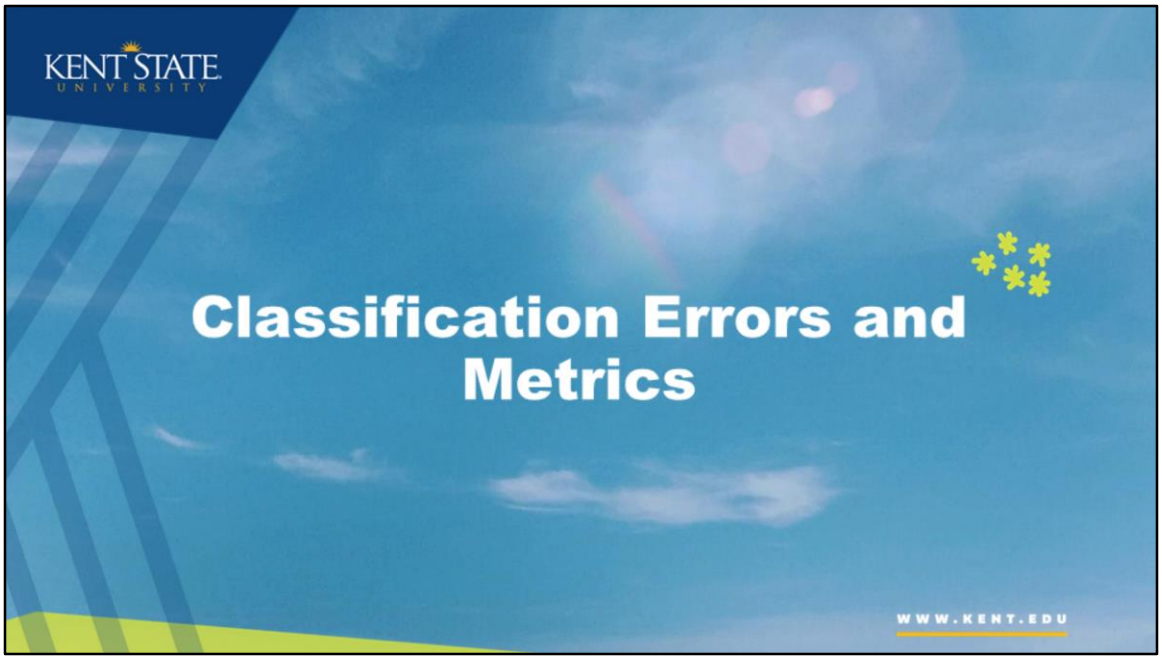
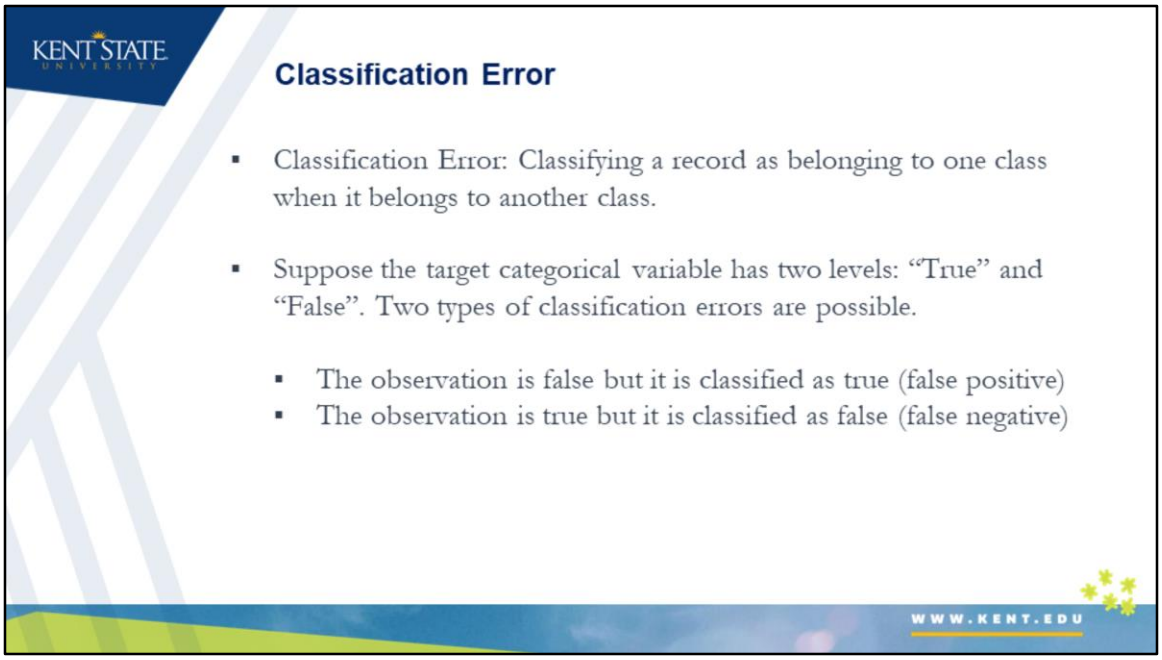


