MWSUG 2016 - Paper RF13

An Animated Guide: Insights into the Logic of the ROC Curve

Russ Lavery, Bryn Mawr, PA

ABSTRACT

The ROC curve is often taught without any explanation. I've seen teachers simply say that the area under the ROC curve for a random process is .5 and a good model has an area greater than .5. This paper suggests that comparing histograms of "successes" and "failures", plotted on the same X axis, can provide insights into the logic, and use, of the ROC curve. Insights can be gleaned from the shape of the ROC Curve.

INTRODUCTION

I think, using ***words,*** and math, to explain ROC curves is confusing. This paper will use pictures and will talk, in an informal way, about ROC Curves. Here is a suggestion for learning the ideas in this paper. First; read the paper fairly quickly to get an exposure to the concepts and do a quick study of the figures. On a second read, having some exposure to all the figures, the words and calculations will make more sense.

ROC curves (Receiver Operator Characteristic) Curves came out of British RADAR research in WWII. They were a way to measure how accurately an operator could distinguish between a plane and a non-plane (or maybe types of planes) on a radar screen (a receiver). The useful part of understanding a ROC curve involves linking a short series (images of a histogram, a table and an ROC curve) into an "Ah-Ha!"

ROC Curves are used to determine how well a model separates two classes (call them "sick" and "healthy" and this paper will be predicting "**sick**"). This paper shows many histograms, and the histograms are the key to understanding ROC curves. Histograms with yellow boxes will represent counts of healthy subjects – one healthy subject to a box and there are twenty-five boxes. The histograms with orange boxes will represent counts of sick subjects – one sick person to a box and there are twenty-five boxes. Fifty observations were used to build the model and ROC Curve.

For presentation clarity, the histograms are separated vertically but both are plotted using the same X axis. The X axis can be anything that a reader might think could separate healthy people from sick people. The X-axis could be the result of a blood test that is being evaluated as a cheaper replacement for a more expensive test (remember, a modeler must have a different way of knowing if people are sick or healthy or SAS® could not make the ROC curve). The ROC curve is always plotting a correct classification versus a mistaken classification. The X axis could also be the probability of being sick; where the analyst used a multi-variable mathematical model to create the probability. Higher probabilities of being sick are to the right side of the X-Axis.

When a researcher ***builds*** a model, she know the outcomes (healthy-sick) for subjects the training data. If a she uses PROC Logistic, the logistic model assigns a probability of sick to every observation - to the healthy subjects as well as the sick subjects. The predicted probability of being sick is the X axis below the yellow and orange boxes in the figures. If a model "predicts well", it assigns the sick subjects a high probability of being sick (puts them to the right side of the X axis) and the healthy subjects a low probability of being sick (they would be plotted towards the left side of the X-axis).

This next idea is important and not often explained well when people talk about ROC Curves. When modeling, a researcher must make a managerial decision on where to set the cut-point that classifies people as being sick or healthy. The researcher must say "people to the left of *this* cut-point will be called healthy and people to the right will be called sick.". The ROC Curve summarizes the results of all possible cut-point decisions on one plot. The ROC curve presents the results of sliding the cut-point to the right or left and seeing how well the model, and the decision, combine to correctly classify people as sick or healthy. Each cut-point, on the X axis, generates one point on the ROC Curve.

As the analyst "slides" the cut-point from left to the right, she stops every time she encounters an observation. At that time, subjects to the left of the cut-point are classified as being healthy. Every time the cut-point encounters an observation and stops sliding, the analyst calculates "cumulative healthy people classed as healthy" and "cumulative sick people classified as healthy". Each stopping of the cut-point generates one point on the ROC curve.

