Gradient Boosting Explained

(Classification Example)

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	12	Blue	Yes
Yes	87	Green	Yes
No	44	Blue	No
Yes	19	Red	No
No	32	Green	Yes
No	14	Blue	Yes

When we use **Gradient Boost for Classification**, the initial **Prediction** for every individual is the **log(odds)**.

I like to think of the **log(odds)** as the **Logistic Regression** equivalent of the average.

$\log(4/2) = 0.7$

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	12	Blue	Yes
Yes	87	Green	Yes
No	44	Blue	No
Yes	19	Red	No
No	32	Green	Yes
No	14	Blue	Yes

…which we will put into our initial leaf.

$$\log\left(\frac{4}{2}\right) = 0.7$$

log(4/2) = 0.7

Probability of **Loving Troll 2** = 0.7

And let's save that up here for now.

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	12	Blue	Yes
Yes	87	Green	Yes
No	44	Blue	No
Yes	19	Red	No
No	32	Green	Yes
No	14	Blue	Yes

Probability of **Loving Troll 2** $= \frac{e^{\log(4/2)}}{1 + e^{\log(4/2)}} = \boxed{0.7}$

log(4/2) = 0.7

Probability of **Loving Troll 2** = 0.7

NOTE: These two numbers, the **log(4/2)** and the **Probability** are the same only because I'm rounding. If I allowed **4** digits passed the decimal place…

$$\log\left(\frac{4}{2}\right) = 0.6931$$

$$\frac{e^{\log(4/2)}}{1 + e^{\log(4/2)}} = 0.6667$$

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	12	Blue	Yes
Yes	87	Green	Yes
No	44	Blue	No
Yes	19	Red	No
No	32	Green	Yes
No	14	Blue	Yes

log(4/2) = 0.7

Probability of **Loving Troll 2** = 0.7

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	12	Blue	Yes
Yes	87	Green	Yes
No	44	Blue	No
Yes	19	Red	No
No	32	Green	Yes
No	14	Blue	Yes

We can measure how bad the initial **Prediction** is by calculating **Pseudo Residuals**, the difference between the **Observed** and the **Predicted** values.

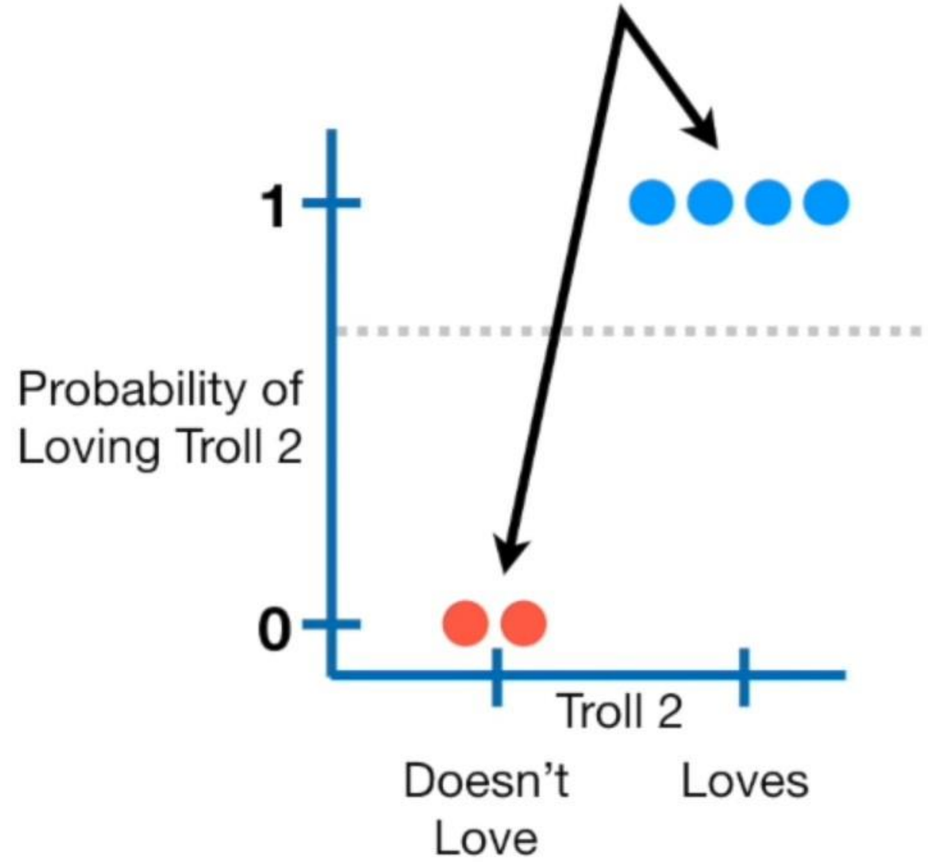
Residual = (**Observed** - **Predicted**)

log(4/2) = 0.7

Probability of Loving Troll 2 = 0.7

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	12	Blue	Yes
Yes	87	Green	Yes
No	44	Blue	No
Yes	19	Red	No
No	32	Green	Yes
No	14	Blue	Yes

In other words, the **Red** and **Blue** dots are the **Observed** values…

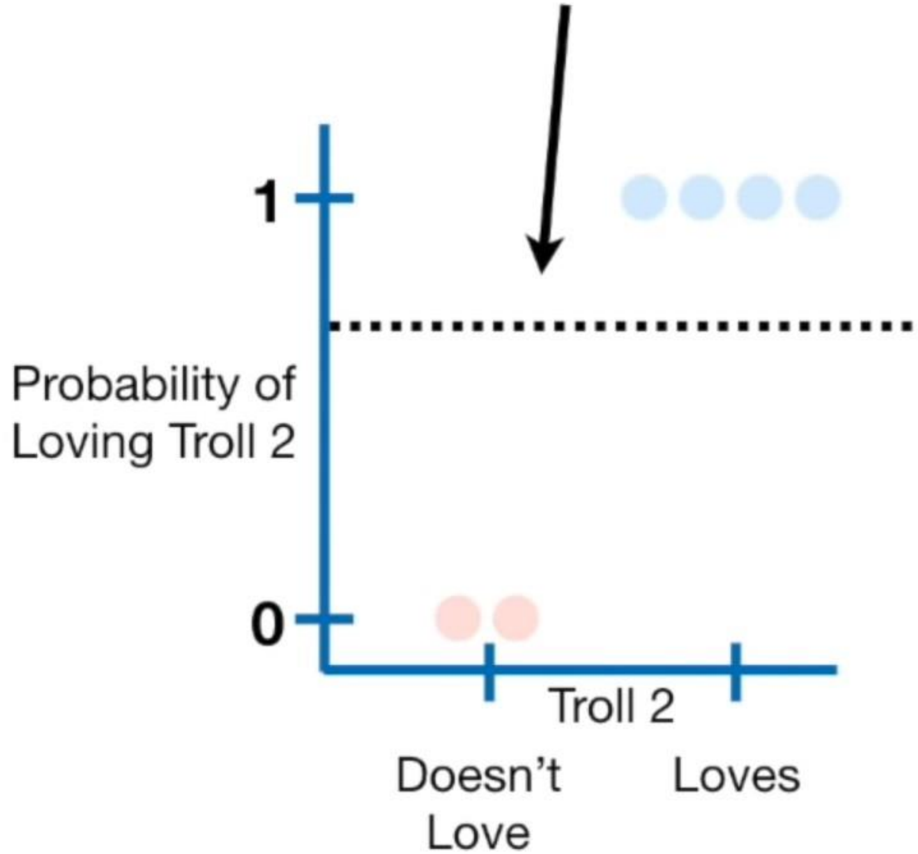


log(4/2) = 0.7

Probability of Loving Troll 2 = 0.7

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	12	Blue	Yes
Yes	87	Green	Yes
No	44	Blue	No
Yes	19	Red	No
No	32	Green	Yes
No	14	Blue	Yes

…and the dotted line is the **Predicted** value.



