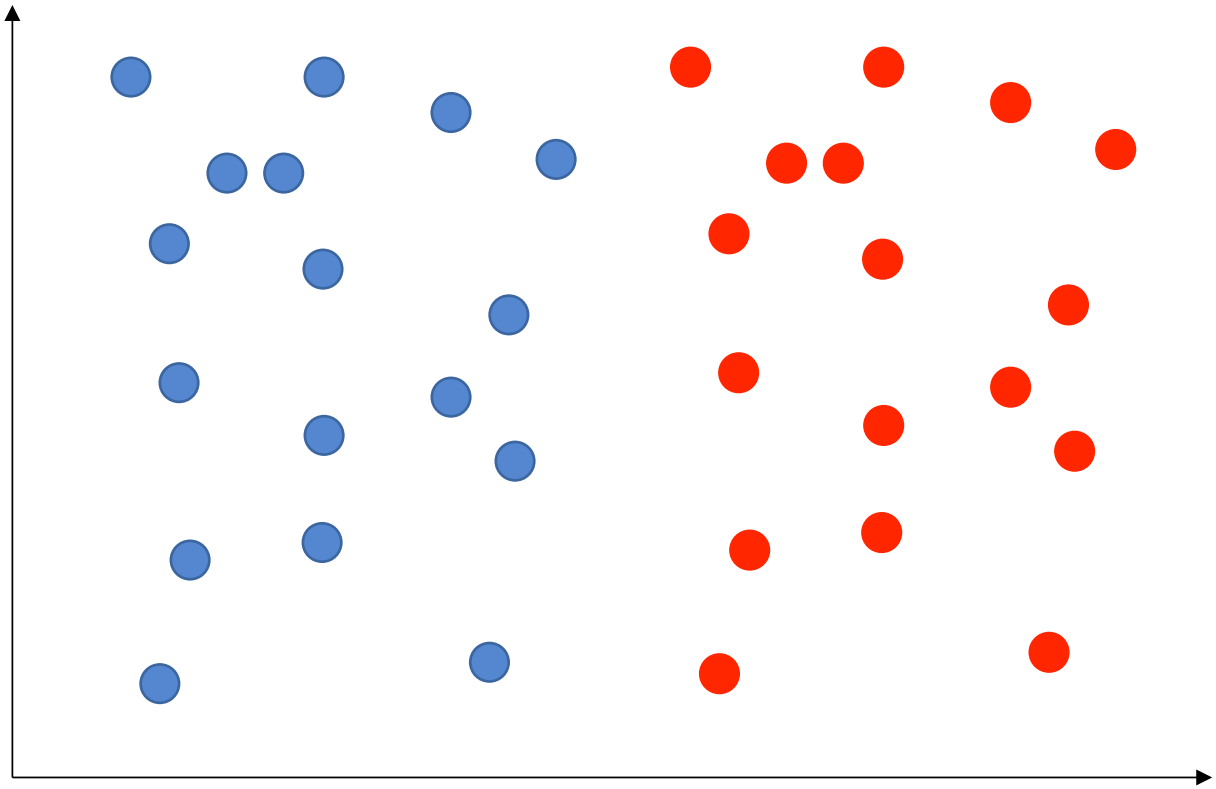
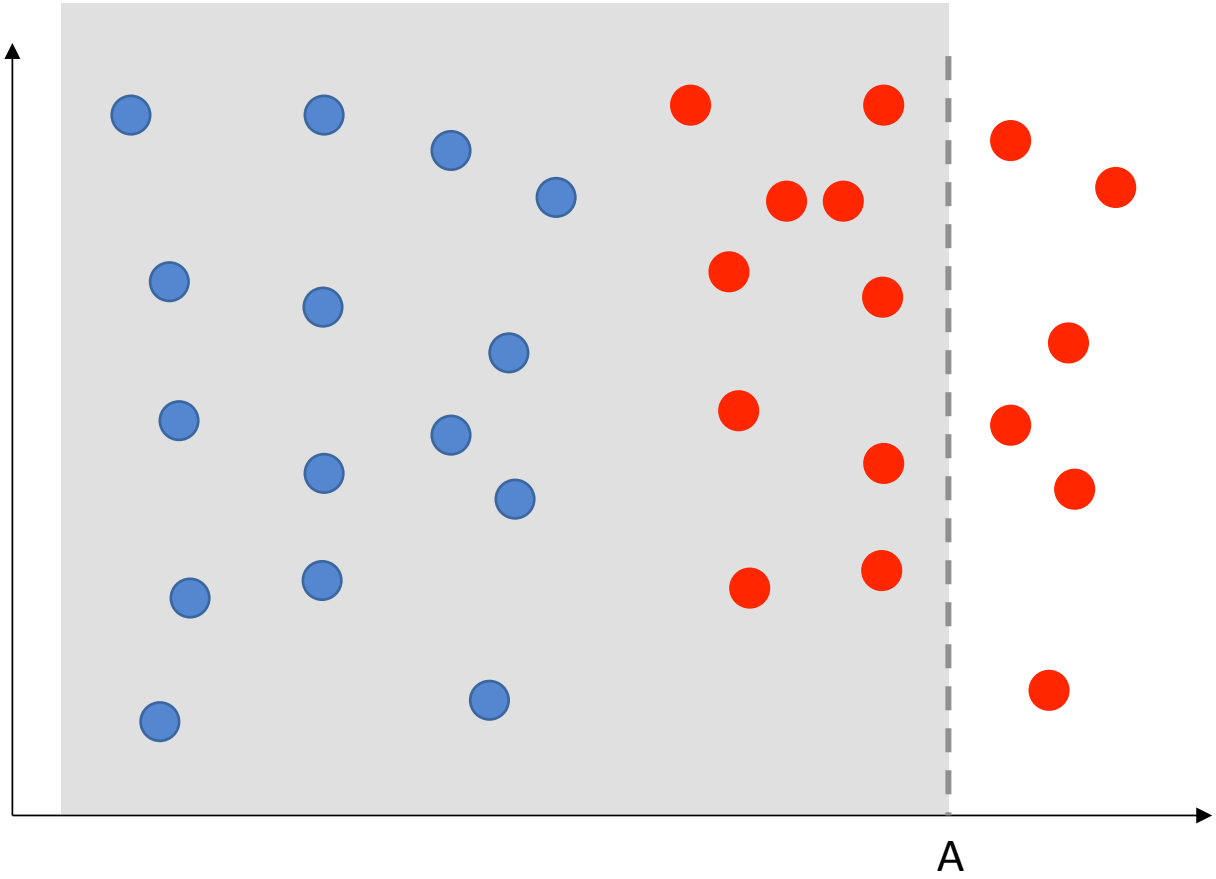
