MYDATAMODELS

AI-Driven Analytics for Every Professional

Data Science Metrics

www.mydatamodels.com

WHITE PAPER

AI claims to improve our lives

It claims it loud and clear in figures: by next year, 80% of emerging technologies will have AI foundations, AI saves Netflix $1 billion each year with 75% of what users watch coming from **AI recommendations**.

To build **AI-based systems** that users can justifiably trust, one needs to understand how accurate machine learning technologies are. A central problem in this context is that **ML predictions' quality** is challenging to measure. Yet evaluations, comparisons, and improvements of ML models require quantifiable measures. That's when **ML** and **AI metrics** come into play.

We can use a variety of **AI metrics** to measure the actual performances of algorithms. We, at **MyDataModels**, would like to share our understanding of these metrics with you throughout this white paper. We hope it can help better understand these metrics and are happy to hear your feedback.

WHY SHOULD WE USE MULTIPLE METRICS?

Both for classification and regression algorithms, TADA proposes **various metrics defined by the Artificial Intelligence scientific community**.

Each metric is a means to evaluate the model's performance from a different aspect.

This slide set presents and explains these various metrics and provides examples of their use.

MYDATAMODELS

Classification Metrics

ACCURACY PARADOX:

Medical Use Case : A model delivers a diagnosis (i.e., prediction) of a rare disease based on a set of symptoms.

- Out of 1000 people tested, 995 are healthy, and five are sick.
- The algorithm misclassifies two healthy people as ill and four sick people as healthy.

PREDICTED VALUE

		Positive	Negative	
ACTUAL VALUE	Positive	1	4	(Sick people)
	Negative	2	993	(Healthy people)
		(Number of people estimated sick by the model)	(Number of people estimated healthy by the model)	

In such a context, the model's Accuracy is computed:

$$\textit{Accuracy} = \frac{1 + 993}{5 + 995} = 0{,}994$$

It looks like the model's Precision is excellent, while the model was wrong four times out of five at identifying a sick person.
Other metrics are available to spot these errors:

$$\textit{Sensitivity} = \frac{1}{1 + 4} = 0{,}20$$

Sensitivity provides a way to understand a model's weakness: the model identifies only 20% of sick patients as actual sick people.

ACCURACY: THE RATIO OF CORRECTLY CLASSIFIED CASES

Medical Use Case:

		PREDICTED VALUE		
		Positive	Negative	
ACTUAL VALUE	Positive	1	4	(Sick people)
	Negative	2	993	(Healthy people)
		(Number of people estimated sick by the model)	(Number of people estimated healthy by the model)	

Pluses:

- This metric is **easy** to compute, **understandable** and **intuitive**.
- **Excellent metric** when the class to predict is well **balanced**.
- Significant metric when **False Negatives** are not more impacting than **False Positives**.

Minuses:

- Might lead to confusion and errors when the classes are **significantly unbalanced**.
- Does not account for the fact that wrongfully classifying positives might have more significant consequences than **wrongfully classifying negatives** (or vice versa).

Accuracy:

$$Acc = \frac{TP + TN}{P + N} = \frac{1 + 993}{5 + 995} = 0{,}994$$

This model seems very accurate.

SENSITIVITY (RECALL): RATIO OF TRUE POSITIVES

Medical Use Case:

		PREDICTED VALUE		
		Positive	Negative	
ACTUAL VALUE	Positive	1	4	(Sick people)
	Negative	2	993	(Healthy people)
		(Number of people estimated sick by the model)	(Number of people estimated healthy by the model)	

Pluses:

- This metric is also called **"Ratio of True Positives."**
- This **metric is understandable** and can detect **imbalances** in the classifications' quality.
- Key metric to **disfavor False Negatives.**

Minuses:

- Less intuitive than **Accuracy.**
- It **is an incomplete representation** of the model's quality. It cannot by itself, represent the complete performance of the model.

Sensitivity:

$$Sensitivity = \frac{TP}{TP + FN} = \frac{1}{1 + 4} = 0{,}20$$

This model performs poorly to classify positive cases.

SPECIFICITY: RATIO OF TRUE NEGATIVES

Medical Use Case:

		PREDICTED VALUE		
		Positive	Negative	
ACTUAL VALUE	Positive	1	4	(Sick people)
	Negative	2	993	(Healthy people)
		(Number of people estimated sick by the model)	(Number of people estimated healthy by the model)	

Pluses:

- This metric is also named **Ratio of True Negatives**.
- This metric is readable and provides a way to detect **imbalances** in the classification quality.
- Significant metric to **emphasize False Negatives.**

Minuses:

- Less intuitive than **Accuracy**.
- It is an **incomplete representation** of the model's quality. It cannot by itself, represent the complete performance of the model.

Specificity:

$$Specificity = \frac{TN}{FP + TN} = \frac{993}{2 + 993} = 0{,}998$$

This model performs well to classify negative cases.

