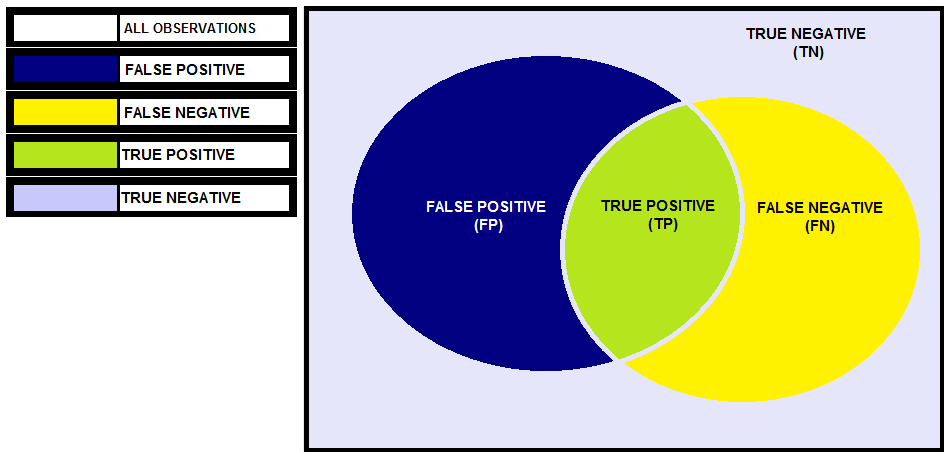
MODEL EVALUATION – CLASSIFICATION MODELS

**V**arious model evaluation techniques can be used under the supervised learning setup that helps us in finding the performance of the model. A very simple method to evaluate a model is by finding the accuracy which is the difference between the predicted and the actual values and when we are working with classification models then by accuracy what we mean is the count of the correct predictions of the class labels, however, it is not a perfect method and can lead to poor decision making. Therefore, we need other measures to evaluate the various models and picking the right measure of evaluation is crucial in selecting and distinguishing the right model from other models.

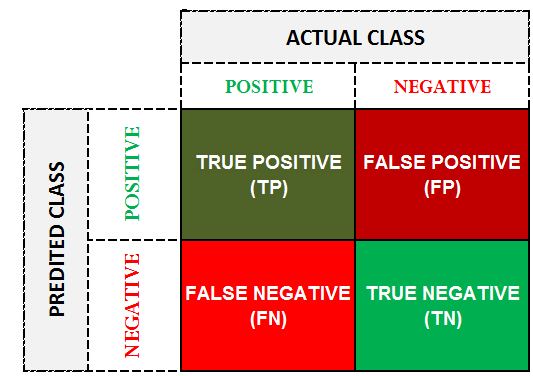
To understand the various metrics that can be used to evaluate a classification problem, we first need to understand what we mean by a Base Rate Model. For example, we have a dataset with 100 records where 90 records belong to class label ‘0’ while the remaining 10 observations belong to class label ‘1’. As the frequency of the class label ‘0’ is higher than class label ‘1’ then a model which predicts all observations to have the class label ‘0’ will have an accuracy of as high as 90%. Therefore, the accuracy here provides us with little information on how actually the model is performing. It is also important for our model to have accuracy at least more than 90% because a model that simply predicts the most frequent class has the accuracy of 90%. This limitation of accuracy asks for more complex measures of evaluation.

**Confusion Matrix**

For example, we have a dataset having class labels 0 and 1 where 0 stands for ‘Non-Defaulters’ while 1 stands for ‘Defaulters’. We build a logistic regression model to predict the class label 1. Below is a Venn diagram where all the observations are in the square box. All the observations that were predicted as 1 by the model are represented as the Blue Circle. All the observations that were actually 1 are represented by the yellow circle. The common area of these two circles is denoted by green and contains the observations that were correctly predicted by the model and are called True Positives. False Positive observations are those observations that were incorrectly predicted as 1 by our model while False Negative are those observations that were incorrectly predicted as having the class label ‘0’ by our model. The remaining observations are the True Negative which are those observations that were correctly predicted as having the class label ‘0’ by the logistic regression model.



Rather than a Venn diagram, we can come up with a confusion matrix to show all these observations. For a two-class classification problem, the confusion matrix will have two rows and two columns where the rows represent the actual values whereas the columns represent predicted values. Our objective is to minimize the False Positives and False Negatives and maximise the True Positives and True Negatives thus we want to maximise the diagonal values.