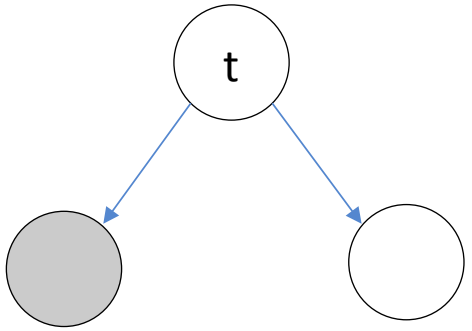
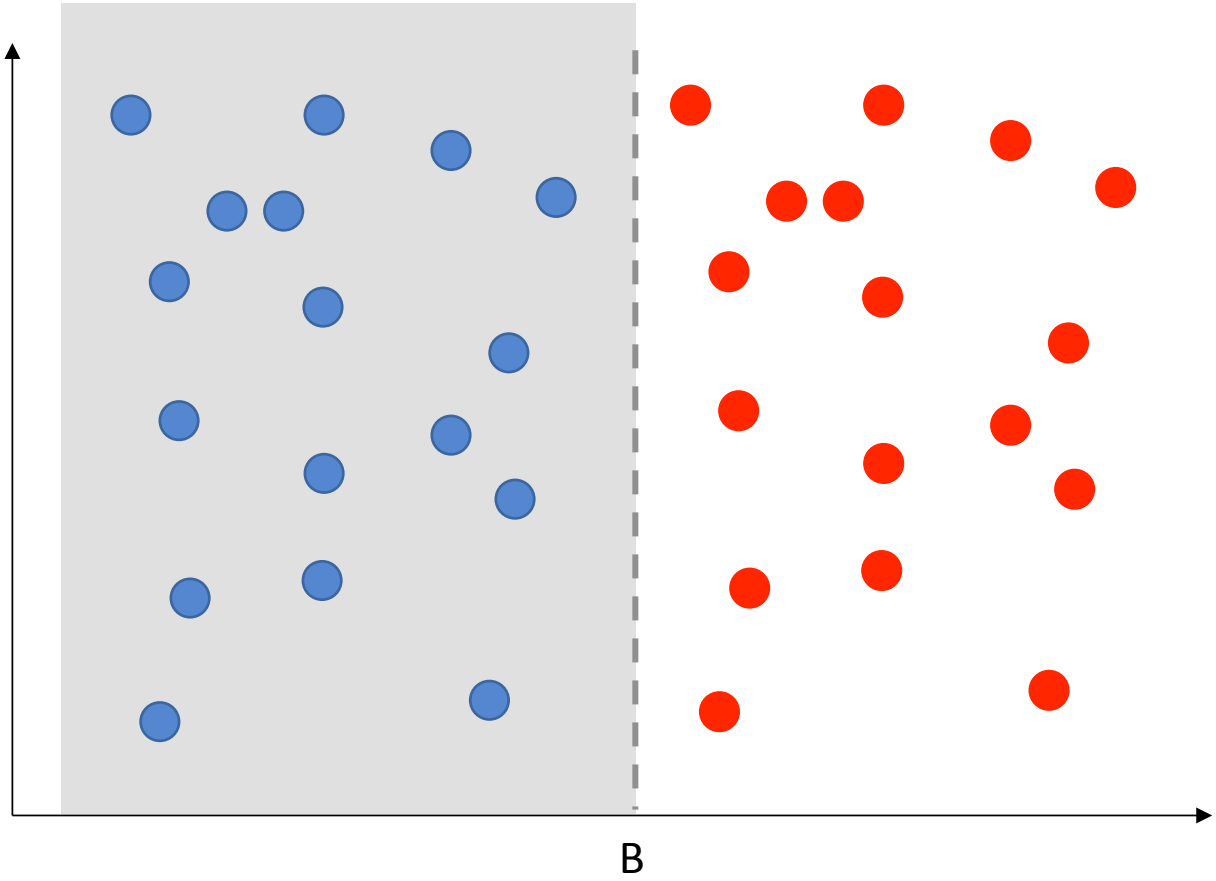
