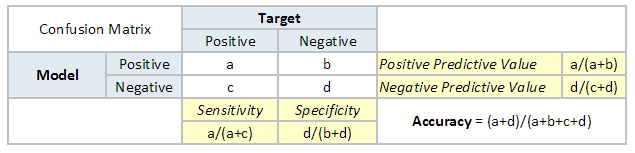
**Confusion Matrix :**

A confusion matrix is an N X N matrix, where N is the number of classes being predicted. For the problem in hand, we have N=2, and hence we get a 2 X 2 matrix.

* **Accuracy** : the proportion of the total number of predictions that were correct.
* **Positive Predictive Value or Precision** : the proportion of positive cases that were correctly identified.
* **Negative Predictive Value** : the proportion of negative cases that were correctly identified.
* **Sensitivity or Recall** : the proportion of actual positive cases which are correctly identified.
* **Specificity** : the proportion of actual negative cases which are correctly identified.



A **Hypothesis** is speculation or theory based on insufficient evidence that lends itself to further testing and experimentation. With further testing, a hypothesis can usually be proven true or false.

A **Null Hypothesis** is a hypothesis that says there is no statistical significance between the two variables in the hypothesis. It is the hypothesis that the researcher is trying to disprove.

* We would always reject the null hypothesis when it is false, and we would accept the null hypothesis when it is indeed true.

**There are two types of errors that can occur.**

* Type 1 errors
* Type II errors.

|  |  |
| --- | --- |
| **Type I Error** | **Type II Error** |
| * equivalent to False Positives(FP). * Let’s go back to the example of a drug being used to treat a disease. If we reject the null hypothesis in this situation, then we claim that the drug does have some effect on a disease. But if the null hypothesis is true, then, in reality, the drug does not combat the disease at all. The drug is falsely claimed to have a positive effect on a disease. | * equivalent to False Negatives(FN). * If we think back again to the scenario in which we are testing a drug, what would a type II error look like? A type II error would occur if we accepted that the drug hs no effect on disease, but in reality, it did. |

### Accuracy :

* Accuracy is a common evaluation metric for classification problems. It’s the number of correct predictions made as a ratio of all predictions made.



* **Misclassification Rate(Error Rate):**

Overall, how often is it wrong. Since accuracy is the percent we correctly classified (success rate), it follows that our error rate (the percentage we got wrong) can be calculated as follows:

Misclassification Rate = (FP+FN)/total

* **Precision :** It is the number of correct positive results divided by the number of positive results predicted by the classifier.



* **Recall :** It is the number of correct positive results divided by the number of ***all*** relevant samples (all samples that should have been identified as positive).

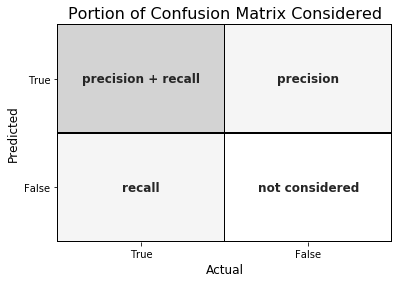
Recall = True Positive/( True Positive + False Negative )

High precision means that an algorithm returned substantially more relevant results than irrelevant ones.

High recall means that an algorithm returned most of the relevant results

# ****F1 Score****

* F1 Score is used to measure a test’s accuracy. F1 Score tries to find the balance between precision and recall.
* The range for F1 Score is [0, 1].
* F1 Score is the Harmonic Mean between precision and recall. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).
* High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify.
* The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as :



***Why harmonic mean?***

* Since the harmonic mean of a list of numbers skews strongly toward the least elements of the list, it tends (compared to the arithmetic mean) to mitigate the impact of large outliers and aggravate the impact of small ones.

An F1 score punishes extreme values more. Ideally, an F1 Score could be an effective evaluation metric in the following classification scenarios:

* *When FP and FN are equally costly — meaning they miss on true positives or find false positives — both impact the model almost the same way, as in our cancer detection classification example*
* *Adding more data doesn’t effectively change the outcome effectively*
* *TN is high (like with flood predictions, cancer predictions, etc.)*

# ****Logarithmic Loss****

* Logarithmic Loss or Log Loss, works by penalising the false classifications.
* It works well for multi-class classification. When working with Log Loss, the classifier must assign probability to each class for all the samples.
* Suppose, there are N samples belonging to M classes, then the Log is calculated as

below :



where,

y\_ij, indicates whether sample i belongs to class j or not

p\_ij, indicates the probability of sample i belonging to class j

* Log Loss has no upper bound and it exists on the range [0, ∞).
* Log Loss nearer to 0 indicates higher accuracy, whereas if the Log Loss is away from 0 then it indicates lower accuracy.

In general, minimising Log Loss gives greater accuracy for the classifier.

# ****Area Under Curve****

* Area Under Curve(AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem.
* AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.
* **True Positive Rate (Sensitivity)** : True Positive Rate is defined as *TP/ (FN+TP)*. True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.