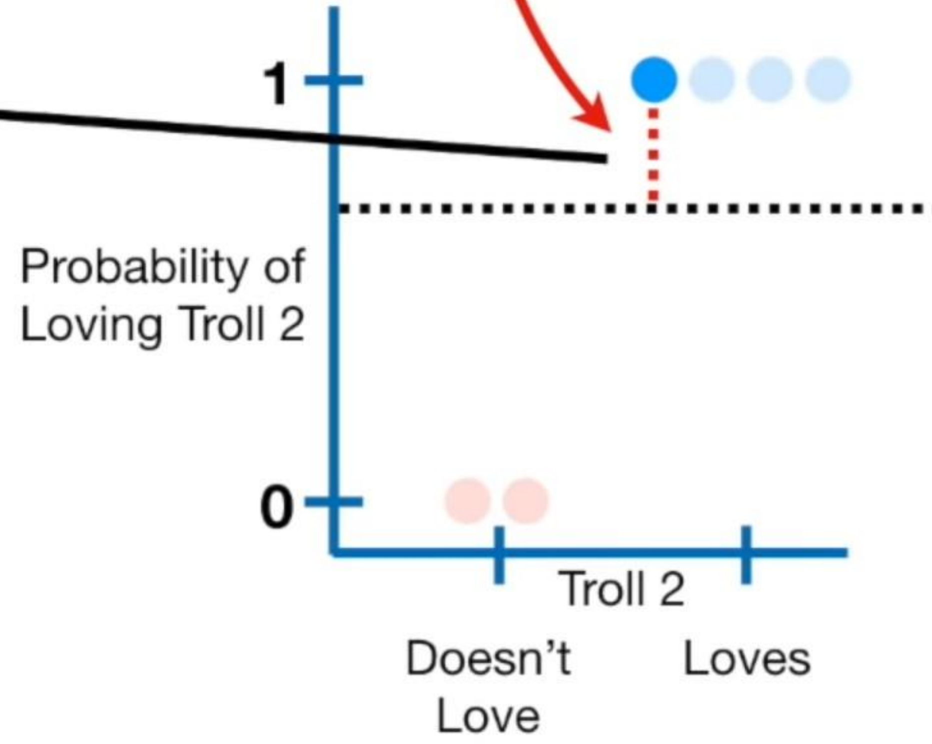
log(4/2) = 0.7

Probability of Loving Troll 2 = 0.7

...and we save the **Residual** in a new column.

Residual = (**1** - **0.7**) = 0.3

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Residual
Yes	12	Blue	Yes	0.3
Yes	87	Green	Yes	
No	44	Blue	No	
Yes	19	Red	No	
No	32	Green	Yes	
No	14	Blue	Yes	



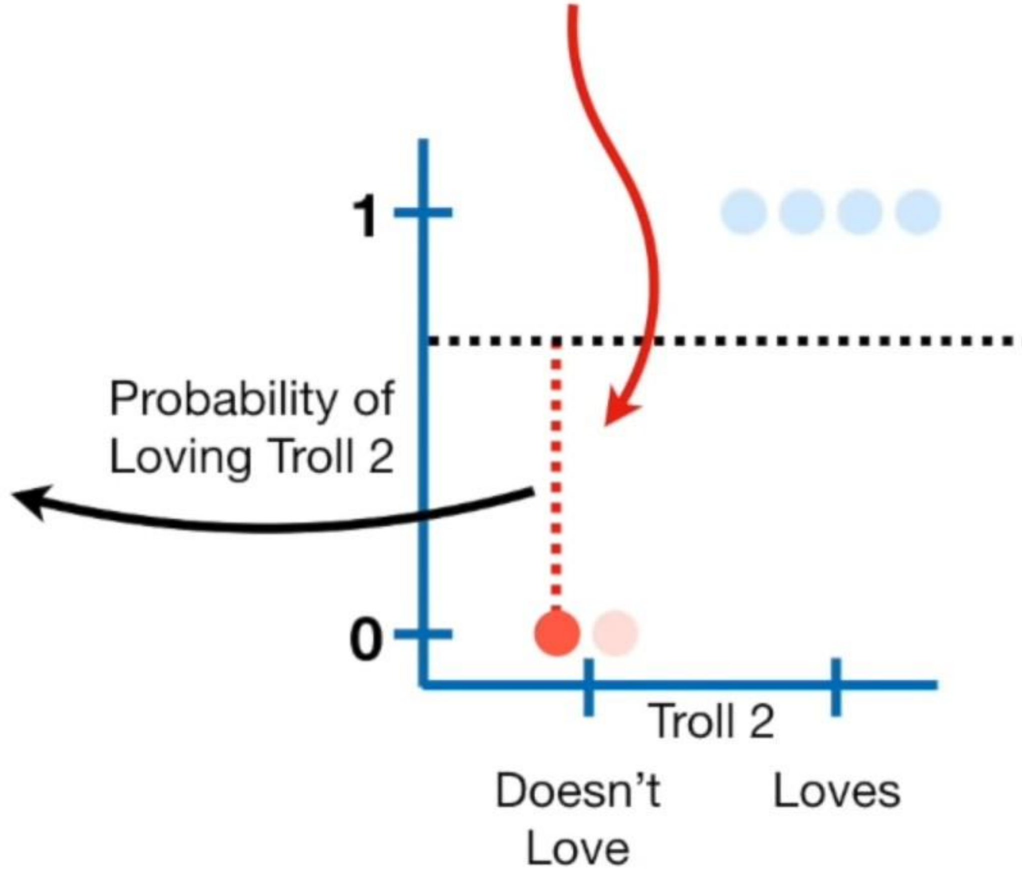
log(4/2) = 0.7

Probability of Loving Troll 2 = 0.7

Then we calculate the rest of the **Residuals**…

Residual = (**0** - **0.7**) = -0.7

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Residual
Yes	12	Blue	Yes	0.3
Yes	87	Green	Yes	0.3
No	44	Blue	No	-0.7
Yes	19	Red	No	
No	32	Green	Yes	
No	14	Blue	Yes	



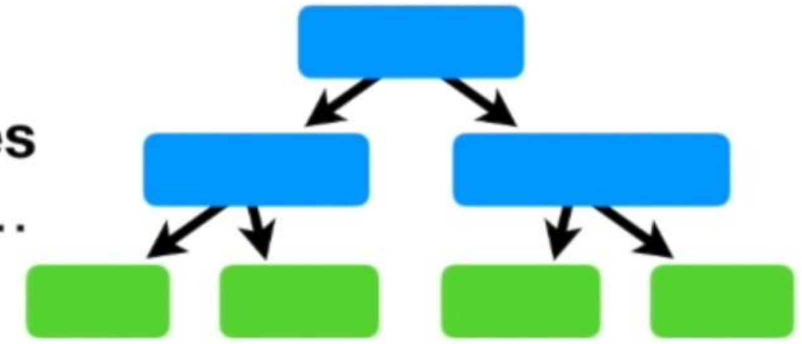
log(4/2) = 0.7

Probability of **Loving Troll 2** = 0.7

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Residual
Yes	12	Blue	Yes	0.3
Yes	87	Green	Yes	0.3
No	44	Blue	No	-0.7
Yes	19	Red	No	-0.7
No	32	Green	Yes	0.3
No	14	Blue	Yes	0.3

Hooray! We've calculated the **Residuals** for the leaf's initial **Prediction**.

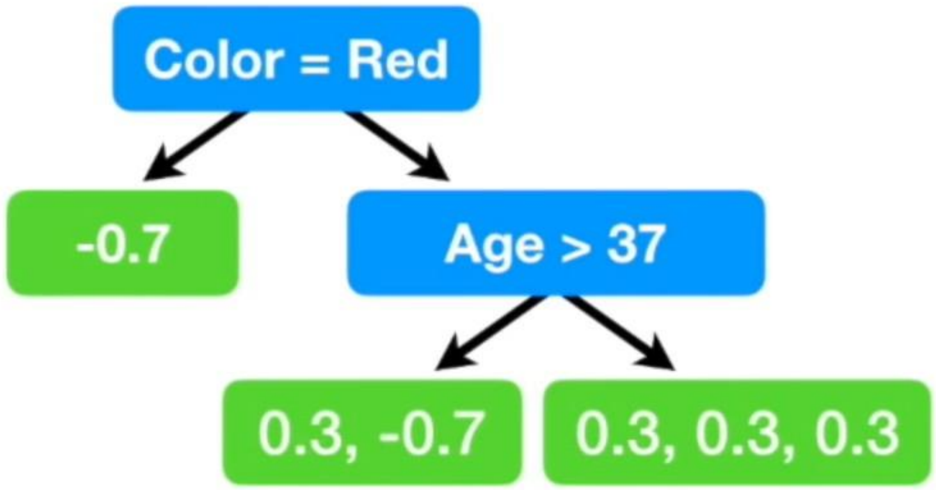
Now we will build a **Tree**, using **Likes Popcorn**, **Age** and **Favorite Color**...