PRECISION: RATIO OF POSITIVES WELL CLASSIFIED

Medical Use Case:

PREDICTED VALUE

		Positive	Negative	
ACTUAL VALUE	Positive	1	4	(Sick people)
	Negative	2	993	(Healthy people)
		(Number of people estimated sick by the model)	(Number of people estimated healthy by the model)	

Pluses:

- This metric enables us to calculate the **ratio of cases classified as positives**, which are really positives. A substantial value for this metric means that the model performs well on positive cases.

Minuses:

- Less intuitive than the **Accuracy**.
- It does not provide information on the n**egative cases classification quality**.

Precision:

$$Precision = \frac{TP}{TP + FP} = \frac{1}{1 + 2} = 0{,}333$$

Note: a model can have an outstanding Precision and a low Sensitivity, which is not accurate enough for positives cases. But when it classifies a case as positive, it is rarely wrong.

This model presents a high rate of False Positives.

F1 SCORE: AVERAGE OF RECALL AND PRECISION

Medical Use Case:

		PREDICTED VALUE		
		Positive	Negative	
ACTUAL VALUE	Positive	1	4	(Sick people)
	Negative	2	993	(Healthy people)
		(Number of people estimated sick by the model)	(Number of people estimated healthy by the model)	

Pluses:

- This metric is a combination of **Recall and Precision**.
- This metric is hard on both **False Negatives** and **False Positives**.

Minuses:

- Not intuitive.
- It does not emphasize models that wrongly classify **negative cases**.

F1 Score:

$$F1 = \frac{2TP}{(2TP + FP + FN)} = \frac{2 * 1}{(2 * 1 + 2 + 4)} = 0{,}25$$

This model performs poorly to classify positive cases.

MCC: MATTHEWS CORRELATION COEFFICIENT

Medical Use Case:

		PREDICTED VALUE		
		Positive	Negative	
ACTUAL VALUE	Positive	1	4	(Sick people)
	Negative	2	993	(Healthy people)
		(Number of people estimated sick by the model)	(Number of people estimated healthy by the model)	

Pluses:

- This metric accentuates **imbalances**, whether on positives or negatives cases.

Minuses:

- Not very intuitive.
- It puts a strong emphasis on classification errors.

Precision:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} = 0{,}255$$

This model is strongly unbalanced.

EXAMPLE 1: Medical Use Case

		PREDICTED VALUE		
		Positive	Negative	
ACTUAL VALUE	Positive	1	4	(Sick people)
	Negative	2	993	(Healthy people)
		(Number of people estimated sick by the model)	(Number of people estimated healthy by the model)	

$$Accuracy = \frac{1 + 993}{5 + 995} = 0{,}994$$

$$Sensitivity = \frac{1}{1 + 4} = 0{,}20$$

$$Specificity = \frac{993}{2 + 993} = 0{,}998$$

$$Precision = \frac{1}{1 + 2} = 0{,}333$$

$$F1 = \frac{2 * 1}{(2 * 1 + 2 + 4)} = 0{,}25$$

$$MCC = 0{,}255$$

In this specific case, the **Accuracy** wrongfully suggests that the model is excellent. The medical testing goal is to make sure that as many sick people as possible are detected. Other metrics help in identifying this asymmetry.

EXAMPLE 2: Medical Use Case

		PREDICTED VALUE		
		Positive	Negative	
ACTUAL VALUE	Positive	10	0	(Sick people)
	Negative	20	970	(Healthy people)
		(Number of people estimated sick by the model)	(Number of people estimated healthy by the model)	

$$\textit{Accuracy} = \frac{10 + 970}{20 + 970} = 0{,}98$$

$$\textit{Sensitivity} = \frac{10}{10 + 0} = 1$$

$$\textit{Specificity} = \frac{970}{20 + 970} = 0{,}9798$$

$$\textit{Precision} = \frac{10}{10 + 20} = 0{,}333$$

$$F1 = \frac{2 * 10}{(2 * 10 + 20 + 0)} = 0{,}50$$

$$\textit{MCC} = 0{,}5717$$

In this configuration, the model does an excellent job since it does not forget any ill person. It erroneously classifies healthy people as unhealthy. It is not a significant flaw when the model is used to select which people will undergo further medical testing (good **Sensitivity**, low **Precision).** However, the model's **Precision** is low since only one-third of positive cases predicted are positives.

EXAMPLE 3: Medical Use Case

		PREDICTED VALUE		
		Positive	Negative	
ACTUAL VALUE	Positive	585	20	(Sick people)
	Negative	25	370	(Healthy people)
		(Number of people estimated sick by the model)	(Number of people estimated healthy by the model)	

$$Accuracy = \frac{585 + 370}{610 + 390} = 0{,}955$$

$$Sensitivity = \frac{585}{585 + 20} = 0{,}967$$

$$Specificity = \frac{370}{25 + 370} = 0{,}937$$

$$Precision = \frac{585}{585 + 25} = 0{,}959$$

$$F1 = \frac{2 * 585}{(2 * 585 + 25 + 20)} = 0{,}963$$

$MCC = 0{,}906$

In this case, both classes are well balanced. All metrics are in the same range. **Accuracy** is a reliable metric.
This balanced case represents the efficiency of the generic metrics commonly used in the context of AI. In contrast, the initial unusual use case illustrates the Accuracy paradox and the weakness of using a single metric.

