Performance Metrics for ML

Asad Jeewa

DS-I Africa Machine Learning Short Course

23/01/2023

Day One

EDGEWOOD CAMPUS | HOWARD COLLEGE CAMPUS | NELSON R MANDELA SCHOOL OF MEDICINE | PIETERMARITZBURG CAMPUS | WESTVILLE CAMPUS

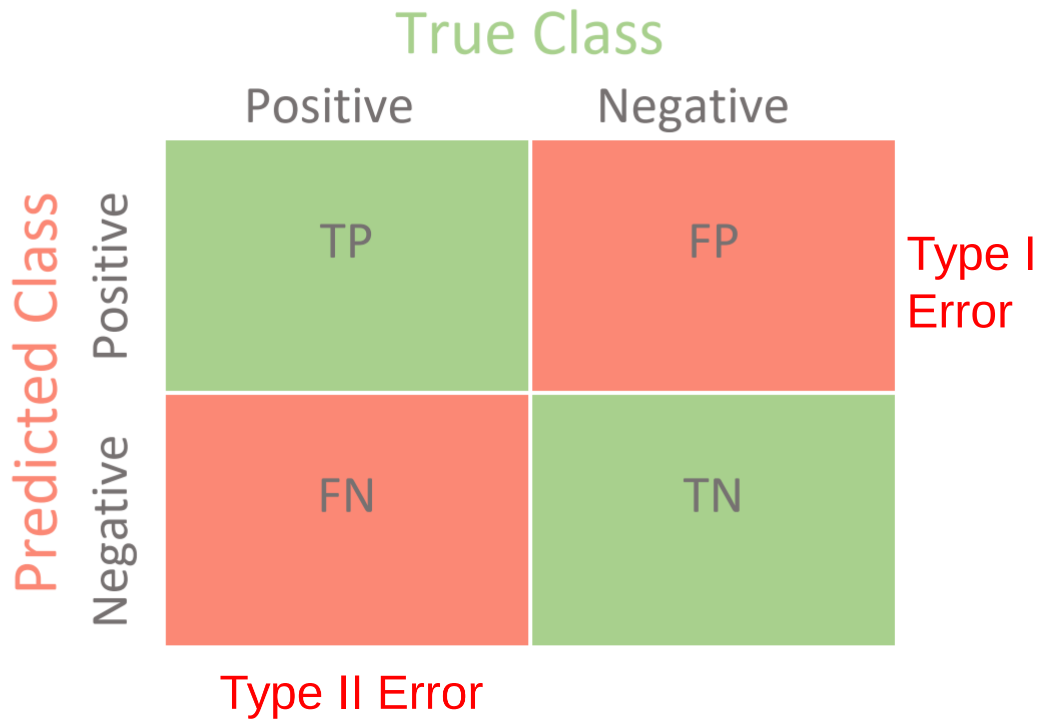
How effective is my model?

- Choice of metric is important
  - Confusion Matrix
  - Accuracy
  - Precision
  - Recall/ Sensitivity
  - Specificity
  - F1 Score
  - AUC (Area Under Curve)

  - MAE (Mean Absolute Error)
  - MSE (Mean Squared Error)

Confusion Matrix

- 1: a person has Cancer; 0: a person does not
- Many metrics are based on the CM
- Minimisation depends on use case
- Extended for multiclass classification

