MILLIMAN WHITE PAPER

Evaluating supervised machine learning classification models in healthcare analytics

Whether you are a hospital administrator looking to improve workflow efficiency, a provider looking to improve patient outcomes, or an insurance administrator looking to decrease the number of fraudulent claims, evaluating a machine learning model is essential to your toolbox

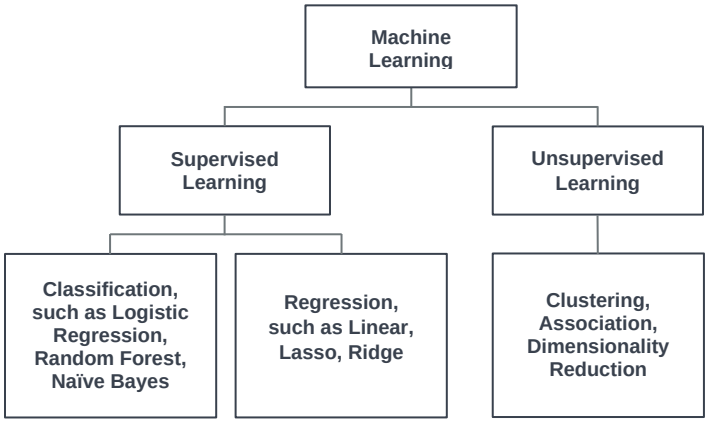
Ketaki Nagarkar, DPT
Ellyn Russo, MS

Milliman

Artificial intelligence (AI) tools, such as machine learning (ML), have the potential to improve healthcare operations and delivery, assist in diagnosis detection, and improve workflows.[1,2]

ML algorithms include both supervised and unsupervised models (see Figure 1 for further detail on these types).[3] Supervised ML algorithms used for predictive analytics learn, or train, from labeled data. This trained model is then used to predict future outcomes based on new, unseen data. An example application of a supervised ML algorithm is predicting the likelihood of a condition, or diagnosis, based on a patient's radiological scan. Conversely, an unsupervised ML algorithm does not have a labeled target outcome. An unsupervised ML algorithm can be used to identify subgroups of patients with similar characteristics.

FIGURE 1: EXAMPLES OF MACHINE LEARNING MODELS



There are several important components to consider when evaluating ML algorithms, including, but not limited to, bias testing and metric selection. Elimination of bias, such as sample bias, prejudicial bias, measurement bias, and algorithmic bias, is critical. Bias testing should be performed on an ongoing basis rather than as a onetime task and requires monitoring over the entire project lifecycle.[4]

We focus the discussion for the remainder of this paper on metric selection and some commonly used tools to evaluate the performance of a supervised binary (two classes) classification model, illustrating them through several example scenarios.

ML model performance evaluation

Common tools and metrics for performance evaluation of supervised classification models (definitions are provided throughout the remainder of the text and in the Appendix) include:

- Confusion matrix
- Receiver operator characteristic (ROC) curve
- Precision recall curve
- Accuracy
- F1 score
- Recall/true positive rate/sensitivity
- Precision/positive predictive value
- True negative rate/specificity

A confusion matrix is a helpful visual of the information needed for model performance evaluation and places true and false class predictions into a two-by-two format based on the model's predicted probabilities at a certain threshold (see Figure 2 for a confusion matrix template).

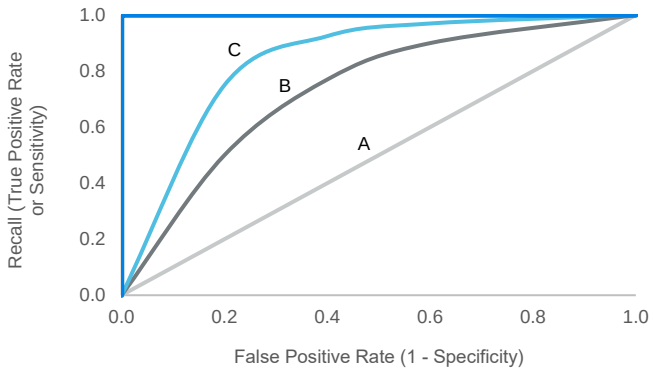
FIGURE 2: CONFUSION MATRIX FOR BINARY CLASSIFICATION MODEL

	ACTUAL POSITIVE	ACTUAL NEGATIVE
Predicted Positive	True Positive (TP)	False Positive (FP; Type I error)
Predicted Negative	False Negative (FN; Type II error)	True Negative (TN)

ROC curves are commonly used to visualize models when assessing both classes (0 and 1). An ROC curve plots recall, or true positive rate or sensitivity, on the y-axis against the false

positive rate, or 1 – specificity, on the x-axis at different thresholds or operation points.[5] The area under the curve (AUC) quantifies performance of the model at possible thresholds: the larger the value of the AUC, the better the overall diagnostic performance of a test. As illustrated in Figure 3, model C has a larger ROC-AUC as compared to B and, thus, a better ability to discriminate.

