neue fische

School and Pool for Digital Talent

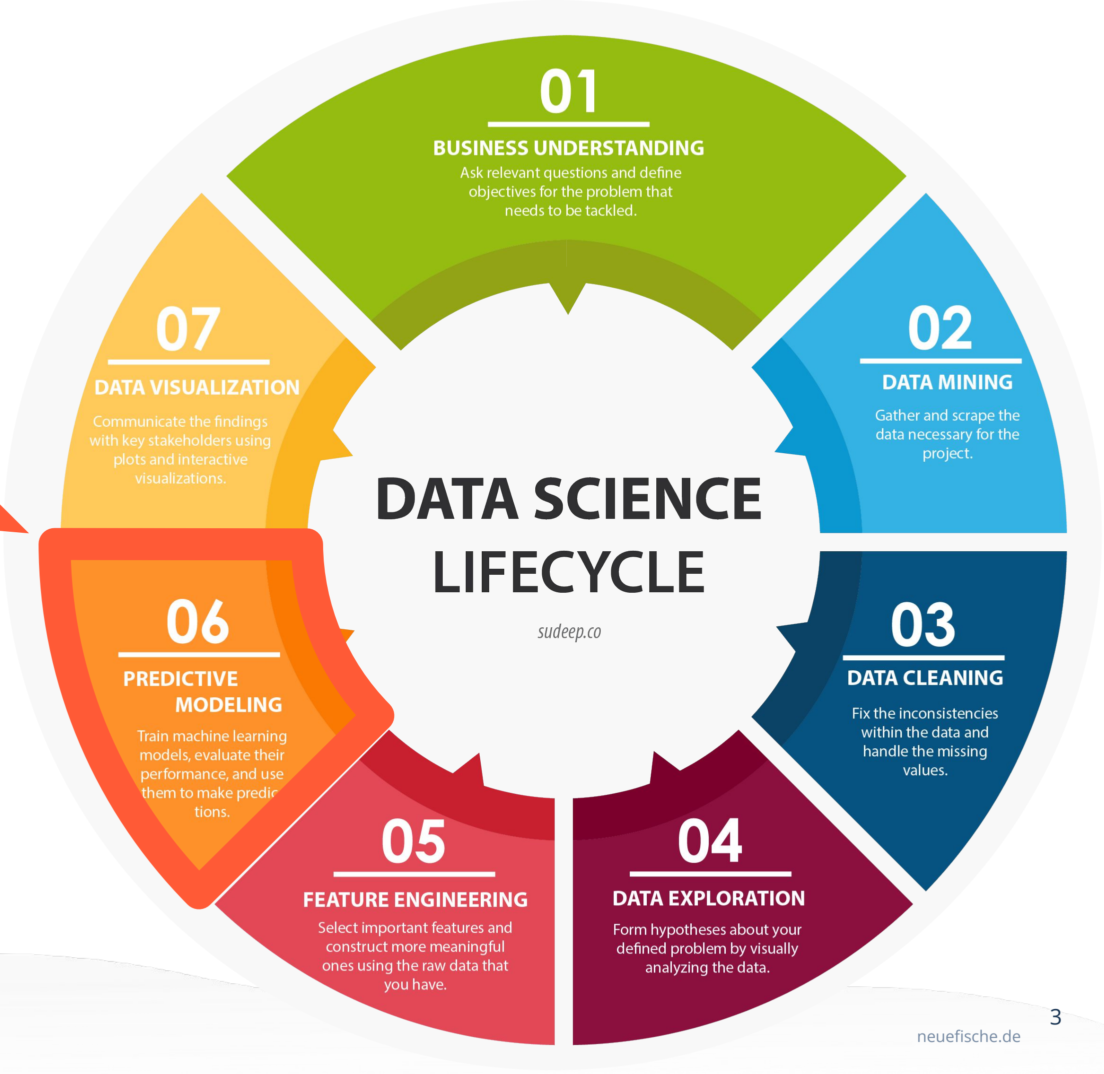
Evaluation Metrics

Orientation

Where are we now?

