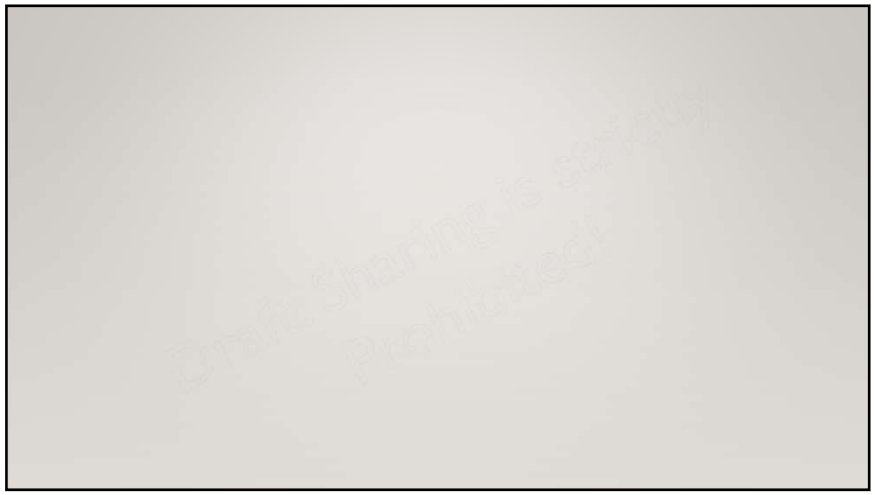
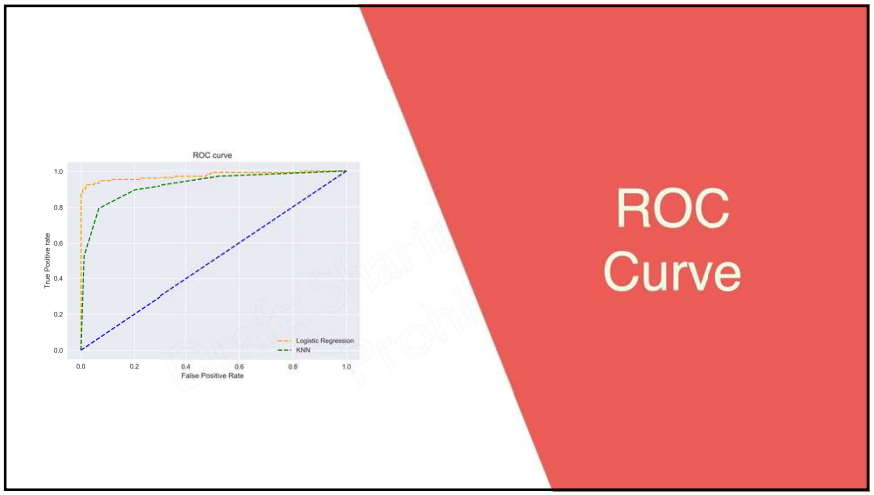
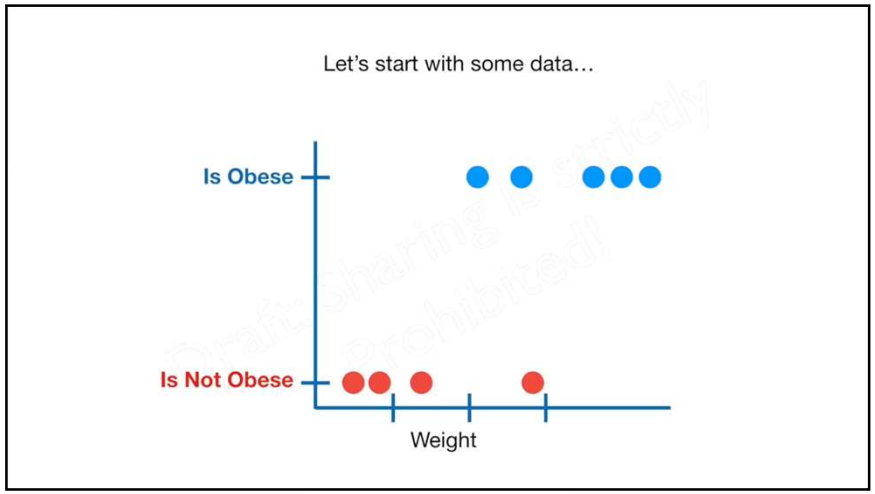
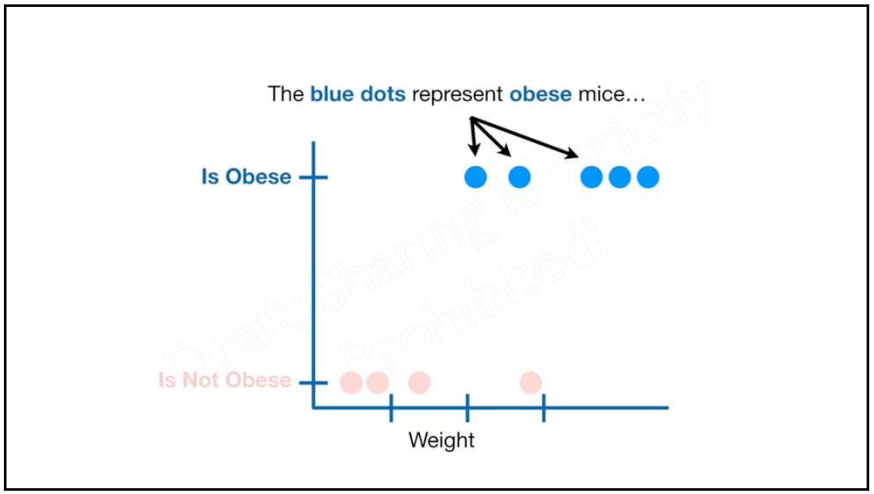
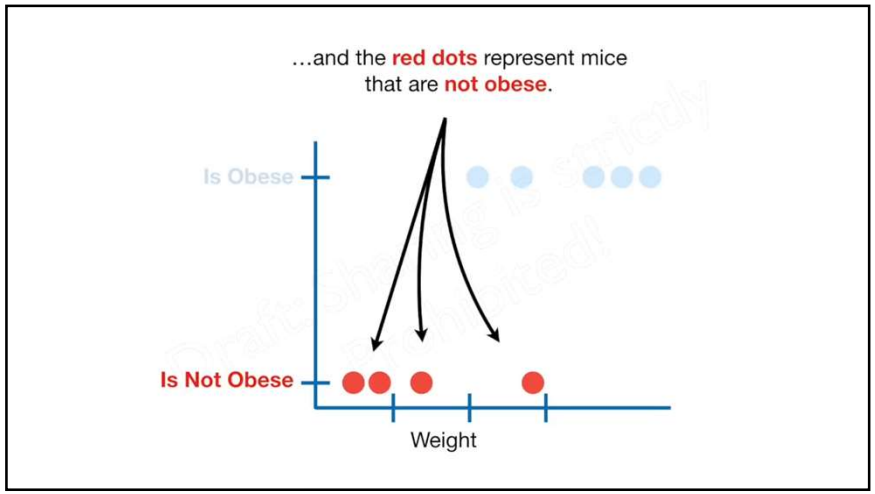
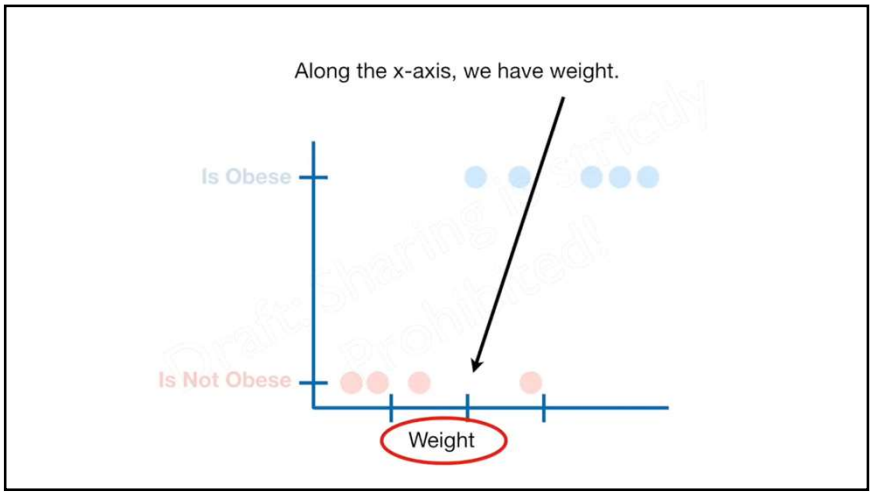


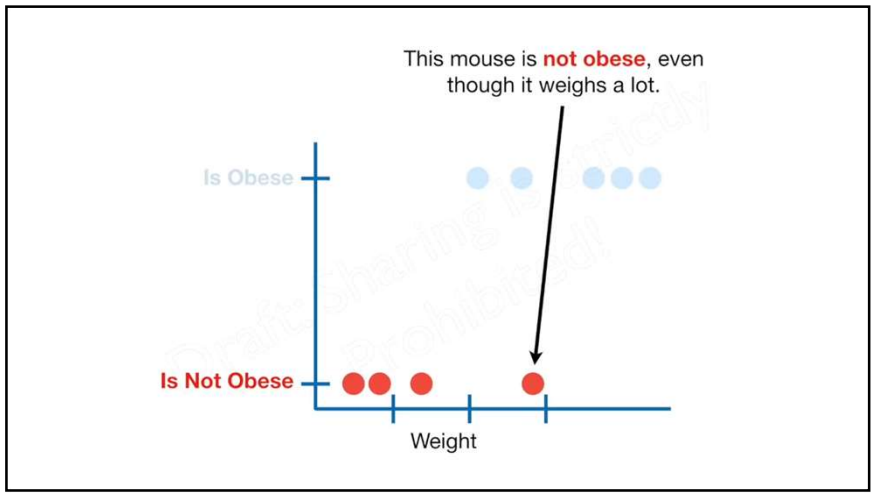

ROC curve
True Positive rate
False Positive Rate
Logistic Regression
KNN
ROC
Curve

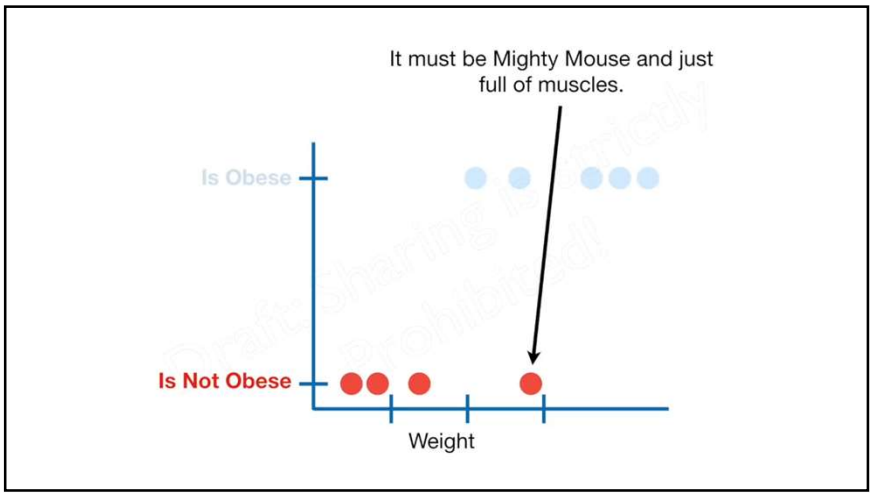

Let's start with some data…
Is Obese
Is Not Obese
Weight

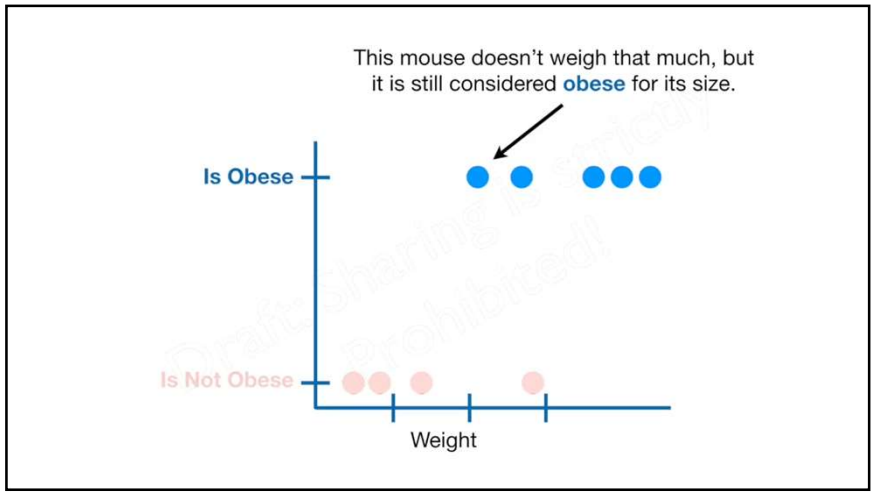

The blue dots represent obese mice…
Is Obese
Is Not Obese
Weight

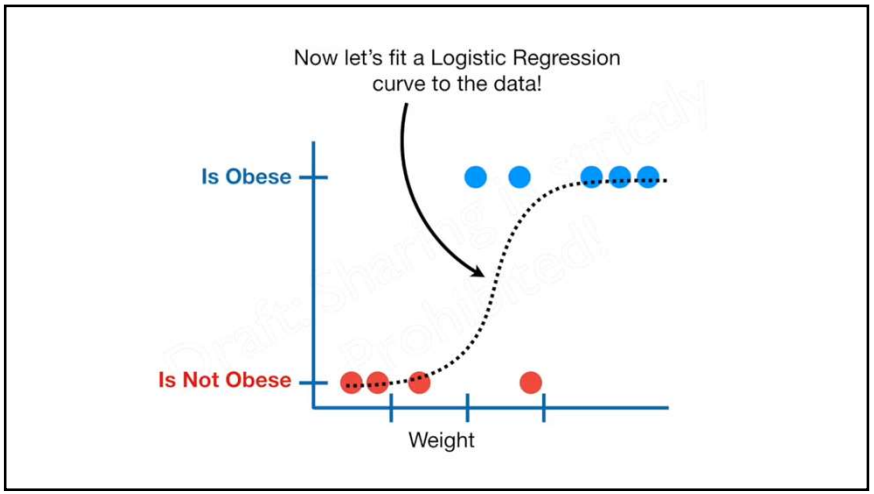

…and the red dots represent mice
that are not obese.
Is Obese
Is Not Obese
Weight

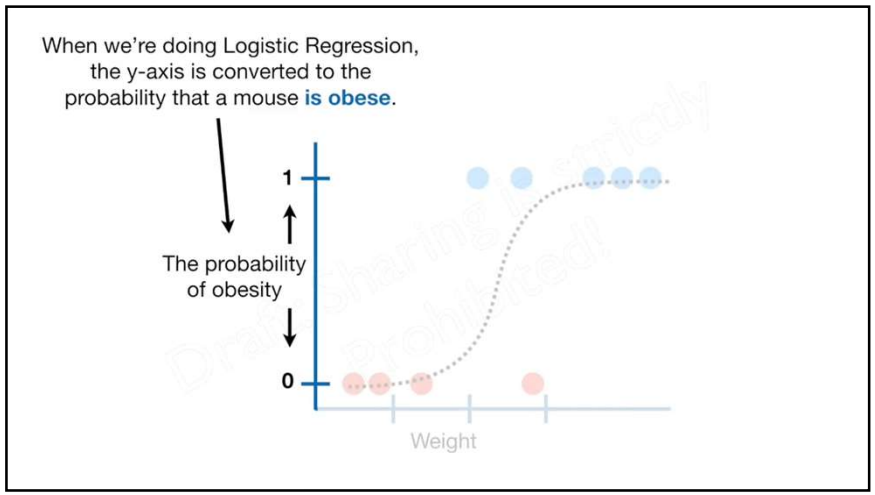

Along the x-axis, we have weight.
Is Obese
Is Not Obese
Weight

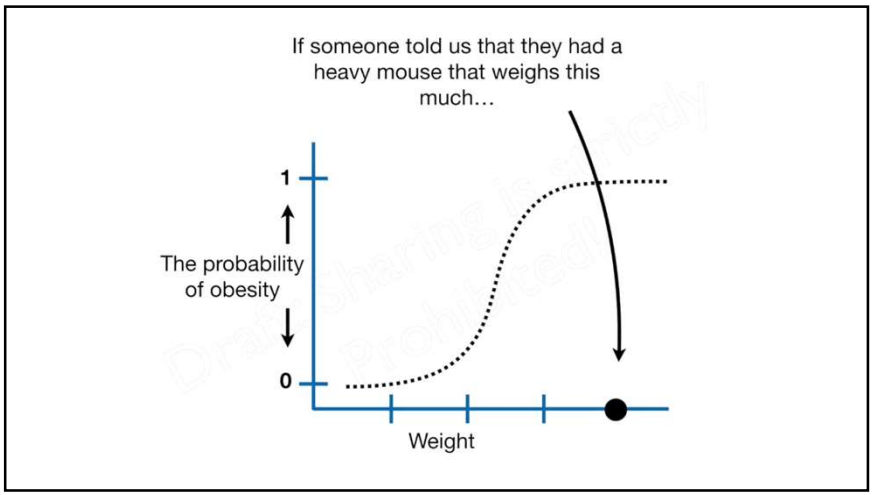

This mouse is not obese, even
though it weighs a lot.
Is Obese
Is Not Obese
Weight