Positive (1) – Defaulter

Negative (0) – Non-Defaulter

True Positive- Model predicted ‘Defaulter’ and were actually ‘Defaulter’

True Negative- Model predicted ‘Non-Defaulter’ and were actually ‘Non-Defaulter’

False Positive- Model predicted ‘Non-Defaulter’ and were actually ‘Defaulter’ (Type I Error)

False Negative- Model predicted ‘Defaulter’ and were actually ‘Non-Defaulter’ (Type II Error)

The real power of confusion matrix lies with the various metrics that can be calculated using the confusion matrix. The most basic measures involve Classification Error and Accuracy.

**Classification Error**

The formula for Classification Error is (FP + FN) ÷ (TP + TN + FP + FN). Here we divide all the errors with all the observations (ERRORS ÷ TOTAL).

**Accuracy**

It is simply 1-Error. Here we divide all the correct predictions with all the observations (CORRECT ÷ TOTAL). The formula for Accuracy is (TP + TN) ÷ (TP + TN + FP + FN)

However, as mentioned earlier, these measures can be very misleading. Thus, we come up with more complex measures to evaluate the model. Some of these measures are mentioned below:

**False Positive Rate aka False Alarm rate aka 1-Specificity aka 1-True Negative Rate**

It provides us with the percentage of negatives that were classified as positive by the model. Here we divide the false positive with the total number of observations that could have come under False Positive making the formula to be FP ÷ (FP + TN)

**False Negative Rate aka Miss rate**

It provides us with the percentage of the positives that were classified as negative by the model. Thus, out of all the positives in the data, the positives that were miss-classified by the model is indicated by the False Negative Rate. The formula for it is FN ÷ (FN + TP)

**Recall aka Sensitivity aka True Positive Rate aka 1-False Negative Rate**

It is the opposite of False Negative Rate. It provides us with the percentage of positives that were classified correctly by the model. The formula for it is = TP ÷ (TP + FN)

**Precision aka Positive Predicted Value**

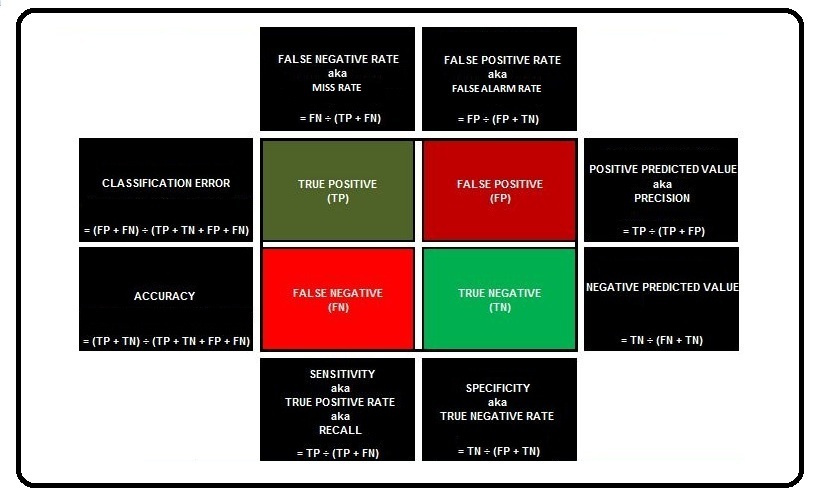
Out of all the predicted positives, the fraction of the actual positive is indicated by precision. Thus, precision provides us with the percentage of the positive out of what the model predicted as positive. Therefore, unlike Recall which represents the percentage of True positive out of all the actual positives, Precision provides us with the percentage of True-positive out of all predicted positives. By precision, we get to know how many of the positives predicted are relevant which help us to evaluate the model. The formula for Precision is = TP ÷ (TP + FP)

**Specificity aka True Negative Rate**

It indicates how good the model is at avoiding the misclassification of the negatives I.e. Out of the predicted Negatives, how many of them are actually negatives. The formula for Specificity is = TN ÷ (FP + TN)

**Negative Predicted Value**

Here we divide the True Negatives with the False Negative and True Negative. The formula for Specificity is = TN ÷ (FN + TN)



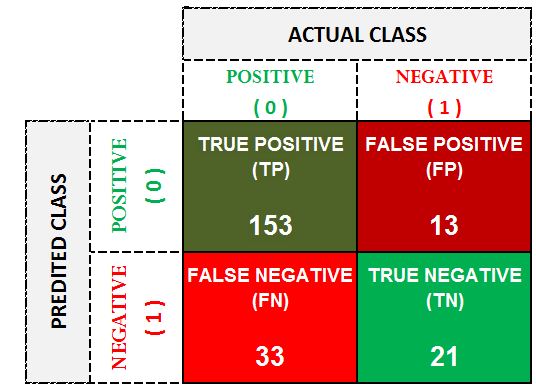
All the above-mentioned metrics are useful, however, they alone cannot provide us with a good evaluation score. For example, if a model predicts all the observations as class label ‘0’ then the False Positive will be nill or a model that predicts class label ‘1’ for all the observations will have a 100% recall. Thus, we need to consider these values in pairs such as Recall-Precision, False Positive-False Negative Rate, True Positive-False Negative Rate etc.

