Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1081

Leveraging Modern Machine Learning to Improve Early Warning Systems and Reduce Chronic Absenteeism in Early Childhood

Tiffany Wu
University of Michigan

Christina Weiland
University of Michigan

This study focuses on improving the predictive power of early warning systems (EWSs) to decrease chronic absenteeism in early childhood. Using a demographically diverse sample of students followed from PreK to third grade in Boston Public Schools (N=6,698), we demonstrate how and why two modern machine learning algorithms—the Synthetic Minority Oversampling Technique (SMOTE) and Extreme Gradient Boosting (XGBoost)—can enhance EWS accuracy. The best-performing XGBoost model with SMOTE achieved a 54 percentage point improvement in accuracy (in terms of recall rate) over the logistic regression model closest to those used in current EWSs, more accurately identifying students would become chronically absent in third grade. Notably, models excluding student demographic information maintained comparable predictive accuracy.

VERSION: August 2025

Suggested citation: Wu, Tiffany, and Christina Weiland. (2025). Leveraging Modern Machine Learning to Improve Early Warning Systems and Reduce Chronic Absenteeism in Early Childhood. (EdWorkingPaper: 24-1081). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/xxvz-cv94

Leveraging Modern Machine Learning to Improve Early Warning Systems and Reduce Chronic Absenteeism in Early Childhood

Tiffany Wu
University of Michigan

Christina Weiland
University of Michigan

Authors' note: This study is funded by the Institute of Education Sciences, grants R305A220036 and R305B200011. The authors thank the Boston Public Schools (BPS), TeeAra Dias, Jason Sachs, the BPS Department of Early Childhood coaches and staff (particularly Yuzhu Xia), the BPS Office of Data and Accountability (particularly Mariana Ronchini, Megan Toney, and Apryl Clarkson), and the Massachusetts Department of Elementary and Secondary Education (particularly Elana McDermott, Brendan Longe, and Kate Sandel) for their support and partnership. Special thanks also to Rebecca Unterman, Anna Shapiro, and Annie Taylor for their support on this dataset, and Hyunwoo Jang, Tianliang Xu, and Chaewon Lim for their early insights helping shape the direction of this study. Lastly, we are grateful to the participants of the University of Michigan's Causal Inference in Education Research Seminar (particularly Brian Jacob and Jordy Berne), Allison Ryan, Kevin Stange, Ben Hansen, Annaliese Paulson, Amanda Weissman, and Salar Fattahi for their helpful feedback. Correspondence concerning this article should be addressed to Tiffany Wu, School of Education, 610 E. University Ave, Ann Arbor, MI 48104, or via e-mail at wutiffa@umich.edu.

Abstract

This study focuses on improving the predictive power of early warning systems (EWSs) to decrease chronic absenteeism in early childhood. Using a demographically diverse sample of students followed from PreK to third grade in Boston Public Schools (*N*=6,698), we demonstrate how and why two modern machine learning algorithms—the Synthetic Minority Oversampling Technique (SMOTE) and Extreme Gradient Boosting (XGBoost)—can enhance EWS accuracy. The best-performing XGBoost model with SMOTE achieved a 54 percentage point improvement in accuracy (in terms of recall rate) over the logistic regression model closest to those used in current EWSs, more accurately identifying students would become chronically absent in third grade. Notably, models excluding student demographic information maintained comparable predictive accuracy.

Keywords: chronic absenteeism, machine learning, early warning system, XGBoost, SMOTE, fairness

Leveraging Modern Machine Learning to Improve Early Warning Systems and Reduce Chronic Absenteeism in Early Childhood

In 2016, the U.S. Department of Education sounded the alarm on chronic absenteeism, labeling it the "hidden educational crisis" (U.S. Department of Education, 2016a). Fast forward to today, this crisis has been thrust into the spotlight due to the unprecedented disruption caused by the Covid-19 pandemic (The White House, 2023; Mervosh, 2023; Oliver, 2023). Numerous correlational studies have highlighted the consequences of chronic absenteeism, linking it to lower academic achievement, diminished socioemotional skills, and an increased likelihood of high school dropout, even when accounting for confounding variables like family income and race (Allensworth et al., 2021; Gottfried, 2014; Romero & Lee, 2007). Despite sustained efforts by researchers, school practitioners, and policymakers over the past two decades, problems with school attendance have persisted (Jacob & Lovett, 2017). However, the integration of modern machine learning methods offers a promising opportunity to enhance our strategies for addressing this deeply entrenched problem.

This study focuses on using modern machine learning (ML) algorithms to improve a key system already in place in many school districts—early warning systems (EWSs)—with the goal of reducing students' risk of becoming chronically absent as early as prekindergarten. Although EWSs have the potential to proactively identify students at risk of chronic absenteeism and facilitate timely supports, especially in the earliest grades, they are often underutilized for this purpose since many EWSs were designed to predict high school dropout (Balfanz & Byrnes, 2019). Furthermore, the effectiveness of current EWS models may be hampered by analytical challenges and by the uncertain impact of including or excluding student demographic information, since these choices can shape predictive accuracy in ways that are not well

understood (Gándara et al., 2024; Sansone, 2019; Yu et al., 2021). These limitations represent a missed opportunity, as early childhood is a critical window for establishing positive attendance patterns and represents a more malleable point in a student's life for intervention (Ansari & Gottfried, 2021; Heckman, 2008). Modern ML techniques hold the potential to overcome these limitations. While ML methods have been commonly employed in other disciplines for classification and prediction, the field of education has been slower to adopt these methods (Weissman, 2022). In this study, we leverage these advanced algorithms, demonstrating how and why they can be used to improve the ability of EWSs to provide more accurate predictions to reduce chronic absenteeism during the earliest years of schooling.

We first give an overview of chronic absenteeism. Then, using a demographically diverse population of students followed from prekindergarten to third grade in the Boston Public Schools, we demonstrate the application of two modern ML algorithms—the Synthetic Minority Oversampling Technique (SMOTE; Chawla et al., 2002) and Extreme Gradient Boosting (XGBoost; Chen & Guestrin, 2016)—for improving EWS accuracy. The analyses show the increase in predictive accuracy each additional year of data offers, addressing a policy-relevant question of how much historical data is needed to make effective intervention decisions in early childhood. We examine whether the inclusion of student demographic information meaningfully improves prediction, and we explain how incorporating varying probability thresholds in our models can cater to districts aiming to leverage machine learning predictions for early intervention while adhering to budgetary or intervention constraints. Overall, we hope this paper can serve as a gentle, clear introduction to machine learning for education researchers hoping to incorporate advanced predictive analytics into their research.

Overview of Chronic Absenteeism

What is Chronic Absenteeism?

Chronic absenteeism is most commonly defined as missing 10% or more of school days for any reason (Allison et al., 2019; Faria et al., 2017). This means that excused absences, unexcused absences, and any days a student may miss for being suspended all count as part of this metric. This commonly coincides with missing at least 18 total school days in the U.S. (i.e., 18 is 10% of a 180-day school year).

In recent years, addressing absenteeism has risen in priority for education policymakers. For the first time in 2014, the U.S. Department of Education's Office for Civil Rights asked schools to report how many students missed 15 or more days of school for its survey. In 2015, the Obama administration announced Every Student, Every Day: A National Initiative to Address and Eliminate Chronic Absenteeism with the goals of better attendance monitoring and decreased rates of chronic absenteeism. The Every Student Succeeds Act (ESSA) of 2015 led many states to redefine how to measure school accountability. By 2018, 36 states and the District of Columbia approved ESSA plans to incorporate school-level chronic absenteeism as an indicator of school performance (Swaak, 2018). In January of 2024, the Biden-Harris administration laid out the Improving Student Achievement Agenda, which emphasized increasing attendance as one of their top three education priorities (The White House, 2024).

