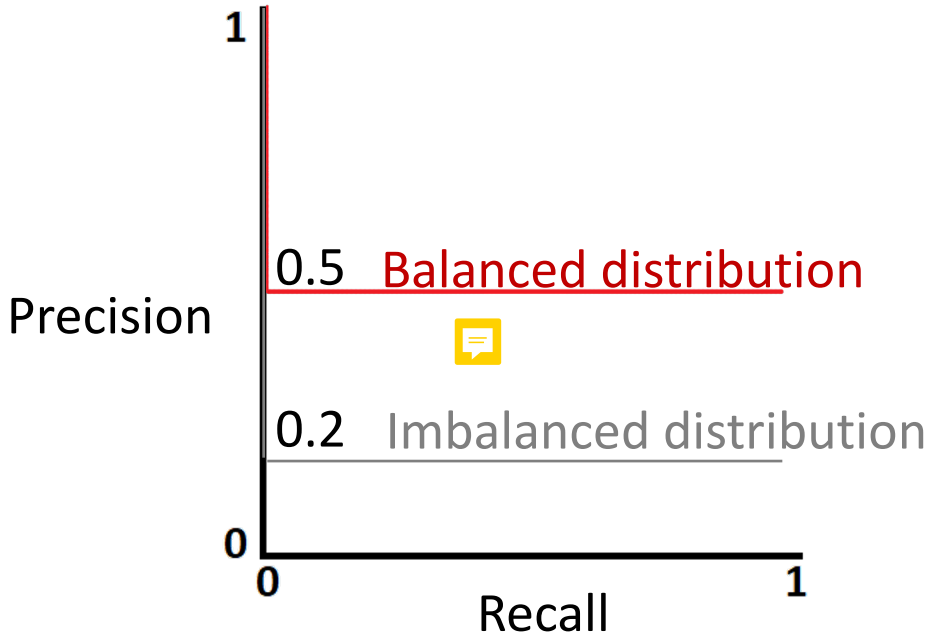
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction



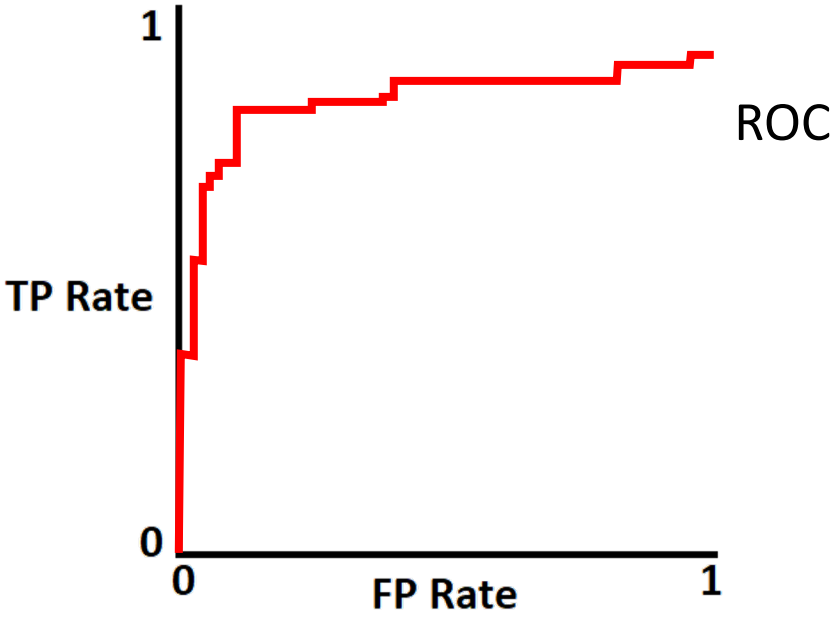
ROC
AUC = 0.5

Worst situation: the model is not able to distiguish the examples, it perform as random guess.