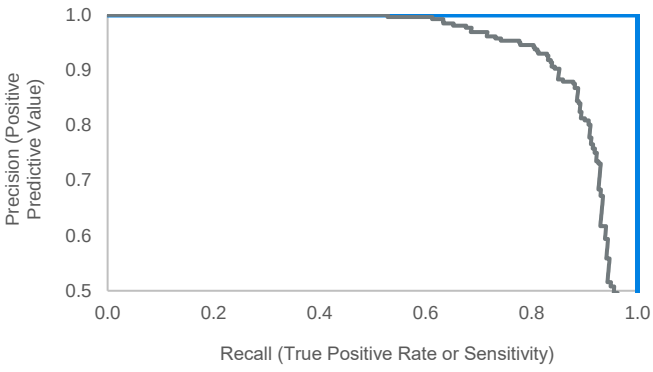
FIGURE 3: EXAMPLE RECEIVER OPERATOR CHARACTERISTIC (ROC) CURVES



Note: The A line represents the outcomes for a random guess (no skill model).

The ROC-AUC is a well-known measure and is well-suited for balancing the trade-off between sensitivity and specificity. The precision recall curve (PRC), however, is useful when evaluating primarily for the positive class in an imbalanced data set.[6] The PRC is obtained by plotting the precision, or positive predictive value, on the y-axis against recall on the x-axis at different thresholds (see Figure 4). It does not consider the true negatives. The PRC average precision (AP) score is a useful metric to evaluate the PRC-AUC. A higher AP score is indicative of a better ability to identify the true positive cases (improved recall) while minimizing false positives (better precision).[7-11]

FIGURE 4: EXAMPLE PRECISION RECALL CURVE (PRC)



Applications of ML model evaluation metrics

To help understand how these metrics are used, we explore the following four example application scenarios.

Scenario 1: Consider 1,000 people, of which 100 have pulmonary hypertension. Here, 90% of the patients are disease-free, and 10% are the (minority) class of interest, a scenario that may be relatable to many healthcare diagnostic tests that detect medical conditions. In this example, what is the consequence of a false positive, that is, if the model falsely predicts a healthy patient as having a condition leading to unneeded treatments (see Figure 5)? On the other hand, what is the penalty of having a false negative, that is, if the model falsely predicts the patient as being healthy when, in fact, this patient has the condition, causing a potential delay in or lack of lifesaving treatment?

FIGURE 5: CONFUSION MATRIX FOR DETECTING MEDICAL CONDITION

	ACTUAL POSITIVE	ACTUAL NEGATIVE
Predicted Positive	TP = condition is correctly detected and treated	FP = test falsely identifies condition resulting in unnecessary treatment
Predicted Negative	FN = condition is present but not detected and treated	TN = condition is not present and no treatment is ordered

If the goal of the model is to avoid false negatives, then an overall classification accuracy score (correct predictions divided by total predictions) of 90% is not necessarily a helpful measure to assess performance for detection of the rare occurrence. A false negative would mean a lifesaving intervention was missed. A false positive would likely mean more expensive testing and unnecessary intervention. In this situation, if minimizing false negatives is more important, then recall, or true positive rate, will be more informative. The precision recall trade-off, discussed further in the Appendix, should be considered along with the business context to achieve the correct balance.[9]

Scenario 2: A mail order pharmacy is trying to increase its market size. The company is looking to identify new members who are most likely to use their service from the total insured population. The class of interest is the new members most likely to use the service and is a percentage of the total insured members. Once identified, marketers inform those members about the mail order service and its benefits.

False negatives, or not identifying members that would use the service, would mean the company may lose out on potential customers, resulting in decreased profits (see Figure 6). False positives may mean that some marketing efforts may go to waste; however, we assume that the cost of marketing is

lower than the cost of losing a potential customer. Thus, recall, or true positive rate, will again be most informative of model performance.

FIGURE 6: CONFUSION MATRIX FOR IDENTIFYING MEMBERS LIKELY TO UTILIZE SERVICE

	ACTUAL POSITIVE	ACTUAL NEGATIVE
Predicted Positive	TP = member received marketing and uses service	FP = member will likely not use service, marketing effort overutilized
Predicted Negative	FN = member will likely use service but not targeted for marketing	TN = member will likely not use service, not targeted for marketing

Scenario 3, Perspective 1: A healthcare organization information systems department is tasked with monitoring spam email detection. The class of interest is the email identified as spam. A false negative is a spam email arriving in the user's inbox who must then use judgment to determine whether it is a legitimate or spam message (see Figure 7). A false positive might, on the other hand, mean that an important email never made it to the user. In this case, a false positive is significantly more detrimental than a false negative. The most informative metric in this scenario is precision, or positive predictive value, as it implies a lower number of false positives.