One of the main reasons it has been so difficult to improve attendance rates is that many of the factors associated with student absenteeism are rooted in systems of inequity. Studies trying to identify the factors driving absenteeism have often linked individual and family characteristics, such as children's race and socioeconomic status, with higher rates of student absenteeism (Allensworth et al., 2021; Klein et al., 2020; Purtell & Ansari, 2022). However, although demographic factors play a crucial role in understanding patterns of student

absenteeism, they are generally not malleable, or at least not easily changed in the short term, making them less actionable for school-based intervention efforts. Schools and districts therefore face limitations to effectively intervening within the constraints of systemic inequity (Gottfried & Hutt, 2019; Kearney et al., 2019).

Early Warning Systems (EWSs)

As a result, one strand of attendance research focuses on discovering strategies individual schools and districts can take in diminishing their own rates of absenteeism (Gottfried & Hutt, 2019). Many districts and states have embraced the implementation of an EWS, a promising, low-cost tool that uses key indicators to identify students at risk of not meeting certain milestones (U.S. Department of Education, 2016b). Until the mid-2010s, most EWSs relied on threshold-based models which would flag a student only when they surpassed a preset threshold, such as missing more than 18 days of school (Christie et al., 2019). In recent years, prediction-based EWSs have become increasingly common since they can proactively predict a student's risk level and identify them for intervention before thresholds are crossed. These prediction-based EWSs are the focus of our study. While these EWSs have been used primarily for predicting the risk of high school dropout, they also hold potential for reducing chronic absenteeism (Balfanz & Byrnes, 2019; Christie et al., 2019).

Despite their promise, current EWSs face notable limitations. Methodologically, EWSs typically rely on risk levels calculated from traditional regression models (Allensworth & Easton, 2007; OECD, 2020; Sansone, 2019). Traditional regression models used in EWSs mostly entail linear and parametric methods, which may not be well-suited for predictions because they typically assume relationships between variables are linear and predefined by the model structure. They also make assumptions about the underlying data such as independence of

observations and no multicollinearity, assumptions that may be violated when using real-world data. Even with interaction terms introducing non-linearity, traditional regressions tend to have poor predictive accuracy compared to more modern machine learning algorithms (Deussen et al., 2017; Sansone, 2019).

These methodological constraints underscore the need for more flexible, non-linear machine learning algorithms for EWSs, algorithms which can more effectively capture the intricate interactions among multiple variables and provide more accurate predictions. A few studies have explored the use of classification and regression tree analysis (CART; Fuchs et al., 2008; Fuchs et al., 2007) or boosting models (Lee & Chung, 2019; Sansone, 2019), but these nonparametric methods remain underutilized in EWSs, especially in the context of early childhood. Additionally, traditional regression models often struggle with class imbalance, a significant challenge when predicting chronic absenteeism due to the disparity in the number of students considered at higher versus lower risk. This imbalance hampers the regression model's ability to accurately predict the minority of students who are genuinely at risk and is best addressed through non-traditional machine learning algorithms, but once again, these methods are currently underutilized (Lee & Chung, 2019).

In addition to methodological limitations, current EWSs are predominantly designed to identify students at risk of high school dropout rather than to pinpoint younger students at risk of absenteeism (Sansone, 2019; Lee & Chung, 2019; Faria et al., 2017). Part of the reason for this is the dearth of data available in early childhood (Ehrlich et al., 2018). EWSs usually rely on a set of strong predictors. While this can differ by school, commonly used predictors include attendance (indicators for missing a preset number of school days), behavior (indicators for being suspended or expelled), and course performance (indicators for failing core courses), which have

become known as the ABC indicators, along with student demographic information (U.S. Department of Education, 2016b). PreK to second grade students may not receive traditional A-F course grades, and their suspension or expulsion rates are substantially lower, with some districts prohibiting such disciplinary actions for younger students (Jacobsen et al., 2019). The limited number of ABC indicators in the early years could consequently result in less reliable predictions, but more research is needed examining the predictive accuracy of early EWSs.

Relatedly, concerns about fairness and equity have become increasingly salient in predictive modeling, particularly when models incorporate demographic features (Baker et al., 2023; Gándara et al., 2025; Yu et al., 2021). Proponents of "fairness through awareness" (Dwork et al., 2012) have argued that incorporating demographic variables can improve predictive accuracy by capturing structural inequalities that disproportionately affect students from historically marginalized groups. Critics, however, caution that such models often reproduce inequities by disproportionately predicting less favorable outcomes for these students (Baker et al., 2023; Gándara et al., 2025). In the context of chronic absenteeism, demographic characteristics are often statistically significant predictors (Ehrlich et al., 2018; Purtell & Ansari, 2022), but because they are not actionable for school-based interventions, their inclusion in an EWS model may shift focus away from supports that schools can directly provide. Attendance rate is the only predictor consistently tracked from early childhood and potentially responsive to school- or district-level intervention. Whether absenteeism EWSs can maintain predictive power while relying solely on early attendance remains an open question; if they can, this approach could allow schools to target supports to factors more within their control while reducing the risk of reinforcing existing disparities.

The research on EWS performance in the early grades remains limited, making it difficult to assess their predictive power when used with the limited data available from students' earliest school years. However, the development of attendance habits starts early, and focusing predominantly on EWSs in high school obscures the critical influence of children's earliest experiences with schooling and absences (Allensworth et al., 2021; Wei, 2024). Since the earliest years are a more malleable point in a student's life for intervention, it would be ideal to begin absenteeism interventions early (Heckman, 2008), and improved EWSs in early childhood could aid in this task.

Attendance in Boston Public Schools and the State of Massachusetts

The Boston Public Schools (BPS) student attendance policy was first established in the 1998-1999 school year, and BPS has worked in past years to update its policy and make it as equitable as possible. When the Every Student Succeeds Act (ESSA) was signed into law in 2015, Massachusetts was one of the 36 states that included chronic absenteeism as a core indicator in its school accountability index, and BPS updated its attendance policy to reflect this (Boston Public Schools, 2022). Under BPS's attendance policy, a student must be at school for at least half the day in order to be counted as "present." A half day is three hours in elementary school, three hours and five minutes in middle school, and three hours and ten minutes in high school. Chronic absenteeism is defined by BPS as missing 10%, or the equivalent of 18 school days, or more of the school year (Boston Public Schools, 2022). BPS currently requires all schools to create a truancy prevention and attendance-promoting plan.

Massachusetts has emphasized the importance of collaboration among families, schools, and the Department of Elementary and Secondary Education (DESE) to improve student attendance. The state provides guidance on attendance policies to support school districts and

educators in promoting consistent engagement. However, in its predictive efforts, Massachusetts has taken a broader academic focus. Its Early Warning Indicator System (EWIS), established in 2011, is designed to identify students who may require additional supports to achieve key academic milestones between 1st and 12th grade (Massachusetts Department of Elementary and Secondary Education, 2023). For early elementary students in first to third grade, the academic milestone is meeting or surpassing expectations on the 3rd-grade ELA Massachusetts State Assessment.

EWIS risk levels are determined through an annually updated multilevel logistic regression model using student demographics, enrollment, attendance, and suspension indicators sourced from existing state-wide collections (Massachusetts Department of Elementary and Secondary Education & American Institutes for Research, 2013; OECD, 2020). While the EWIS is one of the only data-based information systems that uses a statistical model instead of individual indicators, it still faces the common aforementioned limitations associated with traditional linear regression predictions, and there has been no evaluation of how the inclusion of demographic variables affects the accuracy of the model's predictions. Furthermore, a 2019 case study found that understanding of the EWIS among educators and school officials could be further improved, as many were not familiar with how the data system could support their work (OECD, 2020). Notably, the EWIS model does not currently provide specific predictions for students' risk of chronic absenteeism. While focusing on reading proficiency in third grade is important and valuable, it may inadvertently overlook the foundational drivers of academic struggles, such as chronic absenteeism. Chronic absenteeism impacts learning across all subjects and is perhaps the indicator most strongly connected to the development of future early warning indicators like high school dropout (Allensworth et al., 2021; Balfanz & Byrnes, 2019). By

focusing on early identification of absenteeism alongside outcomes like reading proficiency levels, schools may be able to better tackle foundational barriers to learning.