Likes Popcorn	Age	Favorite Color	Loves Troll 2	Residual
Yes	12	Blue	Yes	0.3
Yes	87	Green	Yes	0.3
No	44	Blue	No	-0.7
Yes	19	Red	No	-0.7
No	32	Green	Yes	0.3
No	14	Blue	Yes	0.3

...to **Predict** the **Residuals**.

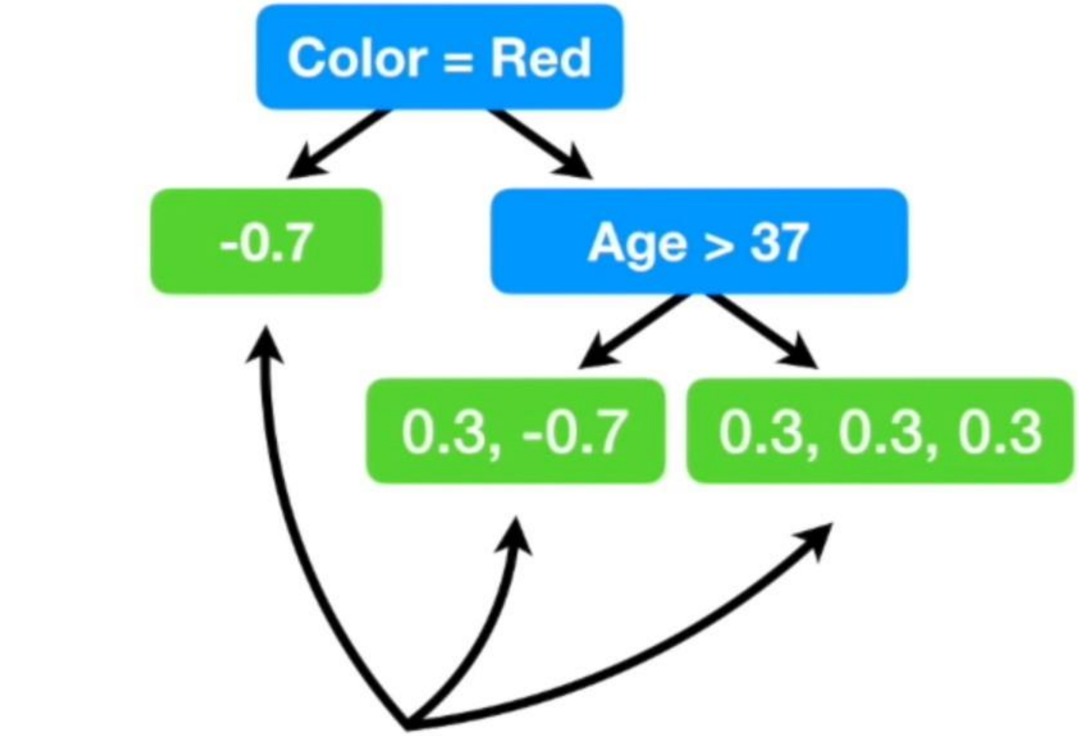
Likes Popcorn	Age	Favorite Color	Loves Troll 2	Residual
Yes	12	Blue	Yes	0.3
Yes	87	Green	Yes	0.3
No	44	Blue	No	-0.7
Yes	19	Red	No	-0.7
No	32	Green	Yes	0.3
No	14	Blue	Yes	0.3



In this simple example, we are limiting the number of leaves to **3**.

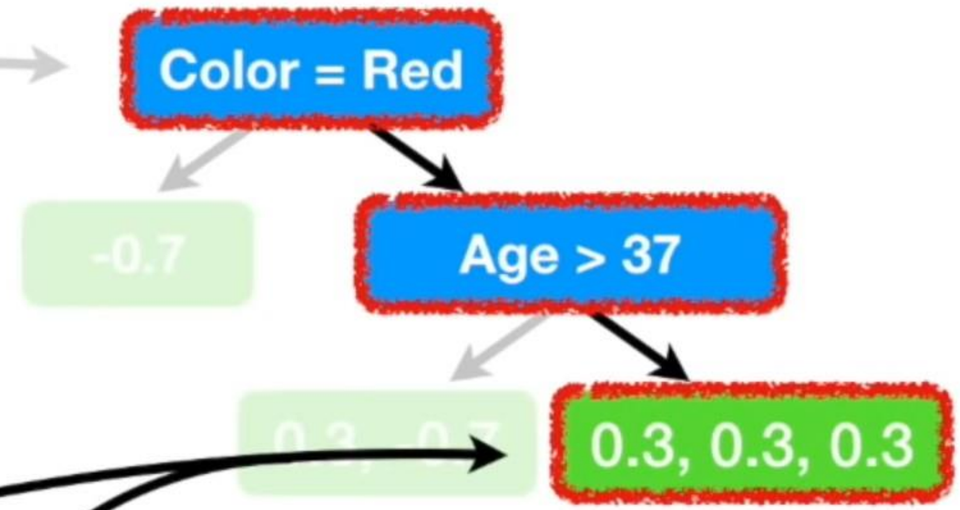
In practice people often set the maximum number of leaves to be between **8** and **32**

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Residual
Yes	12	Blue	Yes	0.3
Yes	87	Green	Yes	0.3
No	44	Blue	No	-0.7
Yes	19	Red	No	-0.7
No	32	Green	Yes	0.3
No	14	Blue	Yes	0.3



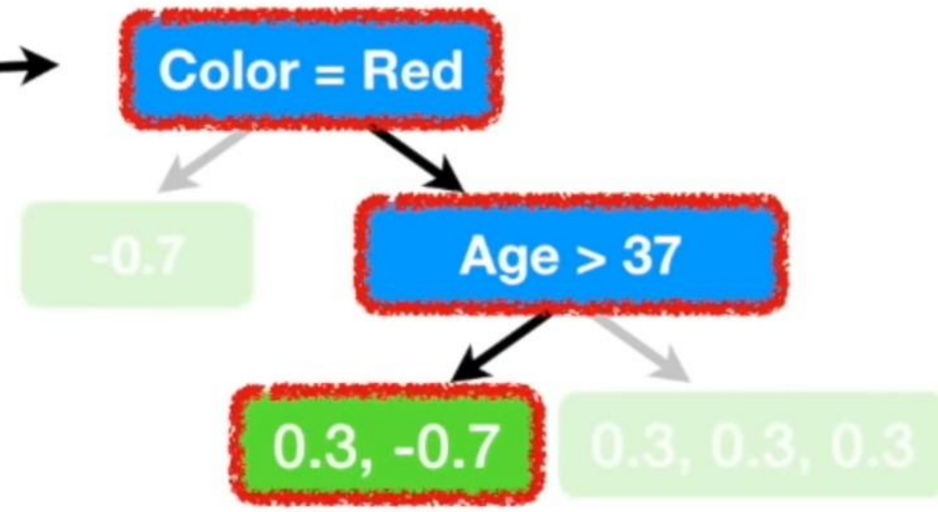
Now let's calculate the **Output Values** for the leaves.

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Residual
Yes	12	Blue	Yes	0.3
Yes	87	Green	Yes	0.3
No	44	Blue	No	-0.7
Yes	19	Red	No	-0.7
No	32	Green	Yes	0.3
No	14	Blue	Yes	0.3



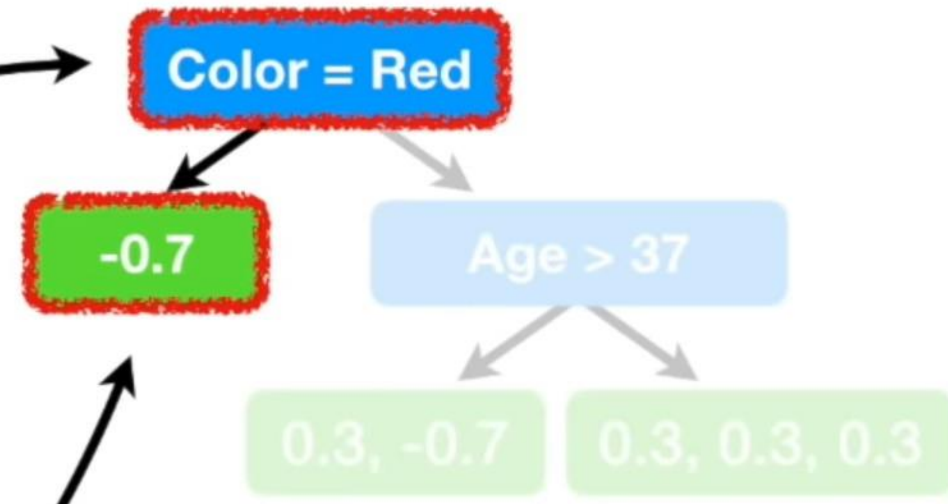
…go to the same leaf.