FIGURE 7: CONFUSION MATRIX FOR SPAM EMAIL DETECTION

	ACTUAL POSITIVE	ACTUAL NEGATIVE
Predicted Positive	TP = spam, user never sees email	FP = not spam, user never sees email
Predicted Negative	FN = spam, user sees email and must use judgement	TN = not spam, user sees email

Scenario 3, Perspective 2: Facing a similar task of identifying spam email as above, the organization is now on high alert for a security threat through email hacking such that a false negative carries more significance (see Figure 7). Here, recall, or true positive rate, instead of precision, or positive predictive value, would be most useful.

Scenario 4: Finally, consider a hospital interested in determining which patients will require a longer-than-average length of stay once admitted (we'll assume the average is five days). This prediction allows the hospital to budget resources efficiently.

A false positive means a patient anticipated to stay more than the five-day threshold is discharged sooner, leaving allocated resources idle, e.g., staff, hospital beds, and supplies (see Figure 8). Conversely, a false negative means a patient predicted to stay fewer than five days remains for a longer stay, requiring unanticipated use of resources and a possible decline in quality of patient care. Both predictive misses are expensive.

FIGURE 8: CONFUSION MATRIX FOR HOSPITAL LENGTH OF STAY

	ACTUAL POSITIVE	ACTUAL NEGATIVE
Predicted Positive	TP = longer-than-average length of stay, resources sufficient	FP = less-than or average length of stay, resources left idle
Predicted Negative	FN = longer-than-average length of stay, resources insufficient, quality lower	TN = less-than or average length of stay, resources sufficient

When both false negatives and false positives matter, the F1 score of the class of interest is a desirable choice of metric for model evaluation. F1 takes into consideration the harmonic mean of precision, or positive predictive value, and recall, or true positive rate, and their relative contributions are equal.

Additional considerations for analyses using healthcare data sets

Exploratory data analysis (EDA) is an important initial step prior to applying ML. Data quality, distributions, types, dimensions (number of input variables), and cardinality (unique values within a categorical variable) will all influence preprocessing, scaling, and feature engineering before the data is ready for predictive analysis and modeling.

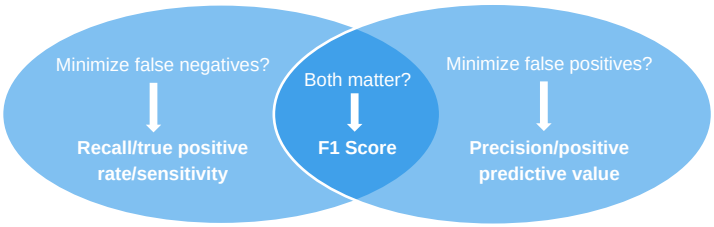
The number of positive observations (outcome of interest) divided by total observations (prevalence) will be close to 50% in a balanced data set. There are techniques to address class imbalance, and improve model performance, such as oversampling of the minority class or under-sampling of the majority class. Synthetic Minority Oversampling Technique (SMOTE) is a popular oversampling approach wherein synthetic samples of the minority class are created.[12-15]

Summary

While there are many different types of ML algorithms and accompanying metrics for evaluation, we have discussed here some of the more commonly used tools and metrics for evaluating a supervised binary classification model. Both a ROC curve and PRC are useful for understanding model performance at different decision thresholds. When evaluating supervised binary classification problems, AUC from the ROC curve is used most frequently. This metric provides overall model performance and is useful when identification of both classes is important.[5] For evaluation of the model to discriminate the positive class in an imbalanced data set, the PRC (AP score) may be a better choice for comparison of model performance.[6-11]

Consideration of the business perspective should contribute to the selection of evaluation metrics and optimization of decision threshold. For metric selection, if it is vital to minimize false negatives, then a focus on recall, or true positive rate, is warranted (see Figure 9). If minimizing false positives is crucial, then precision, or positive predictive value, is more important. When both false negatives and false positives matter, the F1 score will be helpful. Though one metric may be identified as most informative to the model's performance evaluation, it should be explored in the context of other metrics as well.

FIGURE 9: BUSINESS PERSPECTIVE CONSIDERATIONS FOR MODEL EVALUATION METRIC SELECTION



Business partners and model developers should work closely throughout the life cycle of the ML project to mitigate the potential for bias while meeting business objectives.[4] Bias can be present in classification models if use of the model leads to unfair discrimination or disparate outcomes for certain subgroups of the population, such as individuals of certain races, genders, or sexual orientations. The potential for this should be evaluated alongside the model performance that we described.

To understand more about precision-recall trade-off, metric selection, threshold optimization, and tuning model performance, further discussion with a data scientist will be beneficial. A few questions your data science team may pose to you include:

- What is the business question?
- What is the prevalence of the class of interest or condition of interest compared to the total number of observations?
- What perspective matters to your business, or what is the cost of identifying each class incorrectly?