Our study explores a modern machine learning-based EWS that addresses the methodological limitations of the traditional regression models used in current EWSs and examines how accurate EWS algorithms can be for early chronic absenteeism detection. We do not explicitly address the broader issues of EWS implementation within schools, including organizational structure, administrative support, staff training, and the development of effective intervention strategies after students are identified. The practical implementation of EWSs in schools will require a separate, multifaceted effort involving collaboration among stakeholders beyond the scope of this study. Instead, we strive to advance the reliability and accuracy of the EWS algorithm, making it a more valuable and appealing tool for school practitioners to use. When predictions are accurate and actionable, schools can more confidently allocate resources and interventions to the right students at the right time, which boosts the perceived value of the system. This, in turn, fosters greater buy-in among educators.

Present Study

The present study demonstrates the efficacy of modern machine learning techniques, specifically the Synthetic Minority Oversampling Technique (SMOTE) and Extreme Gradient Boosting (XGBoost), in refining EWSs to proactively identify students who, without intervention, have a heightened risk of chronic absenteeism during their early schooling years. In particular, we aim to answer the following research questions:

1. How does the prediction accuracy from using modern machine learning algorithms (like SMOTE and XGBoost) compare to that from more traditional parametric methods (like logistic regression) for use as a proactive early warning system? What

algorithmic structures contribute to performance differences between these approaches?

2. How accurately and early can we identify students who will be chronically absent in 3rd grade? What features, or predictors, contribute most to model predictions?
3. To what extent does student demographic information contribute to overall model performance, and can accurate predictions be achieved using only attendance rates as predictors?
4. How can models be personalized to inform chronic absenteeism intervention while taking into account an institution's financial and resource constraints?

Method

Sample

Our sample for this paper was the population of students who enrolled in the Boston Public Schools (BPS) PreK program for four-year-olds between the 2007-2008 and 2010-2011 school years. The BPS PreK program is a large-scale early childhood education program based entirely in the public schools during our study years (Kabay et al., 2020). We followed students for each focal cohort from their application to PreK to third grade. We combined PreK students from all four cohorts (2007-2010) into our final sample because the third-grade attendance rate distributions for each cohort looked similar (Appendix S1). A total of 12,740 families applied to the BPS PreK program during these four years. For this paper, we included students who got accepted into a BPS PreK program and ended up enrolling in it, attended BPS PreK for greater than or equal to 90 days (out of 180 total school days) in the school they attended the most, and were enrolled in the 3rd grade elementary school they attended the most for 90 or more days. We restricted the sample to only include students who attended for 90 or more days following the

practice set forth by other researchers exploring absenteeism (Weissman, 2022; Chang & Romero, 2008), and we only included students who enrolled in BPS PreK to ensure continuity of student data from PreK to third grade.

We ended with a final analytic sample of 6,698 students. Sample descriptives are in Table 1. On average, our final sample was 51% male and racially diverse (17% White, 28% Black, 42% Hispanic/Latino, 9% Asian, 3% multiracial or other). Almost half of the sample (44%) identified as a dual language learner, 71% were eligible for free or reduced-price lunch, and 17% were eligible for special education services. Of our sample, 6,066 students were not chronically absent in 3rd grade while 632 were. There were statistically significant differences between these two groups for race/ethnicity, eligibility for free or reduced-price lunch and special education services, and chronic absenteeism rates in past school years. Our study subsample contained more students who qualified for free/reduced lunch (71% in study sample compared to 65% in full sample) and more special education students (17% in study sample compared to 13% in full sample) compared to the full sample of 12,740 students. Compared to the broader population of all students in BPS during our study years, our sample includes a lower proportion of Black students (28% vs. ~37%), students who qualified for special education services (17% vs. ~20%), and students who were eligible for free or reduced-price lunch (71% vs. 74%). Our sample includes a slightly higher proportion of Hispanic/Latino students (42% vs. ~39%), White students (17% vs. ~13%), and those who are dual language learners (44% vs. ~40%). A table of student characteristics for the full sample and broader BPS student population is in Appendix S2.

Outcome Variable

The aim for all our models was to predict which students would be chronically absent in third grade. This outcome was a binary variable equal to one if the student was chronically

absent in third grade and a zero if they were not. A student was counted as chronically absent if they had an absence rate of 10% or more during their third-grade year (Allison et al., 2019; Faria et al., 2017). Absence rate was calculated by dividing the number of days a student was absent by the total number of days they were enrolled.

Predictor Variables

We chose predictor variables based on data that school districts in Massachusetts already collected to help our analyses be more easily replicable and accessible to other schools. Our time-varying predictors from PreK to second grade included the attendance rate, number of retentions, number of suspensions, whether the student was eligible for free/reduced priced lunch, whether the student was in special education, and the school attended for the given school year prior to third grade. Additionally, we included a set of student-level covariates using administrative records. We captured students' race/ethnicity using a set of binary variables that identified whether a student was Black, Hispanic, Asian, White, or multiracial/other. We also created binary variables for whether the student was a dual language learner, whether the student was female or male, and which focal cohort the student was in. Not all predictors were used in every model, as we detail in our analytical approach. Additional details for each predictor variable are in Appendix S3.

Machine Learning Methods

We give a brief introduction of the ML algorithms used in our study below. Analyses for implementing the modern machine learning algorithms SMOTE and XGBoost were conducted in Python version 3.9.13, and sample code to run each algorithm is provided in Appendix S8.

Supervised Learning

We focus on the branch of ML called supervised learning in this study. Supervised learning is the machine learning approach that involves training a statistical model using a labeled dataset that contains both dependent and independent variables (Hastie et al., 2009). In our study, each student observation is associated with the outcome variable, third grade chronic absenteeism, in addition to a set of predictor variables. Typically, the outcome variable for supervised learning models is a binary variable like whether a student is chronically absent or not. The goal of supervised learning models is to train a model to predict, or classify, the outcome variable as accurately as possible for new, unseen data. For this reason, supervised learning models are oftentimes called *classifiers* as well. For example, we would like to train a classifier to predict third grade absenteeism in another cohort of students outside our sample.

The most common machine learning classifiers minimize the difference between the predicted and actual values of the outcome by adjusting the model's parameters, like the method of least squares commonly used in linear regressions (Hastie et al., 2009). In fact, traditional linear (ordinary least squares) and logistic regressions both fall into the category of supervised learning. Beyond traditional regressions, more modern machine learning algorithms include ensemble methods like boosting techniques, which we will explain in the XGBoost section.

In order to mitigate overfitting and enhance our classifier's generalizability, it is common in machine learning to divide the data we have into a training set and a test set. We first train our model on the training set, and then we use the data from the testing set to gauge the accuracy of the resulting model. Research indicates optimal outcomes when dedicating 70-80% of the data for training purposes and allocating the remaining 20-30% of the data for testing (Gholamy et al., 2018). Since the 80/20 split is more common, that is how we split our data in this study. As a robustness check, we also implemented a rolling-origin validation approach, where the models

are trained on the first three cohorts and tested on the last cohort. We reached the same conclusion using this splitting method (Appendix S4).

Performance Metrics

How well a supervised learning model performs is determined by how accurate it is. Many popular performance metrics are based on a confusion matrix (Table 2) that gives the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values based on the whether the true and predicted outcome labels match. An example of a true positive (TP) is if a student was chronically absent in third grade (actual label equals 1) and were also predicted by our machine learning model to be chronically absent in third grade (predicted label equals 1).

These confusion matrix values can then be combined to define many performance metrics. Accuracy rate ((TP+TN)/(TP+TN+FP+FN)) is the most common metric used to evaluate a model and gives the number of labels correctly predicted by the model out of the total number of observations. For example, if our model correctly predicts the chronic absenteeism label of 3,000 students out of a total of 6,000 students, then our accuracy rate is 50%. However, accuracy rate can be misleading in imbalanced datasets, as we will later explain.

