Lecture 10: Accuracy metrics and model selection

Why do we, as data scientists, care about the workflow and implementation of the model?

There is a relationship between the implementation workflow (how the model is going to be used), and the accuracy of the model you're implementing.
(how to measure it, and what the threshold is)

Today’s class

- Build a baseline model!
- Accuracy metrics
  - Regression, Classification
  - Risk models

Building a model: start with a baseline

1) The simplest baseline you can think of
2) A slightly more complex model
3) A slightly more complex model

Example

We are trying to create a model to predict 20-year survival of patients with leukemia.

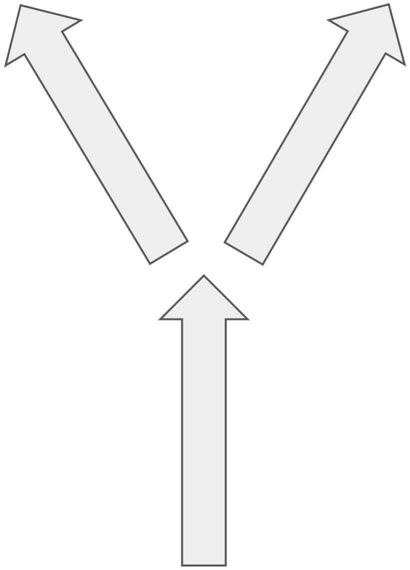
What the most simple model you can think of?

What's the next one in complexity?

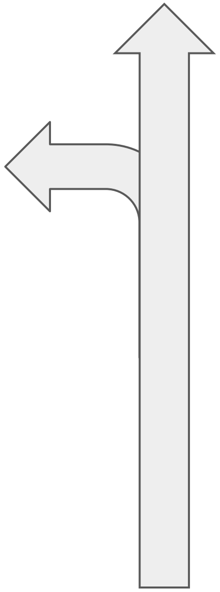
What could be your ideal model?

Considerations related to metrics

What type of decision are you trying to influence?

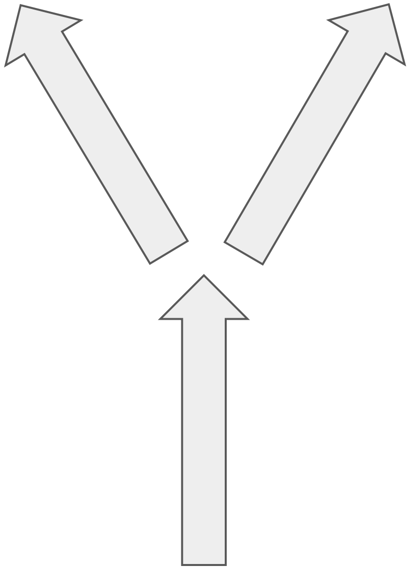


user is stopping to make a decision



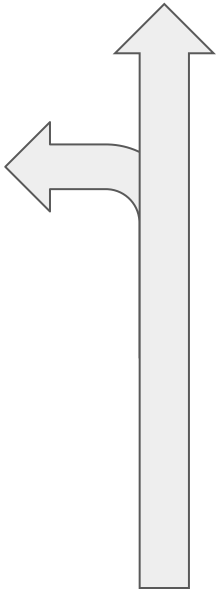
you are interrupting the user’s workflow (e.g. alert)

What type of decision are you trying to influence?



user is stopping to make a decision

(user will see negative and positive results)



you are interrupting the user’s workflow (e.g. alert)

(user will only see positive results)

What is the decision in each direction?

What is the cost of each mistake?

Additional considerations (1)

The model doesn't work in isolation. It works in conjunction with people.

Model accuracy: 85%

    Model accuracy: 90%

Human accuracy: 80%

    Human accuracy: 80%

Combined accuracy?

    Combined accuracy?

Additional considerations (2)

Model accuracy will change through time, as we see changes to:

1) The outcome of interest
2) The underlying data
3) The population that the model is applied to
4) New knowledge (about context, data, etc.)

This is called **drift** (model drift, data drift, concept drift...)

