The Difficulty of Achieving High Precision with Low Base Rates for High-Stakes Intervention

Ryan S. Baker
Graduate School of Education, University of Pennsylvania, ryanshaunbaker@gmail.com

Caitlin Mills
Department of Educational Psychology, University of Minnesota, cmills@umn.edu

Jaeyoon Choi
School of Education, University of California Irvine, jaeyoon.choi@uci.edu

Automated detectors are routinely used in learning analytics for high-stakes, high-risk interventions. Such interventions depend on detectors with a low rate of false positives (i.e., predicting the construct is present when it is not present) in order to avoid giving an intervention where it is not needed, especially when such interventions can be costly or even harmful. This in turn suggests that such a detector needs to have high precision at the cut-off used by the detector for decision-making. However, high precision is difficult to achieve for the common case where the base rate of the target construct is low. In this paper, we demonstrate the difficulty of achieving high precision for low base rates, and demonstrate how other metrics (such as F1, Kappa, Specificity, and AUC ROC) are insufficient for this specific use case and situation, despite their merits and advantages for other use cases and situations.

CCS CONCEPTS

•Human-centered computing •Social and professional topics •**Applied computing~Education~Interactive learning environments**

Additional Keywords and Phrases:

Precision, Prediction Model, Automated Detection, Unbalanced Data

ACM Reference Format:

First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock,

NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1 INTRODUCTION

Our field is, to a large extent, built on detectors – mathematical or computational models that infer something about the learner or the learning situation. We develop dashboards and reporting [1, 2, 3] that often embed inferences made by models and detectors. Such inferences are often used by instructors, teachers, and other stakeholders to make decisions [4, 5], including some high-stakes decisions [6]. Detectors can also be used as the basis of an automated intervention, such as deciding to send the student an automated email warning [7], offering automated scaffolding [8], or selecting which material the student will work on next [9]. Intervention of these types are not the only ways that detectors are used in our field – they are also used as components in analyses to study a range of research questions [10, 11, 12]. But intervention is a highly common use of detectors in LAK.

Interventions are sometimes fail-soft and sometimes high-stakes [13]. In other words, sometimes an intervention can safely be given to a student who doesn't need it, and sometimes an intervention is sufficiently costly or harmful if mis-applied, that we want to avoid giving it to the wrong students. An intervention can be high-stakes either for the receiving student (a risk of harm to the student if mis-applied) or the organization or individual delivering the intervention (costing too much time or resources). It is therefore important to carefully consider detector performance – and the best way to assess anticipated performance – to account for imperfect predictions. However, performance can be evaluated in many ways, and it can be hard to know which way of evaluating performance is best.

This issue becomes even more complicated when we consider that while sometimes the behavior or student state or situation that we are interested in intervening on is common, other times it is uncommon, or even rare. The field of Learning Analytics has traditionally examined constructs with a range of base rates (frequency of the state of interest), and many of the behaviors and student states we are interested in detecting have low base rates – that is to say, those behaviors and student states are uncommon or rare, yet they are still important to detect and respond to from an learning perspective. For example, cognitive-affective states such as curiosity, anxiety, and surprise are reported quite rarely (often less than 5%; [14]). However, each of these represents a theoretically and practically important experience during learning that may be an ideal place to offer an automatic response [15]. Some disengaged behaviors associated with long-term outcomes are similarly rare [12]. States like frustration and mind wandering can also occur with a disproportionately low frequency, on the order of 10-15% of the time [16, 17].

Situations with a combination of a high-stakes intervention and low base rate are seen fairly often in our community, particularly for applications such as at-risk prediction. However, it is not always acknowledged how difficult this task can be: it can be hard to tell if a detector is excellent, good enough with some caveats, or completely missing the mark for deploying an intervention for a behavior that is critical, but simply doesn't occur often. The qualities needed for a detector will vary if the intervention it is combined with is high-risk (harmful if mis-applied) compared to an intervention that is lower-risk. Models are sometimes evaluated with metrics that are not appropriate for their specific application. Previous work has pointed out some of these issues [18, 19, 20]. For example, [20] noted that overprediction of the majority class (i.e., the one that is more common in the data – such as "not confused" occurring more often than "confused") is quite common in existing published work. This is not surprising, yet it has important implications for responding to the target behavior. Along these lines, commonly used metric like accuracy, F1, and kappa are unlikely to be the best metric to guide intervention decisions when data is highly imbalanced since they do not distinguish performance between the majority and minority classes – thus not necessarily picking up on the overprediction mentioned above [19]. Researchers in our community often know about the limitations of accuracy, but other aggregate evaluation metrics (kappa, F1) also have challenges in distinguishing between false positives and false negatives (e.g., consider overpredicting confused vs. not confused), and are likely to be impacted by low base rates [19].

