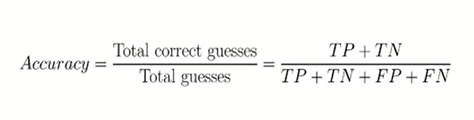
Precison,F1 Score, Recall & Accuracy Metrics: Only Key Points

Accuracy :

* Tells how much we got right answers in both positive and negative classes.

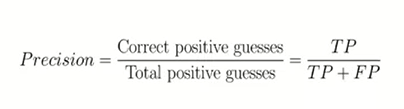


* Should NEVER be used with Biased/Imbalanced Data Sets.

Example: say, a data set with 99% labels(or output variables) as say, negative ‘-‘ . The Model then learns from this data and what happens is that for all possible combinations of input variables, it gets trained to give ‘-‘ as output (b/c of this Biaseness on training data) . **Thus, usually , in such cases, the Accuracy of our Model with even Testing Data is very high (because Testing and Training Data are created usually from the same Dataset. Hence Testing Data also has that same Biasness)**.

* Wherever the costs of having a mis-classified actual positive (or false negative) is very high, ex classifying a person as a Terrorist or Not, or, Identifying Cancer cells or Not, We should be very careful. i.e. **where our Positive Classification has these risky cases i.e. Positive implies a Terrorist or Cancer, then make sure that the dataset is NOT Imbalanced and if possible go for Precision or Recall.**

Precision:



* Tells us how many TP out of total ‘Predictions’ of Positives that our Model made.
* **Precision is a good measure to determine, when the costs of False Positive is high.** For instance, email spam detection. In email spam detection, a false positive means that an email that is non-spam (actual negative) has been identified as spam (predicted spam). The email user might lose important emails if the precision is not high for the spam detection model.
* **Solves issue of High Accuracy for an Imbalanced dataset, by penalizing the Precision to a very small value ( closer to 0).**

Example : Dataset is : 99% people are healthy,

Then, Model starts predicting all (or, say 99% people ) as healthy, even when they are NOT.

Thus, FP ( = 0.99) and TP = 0 . Thus, Precision ~~ 0.00001

* Drawback: Does not paint the complete picture as it does not talk about negative classifications. Thus, we also use Recall with it.
* Drawback: Also, it can cheat the system if the Dataset has just 1 TP and no FP (then Precision =1), and has Many FN(which is a bad case b/c FN is undesirable).

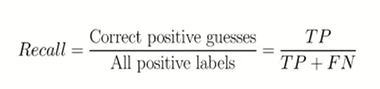
Example: Case in Spam Detection : (FP is Ham classified as Spam) .

i) If FP is High 🡪 Precision is low . (Undesirable , so Model is Not good , better Model must be developed ) Note: Dataset is NOT Biased here.

ii) if FP is low 🡪 Precision is High. (Desirable , so Model is working good .

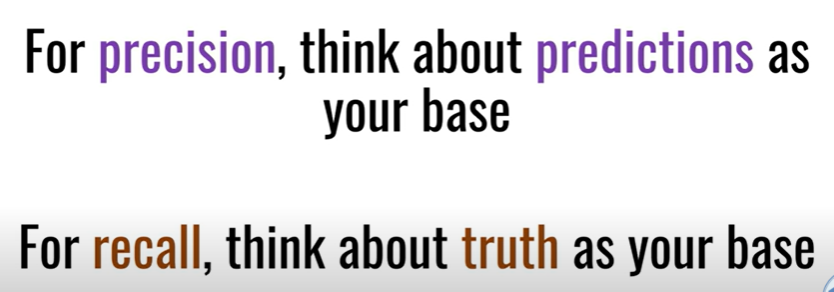
Note: Dataset is considered as NOT biased here.

Recall:



* Recall calculates how many from the Actual Positives our model captured through labeling it as True Positive.
* Applying the same understanding as that of Recall, here, **this shall be the model metric when there is a high cost associated with False Negative. For instance, in fraud detection or sick patient detection. If a fraudulent transaction (Actual Positive) is predicted as non-fraudulent (Predicted Negative), the consequence can be very bad for the bank.** *[ HERE : TP is Fraudulent Transaction and TN = Non-Fraudulent Transaction].*
* Example: Similarly, in sick patient detection. If a sick patient (Actual Positive) goes through the test and predicted as not sick (Predicted Negative). The cost associated with False Negative will be extremely high if the sickness is contagious.

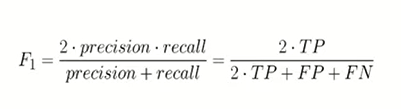
Thus, from their definitions, we can say:



**MIDDLE GROUND:**

* Precision and Recall complement each othe
* If, Prec = 100% , then Recall is 0%, and, if Recall is 100% then Prec
* Both Precision and Recall are capable of cheating the system if the Training Data has data in a certain way. Thus, we combine both the Metrics to get the best of both worlds and avoid cheating. This is where we use F1 score.

F1 score:



* It is the Harmonic mean of both Precision and Recall. F1 Score is needed when you want to seek a balance between Precision and Recall.
* **What is the difference between F1 Score and Accuracy then? EXPLAINATION:**

We have previously seen that accuracy can be largely contributed by a large number of True Negatives which in most business circumstances, we do not focus on much **whereas False Negative and False Positive usually has business costs (tangible & intangible),** thus, **F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall AND there is an uneven class distribution (large number of Actual Negatives) or Our Problem statement is such that Negatives are of more importance than positives (as discussed in Precision and Recall).**

* If the Model is cheating , then, it will have either Precision close to 100 % and Recall close to 0% , OR, it will have Precision close to 0 % and Recall close to 100%. In either of the cases, we will get F1 score = 0 or close to 0. -🡪This shows model is flawed.

**CONCLUSION:**

Thus, the F-1 Score metric is preferable when:

* We have imbalanced class distribution.
* We’re looking for a balanced measure between precision and recall (Type I and Type II errors)
* F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, **but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false** **negatives are very different, it’s better to look at both Precision and Recall.**