Beyond accuracy, there are more nuanced metrics we can look at. The *recall* or *true positive rate* (TPR; TP/(TP+FN)) tells us the percentage of labels our model predicted correctly out of all the students who were actually chronically absent in 3rd grade. This metric is especially helpful for policies and interventions since it focuses on the identification of chronically absent students. The *specificity* or *true negative rate* (TNR; TN/(FP+TN)) tells us the percentage of labels our model predicted correctly out of all the students who were not chronically absent in 3rd grade. The *balanced error rate* (BER; 1-0.5*(TPR+TNR)) takes into account both the TPR and TNR, and considers the tradeoff between the model's ability to classify both chronically absent

and non-chronically absent students. This metric is helpful in cases of class imbalance, which we explain below, and provides a more balanced assessment of the model's overall effectiveness across different classes.

Besides accuracy rate, area under the curve (AUC) is the second most common metric reported for machine learning models. It is often used in conjunction with the Receiver Operating Characteristic (ROC) curve. The ROC curve (Figure 1) is a graphical representation of the trade-off between the True Positive Rate (TP/(TP+FN)) and the False Positive Rate (1-TN/(FP+TN)). The AUC is a single value that summarizes the overall performance of the model represented by the ROC curve. It measures the area under the ROC curve, hence its name, and ranges from 0 to 1, where a value of 1 indicates a perfect classifier (the model makes no prediction mistakes) and a value of 0.5 represents a completely random model (the model's predictions are as good as guessing). In other words, the higher the AUC number, the better the model performance.

Class Imbalance & SMOTE

Since we have many more students who were not chronically absent in third grade (N=6,066) compared to those who were (N=632), we have a *class imbalance* problem. In machine learning, "class" refers to the categorical labels (chronically absent or not) our models predict. The classes are imbalanced when there are many more observations fitting into one class than another. Class imbalance poses a challenge because models trained on an imbalanced dataset tend to have high predictive accuracy for the majority class (students who are not chronically absent) but low predictive accuracy for the minority class (students who are chronically absent). This is a problem because our goal is to identify students who will be in the

minority class. This bias toward the majority class occurs because there are so few instances of the minority class that the model treats these observations as outliers or noise.

Table 3 demonstrates why this is a problem when assessing the accuracy of our models, based on an example by Lee and Chung (2019). The confusion matrix shows an imbalanced distribution of students who are chronically absent (10 students) and not chronically absent (990 students). The hypothetical model predicts nobody will be chronically absent, but remarkably, the model's accuracy rate is 99% (990/1000), concealing its misclassification of all the chronically absent students. That is, Table 3 portrays the case where a model can be deceptively accurate but neglects the misclassification of the minority class. Since our objective is to identify chronically absent students, it is necessary to find a way to properly handle class imbalance when building a predictive model. While not much attention has been paid to class imbalance when developing early warning systems, this problem is well-known in the machine learning community (Lee & Chung, 2019). Therefore, we can borrow methods that have already been proposed to address class imbalance.

Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al., 2002) is a resampling technique widely employed in machine learning to address class imbalance. Rather than duplicating existing minority class observations, which risks overfitting and does not help the model learn broader patterns that characterize the minority class, SMOTE generates synthetic yet plausible examples that lie between real observations. This encourages the model to learn a more generalizable decision boundary that better separates the minority from the majority class.

Figure 2 illustrates the SMOTE process in a hypothetical two-dimensional feature space. Before SMOTE (left panel), the chronically absent X’s are sparse because of the class imbalance. This makes it difficult to classify these points. In the middle panel, the algorithm randomly

identifies a chronically absent student (the first circled X) and selects one of its neighboring points using an algorithm called *k* nearest neighbors (the second circled X). Then, it creates a new synthetic instance (the big orange X) along the line segment connecting the two points. After repeating this process for many minority-class examples, the right panel shows how the chronically absent observations are more evenly distributed across the feature space. After this process, the circular boundary line separating the chronically absent from non-chronically absent students becomes clearer. We will run our models both with and without using SMOTE to demonstrate the importance of addressing class imbalance.

Logistic Regression

We used a logistic regression model as our base comparison model because logistic regressions are often the model of choice for education researchers when analyzing a binary outcome and are the statistical model most commonly used in existing EWSs (Allensworth & Easton, 2007; OECD, 2020; Peng et al., 2002). Logistic regression is a type of parametric model because it makes assumptions about the underlying data, including the independence of observations and linearity in the logit. Because of this, logistic regressions tend to be inflexible and make a linear classification boundary line unless interactions are included.

We first fit the following multilevel, or mixed-effects, logistic regression model, separately for each school year:

$$\text{logit}\left[P\left(Y_{ij}=1\right)\right]=\beta_0+X_{ij}\beta+u_j,\quad u_j\sim N(0,\sigma_u^2)$$

where *i* denotes student and *j* denotes school. Y_{ij} is a binary outcome indicating whether student *i* in school *j* was chronically absent in third grade. β_0 is the overall intercept for the model. X_{ij} is a vector of fixed-effect predictors measured at the student level, and β is the corresponding vector of coefficients. u_j is the random intercept for school *j.* We included random intercepts to account

for the nesting of students within schools and assume that the school random intercepts follow a normal distribution, where σ_u^2 represents the between-school variance in the log-odds of chronic absenteeism. We chose this model specifically because it is the closest to what Massachusetts currently uses in their EWS (Massachusetts Department of Elementary and Secondary Education & American Institutes for Research, 2013). We then also ran models with predictors from all the school years together, a model interacting student demographic characteristics, and a model using only attendance rate without school random intercepts.

XGBoost

XGBoost (Extreme Gradient Boosting; Chen & Guestrin, 2016) is a relatively new machine learning method that has quickly gained popularity among data scientists for building predictive models. According to the Kaggle State of Data Science Survey 2021, nearly 50% of respondents reported using XGBoost, and XGBoost has been the winning model in a majority of Kaggle competitions (Kaggle, 2021). Despite its widespread success in the data science domain, XGBoost has yet to attain comparable popularity or integration in education research. To the best of our knowledge, it has only been previously applied in one other study as a potential EWS algorithm, which focused on predicting high school dropout (Christie et al., 2019). Thus, there remains a gap in understanding XGBoost's potential for enhancing early-grade EWSs.

XGBoost is an ensemble method that creates a sequence of simple classifier models (usually decision trees) that correct the mistakes of the models before it. Ensemble methods are those that combine multiple machine learning algorithms (Zhou, 2012). Analogous to assembling a team of specialists with distinct proficiencies is various domains, ensemble methods amalgamate predictions from simpler models to arrive at more accurate predictions. The central idea hinges on harnessing the diversity of these simpler models and orchestrating their

predictions in a way that capitalizes on their respective strengths while compensating for their weaknesses through iterative refinement.

The objective function (loss function and regularization) that XGBoost minimizes at each iteration *t* is the following:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

where y_i is the true label value (0 if not chronically absent, 1 if chronically absent), $\hat{y_i}$ is the predicted label value at a given iteration *t*-1, and $f_t(x_i)$ is the correction term for each data point at a given iteration. The function *l* measures the dissimilarity between y_i and the corrected predicted label $\hat{y_i}^{(t-1)} + f_t(x_i)$. A small value for *l* means that the corrected predicted label is closer to y_i, leading to better accuracy. The regularization term $\Omega(f_t)$ prevents the trained model from overfitting to the data. The incorporation of this regularization term is especially helpful when the data is scarce or corrupted with noise (Chen & Guestrin, 2016).

Figure 3 illustrates how XGBoost works conceptually, adapted from visualizations by Shah (2020). While a bit oversimplistic and resembling existing visualizations of another boosting method called AdaBoost, it conveys the intuition behind how successive simple classifiers correct the errors of preceding models and why XGBoost, along with other boosting methods, are such powerful predictive tools. At Iteration 1, we see the feature space of the original training dataset, with blue circles representing non-chronically absent students and orange x's representing chronically absent students. The first classifier, a decision tree, is created by making a simple horizontal split, represented by the dotted line. Students above the line are predicted to be blue, and those below the line are predicted to be orange. Misclassified points are circled.