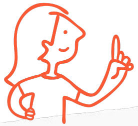
06 PREDICTIVE MODELING:

- select a ML algorithm
- train the ML model
- evaluate the performance
- make predictions



Evaluate model performance

- Regression Metrics (*R-squared, RMSE...*)
- Classification Metrics (*accuracy, precision, recall...*)
- Custom Metrics
  → e.g. based on the worst case scenarios of your product



If you need to present to stakeholders you need a simple metric!
MSE, precision, recall... are too complex to explain

A SHORT RECAP ON REGRESSION METRICS…

What Metrics do we have to evaluate model performance?

- R^2 (R-squared)

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- MSE (**M**ean **S**quare **E**rror)

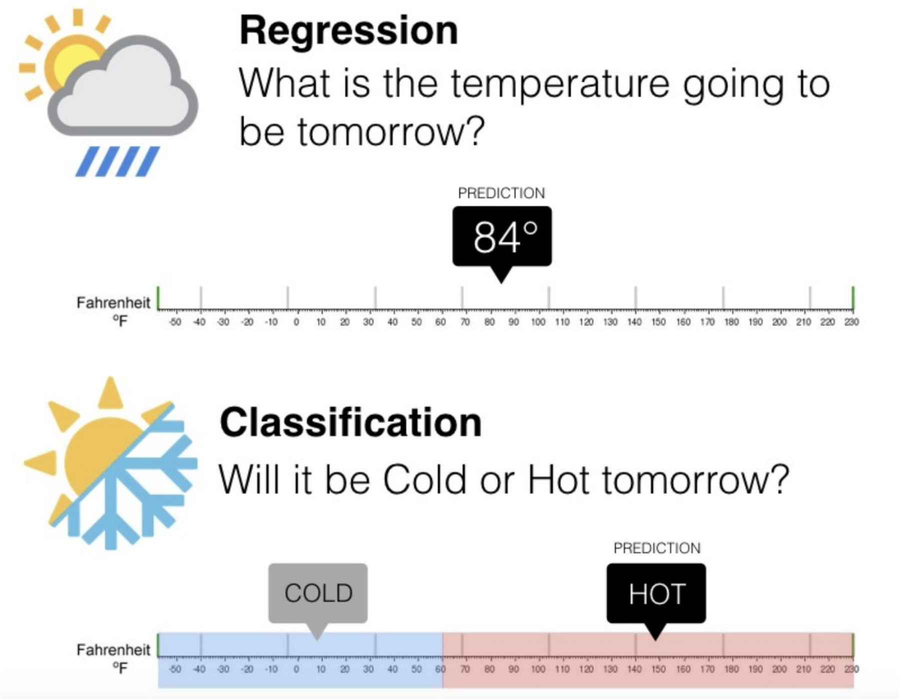
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- RMSE (**R**oot **M**ean **S**quare **E**rror)

$$RMSE = \sqrt{MSE}$$

LET’S TALK ABOUT CLASSIFICATION…

Binary classification



https://towardsdatascience.com/regression-or-classification-linear-or-logistic-f093e8757b9c

Definition

Confusion Matrix

Counts how often the model predicted correctly and how often it got confused

- False Positive: false alarm / type I error
- False Negative: missed detection / type II error

		Predicted	
		Negatives	Positives
Actual	Negatives	TN	FP
	Positives	FN	TP

Definition

- How often the model has been right

$$\frac{Correct}{All} = \frac{TP+TN}{TP+FP+TN+FN}$$

		Predicted: Negatives	Predicted: Positives
Actual	Negatives	TN	FP
	Positives	FN	TP

Drawbacks

- When one class is very rare it leads to false conclusions
- Here, Accuracy is 94 %
- But 5 out of 6 positives have been predicted incorrectly