PERFECTLY WELL-BALANCED CASE:

Medical Use Case:

PREDICTED VALUE

		Positive	Negative	
ACTUAL VALUE	Positive	45	5	(Sick people)
	Negative	5	45	(Healthy people)
		(Number of people estimated sick by the model)	(Number of people estimated healthy by the model)	

$$\textit{Accuracy} = \frac{45 + 45}{50 + 50} = 0{,}9$$

$$\textit{Sensitivity} = \frac{45}{45 + 5} = 0{,}9$$

$$\textit{Specificity} = \frac{45}{5 + 45} = 0{,}9$$

$$\textit{Precision} = \frac{45}{45 + 5} = 0{,}9$$

$$F1 = \frac{2 * 45}{(2 * 45 + 5 + 5)} = 0{,}9$$

$$\textit{MCC} = 0{,}8$$

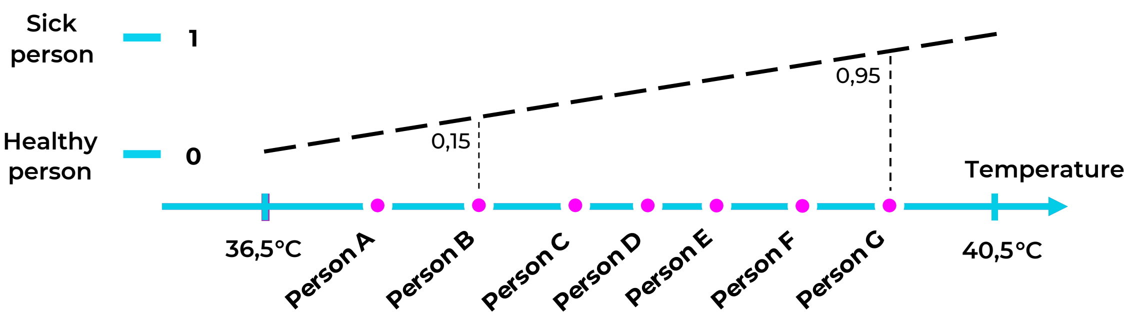
In this well-balanced case, all the metrics are equally interesting to use for interpreting the model’s quality.

ROC CURVE (RECEIVER OPERATOR CHARACTERISTIC) AND AUC:

Medical Use Case: **The ROC curve** is yet another tool for assessing the **performance** of a **binary classifie**r. It is used to calculate the **AUC** (**Area Under Curve**, the geometric area under the ROC curve) proposed by TADA.

Going back to the medical example used so far, it is possible to imagine that the model uses a single measurement (the patient's **body temperature**) to diagnose. A binary classification model often uses several variables; however, we chose to represent only one for simplicity.

To get a diagnosis, **TADA** computes each patient's probability of belonging to one category (healthy or sick).



In this case, person B has a 15% probability of being ill, whereas person G has a 95% probability of being sick.

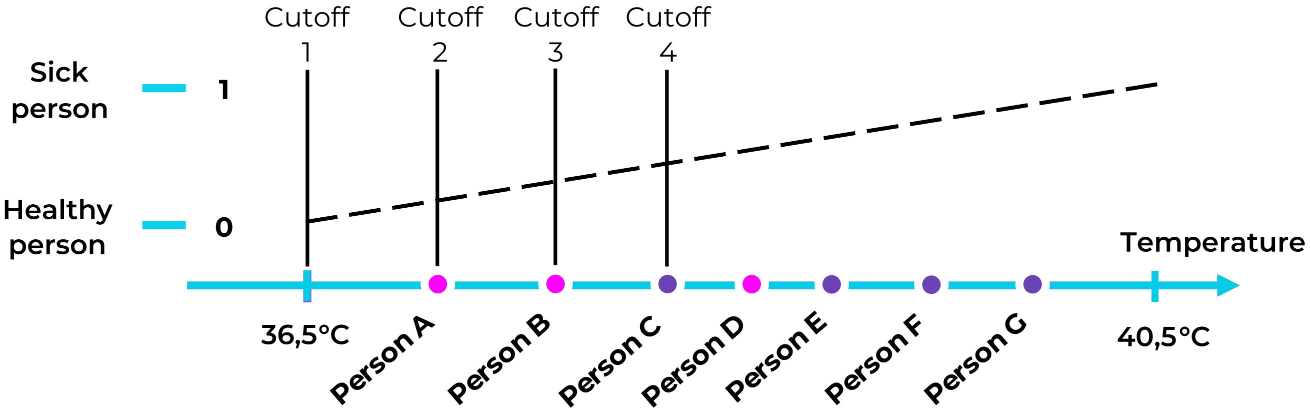
Each prediction allows us to classify a person based on a **cutoff**. The **cutoff** can take any value between 0 and 1. It gives a means to **calibrate the model** by varying the threshold and generate less False Positives at the cost of more False Negatives or vice-versa.

