**FP:** Incorrectly predicted as +ve or Incorrect rejection of Null Hypothesis – Type 1 error

**FN:** Incorrect prediction of –ve or incorrectly retaining the H0 – type 2 error cases

FPR: FP/(total actual Negative cases) = FP/(TN+FP) – also called type-1 error or false threats

FNR = FN/(FN+TP) – also called as type 2 errors or false alarms

TPR = TP/(TP+FN) = 1-FNR = 1-type 2 error

TNR = TN/(TN + FP) = 1-FPR = 1-type 1 error

If reducing type1 error (FPR) is critical for me, we’lll increase the threshold from .5 and if reducing type 2 error is critical for me then we are going to decrease the cutoff . Incase of telecom churn analysis which error is critical? It’s important to correctly predict the high churn cases, i.e incorrect prediction of negative should not happen. To increase my TP cases and reduce the FN cases, we’ll reduce the cut-off.

**ROC Curve, Sensitivity and Specificity**

Sensitivity = TP /(TP+FN) = True Positive Rate = 1- FNR = 1- Type 2 error = %age of actual positives predicted correctly. High sensitivity indicates low FNR or low Type 2 error

When sensitivity is of higher importance, then we want to reduce the threshold to increase the TPR and reduce the FNR

Specificity = TN/(TN + FP) = True Negative Rate = 1-FPR = 1- Type 1 error = %age of actual negatives predicted correctly. High specificity indicates low FPR or low Type 1 error.

When specificity is more important than sensitivity then we are going to increase the cutoff to increase the TNR and reduce the FPR

FNR = same as Type 2 error = %age of actual positives predicted incorrectly

FPR = same as Type 1 error = %age of actual negatives predicted incorrectly

**PRECISION RECALL AND F1 SCORE**

Precision = TP/(TP+FP) – also called as +ve prediction rate – This indicates how precise a model is in it’s positive predictions i.e how many of the positive predictions made are correct. High precision indicates low type 1 error .  
Recall = TP/(TP+FN) – same as sensitivity – high recall indicates low type 2 error.

In both precision and recall, importance has been given to correct prediction of +ves. In addition it talks of which error I am receptive towards? FPR or FNR. In telecom case we want to avoid FNR

F1 score: 2/(1/Precision + 1/Recall)

NPR – Negative prediction Rate = TN/(TN+FN) – i.e what %age of –ve predictions are correct.

**Precision** helps when the costs of false positives are high. So let’s assume the problem involves the detection of skin cancer. If we have a model that has very low precision, then many patients will be told that they have melanoma, and that will include some misdiagnoses. Lots of extra tests and stress are at stake. When false positives are too high, those who monitor the results will learn to ignore them after being bombarded with false alarms.

**Recall** helps when the cost of false negatives is high. What if we need to detect incoming nuclear missiles? A false negative has devastating consequences. Get it wrong and we all die. When false negatives are frequent, you get hit by the thing you want to avoid. A false negative is when you decide to ignore the sound of a twig breaking in a dark forest, and you get eaten by a bear. (A false positive is staying up all night sleepless in your tent in a cold sweat listening to every shuffle in the forest, only to realize the next morning that those sounds were made by a chipmunk. Not fun.) If you had a model that let in nuclear missiles by mistake, you would want to throw it out. If you had a model that kept you awake all night because *chipmunks*, you would want to throw it out, too. If, like most people, you prefer to not get eaten by the bear, and also not stay up all night worried about chipmunk alarms, then you need to optimize for an evaluation metric that’s a combined measure of precision and recall. Enter the F1 score…

**F1 Score**

f1 formula

F1 is an overall measure of a model’s accuracy that combines precision and recall, in that weird way that addition and multiplication just mix two ingredients to make a separate dish altogether. That is, a good F1 score means that you have low false positives and low false negatives, so you’re correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it’s 1, while the model is a total failure when it’s 0.

Remember: All models are wrong, but some are useful. That is, all models will generate some false negatives, some false positives, and possibly both. While you can tune a model to minimize one or the other, you often face a tradeoff, where a decrease in false negatives leads to an increase in false positives, or vice versa. You’ll need to optimize for the performance metrics that are most useful for your specific problem.

**CHANGING THRESHOLD CUT-OFF**

If we reduce the cut-off from .5 to .3, then number of +ve predictions will increase, hence TPR and FPR will also increase, and TNR and FNR – So this is done when it is extremely critical for us to reduce FNR (i.e type 2 error i.e incorrect prediction of -ve). In telecom churn case we have to make sure that high churn cases are predicted correctly i.e the error that is critical for us to avoid: high churn case is not predicted as –ve , so we want to reduce the fnr, because these are the people who need to be retained using measures. But we are okay, if some extra –ve cases are predicted as +ve – so we have higher tolerance toward FPR, but we have low tolerance towards FNR. So here, reducing the threshold will improve our retention.

If we increase the threshold to .7 – we are reducing the +ve predictions and increasing the –ve predictions. Then we are reducing the TPR as well as FPR. So, this is done when reducing the Type 1 error is extremely critical for us. We have a higher tolerance toward FNR (ie incorrect negative predictions). Eg. when we are deciding whether to hang the person or not, here FPR is extremely critical, we have a lil higher receptivity towards FNR; but extremely low tolerance towards hanging an innocent person. So, here we’ll increase the threshold to reduce the +ve predictions.

**ROC Curve is between TPR and FPR.**

At t = 0, all cases are predicted as positive

At t = 1, all cases are predicted as negative

As we increase the threshold from 0 to 1, our prediction of positives decreases, henceTPR and FPR decrease. But a good model is the one where TPR is high and FPR. The worst model is the one where TPR and FPR reduce at the same rate. This indicates the AUC = 50%.

**AUC – Performance Metric**

The best model will be where TPR = 100% and FPR = 0. This model will have AUC = 100%. Generally our models will have AUC around 70 – 80% , any model having AUC > 70% is considered to be a good model. AUC – Area under the curve – it’s an important accuracy metric

As we increase the threshold from 0 to 1, our prediction of negatives increases, hence TNR and FNR increase.

**How to select threshold**

Selecting threshold is a business decision, whether correctly predicting 1 is more important or correctly predicting 0 is more important. Eg. incase of establishing crime / murder, we must not hang an innocent person. So correctly retaining the null is more important, hence TNR is more important or reducing the FPR is more important than predicting the 1s correctly, so in this case we’ll choose a high cut-off (as discussed earlier a high cut-off means more negative predictions)

Q. In the attached ROC graph, find the threshold to minimize FPR, while maximizing TPR. What is the best threshold in the above graph

Ans. .6

Q. Find threshold where 40% of positives are correctly identified, while making as few errors as possible.

Ans. t = .6

**Cohen’s Kappa**

<https://stats.stackexchange.com/questions/82162/cohens-kappa-in-plain-english>