It must be Mighty Mouse and just
full of muscles.
Is Obese
Is Not Obese
Weight

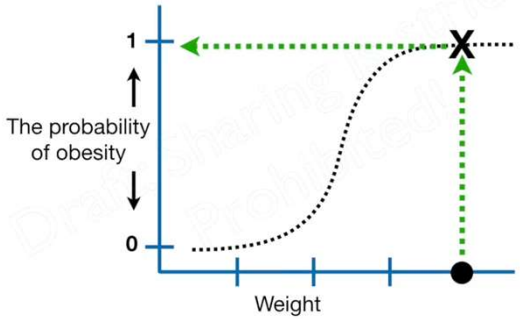

This mouse doesn't weigh that much, but it is still considered obese for its size.
Is Obese
Is Not Obese
Weight


Now let's fit a Logistic Regression curve to the data!
Is Obese
Is Not Obese
Weight

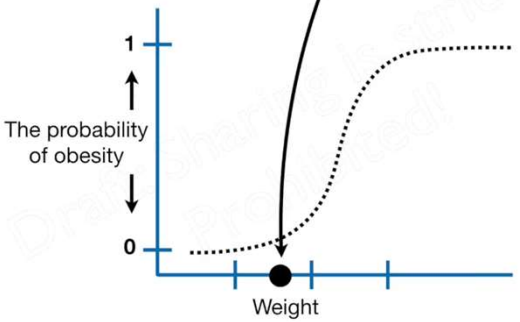

When we're doing Logistic Regression, the y-axis is converted to the probability that a mouse is obese.
1
The probability of obesity
0
Weight


If someone told us that they had a heavy mouse that weighs this much…
1
The probability of obesity
0
Weight

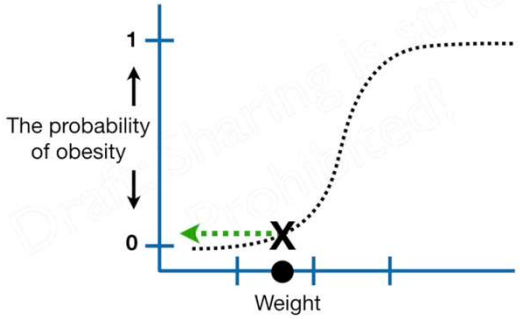
…then the curve would tell us that there is a ***high*** probability that the mouse **is obese**.



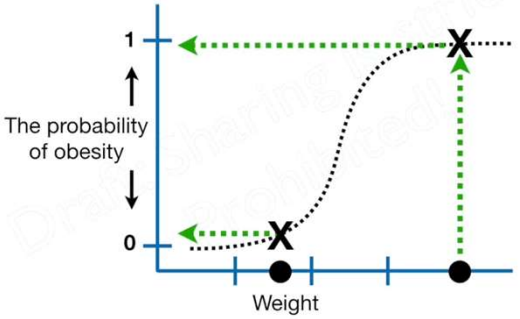

If someone told us that they had a light mouse that weighs this much…



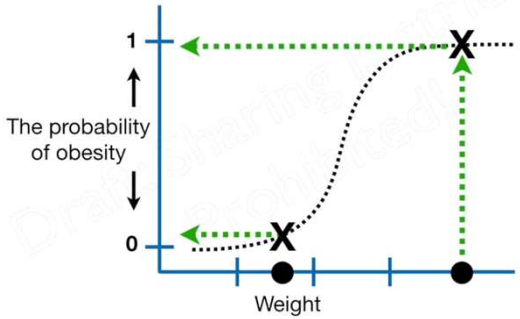

…then the curve would tell us that there is a ***low*** probability that the mouse **is obese**.



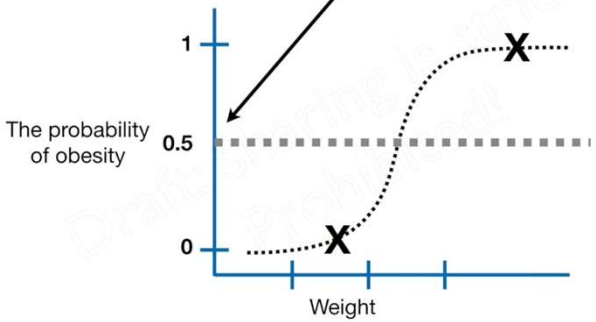

So this Logistic Regression tells us the ***probability*** that a mouse is **obese** based on its weight.



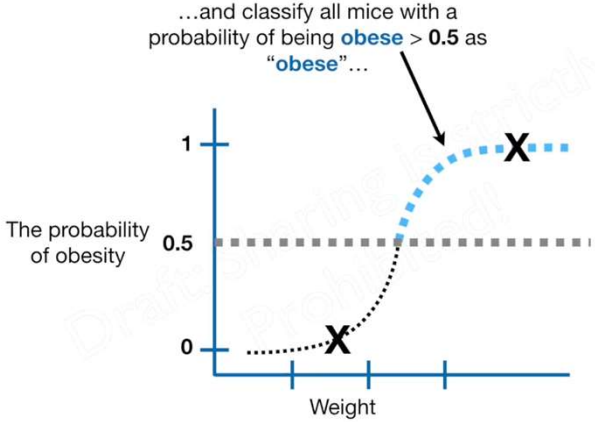
However, if we want to ***classify*** the mice as **obese** or **not obese**, then we need a way to turn probabilities into classifications.



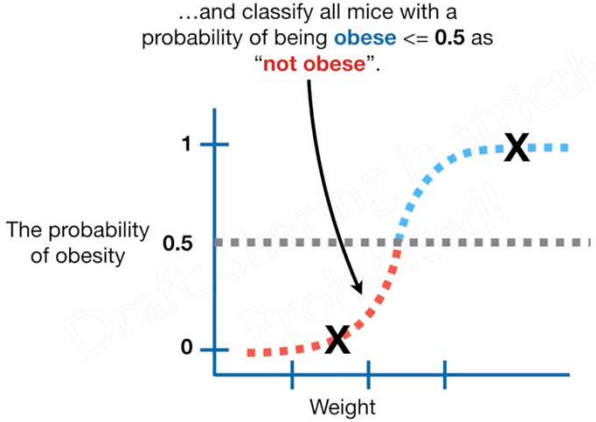

One way to classify mice is to set a threshold at **0.5**…

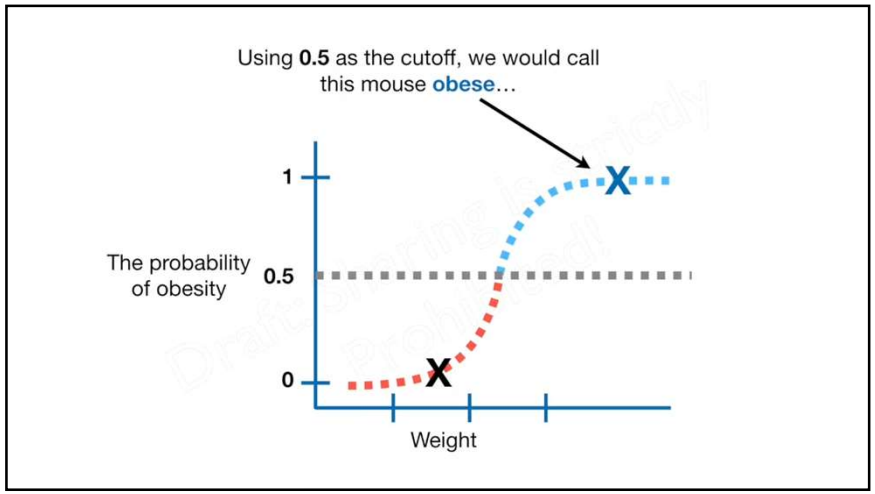


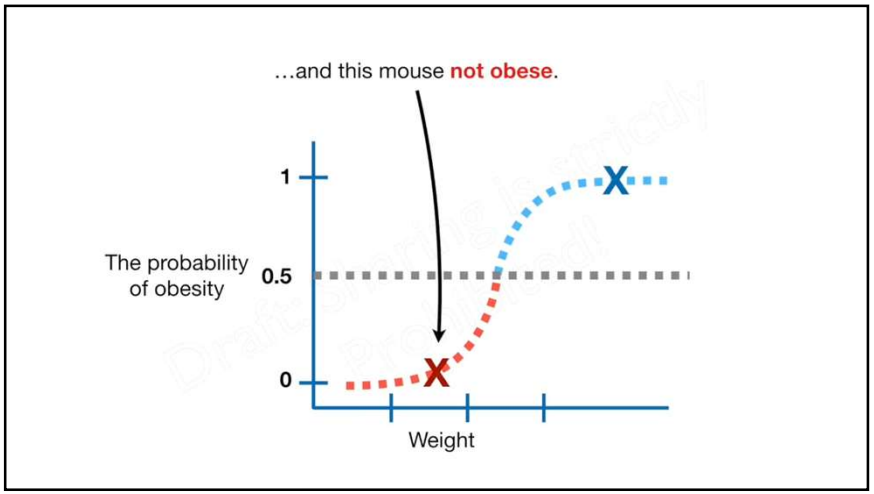

…and classify all mice with a probability of being **obese** > **0.5** as "**obese**"…

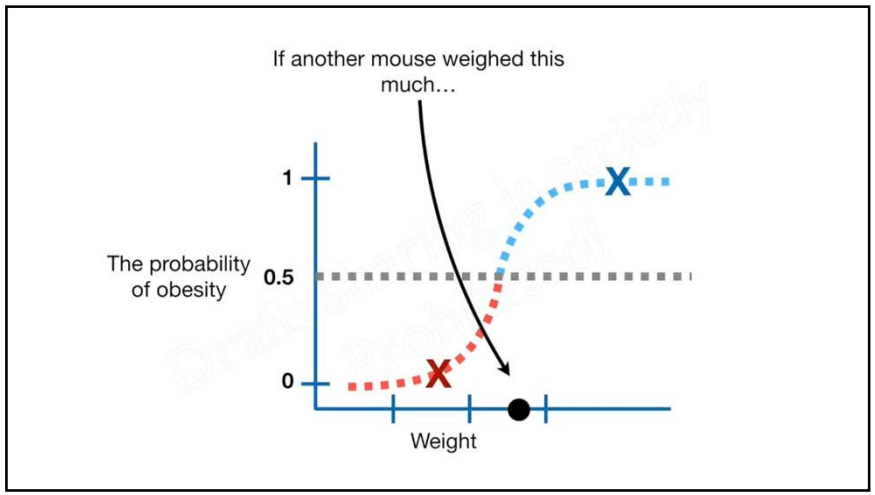


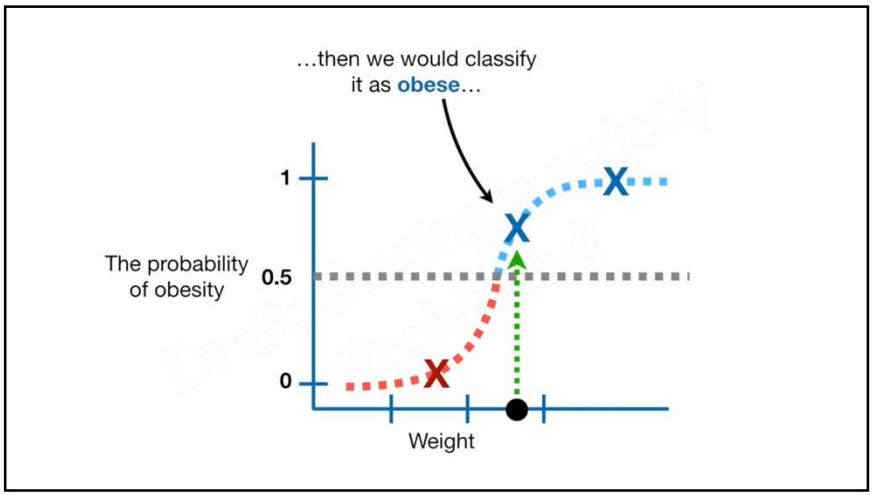

…and classify all mice with a probability of being **obese** <= **0.5** as "**not obese**".

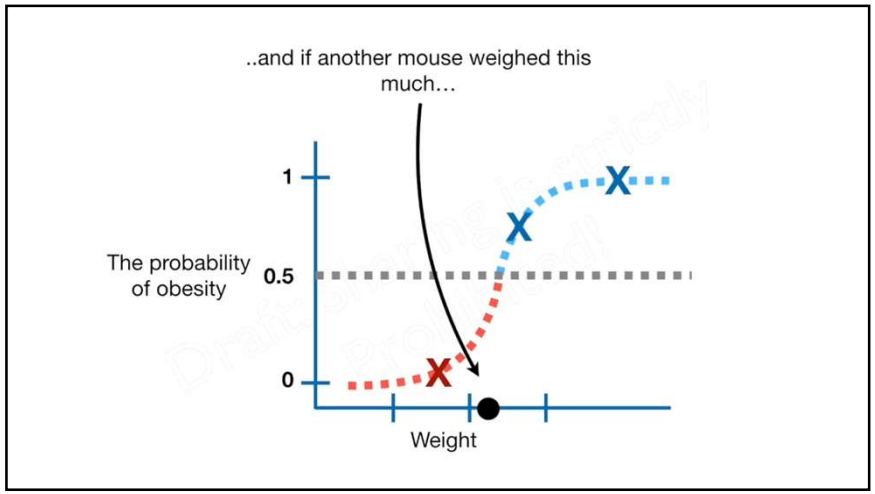


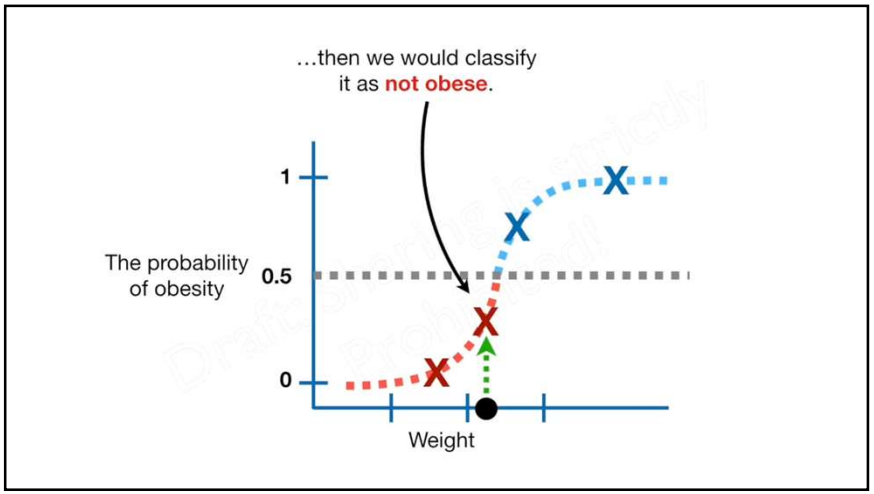

Using 0.5 as the cutoff, we would call this mouse obese…
1
0.5
0
The probability of obesity
Weight