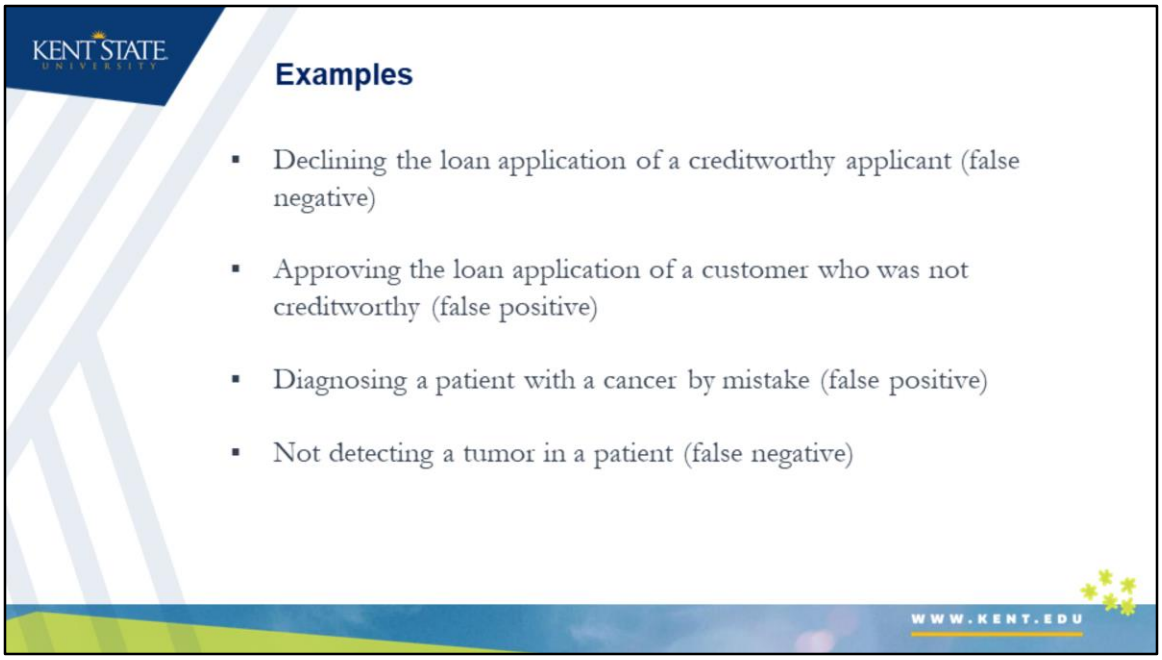

In this lecture, we concentrate on classification model outputs. Specifically, in determining how to define classification performance.




Simply, a classification error is when we classify an observation into a group other than the one it truly belongs. For a two group classification problem, there are four results that can be happen for a particular observation.