Previous discussions related to LA constructs have focused on comparing metrics like F1, kappa, AUC ROC, etc. with a goal of providing recommendations for what metrics to report and use for evaluation; here we suggest that such a generalized recommendation may not be fruitful, and instead argue that there are some contexts when a generally recommended and useful metric would be unhelpful (or even misleading) for the situation when the base rate of the behavior/state we care about supporting is very low. As we will discuss below, in situations where the goal is to apply a single intervention while ensuring it is only given to students who need it, the metric of precision is likely the most appropriate. However, precision is often not considered. Precision – whether everyone who receives an intervention needed it – is important when an intervention is harmful if mis-applied; in those cases, we want to minimize applying interventions when they are not warranted. But insufficient attention is often given to how difficult it can be to achieve good precision when base rates are low. In other words, our community is routinely taking on a challenge (high-stakes intervention for low base rate constructs) that it is mathematically difficult to succeed at. In the following section, we demonstrate this challenge using examples of three different base rates for the behavior that would require an intervention (50%, 10%, 2%). We then follow up by demonstrating how AUC ROC (a metric with many merits and uses [18]) and specificity cannot provide this same information (despite their other merits and uses).

2 ANALYSIS OF PRECISION

Let us start by setting the stage with a situation often seen in learning analytics practice. We have developed an intervention that we expect to have a larger positive impact when given to a student who needs it, but it is costly to provide. Perhaps it is time-consuming for a teacher or school leader to provide. Perhaps it is financially costly. Perhaps it has negative impacts on the student if given to a student who does not need it. In a case like this, we want to avoid giving this intervention to students who do not need it. In other words, we want to capture as many true positive cases (i.e., predicting the behavior correctly when it actually occurred) as possible but avoid false positives (i.e., predicting the behavior incorrectly when it did not actually occur), where a model labels negative cases as positive. In other words, we want our detector driving selection of this intervention to have high precision. A relatively large number of learning analytics interventions have this property, including tutoring programs given to at-risk students [21], interventions given by highly busy academic advisors [7], and intrusive automated responses to disengaged behaviors [22].

This metric, precision, is mathematically defined as $\frac{True\ Positive\ Count}{True\ Positive\ Count+False\ Positive\ Count}$. Let us say, for the sake of our upcoming case studies, that we want a precision of 0.8 or better for our high-stakes intervention. In other words, we want at least 80% of the cases that the detector labels as positives to be true positives.

In such a case, we might be willing to tolerate a lower level of recall, where recall is defined as $\frac{True\ Positive\ Count}{True\ Positive\ Count+False\ Negative\ Count}$. Let us say that we are willing to miss half of the students who are in need (perhaps because we cannot afford to give interventions to all students in need anyway). Making some concrete impact is better than the current state of affairs. In that case, we would be willing to accept a recall of 0.5 or better. In other words, we want at least 50% of the cases that are positives in the data to be labeled as positives by the detector.

In this case, then, it is important that we find a detector with precision of 0.8+ and a recall of 0.5+ for the right class. Sometimes packages provide metrics for the class labeled "1" by default; in other cases, packages provide metrics for the majority class. In general, whichever class is our class of interest – whichever class is the class receiving the intervention – this is the class that we should be calculating precision and recall for. Not the other class; not an average of the two classes. In general, below, for simplicity, we will treat the positive class as the class of interest as we walk through a set of case studies.

Figure 1 shows the false positive rate and Kappa for each of the following case studies. Also, for simplicity, our examples below will each consist of 10,000 data points.

2.1 Case study 1: Base rate of intervention class = 0.5

How does the base rate impact the difficulty of achieving our standard for intervention (precision of 0.8+ and recall of 0.5+)? Let us first consider the easiest case: where the positive class and negative class are evenly balanced 50/50 (i.e., base rate = 0.5).

In this case, we have 5000 positive cases (P) and 5000 negative cases (N). In this case, we can achieve a recall of 50% if we have 2,500 true positives TP (5000 ÷ 2) and therefore 2,500 false negatives FN (5000 - 2500).

Now, we want to calculate the number of false positives FP that gives us a precision of exactly 80%. To have a precision of 80%, we need for TP/(TP+FP) to equal 0.8. By simple algebra, there must be 625 false positives FP to have a precision of 0.8. If there are 625 false positives out of a set of 5000 actual negatives, then the false positive rate FP/(Total N) equals 0.125. In other words, precision of 0.8 and recall of 0.5 is achieved by a true positive rate of 50% and a false positive rate of 12.5%.

