KYTHERA™

DECIPHERING HEALTHCARE

Machine Learning: How Good is Your Model?

A REFERENCE GUIDE

PREFACE

Machine Learning has been widely adopted in a variety of industries. A 2020 Deloitte survey found that 67% of companies use machine learning, and 97% plan to use it. The McKinsey Global Institute estimates that Artificial Intelligence and Machine Learning could generate nearly $100B annually across the US healthcare system when focusing on the pharmaceutical industry.

Although Artificial Intelligence and Machine Learning power business strategy, innovation, and discoveries, Machine Learning is often referred to as a black box. Not only does the lack of clarity about the data required for machine learning models contribute to this situation, but the rules used to construct the models themselves also are not transparent, justifying the "black box" characterization and lack of confidence felt by the end-user.

PREFACE

For Machine Learning models to work, there must not only be sufficient data (quantity), the data must also be as accurate as possible (quality). You may have access to data but not know how complete the data is or if it is biased. Alogrithms that are trained on data sets that are biased will inherit that bias. While no data is perfect, it is important to turn it into its best and most appropriate format. It is also important to explain the model and the rules that were used to generate the output. Without transparency about data and rules, end-user confidence in using the output for business decisions may be limited.

Machine learning is being used for applications as varied as disease diagnosis, identifying patients and prescribers for new therapies and clinical trials, drug discoveries, and creating new business processes in the healthcare and life sciences industries. The growing adoption of AI and ML requires those in the industry to go beyond the latest marketing buzz and have the ability to explain how their model works.

This guide is intended to help data scientists and analysts understand how to assess model performance and provide guidance for explaining model rules for those who depend on the output.

- Assess Model Performance
- Confusion Matrix
- Probability Threshold
- Model Metrics Sensitivity, Specificity, and Accuracy
- Accuracy can be Misleading
- Other Model Metrics

Machine Learning: How Good is Your Model?

Machine learning (ML) **in healthcare** is widely used to predict the risk of illnesses; model disease progression; help radiologists interpret imaging tests; identify populations for specific therapeutic interventions; find patterns used in drug discovery, and its **uses are rapidly expanding**.

When machine learning is applied to healthcare, domain expertise is essential since healthcare data is complex and nuanced. Clinical codes like ICD10 and CPT codes, lab values, clinical notes from electronic health records, and other forms of structured and unstructured data are examples of the types of data currently used for training machine learning models. Data, whether structured or unstructured, requires correction, normalization, cleansing, and enrichment. High-quality data is paramount for machine learning training data. Without clean, consistently formatted data, training sets may be noisy, leading to less accurate predictive models.

It is important to understand how well a model performs when applied to actual healthcare data. Understanding how a machine learning model makes its predictions is important for interpretation, user adoption, and confidence in model predictions.

"Machine learning in healthcare is widely used to predict the risk of illness, model disease progression, help radiologists interpret imaging tests, identify populations for specific therapeutic interventions, find patterns used in drug discovery and its uses are rapidly expanding."

How to Assess Your Model

We need to understand the different ways a model can be right and the different ways it can be wrong. For the sake of simplicity, let's assume our model is to predict either "Yes" or "No." For example, Yes, a patient has diabetes; or, No, this patient does not have diabetes. Because our model predicts one of two outcomes, Yes or No, it is a binomial classifier model.

Our binomial classifier model can be right in two different ways and it can be wrong in two others.

Correct predictions:

- A patient does has diabetes and the model predicts, Yes, the patient has diabetes (true positive (TP)).
- A patient does not have diabetes and the model predicts, No, the patient does not have diabetes (true negative (TN)).

Incorrect predictions:

- A patient has diabetes and the model predicts, No, the patient does not have diabetes (false negative (FN)).
- A patient does not have diabetes and the model predicts, Yes, the patient does have diabetes (false positive (FP)).

We can verify if our model is right or wrong if we have data that includes each patient's true status, i.e., does or does not have diabetes.