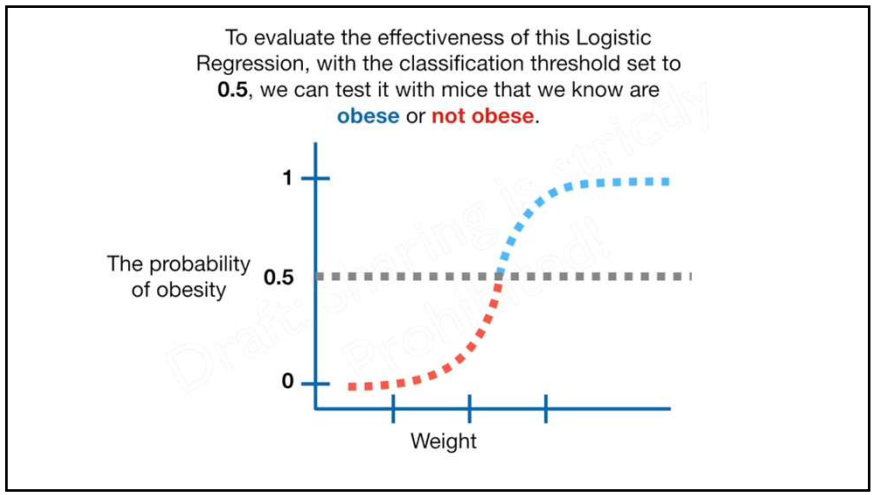

…and this mouse not obese.
1
0.5
0
The probability of obesity
Weight

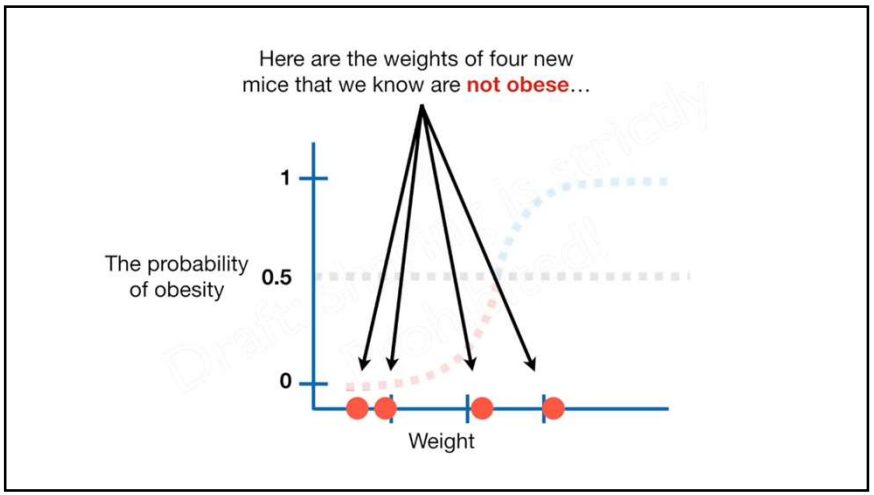

If another mouse weighed this much…
1
0.5
0
The probability of obesity
Weight

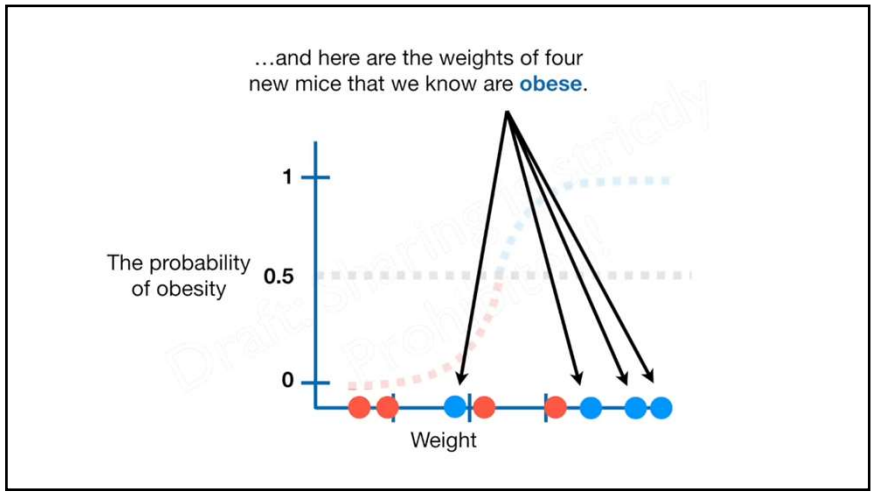

…then we would classify it as obese…
1
0.5
0
The probability of obesity
Weight

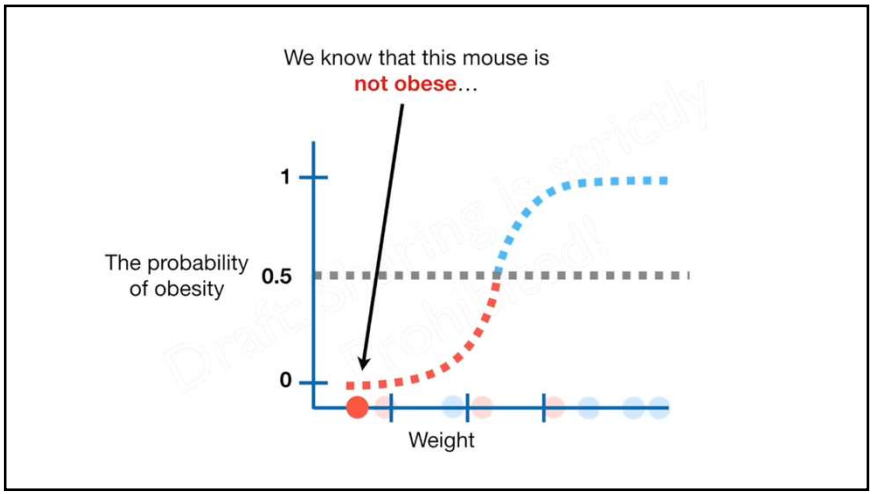

..and if another mouse weighed this much…
1
0.5
0
The probability of obesity
Weight

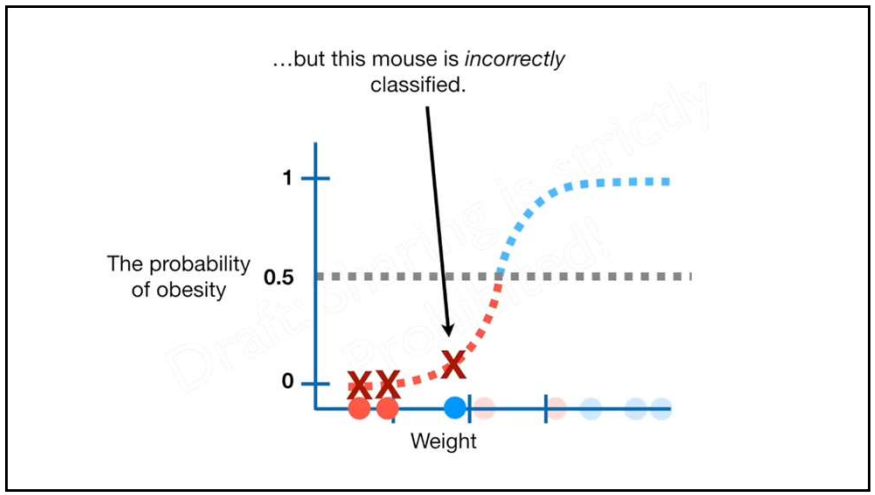

…then we would classify it as not obese.
1
0.5
0
The probability of obesity
Weight

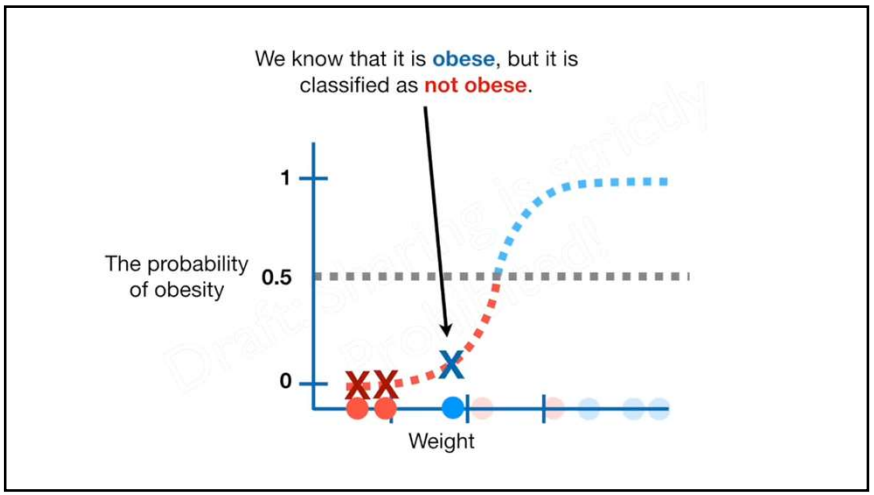

To evaluate the effectiveness of this Logistic Regression, with the classification threshold set to 0.5, we can test it with mice that we know are obese or not obese.
1
0.5
0
The probability of obesity
Weight

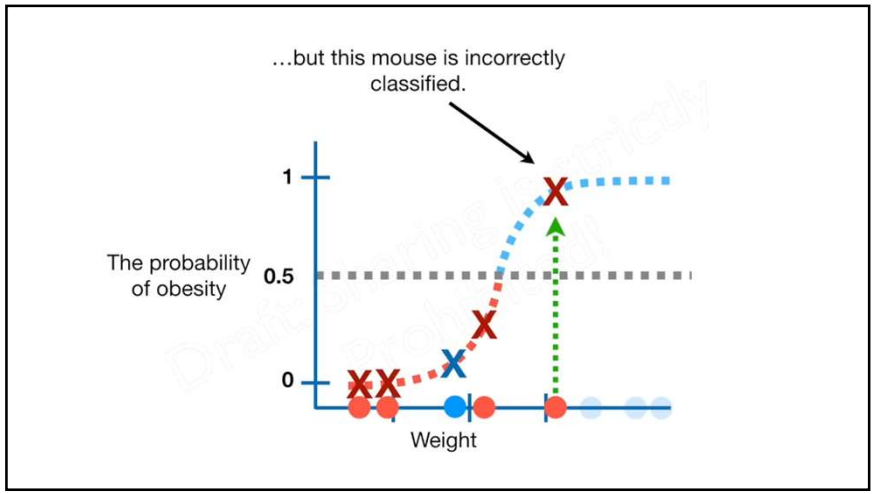

Here are the weights of four new mice that we know are not obese…
1
0.5
0
The probability of obesity
Weight

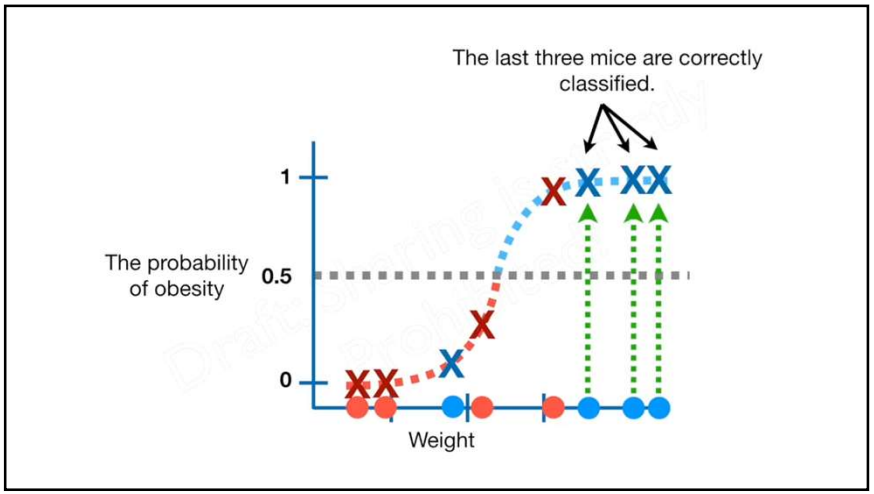

…and here are the weights of four
new mice that we know are obese.
The probability
of obesity
1
0.5
0
Weight

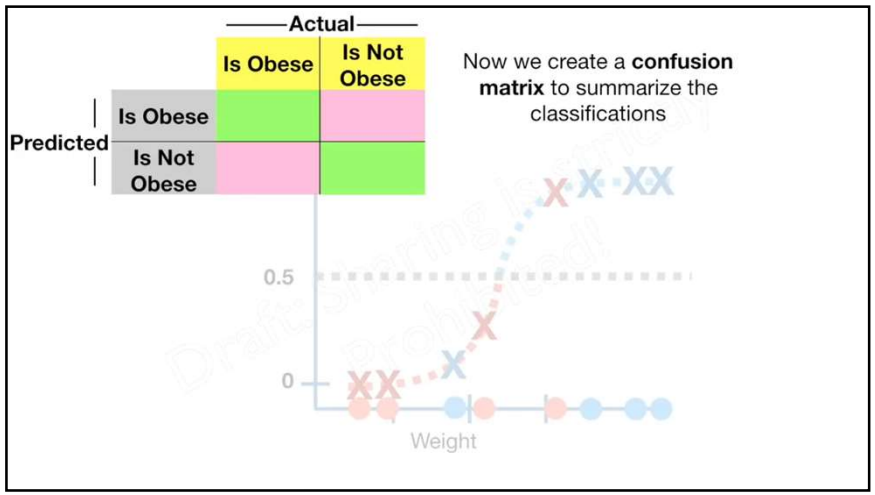

We know that this mouse is
not obese…
The probability
of obesity
1
0.5
0
Weight

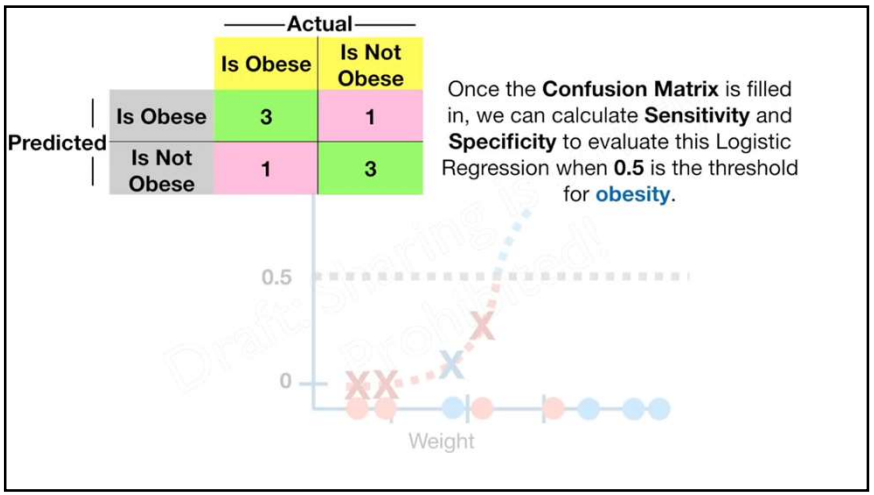

…but this mouse is incorrectly
classified.
The probability
of obesity
1
0.5
0
XX
X
Weight


We know that it is obese, but it is
classified as not obese.
The probability
of obesity
1
0.5
0
XX
X
Weight

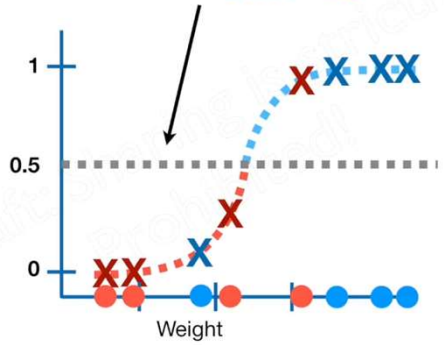

…but this mouse is incorrectly classified.
The probability of obesity
1
0.5
0
Weight


The last three mice are correctly classified.
The probability of obesity
1
0.5
0
Weight

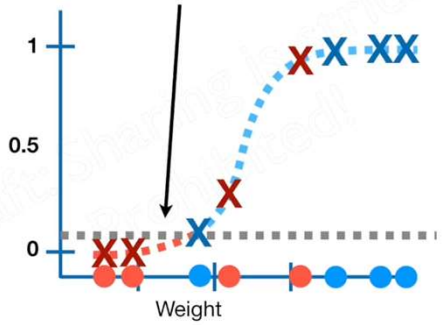

Actual
Is Obese
Is Not Obese
Predicted
Is Obese
Is Not Obese
Now we create a confusion matrix to summarize the classifications
0.5
0
Weight


Actual
Is Obese
Is Not Obese
Predicted
Is Obese
3
1
Is Not Obese
1
3
Once the Confusion Matrix is filled in, we can calculate Sensitivity and Specificity to evaluate this Logistic Regression when 0.5 is the threshold for obesity.
0.5
0
Weight

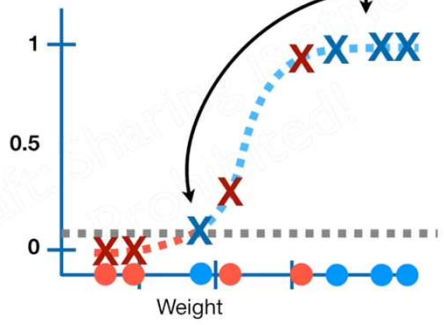
Now let's talk about what happens when we use a different threshold for deciding if a sample is **obese** or **not**.