Predicted

		Negatives	Positives
Actual	Negatives	93	1
	Positives	5	1

Accuracy might not be good enough

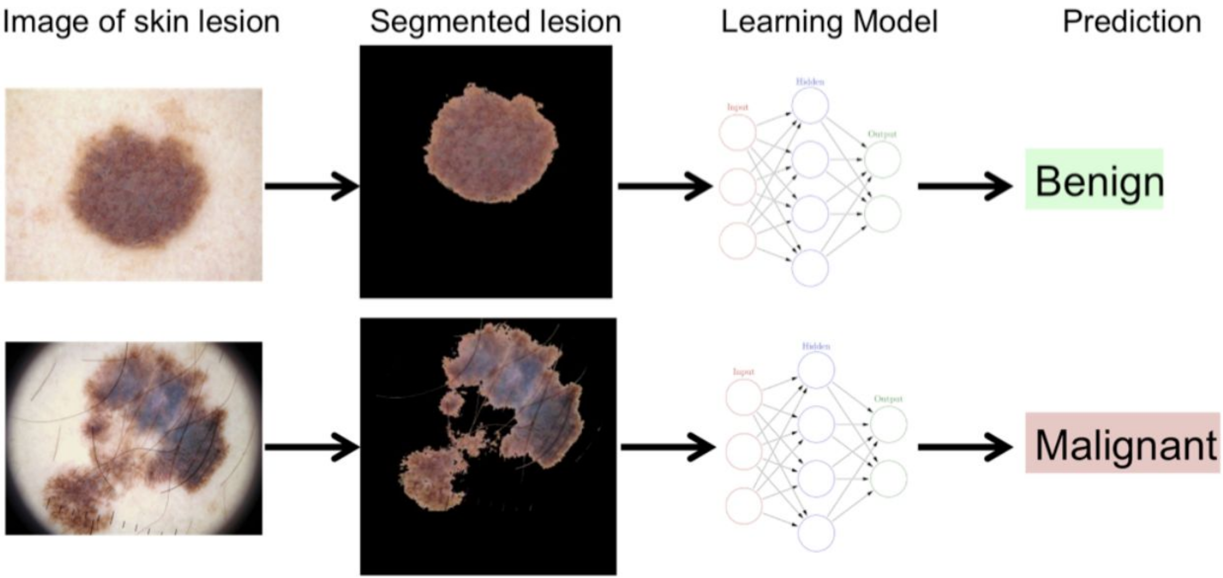


Image source

Precision and Recall

- Focus on positive class
- Number of True Negatives are not taken into account
- When trying to detect a rare event
- The number of negatives is very large

What proportion of actual positives was predicted correctly?

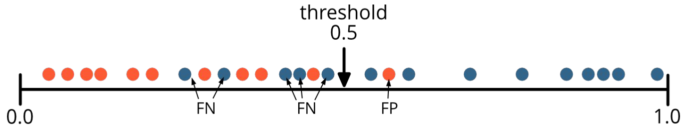
- The TPR is also called *sensitivity* or *recall*
- Here, the True Positive Rate is ⅙ (~16.67%)

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

		Predicted	
		Negatives	Positives
Actual	Negatives	93	1
	Positives	**5**	**1**

Tweaking the model

- Every model has a **threshold** that discerns positive from negative predictions
- Typically, instances will get predicted positive if the probability for that is larger 0.5



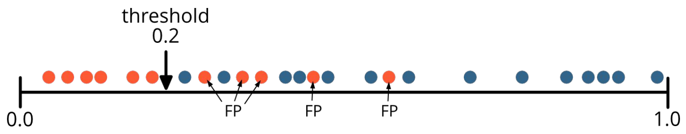
Before tweaking $$\frac{TP}{TP+FN}$$

Predicted

Actual		Negatives	Positives
	Negatives	93	1
	Positives	5	1

Now let's tweak the model

- The lower the threshold the more instances get predicted positive
- This will automatically raise the True Positive Rate (TPR)



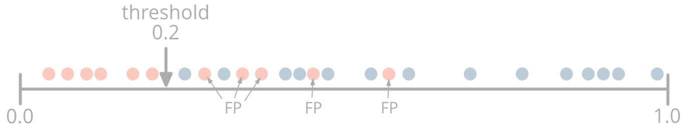
Before tweaking $$\frac{TP}{TP+FN}$$

Predicted

		Negatives	Positives
Actual	Negatives	93	1
	Positives	5	1

Now let's tweak the model

- Now, the True Positive Rate (TPR, recall) is at 100%
- But are we entirely happy?