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Residual
Yes	12	Blue	Yes	0.3
Yes	87	Green	Yes	0.3
No	44	Blue	No	-0.7
Yes	19	Red	No	-0.7
No	32	Green	Yes	0.3
No	14	Blue	Yes	0.3

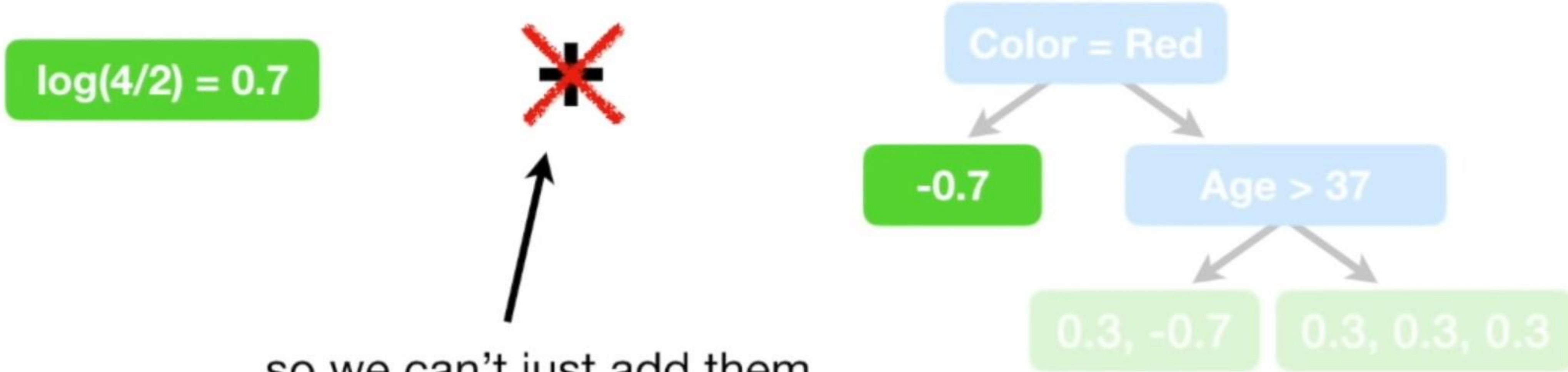


…go to the same leaf.

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Residual
Yes	12	Blue	Yes	0.3
Yes	87	Green	Yes	0.3
No	44	Blue	No	-0.7
Yes	19	Red	No	-0.7
No	32	Green	Yes	0.3
No	14	Blue	Yes	0.3



…goes to its own leaf.



…so we can't just add them together to get a new **log(odds) Prediction** without some sort of transformation.

When we use **Gradient Boost** for **Classification**, the most common transformation is the following formula.

$$\frac{\sum \text{Residual}_i}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]}$$

…and we end up with -3.3 as the new **Output Value** for this leaf.

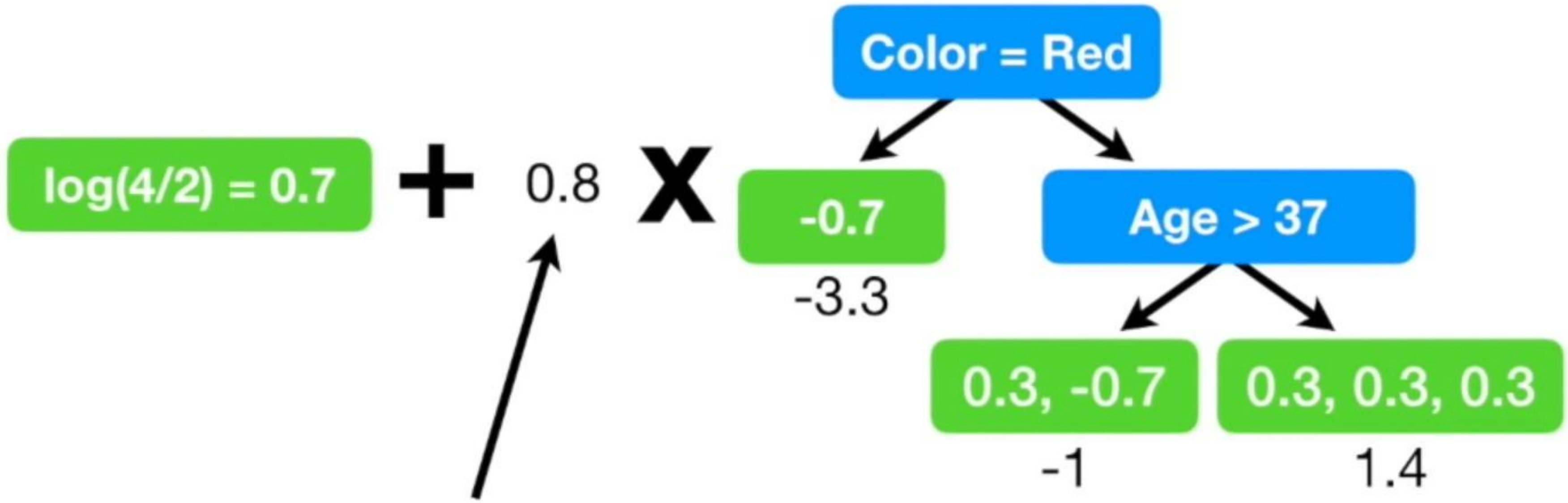
$$\frac{-0.7}{0.7 \times (1 - 0.7)} = \boxed{-3.3}$$

$$\frac{0.3 + -0.7}{2(0.7 \times (1 - 0.7))} = \boxed{-1}$$

...and the **Output Value** for this leaf is **-1**.

$$\frac{0.3}{(0.7 \times (1 - 0.7))} = \boxed{1.4}$$

...and the **Output Value**
for this leaf is **1.4**.



NOTE: Just like before, the new tree is scaled by a **Learning Rate**.

This example uses a relatively large **Learning Rate** for illustrative purposes. However, **0.1** is more common.

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	12	Blue	Yes
Yes	87	Green	Yes
No	44	Blue	No
Yes	19	Red	No
No	32	Green	Yes
No	14	Blue	Yes

…and the new **log(odds)**
Prediction = **1.8**.

log(odds) Prediction = 0.7 + (0.8 × 1.4) = $\boxed{1.8}$

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	12	Blue	Yes
Yes	87	Green	Yes
No	44	Blue	No
Yes	19	Red	No
No	32	Green	Yes
No	14	Blue	Yes

Now we convert the new **log(odds) Prediction** into a **Probability**…

$$\text{Probability} = \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}}$$

$$\log(\text{odds})\ \text{Prediction} = 0.7 + (0.8 \times 1.4) = 1.8$$

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Predicted Prob.
Yes	12	Blue	Yes	$\boxed{0.9}$
Yes	87	Green	Yes	
No	44	Blue	No	
Yes	19	Red	No	
No	32	Green	Yes	
No	14	Blue	Yes	

We save the new **Predicted Probability** here.