This model achieves a fairly modest Cohen's Kappa of 0.37. This is a good model. It does not seem particularly hard to achieve, although we already – even with fully balanced data – have a ratio of true positives to false positives of 4:1 (mathematically the same as that 80% precision – 80%/20% is 4:1).

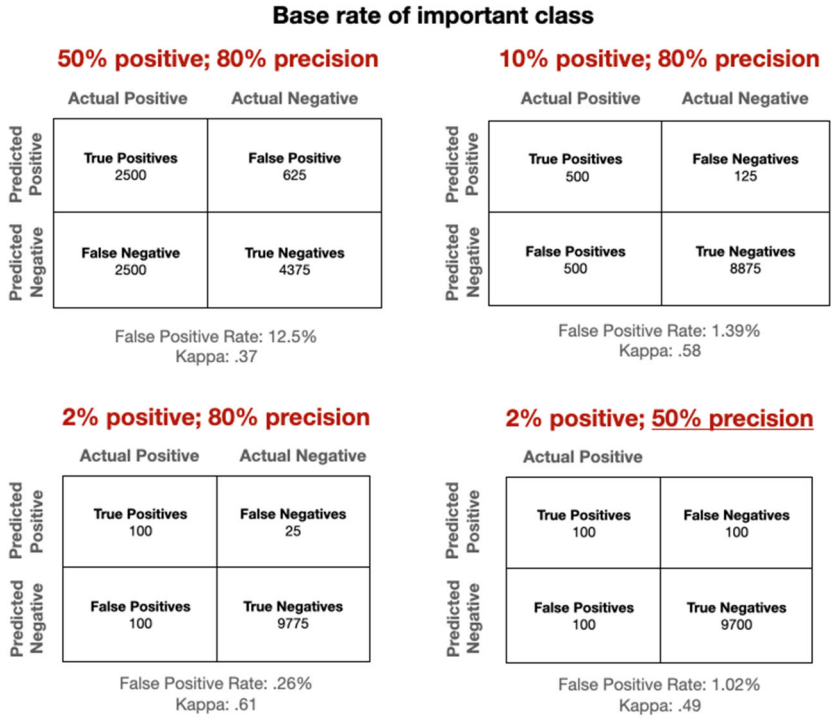


Figure 1: False Positive Rate and Kappa for Different Base Rates and Precision

2.2 Case study 2: Base rate = 0.1

So now let us consider a case where the base rate is a realistic but more challenging 10% (i.e., base rate = 0.1), a common base rate in learning analytics applications. In this case, we have 1000 positive cases P and 9000 negative cases N. In this case, we can achieve a recall of 50% if we have 500 true positives TP (500 ÷ 2) and therefore 500 false negatives FN (1000 - 500).

Now, we want to calculate the number of false positives FP that gives us a precision of exactly 80%. To have a precision of 80%, we again need for TP/(TP+FP) to equal 0.8. By simple algebra, there must be 125 false positives FP to have a precision of 0.8. If there are 125 false positives out of a set of 9000 actual negatives, then the false positive rate FP/(Total N) equals 0.0139. In other words, precision of 0.8 and recall of 0.5 is achieved by a true positive rate of 50% and a false positive rate of 1.39%.

This model achieves a better Cohen's Kappa of 0.58. It has to have a better Kappa; it has to do a much better job of avoiding false positives than the balanced model in order to still achieve a precision of 0.8.

Thus, this model has to be an excellent model to achieve similar results to the example above. It is only mis-classifying a negative example as positive 1.39% of the time. How many models in our field successfully achieve this? Is this a reasonable standard?

2.3 Case study 3: Base rate = 0.02

Now let us consider a case where the base rate is a very low 2% (i.e., base rate = 0.02). This may seem extremely low, but constructs of interest to learning analytics research and practice can often be rare, such as specific forms of gaming the system [23], affective states such as frustration [24], and school events such as a student being suspended [25]. The detection of increasingly rare constructs becomes more feasible with the advent of increasingly massive educational data sets [26, 27, 28]. A construct can be rare but highly important to student outcomes (as both gaming the system and school suspension have been found to be), yet the way we evaluate our detectors can play a huge role in whether we catch the actual instances of the behavior/state vs. simply (and accurately) predicting that it mostly will not happen. We can see this by examining an example with an extremely low base rate of 2%.

In this case, we have 200 positive cases P and 9800 negative cases N. In this case, we can achieve a recall of 50% if we have 100 true positives TP (200 ÷ 2) and therefore 500 false negatives FN (200 - 100).