* **True Negative Rate (Specificity)** : True Negative Rate is defined as *TN / (FP+TN)*. False Positive Rate corresponds to the proportion of negative data points that are correctly considered as negative, with respect to all negative data points.



* **False Positive Rate(1 -Specificity)** : False Positive Rate is defined as *FP / (FP+TN)*. False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.



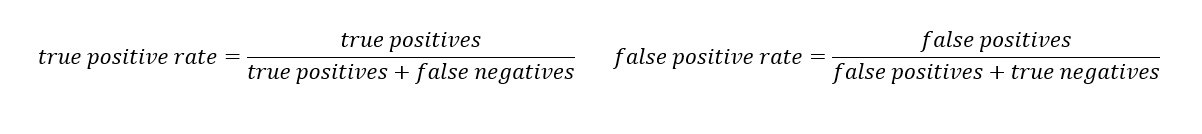
* *False Positive Rate* and *True Positive Rate* both have values in the range **[0, 1]**.
* *FPR* and *TPR* both are computed at varying threshold values such as (0.00, 0.02, 0.04, …., 1.00) and a graph is drawn.
* *AUC* is the area under the curve of plot *False Positive Rate vs True Positive Rate* at different points in **[0, 1]**.
* AUC *has a range of [0, 1]. The greater the value, the better is the performance of our model.*

AUC is desirable for the following two reasons:

* AUC is **scale-invariant**. It measures how well predictions are ranked, rather than their absolute values.
* AUC is **a classification-threshold-invariant**. It measures the quality of the model’s predictions irrespective of what classification threshold is chosen.

### Decision Threshold & Receiver Operating Characteristic (ROC) curve :

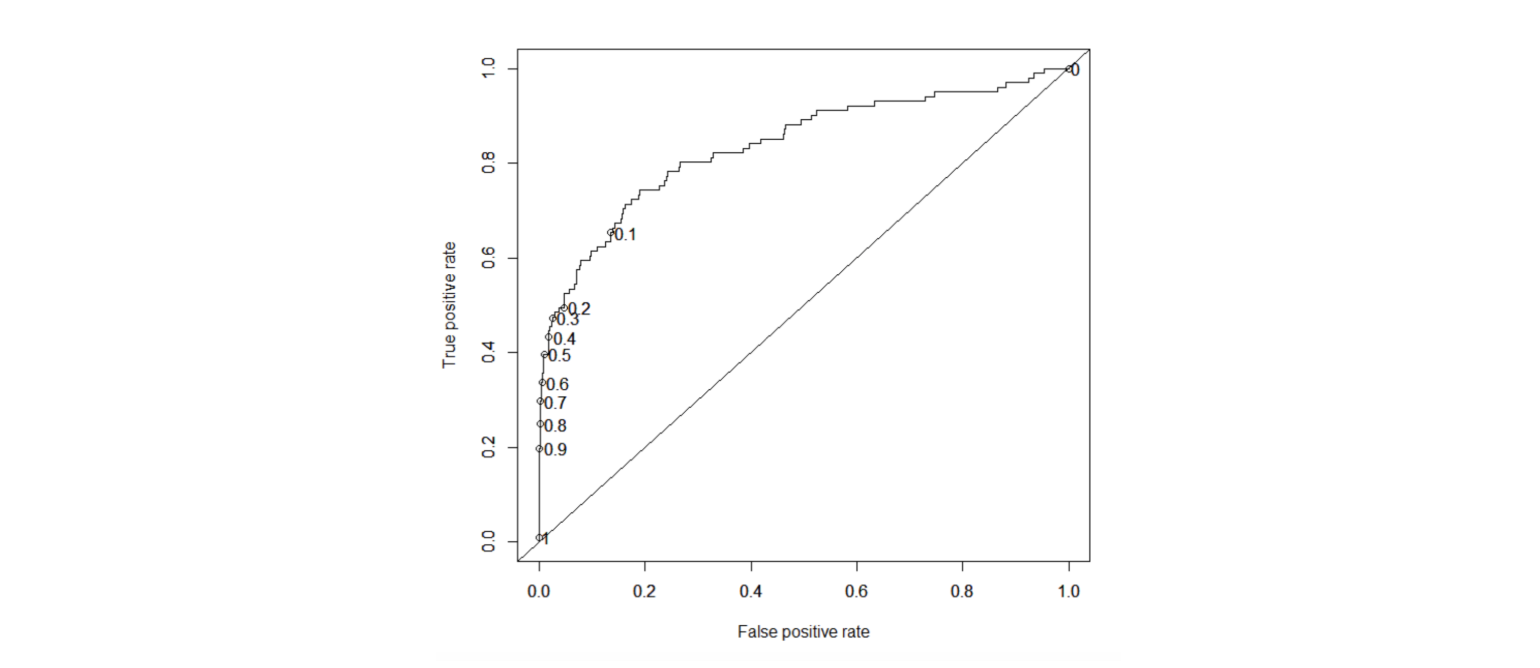
ROC is a major visualization technique for presenting the performance of a classification model. It summarizes the trade-off between the true positive rate (tpr) and false positive rate (fpr) for a predictive model using different probability thresholds.