1. The observation is correctly grouped
    1. If the observed value is positive, and we correctly predict it as positive, we say that it is True Positive
    2. If the observed value is negative, and we correctly predict it as negative, we that that it is True Negative
2. The observation is incorrectly identified
    1. If the observed value is positive, but we predict it as negative, then we say it is False Negative
    2. If the observed value is negative, but we predict it as positive, then we say it is False Positive

What would you use as an objective to determine classifier performance?




Here are some examples of false positive and false negative. Note that the designation of positive and negative are completely arbitrary and depends on the problem context. As a modeler, you can define which outcome is positive, and which is negative. But, as we will discover later, there are hidden assumptions to the classification results presentation, and you should be aware of them when designating positive and negative. For the moment, the positive and negative labels on our examples should be self explanatory.

KENT STATE UNIVERSITY

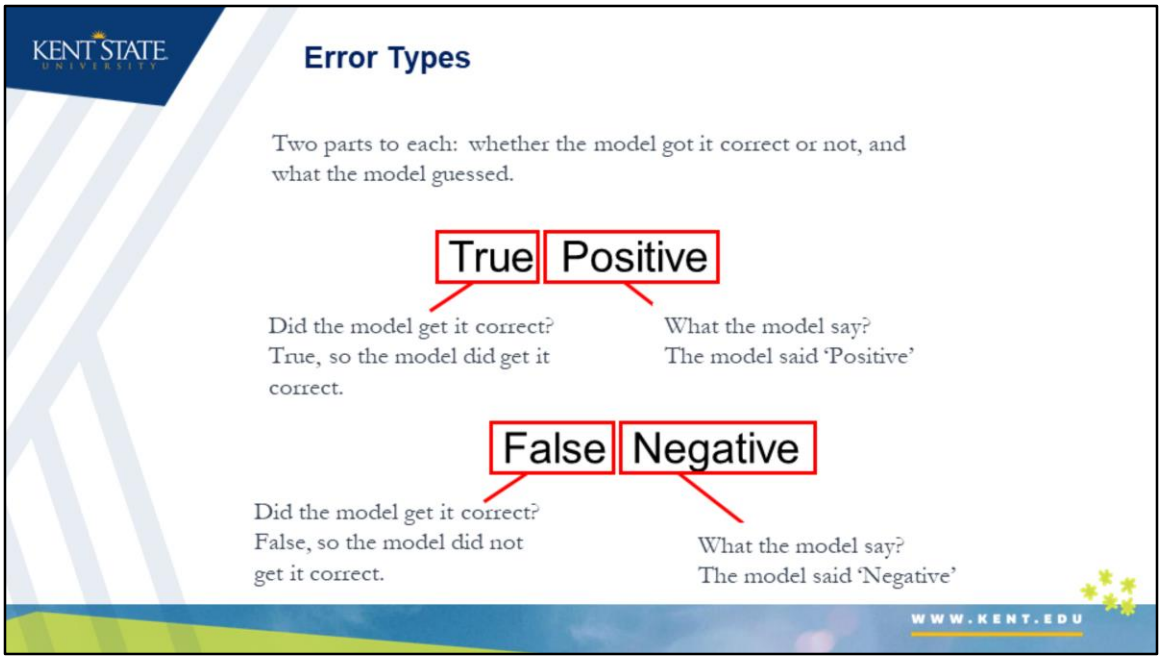
## Confusion Matrix

- **TN** is the number of correct predictions that an instance is negative (True Negative),
- **FP** is the number of incorrect predictions that an instance is positive,
- **FN** is the number of incorrect of predictions that an instance negative, and
- **TP** is the number of correct predictions that an instance is positive.