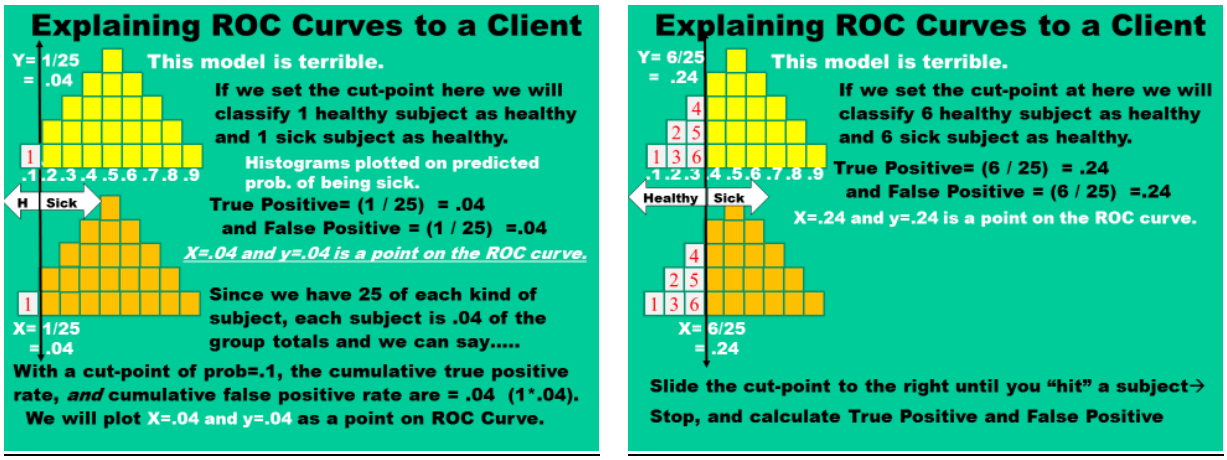


Figure 1

Therefore, just ROC curve summarizes information about all possible cut-points – and, after a bit of thinking, can be used to evaluate a model.

When the probability curves (histograms) for healthy and sick overlap a researcher will always make some mistakes in classification – no matter where the cut-point is put. Because of the overlap, no matter where a researcher "sets" the cut-point, she will call some people healthy, who are in fact sick, and some sick, who are in fact healthy. Where a researcher places the cut-point affects the percentage of, and types of (type I or type II), errors she makes.

The axes on an ROC curve are: y= cumulative percent of true positive classifications for a cut-point and x=Cumulative percent of false positive classifications ***for a cut-point***.

THE ROC FOR A MODEL THAT CAN NOT OUTPERFORM CHANCE

Figure 2 provides the first example of a model and ROC curve.

This model performs very poorly. We say that because the model does not separate the histograms when they are plotted on the same X axis.

The X axis can be thought of as either the predicted probability of being sick, or some x variable (like a blood test) that a researcher hopes will separate the groups.

Maybe this researcher tried to model sick-healthy as a function of a really foolish variable.

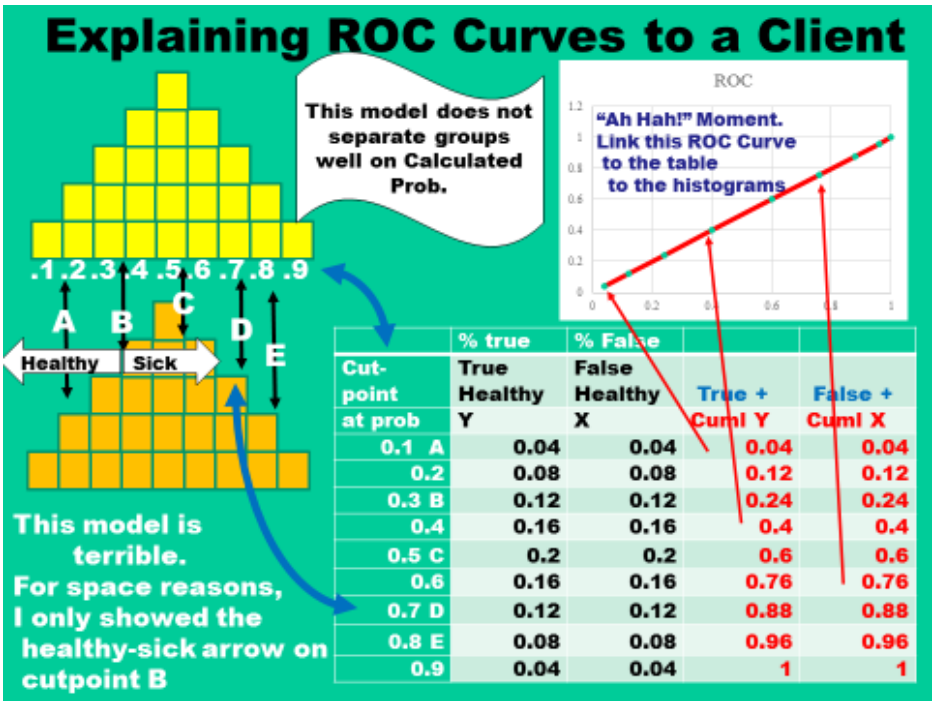
Using the last two digits of a zip-code, as an X variable, would be a poorly predicting variable.