In subsequent iterations, XGBoost refines the model using the errors, or residuals, from the previous iteration. Specifically, the algorithm will fit new decision trees to predict the residuals. This means that, at each iteration, XGBoost attempts to minimize the residuals from the previous model by focusing on the patterns of the errors. For example, in the second iteration, the algorithm may make a vertical split, trying to correctly classify points that were previously misclassified. This process continues iteratively, with each new tree correcting residual errors from the previous one. Once all iterations are complete, XGBoost combines the predictions from all the individual decision trees into a final prediction through a weighted sum of the outputs, scaled by how much each tree's prediction contributes to the final model.

This boosting approach, along with XGBoost's regularization techniques, makes it highly powerful and effective in many prediction tasks. The iterative progression of XGBoost makes it well-suited for modelling non-linear relationships and complicated interactions in the data. XGBoost can also easily incorporate predictors from multiple school years together in the same model. These models are more robust against overfitting and outliers due to their ability to combine models and adjust hyperparameters.

Although XGBoost does not produce traditional coefficients, we can use SHAP (SHapley Additive exPlanations; Lundberg & Lee, 2017) values to offer practical interpretability and quantify each predictor's (or feature's) contribution. SHAP values are grounded in cooperative game theory's fair allocation principles and indicate how much each predictor shifts an individual student's predicted risk of chronic absenteeism upward or downward relative to the baseline risk (the average risk across all students). They are calculated by averaging a feature's marginal contribution to the prediction across all permutations of input features. For example, to compute the SHAP value for attendance rate, SHAP computes how much attendance contributes,

on average, to a student's prediction, across all the different ways it could have been added to the model with the other features. This approach allows us to generate SHAP beeswarm plots, which visually summarize three key elements: feature importance (which predictors had the greatest average influence), direction of effect (whether a feature increased or decreased predicted risk), and distribution of effects (how much a feature's influence varied across students). This enhances the transparency and understanding of the XGBoost predictions.

Importantly, in preliminary analyses, we benchmarked XGBoost against several alternative machine learning algorithms, including random forest, linear and radial basis function support vector machines, AdaBoost, LightGBM, lasso regression, and a stacked lasso and random forest model. These models were chosen because they are widely regarded as strong-performing classifiers in predictive analytics (Bertsimas & Dunn, 2017; Cortes & Vapnik, 1995; Freund & Schapire, 1997; Fernández-Delgado et al., 2014; Ke et al., 2017). XGBoost consistently achieved the best balance of high recall, low BER, and strong AUC compared to the alternatives, even under a temporal test-train split strategy. These results, along with additional details on our benchmarking process and rationale for selecting XGBoost as our primary method, are included in Appendix S4 for transparency. We chose to place these more technical benchmarking results in the Appendix to maintain the main text's focus of making modern machine learning methods more approachable for education researchers more broadly.

Analytical Process

We trained a total of 24 models across four different types of classifiers: logistic regression, logistic regression with SMOTE, XGBoost, and XGBoost with SMOTE. Inputs for each model are detailed in Table 4. Preprocessing, SMOTE, and XGBoost algorithms were coded using the Scikit-learn, Imbalanced-learn, and XGBoost libraries in Python, respectively.

For preprocessing, the original dataset (N=6,698) was divided into training (80%; N=5,358) and test (20%; N=1,340) datasets in order to both train and evaluate our models. For the models using SMOTE, we used a variant of SMOTE called SMOTE-Nominal Continuous (SMOTE-NC; Chawla et al., 2002) to preprocess the training dataset since our features contained both continuous and categorical variables. Empirical verification that SMOTE instances are comparable to our true minority instances is in Appendix S5. All missing features were imputed using MICE (Multivariate Imputation by Chained Equations; Van Buuren & Groothuis-Oudshoorn, 2011).

We tuned each XGBoost model via a grid search within 5-fold cross-validation for four core hyperparameters described by Chen & Guestrin (2016). Hyperparameters are settings that control how a model learns from the data. In machine learning, hyperparameters determine aspects such as how complex the model can become, how fast it learns, and how much it tries to avoid overfitting. A useful analogy is to think of hyperparameters like the conductor's instructions to an orchestra: They set the tempo, dynamics, and interpretation before the music begins, guiding the musicians but not playing the instruments themselves. Specifically, we tuned the number of trees (more trees can improve learning but risk overfitting), the depth of trees (deeper trees can capture more complex patterns but also risk overfitting), the learning rate (how quickly the model adjusts to errors from one iteration to the next), and gamma (a regularization term that penalizes overly complex trees). Optimal hyperparameter values for each model along with additional details and rationale for the chosen hyperparameters are in Appendix S6.

Trained models were evaluated on the testing set using accuracy rate, recall/TPR, specificity/TNR, BER, and AUC, and all these performance metrics are reported in our results. However, we paid particular attention to recall/TPR and BER as the main performance metrics

when assessing our different models. We emphasized recall/TPR because of our focus on predicting students who will be chronically absent and BER because of how it balances performance across the positive and negative classes, making it more robust for evaluating model performance datasets with class imbalance. We compared the performance metrics of all the models, reflected on how algorithmic structures could contribute to the performance patterns, and then examined the SHAP beeswarm for the best-performing predictive model.

To test whether the inclusion of demographic information meaningfully improved model performance beyond what can be achieved using attendance predictors alone, we compared the performance metrics of "full" models that included both time-varying and non-time-varying demographic and behavioral predictors, consistent with many current early warning models, with models that relied solely on attendance rate. This comparison enabled us to assess how much the exclusion of demographic characteristics, along with other time-varying predictors which may implicitly encode demographic information (e.g., suspension history or students' school), would impact model performance. Although attendance rate may correlate with these other predictors, it is distinct in that it is consistently tracked from early childhood and likely most responsive to school-based interventions. While this initial comparison does not constitute a full fairness audit, it serves as a critical first step in understanding how the inclusion of sensitive demographic variables affects model behavior. Given the importance of this issue, we are developing a separate follow-up study focused on examining advanced techniques for algorithmic fairness. For readers interested in a more in-depth discussion of fairness in predictive analytics, we recommend Gándara et al. (2025) as a helpful primer.

Finally, we explored how the performance metrics of our predictive models changed when we altered the probability threshold used to classify students as at risk of chronic

absenteeism. For all models, the default classification threshold was 0.5, meaning that students with predicted probabilities greater than or equal to 50% were considered at risk of chronic absenteeism. We systematically varied this threshold in increments of 0.1, from 0.1 to 0.9, and recalculated accuracy, recall, specificity, BER, and AUC at each step. This analysis allowed us to assess how different thresholds influenced model performance and to identify threshold patterns that could be better aligned with school and district priorities.

Results

RQ1: Prediction Accuracy from SMOTE and XGBoost Versus Logistic Regression

Table 4 presents the performances of each machine learning algorithm we tested. The first seven rows are the results for the parsimonious logistic regression model, each with different specifications. The outlined fourth row represents the model closest to that used in Massachusetts's current EWSs for third grade, and we refer to this model as the "baseline" model for the results. This is followed by the set of results for logistic regression with SMOTE, for XGBoost without SMOTE, and finally for XGBoost with SMOTE.

For accuracy rate, the XGBoost model without SMOTE with predictors from school years PreK-2nd grade (row 18) performed the best with 91.6% accuracy. However, the accuracy rates for most models hovered around 90%. For recall/TPR, the XGBoost + SMOTE model using predictors from all school years PreK-2nd grade (row 23) had the highest rate at 64.3%. For specificity/TNR, the logistic regression using kindergarten predictors (row 2) performed the best at 0.995. For BER, the XGBoost + SMOTE model using predictors for all school years (row 23) performed best at 0.227. The model with the highest AUC at 0.890 was also the XGBoost + SMOTE model using predictors for all school years (row 23).