Accuracy metrics

Accuracy metric types

Regression

Classification

Regression metrics

Absolute measures
(with unit)

Mean square error (MSE, RMSE)

Mean Absolute error

Relative measures
(unitless)

Normalized RMSE

Mean absolute percent error

R-squared

Classification metrics

Precision vs recall

Accuracy

ROC curve

What is it?

Remember the 2x2 table and sensitivity, specificity, positive predictive value, etc.?

Try to recreate it on paper

Classification metrics

Metrics **independent** of prevalence

Sensitivity (recall)

Specificity

Measures **dependent** on prevalence

Positive Predictive Value (precision)

Negative Predictive Value

ROC curve

PR curve

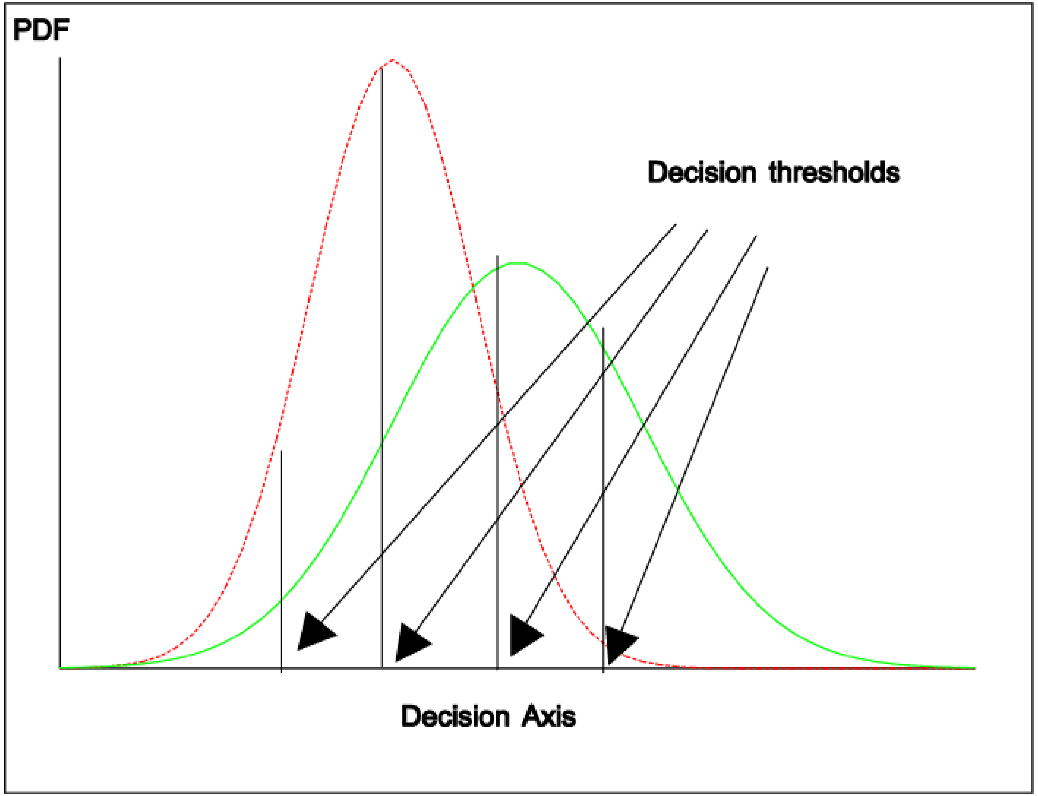
Traditionally...

