Picking a good threshold value is often challenging.

A Receiver Operator Characteristic curve, or ROC curve, can help you decide which value of the threshold is best.

The ROC curve for our problem is shown on the right of this slide.

The sensitivity, or true positive rate of the model, is shown on the y-axis.

And the false positive rate, or 1 minus the specificity, is given on the x-axis.

The line shows how these two outcome measures vary with different threshold values.

The ROC curve always starts at the point (0, 0).

This corresponds to a threshold value of 1.

If you have a threshold of 1, you will not catch any poor care cases, or have a sensitivity of 0.

But you will correctly label of all the good care cases, meaning you have a false positive rate of 0.

The ROC curve always ends at the point (1,1), which corresponds to a threshold value of 0.

If you have a threshold of 0, you'll catch all of the poor care cases, or have a sensitivity of 1, but you'll label all of the good care cases as poor care cases too, meaning you have a false positive rate of 1.

The threshold decreases as you move from (0,0) to (1,1).

At the point (0, 0.4), or about here, you're correctly labeling about 40% of the poor care cases with a very small false positive rate.

On the other hand, at the point (0.6, 0.9), you're correctly labeling about 90% of the poor care cases, but have a false positive rate of 60%.

In the middle, around (0.3, 0.8), you're correctly labeling about 80% of the poor care cases, with a 30% false positive rate.

The ROC curve captures all thresholds simultaneously.

The higher the threshold, or closer to (0, 0), the higher the specificity and the lower the sensitivity.

The lower the threshold, or closer to (1,1), the higher the sensitivity and lower the specificity.

So which threshold value should you pick?

You should select the best threshold for the trade-off you want to make.

If you're more concerned with having a high specificity or low false positive rate, pick the threshold that maximizes the true positive rate while keeping the false positive rate really low.

A threshold around (0.1, 0.5) on this ROC curve looks like a good choice in this case.

On the other hand, if you're more concerned with having a high sensitivity or high true positive rate, pick a threshold that minimizes the false positive rate but has a very high true positive rate.

A threshold around (0.3, 0.8) looks like a good choice in this case.

You can label the threshold values in R by color-coding the curve.

The legend is shown on the right.

This shows us that-- say we want to pick a threshold value around here.

This corresponds to between the aqua color and the green color.

Or it looks like about a threshold of 0.3.

Instead, if we wanted to pick a threshold around here, this looks like the start of the darker blue color, and looks like it's probably a threshold around 0.2.

We can also add specific threshold labels to the curve in R. This helps you see which threshold value you want to use.

Let's go into R and see how to generate these ROC curves.

To generate ROC curves in R, we need to install a new package.

We'll use the same two commands as we did earlier in the lecture, but this time the name of the package is ROCR.

So first type install.packages("ROCR"), and hit Enter.

Since we picked a CRAN mirror already in this R session, we shouldn't have to pick it again.

You know the package is done installing when you see the arrow and blinking cursor in your R console.

Now let's load the package using the library function.

Recall that we made predictions on our training set and called them predictTrain.

We'll use these predictions to create our ROC curve.

First, we'll call the prediction function of ROCR.

We'll call the output of this function ROCRpred, and then use the prediction function.

This function takes two arguments.

The first is the predictions we made with our model, which we called predictTrain.

The second argument is the true outcomes of our data points, which in our case, is qualityTrain$PoorCare.

Now, we need to use the performance function.

This defines what we'd like to plot on the x and y-axes of our ROC curve.

We'll call the output of this ROCRperf, and use the performance function, which takes as arguments the output of the prediction function, and then what we want on the x and y-axes.

In this case, it's true positive rate, or "tpr", and false positive rate, or "fpr".

Now, we just need to plot the output of the performance function, ROCRperf.

You should see the ROC curve pop up in a new window.

This should look exactly like the one we saw on the slides.

Now, you can add colors by adding one additional argument to the plot function.

So in your R console, hit the up arrow, and after ROCRperf, type colorize=TRUE, and hit Enter.

If you go back to your plot window, you should see the ROC curve with the colors for the threshold values added.

Now finally, let's add the threshold labels to our plot.

Back in your R console, hit the up arrow again to get the plot function, and after colorize=TRUE, we'll add two more arguments.

The first is print.cutoffs.at=seq(0,1,0.1), which will print the threshold values in increments of 0.1.

If you want finer increments, just decrease the value of 0.1.

And then the final argument is text.adj=c(-0.2,1.7), and hit enter.

If you go back to your plot window, you should see the ROC curve with threshold values added.

Using this curve, we can determine which threshold value we want to use depending on our preferences as a decision-maker.

In the next video, we'll discuss how to assess the strength of our model.