Confusion Matrix

A binomial classifier model's two ways to be right and the two ways to be wrong are typically depicted as a 2 x 2 confusion matrix (see Table 1). A confusion matrix shows how a model's predictions are categorized as correct or incorrect *[a]*.

		Model Prediction	
		Predicted Yes	Predicted No
Actual / Truth	Yes, at risk	TP	FN
	No, not at risk	FP	TN

Table 1: 2 x 2 confusion matrix.
TP: True positive. FN: False negative. TN: True negative. FP: False positive.

Each cell of the 2 x 2 confusion matrix contains the number of patients who do (or do not) have diabetes and were correctly (or incorrectly) predicted as having diabetes. The 2 x 2 confusion matrix provides a more complete picture of how well our model performs because not all errors are equal (more details below).

"Understanding how a machine learning model makes its predictions is important for interpretation user adoption and confidence in model predictions."

RIGHT AND WRONG PREDICTIONS

As an example, suppose we have ten (10) patients where four (4) do have diabetes and six (6) do not, as shown in column "Actual/True Risk" in Table 2, below. The model predicts some patients correctly and others incorrectly, shown in column "Predicted Risk".

Patient ID	Actual / True Risk	Predicted Risk	Label
1	Yes	Yes	TP
2	Yes	Yes	TP
3	Yes	Yes	TP
4	Yes	No	FN
5	Yes	No	TN
6	Yes	No	TN
7	**No**	**Yes**	**FP**
8	No	Yes	FP
9	No	Yes	FP
10	No	Yes	FP

Table 2: Patient ID, actual/true risk, predicted risk, and confusion matrix label.

"We need to understand the different ways a model can be right and the different ways it can be wrong."

RIGHT AND WRONG PREDICTIONS

The labels -- TP, FN, TN, FP -- are based on the model's predictions (true v. false) and patients' status (positive/Yes v. negative/No).

We interpret Patient 4 (in purple) as the model incorrectly (falsely) predicts, No, this patient does not have diabetes (negative) when, in truth, this patient does have diabetes. This is a false negative (FN). We interpret Patient 7 (in red and bold text) as the model incorrectly (falsely) predicts, Yes, this patient does have diabetes (positive) when, in truth, this patient does not have diabetes. This is a false positive (FP).

Table 3 shows the 2 x 2 confusion matrix for these ten patients.

		Model Prediction	
		Predicted Yes	Predicted No
Actual / Truth	Yes, at risk	3 TP	1 FN
	No, not at risk	4 FP	2 TN

Table 3: 2 x 2 confusion matrix for 10 patients listed in Table 2.

"When evaluating a model ask "what are the business objectives that the model informs."

RIGHT AND WRONG PREDICTIONS

Note: sometimes the layout of the confusion matrix is flipped with "Actual/Truth" as columns and "Model Prediction" as the rows. It may also be ordered with "Yes" as the second row (or column) rather than the first. The ordering is very important when interpreting the cells' labels. We must always be sure to note the ordering of any confusion matrix we review.

We simplified this discussion by starting with a binomial classifier model. We can expand the 2 x 2 confusion matrix to describe models that predict more than two categories, e.g., "Low", "Medium", and "High", which produces a 3 x 3 confusion matrix.

QUESTION: When evaluating a model, ask: May we review its confusion matrix?

QUESTION: When evaluating a confusion matrix, ask: What is the layout of the matrix? What do the rows and columns represent?

So far, we have discussed these aspects of assessing how well our model performs:

- Our binomial classifier predicts one of two values: "Yes" or "No".
- Our binomial classifier can be correct in two ways and incorrect in two ways.
- We populate patient counts in the 2 x 2 confusion matrix by comparing the model's predicted "Yes" or "No" to each patient's actual outcome, "Yes" or "No".

Probability Threshold

In actuality, our binomial classifier model does not return “Yes” and “No”. Rather, this model returns a probability, a number between 0 and 1, inclusive. (Some models may return probabilities that are not well calibrated, therefore please consider these posts *[b]*.) A probability of 1 means that the model is completely confident that this patient has diabetes. A probability of 0 means that the model has no confidence that this patient has diabetes. A probability of 0.5 means that the model is equally confident that this patient does and does not have diabetes, i.e, the model is not confident at all.