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

Predicted

Actual		Negatives	Positives
	Negatives	14	80
	Positives	0	6

Precision

What proportion of positive predictions are actually correct?

- Here, Precision is 6/86 (~6.97 %), even lower as recall before tweaking!
- But: if it's too low or acceptable depends on the business case
- For detecting cancer it might be okay for the stakeholders
$\rightarrow$ still, costs for screening millions of people might be very high

After tweaking

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

Predicted

		Negatives	Positives
Actual	Negatives	14	**80**
	Positives	0	**6**

Summary

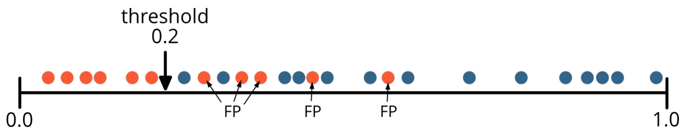
N_+ : the number of positives
N_- : the number of negatives
n = # observations

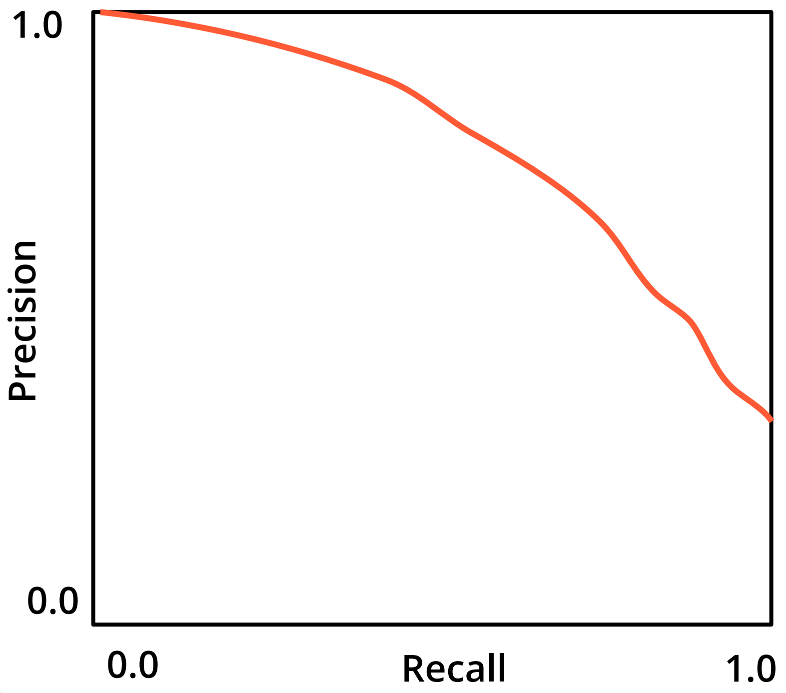
		Predicted		
		Negatives / 0	Positives / 1	∑
Actual	Negatives / 0	TN	FP	N_- = FP + TN
	Positives / 1	FN	TP	N_+ = TP + FN
	∑	$\hat{N}_-$ = FN + TN	$\hat{N}_+$ = TP + FP	n = TP + FP + FN + TN

Trade-off

Precision-Recall Curve

- Plots Precision vs. Recall depending on the **threshold**
- If threshold is high:
  → Precision is close to 1
  → Recall will be very low