This (nonsensical - zip-code based) model assigns "probabilities of being sick" to both the sick and the healthy – all models do this. Figure 2 displays the two distributions arranged one above one another so they both share the same X axis. The X axis shows the probability of being sick as assigned by the model.

Since there are 25 subject in each group, counting boxes simplifies math. Each box is .04 of the total sick or healthy subjects.

Imagine sliding the "healthy-sick cut-point" from left to right and stopping each time the cut-point encounters an observation (sick or healthy).

When the cut-point stops at A (prob=.1) the researcher will classify 1 healthy subject as healthy and 1 sick subject as healthy. Since there are 25 of each kind of subject, a single subject will be .04 of the group totals.



	% true	% False		
Cut-point at prob	True Healthy Y	False Healthy X	True + Cuml Y	False + Cuml X
0.1 A	0.04	0.04	0.04	0.04
0.2	0.08	0.08	0.12	0.12
0.3 B	0.12	0.12	0.24	0.24
0.4	0.16	0.16	0.4	0.4
0.5 C	0.2	0.2	0.6	0.6
0.6	0.16	0.16	0.76	0.76
0.7 D	0.12	0.12	0.88	0.88
0.8 E	0.08	0.08	0.96	0.96
0.9	0.04	0.04	1	1

Figure 2

If the researcher sets the cut-point at probability of being sick=.2, the researcher will, in total, classify 3 healthy, and 3 sick, people as being healthy. The cumulative true positive rate will be .12 and the cumulative false positive rate will be .12. This point can be plotted on the ROC Curve.

If the researcher sets the cut-point at probability of being sick=.3 (The B arrow), the researcher will, in total, classify 6 healthy, and 6 sick, people as being healthy. The cumulative true positive rate will be .24 and the cumulative false positive rate will be .24. This point can be plotted on the ROC Curve.

A researcher can calculate as above, for all the other possible cut-points in the histogram and produce the table and curve in Figure 2. That curve has summarized all the data points, and cut-points, into one plot. Note that the plot is at a 45 degree angle and ***that is a characteristic of a model that predicts terribly***. Please note that the histograms do not "separate" at all they are right above each other.

This model is terrible and predicts only as well as a fair coin. Here is the "Ah-Hah" moment of the paper. Please make a mental link between the histograms and the ROC curve shown in Figure 1. The histograms in figure 2 do not separate at all and the ROC curve is at 45 degrees. An ROC curve on a 45 degree indicates a worthless model. The key to understanding how ROC curves can be used to judge model worth is to "link" the picture of the distributions – through the table of calculations - to the ROC Plot - to the area under the blue curve (Called AUC or **A**rea **U**nder the **C**urve). A terrible model has an AUC of .5 because it predicts with the power of "chance" and the histograms "do not separate".