**The true positive rate (tpr)** is the recall and the **false positive rate (FPR)** is the probability of a false alarm.

* By plotting the true positive rate (sensitivity) versus the false-positive rate (1 — specificity), we get the **Receiver Operating Characteristic** (**ROC**) **curve**.
* This curve allows us to visualize the trade-off between the true positive rate and the false positive rate.

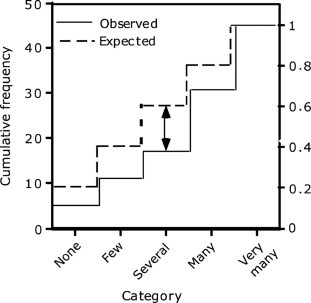
1. A ROC curve plots the true positive rate (tpr) versus the false positive rate (fpr) as a function of the model’s threshold for classifying a positive.
2. Given that **c**is a constant known as decision threshold, the below ROC curve suggests that by default c=0.5, when c=0.2, both tpr and fpr increase. When c=0.8, both tpr and fpr decrease. In general, tpr and fpr increase as c decrease. In the extreme case when c=1, all cases are predicted as negative; tpr=fpr=0. On the other hand, when c=0, all cases are predicted as positive; tpr=fpr=1.



### Kolomogorov Smirnov chart

K-S or Kolmogorov-Smirnov chart measures the performance of classification models. More accurately, K-S is a measure of the degree of separation between positive and negative distributions.

The K-S is 100 if the scores partition the population into two separate groups in which one group contains all the positives and the other all the negatives. On the other hand, If the model cannot differentiate between positives and negatives, then it is as if the model selects cases randomly from the population. The K-S would be 0.



In most classification models the K-S will fall between 0 and 100, and that the higher the value the better the model is at separating the positive from negative cases.

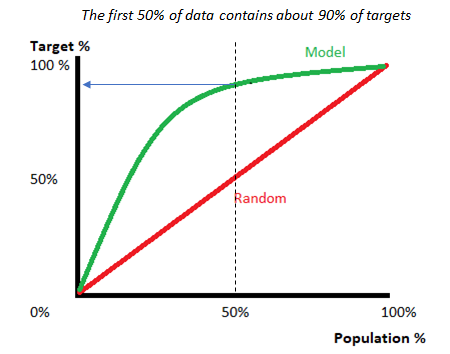
The K-S may also be used to test whether two underlying one-dimensional probability distributions differ. It is a very efficient way to determine if two samples are significantly different from each other.

### Gain and Lift Chart

* Gain or Lift is a measure of the effectiveness of a classification model calculated as the ratio between the results obtained with and without the model.
* Gain and lift charts are visual aids for evaluating the performance of classification models.
* However, in contrast to the confusion matrix that evaluates models on the whole population gain or lift chart evaluates model performance in a portion of the population.

The higher the lift (i.e. the further up it is from the baseline), the better the model.

The following gains chart, run on a validation set, shows that with 50% of the data, the model contains 90% of targets, Adding more data adds a negligible increase in the percentage of targets included in the model.



### Gini Coefficient :

* The Gini coefficient or Gini Index is a popular metric for imbalanced class values.
* The coefficient ranges from 0 to 1 where 0 represents perfect equality and 1 represents perfect inequality. Here, if the value of an index is higher, then the data will be more dispersed.

Gini coefficient can be computed from the area under the ROC curve using the following formula:

**Gini Coefficient = (2 \* ROC\_curve) — 1**

## Concordant – Discordant ratio

This is again one of the most important metric for any classification predictions problem. To understand this let’s assume we have 3 students who have some likelihood to pass this year. Following are our predictions :

A – 0.9

B – 0.5

C – 0.3

Now picture this. if we were to fetch pairs of two from these three student, how many pairs will we have? We will have 3 pairs : AB , BC, CA. Now, after the year ends we saw that A and C passed this year while B failed. No, we choose all the pairs where we will find one responder and other non-responder. How many such pairs do we have?

We have two pairs AB and BC. Now for each of the 2 pairs, the concordant pair is where the probability of responder was higher than non-responder. Whereas discordant pair is where the vice-versa holds true. In case both the probabilities were equal, we say its a tie. Let’s see what happens in our case :

AB  – Concordant

BC – Discordant

Hence, we have 50% of concordant cases in this example. Concordant ratio of more than 60% is considered to be a good model. This metric generally is not used when deciding how many customer to target etc. It is primarily used to access the model’s predictive power. For decisions like how many to target are again taken by KS / Lift charts.

<https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html>

<https://www.kdnuggets.com/2020/04/performance-evaluation-metrics-classification.html>

<https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>

<https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/>

<https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b>

<https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm>

<https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/>

<https://heartbeat.fritz.ai/evaluation-metrics-for-machine-learning-models-d42138496366>