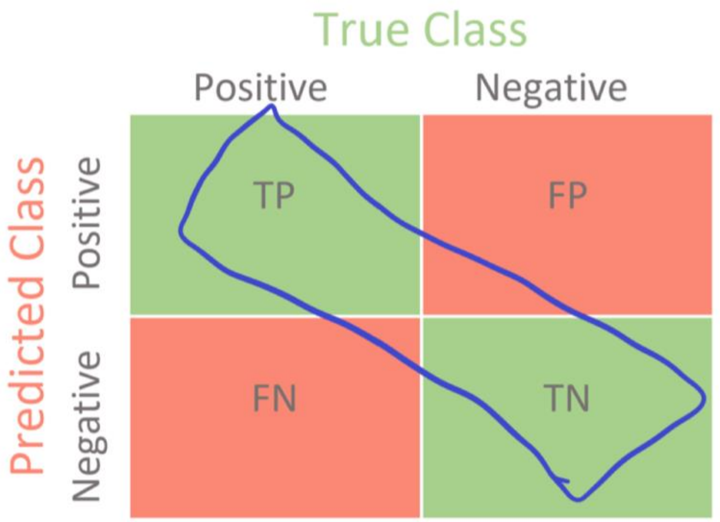


https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

Accuracy

- "How much did I get correct?"
- Useful when dataset labels are balanced: not the case in the real world!

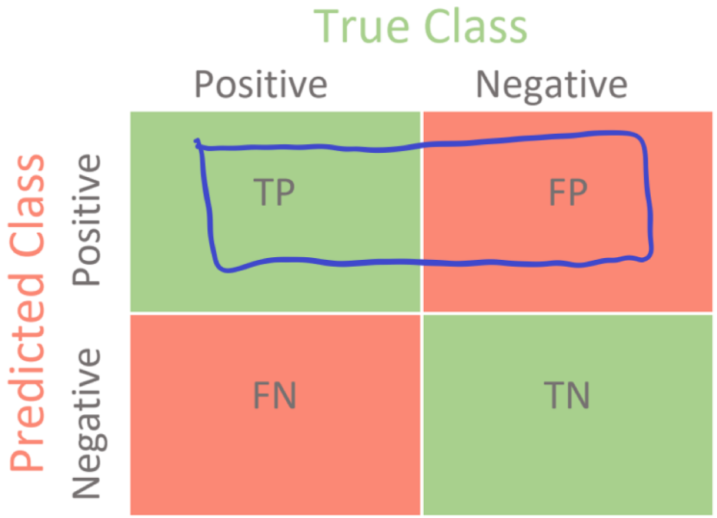
$$\frac{TP + TN}{TP + FN + TN + FP}$$



Precision

- Proportion of patients predicted to have cancer, that actually have cancer
- “How much did we catch”
- Goal: Minimise False Positives

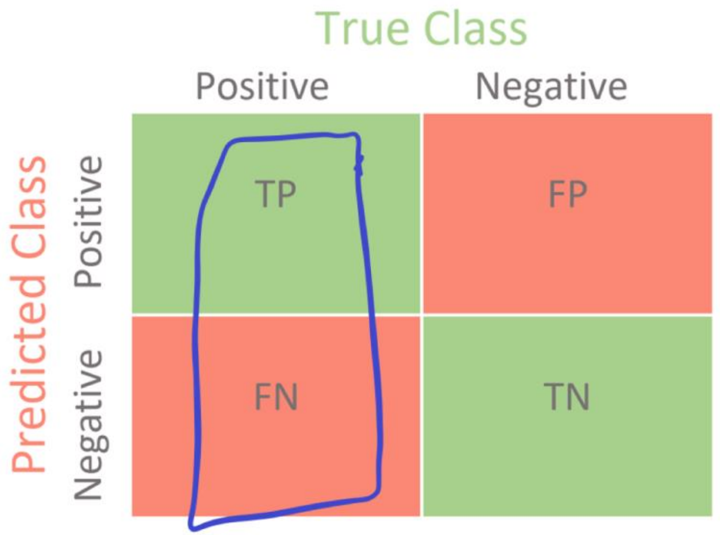
$$\frac{TP}{TP + FP}$$



Recall

- Proportion of patients that actually have cancer, predicted by model
- “How much did we miss”
- Goal: Minimise false negatives