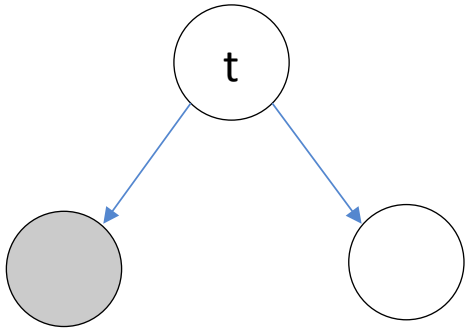
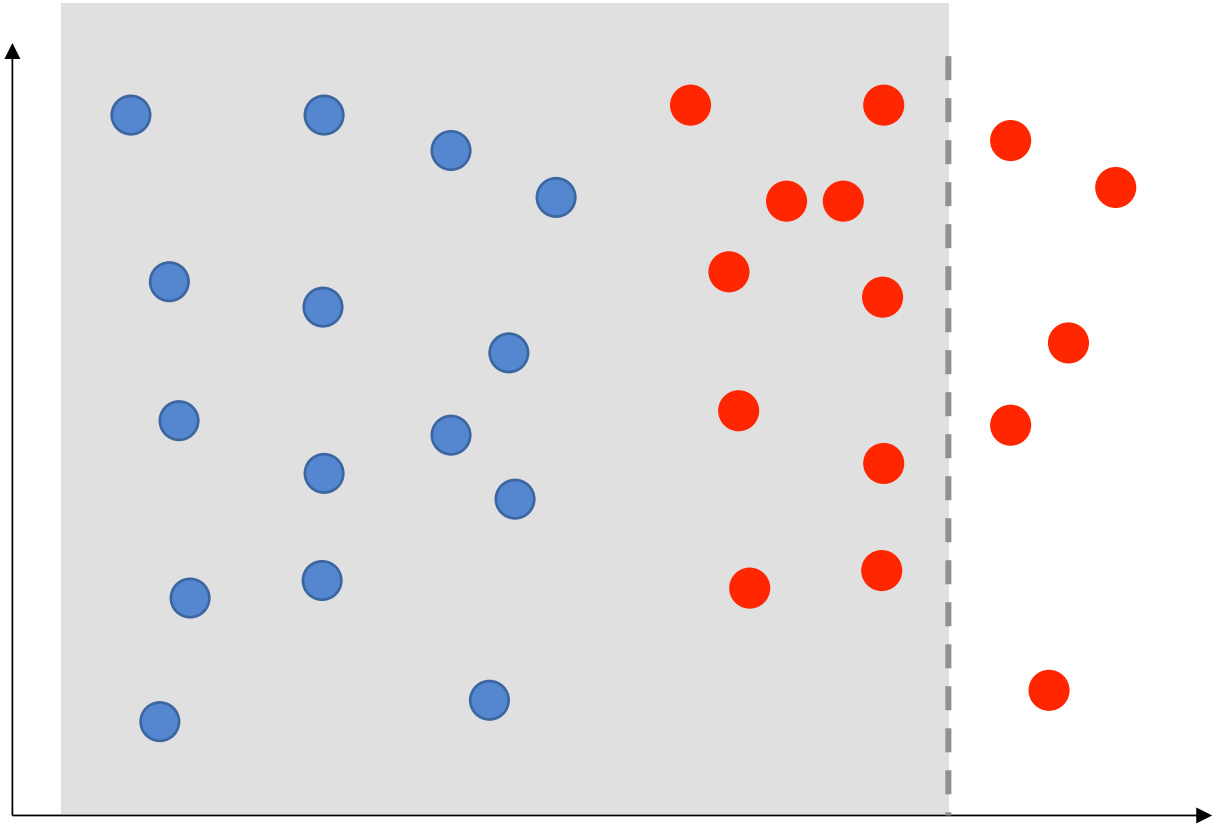