For example, if it was super important to correctly classify every **obese** sample, we could set the threshold to **0.1**…

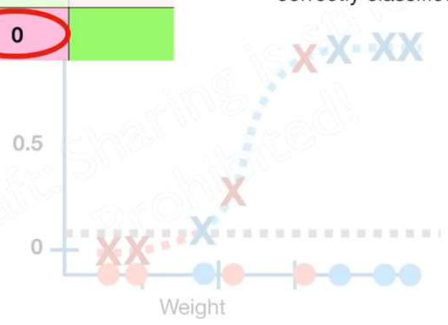


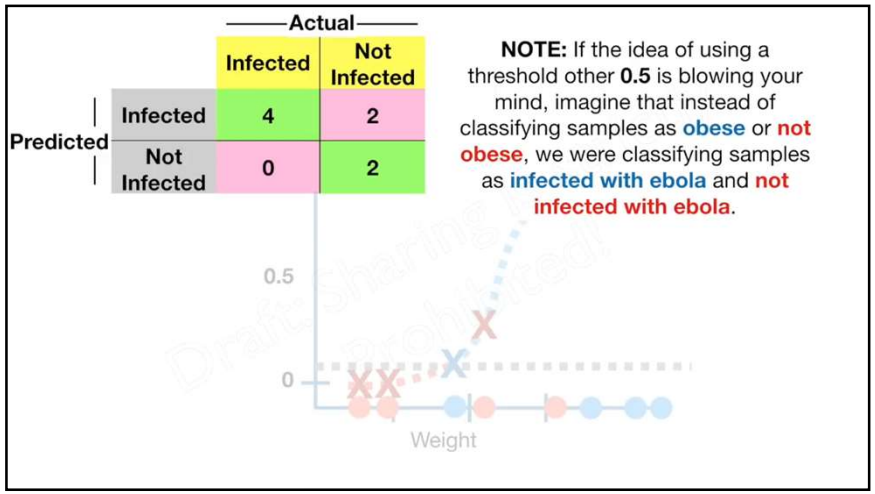

This would result in correct classifications for all **4** **obese** mice…

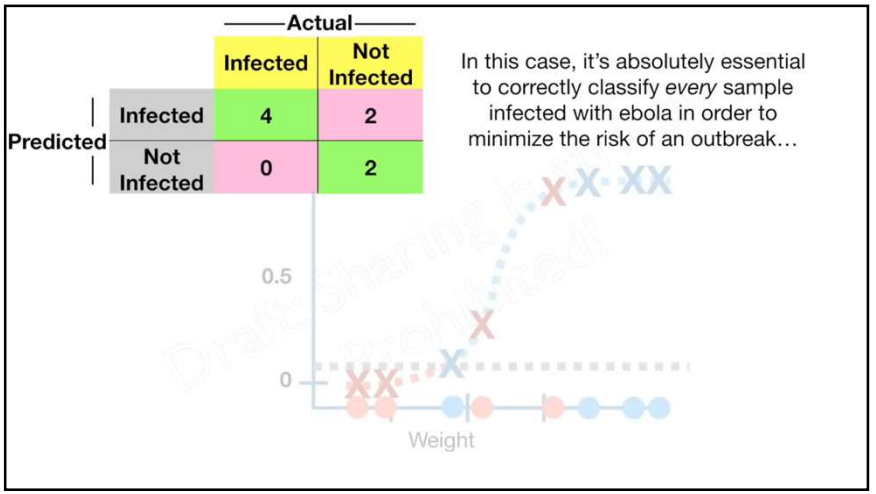


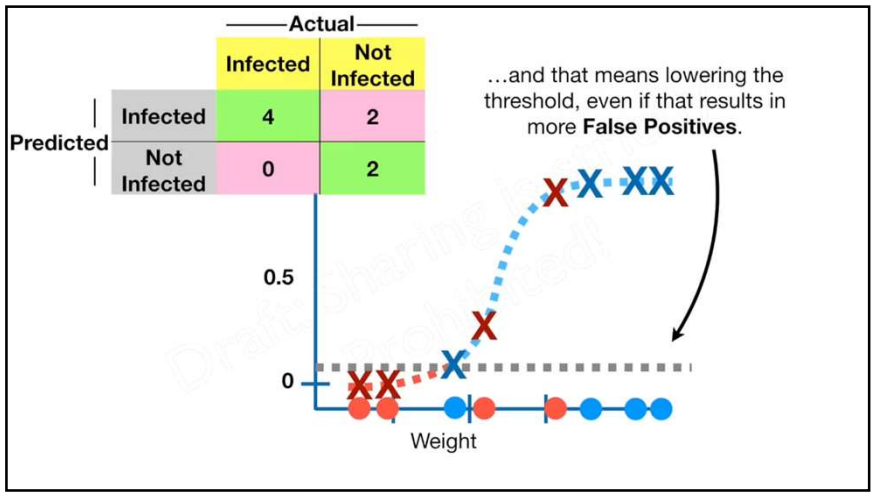

| | | Actual: Is Obese | Actual: Is Not Obese |
|---|---|---|---|
| Predicted | Is Obese | 4 | 2 |
| | Is Not Obese | 0 | |

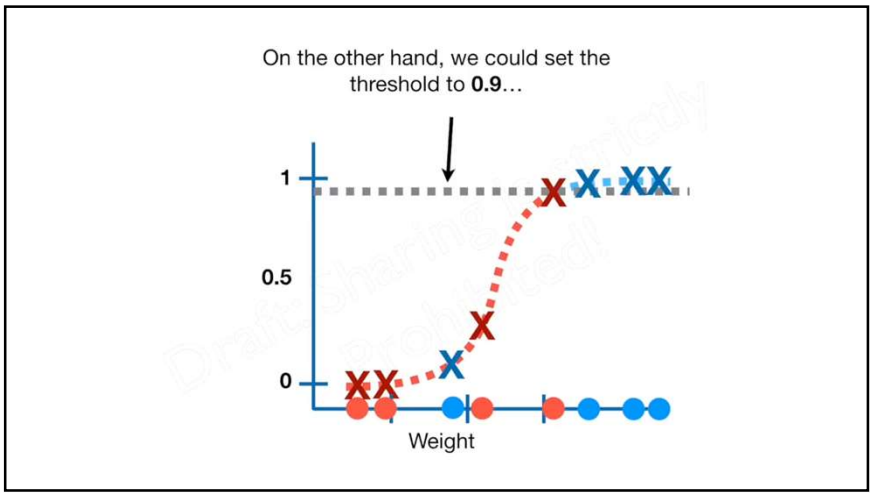
The lower threshold would also reduce the number of **False-Negatives**, because all of the **obese** mice were correctly classified…



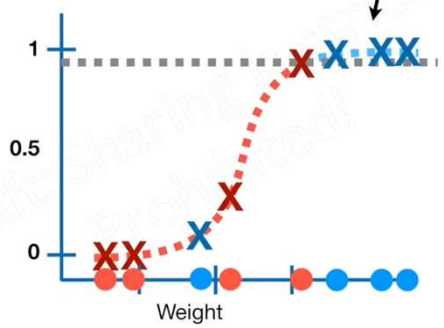

Actual
Infected
Not Infected
Predicted
Infected
4
2
Not Infected
0
2
NOTE: If the idea of using a threshold other 0.5 is blowing your mind, imagine that instead of classifying samples as obese or not obese, we were classifying samples as infected with ebola and not infected with ebola.
0.5
0
Weight
Draft. Sharing is strictly prohibited.


Actual
Infected
Not Infected
Predicted
Infected
4
2
Not Infected
0
2
In this case, it's absolutely essential to correctly classify *every* sample infected with ebola in order to minimize the risk of an outbreak…
0.5
0
Weight
Draft. Sharing is strictly prohibited.


Actual
Infected
Not Infected
Predicted
Infected
4
2
Not Infected
0
2
…and that means lowering the threshold, even if that results in more False Positives.
0.5
0
Weight
Draft. Sharing is strictly prohibited.

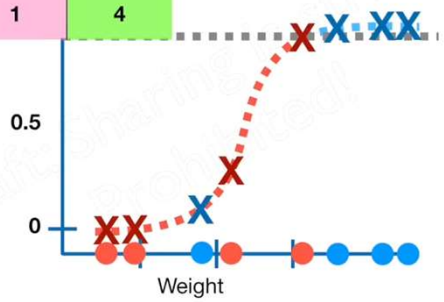

On the other hand, we could set the threshold to 0.9…
1
0.5
0
Weight
Draft. Sharing is strictly prohibited.

In this case, we would correctly classify the same number of **obese** samples as when the threshold was set to **0.5**…

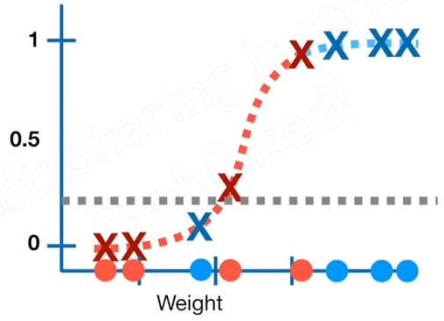



| | | Actual: Is Obese | Actual: Is Not Obese |
|---|---|---|---|
| Predicted | Is Obese | 3 | 0 |
| | Is Not Obese | 1 | 4 |

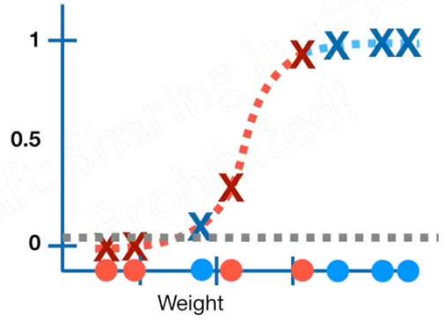
With this data, the higher threshold does a better job classifying samples as **obese** or **not obese**…



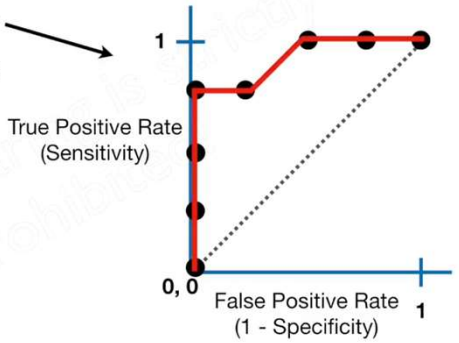

…but the threshold could be set to anything between 0 and 1.



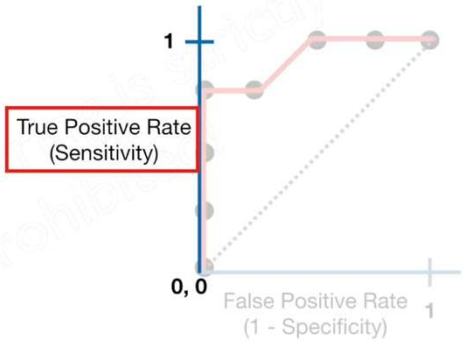

How do we determine which threshold is the best?



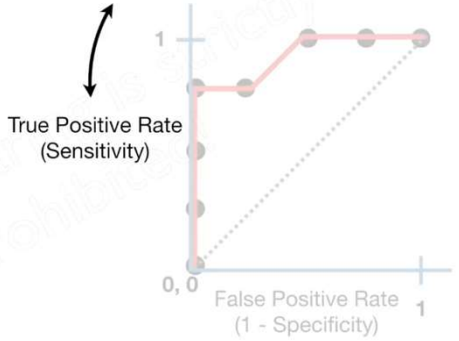
So instead of being overwhelmed with confusion matrices, **Receiver Operator Characteristic** (**ROC**) graphs provide a simple way to summarize all of the information.




The y-axis shows the **True Positive Rate**, which is the same thing as **Sensitivity**.




$$\text{True Positive Rate} = \text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

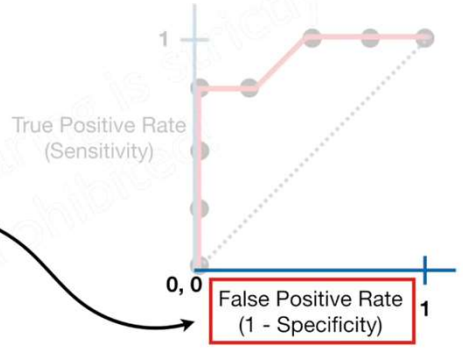



$$\text{True Positive Rate} = \text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

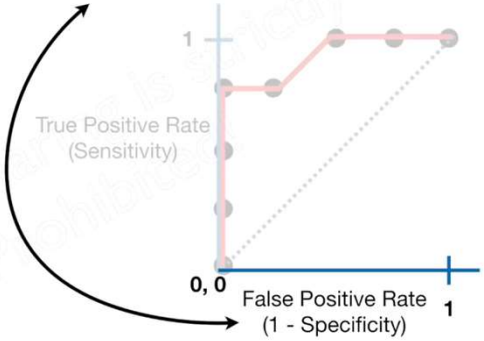
| | | Actual: Is Obese | Actual: Is Not Obese |
|---|---|---|---|
| Predicted | Is Obese | True Positives | False Positives |
| Predicted | Is Not Obese | False Negatives | True Negatives |

The **True Positive Rate** tells you what proportion of **obese** samples were correctly classified.