**ROC Curve**

For understanding the need of ROC Curve, we first need to understand the role played by the threshold in the classification models.

For example, we have a dataset with 220 observations having demographic, financial and other information about people who have taken a loan from XYZ Bank. The observations have two classes, class labels ‘0’ and ‘1’ where ‘0’ denotes Non-Defaulters i.e. those people who were able to repay their loans while ‘1’ indicates ‘defaulters’ i.e. those people who failed to repay their loan.

We fit a logistic regression model on this data to solve this binary classification problem and predict the class label ‘0’ and come up with the following confusion matrix.



With the following confusion matrix, we know that the actual number of non-defaulters is 186 (TP + FN) while the actual number of defaulters is 33 (FP + TN). The Total Number of Predicted Non-Defaulters is 165 (TP + FP) while the Predicted Defaulters are 54 (FP + TN). Now if we concentrate on the predicted negatives then we can calculate specificity which tells us out of the predicted Negatives, how many of them are actually negatives and from the above confusion matrix we get specificity to be at 64% (the formula for Specificity is TN ÷ (FP + TN)).

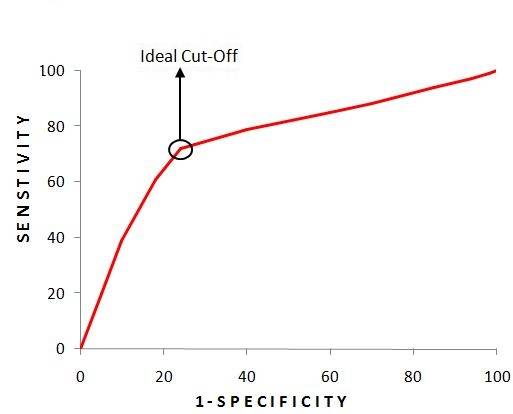
This means that as far as the defaulters are concerned we are 64% accurate in classifying the defaulters. However, many of the classification techniques such as Logistic Regression, Decision Trees, Naive Bayes doesn’t explicitly provide us with class labels rather they compute the probability of an observation belonging to a certain class. By default this threshold is set at 50% which means in our case any observation that has the probability of less than 50% will be provided with the class label ‘1’ i.e. will be labelled as a defaulter. As mentioned above, the various measures discussed under confusion matrix can be easily manipulated as a model predicting all observations as ‘1’ will increase the Specificity, thus, we will be able to classify all the defaulters correctly, however, by doing so we will end up labeling a lot of non-defaulters as defaulters and we will end up increasing the False Negatives. This decision of how much of False Negative or False Positive is admissible is very domain specific. For example, if the Bank XYZ is liberal and can tolerate bad debts and wants to grant the loan to more people then the bank will want to have high Sensitivity and will tolerate low Specificity. However, if the bank is stringent and wants to be very cautious in dispersing loans then it won’t bother low Sensitivity and will want to have high Specificity. Now all these measures depend on the values that we have as True Positives, True Negatives, False Positive and False Negatives which can be manipulated by altering the threshold value. For example, if we raise the threshold value to 90% then it will become tough for an observation to be categorised as ‘Non-Defaulter’ and this will affect the various measures.

Thus, we can say that the True Positive and False Positive Rates are simply a function of the threshold and for different values of threshold we will have different values of such measures. This is where we take help of Receiver Operating Characteristic Curve or ROC Curve which is originally based on a military communication device which was used to communicate through radio and morse code. With the help of ROC Curve, we consider different threshold values and get to know the different values of True Positive Rate (Sensitivity) and False Positive Rate (1-Specificity).

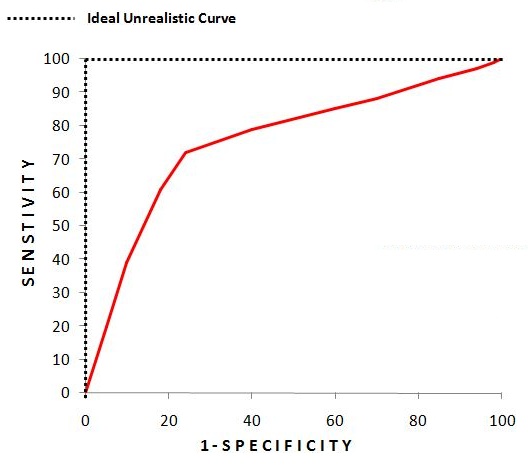
**Sensitivity v/s 1-Specificity**

On one hand, we have Sensitivity which provides us with the percentage of positives that were classified correctly by the model and naturally we want this number to be high. On the other hand, we have Specificity which indicates how good the model is at avoiding the misclassification of the negatives i.e. out of the predicted negatives, how many of them are actually negatives and naturally we want this number to be as high as possible. Now if these values are high then it means that the diagonal values in the confusion matrix are high indicating that the number of False Positive and Negative observations are low. Now we want to see for which value of threshold these two numbers are the highest, however, as we want both Sensitivity and Specificity to be as high as possible, it makes it difficult to plot their values on a graph for different values of thresholds. Thus, rather than looking for very high Specificity we look for very low False Positive Rate which indicates the percentage of negatives that were classified as positive by the model (basically an opposite of Specificty i.e. 1-Specificity).