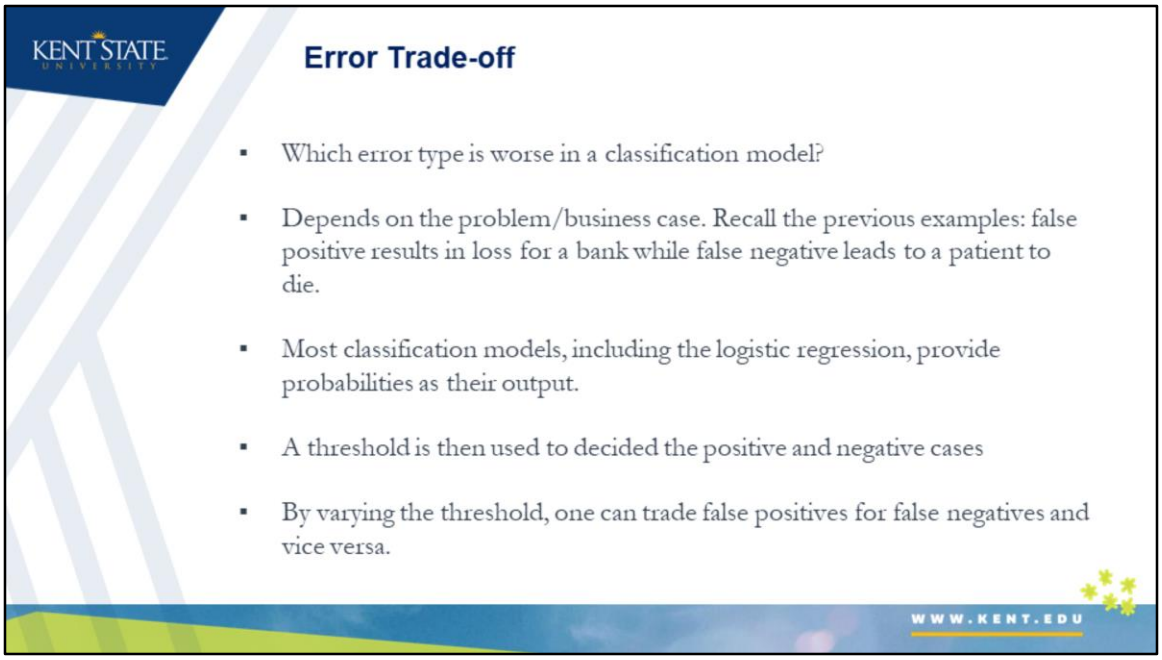
| | | Predicted | |
|---|---|---|---|
| | | Negative | Positive |
| Actual | Negative | **TN** | **FP** |
| | Positive | **FN** | **TP** |

WWW.KENT.EDU

We can represent the four outcomes of our prediction with the actual by means of a confusion matrix. Note that this confusion matrix can be easily calculated regardless of the number of groups in our classification problem. Some of these outcomes have special meaning when doing classification, and we explore that now.




As we saw earlier, for any observed value, our model can either predict it correctly, or incorrectly. As such, there are two correct outcomes, and two incorrect outcomes. The primary question is "which is more important to control and observe?"



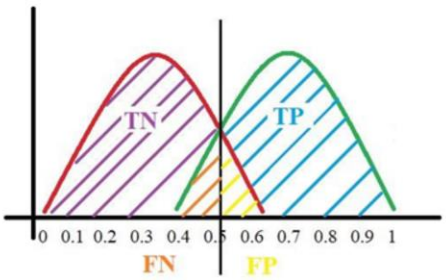

Remember our earlier comment on what we should designate as positive and negative? That affects how we make decisions in determining the severity of the errors? Which is more harmful, a false negative, or a false positive?

Some ML models, instead of providing a strict classification actually provide a probability of the observation being positive. That allows some control. So, instead of indicating whether a particular observation of cancer is malignant, we might instead get a probability of it being malignant. Thus two different decision makers, depending on their level of risk, might take different actions for the same value of probability of malignancy. Obviously, varying the cutoff probability for positive and negative classification will result in a tradeoff between false positives and false negatives.