The x-axis shows the **False Positive Rate**, which is the same thing as **1 - Specificity**.




$$\text{False Positive Rate} = (1 - \text{Specificity}) = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

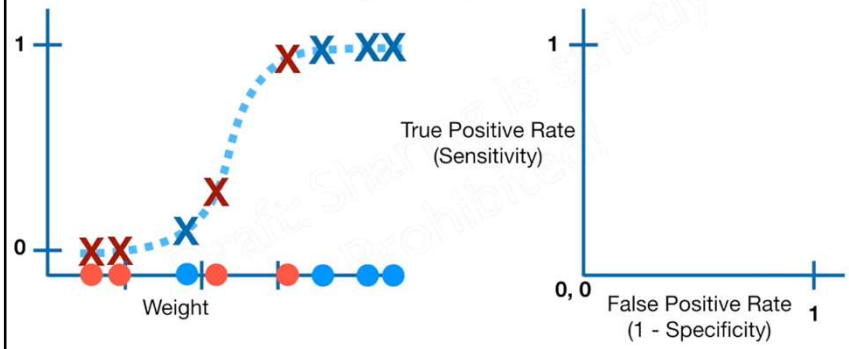


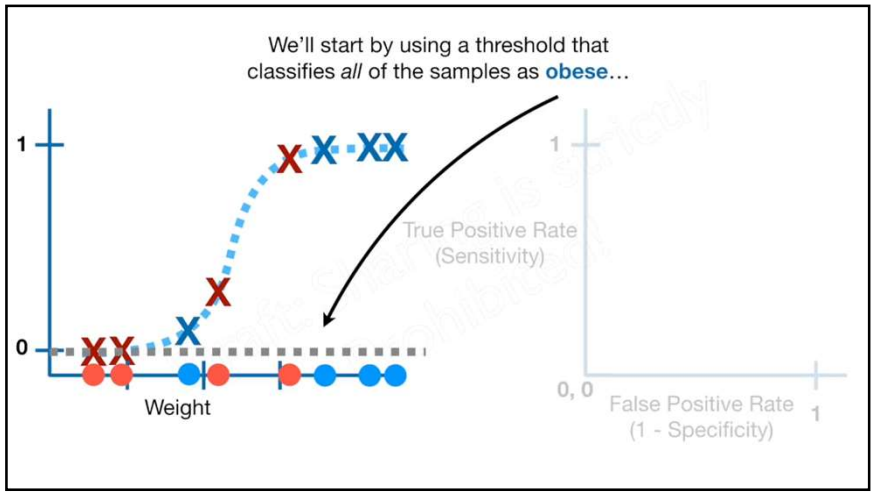

$$\text{False Positive Rate} = (1 - \text{Specificity}) = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

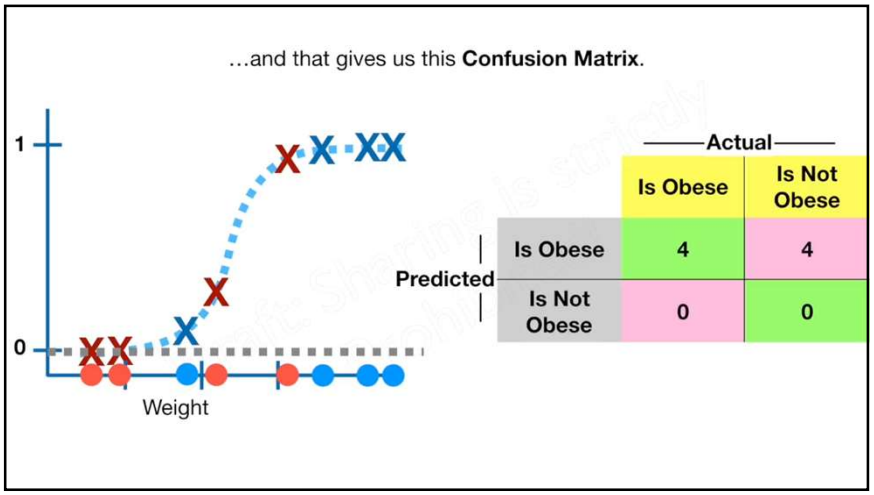
| | | Actual: Is Obese | Actual: Is Not Obese |
|---|---|---|---|
| Predicted | Is Obese | True Positives | **False Positives** |
| Predicted | Is Not Obese | False Negatives | True Negatives |

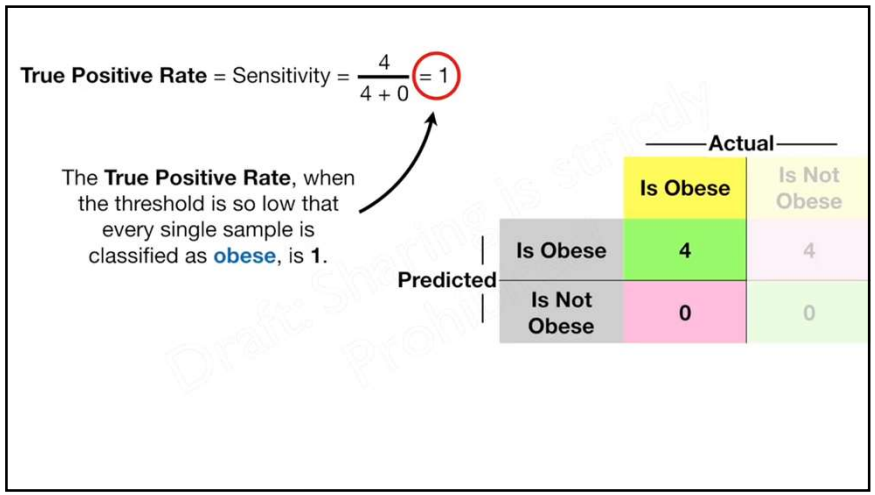
The **False Positive Rate** tells you the proportion of **not obese** samples that were incorrectly classified and are **False Positives**.

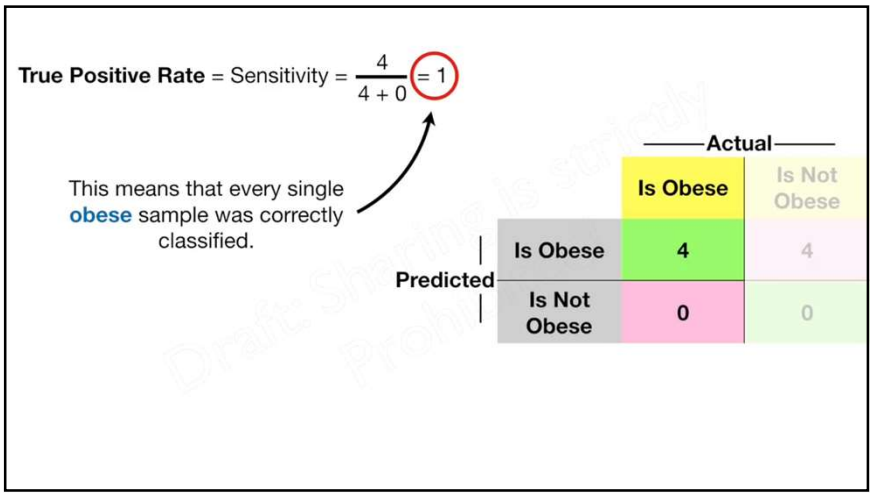
To get a better sense of how the **ROC** works, let's draw one from start to finish using our example data.




We'll start by using a threshold that classifies *all* of the samples as **obese**...
1
0
Weight
True Positive Rate (Sensitivity)
1
0, 0
False Positive Rate (1 - Specificity)
1


...and that gives us this **Confusion Matrix**.
1
0
Weight
Actual
Is Obese
Is Not Obese
Predicted
Is Obese
4
4
Is Not Obese
0
0


**True Positive Rate** = Sensitivity = $\frac{4}{4+0} = 1$
The **True Positive Rate**, when the threshold is so low that every single sample is classified as **obese**, is **1**.
Actual
Is Obese
Is Not Obese
Predicted
Is Obese
4
4
Is Not Obese
0
0


**True Positive Rate** = Sensitivity = $\frac{4}{4+0} = 1$
This means that every single **obese** sample was correctly classified.
Actual
Is Obese
Is Not Obese
Predicted
Is Obese
4
4
Is Not Obese
0
0

**True Positive Rate** = Sensitivity = $\frac{4}{4+0}$ = 1

**False Positive Rate** = 1 - Specificity = $\frac{4}{4+0}$ = 1

The **False Positive Rate**, when the threshold is so low that every single sample is classified as **obese**, is also **1**.

| | | Actual: Is Obese | Actual: Is Not Obese |
|---|---|---|---|
| Predicted | Is Obese | 4 | 4 |
| | Is Not Obese | 0 | 0 |



**True Positive Rate** = Sensitivity = $\frac{4}{4+0}$ = 1

**False Positive Rate** = 1 - Specificity = $\frac{4}{4+0}$ = 1

This means that every single sample that was **not obese** was *incorrectly* classified as **obese**.

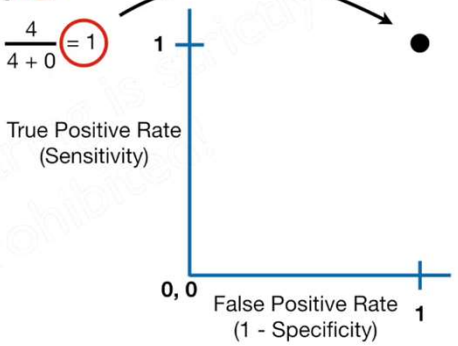
| | | Actual: Is Obese | Actual: Is Not Obese |
|---|---|---|---|
| Predicted | Is Obese | 4 | 4 |
| | Is Not Obese | 0 | 0 |

Draft. Sharing is strictly Prohibited.

**True Positive Rate** = Sensitivity = $\frac{4}{4+0}$ = 1

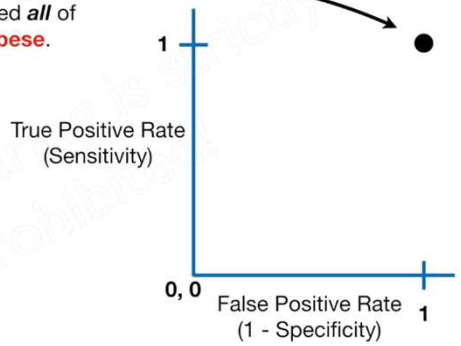
**False Positive Rate** = 1 - Specificity = $\frac{4}{4+0}$ = 1

Now plot a point at **1, 1**.



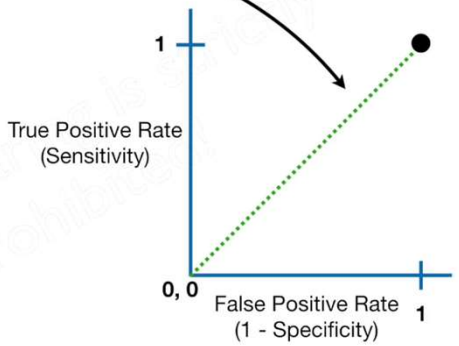

Draft. Sharing is strictly Prohibited.

A point at **1, 1** means that even though we correctly classified ***all*** of the **obese** samples, we *incorrectly* classified ***all*** of the samples that were **not obese**.




Draft. Sharing is strictly Prohibited.

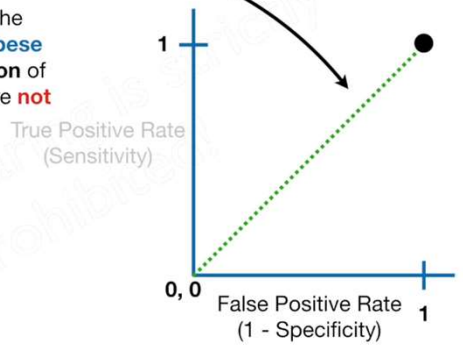
This **green diagonal line** shows where the **True Positive Rate** = **False Positive Rate**





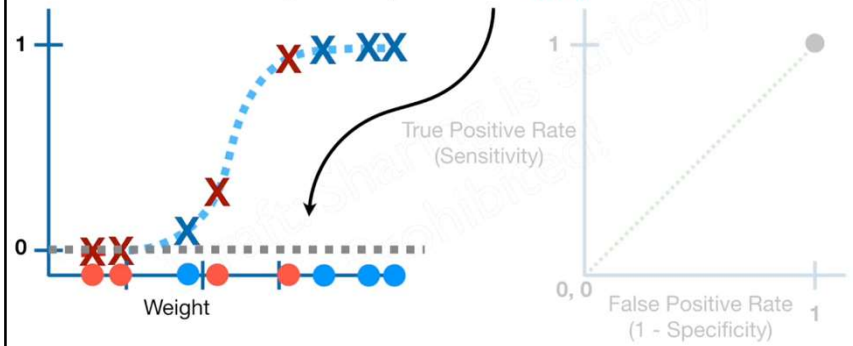

This **green diagonal line** shows where the **True Positive Rate** = **False Positive Rate**

Any point on this **line** means that the **proportion** of ***correctly*** classified **obese** samples is the same as the **proportion** of ***incorrectly*** classified samples that are **not obese**.



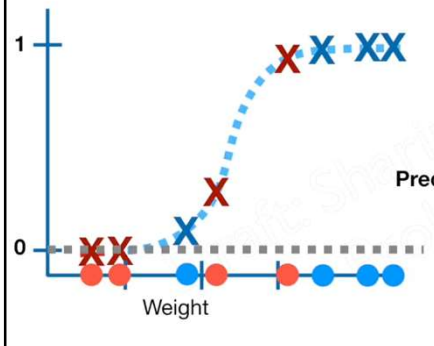



Going back to the Logistic Regression, let's increase the threshold so that *all but the lightest sample* are called **obese**.






The new threshold gives us this **Confusion Matrix**.

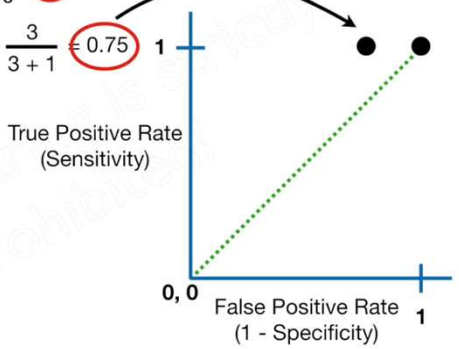



| | | Actual: Is Obese | Actual: Is Not Obese |
|---|---|---|---|
| Predicted | Is Obese | 4 | 3 |
| Predicted | Is Not Obese | 0 | 1 |

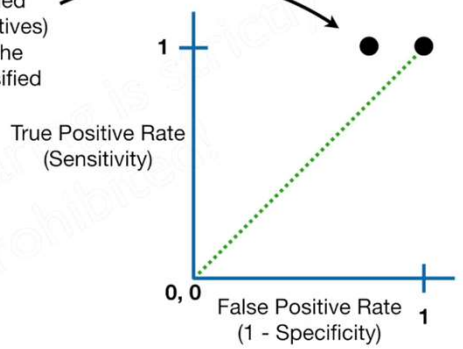
**True Positive Rate** = Sensitivity = $\frac{4}{4+0}$ = 1

**False Positive Rate** = 1 - Specificity = $\frac{3}{3+1}$ = 0.75

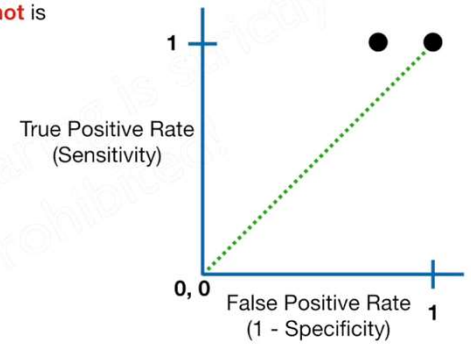
…and plot a point at **0.75, 1**.



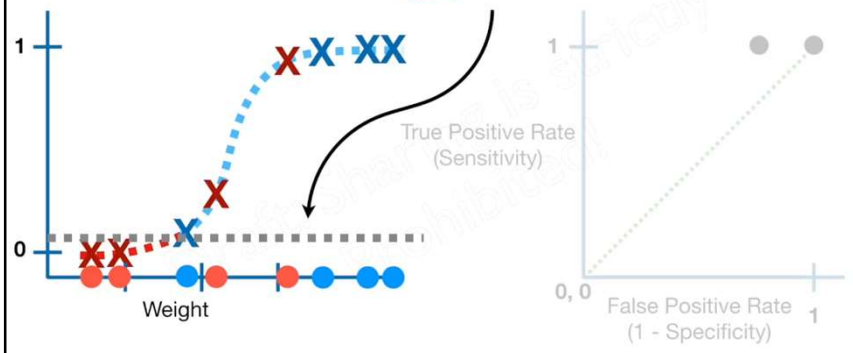

Since the new point (**0.75, 1**) is to the left of the **dotted green line**, we know that the proportion of correctly classified samples that were **obese** (true positives) *is greater* than the proportion of the samples that were incorrectly classified as **obese** (false positives).

