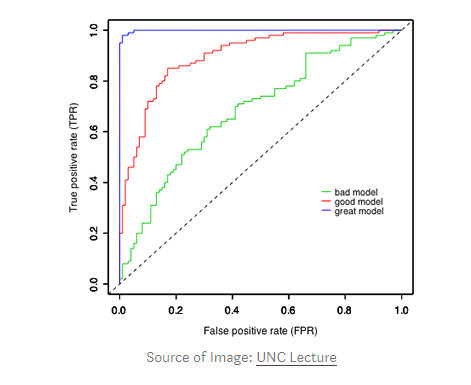
## ROC-AUC Score

The probabilistic interpretation of ROC-AUC score is that if you randomly choose a positive case and a negative case, the probability that the positive case outranks the negative case according to the classifier is given by the AUC. Here, rank is determined according to order by predicted values.

****

*Mathematically, it is calculated by area under curve of sensitivity (TPR) vs.  
FPR(1-specificity). Ideally, we would like to have high sensitivity & high specificity, but in real-world scenarios, there is always a tradeoff between sensitivity & specificity.*

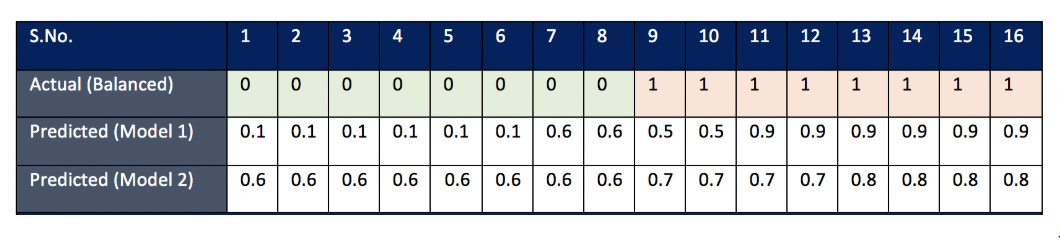
Some important characteristics of ROC-AUC are-

* The value can range from 0 to 1. However auc score of a random classifier for balanced data is 0.5
* ROC-AUC score is independent of the threshold set for classification because it only considers the rank of each prediction and not its absolute value. The same is not true for F1 score which needs a threshold value in case of probabilities output
* **Log-Loss**
* Log-loss is a measurement of accuracy that incorporates the idea of probabilistic confidence given by following expression for binary class:

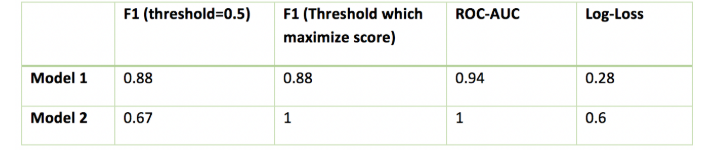
https://miro.medium.com/max/263/0*Uig_BbP8Pe1vd0q4.png

* It takes into account the uncertainty of your prediction based on how much it varies from the actual label. In the worst case, let’s say you predicted 0.5 for all the observations. So log-loss will become -log(0.5) = 0.69. Hence, we can say that anything above 0.6 is a very poor model considering the actual probabilities.

**Comparison of Log-loss with ROC & F1**



Consider Case 1 (Balanced Data), it looks like model 1 is doing a better job in predicting the absolute probabilities whereas model 2 is working best in ranking observations according to their true labels. Let’s verify with the actual score:

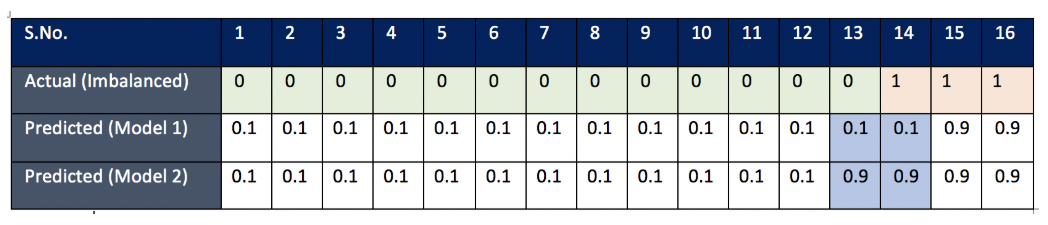


If you consider log-loss, Model 2 is worst giving a high value of log-loss because the absolute probabilities have big difference from actual labels. But this is in complete disagreement with F1 & AUC score, according to which Model 2 has 100% accuracy. Also, you would like to note that with different thresholds, F1 score is changing, and preferring model 1 over model 2 for default threshold of 0.5.