KENT STATE UNIVERSITY

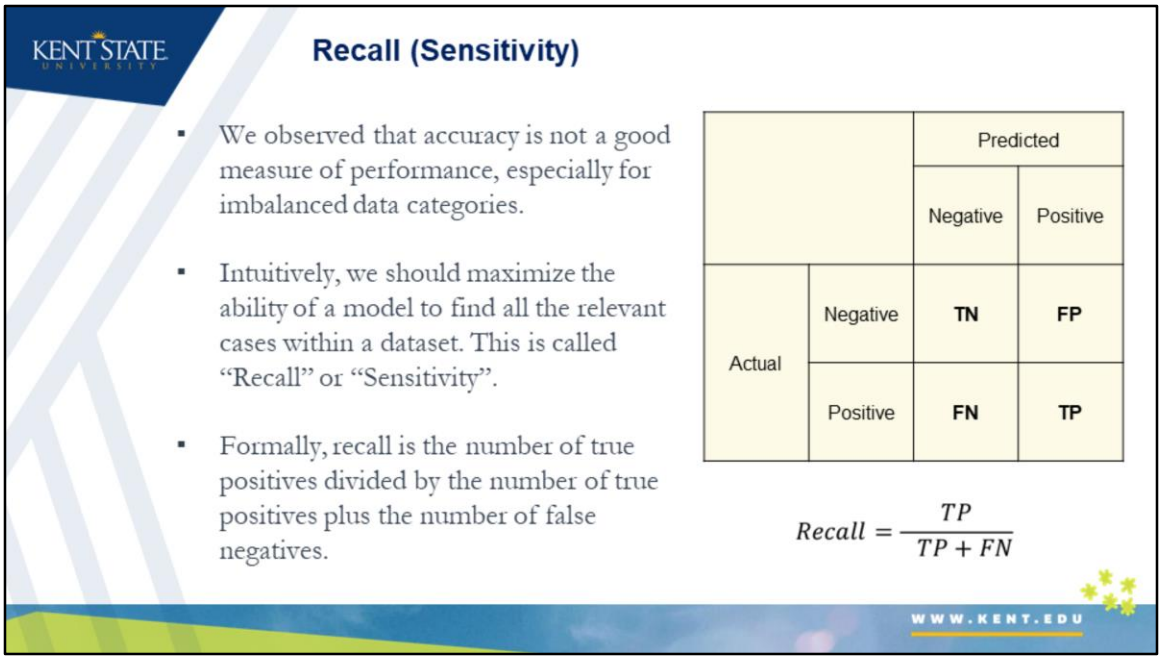
## Error Trade-off continued

- Increasing the threshold value means that we will be more strict in classifying an observation as positive.
- This results in lowering the false positives and true positives and increasing true and false negatives.
- On the other hand, lowering the threshold result in decreased false and true negatives and increased false and true positives.