Overall, the models using SMOTE performed better than the ones not using SMOTE, and the XGBoost models performed better than the logistic regression models. The best model performance based on recall rate, BER, and AUC was the XGBoost + SMOTE model with predictors from all grades (row 23).

Regarding overall patterns in the results, the first four rows of logistic regression models in Table 4 had a high overall accuracy rate (around 90%) but hovered around 5-10% for recall, even when we added covariates from all school years PreK-2nd grade (row 6). This means that only a small percentage of students in the test sample who ended up being chronically absent in 3rd grade were identified as having a high risk of chronic absenteeism. The performance of these models, in particular the baseline model (row 4), is closest to approximating the performance of existing EWSs. From this, we can infer that current EWSs that rely solely on logistic regressions may not have high recall for identifying students who will be chronically absent, even if the model includes predictors from all of the students' past school years (row 6) or interactions (row 7).

The low recall rate for this first set of logistic regression results is likely due to two main reasons. First, the minority class of chronically absent students was not well-represented in our original dataset. We see that the recall rate increased from around 5-10% to approximately 25-30% in the second set of models in Table 4 (rows 8-13) when we use SMOTE to address class imbalance in addition to the logistic regression. This is evidence that synthetically increasing the number of minority class samples did make a positive impact on our predictive ability. The tradeoff is that the specificity rate decreased by approximately 4-5 percentage points when using SMOTE, so the accuracy decreased for identifying students who would not be chronically absent. Nevertheless, the specificity rate remained high at 94-95%. We also see a boost in the

recall rate in the XGBoost models when using SMOTE, giving further evidence of the utility of using the SMOTE algorithm to train predictive models for a EWSs. This result is consistent with recent studies on the impact of class rebalancing techniques like SMOTE on the performance of predictive models (Tantithamthavorn et al., 2018).

The second reason for the low recall rate is that logistic regression models without an interaction term make a linear classification boundary, which do not do well separating the two classes if the true classification boundary is non-linear. A visualization of a 2-D example is shown in Figure 4. Using the class imbalance graph, we see that a linear model like a logistic regression cannot accurately separate out the chronically absent points from the non-chronically absent points if the true shape of the data is non-linear. A possible fix would be to run a non-linear logistic regression by including interaction terms. We do this in row 7 in Table 4, interacting all non-time-varying student characteristic variables. We see that this improves the recall rate to 0.151, providing evidence for our theory that the true classification boundary is non-linear, at least to a degree. While we would ideally like to have interacted all the predictors together to test the extent of non-linearity, we were only able to interact non-time-varying predictors. Including interactions for the time-varying along with the non-time-varying predictors led to non-convergence errors in the regression model. Including interaction terms in the logistic regression model with SMOTE also led to non-convergence errors (row 13), likely because of multicollinearity issues between the synthetic samples generated and our actual sample. This demonstrates the limits of using logistic regression for prediction non-linear boundaries.

XGBoost models, on the other hand, do not run into this limitation. The higher recall rates from the XGBoost models (rows 14-24) provide evidence supporting the need for a more

non-linear, non-parametric classification boundary line than a logistic regression is able to produce. The best XGBoost model (row 23) had the highest recall rate (0.643) of all the models, the lowest BER (0.227), and the highest AUC (0.890). While there is no specific threshold for what is considered a good AUC score, models with an AUC of 0.8-0.9 are generally considered excellent classifiers (Hosmer & Lemeshow, 2000).

RQ2: Accuracy, Timeliness, and Interpretation of Predictions

As aforementioned, the best model (XGBoost + SMOTE in row 23 of Table 4) predicted third grade chronic absenteeism status with a recall rate of 64.3%, which is 33 percentage points higher than the recall rate of the top logistic regression model with SMOTE (row 11), approximately a 50% improvement, and 54 percentage points higher than the recall rate of the baseline logistic regression model (row 4) that best approximates the model used in EWSs today. The usage of predictors spanning all grades from PreK-2 implies that having more data from more years of schooling bolsters the accuracy of the XGBoost + SMOTE models in early childhood, but not by a wide margin.

Surprisingly, relying solely on PreK data in the XGBoost + SMOTE model (row 19) already yielded a recall rate of 58.7%, albeit at the cost of decreased specificity, resulting in an overall accuracy of only 81.4%. This suggests that even data from children's earliest schooling experience has predictive power. Overall, the results from the models with just one grade level of data indicate that even if a school or district only possesses data from one school year, XGBoost + SMOTE models can still harness it to identify students at a heightened risk of chronic absenteeism more accurately compared to logistic regressions.

To better understand what drove predictions in the top-performing XGBoost + SMOTE model (row 23 of Table 4), we used a SHAP beeswarm plot (Figure 5). A step-by-step guide to

interpreting the beeswarm plot is included in the note under Figure 5. Based on the plot, second grade attendance rate emerged as the most influential predictor. High second grade attendance rates were associated with reduced risk of chronic absenteeism (red dots cluster on the left with negative SHAP values), and low attendance rates were associated with increased risk (blue dots cluster on the right with positive SHAP values). The wide spread of SHAP values for this feature suggests that its impact varied substantially across students, likely due to interactions with other variables in the model. Following 2nd grade attendance rate, the next three most important features were the attendance rates from kindergarten, 1st grade, and PreK. These followed a similar pattern to 2nd grade attendance but exhibited slightly less variance in their SHAP values, likely reflecting more modest though still meaningful contributions to the predictions. The prominence of attendance rates as the most influential predictors in the XGBoost + SMOTE model is encouraging, as these are malleable indicators that schools can act on directly. In contrast, some of the next features, such as race and dual language learner status, were fixed demographic characteristics. However, these variables had narrower SHAP value distributions, suggesting that while they contributed to model predictions, their overall influence was more limited.

In tree-based models like XGBoost, the SHAP values used to calculate feature importance can favor continuous variables, which have more potential split points than binary variables (Lundberg & Lee, 2017). To assess whether the high importance of attendance rate predictors reflected meaningful contributions rather than this modeling artifact, we computed SHAP interaction values as a robustness check (additional details in Appendix S7). We found that attendance rate features ranked high in their SHAP interaction values as well, providing

further evidence that their influence in the XGBoost model stems from substantive relationships in the data, not just structural favoritism toward continuous variables.

RQ3: The Impact of Student Demographic Information

While attendance rates emerged as the strongest predictors in our XGBoost + SMOTE model, as shown in the SHAP beeswarm plot (Figure 5), several demographic characteristics also contributed to predictions, albeit with smaller overall impact. To evaluate how including student demographic information affected model accuracy, we compared the performance metrics of the "full" baseline logistic and best-performing XGBoost + SMOTE models to pared-down versions using only attendance rate predictors (Table 4, rows 5 and 24). For the logistic regression model with only attendance rate predictors (row 5), recall was 0.079 and BER was 0.463, compared to a recall of 0.103 and BER of 0.452 for the full model (row 4). The best-performing XGBoost + SMOTE model with only attendance predictors (row 24) achieved a recall of 0.627 and BER of 0.236, compared to 0.643 recall and 0.227 BER in the full model (row 23).

In both modeling approaches, using only attendance rate as a predictor resulted in only modest drops in performance. The fact that attendance-only models yielded comparable performance suggests that demographic information is not necessarily essential for achieving strong model predictions, and that it may be possible to design accurate EWS algorithms without relying on sensitive student attributes. These findings reinforce the results from the SHAP beeswarm plot, which ordered attendance rates as the most influential features; adding additional demographic predictors offered limited gains in overall predictive power.

RQ4: Personalization of Machine Learning Models

Table 5 displays the performance metrics of four models across a range of probability thresholds: the baseline logistic regression model, the best-performing logistic model with

SMOTE, the best-performing XGBoost + SMOTE model, and the XGBoost + SMOTE model using only attendance rate predictors. The default probability threshold for all models in Table 4 was 0.5. However, there are cases when educational institutions may want to be more or less stringent with the threshold. For example, a district facing severe budget constraints may want to implement an intensive intervention only for students who have a very high likelihood of being chronically absent in the next year without the intervention. In this case, the district may want to use a higher probability threshold of 0.8 or 0.9. Conversely, a district considering a low-cost text messaging intervention (Heppen et al., 2020; Rogers, 2018) may opt to use a lower probability threshold of 0.3 or 0.4 to reach a wider range of students. By combining predictive models with strategic threshold-setting, districts can respond more dynamically to patterns of absenteeism using the data they already have on hand.