How do we convert a predicted probability to “Yes” or “No”? We apply a threshold to the predicted probability in the following way:

Any prediction that is greater than or equal to the threshold is converted to 1 which is interpreted as “Yes”. Any prediction that is less than the threshold is converted to 0 which is interpreted as “No”.

- Model prediction >= probability threshold → 1/”Yes”
- Model prediction < probability threshold → 0/”No”

What probability threshold should we use? This depends upon our business objectives. An often-used threshold is 0.50, i.e., rounding the predicted probability. A probability threshold of 0.50 may not be the best threshold; in fact, it is arbitrary.

Sensitivity and Specificity

The choice of threshold is not solely a mathematical decision chosen by a data scientist; it is a decision based on business objectives. These objectives, in turn, speak to the ideas of sensitivity and specificity *[c]*. Sensitivity and specificity are two different measurements for the two different types of correct predictions our binomial classifier model can make.

Sensitivity

Rather than predict diabetes status, suppose our model is to predict whether or not a person has a deadly and communicable disease, e.g., smallpox. (Note that in 1980 the World Health Organization declared that smallpox had been eradicated.) It is critically important for our model to identify as many people with smallpox as possible. This -- the model predicting “Yes” for those who do, in truth, have smallpox -- is the definition of sensitivity. One way to increase the sensitivity of a model’s predictions is to lower the probability threshold. We may choose a probability threshold of, say, 0.30 so that more “smallpox-like” patients are identified by the model. We can think of this as “casting a wider net”.

However, there is no free lunch: casting a wider net -- increasing the model's sensitivity by reducing the probability threshold -- often results in "capturing" more patients who do not, in truth, have smallpox. Reducing the probability threshold typically results in increased false positives. In our example of smallpox, an increase in false positives may be acceptable (and necessary) in order to increase sensitivity (identifying more true positives).

Specificity

Let's return to our original example of predicting diabetes status. In tandem with sensitivity, specificity measures the model's ability to predict patients as "No" who do not, in truth, have diabetes. As we've seen above, changing the probability threshold also changes the number of patients that the model predicts as No.

The choice of probability threshold is a "knob" we can turn that determines the number and types of correct and incorrect predictions a model produces.

Synonymous terms for **sensitivity** are: recall, hit rate, and true positive rate (**TPR**).	Synonymous terms for **specificity** are: selectivity and true negative rate (**TNR**).

Wikipedia's Confusion matrix entry is an excellent resource for various model metrics.

"High quality data is paramount for machine learning training data. Without clean consistently formatted data, training sets may be noisy leading to less accurate models."

MODEL METRICS

QUESTION: When evaluating a model, ask: What are the business objectives that this model informs?

QUESTION: When evaluating a model, think about sensitivity: How important is it for the model to predict "Yes" for those who do have diabetes? What happens if the model incorrectly predicts "No" for these patients?

QUESTION: When evaluating a model, think about specificity: How important is it for the model to predict "No" for those who do not have diabetes? What happens if the model incorrectly predicts "Yes" for these patients?

Business partners may not be familiar with the impact of model metrics (e.g., sensitivity and specificity) with respect to actual patient counts. Therefore a data scientist should provide a table generated from a range of probability thresholds that includes: patient counts from each of the cells in the confusion matrix and the accompanying sensitivity and specificity values (as well as other model metrics which we will discuss later). The business objectives for our model and its sensitivity and specificity values will help us choose the most appropriate probability threshold.

In the following table of (fictitious) patient counts and (fictitious) probability thresholds, note the range of sensitivity, specificity, and accuracy values.

"All models are wrong, but some are useful."

– George Box *[d]*.