Now, we want to calculate the number of false positives FP that gives us a precision of exactly 80%. To have a precision of 80%, we again need for TP/(TP+FP) to equal 0.8. By simple algebra, there must be 25 false positives FP to have a precision of 0.8. If there are 25 false positives out of a set of 9800 actual negatives, then the false positive rate FP/(Total N) equals 0.0026. In other words, precision of 0.8 and recall of 0.5 is achieved by a true positive rate of 50% and a false positive rate of 0.26%.

This model achieves a slightly better Cohen's Kappa of 0.61. At this point, the majority class is so dominant that reducing the false positive rate by around 80% doesn't budge Kappa much (roughly a .03 increase).

Do we actually expect any learning analytics model to achieve a false positive rate of 0.26%?

2.4 Case study 4: Base rate = 0.02, precision = 0.5

Perhaps, you say, a precision of 80% is simply unrealistic for a base rate of 2%. Perhaps you are willing to accept lower precision – 50%, say. Lower precision may be more acceptable when the intervention applied is relatively fail-soft – where intervention is not costly and students will not be harmed by the intervention. Does lowering the precision rate that we are willing to accept make things a bit more feasible? Let's take a look.

As with case study 3, we have 200 positive cases P and 9800 negative cases N. Again, we can achieve a recall of 50% if we have 100 true positives TP (200 ÷ 2) and therefore 500 false negatives FN (200 - 100).

Now, we want to calculate the number of false positives FP that gives us a precision of exactly 50%. To have a precision of 50%, we again need for TP/(TP+FP) to equal 0.5. By simple algebra, there must be 100 false positives FP (the exact same number as TP) to have a precision of 0.5. If there are 100 false positives out of a set of 9800 actual negatives, then the false positive rate FP/(Total N) equals 0.0102. In other words, precision of 0.5 and recall of 0.5 is achieved by a true positive rate of 50% and a false positive rate of 1.02%.

That's not all that much easier; it's still slightly harder than achieving a precision of 0.8 for a base rate of 0.1. Once base rates get very low, it's a rather difficult proposition to achieve even fairly modest precision, highlighting

why it is important to consider when attempting to respond to behaviors and states that are rare, yet important to address.

2.5 Summary so far: Implications for precision

As the previous section demonstrates, good precision is difficult to achieve for minority classes in imbalanced data. What are the implications of that? For one thing, it suggests that a certain class of intervention – interventions that have significant benefits if applied properly, but that have significant costs if mis-applied – may be infeasible for the level of quality of models and base rates we typically see at LAK, EDM, and related communities. This type of intervention depends on high precision, which is difficult to achieve with low base rates.

However, that does not mean that all of our years of modeling effort is futile. There remain many uses of models that are not problematic: First, many interventions are fail-soft [13]: beneficial when applied correctly, but not particularly expensive or problematic if mis-delivered. Examples of fail-soft interventions include email reminders to start assignments [29], mastery learning-based problem selection within intelligent tutoring systems where giving the student an extra problem or two involves limited time cost [9], and supportive automated messages for student affect [30]. In these cases, even fairly low precision is acceptable, so long as recall is high, but it is still more advantageous on the whole to use an algorithm than to simply provide the intervention to everyone.

Second, not all interventions use binary differentiation. Many at-risk prediction systems used at scale in the real world instead provide probabilities or ordinal risk buckets to teachers and school leaders (e.g., [6]), a form of reporting that reduces the risks of making decisions based on a single threshold with low precision. In particular, even if the highest-risk bucket has a low base rate, middle levels of risk will be associated with higher base rates, reducing the likelihood of very low precision while still representing a valuable differentiation from the lower levels of risk. In general, this form of reporting requires greater analytics literacy from its users and stakeholders than a single pre-chosen threshold, but in turn offers greater support for making decisions informed by considerations of precision, recall, benefit and cost.

Third, automated detectors and models are often used in our field within discovery with models analyses [31] (e.g., [10, 11, 12]), where the algorithm's outputs are used as a component in another analysis for basic discovery – such as correlation or statistical models or cluster analysis. In these cases, probabilities are often retained rather than thresholding to binaries, so the subtle differences captured by differences in probability are retained, and therefore the challenge of minimizing false positive rates with highly imbalanced data is bypassed. We return to this case at the end of the next section.

3 ANALYSIS OF AUC-ROC

Let us turn now to an alternate metric: AUC ROC [32]. AUC ROC is both the area under the ROC curve (which shows the trade-off between false positive rates and true positive rates for all possible cut-off thresholds) and the probability that if the model is given one positive case and one negative case, it will accurately distinguish which is which [33].

