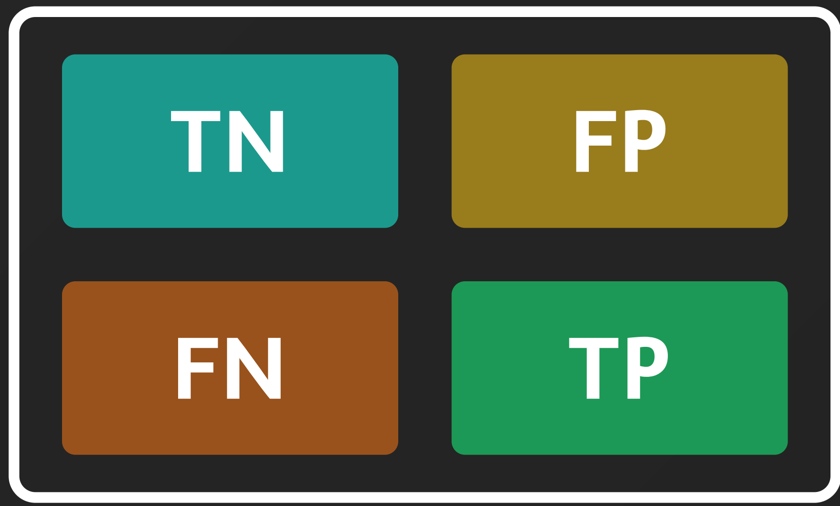
Confusion Matrix



and Classification Evaluation Metrics

Trust is a must when a decision-maker's judgment is critical. To give such trust, we summarize all possible decision outcomes into four categories: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) to serve an outlook of how confused their judgments are, namely, the confusion matrix. From the confusion matrix, we calculate different metrics to measure the quality of the outcomes. These measures influence how much trust we should give to the decision-maker (classifier) in particular use cases. This document will discuss the most common classification evaluation metrics, their focuses, and their limitations in a straightforward and informative manner.

Author: Yousef Alghofaili

Confusion Matrix

Predicted Value

Actual Value	Negative	Positive
Negative	TN	FP
Positive	FN	TP

Guess	→	Fact	
This car is NOT red	→	This car is NOT red	TN
This car is red	→	This car is red	TP
This car is NOT red	→	This car is red	FN
This car is red	→	This car is NOT red	FP

Positive Predictive Value

Precision

$$\frac{TP}{TP + FP}$$

Sensitivity or True Positive Rate

Recall

$$\frac{TP}{TP + FN}$$

True Negative Rate

Specificity

$$\frac{TN}{TN + FP}$$

(Negative Predictive Value)

NPV

$$\frac{TN}{TN + FN}$$

Accuracy

$$\frac{TP + TN}{FP + TP + TN + FN}$$

F1-Score

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Balanced Accuracy

$$\frac{\text{Sensitivity} + \text{Specificity}}{2}$$

Matthews Correlation Coefficient (MCC)

$$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Author: Yousef Alghofaili

Precision & Recall

Common Goal

We use both metrics when actual negatives are less relevant. For example, googling "Confusion Matrix" will have trillions of unrelated (negative) web pages, such as the "Best Pizza Recipe!" web page. Accounting for whether we have correctly predicted the latter webpage and alike as negative is impractical.

Precision Goal

This customer loves steak! — No, this customer is vegan. FP

Bad Product Recommendation → Less Conversion → Decrease in Sales

Recall Goal

The product has no defects. — A customer called... He's angry. FN

Bad Defect Detector → Bad Quality → Customer Dissatisfaction

Author: Yousef Alghofaili

Specificity & NPV

Common Goal

We use both metrics when actual positives are less relevant. In essence, we aim to rule out a phenomenon. For example, we want to know how many healthy people (no disease detected) there are in a population. Or, how many trustworthy websites (not fraudulent) is someone visiting.

Specificity Goal

This person is a criminal. — They were detained for no reason. **FP**

Bad Predictive Policing → Injustice

NPV Goal

They don't have cancer. — No, they should be treated! **FN**

Bad Diagnosis → No Treatment → Consequences

Author: Yousef Alghofaili

Hacks

Previously explained evaluation metrics, among many, are granular, as they focus on one angle of prediction quality which can mislead us into thinking that a predictive model is highly accurate. Generally, these metrics are not used solely. Let us see how easy it is to manipulate the aforementioned metrics.

Precision Hacking

Precision is the ratio of correctly classified positive samples to the total number of positive predictions. Hence the name, Positive Predictive Value.