MODEL METRICS

Probability Threshold	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
[0,...,1]	Counts				Percent		
Thresh1	3	2	4	1	75.0	33.3	50.0
Thresh2	2	3	4	1	66.7	42.9	50.0
Thresh3	2	3	3	2	50.0	50.0	50.0
Thresh4	4	1	5	0	100.0	16.7	50.0
Thresh5	4	1	0	5	44.0	100	50.0

Table 4: Probability thresholds used, resulting patient counts in the 2 x 2 confusion matrix, and model metrics.

QUESTION: When evaluating a model, ask: What threshold was applied to the predicted probabilities?

QUESTION: When evaluating a model, ask: Who chose the probability threshold? How?

QUESTION: When evaluating a model, ask: What objectives were considered when choosing the probability threshold? Why?

QUESTION: When evaluating a model, ask for a table of actual counts (true positives; false positives; true negatives; false negatives) and their accompanying model metrics (e.g., sensitivity, specificity, accuracy, etc.) at various probability thresholds.

"Accuracy is an important model metric. However it is only one metric of many that indicates usefulness of our model."

Accuracy

When reviewing Table 4 on the previous page, notice that model accuracy is 50% for all (fictitious) probability thresholds. In this example, when the threshold changed, the sensitivity and specificity changed but accuracy did not. Why?

Accuracy is based only on predictions that are correct. Accuracy does not consider incorrect predictions.

- *Accuracy = (True Positives + True Negatives) / Total Number of Patients*

If we look at the TP and TN columns in Table 4, we see that for each row, they add up to 5. Since the total number of patients is 10, accuracy is always 5/10 * 100 = 50%. Accuracy does not consider false positive (FP) nor false negative (FN) predictions. But our business objectives do! We should assess a model's usefulness by considering its incorrect predictions, too.

Accuracy is an important model metric. However, it is only one metric -- of many -- that indicates the usefulness of our model. When evaluating a model, if we look only at accuracy, we may draw an incomplete -- and possibly misleading -- conclusion, as discussed next.

"Accuracy is based only on predictions that are correct. Accuracy does not consider incorrect predictions."

Accuracy Alone Can Be Misleading

Often, accuracy alone is misleading when our training dataset is highly imbalanced. One example of imbalanced data is when a dataset is made up of very few patients who, for example, have a rare disease, like Hutchinson-Gilford Syndrome. Suppose we have 1000 patients and only 10 have Hutchinson-Gilford Syndrome. If we train a model that does not consider this imbalance, then this model may predict everything as No, no one has Hutchinson-Gilford Syndrome. In this case, the model's accuracy is 99%!

- *Accuracy = (True Positives + True Negatives) / Total Number of Patients*
- *Accuracy = (0 + 990) / 1000 = 0.99*

Additionally, this model's sensitivity is 0% and its specificity is 100%.

- *Sensitivity = True Positives / (True Positives + False Negatives)*
- *Sensitivity = 0 / (0 + 10) = 0*
- *Specificity = True Negatives / (False Positives + True Negatives)*
- *Specificity = 990 / (0 + 990) = 1*

Probability Threshold	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
[0,...,1]	Counts				Percent		
ThreshA	0	990	0	10	0	100	99

Table 5: Metrics for a model trained on an imbalanced dataset. Accuracy is high and sensitivity is low.

We interpret a sensitivity of 0 as: for those patients who do have Hutchinson-Gilford Syndrome, our model identified no patients. This model is useless in predicting rare Hutchinson-Gilford Syndrome, in spite of its high accuracy.

Fortunately, there are ways to deal with training data that is imbalanced *[e]*.

QUESTION: When evaluating a model, ask: What are the other model metrics in addition to accuracy, e.g., sensitivity and specificity?

QUESTION: When evaluating a model, ask: What do these model metrics mean with respect to our business objectives?

Other Model Metrics

There are other metrics besides accuracy, sensitivity, and specificity that we can use to assess how good or useful a model is. As discussed above, sensitivity and specificity measure the two ways a model's predictions are correct. But what about the two ways a model can be incorrect?

False Positive Rate and False Negative Rate

In brief, false positive rate (FPR) measures one type of incorrect prediction: of the number of patients who do not have diabetes, how many did the model incorrectly predict as Yes, has diabetes? False negative rate (FNR) measures the other type of incorrect prediction: of the number of patients who do have diabetes, how many patients did the model incorrectly predict as No, do not have diabetes?