t


A


t

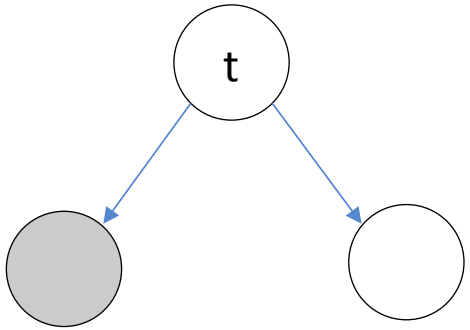

B


t



$pB\square A = 3/5$
p(R|A)=2/5

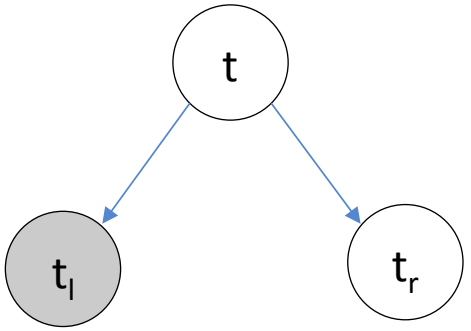
$pB\square A = 0$
p(R|A)=1



$p\downarrow l = N\downarrow l / N = 5/6$

$N\downarrow l / N = 1/6$

Impurity, Splits and the Gini Index



- Gini Index: $i\downarrow G\ (t)=\sum\uparrow\blacksquare p(c\downarrow k\,|\,t)(1-\ p(c\downarrow k\,|\,t))$