This **cutoff** concept is what allows for the generation and the creation of the ROC and the AUC.

Please note that the example curve has an arbitrary shape. Actual curves are often more complicated than a linear curve.

ROC CURVE (RECEIVER OPERATOR CHARACTERISTIC) AND AUC:

Let's assume that we know persons A, B, and D are **healthy**, and persons C, E, F, and G are **sick**.



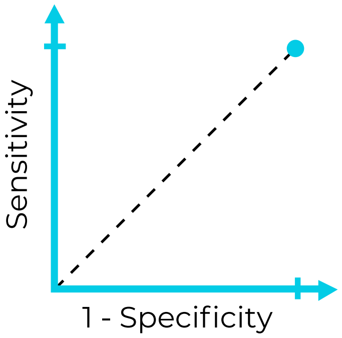
The ROC curve is generated by **varying the cutoff** and by including, with each new iteration, one more patient. For each cutoff, we draw the Sensitivity versus the **Specificity** (more precisely, we use 1 - Specificity).

With a zero cutoff, each person is classified as sick. We have:

4 True Positives -> **Sensitivity** = 1

3 False Positives -> 1 – **Specificity** = 1

We can draw the first point of the curve.

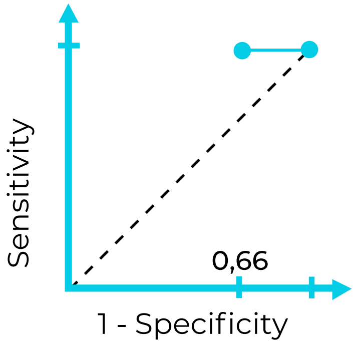


With a cutoff of 2, all persons from B to G are classified as sick.

4 True Positives -> **Sensitivity** = 1

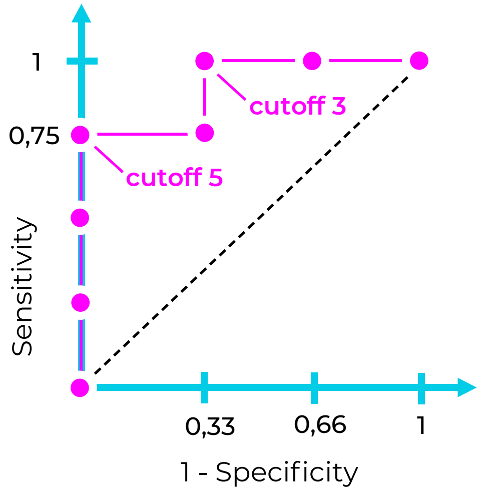
2 False Positives -> 1 - **Specificity** = 0.66

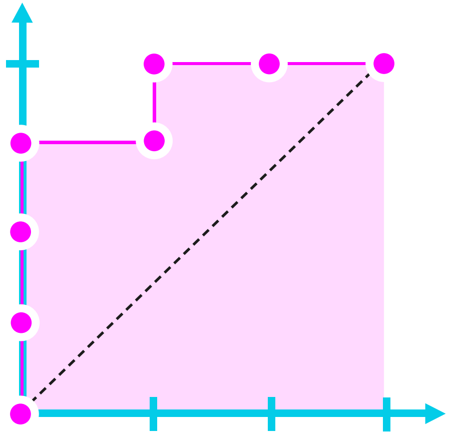
The second point can be drawn on the curve.



ROC CURVE (RECEIVER OPERATOR CHARACTERISTIC) AND AUC:

We **reiterate this process** until the whole curve is drawn (until the last **cutoff**, which puts all the subjects in the healthy category).





All these points make up the **ROC curve**. It can be used to identify the **right cutoff**.

There is no perfect cutoff. It is a tradeoff since some allow for fewer **False Positives** or **False Negatives**. The 'right' cutoff does not exist and should be selected based on one's needs.

In our example, **cutoff 3** and **5** are the most interesting choices.

Please note that the dotted line displays the efficiency of a **random classification**. A good model will always be above this line.

In a nutshell, the **ROC curve** and the **AUC** (area under this curve) are performant metrics to estimate the model's quality.

The Area Under the Curve is a figure between 0 and 1. It allows for easy model performance comparison instead of visually comparing two curves.

Please note that a random classifier will generate an **AUC** of 0.5. Therefore, we expect to always get an AUC above 0.5 for our models.

TAKEAWAYS

A single metric is usually not sufficient to give a **complete picture of a model's performance**.

The expectations regarding the model's performance should be **clarified ahead of analyzing the actual metrics and results**.
Is the goal to be very accurate with positive cases? Or is it to be as precise as possible in a comprehensive way?