Note that there are other metrics that measure correct and incorrect model predictions, each from a somewhat different perspective. The best way to assess how good or useful our model is is to interpret multiple model metrics in light of our business objectives.

"The Wayfinder platform can dramatically reduce the time spent cleaning data and transforming it into a format so your model can interpret the data."

ROC AUC and AUC of Precision/Recall Curve

Another common model metric is the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. We won't go into any detail here (here are a few good resources for your consideration: *[f][g][h]*). However, we mention this metric because the same considerations should be given to ROC AUC as we've given to accuracy. ROC AUC is an important and often-used model metric.

However, evaluating a model based on ROC AUC alone can be misleading. Like accuracy, ROC AUC is only one metric that indicates the usefulness of our model. When evaluating a model, we should also consider the AUC of the precision/recall curve *[i]*.

We have discussed these aspects of assessing how well our model performs:

- Our binomial classifier model generates a probability between 0 and 1.
- We apply the threshold to convert a probability to "Yes" or "No".
- We select a probability threshold according to business objectives.
- We populate patient counts in the 2 x 2 confusion matrix by comparing the model's predicted "Yes" or "No" to each patient's actual outcome, "Yes" or "No".
- We compute model metrics from the confusion matrix.
- We should review a table of patient counts and model metrics based on a range of probability thresholds to help choose the best threshold.
- We recognize that a single metric provides an incomplete assessment.
- Accuracy alone can be misleading.
- We interpret our model metrics with respect to our business objectives.
- A good model supports our business objectives.

"Platforms like Wayfinder have embedded analytics with tested machine learning techniques."

OTHER MODEL METRICS

While the idea of developing algorithms and applying data science techniques may be daunting for some organizations, platforms like Wayfinder have embedded analytics with tested machine learning techniques. Having a technology platform to help you minimize time spent on data wrangling can greatly optimize your process. Our Wayfinder platform can dramatically reduce the time spent cleaning data and transforming it into a format so your model can interpret the data. We also have machine learning data assets that are available on the Wayfinder platform so that, even without a team of data scientists, organizations can begin to use machine learning to address their business challenges.

if you have any questions or comments about this guide or to learn more about Wayfinder, please reach out by emailing us at seek@kytheralabs.com.

BettyAnn Chodkowski, PhD

Kythera Senior Manager of Data Science

linkedin.com/in/bettyannc

BettyAnn has over 20 years of data science experience working with healthcare data to improve healthcare outcomes and a diverse background that includes biostatistics, medical engineering, computer science, and chemical & physical biology.

References

[a] Confusion matrix. From Wikipedia, the free encyclopedia. This Wikipedia entry is an excellent resource for all model metrics discussed here.

[b] For the sake of completeness, see these discussions:

- Are you sure that's a probability? Damien Martin, 23 May 2019.
- Python's «predict_proba» doesn't actually predict probabilities (and how to fix it). Samuele Mazzanti, 06 May 2021.

[c] Sensitivity and specificity. From Wikipedia, the free encyclopedia.

[d] All models are wrong. From Wikipedia, the free encyclopedia.

[e] A Gentle Introduction to Imbalanced Classification. Machine Learning Mastery. Jason Brownlee, 14 Jan 2020.

[f] Classification: ROC Curve and AUC. Machine Learning Crash Course: ML Concepts. 10 Feb 2020.

[g] How to Use ROC Curves and Precision-Recall Curves for Classification in Python. Machine Learning Mastery. Jason Brownlee, 13 Jan 2021.

[h] ROC Curves and Precision-Recall Curves for Imbalanced Classification. Machine Learning Mastery. Jason Brownlee, 16 Sep 2020.

[i] Classification: Precision and Recall. Machine Learning Crash Course: ML Concepts. 10 Feb 2020.

KYTHERA™

DECIPHERING HEALTHCARE

WWW.KYTHERALABS.COM