Decision Trees and Random Forests

What is a Decision Tree?

- Supervised machine learning
- Flow chart the machine takes to sort things into categories (making *decisions* along the way)
- For classification

Parts of a Decision Tree

- Node – each decision point
  - Root Node – starting decision point
- Edge – The path between nodes
- Leaves – Possible outcomes at the end (categories)

What is a Random Forest?

- Decision tree on steroids!

- Test every combination of nodes to find the best place to put it

Putting it all together…

	Supervised?	Classifying?	True x and y?
Linear regression	Y	N	Y
K-means	N	N	N
K-nearest neighbors	Y	Y	Y
Decision trees	Y	Y	Y
Random forests	Y	Y	Y

General Steps for Decision Trees & RF

- Wrangle the data
- Split into training and testing sets
- Create the initial model
- Assess the fit of the model

Assessing Model Fit

Model Accuracy

Recall

- Ability to find all relevant cases within the dataset

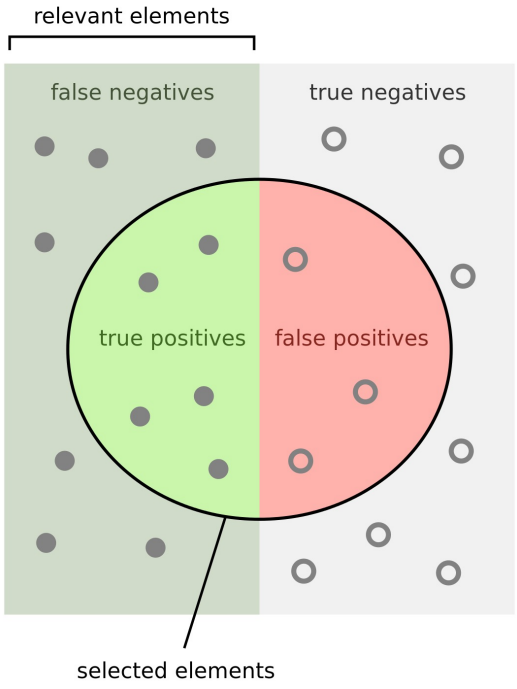
$$Recall = \frac{TP}{TP + FN}$$

Precision

- Ability to identify only the relevant data points

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

In graphical form...



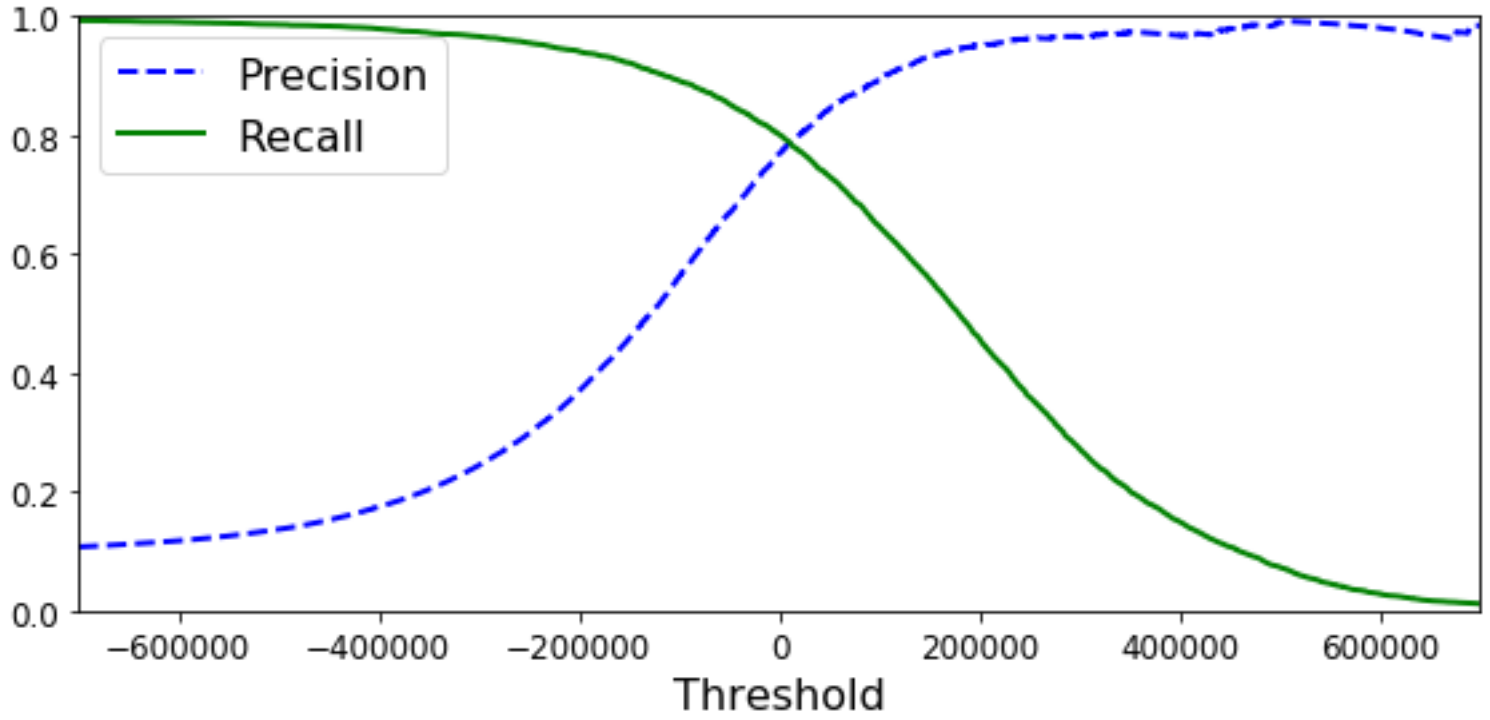
How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

Inversely Related



F1 Score

- Takes both precision and recall into account

$$F1 = \frac{2\ x\ precision\ x\ recall}{precision + recall}$$

Why are there three?!

- Recall – when you can’t afford to miss anything
  - Disease screening
  - Terrorists
- Precision – when the consequences of mislabeling are high
  - Administering very expensive treatments
  - Putting people in prison for life
- F1 – when you need a good mix of both