$$\text{Probability} = \frac{e^{1.8}}{1 + e^{1.8}} = \boxed{0.9}$$

$$\text{log(odds) Prediction} = 0.7 + (0.8 \times 1.4) = 1.8$$

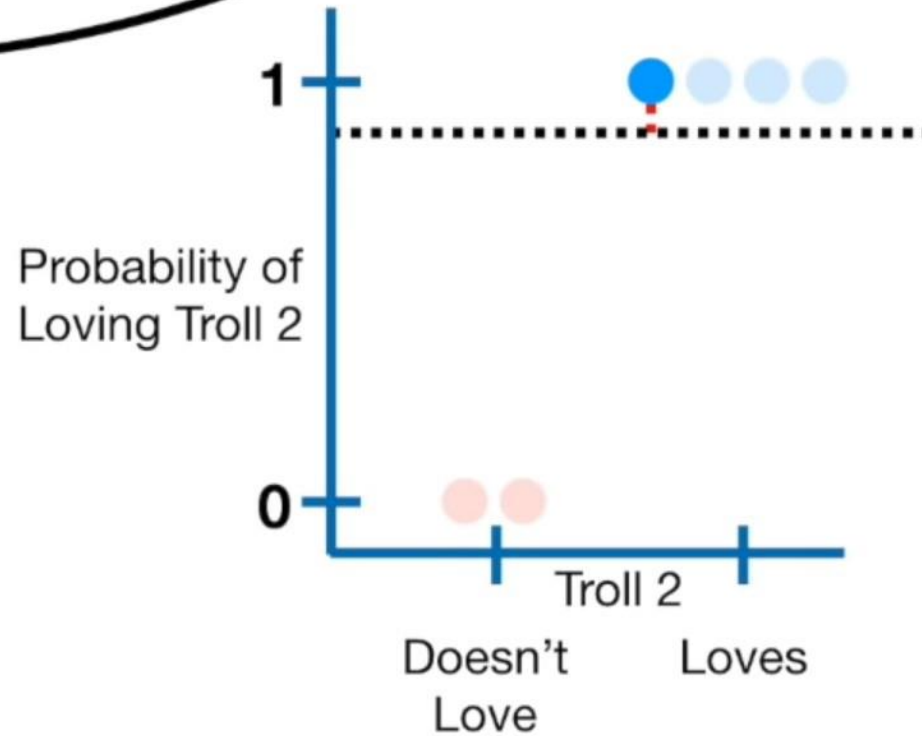
Likes Popcorn	Age	Favorite Color	Loves Troll 2	Predicted Prob.
Yes	12	Blue	Yes	0.9
Yes	87	Green	Yes	0.5
No	44	Blue	No	0.5
Yes	19	Red	No	0.1
No	32	Green	Yes	0.9
No	14	Blue	Yes	0.9

Then we calculate the **Predicted Probabilities** for the remaining people.

And we save that value here.

$$\textbf{Residual} = (\mathbf{1} - \mathbf{0.9}) = \boxed{0.1}$$

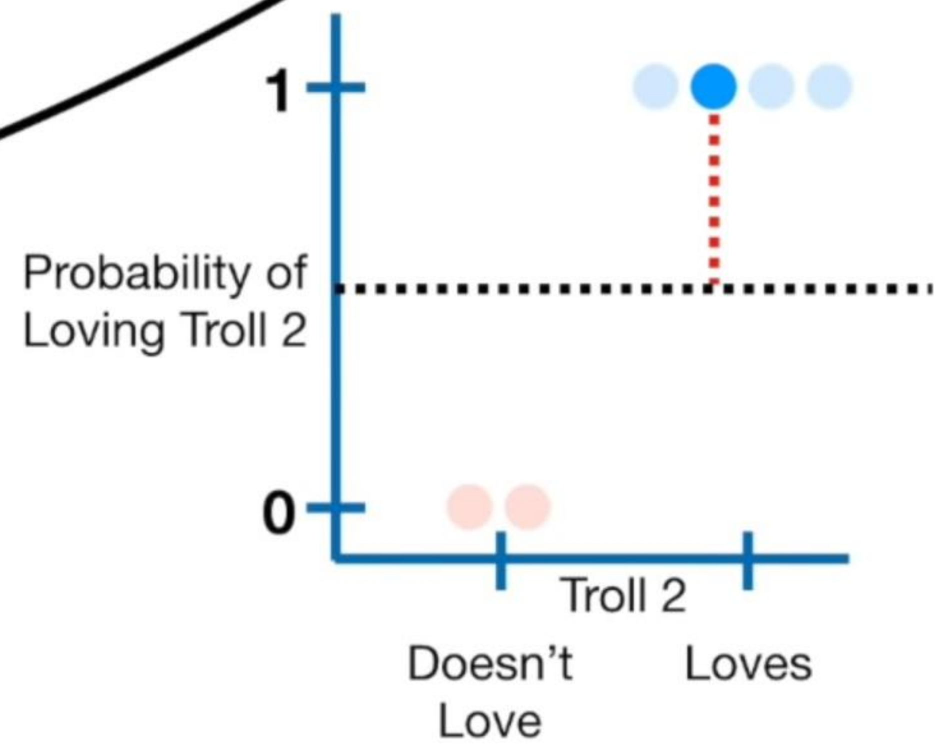
Likes Popcorn	Age	Favorite Color	Loves Troll 2	Predicted Prob.	Residual
Yes	12	Blue	Yes	0.9	0.1
Yes	87	Green	Yes	0.5	
No	44	Blue	No	0.5	
Yes	19	Red	No	0.1	
No	32	Green	Yes	0.9	
No	14	Blue	Yes	0.9	



And we save that value here.

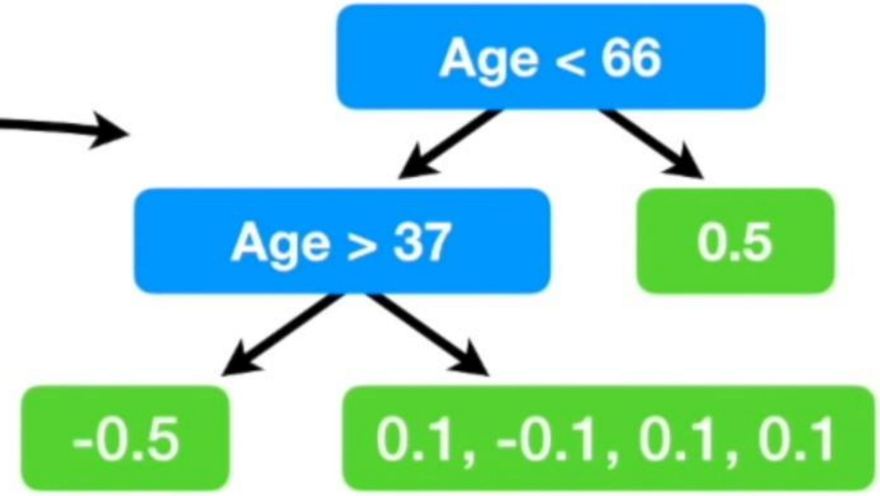
Residual = (**1** - **0.5**) = $\boxed{0.5}$

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Predicted Prob.	Residual
Yes	12	Blue	Yes	0.9	0.1
Yes	87	Green	Yes	0.5	0.5
No	44	Blue	No	0.5	
Yes	19	Red	No	0.1	
No	32	Green	Yes	0.9	
No	14	Blue	Yes	0.9	



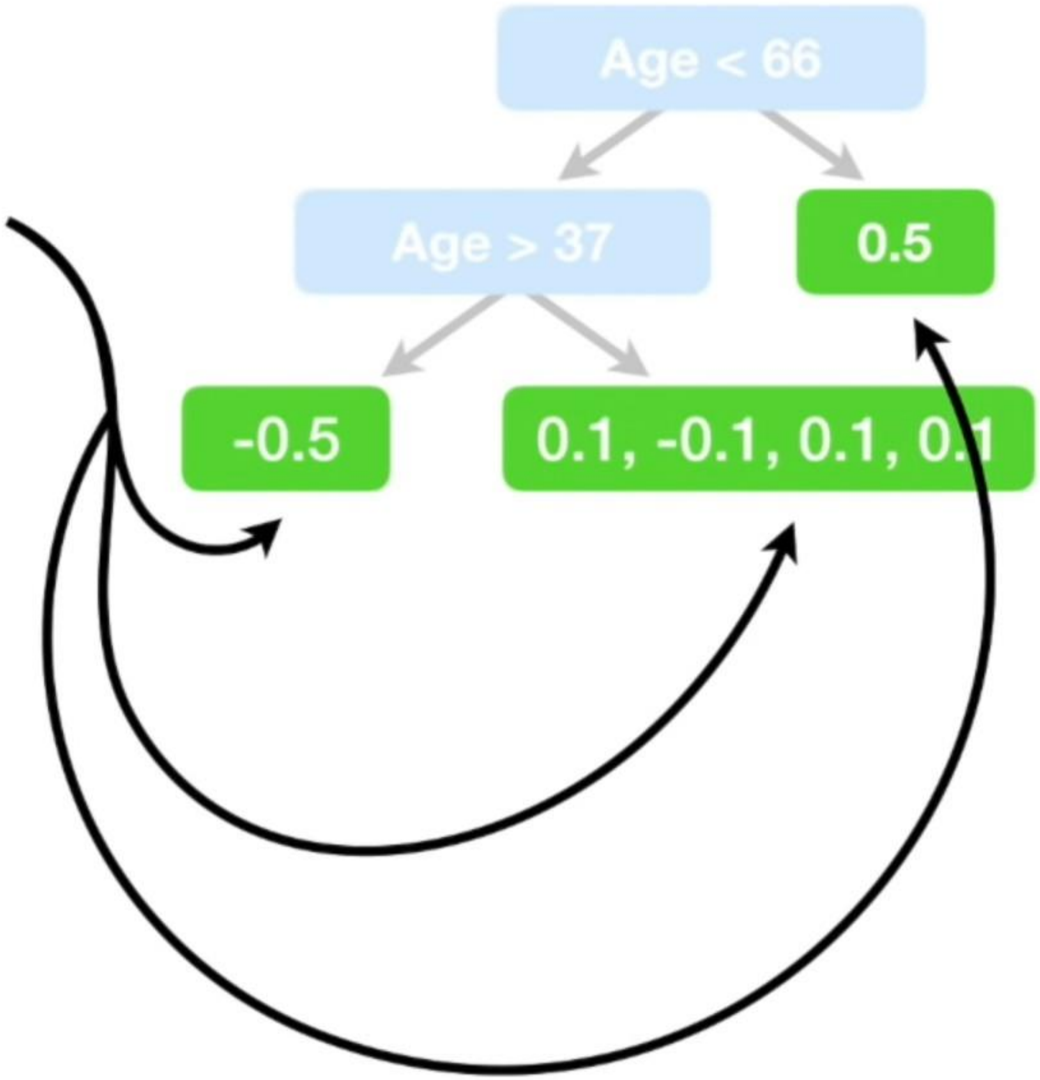
Now that we have the **Residuals**, we can build a new tree…