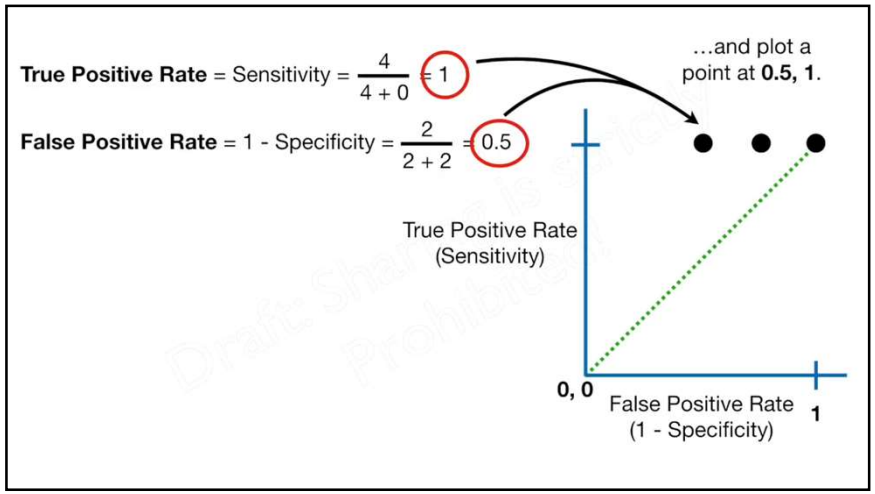


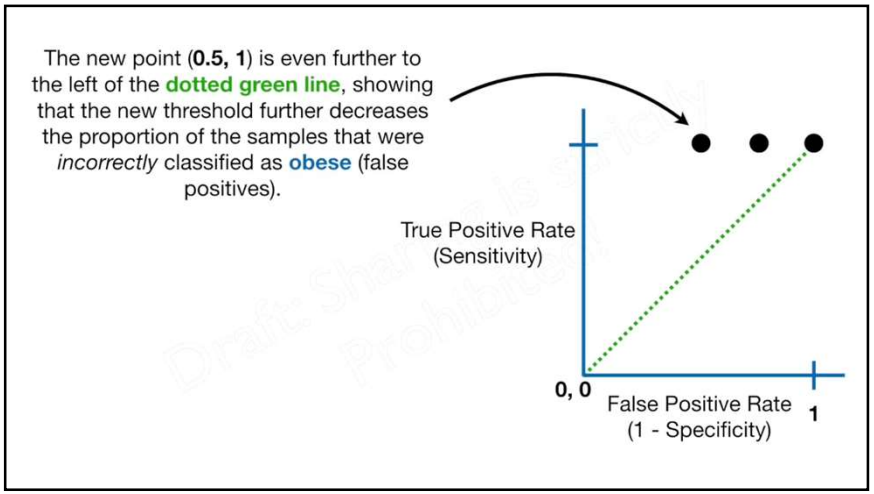

In other words, the new threshold for deciding if a sample is **obese** or **not** is better than the first one.

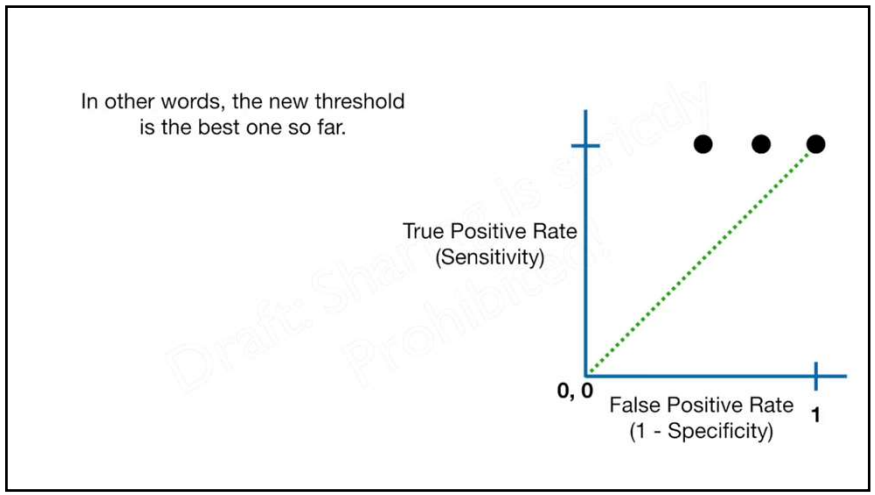


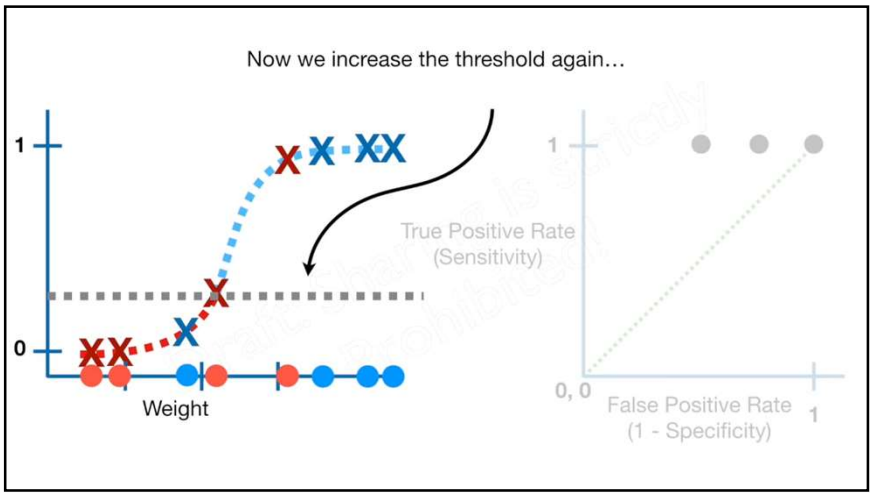

Now let's increase the threshold so that all but the two lightest samples are called **obese**.




True Positive Rate = Sensitivity = 4 / (4 + 0) = 1
False Positive Rate = 1 - Specificity = 2 / (2 + 2) = 0.5
…and plot a point at 0.5, 1.
True Positive Rate (Sensitivity)
0, 0
False Positive Rate (1 - Specificity)
1


The new point (0.5, 1) is even further to the left of the dotted green line, showing that the new threshold further decreases the proportion of the samples that were incorrectly classified as obese (false positives).
True Positive Rate (Sensitivity)
0, 0
False Positive Rate (1 - Specificity)
1


In other words, the new threshold is the best one so far.
True Positive Rate (Sensitivity)
0, 0
False Positive Rate (1 - Specificity)
1


Now we increase the threshold again…
1
0
Weight
True Positive Rate (Sensitivity)
0, 0
False Positive Rate (1 - Specificity)
1

**True Positive Rate** = Sensitivity = $\frac{3}{3+1}$ = 0.75

**False Positive Rate** = 1 - Specificity = $\frac{1}{1+3}$ = 0.25

…calculate the **True Positive Rate** and the **False Positive Rate**…

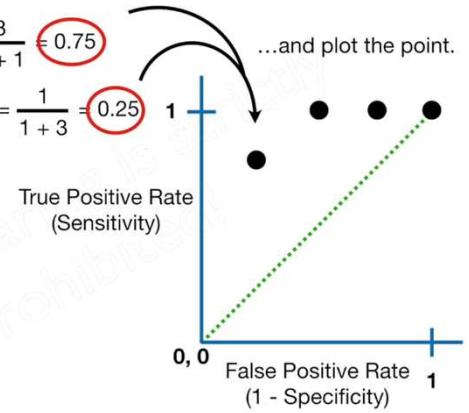
| | | Actual: Is Obese | Actual: Is Not Obese |
|---|---|---|---|
| Predicted | Is Obese | 3 | 1 |
| Predicted | Is Not Obese | 1 | 3 |



**True Positive Rate** = Sensitivity = $\frac{3}{3+1}$ = 0.75

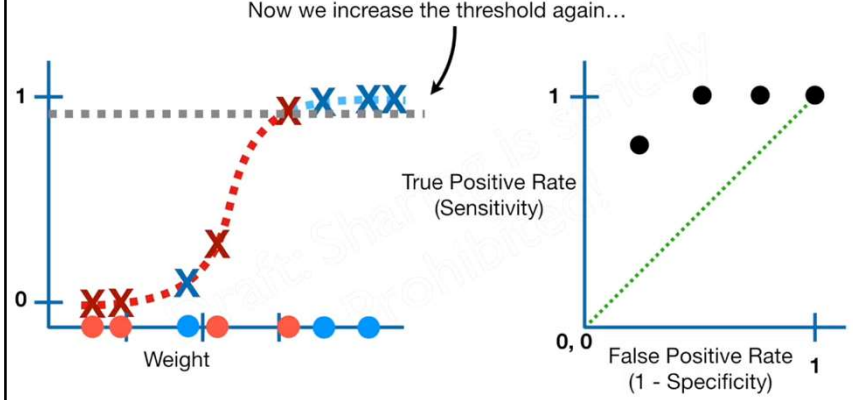
**False Positive Rate** = 1 - Specificity = $\frac{1}{1+3}$ = 0.25

…and plot the point.






Now we increase the threshold again…



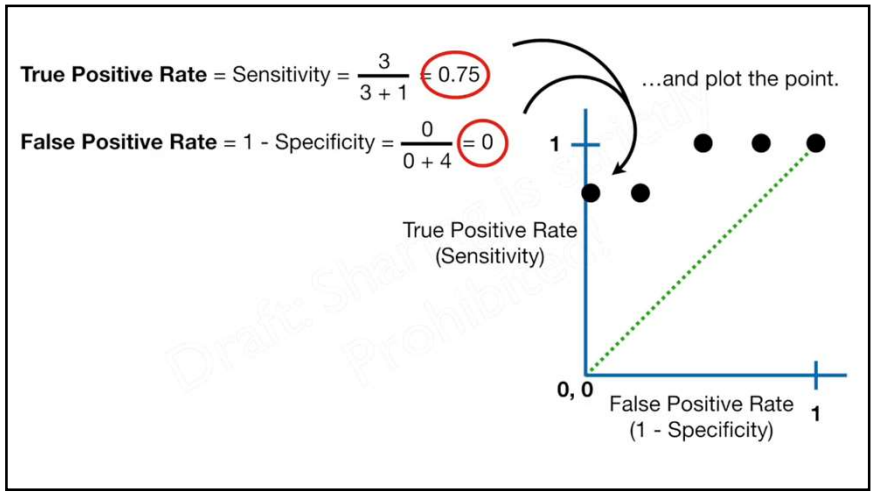


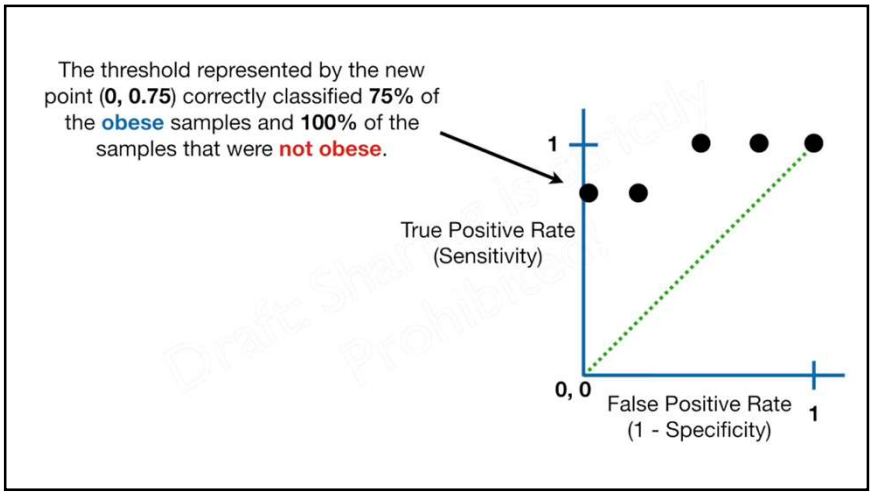

**True Positive Rate** = Sensitivity = $\frac{3}{3+1}$ = 0.75

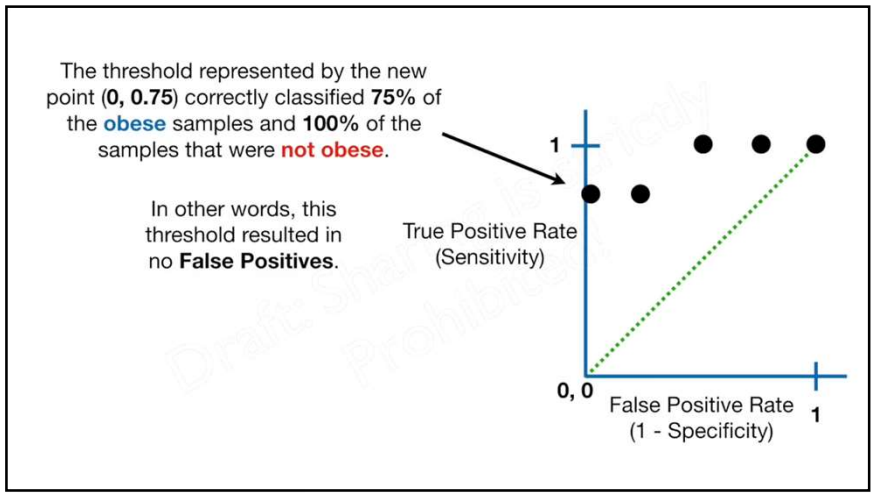
**False Positive Rate** = 1 - Specificity = $\frac{0}{0+4}$ = 0

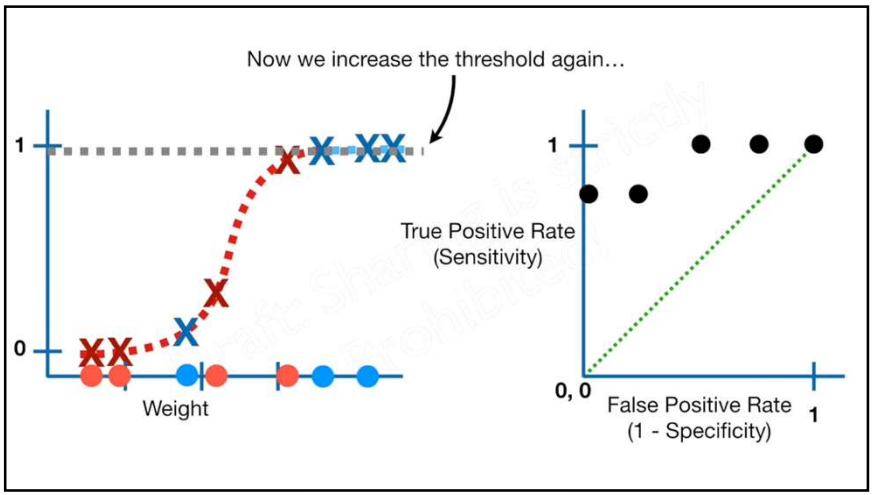
…calculate the **True Positive Rate** and the **False Positive Rate**…

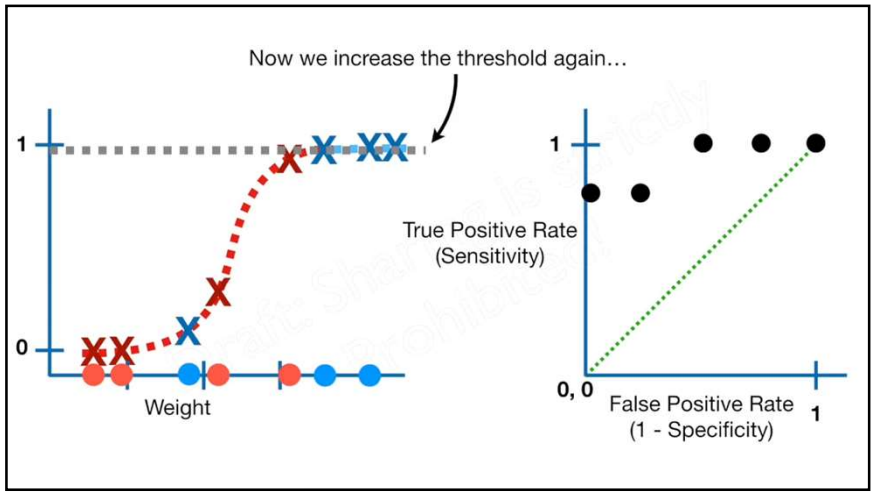
| | | Actual: Is Obese | Actual: Is Not Obese |
|---|---|---|---|
| Predicted | Is Obese | 3 | 0 |
| Predicted | Is Not Obese | 1 | 4 |