Dataset

50 Positive Samples	50 Negative Samples

Get at least one positive sample correctly

Predict almost all samples as negative

Positives	**Negatives**
FN≈49 / TP≈1	**TN≈50**

$$\frac{TP \approx 1}{TP \approx 1 + FP \approx 0} \approx 100\%$$

Predicting positive samples with a high confidence threshold would potentially bring out this case. In addition, when positive samples are disproportionately higher than negatives, false positives will probabilistically be rarer. Hence, precision will tend to be high.

Author: Yousef Alghofaili

Recall Hacking

Recall is the ratio of correctly classified positive samples to the total number of actual positive samples. Hence the name, True Positive Rate.

Dataset

50 Positive Samples	50 Negative Samples

Predict all samples as positive

Positives	**Negatives**
TP=50	**FP=50**

$$\frac{TP=50}{TP=50 + FN=0} = 100\%$$

Similar to precision, when positive samples are disproportionately higher, the classifier would generally be biased towards positive class predictions to reduce the number of mistakes.

Specificity Hacking

Specificity is the ratio of correctly classified negative samples to the total number of actual negative samples. Hence the name, True Negative Rate.

Dataset

50 Positive Samples	50 Negative Samples

Predict all samples as negative

Positives	Negatives
FN=50	**TN=50**

$$\frac{TN=50}{TN=50 + FP=0} = 100\%$$

Contrary to Recall (Sensitivity), Specificity focuses on the negative class. Hence, we face this problem when negative samples are disproportionately higher. Notice how the Balanced Accuracy metric intuitively solves this issue in subsequent pages.

Author: Yousef Alghofaili

NPV Hacking

Negative Predictive Value is the ratio of correctly classified negative samples to the total number of negative predictions. Hence the name.

Dataset

50 Positive Samples	50 Negative Samples

Predict almost all samples as positive

Get at least one negative sample correctly

Positives

TP≈50

Negatives

FP≈49

TN≈1

$$\frac{TN\approx 1}{TN\approx 1 + FN\approx 0} \approx 100\%$$

Predicting negative samples with a high confidence threshold has this case as a consequence. Also, when negative samples are disproportionately higher, false negatives will probabilistically be rarer. Thus, NPV will tend to be high.

Author: Yousef Alghofaili

Comprehensive Metrics

As we have seen above, some metrics can misinform us about the actual performance of a classifier. However, there are other metrics that include more information about the performance. Nevertheless, all metrics can be "hacked" in one way or another. Hence, we commonly report multiple metrics to observe multiple viewpoints of the model's performance.

Accuracy

Accuracy treats all error types (false positives and false negatives) as equal. However, equal is not always preferred.

$$\frac{TP + TN}{FP + TP + TN + FN}$$

Accuracy Paradox

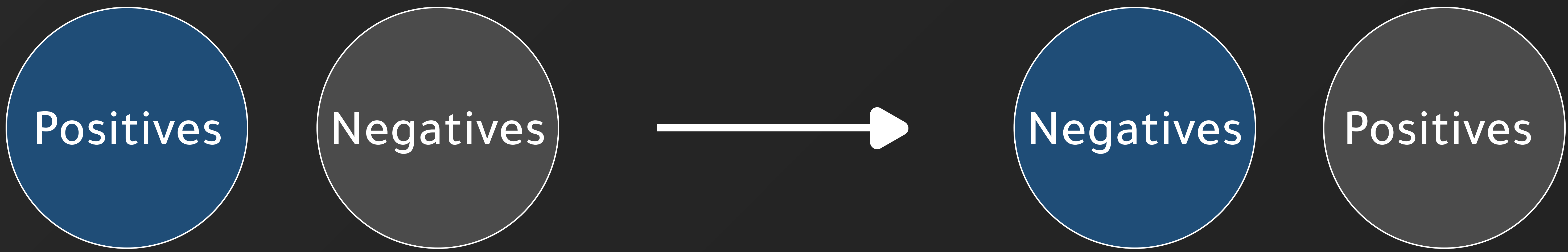
Since accuracy assigns equal cost to all error types, having significantly more positive samples than negatives will make accuracy biased towards the larger class. In fact, the Accuracy Paradox is a direct "hack" against the metric. Assume you have 99 samples of class 1 and 1 sample of class 0. If your classifier predicts everything as class 1, it will get an accuracy of 99%.

F1-Score

F1-Score will combine precision and recall in a way that is sensitive to a decrease in any of the two (Harmonic Mean). Note that the issues mentioned below do apply to F_β score in general.

$$2 \times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$$

Asymmetric Measure



F1-Score is asymmetric to the choice of which class is negative or positive. Changing the positive class into the negative one will not produce a similar score in most cases.