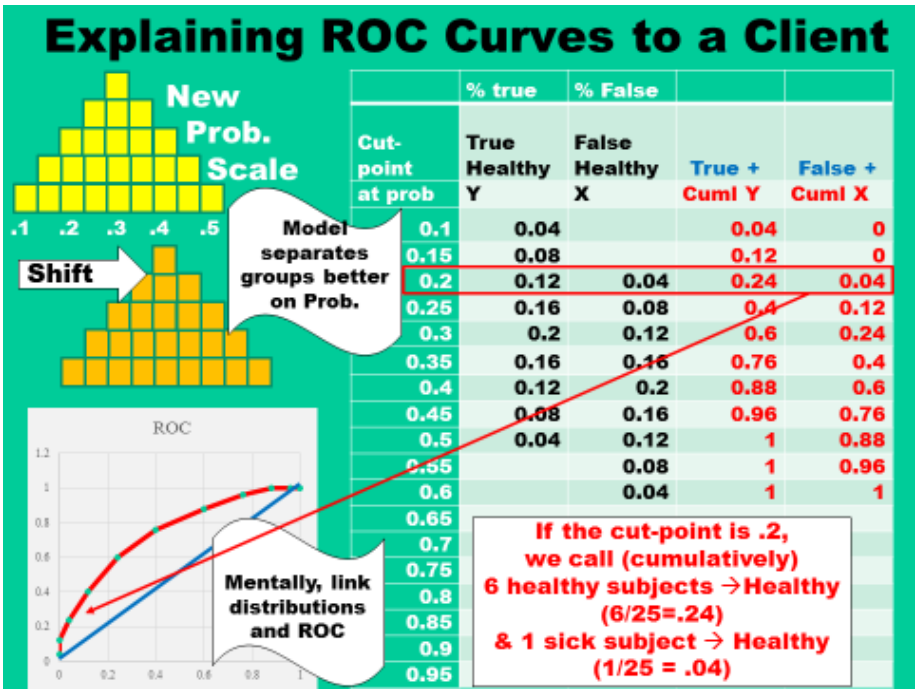
TWO SLIGHTLY BETTER MODELS

Figure 3 uses a better x variable (maybe a better blood test) or a better mathematical model.

It predicts better than the model in figure 1.

We say that it predicts better because the distribution of sick people is ***shifted to the right.***

Because the model assigns the known-sick subjects a higher probability of being sick, it causes the histograms to start to separate.



Cut-point at prob	% true True Healthy Y	% False False Healthy X	True + Cuml Y	False + Cuml X
0.1	0.04		0.04	0
0.15	0.08		0.12	0
0.2	0.12	0.04	0.24	0.04
0.25	0.16	0.08	0.4	0.12
0.3	0.2	0.12	0.6	0.24
0.35	0.16	0.16	0.76	0.4
0.4	0.12	0.2	0.88	0.6
0.45	0.08	0.16	0.96	0.76
0.5	0.04	0.12	1	0.88
0.55		0.08	1	0.96
0.6		0.04	1	1
0.65				
0.7				
0.75				
0.8				
0.85				
0.9				
0.95				

Figure 3

The table in figure 3 calculates the "cumulative true positive percentage" and "cumulative false positive percentage" at various cut-points in figure 3. Notice that the X-axis, the probability of being sick is not the same scale that was used in figure 2. The scale was changed to make things fit on the slide better.

Remember, since there are 25 subjects in each group, each subject is .04 of the total number in a group.

If the researcher sets the cut-point at .2 she will, in total, classify 6 (=.24/.04) people as healthy and 1 sick person as healthy. The cumulative true positive rate will be .24 and the cumulative false positive rate will be .04. The coordinates (x=.04 , y=.24) is a point on the ROC curve. If she sets the cut-point at .3 she will, in total, classify 15 (=.60/.04) subjects as healthy and 6 sick subjects as healthy. The cumulative true positive rate will be .60 and the cumulative false positive rate will be .24. This point can be plotted and note that, on the ROC Curve. When X is .24, Y is .60,

Calculations proceed in a similar manner for all the other cut-points and will produce the curve in Figure 3. That curve summarizes all the data, and cut-points, into one plot. Note this plot (blue line) is above the 45 degree line. Having an ROC above the 45 degree line is a characteristic of a model that predicts better than a fair coin. This model has "OK" predictive power. Please make a mental link between the histogram, the table of calculations and the ROC curve in figure3 – "shifted" histograms result in an ROC curve above a 45 degree line.

As a model predicts better and better, histograms will "shift" more and the ROC curve will move farther and farther away from a 45 degree line. The paper will present an even better model in the next example.

Please note the small vertical rise at the left hand side of the red line. The paper will discuss that characteristic more in future examples, where it is easier to see.

The model in figure 4 predicts better than the model in figure 3.

It must come from an even better X variable or model.

The histogram of sick people is shifted more to the right, than in previous figures, because the model predicts better.

The model separates the histograms more than previous models.