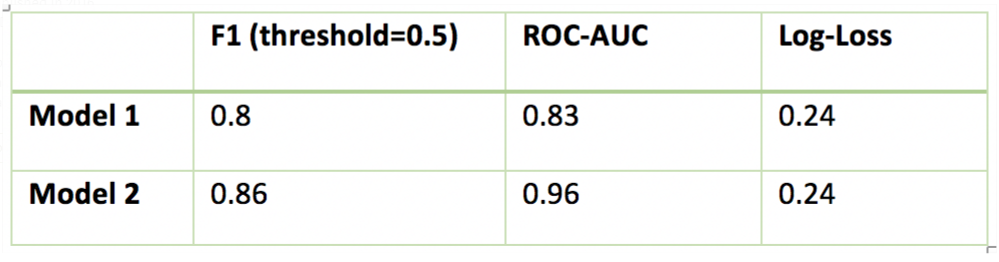
*Inferences drawn from the above example (balanced):  
- If you care for absolute probabilistic difference, go with log-loss  
- If you care only for the final class prediction and you don’t want to tune threshold, go with AUC score  
-F1 score is sensitive to threshold and you would want to tune it first before comparing the models*

# Case 2

## How each of them deals with class imbalance?



The only difference in the two models is their prediction for observation 13 & 14. Model 1 is doing a better job in classifying observation 13 (label 0) whereas Model 2 is doing better in classifying observation 14 (label 1). The goal is to see which model actually captures the difference in classifying the imbalanced class better (class with few observations, here it is label 1). In problems like fraud detection/spam mail detection, where positive labels are few, we would like our model to predict positive classes correctly and hence we will sometime prefer those model who are able to classify these positive labels

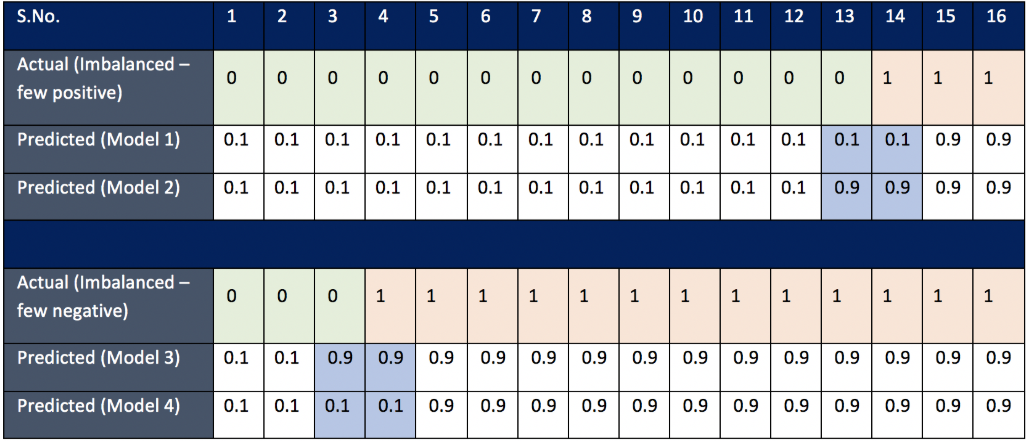


*Clearly log-loss is failing in this case because according to log-loss both the models are performing equally. This is because log-loss function is symmetric and does not differentiate between classes .*

Both F1 score and ROC-AUC score is doing better in preferring model 2 over model 1. So we can use both these methods for class imbalance. But we will have to dig further to see how differently they treat class imbalance.

ROC-AUC score handled the case of few negative labels in the same way as it handled the case of few positive labels. An interesting thing to note here is that F1 score is pretty much same for both Model 3 & Model 4 because positive labels are large in number and it cares only for the misclassification of positive labels.

*Inferences drawn from above example:  
- If you care for a class which is smaller in number independent of the fact whether it is positive or negative, go for ROC-AUC score.*



ROC-AUC score handled the case of few negative labels in the same way as it handled the case of few positive labels. An interesting thing to note here is that F1 score is pretty much same for both Model 3 & Model 4 because positive labels are large in number and it cares only for the misclassification of positive labels.

*Inferences drawn from above example:  
- If you care for a class which is smaller in number independent of the fact whether it is positive or negative, go for ROC-AUC score.*

## When will you prefer F1 over ROC-AUC?

When you have a small positive class, then F1 score makes more sense. This is the common problem in fraud detection where positive labels are few. We can understand this statement with the following example.