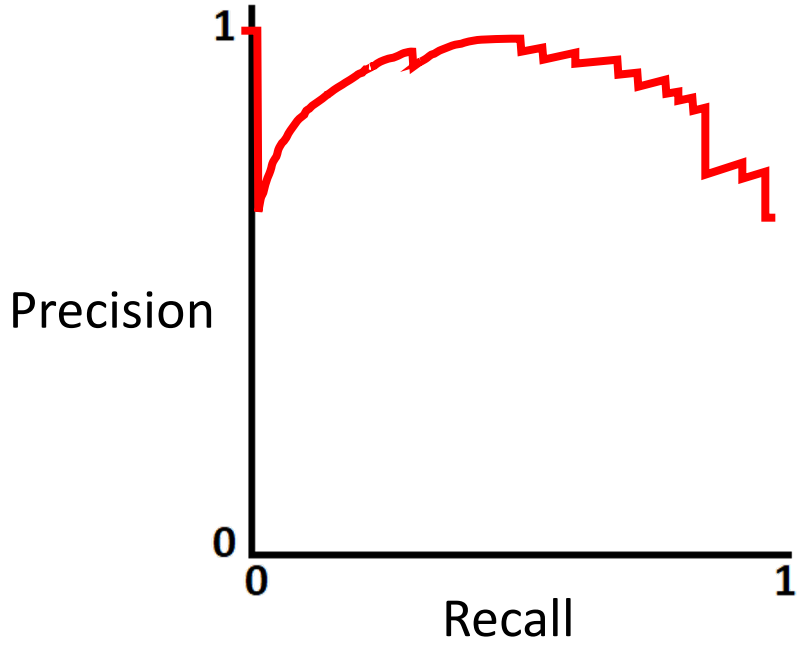


Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction



The distributions overlap.



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

- **With unbalanced data it becomes pretty easy for any model to correctly predict negatives.**
- Because the ROC curve in part plots false positive rates that are calculated with the resulting large number of true negatives in the denominator, by that metric we may seem to be doing pretty well.

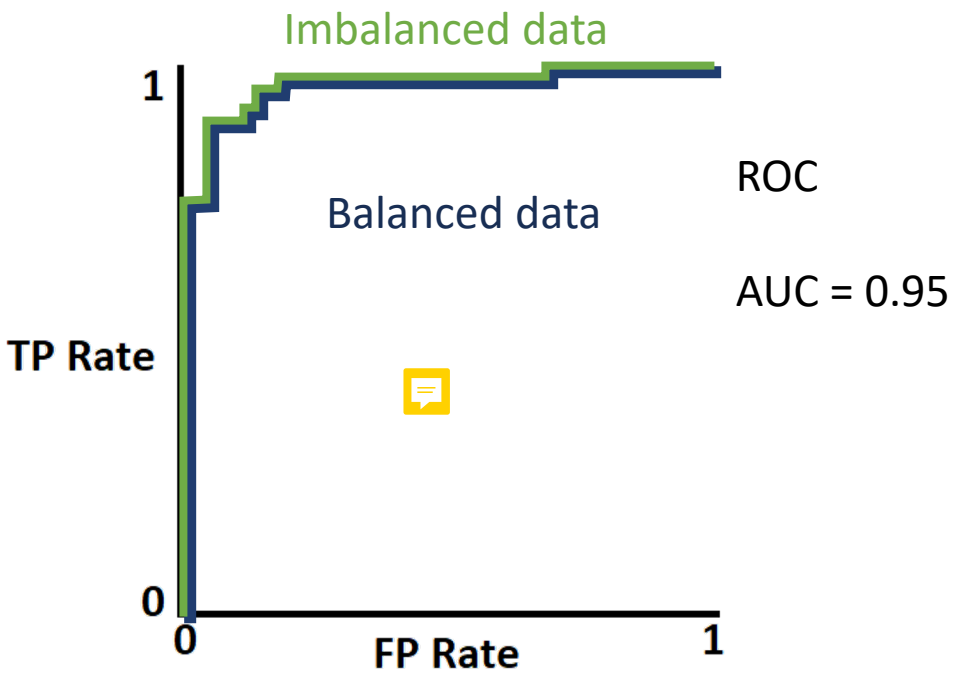
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction

- Reviewing both precision and recall is useful in cases where there is an imbalance in the observations between the two classes
  - More negative examples (class 0) and only a few positive examples (class 1).
  - In this case we are usually more interested in the capability of the model to correctly classify class 1 (remember e.g. imbalanced results of accuracy when the model classifies every example as 0).
  - **Precision and Recall do not use true negatives, but only concern with the correct prediction of the minority class 1.**

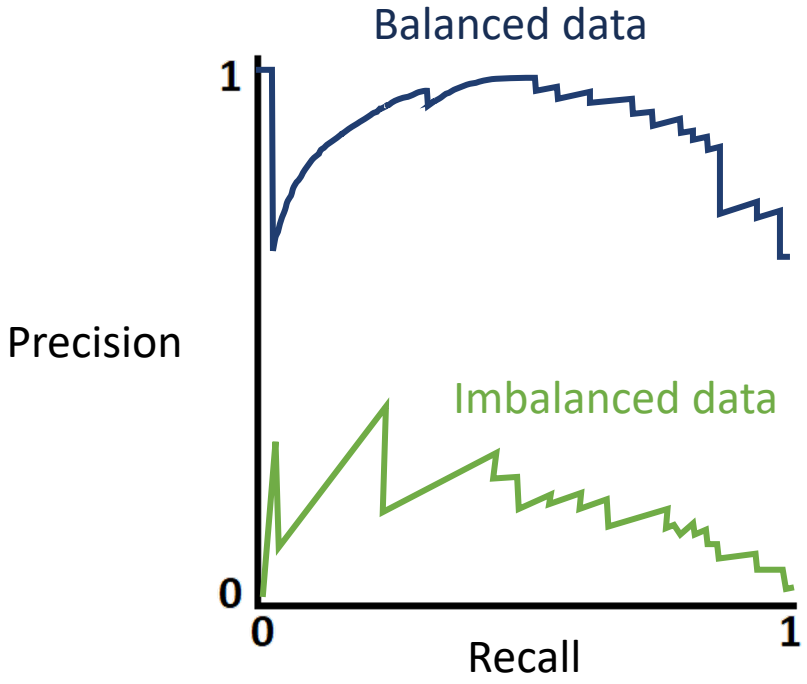
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for ranking prediction