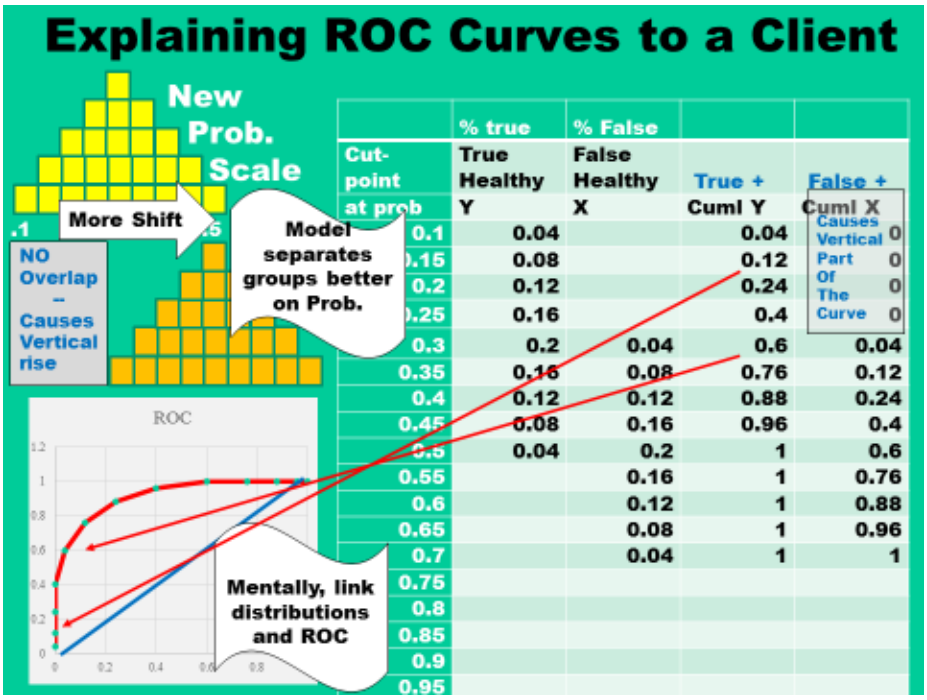


Figure 4

This model assigns sick subjects a higher probability of being sick than did the model in figures 2 or 3. The paper will now calculate the true positive rate and false positive rate at various cut-points. Notice that the X-axis, the probability of being sick is not the same as in previous figures.

If a researcher sets the cut-point at .2, she will, classify 6 healthy and 0 sick people as being healthy. The true positive rate will be .24 and the false positive rate will be .0 and this can be plotted.

If the researcher sets the cut-point at .3, she will, classify 15 healthy and 1 sick person as being healthy. The true positive rate will be .6 and the false positive rate will be .04 and this can be plotted.

If the researcher sets the cut-point at .5 she will, classify 25 healthy and 15 sick people as being healthy. The cumulative true positive percentage will be 1.0 and the cumulative false positive percentage will be .60. At this cut-point, all healthy people have been classified as healthy and the curve turns horizontal

The curve starts out as a vertical line because, until cut-point=prob=.3 there are no false positives. At cut-point=prob=.5 there are no more healthy people to be "classified", so the curve flattens out.

Here is an insight into the ROC curve. A vertical part of the ROC curve, or a horizontal part of the ROC curve indicates that the distributions are not overlapping. Please see the shaded boxes in Figure 4 and Figure 5.

We can proceed in a similar manner for all the other cut-points and we get the curve in Figure 4. That curve has summarized all the possible cut-points into one plot. Note, this line-plot is above the 45 degree line and being above the 45 degree line a characteristic of a model that predicts better than a fair coin. This is a pretty good model and please make a mental link between the histograms, the table of calculations and this curve.

We now have the logic for the rule that most researchers use. We want a model that has an ROC curve above the 45 degree line and the greater the area under the ROC curve the better the model. Let's look at three more curves and see if we can learn more from the shape.

AN EXCELLENT MODEL

In Figure 5, the model is a very good model. This is the ROC plot all modelers ***hope to see***. It indicates a predictive model with almost no misclassifications.

Please note how the curve starts out as vertical because, until cut-point=.5 (prob=.5), it does not have any false positives (note gray boxes).

A perfect model will go straight up and then take a 90 degree turn.