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Predicted Prob.	Residual
Yes	12	Blue	Yes	0.9	0.1
Yes	87	Green	Yes	0.5	0.5
No	44	Blue	No	0.5	-0.5
Yes	19	Red	No	0.1	-0.1
No	32	Green	Yes	0.9	0.1
No	14	Blue	Yes	0.9	0.1

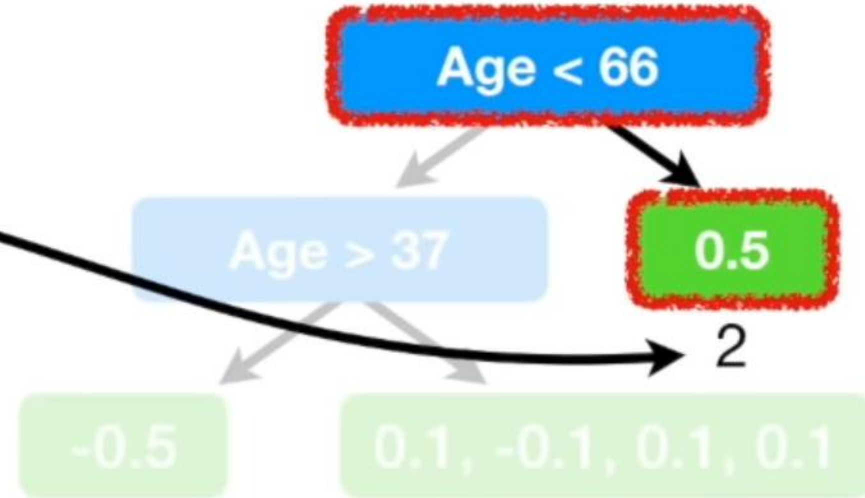


…and then we need to calculate the **Output Values** for each leaf.

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Predicted Prob.	Residual
Yes	12	Blue	Yes	0.9	0.1
Yes	87	Green	Yes	0.5	0.5
No	44	Blue	No	0.5	-0.5
Yes	19	Red	No	0.1	-0.1
No	32	Green	Yes	0.9	0.1
No	14	Blue	Yes	0.9	0.1



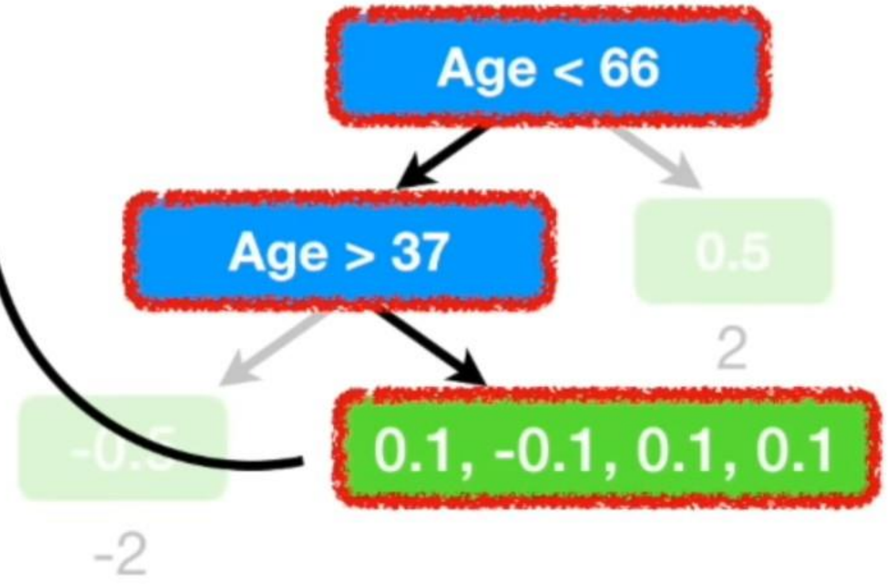
$$\frac{0.5}{0.5 \times (1 - 0.5)} = \boxed{2}$$



Likes Popcorn	Age	Favorite Color	Loves Troll 2	Predicted Prob.	Residual
Yes	12	Blue	Yes	0.9	0.1
Yes	87	Green	Yes	0.5	0.5
No	44	Blue	No	0.5	-0.5
Yes	19	Red	No	0.1	-0.1
No	32	Green	Yes	0.9	0.1
No	14	Blue	Yes	0.9	0.1

…and the **Output Value** for this leaf is **2**.

$$\frac{\sum \text{Residual}_i}{\sum \text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)}$$



Likes Popcorn	Age	Favorite Color	Loves Troll 2	Predicted Prob.	Residual
Yes	12	Blue	Yes	0.9	0.1
Yes	87	Green	Yes	0.5	0.5
No	44	Blue	No	0.5	-0.5
Yes	19	Red	No	0.1	-0.1
No	32	Green	Yes	0.9	0.1
No	14	Blue	Yes	0.9	0.1

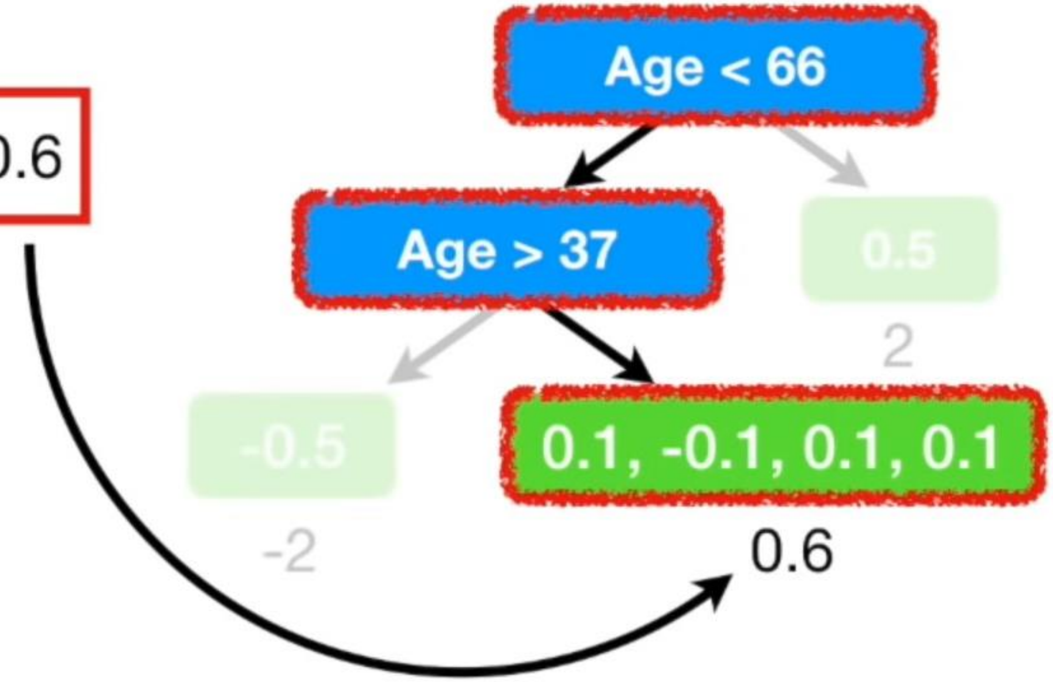
So we plug the **Residuals** into the formula for the **Output Values**…

$$\frac{0.1 + -0.1 + 0.1 + 0.1}{(0.9 \times (1 - 0.9)) + (0.1 \times (1 - 0.1)) + (0.9 \times (1 - 0.9)) + (0.9 \times (1 - 0.9))}$$