- Impurity decrease due to a split s applied to a node t:

$$\Delta i(s,t)=i(t)-p\downarrow l\ i(t\downarrow l\ )-p\downarrow r\ i(t\downarrow r\ )$$

- Choose splits to **maximize** the decrease in impurity.

Exercise

X1	X2	X3	Y
0	0	0	c1
0	0	1	c1
0	1	0	c2
0	1	1	c2
0	1	1	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	1	1	c2

The features are all binary, and there
2 classes (c1 and c2).
Which feature should I split on first?

- Gini Index: $i\downarrow G\ (t)=\sum\uparrow\blacksquare p(c$
  $c\downarrow k\,|\,t))$

- Impurity decrease due to a split s applied to a

$$\Delta i(s,t)=i(t)-p\downarrow l\ i(\ t\downarrow l\ )$$

Exercise

X1	X2	X3	Y
0	0	0	c1
0	0	1	c1
0	1	0	c2
0	1	1	c2
0	1	1	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	1	1	c2

The features are all binary, and there a
2 classes (c1 and c2).
Which feature should I split on first?

Step 1: Calculate Gini impurity of the original node.

$i\downarrow G\ (t)=\sum\uparrow\blacksquare p(c\downarrow k\,|t)(1-\ p(c\downarrow k\,|t)$

$i\downarrow G\ (t)=1/5\ 4/5\ +4/5\ 1/5$

- Gini Index: $i\downarrow G\ (t)=\sum\uparrow\blacksquare p($
  $c\downarrow k\,|t))$

- Impurity decrease due to a split s applied to a

$\Delta i(s,t)=\boldsymbol{i(t)}-p\downarrow l\ i(\ t\downarrow l\ )$

Exercise

X1	X2	X3	Y
0	0	0	c1
0	0	1	c1
0	1	0	c2
0	1	1	c2
0	1	1	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	1	1	c2

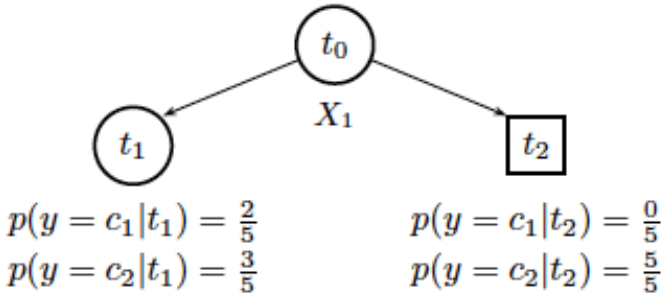
The features are all binary, and there a
2 classes (c1 and c2).
Which feature should I split on first?

Step 1: Calculate Gini impurity of the original node.

$i{\downarrow}G\ (t)=\sum{\uparrow}\ p(c{\downarrow}k\,|t)(1-\ p(c{\downarrow}k\,|t)$

$i{\downarrow}G\ (t)=1/5\ 4/5\ +4/5\ 1/5$

Step 2: Calculate Gini impurity for the new nodes.