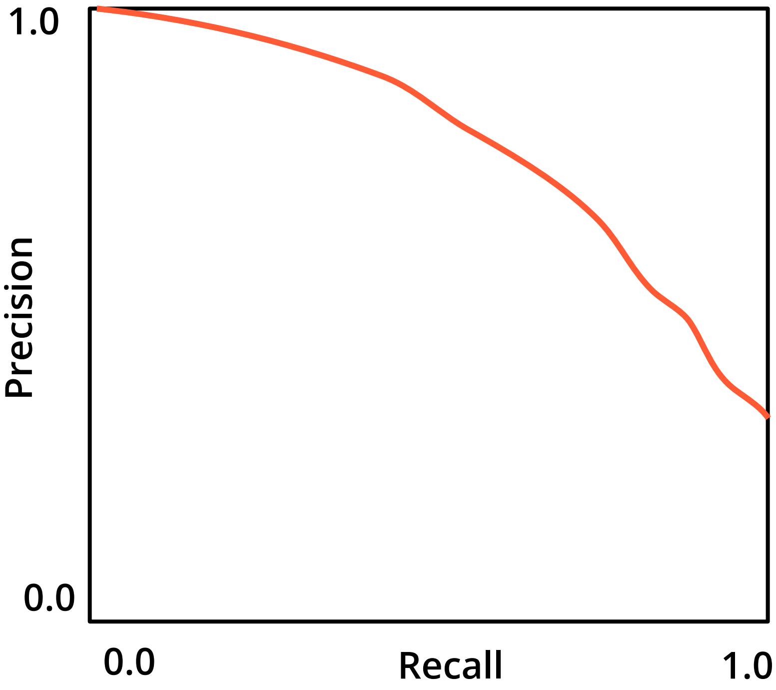
Trade-off

Precision-Recall Curve

- If threshold is effectively zero:
  → predicting all instances as positives
  → Recall will be 1
  → Precision is equal to the share of positives

- Goal: Get a threshold the stakeholder agrees on

- Starting point might be estimation of economic benefit and cost



F1-Score

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- Harmonic mean of precision and recall

		Predicted	
		Negatives	Positives
Actual	Negatives	14	80
	Positives	0	6

F1-Score

- The harmonic mean punishes low rates.

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Precision	Recall	F1-score
5%	50%	9%
90%	90%	90%
30%	60%	40%

Let's take negatives into the equation

- The amount of correct negative predictions is sometimes just as important
- Spam vs. ham is just one example (email spam detection)

Image source

False Positive Rate

What proportion of actual negatives was predicted as positives?

- Here, FPR is 80/(80+14) = 85.11%

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

		Predicted	
		Negatives	Positives
Actual	Negatives	14	80
	Positives	0	6

True Negative Rate

What proportion of actual negatives was predicted correctly?

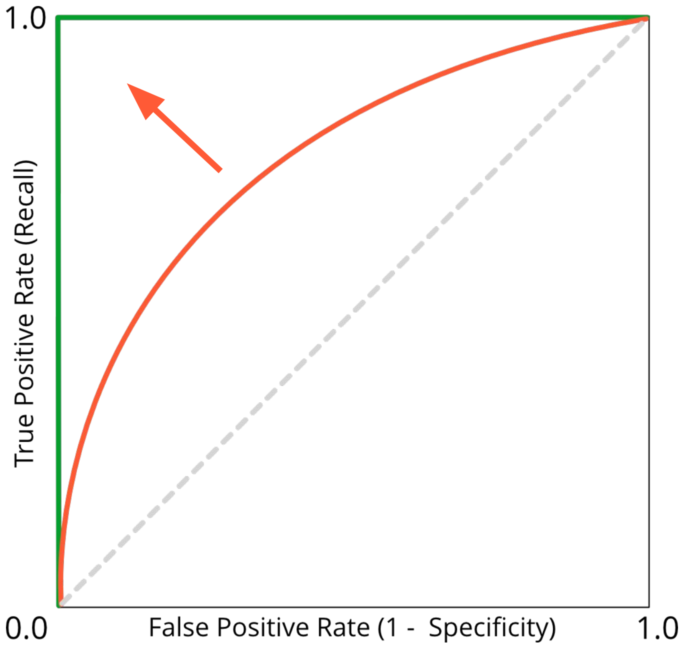
- Here, TNR is 14/(80+14) = 14.89%
- TNR is also called *specificity*
- FPR = 1 - specificity