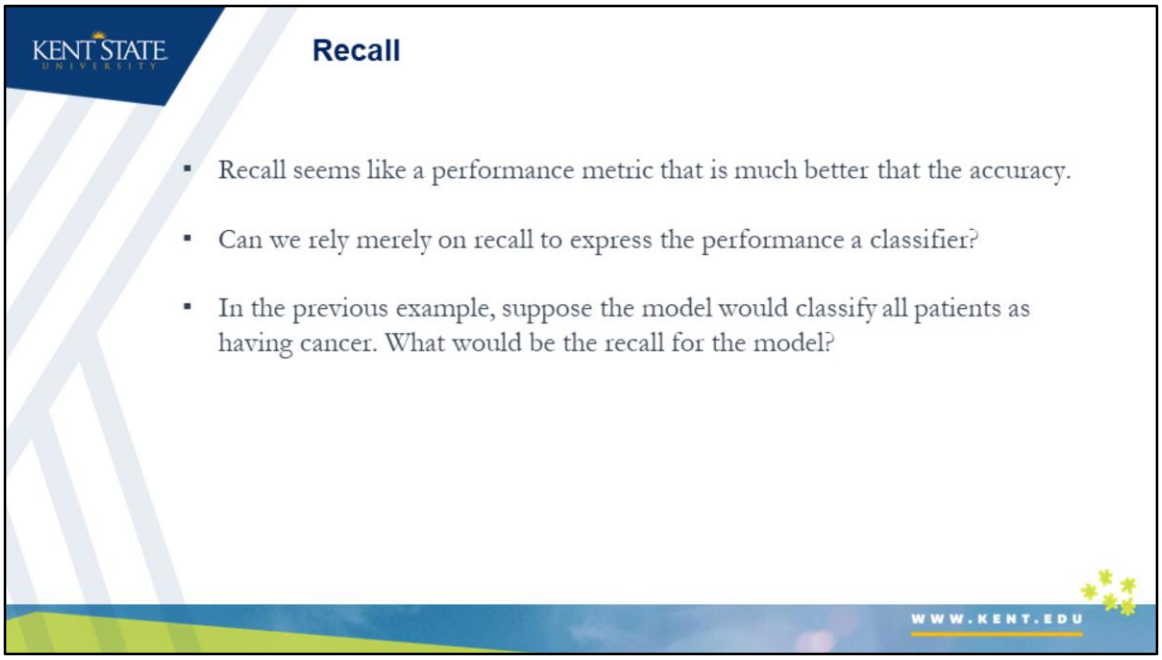



WWW.KENT.EDU

As you can see in the graph above, depending on where we create the cutoff (the black vertical line), the amount of false negatives and false positives vary. For those of you familiar with hypothesis testing in statistics, this is similar to the relationship between Type I and II errors.