Across all thresholds, both XGBoost + SMOTE models consistently outperformed the logistic regression models in terms of recall, BER, and AUC. Even at a high threshold of 0.8, where models become more conservative in flagging students as at risk of future chronic absenteeism, the best-performing XGBoost + SMOTE model identified nearly three times as many chronically absent students (recall of 0.349) as the best-performing logistic + SMOTE model (recall or 0.127) and more than ten times as many chronically absent students as the baseline logistic model (recall of 0.032). The best-performing XGBoost + SMOTE model also maintained a lower BER compared to both logistic models at this threshold.

Comparing the two XGBoost + SMOTE models—the full model using all predictors and the one using only attendance rates—reveals once again that attendance alone accounts for much of the model's predictive power even across probability thresholds. At every threshold, the performance gap between the full model and the attendance-only model was minimal, with

differences in recall and BER typically within 1 to 2 percentage points. In fact, at certain thresholds, such as 0.3 and 0.8, the attendance-only model slightly outperformed the full model on recall (e.g., recall of 0.373 in the attendance-only model versus 0.349 at the 0.8 threshold). This supports the above RQ3 finding, suggesting that while adding demographic or contextual variables may offer marginal gains, recent attendance patterns are by far the most powerful and consistent indicators of future absenteeism. These findings reinforce the feasibility of designing strong early warning models that prioritize behavioral indicators. Overall, Table 5 demonstrates the ability to personalize ML models to inform school and district policies based on their specific needs, including accounting for financial or resource constraints.

Discussion & Conclusion

Overall, this present study illustrates the utility of two modern machine learning algorithms, SMOTE and XGBoost, in enhancing early warning systems for the proactive identification of students at heightened risk of chronic absenteeism during early childhood. The top-performing XGBoost model with SMOTE outperformed the logistic regression model closest to that in current EWSs by 54 percentage points and the best logistic regression model with SMOTE by approximately 33 percentage points in accurately forecasting chronic absenteeism among third grade students. This finding aligns with previous research (Lee & Chung, 2019; Sansone, 2019) concluding that modern machine learning tools provide more accurate predictions compared to the logistic regression models used in many parsimonious EWSs today.

Our findings also suggest that school districts do not necessarily need four years of longitudinal data to meaningfully improve EWS performance in early childhood. Even when using only PreK data, for example, our XGBoost models achieved substantial gains in recall and BER relative to baseline logistic regression model. This suggests that districts with limited data

can still strengthen their ability to identify students at risk of chronic absenteeism in early childhood by adopting flexible machine learning approaches. Consistent with prior research (Chang & Romero, 2008; Ehrlich et al., 2013), our results show that PreK attendance can serve as an early and meaningful signal of school disengagement and a foundation for predictive EWSs. This has policy implications: PreK is often the first sustained point of contact between families and the school system, offering a critical opportunity to identify and support students before patterns of absenteeism become entrenched. In Massachusetts, early attendance is viewed as a key metric of school readiness (Kane, 2024), and although attending PreK is optional, Boston Public Schools has invested heavily in universal PreK access as part of its long-term strategy for improving educational outcomes. The fact that PreK attendance emerged as the fourth most important predictor in our SHAP analysis and appeared in multiple high-value interactions underscores its potential as an early intervention point.

Importantly, the XGBoost + SMOTE model using only attendance rate predictors performed comparably to those using all available predictors, even across varying probability thresholds. Recent work by Yu et al. (2021) and Bird et al. (2021) similarly found that the exclusion of sensitive demographic attributes had little impact on model accuracy. While their studies focused on predicting college student outcomes, our findings in an early childhood and absenteeism prediction context suggest this insight may generalize: comparable predictive performance for EWSs may be possible without relying on sensitive student attributes. These results, alongside the SHAP beeswarm plot which identified attendance rates as the most influential features, underscore the central role of attendance patterns more broadly in predicting future chronic absenteeism, even relative to broader contextual or demographic factors. This finding aligns with a robust body of research demonstrating the predictive value of attendance,

not only for early attendance patterns (Ehrlich et al., 2018) but also for later academic achievement, high school graduation, college enrollment, and workforce participation (Allensworth & Easton, 2007; Balfanz & Byrnes, 2012; Klein et al., 2024; Wei, 2024; Wu et al., 2025). Our study extends this literature by using modern machine learning methods rather than traditional regression approaches, offering additional support for attendance as a powerful and scalable early indicator. Together, these findings suggest that EWSs can be designed to rely primarily on malleable, behavioral indicators like attendance, reducing dependence on demographic predictors that are not directly actionable for schools.

At the same time, longstanding concerns remain that EWSs may inadvertently reinforce biases (Baker et al., 2023; Gándara et al., 2025; Yu et al., 2021). While our findings offer a compelling case for ML algorithms, responsible use of EWS models requires further attention to fairness. Even when overall model performance is comparable between models that include or exclude student demographic information, subgroup-level disparities in prediction can persist. Future work should focus on assessing model performance using algorithmic fairness metrics such as demographic parity, equalized odds, equal opportunity, and disparate impact (Hardt et al., 2016; Feldman et al., 2015). These metrics can help ensure that the model predicts chronic absenteeism at similar rates for all groups and highlight whether any group is disproportionately affected by incorrect predictions. Prior research has found that including sensitive student attributes yields only marginal gains in algorithmic fairness (Yu et al., 2021). Therefore, if unfairness is detected, strategies beyond including or excluding demographic predictor variables, such as reweighting the training data or applying fairness constraints during model training should be examined to mitigate bias in predictions (Baker et al., 2023; Gándara et al., 2024; Kamiran & Calders, 2012; Zafar et al., 2017). In addition, while the best-performing XGBoost +

SMOTE model achieved a recall rate of 64.3%, a substantial improvement over the baseline logistic regression model, it still leaves considerable room for improvement. Future research should explore the use of additional predictors; given that demographic characteristics added minimal predictive value, it is particularly important to identify alternative indicators that can further enhance model accuracy without introducing additional fairness concerns. There are also more hyperparameters we could tune for our XGBoost model that could enhance model performance, such as the maximum delta step (how much a tree's prediction value can change at each iteration) and the lambda and alpha regularization terms (both try to encourage simpler, more generalizable models). In this study, we prioritized tuning core parameters such as tree depth and number of estimators because they are the parameters commonly recommended as the first to tune in XGBoost tutorials (Chen & Guestrin, 2016). Future studies should explore additional hyperparameters that could help reduce overfitting and help the algorithm run faster.

As noted in Appendix S4, while we selected XGBoost as our final model to examine in depth due to its strong performance, several other modern machine learning algorithms, such as AdaBoost, also outperformed logistic regression and achieved performance metrics comparable to XGBoost. Our decision to highlight XGBoost was not based on the assumption that it will always outperform other algorithms, but by a desire to provide a detailed, pedagogically clear example of a widely-used ML model that could perform well across other education use cases as well. With additional tuning or in a different sample, it is entirely possible that AdaBoost or another modern ML method could match or exceed XGBoost's performance. Future research should systematically compare these models across settings to assess tradeoffs in predictive accuracy, fairness, and computational efficiency.

Lastly, we focus on recall rate and BER as our primary performance metrics, given our goal of identifying students at risk of chronic absenteeism as accurately as possible. While no single metric can fully capture model performance, recall is particularly important in this context because it reflects the proportion of students correctly identified as at risk. In districts where early intervention is possible, missing these students (false negatives) can undermine the impact of support efforts: the longer a student stays off track, the harder and more resource-intensive it becomes to re-engage them. Prioritizing recall thus increases the likelihood that students in need of support are flagged early. That said, there may be situations where other metrics warrant emphasis. For instance, if a district could only intervene with a small number of students, it might prioritize precision (also known as positive predictive value; TP/(TP+FP)) to ensure that flagged students are truly at risk, even if that means missing some who also need support.

