Now let's look at the ROC curve so we can understand the performance of our model at different cutoffs.

We'll first need to load the ROCR package with a library(ROCR).

Next, we'll build our ROCR prediction object.

So we'll call this object predROCR = prediction(pred.prob, test$responsive).

All right.

So now we want to plot the ROC curve so we'll use the performance function to extract the true positive rate and false positive rate.

So create something called perfROCR = performance(predROCR, "tpr", "fpr").

And then we'll plot(perfROCR, colorize=TRUE), so that we can see the colors for the different cutoff thresholds.

All right.

Now, of course, the best cutoff to select is entirely dependent on the costs assigned by the decision maker to false positives and true positives.

However, again, we do favor cutoffs that give us a high sensitivity.

We want to identify a large number of the responsive documents.

So something that might look promising might be a point right around here, in this part of the curve, where we have a true positive rate of around 70%, meaning that we're getting about 70% of all the responsive documents, and a false positive rate of about 20%, meaning that we're making mistakes and accidentally identifying as responsive 20% of the non-responsive documents.

Now, since, typically, the vast majority of documents are non-responsive, operating at this cutoff would result, perhaps, in a large decrease in the amount of manual effort needed in the e-discovery process.

And we can see from the blue color of the plot at this particular location that we're looking at a threshold around maybe 0.15 or so, significantly lower than 50%, which is definitely what we would expect since we favor false positives to false negatives.

So lastly, we can use the ROCR package to compute our auc value.

So, again, call the performance function with our prediction object, this time extracting the auc value and just

grabbing the y value slot of it.

We can see that we have an auc in the test set of 79.4%, which means that our model can differentiate between a randomly selected responsive and non-responsive document about 80% of the time.