Generally we have loss of precision as we move to more imbalanced data.

Consider two models having the same AUC ROC, one considering balanced data and one with imbalanced data.



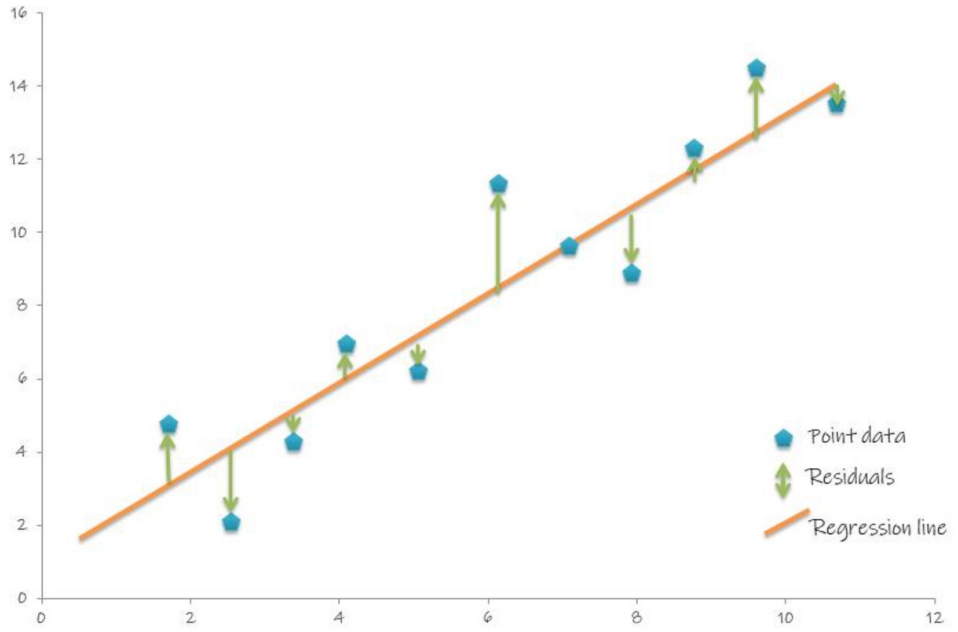
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for Regression

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for regression

- **Mean Squared Error**
  - It shows how far predictions fall from true values using Euclidean distance (*residuals or prediction errors*)
  - It tells how concentrated the data is around the line of best fit
  - It measures the variance of the residuals



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Regressions

- Some types of regressions:
  - Linear regression

$$\widehat{y_i} = E(X) = \beta X = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}$$

  - Polinomial regression

$$\widehat{y_i} = E(X) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \ldots + \beta_n x_i^n$$

  - Logistic regression

$$\widehat{y_i} = E(X) = \frac{1}{1 - e^{-(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in})}}$$

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for regression

- **Mean Squared Error (MSE)**
  - Computes the **average** of the **squared differences** between the original y_i and the predicted $\hat{y}_i$ value over all N examples in the test set

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

  - The use of the square ensures that larger errors are taken into account more than smaller errors
  - It is always non–negative and **close to zero is better**

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for regression

- **Root Mean Squared Error (RMSE)**

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

  - It measures the *standard deviation* of residuals
  - It is always non–negative and **close to zero is better**

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for regression

- **Mean Absolute Error (MAE)**
  - Computes the average of the absolute values of the difference between the original y_i and the predicted $\hat{y}_i$ over all instances in the test set
  - We use the absolute value of the distances so that negative errors are accounted properly

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

  - It is always non–negative and **close to zero is better**
  - It measures the *average* of the residuals in the dataset

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for regression

- MSE is a differentiable function that makes it easy to perform mathematical operations in comparison to a non-differentiable function like MAE
- In MSE higher errors (or distances) weigh more in the metric than lower ones, due to the nature of the power function
- In MSE the unit of the metric is also squared, so if the model tries to predict price in US$, the MSE will yield a number with unit $(US\$)^2$ which does not make sense
- RMSE is used then to return the MSE error to the original unit by taking the square root of it, while maintaining the property of penalizing higher errors
- For comparing the accuracy among different linear regression models, RMSE is a better choice

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021