We compare to AUC ROC, as it is very commonly used in our community and many have argued that it is the most informational metric overall [34, 18]. Further, AUC ROC has been argued to be particularly relevant for evaluating at-risk prediction [32], a key application of learning analytics that can vary considerably in its base rate [35]. In line with our discussion about precision above, we consider whether the generally recommended use of AUC ROC may have similar issues when the intervention behavior only occurs at a low probability.

To be clear, we believe that AUC ROC is a valuable metric. However, it provides information on something very different from precision; it is relevant to different model uses than precision, and an excellent AUC ROC is not indicative of an excellent precision (and vice-versa). In particular, AUC ROC cares equally about the entire distribution of data and probability, whereas precision represents a single threshold. For predicting the minority

class for a low base rate, that decision threshold must necessarily be near the end of the probability spectrum for the model – the goal is to capture a small number of positive cases (high proportion true positives) and avoid capturing more than a small number of negative cases (very low proportion false positives).

Take, for instance, the case where base rate = 0.1, precision = 0.8, and recall = 0.5. From above, we know that to achieve precision of 0.8 and recall of 0.5 for a base rate of 0.1, we need to capture 500 true positives but no more than 175 false negatives. This represents only 6.25% of the overall 10,000 data points, creating an ROC curve that anchors at an X value of 1.25% (125 false negatives is 1.25% of the overall 10,000 data points). All data to the right of there on the graph – with probabilities lower than the cut-off threshold – are irrelevant to the calculation of precision, but represent 93.75% of the data and 98.75% of the X values that area can be determined for. As Figure 2 shows, a broad range of AUC values can be obtained for this one anchor point: from barely above chance to barely below perfect. As such, AUC ROC is telling us something fundamentally different than precision; it tells us how a model functions across all prediction thresholds. As such, it is useful for applications where multiple prediction thresholds are used (e.g., [6]) or for correlational or cluster analysis using predictions (e.g., [10, 11, 12]), but not when a single decision threshold is used and we care about that single decision threshold's precision.

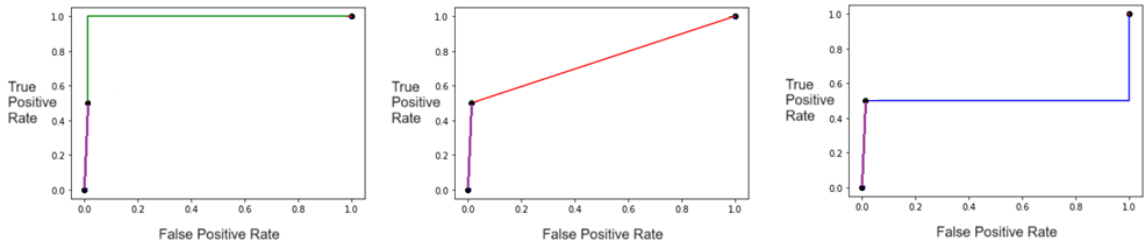


Figure 2: Possible AUC ROC curves for precision of 0.8 and recall of 0.5 for a base rate of 0.1. The left curve (highest AUC ROC) is very close to perfect performance; the right curve (lowest AUC ROC) is very close to chance performance; the middle curve (halfway in between) achieves an AUC ROC of 0.74. The three curves are identical at the start.

This is even more extreme for a base rate of 0.02. In this case, to achieve a precision of 0.8 and recall of 0.5, we need to capture 100 true positives but no more than 25 false negatives. This represents only 1.25% of the overall 10,000 data points, creating an ROC curve that anchors at an X value of 0.25% (25 false negatives is 0.25% of the overall 10,000 data points). All data to the right of there on the graph – with probabilities lower than the cut-off threshold – are irrelevant to the calculation of precision, but represent 98.75% of the data and 99.75% of the X values where area can be determined.

Keeping this base rate and recall but reducing expected precision to 0.5 makes almost no difference. To achieve this, we need to capture 100 true positives and no more than 100 false negatives. This represents only 2.0% of the overall 10,000 data points, creating an ROC curve that anchors at an X value of 1.0% (100 false negatives is 1% of the overall 10,000 data points). All data to the right of there on the graph – with probabilities lower than the cut-off – are irrelevant to the calculation of precision, but represent 98% of the data and 99% of the X values that area can be determined for.

AUC ROC, by design, is focused on model performance across the entire spectrum of probability values and thresholds. It can tell its user how well a model performs overall. Ironically, it is often considered more robust to class imbalance issues than metrics like precision, F1, kappa, or accuracy, as those other metrics are highly impacted by skew [19]. In that paper, they briefly state that "while ROC was unaffected by skew, the precision-recall curves suggest that ROC may mask poor performance in some cases" [19 p.249] but do not offer further detail. Here we see that low precision can occur for high or low AUC ROC – for a low base rate, the two metrics measure very different things.