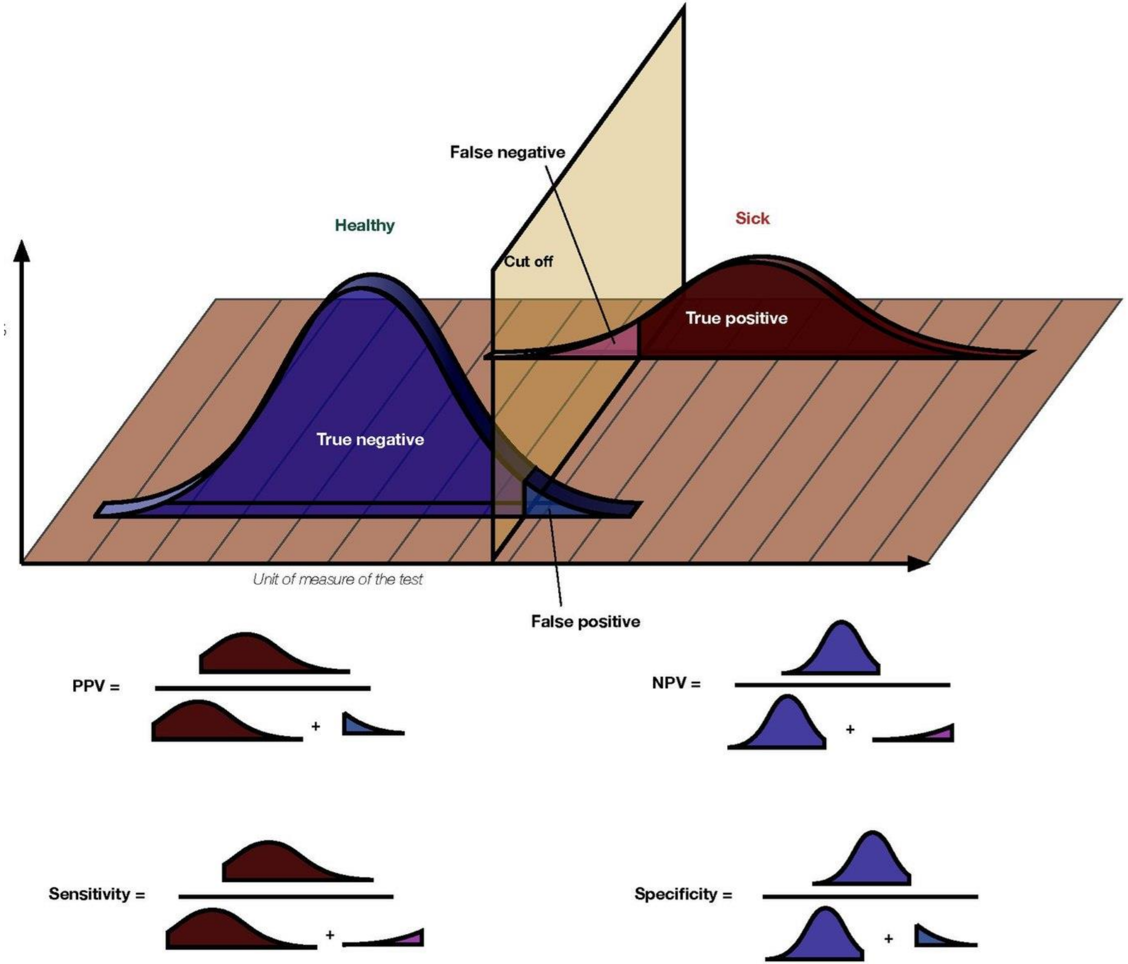
Tests were ordered by physicians, so it was hard to know the base prevalence of the population in which they were used.

This made prevalence-independent metrics generally preferable.

ML models are usually automated, so we know the base prevalence.