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

		Predicted: Negatives	Predicted: Positives
Actual	Negatives	14	80
Actual	Positives	0	6

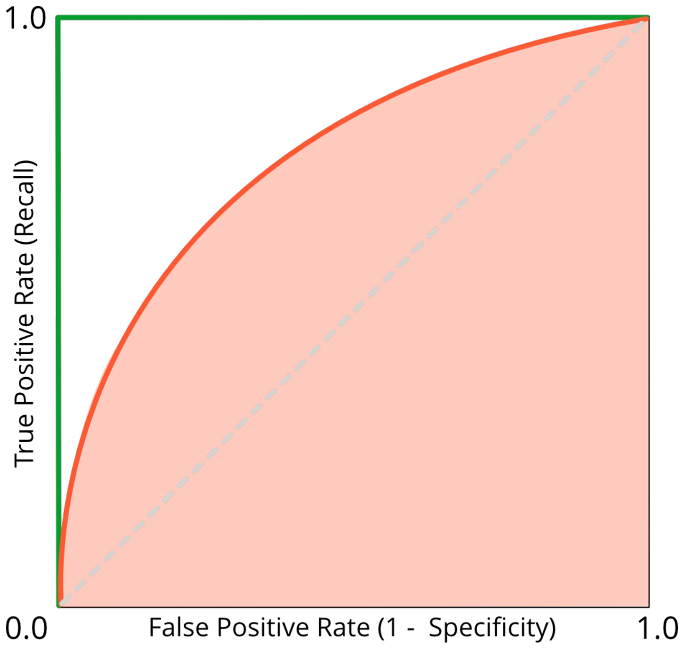
Receiver Operating Characteristic Curve (ROC Curve)

- Plots TPR vs. FPR for different **thresholds**
- The 45° line is equivalent to throwing a coin
- If all positives are correctly predicted and no negative is incorrectly predicted ROC curve would be the green curve
- Aim: ROC curve as closely as possible to (0,1)

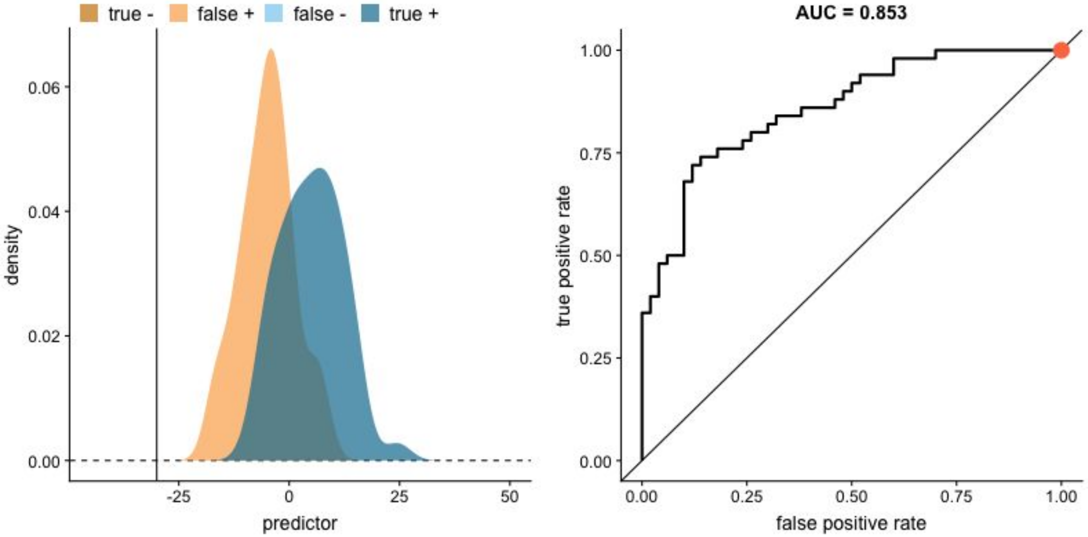


ROC and the Area Under the Curve (ROC AUC)