AUC ROC's properties are relevant for determining if a model is good for some common uses in our community. To be a bit tautological, assessing model performance across the entire spectrum of probability

values and thresholds is good if it is the probability values we care about, and we care equally about the entire spectrum of those values.

As mentioned in the previous section, a common case where that is what we care about is in the use of the model's outputs in discovery with models types of analyses, where we conduct some data mining or statistical procedure using the probabilities. This can include correlation mining, statistical tests, structural equation modeling, and use as covariates or matching in quasi-experimental statistical methods, for instance. It can include using probabilities of a detector as a feature within another detector or model, as well, although the degree to which the limitations of a model either propagate or are controlled for by the follow-on model is still poorly understood. In many such analyses, we care about overall ability to distinguish categories, and care equally about performance across the entire spectrum of probability -- because that information is being used, rather than being most ignored by an intervention selecting only a very small proportion of the cases. In addition, as mentioned in the previous section, AUC ROC may also be appropriate for intervention and reporting approaches that do not involve a single binary threshold.

4 ANALYSIS OF SPECIFICITY

Specificity is another metric sometimes seen in learning analytics and related communities. Specificity measures the proportion of true negative results among all cases where the construct is not present: TN/(TN+FP). Specificity provides valuable information about the ability of a measure to correctly identify cases where the construct is not present. However, it is still vulnerable to differences in base rates, and specificity can be misleading when prevalence is low, since it can remain very high even when precision is very low.

Take for example, our scenario presented earlier where the base rate is 0.02 and the recall is 0.5. If a respectable specificity of 0.8 is achieved, then with 1000 data points, and a base rate of 0.02, there are 20 positives and 980 negatives. Recall of 0.5 means 10 TP and 10 FN. Specificity – TN/(TN+FP) -- of 0.8 mathematically produces a 4:1 ratio which is 784 TN and 196 FP. These numbers give us a precision – TP/(TP+FP) -- of 10/(10+196) which equals 0.049. Hence, both sensitivity and specificity can achieve respectably high values, but when base rate is low, precision can be quite poor.

5 DISCUSSION AND CONCLUSIONS

Some of the most important events and states during a learner's experience may be quite uncommon. But they may be highly important for the learner's eventual outcome. Some of the most widely used metrics paint a murky picture of how well a given model is performing at identifying uncommon and rare constructs. In this paper, we have argued that high precision is essential when delivering high-risk interventions for high-stakes, yet low occurrence events, but also that high precision is very difficult to achieve for constructs with low base rates. We have also demonstrated how another popular metric, AUC ROC, is not informative about model performance for this use case, though – as we discuss – it remains highly meaningful for other use cases.

The information presented in this paper falls into a curious category of knowledge. It is not particularly challenging to discover the properties shown in this paper. Related claims are made in passing in other papers (such as [19 p.249] noting that "precision-recall curves suggest that ROC may mask poor performance in some cases"). Introductory machine learning courses discuss the importance of precision (including popular learning analytics courses [34]).

And yet, popular articles recommending practices for intervention [32] and these same introductory courses do not specifically discuss the challenges presented here and often make recommendations contrary to what is shown here. More importantly, it is not unusual in our field to see papers attempting to develop detectors for high-stakes interventions for low base rate constructs, or failing to report precision when reporting detectors intended for intervention. As such, we present this article to crystallize understanding around these issues, and drive towards more thoughtful consideration of the difficulty of achieving good precision for rare constructs, and the necessity of doing so for certain types of intervention.

Detection is complex, and the quality of a detector is multi-faceted. No single metric captures all aspects of quality. In this article, we argue that a specific metric – precision – is particularly important for high-stakes intervention, and it is particularly difficult for rare constructs. We hope that through considering this carefully, learning analytics researchers and practitioners can clearly discern whether their detector is "good enough" for their desired application by considering precision more often, and can identify cases in advance where it is unlikely that the current state of the art will achieve performance that is good enough for actual use.

ACKNOWLEDGMENTS

We thank Chelsea Porter and ChatGPT for manuscript preparation assistance. This work was supported by LEVI Hub.