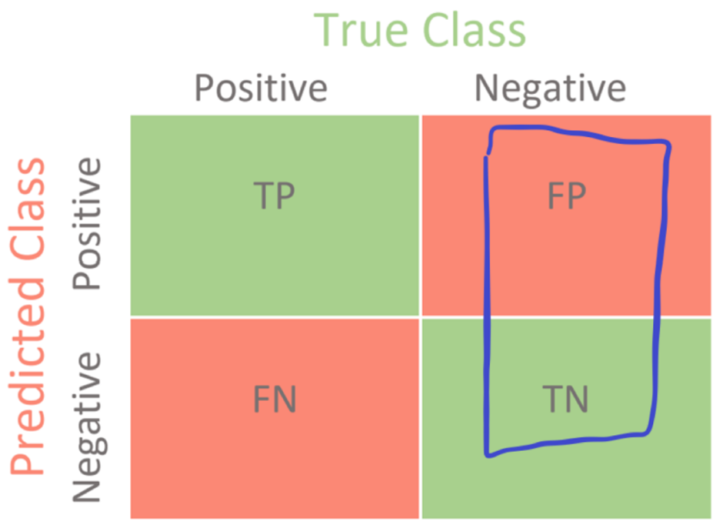
$$\frac{TP}{TP + FN}$$



Specificity

- Proportion of patients that actually do NOT have cancer, predicted by model
- Exact opposite of Recall

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



F1 Score

- Difficult to compare models with low precision, high recall (vice versa)
- F1-score: best of both worlds
- Harmonic Mean to address precision-recall imbalance
- Punishes the extreme values more

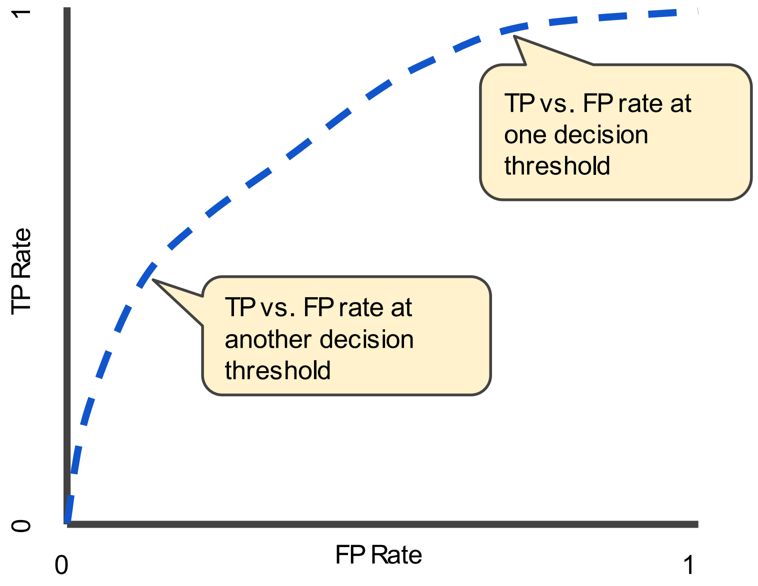
$$\frac{2 \times precision \times recall}{precision + recall}$$

ROC Curve (Receiver operating characteristic)

- Recall: Logistic Regression returns a probability
- Interpreted with a Classification/ Decision Threshold
- Threshold is problem dependent
- **TPR vs FPR at different classification thresholds**