If the classes are well balanced, and the False Negatives are just as significant as the False Positives, then the **Accuracy is the right criteria**.

A single high value in a metric does not necessarily mean that the model is well-performing; neither does a low figure mean that the model's performance is weak.

MYDATAMODELS

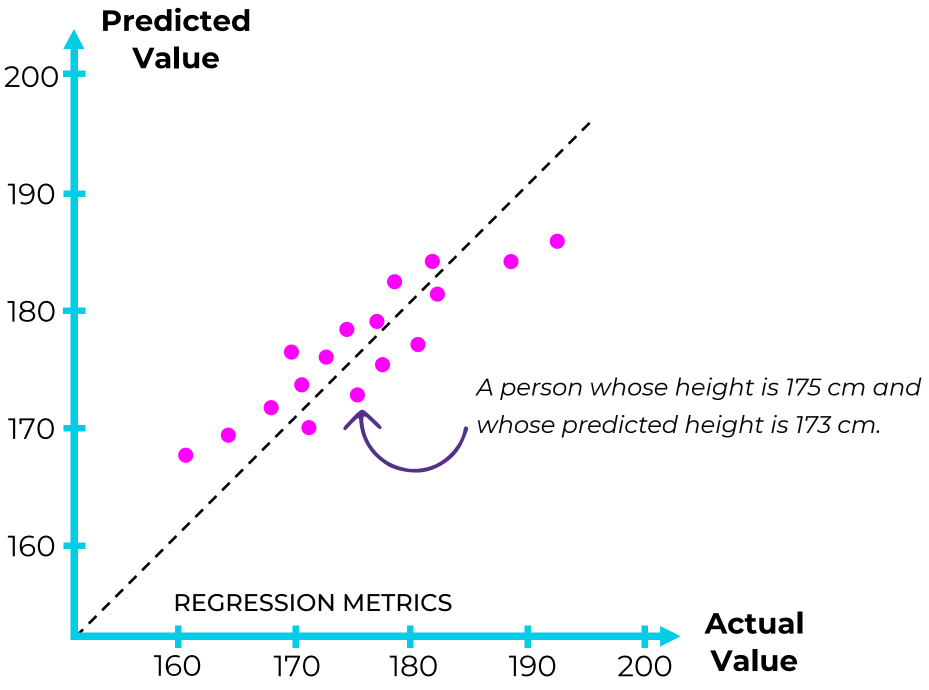
Regression Metrics

©2020 MyDataModels, Inc. | www.mydatamodels.com | contact@mydatamodels.com | @mydatamodels

GENERAL CASE

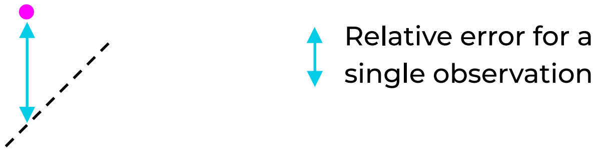
Use case: the model predicts a person's height based on known data (age, weight, sex, etc.). Then the model's predictions and actual observations are compared:

Any point positioned precisely on this line indicates that the model's prediction was perfect since the predicted value and the real value are identical.



A person whose height is 175 cm and whose predicted height is 173 cm.

Predictions are continuous-valued variables which match more or less the real values. It is possible to measure the error between the predicted value and the actual value as the distance between **the point and the line**.



Relative error for a single observation

Summing all the errors might not be the best approach since some errors might compensate each other. Hence, computing several **metrics** and analyzing them gives a better understanding of the model's strengths and flaws.

Please note that the model presented here as an example is weak for extreme values.
It overestimates small sizes and underestimates big sizes.

MEAN ABSOLUTE ERROR (MAE) / MEAN ABSOLUTE PERCENTAGE ERROR (MAPE):

Use case: we estimate a person's height based on known data (age, weight, sex, etc.). Here the results are displayed in a table:

Actual Value	Predicted Value
167	170
168	169
170	173
172	171
175	176
178	175
179	181
181	182
185	183
187	184

→

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right|$$

With:

$$\begin{cases} Y_i & \text{the ith actual value} \\ \hat{Y}_i & \text{the ith predicted value} \end{cases}$$

It is the mean of the absolutes of the deviations. The larger the deviation, the greater the error.

Please note that the MAPE uses the same principle but to the ratio of the actual value:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right|$$

→

In our case :

MAE = 2,00 cm

Unlike most metrics, the MAE cannot serve as a comparison across cases.
For instance:
Prediction of a **person's heigh**t:

MAE ≈ 2 cm

Prediction of a **car's price**:

MAE ≈ 500 €

In our case:

MAPE = 1,13 %

MAPE, however, can serve as a comparison between cases.