Likes Popcorn	Age	Favorite Color	Loves Troll 2	Predicted Prob.	Residual
Yes	12	Blue	Yes	0.9	0.1
Yes	87	Green	Yes	0.5	0.5
No	44	Blue	No	0.5	-0.5
Yes	19	Red	No	0.1	-0.1
No	32	Green	Yes	0.9	0.1
No	14	Blue	Yes	$\boxed{0.9}$	0.1

…and we plug in the **Predicted Probability** for each individual in the leaf…

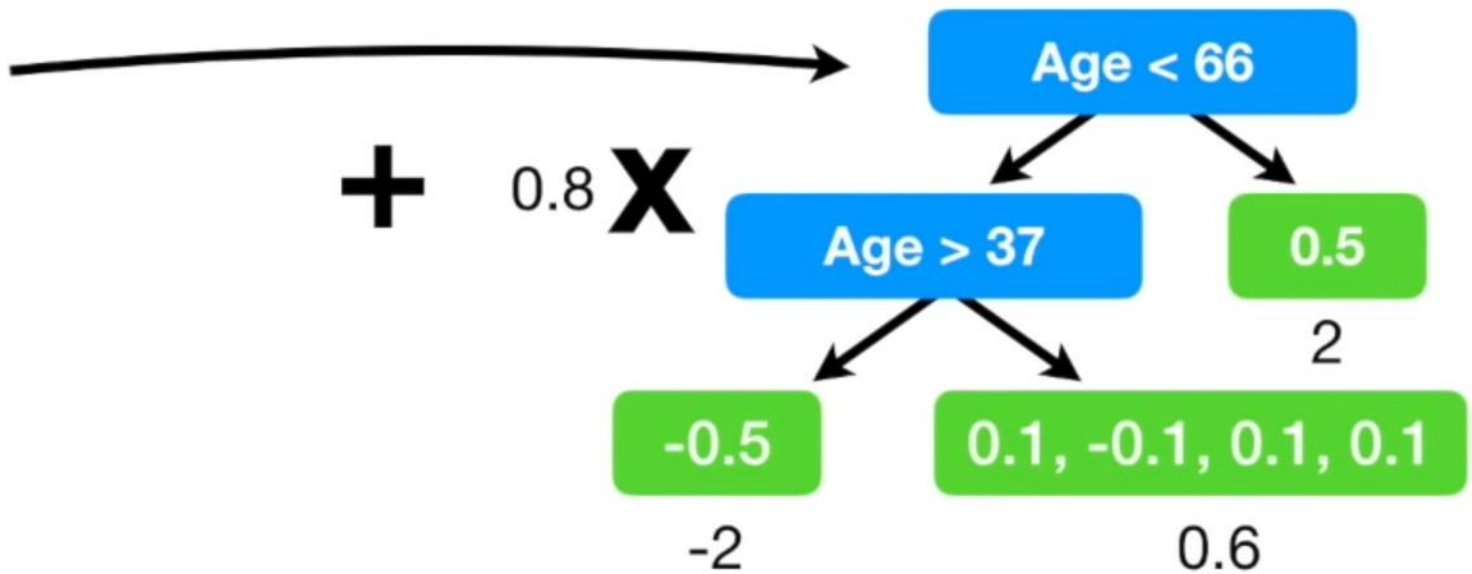
$$\frac{0.2}{0.09 + 0.09 + 0.09 + 0.09} = \boxed{0.6}$$

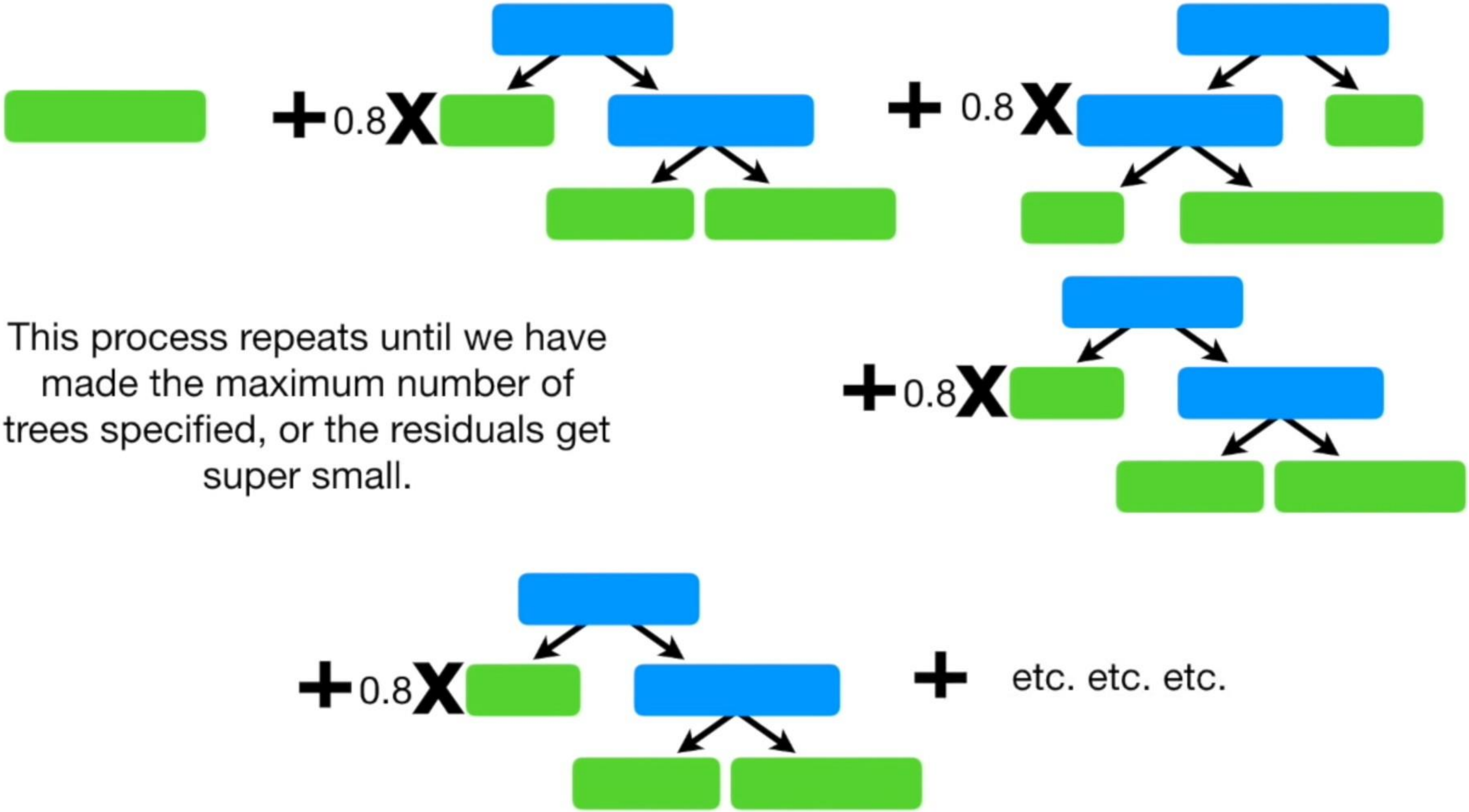


...and the **Output Value** for this leaf is **0.6**.