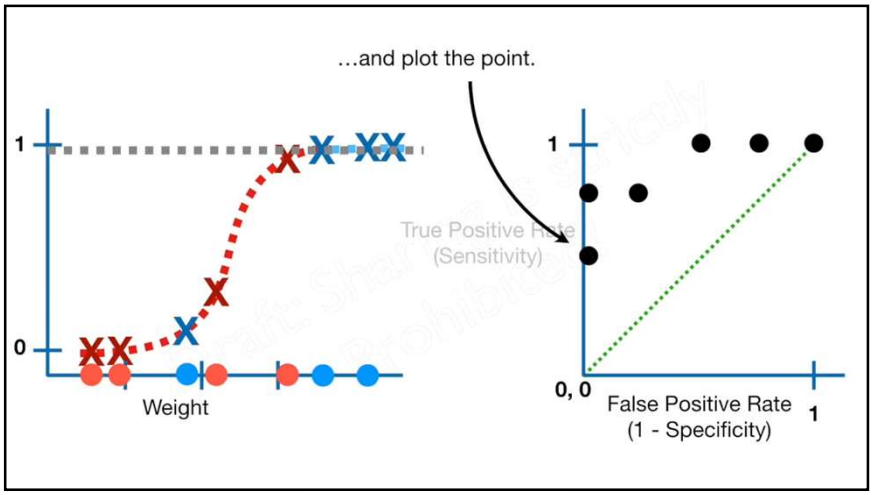

True Positive Rate = Sensitivity = 3 / 3 + 1 = 0.75
False Positive Rate = 1 - Specificity = 0 / 0 + 4 = 0
...and plot the point.
1
True Positive Rate
(Sensitivity)
0, 0
False Positive Rate
(1 - Specificity)
1

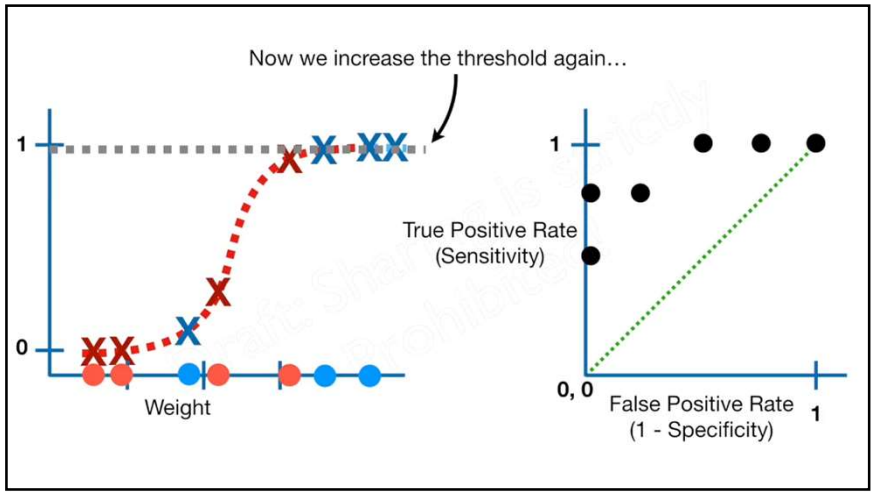

The threshold represented by the new point (0, 0.75) correctly classified 75% of the obese samples and 100% of the samples that were not obese.
1
True Positive Rate
(Sensitivity)
0, 0
False Positive Rate
(1 - Specificity)
1

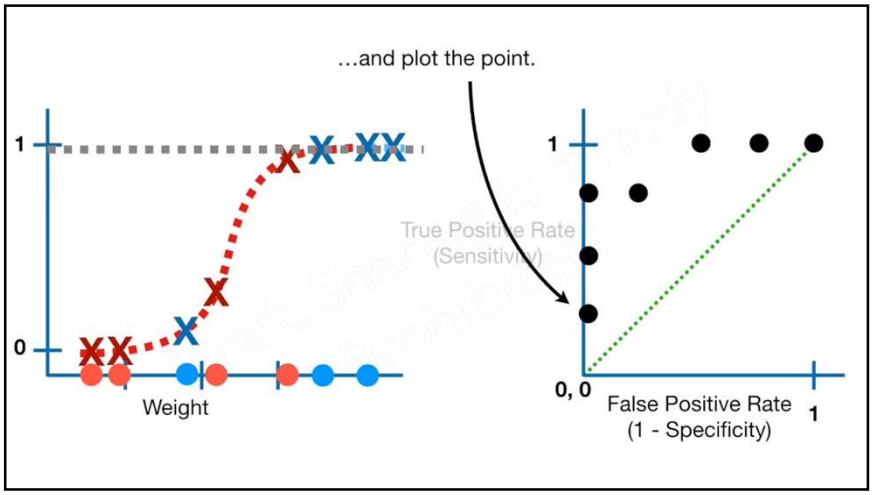

The threshold represented by the new point (0, 0.75) correctly classified 75% of the obese samples and 100% of the samples that were not obese.
In other words, this threshold resulted in no False Positives.
1
True Positive Rate
(Sensitivity)
0, 0
False Positive Rate
(1 - Specificity)
1

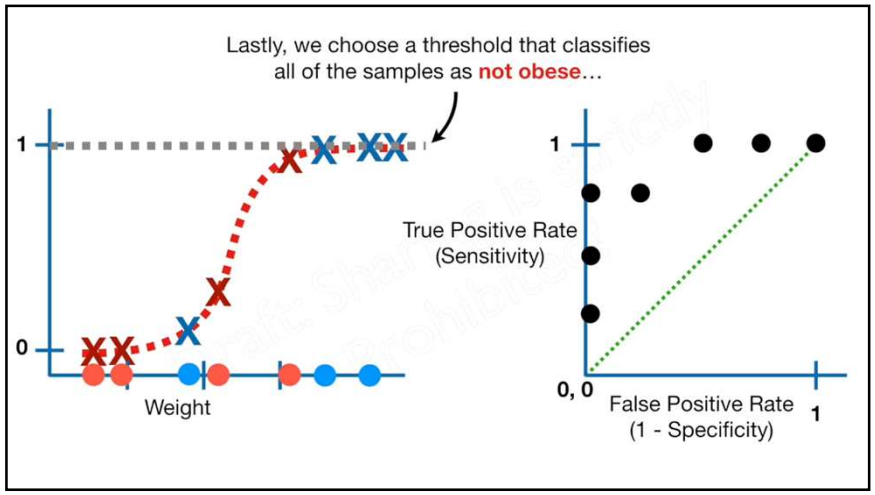

Now we increase the threshold again...
1
0
Weight
True Positive Rate
(Sensitivity)
0, 0
False Positive Rate
(1 - Specificity)
1

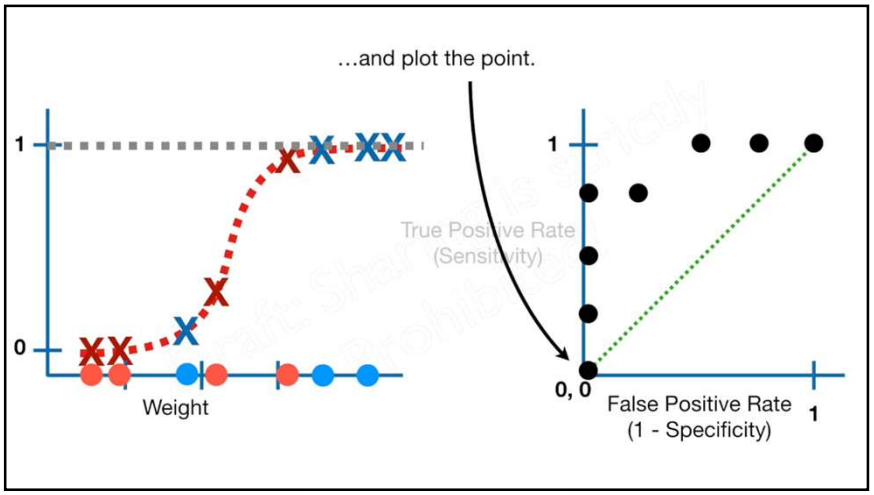

Now we increase the threshold again…
1
0
Weight
True Positive Rate
(Sensitivity)
1
0, 0
False Positive Rate
(1 - Specificity)
1

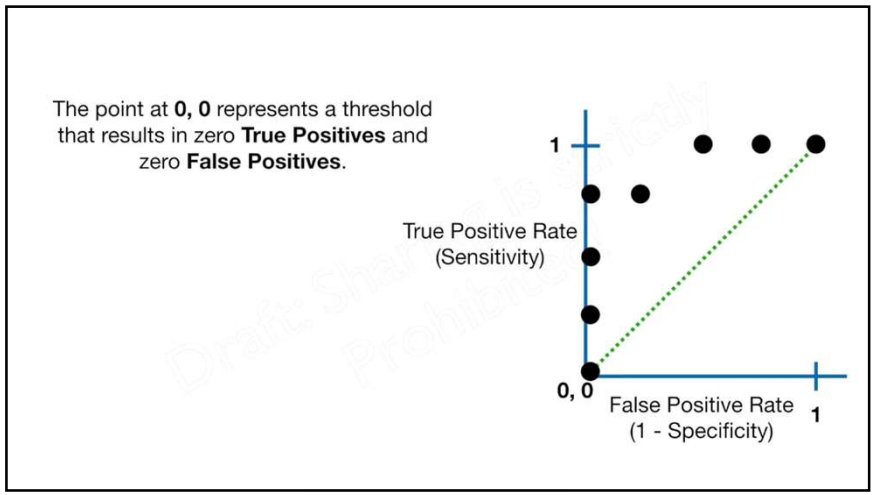

…and plot the point.
1
0
Weight
True Positive Rate
(Sensitivity)
1
0, 0
False Positive Rate
(1 - Specificity)
1

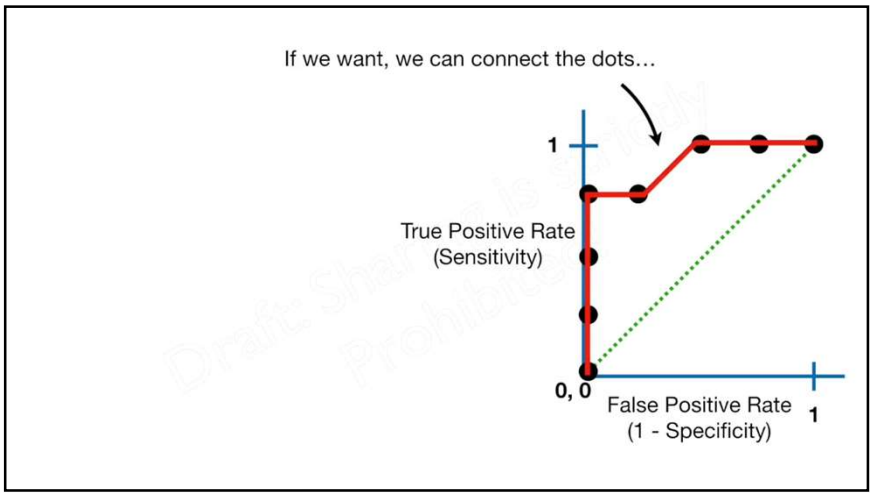

Now we increase the threshold again…
1
0
Weight
True Positive Rate
(Sensitivity)
1
0, 0
False Positive Rate
(1 - Specificity)
1


…and plot the point.
1
0
Weight
True Positive Rate
(Sensitivity)
1
0, 0
False Positive Rate
(1 - Specificity)
1

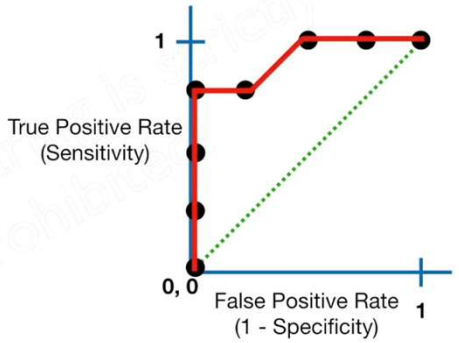

Lastly, we choose a threshold that classifies all of the samples as not obese…
1
0
Weight
True Positive Rate (Sensitivity)
1
0, 0
False Positive Rate (1 - Specificity)
1


…and plot the point.
1
0
Weight
True Positive Rate (Sensitivity)
1
0, 0
False Positive Rate (1 - Specificity)
1


The point at 0, 0 represents a threshold that results in zero True Positives and zero False Positives.
True Positive Rate (Sensitivity)
1
0, 0
False Positive Rate (1 - Specificity)
1

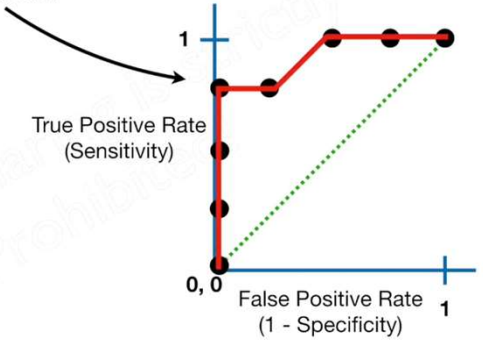

If we want, we can connect the dots…
True Positive Rate (Sensitivity)
1
0, 0
False Positive Rate (1 - Specificity)
1

…and that gives us an **ROC** graph.



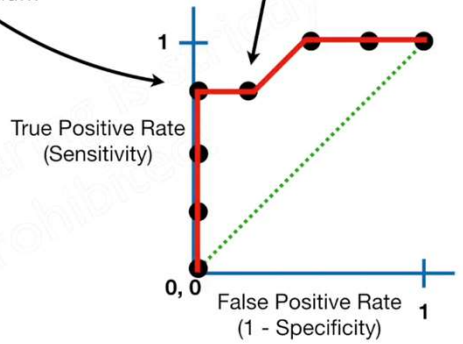

The **ROC** graph summarizes all of the confusion matrices that each threshold produced.