- Metric to compare different classifiers
- Random classifier:
  → ROC AUC is 0.5
  → ROC curve is on the 45° line
- Perfect classifier:
  → ROC AUC is 1



Explanation of ROC curve



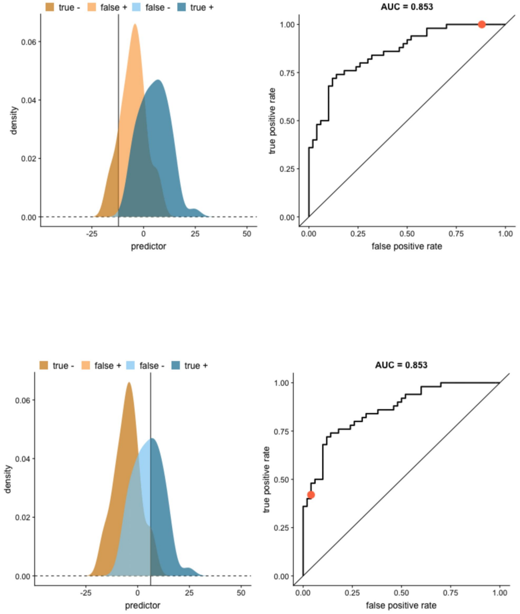
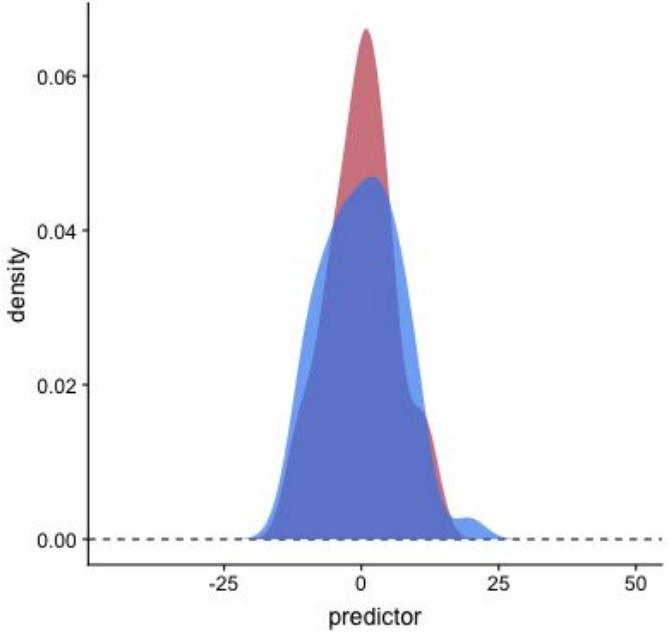