REFERENCES

<bib id="bib1"><number>[1]</number>June Ahn, Fabio Campos, Maria Hays, and Daniela DiGiacomo. 2019. Designing in Context: Reaching beyond Usability in Learning Analytics Dashboard Design. Journal of Learning Analytics 6, 2, 70-85.</bib>
<bib id="bib2"><number>[2]</number>Katrien Verbert, Xavier Ochoa, Robin De Croon, Raphael A. Dourado, and Tinne De Laet. 2020. Learning analytics dashboards: The past, the present and the future. In Proceedings of the tenth international conference on learning analytics & knowledge, 35-40.</bib>
<bib id="bib3"><number>[3]</number>Stephen J. Aguilar, Stuart A. Karabenick, Stephanie D. Teasley, and Clare Baek. 2021. Associations between learning analytics dashboard exposure and motivation and self-regulated learning. Computers & Education 162, 104085.</bib>
<bib id="bib4"><number>[4]</number>Stanislav Pozdniakov, Roberto Martinez-Maldonado, Yi-Shan Tsai, Mutlu Cukurova, Tom Bartindale, Peter Chen, Harrison Marshall, Dan Richardson, and Dragan Gasevic. 2022. The question-driven dashboard: how can we design analytics interfaces aligned to teachers' inquiry?. In LAK22: 12th international learning analytics and knowledge conference, 175-185.</bib>
<bib id="bib5"><number>[5]</number>Carolien A.N. Knoop-van Campen, Alyssa Wise, and Inge Molenaar. 2023. The equalizing effect of teacher dashboards on feedback in K-12 classrooms. Interactive Learning Environments 31, 6, 3447-3463.</bib>
<bib id="bib6"><number>[6]</number>Leif Esbenshade, Ryan S. Baker, and Jonathan Vitale. 2023. From a Prediction Model to Meaningful Reports in School. In Education Leadership Data Analytics (ELDA) 2023 Conference. New York, NY.</bib>
<bib id="bib7"><number>[7]</number>Lisa-Angelique Lim, Shane Dawson, Dragan Gašević, Srecko Joksimović, Abelardo Pardo, Anthea Fudge, and Sheridan Gentili. 2021. Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: An exploratory study of four courses. Assessment & Evaluation in Higher Education 46, 3, 339-359.</bib>
<bib id="bib8"><number>[8]</number>Janice D. Gobert, Raha Moussavi, Haiying Li, Michael Sao Pedro, Rachel Dickler. 2018. Real-time scaffolding of students' online data interpretation during inquiry with Inq-ITS using educational data mining. Cyber-physical laboratories in engineering and science education, 191-217.</bib>
<bib id="bib9"><number>[9]</number>Shayan Doroudi. 2020. Mastery learning heuristics and their hidden models. In Artificial Intelligence in Education: 21st International Conference (AIED 2020). Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21, pp. 86-91. Springer International Publishing, 2020.</bib>
<bib id="bib10"><number>[10]</number>Shane Dawson, Leah Macfadyen, Lori Lockyer, and David Mazzochi-Jones. 2011. Using social network metrics to assess the effectiveness of broad based admission practices. Australasian Journal of Educational Technology 27, 1.</bib>
<bib id="bib11"><number>[11]</number>Dragan Gasevic, Jelena Jovanovic, Abelardo Pardo, and Shane Dawson. 2017. Detecting learning strategies with analytics: Links with self-reported measures and academic performance. Journal of Learning Analytics 4, 2, 113-128.</bib>
<bib id="bib12"><number>[12]</number>Seth A. Adjei, Ryan S. Baker, and Vedant Bahel. 2021. Seven-year longitudinal implications of wheel spinning and productive persistence. In International Conference on Artificial Intelligence in Education, 16-28.</bib>
<bib id="bib13"><number>[13]</number>Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI conference on human factors in computing systems, 1-16. 2019.</bib>
<bib id="bib14"><number>[14]</number>Nigel Bosch, Sidney D'Mello, and Caitlin Mills. 2013. What emotions do novices experience during their first computer programming learning session?. In International Conference on Artificial Intelligence in Education, 11-20.</bib>
<bib id="bib15"><number>[15]</number>Kristina Loderer, Reinhard Pekrun, and James C. Lester. 2020 Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments. Learning and instruction 70, 101162.</bib>
<bib id="bib16"><number>[16]</number>Caitlin Mills and Sidney D'Mello. 2015. Toward a Real-Time (Day) Dreamcatcher: Sensor-Free Detection of Mind Wandering during Online Reading. Proc. of the International Conference on Educational Data Mining.</bib>
<bib id="bib17"><number>[17]</number>Ryan S.J.D. Baker, Sidney K. D'Mello, Ma Mercedes T. Rodrigo, and Arthur C. Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. International Journal of Human-Computer Studies 68, 4, 223-241.</bib>
<bib id="bib18"><number>[18]</number>Ryan S. Baker, Nigel Bosch, Stephen Hutt, Andres F. Zambrano, and Alex J. Bowers. 2024. On Fixing the Right Problems in Predictive Analytics: AUC Is Not the Problem. Technical Report. Philadelphia, PA: University of Pennsylvania Center for Learning Analytics.</bib>
<bib id="bib19"><number>[19]</number>László A. Jeni, Jeffrey F. Cohn, and Fernando De La Torre. 2013. Facing imbalanced data--recommendations for the use of performance metrics. In 2013 Humaine association conference on affective computing and intelligent interaction. IEEE, 245-251.</bib>
<bib id="bib20"><number>[20]</number>Nigel Bosch, Luc Paquette. 2018. Metrics for discrete student models: Chance levels, comparisons, and use cases. J. Learning Analytics 5, 2, 86-104.</bib>
<bib id="bib21"><number>[21]</number>Fredys Simanca, Rubén Gonzalez Crespo, Luis Rodríguez-Baena, and Daniel Burgos. 2019. Identifying students at risk of failing a subject by using learning analytics for subsequent customised tutoring. Applied Sciences 9, 3, 448.</bib>