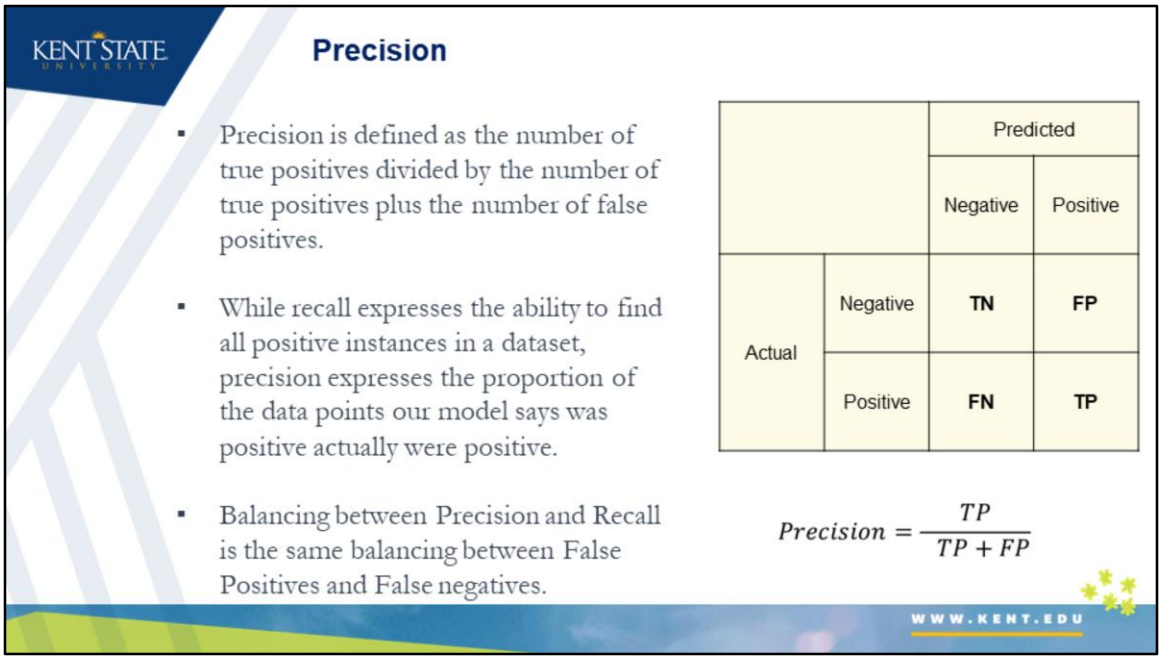



As you might have gathered, accuracy, which is just the measure of true classifications, is not necessarily a good measure by itself, especially when the categories are imbalanced. For example for a two-group classification with 90% in one group and 10% in another group, even a baseline measure for a classifier will achieve 90% accuracy. A more relevant measure may be *sensitivity*, which is the proportion of positives correctly classified. It is defined as the ration of TP to the sum of TP and FN. The denominator is nothing but the actual number of positive items in your sample.

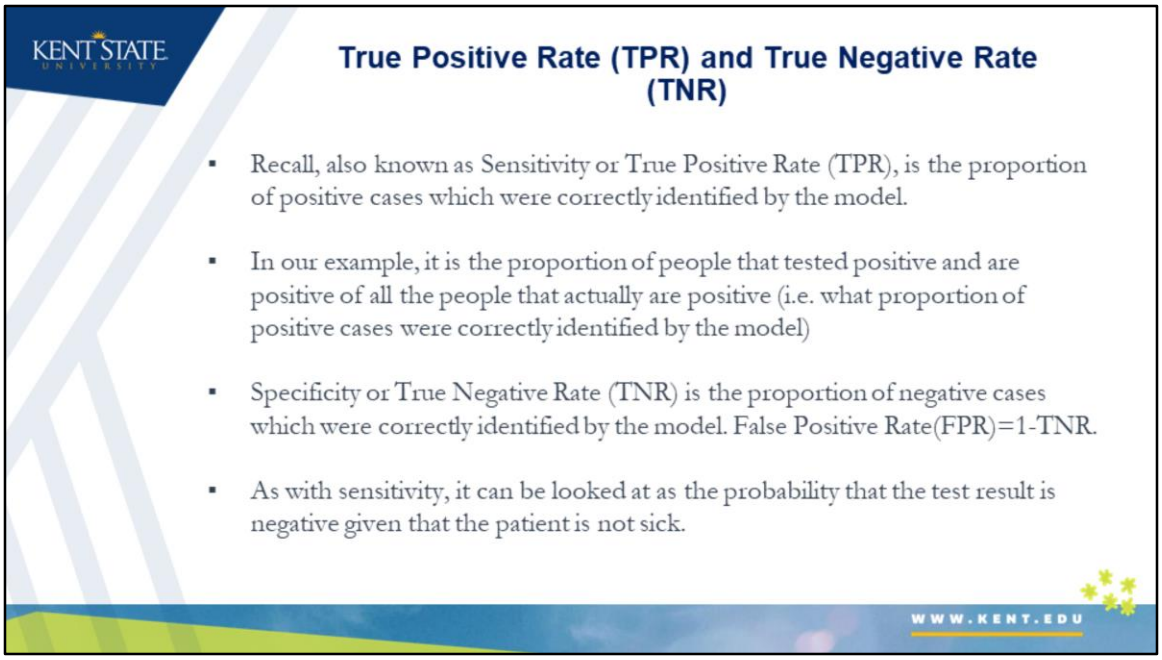



But, recall or sensitivity may also not be sufficient as a measure of performance.

To answer the question, the recall for the model would be 1 since there was no negative prediction and so TN=0 and FN=0. Therefore recall would be TP/(TP+0)=1. As such, recall is not fully sufficient as a measure of performance.




A complementary measure to Recall or Sensitivity is Precision, which is defined as the ration of TP to the sum of TP and FP. In other words, Precision indicates the proportion of data points that were classified as positive and were in reality positive. Convince yourself that the tradeoff between FP and FN that we saw earlier holds true here also.