True Negatives Absence

F1-Score does not account for true negatives. For example, correctly diagnosing a patient with no disease (true negative) has no impact on the F1-Score.

Extra: Know more about F_β Score on Page 18
Author: Yousef Alghofaili

Balanced Accuracy

Balanced Accuracy accounts for the positive and negative classes independently using Sensitivity and Specificity, respectively. The metric partially solves the Accuracy paradox through independent calculation of error types and solves the true negative absence problem in F_β-Score through the inclusion of Specificity.

$$\frac{\textbf{Sensitivity} + \textbf{Specificity}}{2}$$

Relative Differences in error types

TN 9	FP 1
FN 1000	TP 9000

TN 9000	FP 1000
FN 1	TP 9

Balanced Accuracy is commonly robust against imbalanced datasets, but that does not apply to the above-illustrated cases. Both models perform poorly at predicting one of the two (positive P or negative N) classes, therefore unreliable at one. Yet, Balanced Accuracy is 90%, which is misleading.

Matthews Correlation Coefficient (MCC)

MCC calculates the correlation between the actual and predicted labels, which produces a number between -1 and 1. Hence, it will only produce a good score if the model is accurate in all confusion matrix components. MCC is the most robust metric against imbalanced dataset issues or random classifications.

$$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

MCC faces an issue of it being undefined whenever a full row or a column in a confusion matrix is zeros. However, the issue is outside the scope of this document. Note that this is solved by simply substituting zeros with an arbitrarily small value.

Conclusion

We have gone through all confusion matrix components, discussed some of the most popular metrics, how easy it is for them to be "hacked", alternatives to overcome these problems through more generalized metrics, and each one's limitations. The key takeaways are:

- Recognize the hacks against granular metrics as you might fall into one unintentionally. Although these metrics are not solely used in reporting, they are heavily used in development settings to debug a classifier's behavior.

- Know the limitations of popular classification evaluations metrics used in reporting so that you become equipped with enough acumen to decide whether you have obtained the optimal classifier or not.

- Never get persuaded by the phrase "THE BEST" in the context of machine learning, especially evaluation metrics. Every metric approached in this document (including MCC) is the best metric only when it best fits the project's objective.

References

- Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, 14(1), 1-22.
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1), 1-13.
- Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData mining*, 10(1), 1-17.
- Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2), 1.
- Jurman, G., Riccadonna, S., & Furlanello, C. (2012). A comparison of MCC and CEN error measures in multi-class prediction.
- Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing education in anaesthesia critical care & pain*, 8(6), 221-223.
- Hull, D. (1993, July). Using statistical testing in the evaluation of retrieval experiments. *In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 329-338).

Author

Yousef Alghofaili

AI solutions architect and researcher who has studied at KFUPM and Georgia Institute of Technology. He worked with multiple research groups from KAUST, KSU, and KFUPM. He has also built and managed noura.ai data science R&D team as the AI Director. He is an official author at Towards Data Science Publication and developer of KMeansInterp Algorithm.

Reviewer

Dr. Motaz Alfarraj

Assistant Professor at KFUPM, and the Acting Director of SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI). He has received his Bachelor's degree from KFUPM, and earned his Master's and Ph.D. degrees in Electrical Engineering, Digital Image Processing and Computer Vision from Georgia Institute of Technology. He has contributed to ML research as an author of many research papers and won many awards in his field.

THANK YOU!

For any feedback, issues, or inquiries, contact yousefalghofaili@gmail.com

F_β Score

F_β Score is the generalized form of F1 Score ($F_{\beta=1}$ Score) where the difference lies within the variability of the β Factor. The β Factor skews the final score into favoring recall β times over precision, enabling us to weigh the risk of having false negatives (Type II Errors) and false positives (Type I Errors) differently.

$$\frac{(1+\beta^2) \times \textbf{Precision} \times \textbf{Recall}}{(\beta^2 \times \textbf{Precision}) + \textbf{Recall}}$$

β	Precision is β times **Less** important than Recall	β
β = 1	Balanced F1-Score	**β = 1**
β	Precision is β times **More** important than Recall	β

F_β Score has been originally developed to evaluate Information Retrieval (IR) systems such as Google Search Engine. When you search for a webpage, but it does not appear, you are experiencing the engine's low Recall. When the results you see are completely irrelevant, you are experiencing its low Precision. Hence, search engines play with the β Factor to optimize User Experience by favoring one of the two experiences you have had over another.