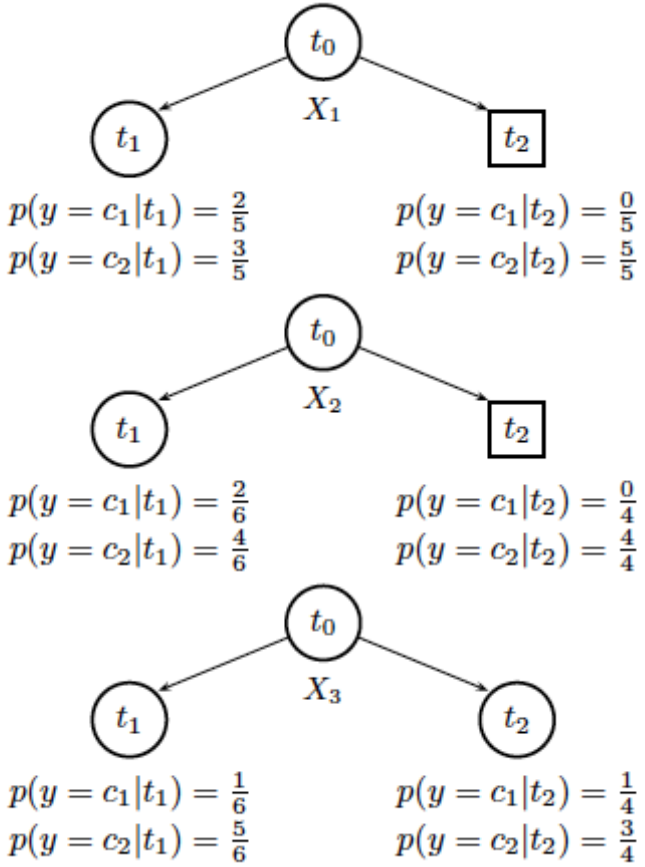
- Gini Index: $i{\downarrow}G\ (t)=\sum{\uparrow}\ p(c{\downarrow}k\,|t))$

- Impurity decrease due to a split s applied to a

$$\Delta i(s,t)=i(t)-p{\downarrow}l\ i(t{\downarrow}l)$$

Exercise

X1	X2	X3	Y
0	0	0	c1
0	0	1	c1
0	1	0	c2
0	1	1	c2
0	1	1	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	1	1	c2

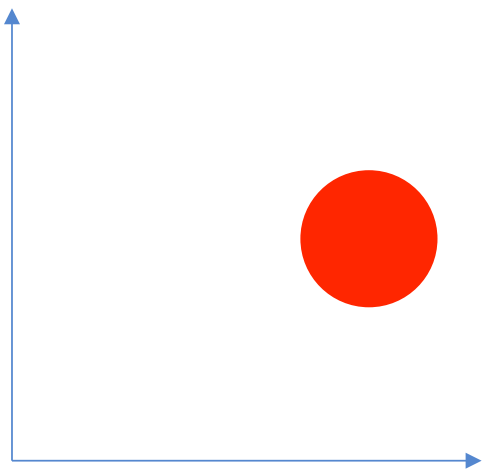


- Gini Index: $i \downarrow G\ (t) = \sum \uparrow \blacksquare p(c$ $c \downarrow k\ |t))$

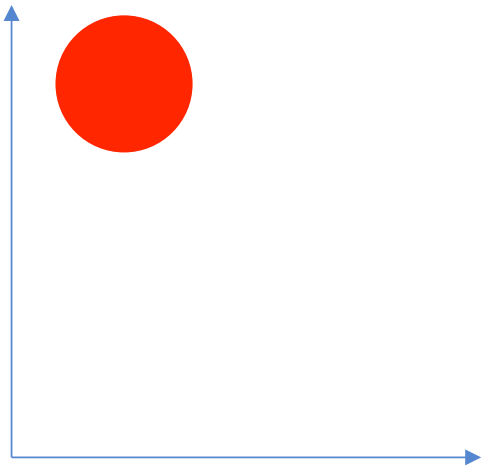
- Impurity decrease due to a split s applied to a