Appendix

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Accuracy = \frac{True\ Positives + True\ Negatives}{All\ Positives + All\ Negatives}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Average Precision "summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight."[12]

$$AP = \sum_n (R_n - R_{n-1}) P_n$$

Threshold, or decision boundary, is a term used for the parameter that converts a predicted probability (or scoring) into a class label.[6] For a binary classification, a label of 1 indicates the positive class or the class of interest. If the decision boundary is set at 0.5, a predicted probability of <0.5 indicates class 0 and ≥0.5, class 1. By increasing the threshold, the number of false positives will decrease while false negatives will increase, and vice versa.

Precision Recall tradeoff is an important consideration prior to threshold optimization. The business context should inform an acceptable level of false negatives.[9]

Milliman

CONTACT

Ketaki Nagarkar
Healthcare Analyst
ketaki.nagarkar@milliman.com

Ellyn Russo
Healthcare Consultant
ellyn.russo@milliman.com

ENDNOTES

[1] Bohr, A. and Memarzadeh, K. The Rise of Artificial Intelligence in Healthcare Applications. Artificial Intelligence in Healthcare 2020; 25-60. Retrieved December 23, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/.

[2] Marr, B. How Is AI Used in Healthcare - 5 Powerful Real-World Examples That Show the Latest Advances. Forbes 2018. Retrieved December 23, 2022, from https://www.forbes.com/sites/bernardmarr/2018/07/27/how-is-ai-used-in-healthcare-5-powerful-real-world-examples-that-show-the-latest-advances/?sh=2d41408d5dfb.

[3] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science 2021;2(3):160. Retrieved December 23, 2022, from https://link.springer.com/article/10.1007/s42979-021-00592-x.

[4] Jones, T.M. Machine Learning and Bias: Look at the Impact of Bias and Explore Ways of Eliminating Bias from Machine Learning Models. IBM Developer. 2019. Retrieved December 23, 2022, from https://developer.ibm.com/articles/machine-learning-and-bias/.

[5] Park S.H., Goo J.M., and Jo C.H. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. Korean Journal of Radiology 2004;5(1):11-8. Retrieved December 23, 2022, from https://synapse.koreamed.org/articles/1027596.

[6] Saito T. and Rehmsmeier M. The Precision-Recall Plot is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS One 2015;10(3):e0118432. Retrieved December 23, 2022, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432.

[7] Scikit-learn developers. Sklearn.metrics.average_precision_score. Retrieved December 23, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html.

[8] Dankers F.J.W.M., Traverso A., Wee L., et al. Chapter 8 Prediction Modeling Methodology. In: Kubben P., Dumontier M., Dekker A., editors. Fundamentals of Clinical Data Science. Springer 2019. Retrieved December 23, 2022, from https://www.ncbi.nlm.nih.gov/books/NBK543534/.

[9] Hillis, S. and Hoormann, S. Precision and Recall: Understanding the Trade-Off. The Opex Analytics Blog 2016. Retrieved December 23, 2022, from https://medium.com/opex-analytics/why-you-need-to-understand-the-trade-off-between-precision-and-recall-525a33919942.

[10] Draelos, R. Measuring Performance: AUPRC and Average Precision. Glass Box Machine Learning and Medicine. Retrieved December 23, 2022, from https://glassboxmedicine.com/2019/03/02/measuring-performance-auprc/.

[11] Boyd, K., Eng, K.H., Page, C.D. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD Lecture Notes in Computer Science 2013;8190:451-66. Springer, Berlin, Heidelberg. Retrieved December 23, 2022, from https://pages.cs.wisc.edu/~boyd/aucpr_final.pdf.

[12] Chawla, N.V., Bowyer, K.W., Hall, L.O., et al. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 2002;16:321-57. Retrieved December 23, 2022, from https://www.jair.org/index.php/jair/article/view/10302.

[13] Rahman, M.M. and Davis, D.N. Addressing the Class Imbalance Problem in Medical Datasets. International Journal of Machine Learning and Computing 2013;3(2):224-8. Retrieved December 23, 2022, from https://www.researchgate.net/publication/239608168_Addressing_the_Class_Imbalance_Problem_in_Medical_Datasets.

[14] Zhao Y., Wong Z.S., and Tsui, K.L. A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-alike Sound-alike Mix-up Incident Detection. Journal of Healthcare Engineering 2018. Retrieved December 23, 2022, from https://www.hindawi.com/journals/jhe/2018/6275435/.

[15] Fujiwara K., Huang Y., Hori K., et al. Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis. Frontiers in Public Health 2020;8:178. Retrieved December 23, 2022, from https://www.frontiersin.org/articles/10.3389/fpubh.2020.00178/full.