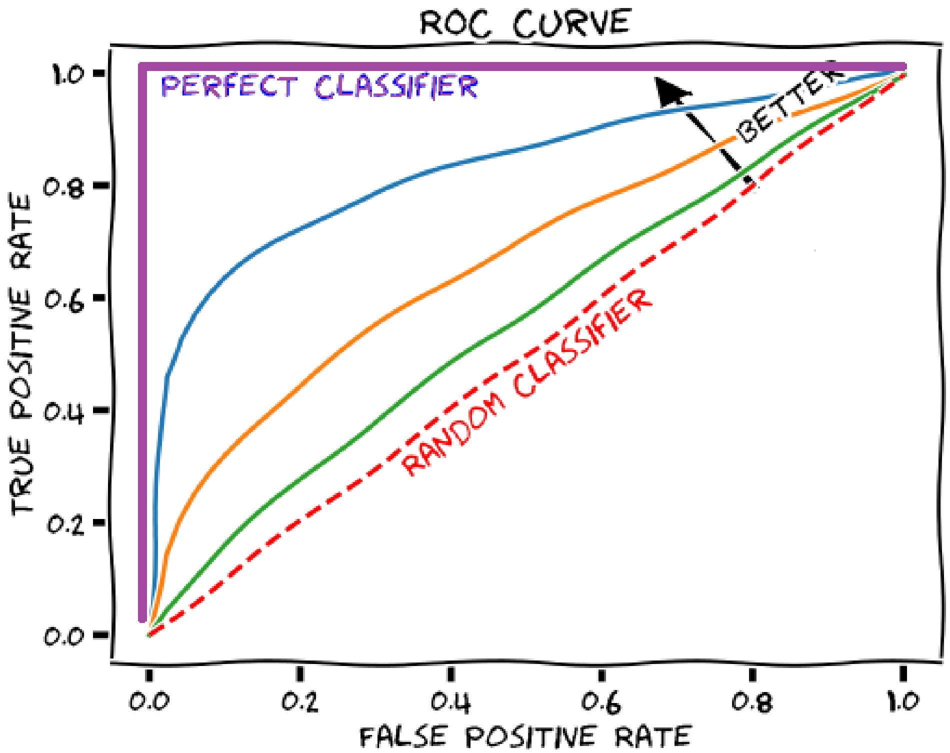
How do we calculate an ROC curve?