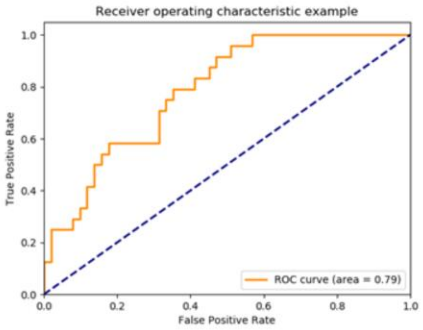



Similar to Sensitivity, we can define a measure called Specificity that is applied to the negatives. Specificity is defined as the proportion of negative cases correctly identified as negative. In notation, TNR = TN / (TN + FP). Thus, the FPR = 1 – TNR.

KENT STATE UNIVERSITY

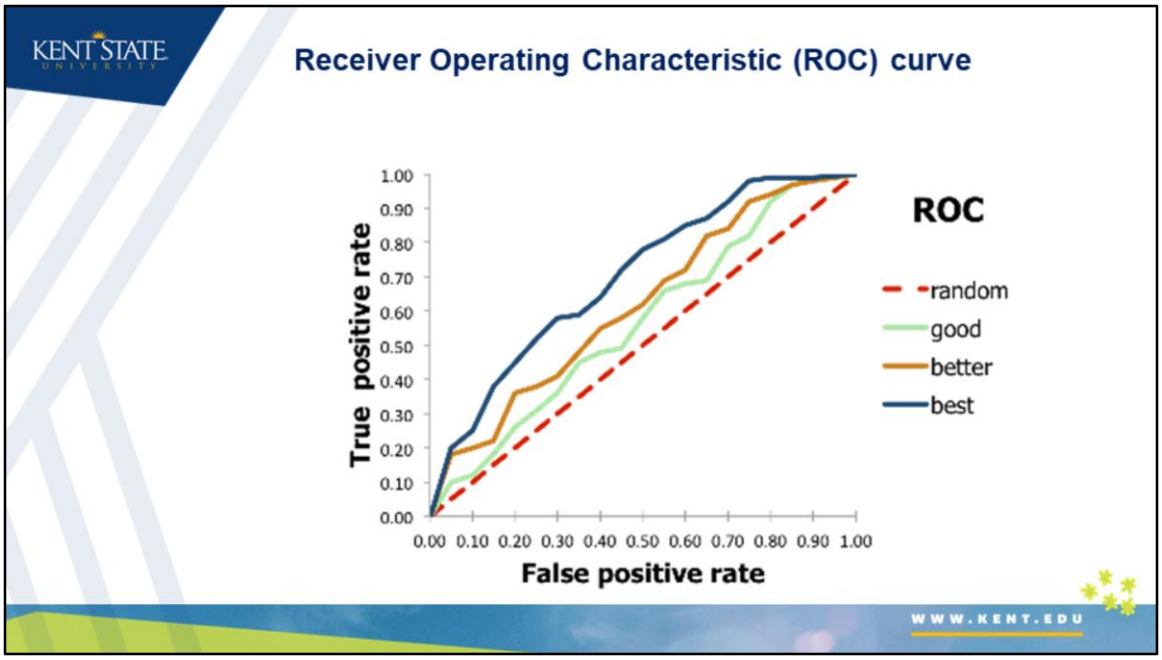
## Receiver Operating Characteristic (ROC) curve

- ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis.
- This means that the top left corner of the plot is the "ideal" point - a false positive rate of zero, and a true positive rate of one.
- This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better.




WWW.KENT.EDU

We now have a visual way of representing our performance measures. This is call ROC curves. We plot the Sensitivity, which is the proportion of positives classified correctly as positive, on the Y axis, versus, 1- Specificity, which is the FPR, on the X axis. Ideally, we want the classifier to occupy the top left point, where Sensitivity = 1, and FPR = 0. In reality, that is rarely the case. We can measure the quality of the classifier by its ROC curve, which has a maximum of 1. The closer the value is to 1, the better the classifier.




The diagonal line is the classifier that we can get on chance. That is our base line performance. Hopefully, any classifier we build will do much better than chance. This concludes our lecture on classification errors.