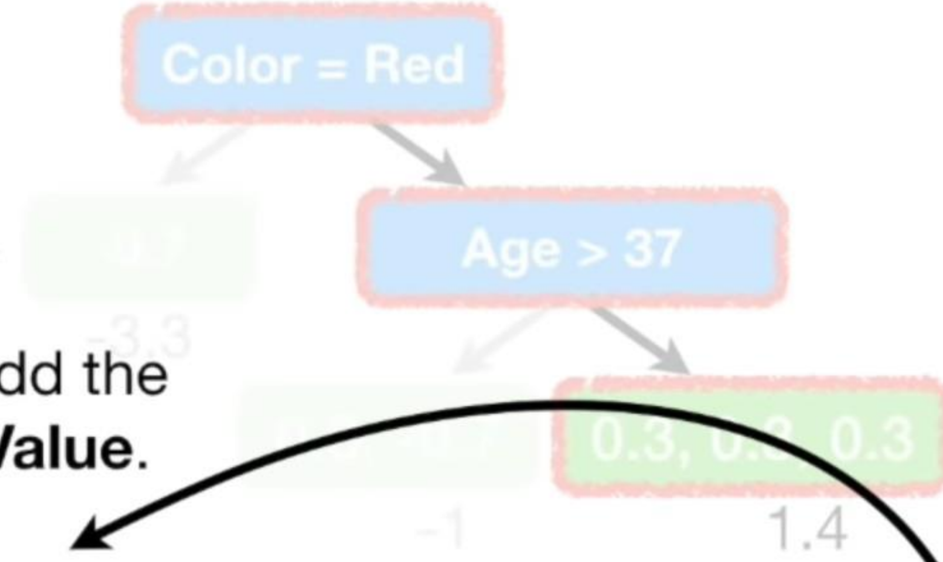
Likes Popcorn	Age	Favorite Color	Loves Troll 2	Predicted Prob.	Residual
Yes	12	Blue	Yes	0.9	0.1
Yes	87	Green	Yes	0.5	0.5
No	44	Blue	No	0.5	-0.5
Yes	19	Red	No	0.1	-0.1
No	32	Green	Yes	0.9	0.1
No	14	Blue	Yes	0.9	0.1

Then we built another tree based on the new **Residuals**, the difference between the **Observed** values and the values **Predicted** by the leaf ***and*** the first tree…




+ 0.8 X
+ 0.8 X
This process repeats until we have made the maximum number of trees specified, or the residuals get super small.
+ 0.8 X
+ 0.8 X
+ etc. etc. etc.

log(4/2) = 0.7 + 0.8 X

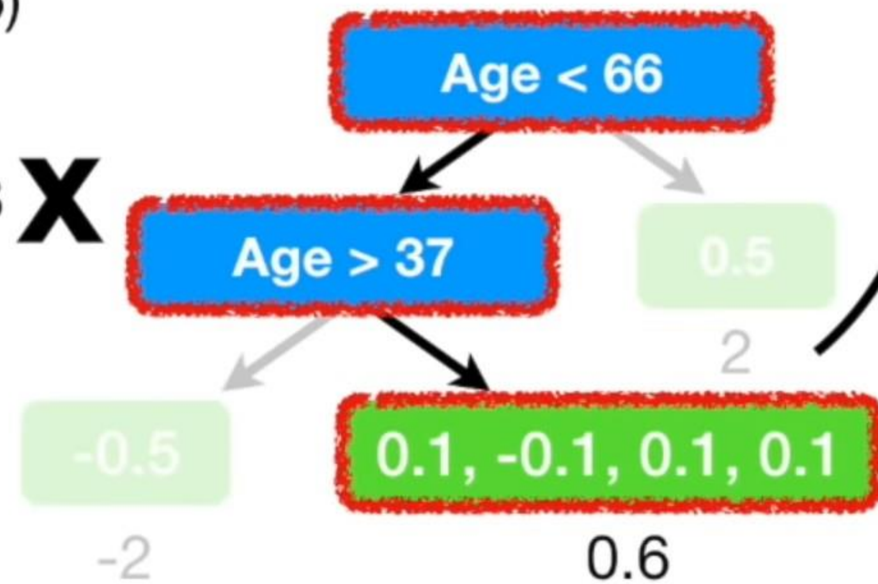


...and then we add the scaled **Output Value**.

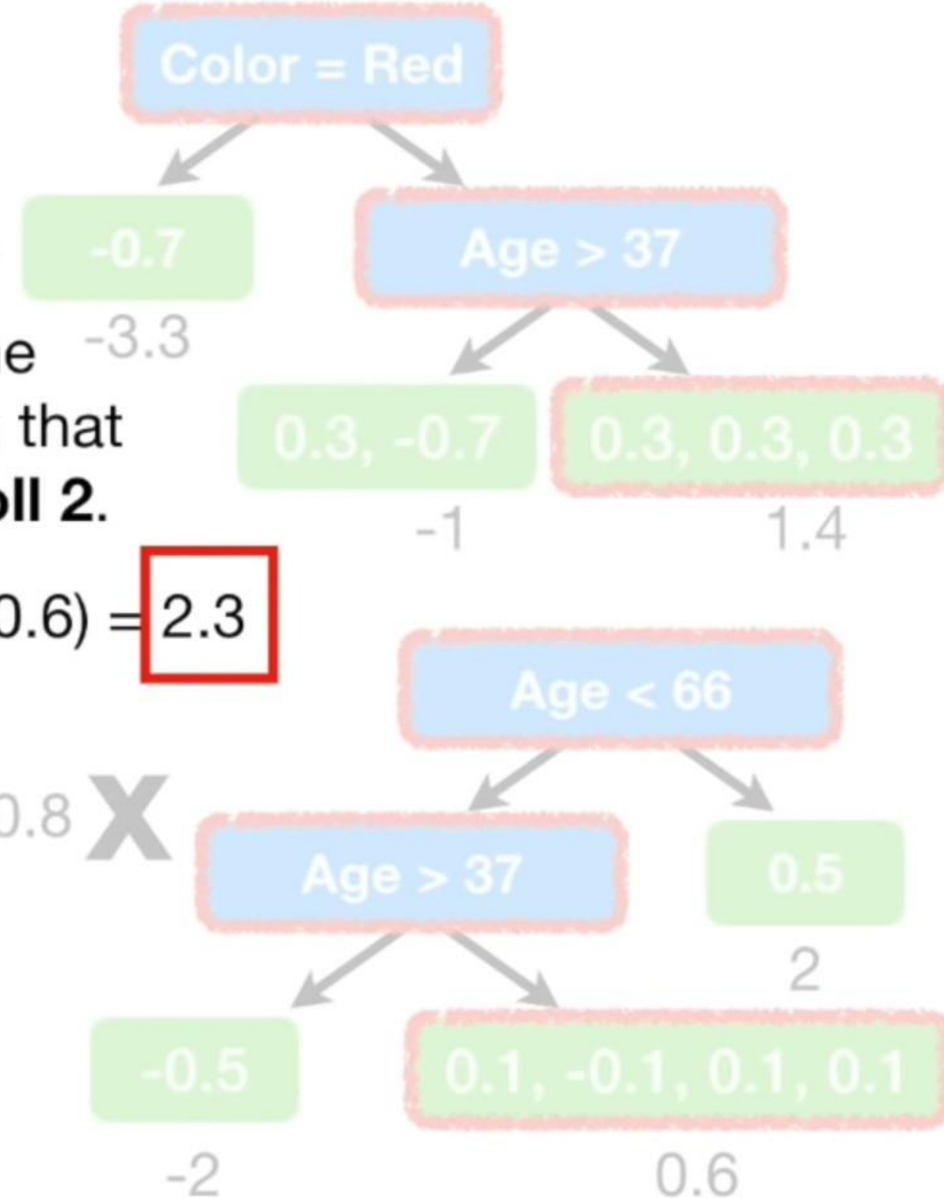
Log(odds) Prediction that someone **Loves Troll 2**:

$$= 0.7 + (0.8 \times 1.4) + (0.8 \times 0.6)$$

+ 0.8 X



Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	25	Green	**???**



…and get **2.3** as the **Log(odds) Prediction** that this person **Loves Troll 2**.

Log(odds) Prediction that someone **Loves Troll 2**:

$$= 0.7 + (0.8 \times 1.4) + (0.8 \times 0.6) = \boxed{2.3}$$

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	25	Green	**???**

Now we need to convert this **Log(odds)** into a **Probability**.

Log(odds) Prediction that someone **Loves Troll 2**: $= 0.7 + (0.8 \times 1.4) + (0.8 \times 0.6) = \boxed{2.3}$

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	25	Green	**???**

…and the **Predicted Probability** that this individual will **Love Troll 2** is **0.9**.

Log(odds) Prediction that someone **Loves Troll 2**: $= 0.7 + (0.8 \times 1.4) + (0.8 \times 0.6) = 2.3$

$$\text{Probability} = \frac{e^{2.3}}{1 + e^{2.3}} = \boxed{0.9}$$

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	25	Green	**???**

…we will **Classify** this person as someone who **Loves Troll 2**.

Log(odds) Prediction that someone **Loves Troll 2**: $= 0.7 + (0.8 \times 1.4) + (0.8 \times 0.6) = 2.3$

$$\text{Probability} = \frac{e^{2.3}}{1 + e^{2.3}} = 0.9$$

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	25	Green	**YES!!!**