$$\Delta i(s,t) = i(t) - p \downarrow l\ i(t \downarrow l)$$

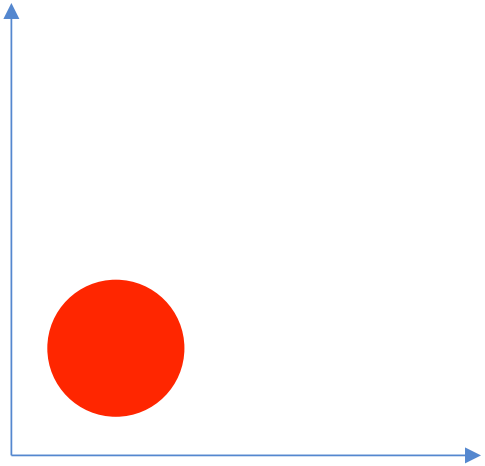
Random Forests: Intuition



Model 1

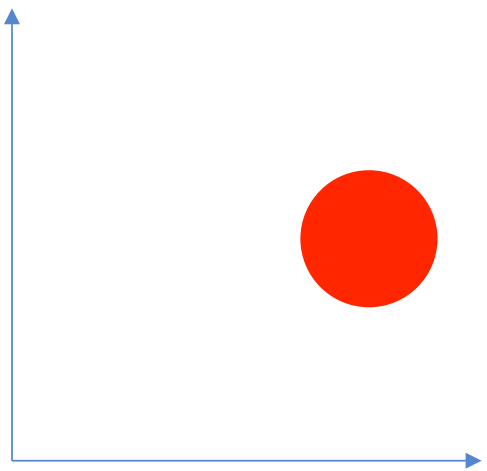


Model 2

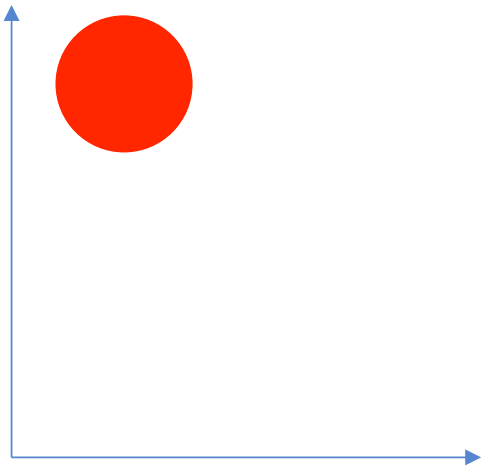


Model 3

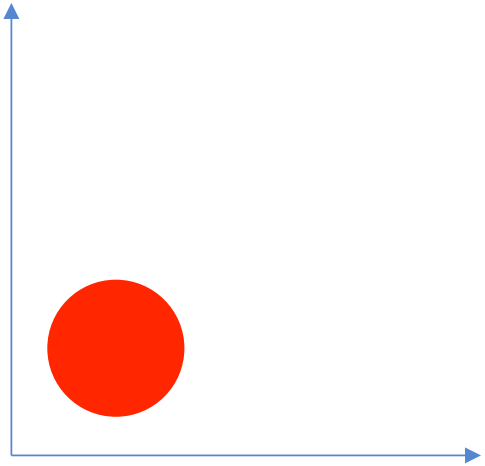
Random Forests: Intuition



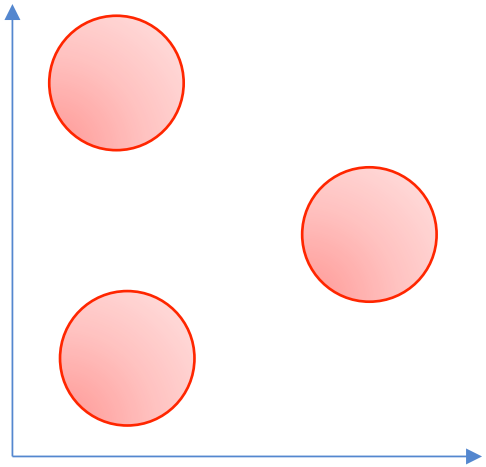
Model 1



Model 2



Model 3



Model Ensemble