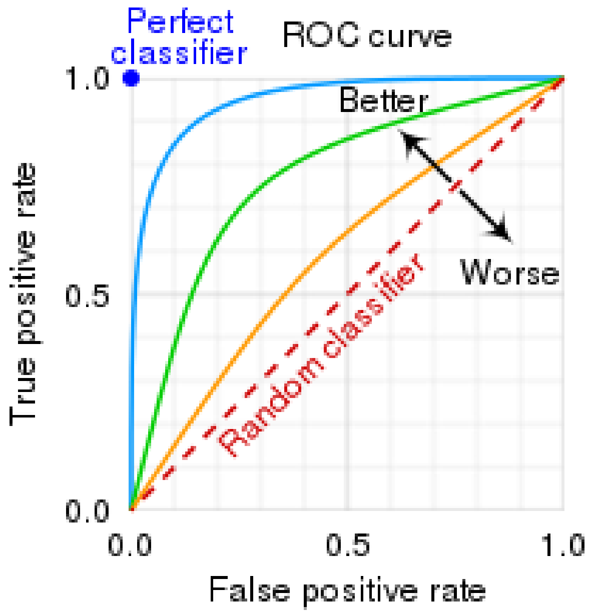
$$TPR = \frac{TP}{TP + FN} = Recall$$

$$FPR = \frac{FP}{FP + TN} = 1 - Specificity$$

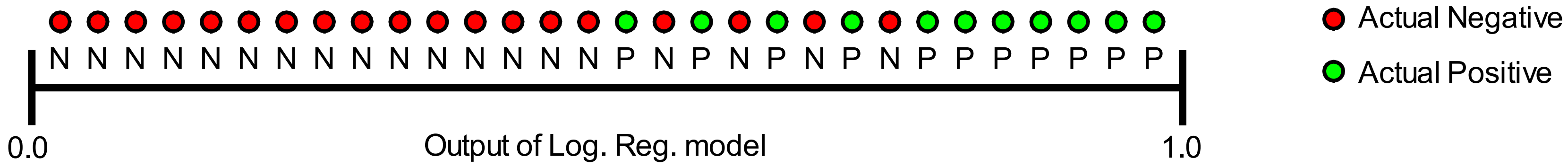


AUC: Area under ROC curve

- Lowering classification threshold classifies more items as positive
- Model performance across all classification thresholds
- Degree of separability of model
- AUC = 1: 100% correct predictions



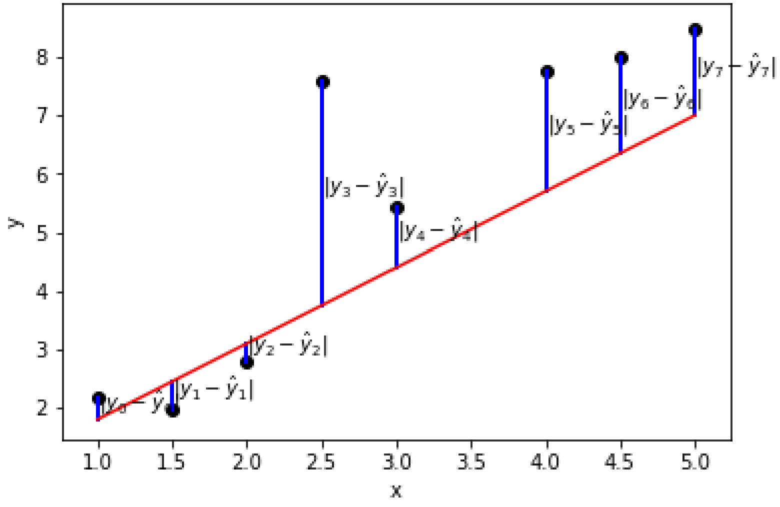
"Probability that the model ranks a random positive example more highly than a random negative example"



Mean Absolute Error

- Used for regression models
- Average difference between original and predicted values
- No indication of direction of error

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



https://jmlb.github.io/flashcards/2018/07/01/mae_vs_rmse/

Mean Squared Error

- Effect of larger errors are more pronounced
- RMSE: Root MSE: useful when larger errors are undesirable

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

CASE 1: Evenly distributed errors

ID	Error	|Error|	Error^2
1	2	2	4
2	2	2	4
3	2	2	4
4	2	2	4
5	2	2	4
6	2	2	4
7	2	2	4
8	2	2	4
9	2	2	4
10	2	2	4

MAE	RMSE
2.000	2.000

CASE 2: Small variance in errors

ID	Error	|Error|	Error^2
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	3	3	9
7	3	3	9
8	3	3	9
9	3	3	9
10	3	3	9

MAE	RMSE
2.000	2.236

CASE 3: Large error outlier

ID	Error	|Error|	Error^2
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	20	20	400

MAE	RMSE
2.000	6.325

https://www.slideshare.net/kushkul/performance-metrics-for-machine-learning-algorithms

Resources

- https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative
- https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide
- https://www.kdnuggets.com/2018/04/right-metric-evaluating-machine-learning-models-1.html

Discussion:

What makes a good ML model?