<bib id="bib22"><number>[22]</number>Ryan S.J.D. Baker, Albert T. Corbett, Kenneth R. Koedinger, Shelley Evenson, Ido Roll, Angela Z. Wagner, Meghan Naim, Jay Raspat, Daniel J. Baker, and Joseph E. Beck. 2006. Adapting to when students game an intelligent tutoring system. In Proceedings of Intelligent Tutoring Systems.</bib>
<bib id="bib23"><number>[23]</number>Luc Paquette and Ryan S. Baker. 2017. Variations of gaming behaviors across populations of students and across learning environments. In Proceedings of Artificial Intelligence in Education: 18th International Conference (AIED 2017). Wuhan, China, June 28–July 1, 2017, 274-286.</bib>
<bib id="bib24"><number>[24]</number>Nigel Bosch, Sidney K. D'Mello, Ryan S. Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2016. Detecting student emotions in computer-enabled classrooms. In IJCAI, vol. 16, 4125-4129.</bib>
<bib id="bib25"><number>[25]</number>NaYoung Hwang. 2018. Suspensions and achievement: Varying links by type, frequency, and subgroup. Educational Researcher 47, 6, 363-374.</bib>
<bib id="bib26"><number>[26]</number>Douglas Selent, Thanaporn Patikorn, and Neil Heffernan. 2016. Assistments dataset from multiple randomized controlled experiments. In Proceedings of the Third (2016) ACM Conference on Learning@ Scale, 181-184.</bib>
<bib id="bib27"><number>[27]</number>Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. 2017. Open university learning analytics dataset. Scientific data 4, 1, 1-8.</bib>
<bib id="bib28"><number>[28]</number>Zadjiichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E. Turne, Richard G. Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. 2020. Instructions and guide for diagnostic questions: The neurips 2020 education challenge. arXiv preprint arXiv:2007.12061 .</bib>
<bib id="bib29"><number>[29]</number>Iryna Nikolayeva, Amel Yessad, Bertrand Laforge, and Vanda Luengo. 2020. Does an e-mail reminder intervention with learning analytics reduce procrastination in a blended university course?. In Addressing Global Challenges and Quality Education: 15th European Conference on Technology Enhanced Learning (EC-TEL 2020). Heidelberg, Germany, September 14–18, 2020, Proceedings 15, 60-73.</bib>
<bib id="bib30"><number>[30]</number>Sidney D'Mello, Blair Lehman, Jeremiah Sullins, Rosaire Daigle, Rebekah Combs, Kimberly Vogt, Lydia Perkins, and Art Graesser. 2010. A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In Proc. of Intelligent Tutoring Systems, 245-254.</bib>
<bib id="bib31"><number>[31]</number>Ryan Baker, Kalina Yacef. 2009. The state of educational data mining in 2009: A review and future visions. J. Educational Data Mining 1, 1, 3-17.</bib>
<bib id="bib32"><number>[32]</number>Alex J. Bowers and Xiaoliang Zhou. 2019. Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. Journal of Education for Students Placed at Risk (JESPAR) 24, 1, 20-46.</bib>
<bib id="bib33"><number>[33]</number>James A. Hanley and Barbara J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 1, 29-36.</bib>
<bib id="bib34"><number>[34]</number>Ryan S. Baker. 2024. Big Data and Education. 8th Edition. Philadelphia, PA: University of Pennsylvania.</bib>
<bib id="bib35"><number>[35]</number>Russell W. Rumberger and Scott L. Thomas. 2000. The distribution of dropout and turnover rates among urban and suburban high schools. Sociology of education, 39-67.</bib>