ROOT MEAN SQUARE ERROR (RMSE):

Use case: we estimate a person's height based on known data (age, weight, sex, etc.). Here the results are displayed in a table:

Actual Value	Predicted Value
167	170
168	169
170	173
172	171
175	176
178	175
179	181
181	182
185	183
187	184

→

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

With:

$$\begin{cases} Y_i & \text{the ith actual value} \\ \hat{Y}_i & \text{the ith predicted value} \end{cases}$$

The RMSE eliminates the error's sign by computing the error's square value, instead of its absolute value in MAE.
As a consequence, this metric accentuates higher errors.

→

In our case:

RMSE = 2,19 cm

The RMSE is use case dependant, just as the MAE. The same unit value applies to the RMSE and the predicted value, making it easier for interpretation. Therefore, there is no "good" or "bad" RMSE in general.

Please note that TADA also provides the Residual Standard Deviation, which is very similar to the RMSE.

$$RSD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

MAXIMUM ERROR:

Use case: we estimate a person's height based on known data (age, weight, sex, etc.). Here the results are displayed in a table:

Actual Value	Predicted Value
167	170
168	169
170	173
172	171
175	176
178	175
179	181
181	182
185	183
187	184

→

$$Max.error = \max_{i} \left| Y_i - \hat{Y}_i \right|$$

With:

$$\begin{cases} Y_i & \text{the ith actual value} \\ \hat{Y}_i & \text{the ith predicted value} \end{cases}$$

It is the most significant error value in the model. This metric points out when the model is not well suited for outliers.

→

In our case:

Max.error = 3 cm

The point in using this metric is that it indicates very rapidly if the model performs well on **outliers**. For instance, if the **Max Error** is much higher than the **RMSE**, a rationale might be that the model did not learn properly to predict extreme values.

For instance :
Let's assume that we add one more line to the table describing a person whose height is 196 cm when the model predicted 186 cm. The following values result from this change:

RMSE = 3,71 cm

The Max Error is much more significant than the RMSE.

R²: COEFFICIENT OF DETERMINATION

Use case: we estimate a person's height based on known data (age, weight, sex, etc.). Here the results are displayed in a table:

Actual Value	Predicted Value
167	170
168	169
170	173
172	171
175	176
178	175
179	181
181	182
185	183
187	184

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

With:

- Y_i the ith actual value
- $\hat{Y}_i$ the ith predicted value
- $\bar{Y}_i$ the average of the measured values

Please note that the R2 is the ratio between the variance determined by the model and the real values' actual variance.

In our case:

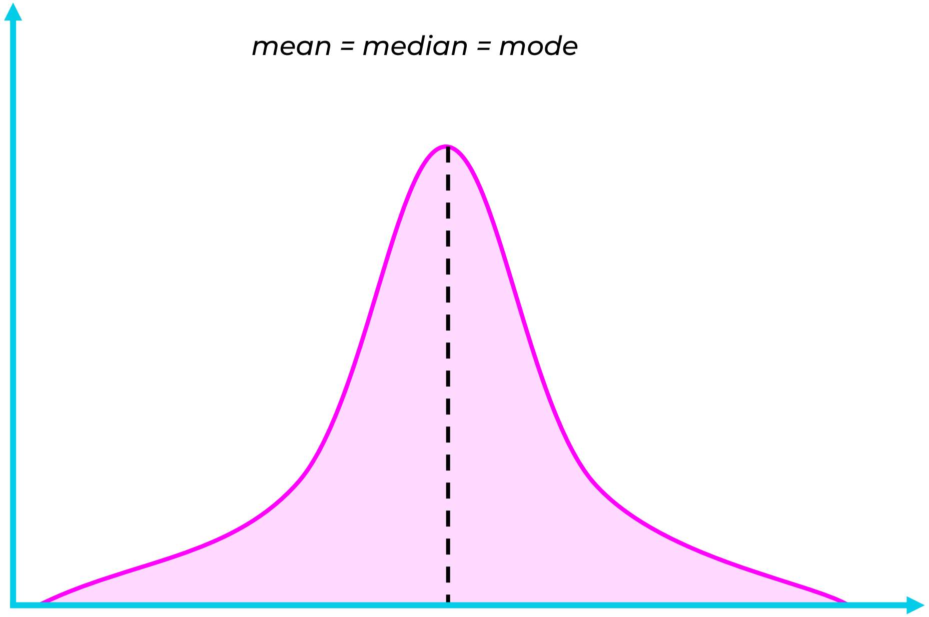
$R^2 = 0,89$

The model accounts for 89% of the observed phenomena. This metric varies from -infinity to 1.

1: the model is entirely accurate.
0: the model is not better than a random one.
Negative value: the model is erroneous.
The expected R2 depends on the use case.
In practice, the goal is to be as close as possible to 1.

EXAMPLE 1: BELL CURVE

Gaussian curve: This very typical case often occurs in biology, for instance.



The density of the variable considered.