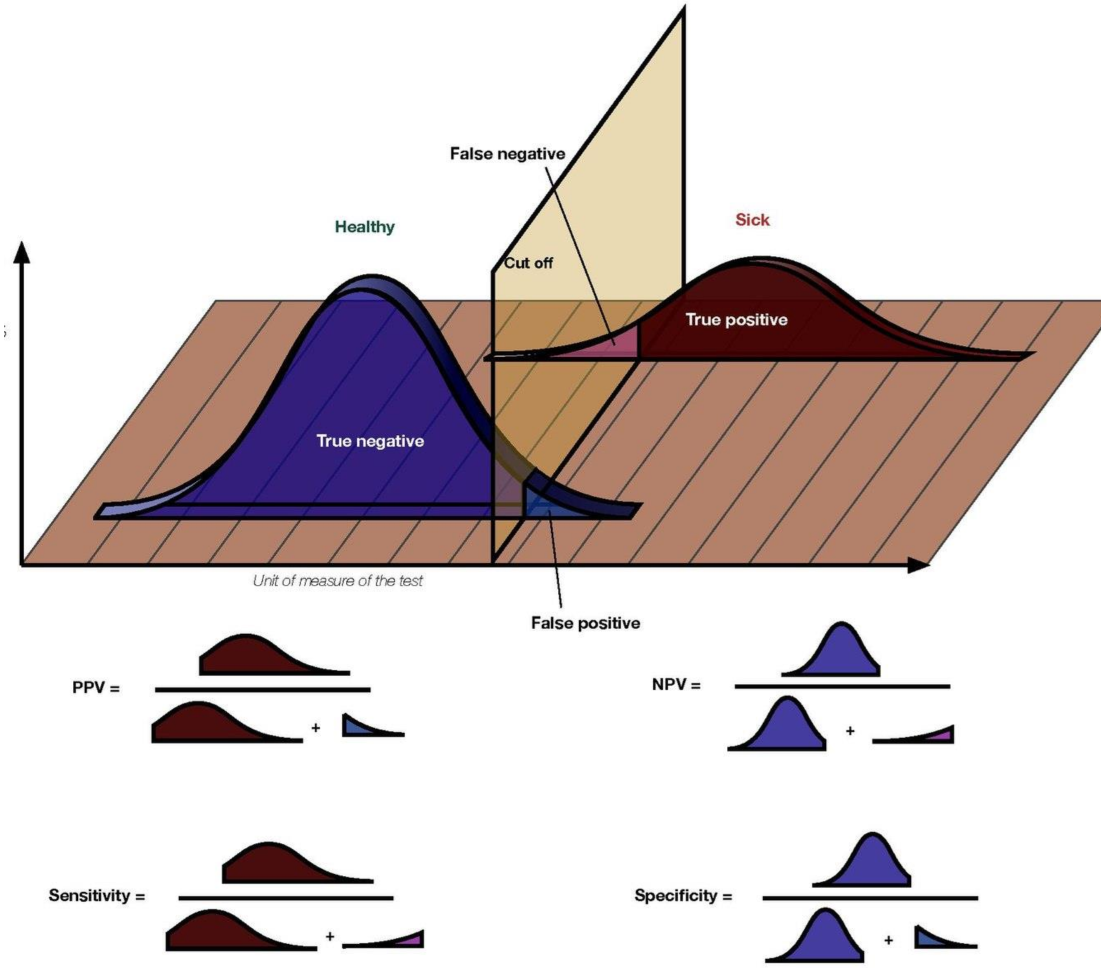

False negative
Healthy
Sick
Cut off
True positive
True negative
Unit of measure of the test
False positive
PPV =
NPV =
Sensitivity =
Specificity =

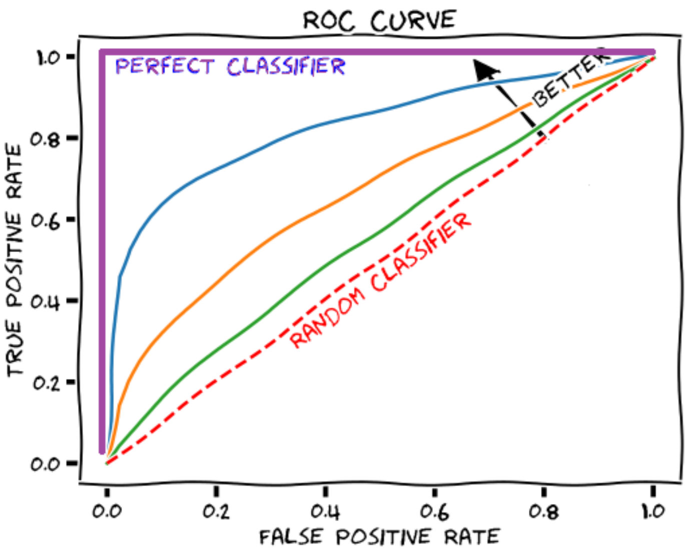
ROC

Sensitivity

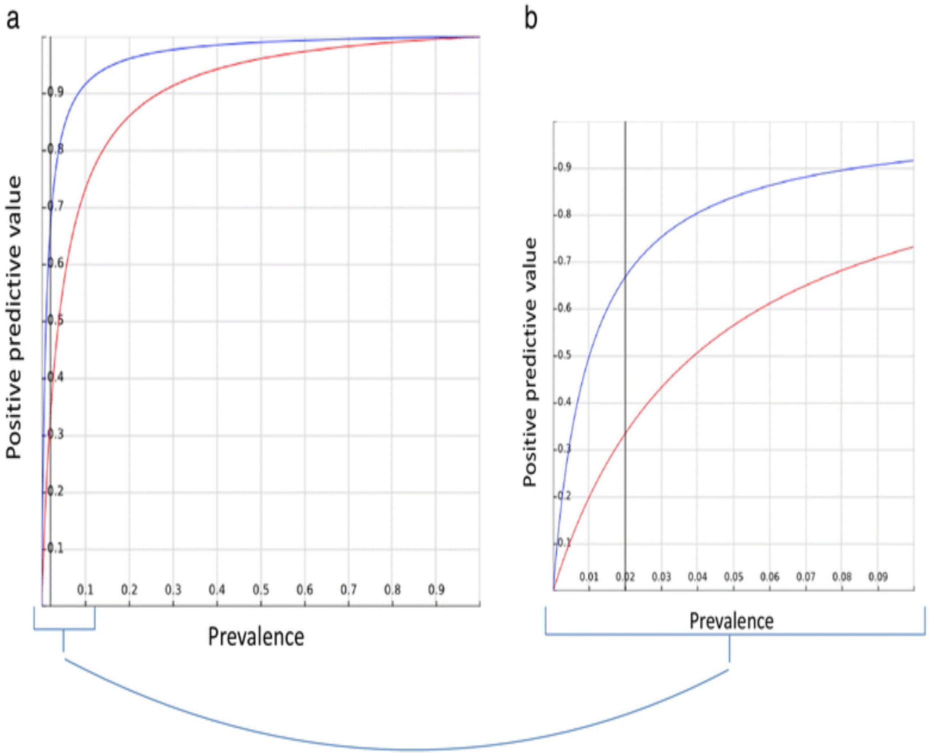


1 - Specificity


False negative
Healthy
Sick
Cut off
True positive
True negative
Unit of measure of the test
False positive
PPV =
NPV =
Sensitivity =
Specificity =


