Evaluation Metrics for Classification Models

Dr. Pei Xu
Auburn University

Outline

- Confusion Matrix
- Accuracy
- Sensitivity (Recall)
- Precision
- F1-Score
- ROC & AUC

Confusion Matrix

The performance of a classification model can be evaluated by comparing the predicted labels against the true labels of instances.

This information can be summarized in a table called a confusion matrix.

	true class: Positives	***true class***: Negatives
predicted as: positives	**True Positives (TP)**	**False Positives (FP)**
predicted as: negatives	**False Negatives (FN)**	**True Negatives (TN)**

T. Fawcett, *Introduction to ROC analysis*, Pattern Recognition Letters 27 (2006) 861. doi:10.1016/j.patrec.2005.10.010

Consider the Stock market dataset

True Positive - The model correctly predicted that the market went up.
Said another way, the model predicted that the label would be Positive, and that ended up being True.

True Negative - The model correctly predicted that the market went down.
Said another way, the model predicted that the label would be Negative, and that ended up being True.

False Positive - The model incorrectly predicted that the market went up even though the market went down.
Said another way, the model predicted that the label would be Positive, but that was False (the actual label was True).

False Negative - The model incorrectly predicted that the market went down even though the market went up.

Accuracy

- The simplest way to determine the effectiveness of a classification model is prediction accuracy.

- Accuracy helps us answer the question:
  What fraction of the predictions were correct (actual label matched predicted label)?

$$Accuracy = \frac{\#\ of\ Correctly\ Predicted}{\#\ of\ Observations}$$

- Regarding the Stock Market Dataset:
  What is the proportion of days that actually went up or went down?

Accuracy

Is accuracy an adequate metric?

Would you believe someone who claimed to create a model entirely in their head to identify terrorists trying to board flights with greater than 99% accuracy?

Sensitivity (also called recall)

- Sensitivity helps us answer the question:
  How effective is this model at identifying positive outcomes?

$$Sensitivity = True\ Positive\ Rate = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- Regarding the Stock Market Dataset:
  What is the ability to predict days that actually went up?

Specificity

- Specificity helps us answer the question:
  How effective is this model at identifying negative outcomes?

$$Specificity = True\ Negative\ Rate = \frac{True\ Negatives}{False\ Positives + True\ Negatives}$$

- Regarding the Stock Market Dataset:
  What is the ability to predict days that actually went down?

Precision

- Specificity helps us answer the question:
  How effective is this model at identifying only positive outcomes?

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positive}$$

- Regarding the Stock Market Dataset:
  What is the proportion of days that predicted as up days actually went up?

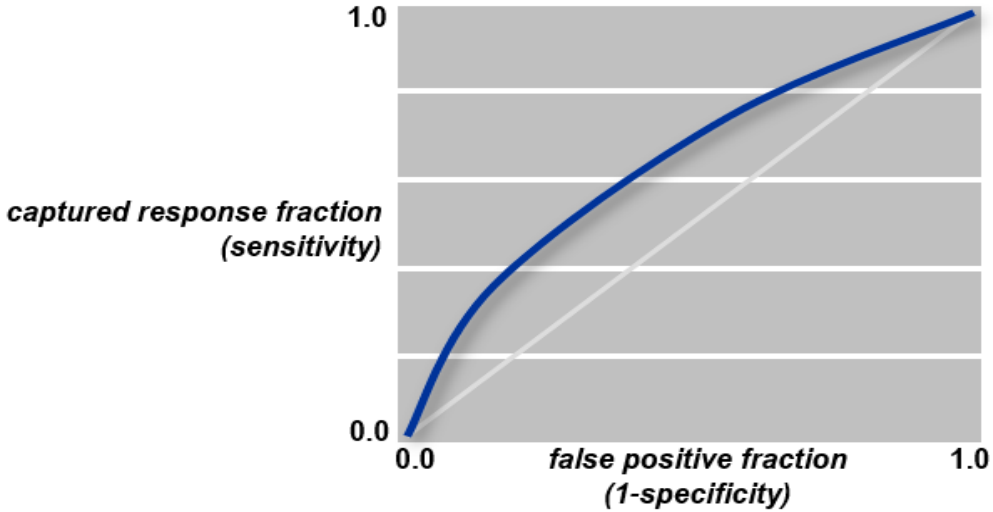
F1-Score

- F1-Score is the harmonic mean of precision and recall
- F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0

$$F1\text{-}Score = 2* \frac{Precision * Recall}{Precision + Recall}$$

ROC Curve

- ROC curve, is a graphical plot that illustrates the performance of a binary classifier system.
- The curve is created by plotting the true positive rate (*Sensitivity*) against the false positive rate (1- specificity), at different discrimination thresholds.



The ROC chart illustrates a tradeoff between a captured response fraction and a false positive fraction.

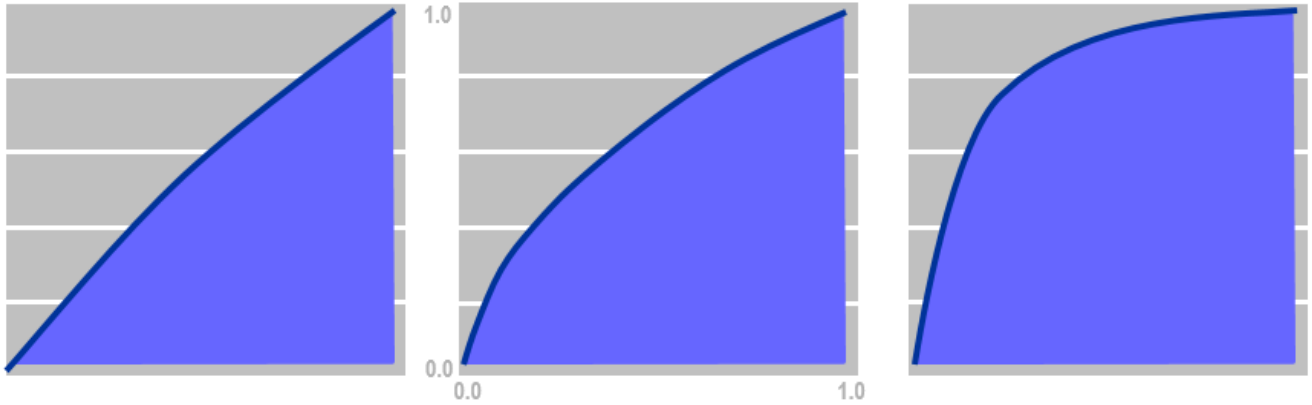
ROC Curve

If we evaluate thresholds from 0.0 to 1.0 in increments of 0.1, at each step calculating the precision, recall, F1, and location on the ROC curve. The classification outcomes at each threshold could be like this:

Threshold	TP	FP	TN	FN
0.0	50	50	0	0
0.1	48	47	3	2
0.2	47	40	9	4
0.3	45	31	16	8
0.4	44	23	22	11
0.5	42	16	29	13
0.6	36	12	34	18
0.7	30	11	38	21
0.8	20	4	43	33
0.9	12	3	45	40
1.0	0	0	50	50

AUC (Area under the ROC curve)

metric to calculate the overall performance of a classification model based on area under the ROC curve



– 1.0: perfect prediction
– 0.9: excellent prediction
– 0.8: good prediction
– 0.7: mediocre prediction
– 0.6: poor prediction
– 0.5: random prediction
– <0.5: something wrong!

Reference

- Textbook: introduction to Data Mining (Chapter 3)

- Beyond Accuracy: Precision and Recall.
https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c