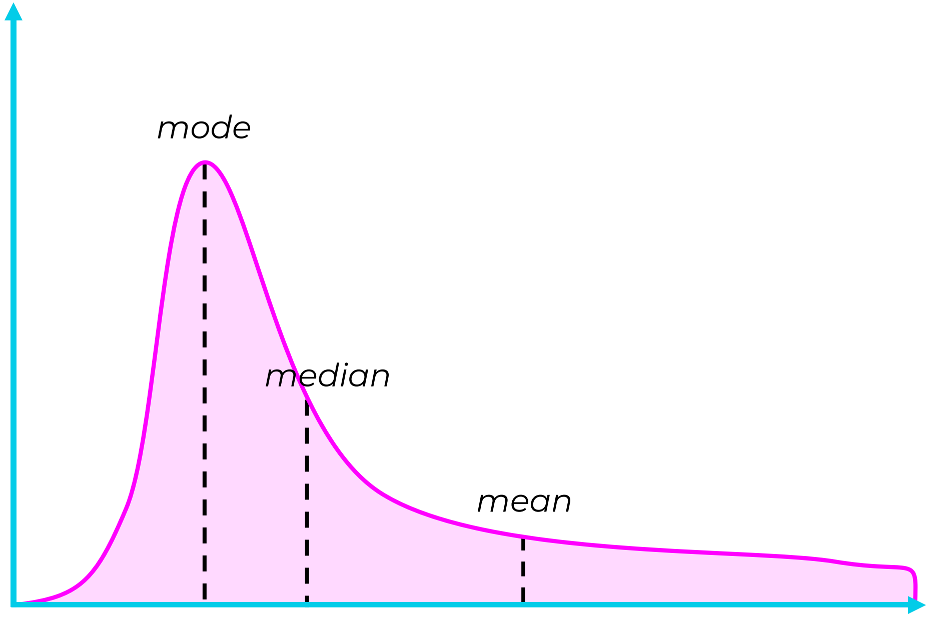
With a Gaussian distribution representing the variable under study, most of the observations are around the mean.

The **outliers** are scarce and quite close to the mean.
Example: we observe the size of a population.
The average size is 175 cm, some people have a height above, but nobody is taller than 2,5 m.

In this case, the **RMSE** is a perfect metric.
We typically expect that the **RMSE** is smaller than the variance of the observations.

EXAMPLE 2: CURVE WITH POSITIVE DEVIATION

Case of an unbalanced curve: this is a very classic case often encountered in economics.



The density of the variable to explain.

Some other classical cases may have a distribution with a **substantial deviation**. They may represent economic or financial variables.

These repartitions have a significant number of **outliers** and **extreme values**.

Examples: housing prices, salaries among a population (or any bounded variable on one side which can take significant values on the other).

In this case, the **RMSE** and the **MAPE** can be biased by a weakness on the **outliers**. It should be taken into consideration.

MYDATAMODELS

Multi-class Classification Metrics

BINARY CASE GENERALIZATION

For a multi-class **classification**, the metrics applied previously (**Accuracy**, **Precision**, **MCC**, **F1**, etc....) can be **generalized** and used in the same manner.

PREDICTED VALUE

		Healthy	Cold	Flu
ACTUAL VALUE	Healthy	7	6	2
	Cold	2	32	5
	Flu	1	3	16

Let's consider once more the example of diagnosing a disease based on a set of symptoms. Still, this time, instead of diagnosing a single illness, the model can identify different diseases.

One reads this multi-class **matrix** the same way as in the binary case.

The **diagonal** represents the persons for which the diagnose of the model was **correct**. The others **are False Negatives** and **False Positives**, depending on the reference class.

Some metrics are computed on the **global** matrix. For the others, the "**sub**" metrics are calculated separately for each class with regards to the others, followed by a **weighted average**.

ACCURACY: PROPORTION OF CORRECTLY CLASSIFIED VALUES

Medical Use Case:

		PREDICTED VALUE		
		Healthy	Cold	Flu
ACTUAL VALUE	Healthy	7	6	2
	Cold	2	32	5
	Flu	1	3	16

$$ACC = \frac{7 + 32 + 16}{(\ 7 + 6 + 2) + (2 + 32 + 5) + (1 + 3 + 16\ )} = 0{,}74$$

The **Accuracy** is a global calculation. It represents the proportion of well-classified individuals in comparison to the total.

The Accuracy has the same strengths and weaknesses in multi-class as in binary classification.