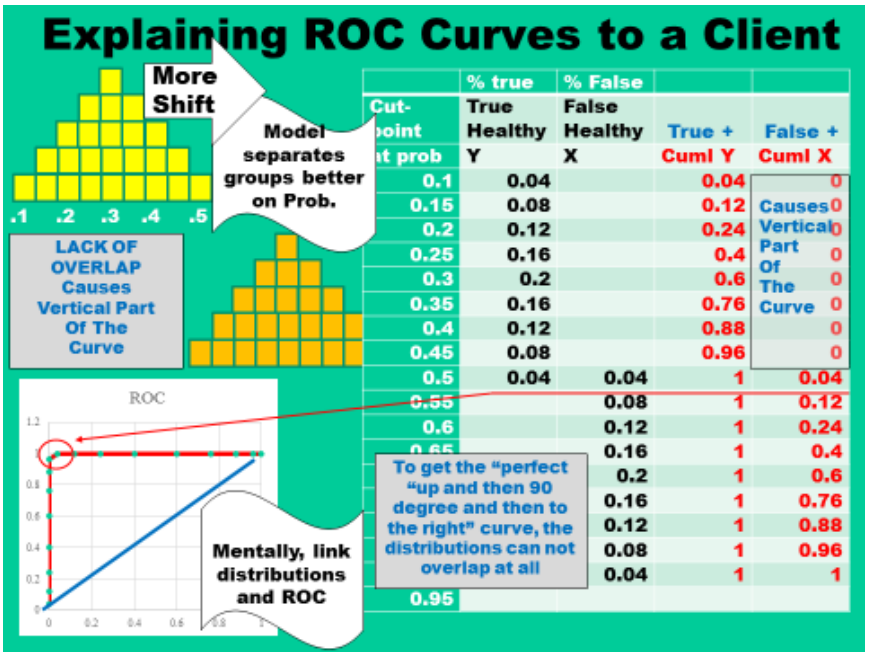


Figure 5

At cut-point .5, we have one true positive and one false positive and the ROC curve, for a short time, has a slope of 45 degrees. Above cut-point =.5 there are no more healthy subjects, so the curve flattens out.

Previous histograms have been smooth. The paper now considers situations where the probability curves are not smooth and mound shaped. In the next figure there are no healthy subjects with prob=.35.

INSIGHTS INTO SHAPES OF THE ROC CURVE

The "kink" we see in this ROC curve, is caused by a data collection issue.

N***o healthy subjects had the combination of X values that would make the model predict .35.***

This "data quirk" makes the ROC curve "jut in" (arrows).

The ROC curve gets closer to the 45 degree line and being close to the 45 degree line usually means that we do not predict very well at that cut-point. However; this "getting closer" is caused by a "data quirk in healthy subjects.

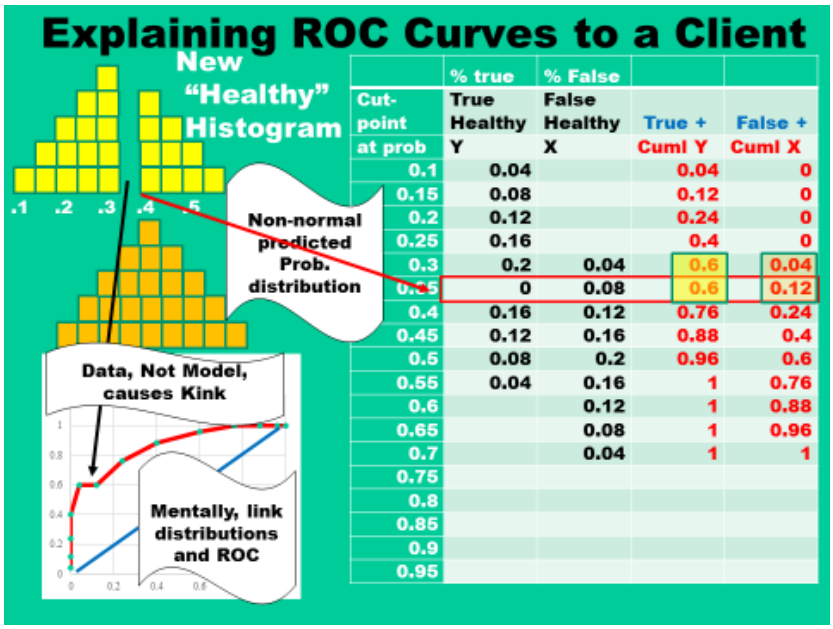


Figure 6

ROC curves can have multiple places where the curve "juts in" towards the 45 degree line. These kinks in the curve are often caused by data collection issues - by lack of data (no observation with the proper combination of X values that would cause the model to predict certain probabilities at that cut-point). **Kinks are problems in how the data was collected and not in the ability of the model to perform.**

Figure 7, shows very flat histograms.