ROC CURVE
PERFECT CLASSIFIER
BETTER
RANDOM CLASSIFIER
TRUE POSITIVE RATE
FALSE POSITIVE RATE
0.0
0.2
0.4
0.6
0.8
1.0

The problem with low incidence



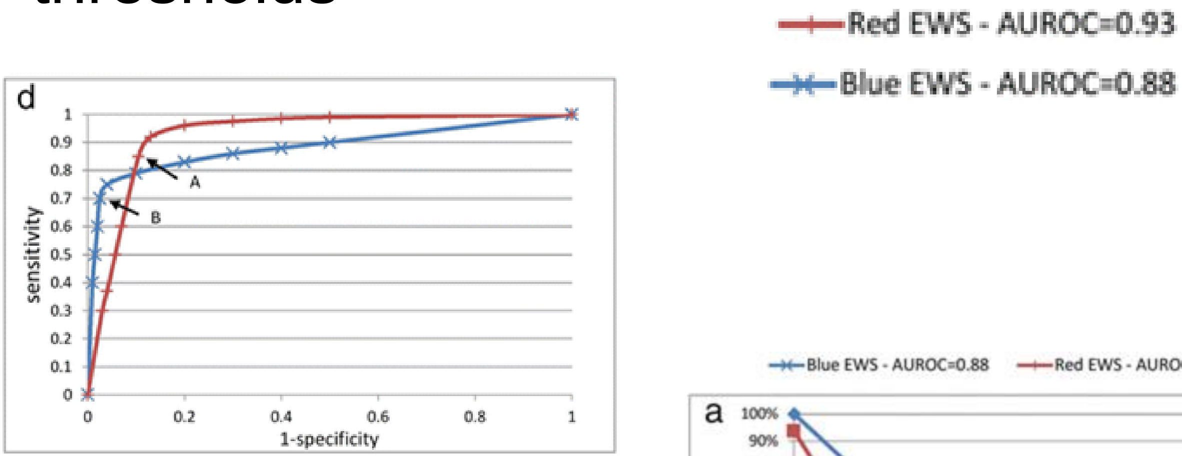
Blue:
Sens 99%,
spec 99%

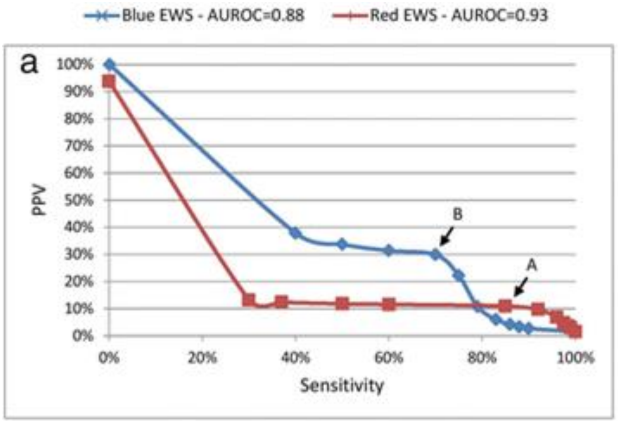
Red:
Sens 99%,
spec 9**6**%

1. Romero-Brufau, Santiago, et al. "Why the C-statistic is not informative to evaluate early warning scores and what metrics to use." *Critical Care* 19.1 (2015): 285.

1. **Metric selection**
2. Context-based data quality
3. Model selection
4. Human-computer synergy
5. System redesign
6. Change management
7. Measuring effect

Need to evaluate all EWS thresholds





1. **Metric selection**
2. Context-based data quality
3. Model selection
4. Human-computer synergy
5. System redesign
6. Change management
7. Measuring effect

So what do we do to compare models?

Partial ROC

Selecting a range, manually comparing performance