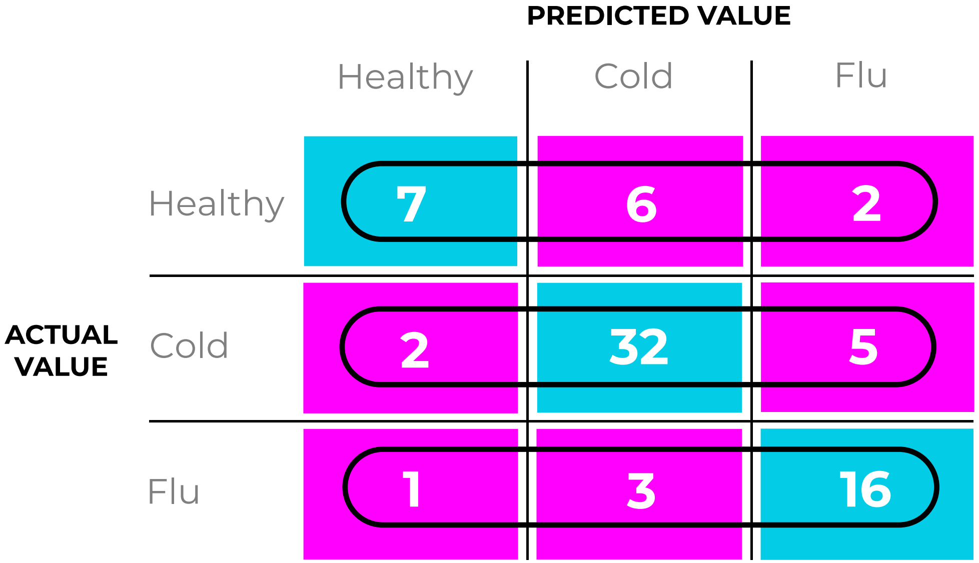
- It is easy to understand, interpret and compute.
- It is very efficient in well-balanced contexts.
- In imbalanced contexts, it becomes misleading.

Accuracy:

$$Acc = \frac{\sum \textit{well classified}}{\textit{Total}}$$

MACRO-RECALL: RATIO OF TRUE POSITIVES

Medical Use Case:



PREDICTED VALUE

ACTUAL VALUE

	Healthy	Cold	Flu
Healthy	7	6	2
Cold	2	32	5
Flu	1	3	16

$$Recall = \frac{1}{3}\left(\frac{7}{7+6+2} + \frac{32}{2+32+5} + \frac{16}{1+3+16}\right) = 0{,}70$$

We calculate the partial **Macro-Recall** the same way we do for binary classification, on a per-class basis (one class against all the others). And the overall **Macro-Recall** is the average of the three partial Macro-Recalls.

Please note that there are other means of calculating the Recall in a multi-class configuration. For instance, a weighted average is also a possibility.

The other similar metrics (**MCC**, **Macro-Precision** and, **F1**) are also calculated using the same approach and are interpreted in the same way as their binary counterparts.

Recall:

$$Recall_{tot} = \frac{1}{\textit{number of classes}} \sum Recall_{\,class}$$

KAPPA: LEVEL OF CONSISTENCY BETWEEN THE MODEL AND REALITY

Medical Use Case:

		PREDICTED VALUE			
		Healthy	Cold	Flu	
ACTUAL VALUE	Healthy	7	6	2	15
	Cold	2	32	5	39
	Flu	1	3	16	20
		10	41	23	74

The **Kappa** has the peculiarity of taking into consideration the predictions correct by **chance.** Therefore, this random part of the predictions has to be calculated.

prob actual healthy $\frac{15}{74}$

prob predicted healthy $\frac{10}{74}$

We deduce from this the probability that the prediction and reality match by chance:

prob(healthy) = prob_actual(healthy)*prob_predicted(healthy) = 0,027

Computing in the same way for cold and flu, we get the concordance of overall probabilities as follows:

Prob_total=prob(healthy)+prob(cold)+prob(flu) = 0,40

The **Kappa** is given by:

$$kappa = \frac{\frac{7 + 32 + 16}{74} - 0,40}{1 - 0,40} = 0,57$$

KAPPA: LEVEL OF CONSISTENCY BETWEEN THE MODEL AND REALITY

Medical Use Case:

		PREDICTED VALUE			
		Healthy	Cold	Flu	
ACTUAL VALUE	Healthy	7	6	2	15
	Cold	2	32	5	39
	Flu	1	3	16	20
		10	41	23	74

kappa = 0,57

Cohen's **Kappa** explains whether the class predictions match the real observations while accounting for randomness.

If the model correctly **identifies** all the classes, **the value gets close to** 1.

If the model does not recognize the classes, the Kappa **gets close to 0**.

Kappa measures whether the model correctly identifies the classes.

In real life, a Kappa **above 0.8** is considered **excellent**.

Conclusion

Evaluate, assess, compare and then decide. Wether it is in medecine, business or industry, metrics are the **energy of decision**. All these metrics can be found, and easily explained, in **our AI-Driven Software**. It is MyDataModels' DNA to empower business experts with the ability to take fully enlightened and meaningful decisions.

MyDataModels aims at making Artificial Intelligence accessible to all. Its flagship product is an Analytics Platform driven by Artificial Intelligence called **TADA**. Powerful and easy to use, TADA's Artificial Intelligence **helps any professional to analyze in depth its data and extract key insights**.

MYDATAMODELS

www.mydatamodels.com

Do not hesitate to send us a message if you
have any questions.
Our data science experts will be happy to help!

contact@mydatamodels.com

WHITE PAPER