IN THIS EXAMPLE, THE NUMBER OF SUBJECTS HAS CHANGED from 25 TO 10 AND EACH SUBJECT IS .1 OF THE TOTAL.

Modelers want ROC curves above the 45 and this example points out a complication. The AUC in Figure 7 is greater than .5 but the model only really predicts well for low probabilities. After probability = .40 the model only predicts as well a chance.

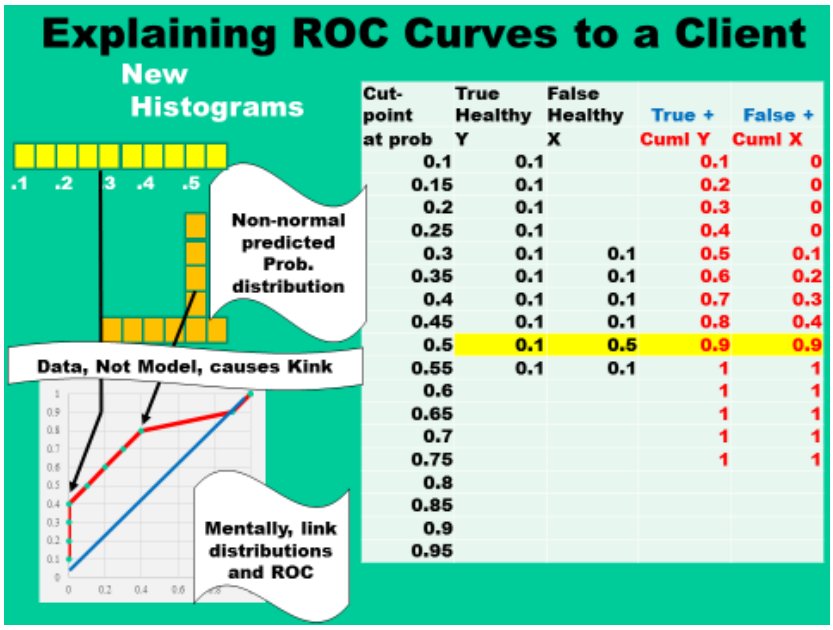


Figure 7

This model predicts the "first" 40 percent of the healthy subjects perfectly and then is a failure. The failure is not gradual. The model, suddenly, goes from predicting perfectly to predicting like a "fair coin".

Figure 8 explores ROC curves where the slope of the curve is 45 degrees.

These ROC curves show vertical and horizontal sections caused by lack of overlap of between healthy and sick.

Note: histograms with very different shapes can all generate a 45 degree line.

A 45 degree line is caused by the curves having the same shape and height.