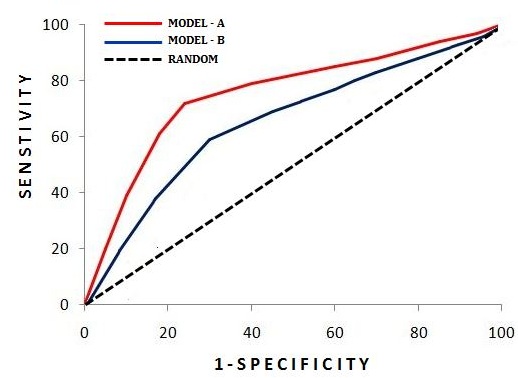
Now for different levels of Threshold when we plot the values of Sensitivity and 1-Specificity we get the following graph.



Here for the Logistic Regression Model, we get different values of specificity and sensitivity for different values of the thresholds. The ideal scenario will be where the value of threshold that provides us with 100 sensitivity and 0 1-specificity, however, as it is the most ideal case which is generally not possible, we pick the next best thing which is the ideal cutoff where the difference between the two is highest i.e. we pick that value of threshold where we can have the maximum sensitivity with minimum 1-specificity. (The point closest from the upper left corner, the probability corresponding to it gives the best cutoff)

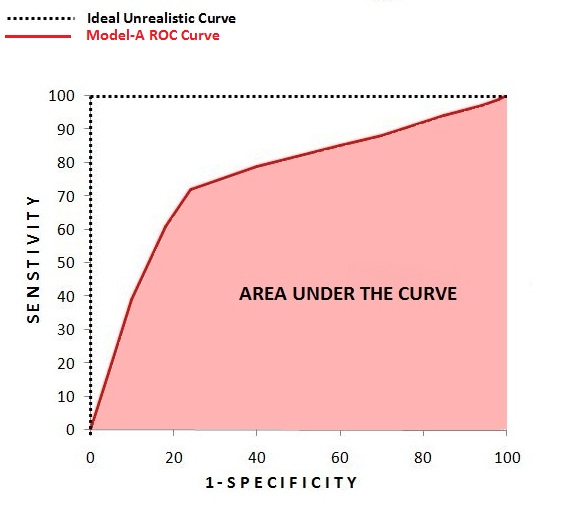


ROC Curve can also be used to compare two models as stand-alone values of different measures cannot make much sense to compare two models. Below we see for different values of the threshold i.e. the values of Sensitivity and 1-Specificity for Model-A (Logistic Regression) and Model-B (Decision Trees). Here no matter which threshold we consider for Model-A, it will provide better results than results provided by the ideal cut-off of Model-B. The dotted black line represents a ROC curve of a completely random classifier. For example, if we have a dataset which has an equal number of class label ‘0’ and ‘1’ then a model which predicts all of them as 0 will have this dotted line as shown in the ROC curve.



**AUC Curve**

It is a single number metric which is an extension of the ROC Curve. Here we measure the area under the curve. For the perfect unrealistic curve, the area under the curve will be 100%. While for the completely random model the area under the curve will be 50%. A model which performs worse than a random model will have an area below 50%. Here Model-A will be more accurate than Model-B as the area under the curve is more for Model-A than it is for Model-B. Thus, AUC Curve provides us with a single value metric which quickly helps us to understand the performance of the classification model.



If we compare a base rate model with a model which is more accurate then the difference between their AUC score will be starker than the difference between their Sensitivity.

Thus, AUC score provides us with a better idea of how the model is working as with the ROC curve sometimes the difference between the two models may not be very visible, with the ROC curves of different models intersecting each other with models performing better in some regions than other. AUC score provides us with good overall evaluation.

All the metrics discussed so far played a crucial role in order to evaluate the performance of a model. We use these metrics during the training phase and after evaluating a model’s performance on the training dataset, we apply the model on the test set and again conduct a series of tests using the above-mentioned measures to evaluate the model. However, we need to validate a model in order to properly understand how well it performs with unseen or test data. To evaluate model’s performance with unseen data, we can perform various model validation methods.