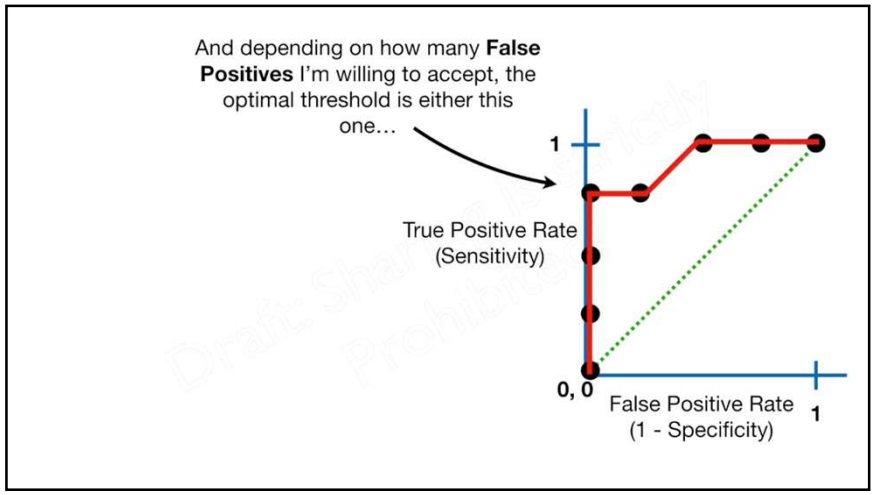

Without having to sort through the confusion matrices, I can tell that this threshold…

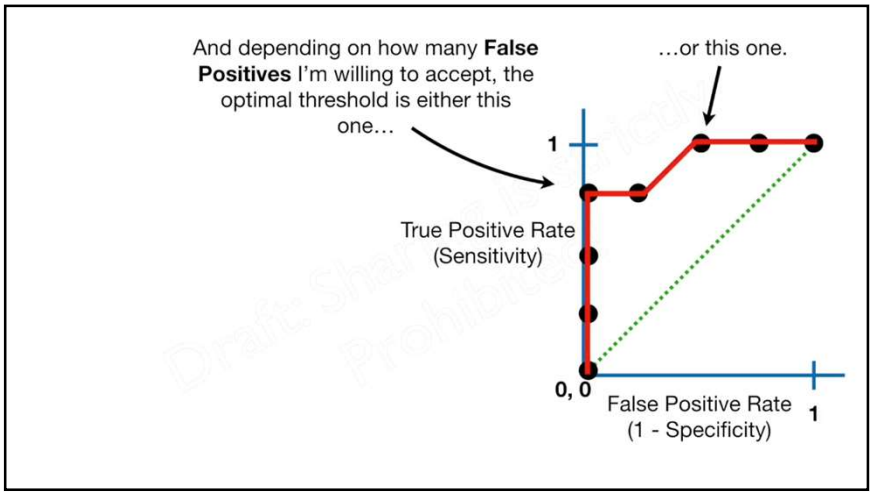


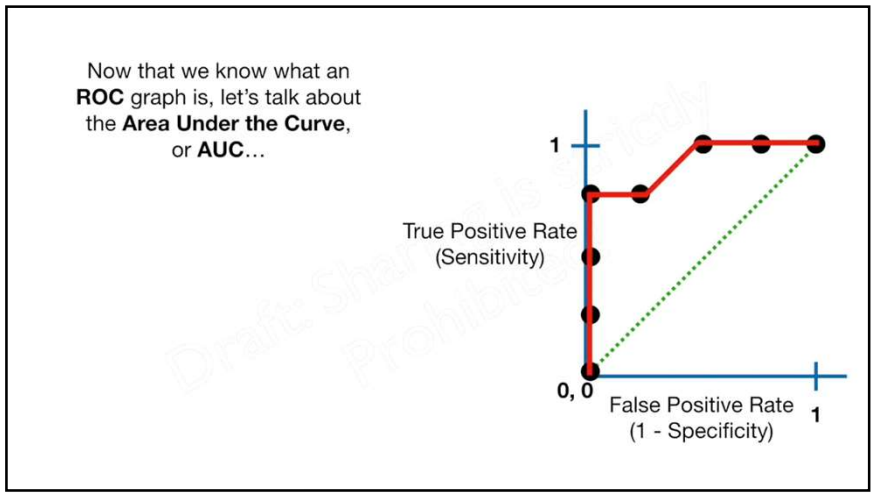

Without having to sort through the confusion matrices, I can tell that this threshold…

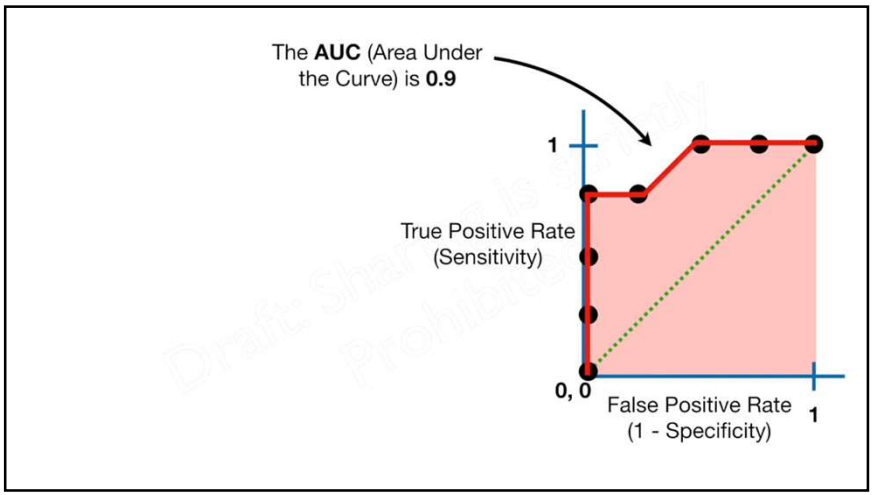
…is better than this threshold.



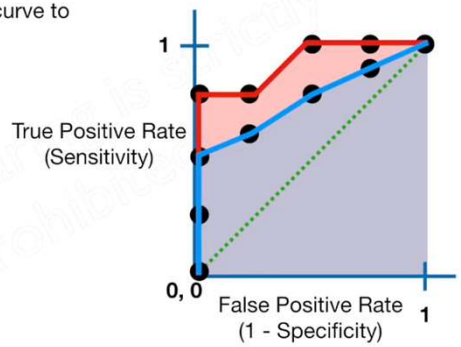

And depending on how many **False Positives** I'm willing to accept, the optimal threshold is either this one…
1
True Positive Rate (Sensitivity)
0, 0
False Positive Rate (1 - Specificity)
1


And depending on how many **False Positives** I'm willing to accept, the optimal threshold is either this one…
…or this one.
1
True Positive Rate (Sensitivity)
0, 0
False Positive Rate (1 - Specificity)
1

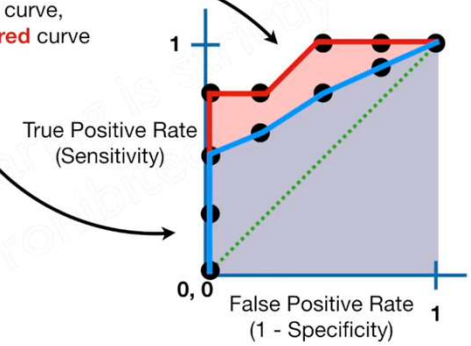

Now that we know what an **ROC** graph is, let's talk about the **Area Under the Curve**, or **AUC**…
1
True Positive Rate (Sensitivity)
0, 0
False Positive Rate (1 - Specificity)
1


The **AUC** (Area Under the Curve) is **0.9**
1
True Positive Rate (Sensitivity)
0, 0
False Positive Rate (1 - Specificity)
1

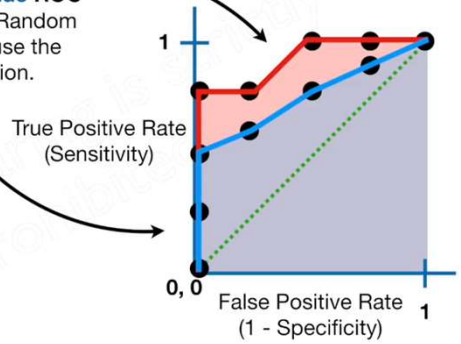
The **AUC** makes it easy to compare one **ROC** curve to another.



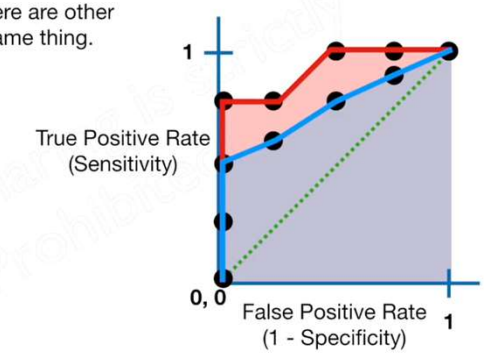

The **AUC** for the **red ROC** curve is greater than the **AUC** for the **blue ROC** curve, suggesting that the **red** curve is better.



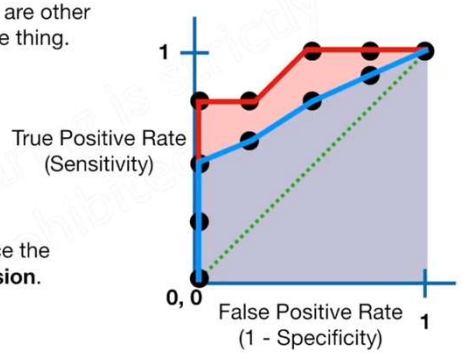

So if the **red ROC** curve represented Logistic Regression and the **blue ROC** curve represented a Random Forest, you would use the Logistic Regression.




Although **ROC** graphs are drawn using **True Positive Rates** and **False Positive Rates** to summarize confusion matrices, there are other metrics that attempt to do the same thing.



Although **ROC** graphs are drawn using **True Positive Rates** and **False Positive Rates** to summarize confusion matrices, there are other metrics that attempt to do the same thing.




For example, people often replace the **False Positive Rate** with **Precision**.

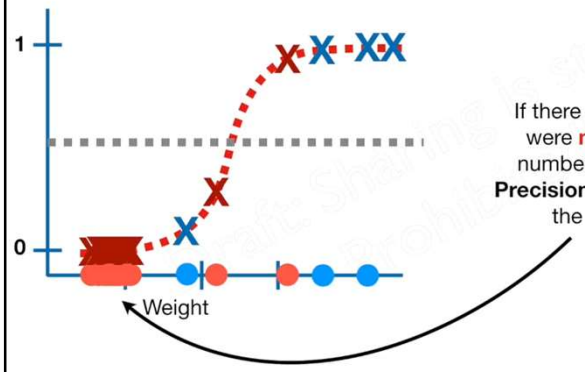
$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

| | | Actual: Is Obese | Actual: Is Not Obese |
|---|---|---|---|
| Predicted | Is Obese | True Positives | False Positives |
| | Is Not Obese | False Negatives | True Negatives |

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}} =$$

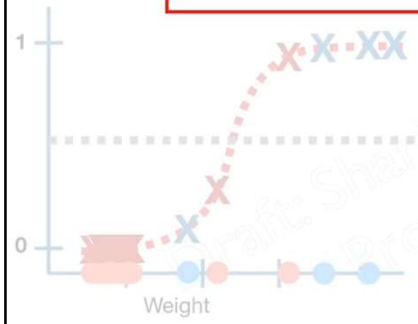
**Precision** is the proportion of positive results that were correctly classified

| | | Actual: Is Obese | Actual: Is Not Obese |
|---|---|---|---|
| Predicted | Is Obese | True Positives | False Positives |
| | Is Not Obese | False Negatives | True Negatives |




If there were lots of samples that were **not obese** relative to the number of **obese** samples, then **Precision** might be more useful than the **False Positive Rate**.

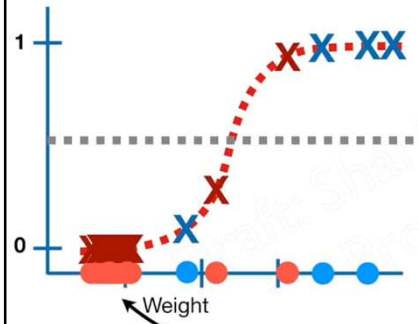
$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$




This is because **Precision** does not include the number of **True Negatives** in its calculation, and is not effected by the imbalance.



$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$



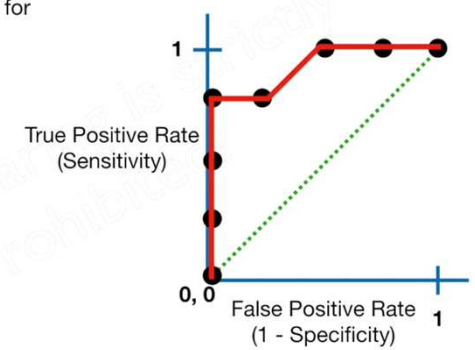

In practice, this sort of imbalance occurs when studying a rare disease. In this case, the study will contain many more people without the disease than with the disease.

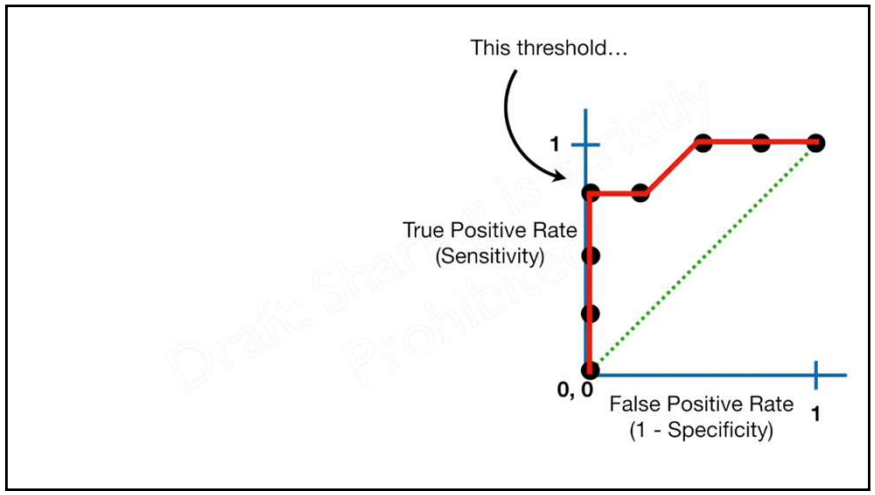


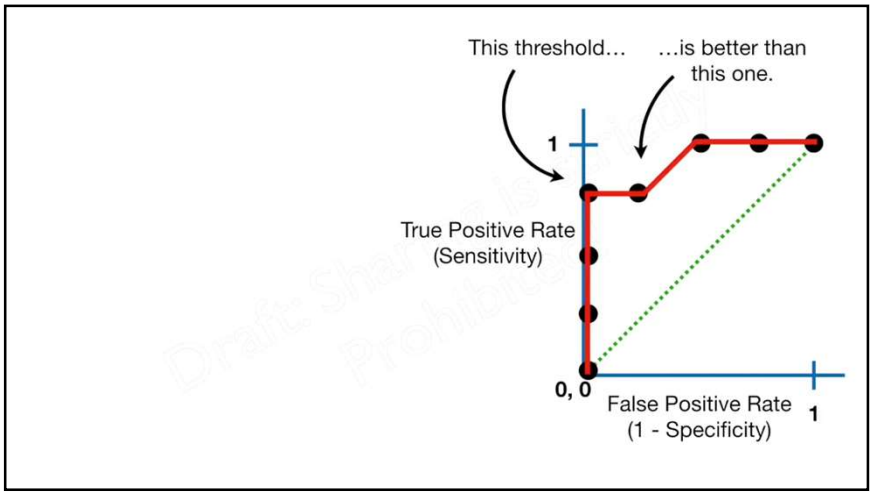
In Summary…



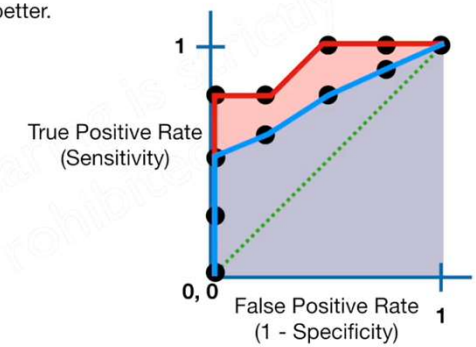
**ROC** curves make it easy to identify the best threshold for making a decision…

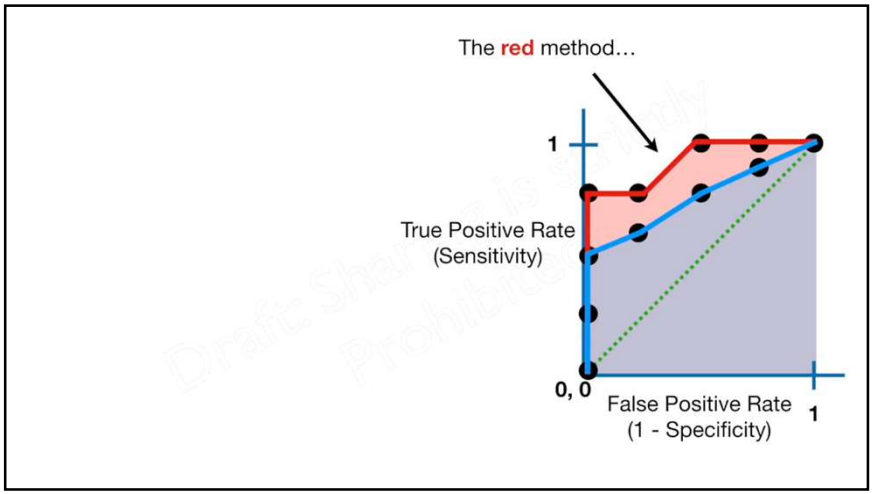





…and the **AUC** can help you decide which categorization method is better.



THANK YOU!