Image source

Explanation of ROC curve



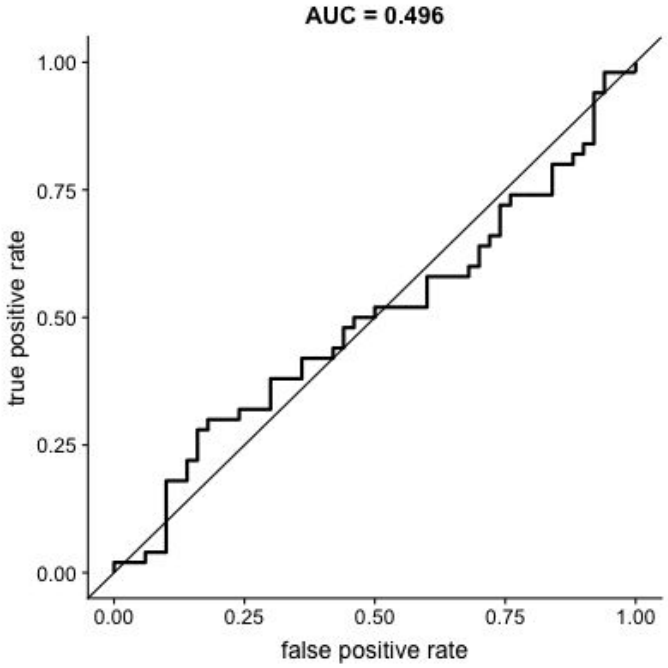


Image source

Trade-off

Imbalanced classes

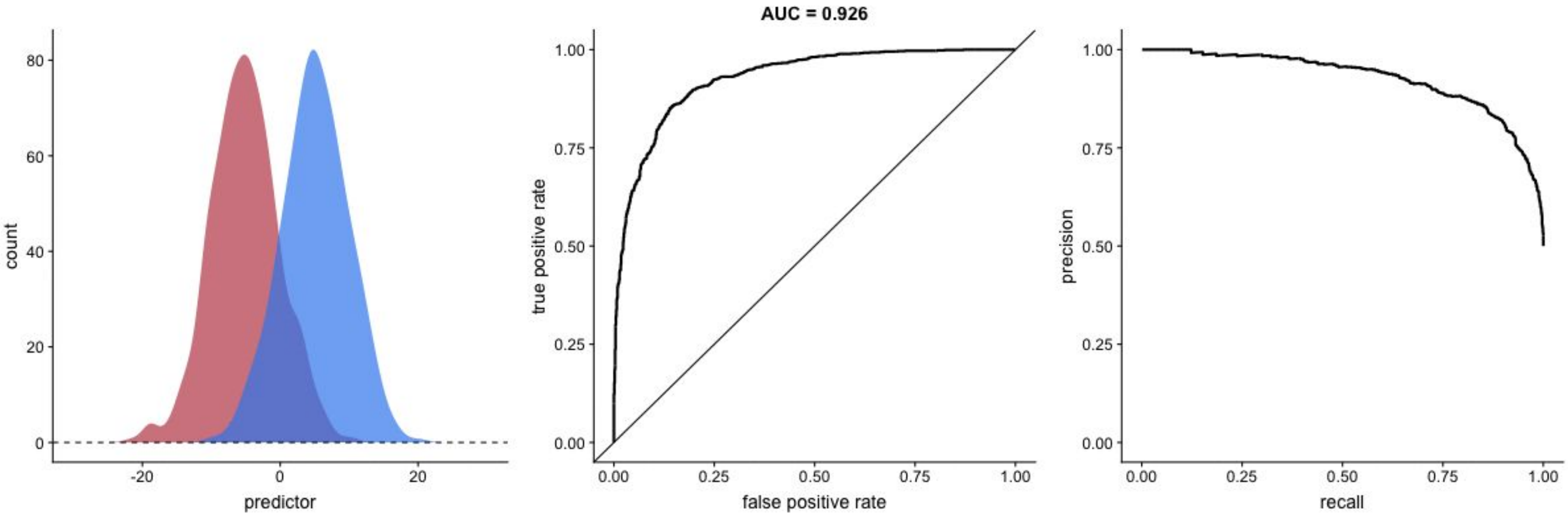


Image source

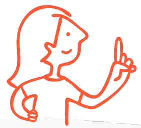
Going beyond aggregated metrics

All the performance metrics we've seen today are aggregated metrics.

They help determine whether a model has learned well from a dataset or needs improvement.

Next step:
examine results and errors to understand why and how the model is failing or succeeding

Why: validation and iteration



Performance metrics can be deceptive, on highly imbalanced datasets a classifier can reach very high accuracy without any predictive power

Next Steps - Error Analysis

Validated your model -
inspect how it is performing

There are a lot of way to do this. You want to contrast data (target and/or features) and predictions

- **Regression**:
  looking at residuals, for example doing EDA on residuals and inspecting the outliers

- **Classification**:
  one can start with a confusion matrix, breaking results in true class and predictions

Resources

https://paulvanderlaken.com/2019/08/16/roc-auc-precision-and-recall-visually-explained/

Building Machine Learning Powered Applications - Emmanuel Ameisen