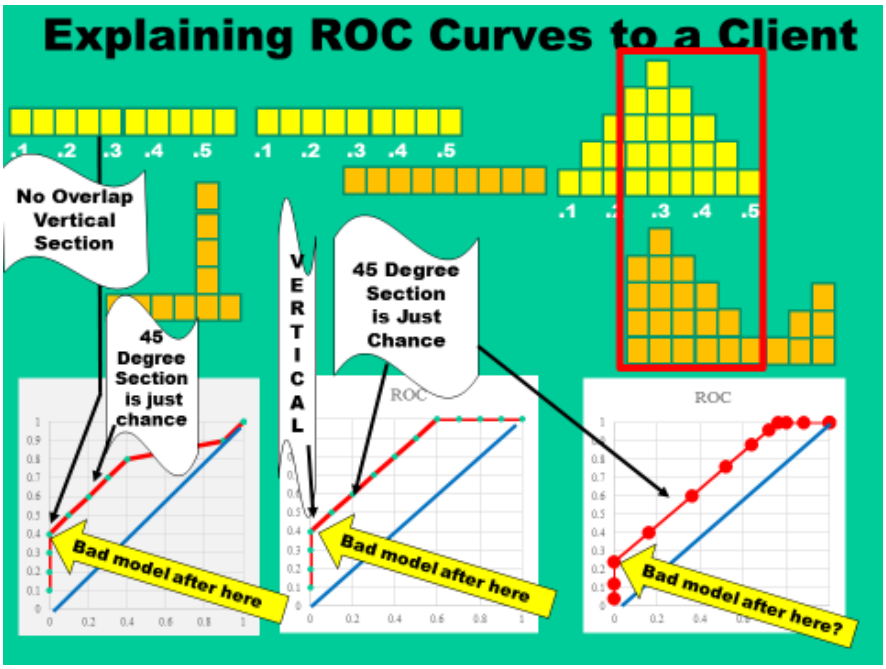


Figure 8

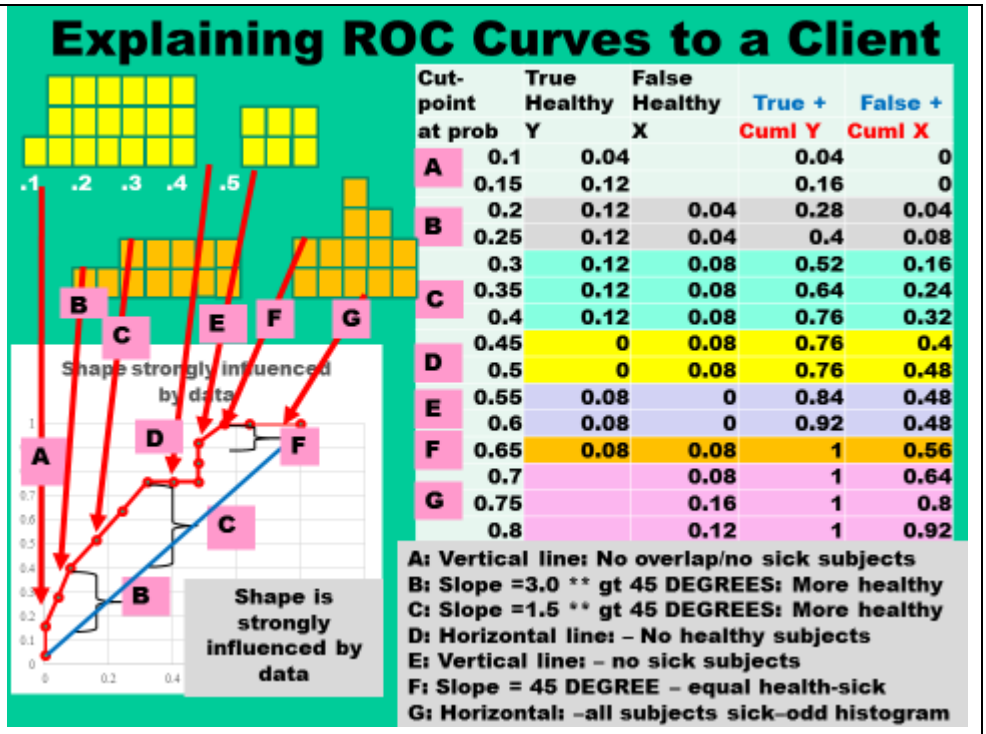
The red box emphasizes that the way the data was collected (the distribution of the X variables for the subjects) caused the both distributions to have the same shape and produce a 45 degree ROC curve.

SUMMARY

This final example is a summary.

It intends, on a very cluttered slide, to show all the points developed above.

Each effect that a reader should understand has been assigned a letter and explanations are on the slide.



	Cut-point at prob	True Healthy Y	False Healthy X	True + Cuml Y	False + Cuml X
A	0.1	0.04		0.04	0
	0.15	0.12		0.16	0
B	0.2	0.12	0.04	0.28	0.04
	0.25	0.12	0.04	0.4	0.08
	0.3	0.12	0.08	0.52	0.16
C	0.35	0.12	0.08	0.64	0.24
	0.4	0.12	0.08	0.76	0.32
	0.45	0	0.08	0.76	0.4
D	0.5	0	0.08	0.76	0.48
E	0.55	0.08	0	0.84	0.48
	0.6	0.08	0	0.92	0.48
F	0.65	0.08	0.08	1	0.56
	0.7		0.08	1	0.64
G	0.75		0.16	1	0.8
	0.8		0.12	1	0.92

Figure 9

ACKNOWLEDGMENTS

Thanks to Peter Flom for listening to early thoughts.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Russ Lavery Russ.lavery@verizon.net