In sum, the application of the modern machine learning algorithms, namely XGBoost and SMOTE, in EWSs could lead to a substantial increase in schools' ability to detect students who have a higher risk of becoming chronically absent and, consequently, to mitigate chronic absenteeism during elementary school years. The findings have implications for future education research grappling with the consequences of class imbalance, algorithmic fairness, and leveraging predictive analytics for outcomes beyond chronic absenteeism. It illuminates the advantages of integrating modern ML algorithms into the field of education, and we hope this paper serves as a valuable introduction for education researchers hoping to incorporate these techniques into their own research.

References

Allensworth, E., Balfanz, R., Rogers, T., & Demarzi, J. (2021). *Absent from school: Understanding and addressing student absenteeism*. Harvard Education Press.

Allensworth, E. M., & Easton, J. Q. (2007). What Matters for Staying On-Track and Graduating in Chicago Public High Schools: A Close Look at Course Grades, Failures, and Attendance in the Freshman Year. Research Report. *Consortium on Chicago School Research*.

Allison, M. A., Attisha, E., Lerner, M., De Pinto, C. D., Beers, N. S., Gibson, E. J., ... & Weiss-Harrison, A. (2019). The link between school attendance and good health. *Pediatrics*, *143*(2), 1-13.

Ansari, A., & Gottfried, M. A. (2018). Early childhood educational settings and school absenteeism for children with disabilities. *AERA Open*, *4*(2), 1-15.

Ansari, A., & Gottfried, M. A. (2021). The grade-level and cumulative outcomes of absenteeism. *Child Development*, *92*(4), e548-e564.

Ansari, A., & Pianta, R. C. (2019). School absenteeism in the first decade of education and outcomes in adolescence. *Journal of School Psychology*, *76*, 48-61.

Baker R. S., Esbenshade L., Vitale J., Karumbaiah S. (2023). Using demographic data as predictor variables: A questionable choice. *Journal of Educational Data Mining, 15*(2), 22-52.

Balfanz, R. (2016). Missing school matters. *Phi Delta Kappan*, *98*(2), 8-13.

Balfanz, R., & Byrnes, V. (2012). The importance of being in school: A report on absenteeism in the nation's public schools. *The Education Digest*, *78*(2), 4-9.

Balfanz, R., & Byrnes, V. (2013). Meeting the challenge of combating chronic absenteeism. *Everyone Graduates Center at Johns Hopkins University School of Education*, 1-2.

Balfanz, R., & Byrnes, V. (2019). Early warning indicators and intervention systems: State of the field. *Handbook of Student Engagement Interventions*, 45-55.

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning*, *106*(7), 1039–1082.

Bird, K. A., Castleman, B. L., Mabel, Z., & Song, Y. (2021). Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education. *Aera Open*, *7*.

Boston Public Schools. (2022). *Attendance and punctuality policies and procedures.* Superintendent's Circular: School Year 2022-2023.

Chang, H. N., & Romero, M. (2008). Present, Engaged, and Accounted for: The Critical Importance of Addressing Chronic Absence in the Early Grades. Report. *National Center for Children in Poverty*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785-794.

Childs, J., & Scanlon, C. L. (2022). Coordinating the mesosystem: An ecological approach to addressing chronic absenteeism. *Peabody Journal of Education*, *97*(1), 74-86.

Christie, S. T., Jarratt, D. C., Olson, L. A., & Taijala, T. T. (2019). Machine-Learned School Dropout Early Warning at Scale. *International Educational Data Mining Society*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Davis, M. H., Mac Iver, M., Balfanz, R., Stein, M, & Fox, J. (2018). Implementation of an early warning indicator and intervention system. *Preventing School Failure, 63*(1), 77-88.

Deussen, T., Hanson, H., & Bisht, B. (2017). Are two commonly used early warning indicators accurate predictors of dropout for English learner students? Evidence from six districts in Washington State. REL 2017-261. *Regional Educational Laboratory Northwest*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

Ehrlich, S. B., Gwynne, J. A., & Allensworth, E. M. (2018). Pre-kindergarten attendance matters: Early chronic absence patterns and relationships to learning outcomes. *Early Childhood Research Quarterly*, *44*, 136-151.

Faria, A. M., Sorensen, N., Heppen, J., Bowdon, J., Taylor, S., Eisner, R., & Foster, S. (2017). Getting students on track for graduation: Impacts of the early warning intervention and monitoring system after one year. REL 2017-272. *Regional Educational Laboratory Midwest*, 1-82.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259-268.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15,* 3133–3181.

Frazelle, S., & Nagel, A. (2015). A practitioner's guide to implementing early warning systems (REL 2015–056). Washington, D.C.: *U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northwest.*

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Fuchs, D., Compton, D. L., Fuchs, L. S., Bryant, J., & Davis, G. N. (2008). Making "secondary intervention" work in a three-tier responsiveness-to-intervention model: Findings from the first-grade longitudinal reading study of the National Research Center on Learning Disabilities. *Reading and Writing, 21,* 413-436.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (2007). Using curriculum-based measurement to inform reading instruction. *Reading and Writing, 20(6),* 553-567.

Gándara, D., Anahideh, H., Ison, M. P., & Picchiarini, L. (2024). Inside the Black Box: Detecting and Mitigating Algorithmic Bias Across Racialized Groups in College Student-Success Prediction. *AERA Open, 10*. https://doi.org/10.1177/23328584241258741

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). *Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation* (Technical Report UTEP-CS-18-09). University of Texas at El Paso. https://scholarworks.utep.edu/cs_techrep/1209/

Gottfried, M. A. (2014). Chronic absenteeism and its effects on students' academic and socioemotional outcomes. *Journal of Education for Students Placed at Risk (JESPAR)*, *19*(2), 53-75.

Gottfried, M. A., & Hutt, E. L. (2019). Addressing Absenteeism: Lessons for Policy and Practice. *Policy Analysis for California Education, PACE*.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems,* 29, 3315-3323.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Heckman, J. J. (2008). The case for investing in disadvantaged young children. CESifo DICE Report, *6*(2), 3-8.

Heppen, J. B., Kurki, A., & Brown, S. (2020). Can Texting Parents Improve Attendance in Elementary School? A Test of an Adaptive Messaging Strategy. Appendix. NCEE 2020-006a. *National Center for Education Evaluation and Regional Assistance*.

Hosmer, D.W., & Lemeshow, S. (2000). *Applied Logistic Regression.* John Wiley and Sons, New York.

Huang, Y., Alvernaz, S., Kim, S. J., Maki, P., Dai, Y., & Bernabé, B. P. (2024). Predicting prenatal depression and assessing model bias using machine learning models. *Biological Psychiatry Global Open Science*, 100376.

Jacob, B., & Lovett, K. (2017). Chronic absenteeism: An old problem in search of new answers. *Brookings Institution*. https://www.brookings.edu/articles/chronic-absenteeism-an-old-problem-in-search-of-new-answers/

Jacobsen, W. C., Pace, G. T., & Ramirez, N. G. (2019). Punishment and inequality at an early age: Exclusionary discipline in elementary school. *Social Forces, 97*(3), 973–998.

Kabay, S., Weiland, C., & Yoshikawa, H. (2020). Costs of the Boston public prekindergarten program. *Journal of Research on Educational Effectiveness*, *13*(4), 574-600.

Kaggle. (2021). *State of Data Science and Machine Learning 2021.* https://www.kaggle.com/kaggle-survey-2021

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1-33.

Kane, D. (2024). *Reducing chronic absenteeism in our schools*. Massachusetts Education-to-Career Research and Data Hub. https://educationtocareer.data.mass.gov/stories/s/Reducing-chronic-absenteeism-in-our-schools/vuut-f46x/

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 3146–3154).

Kearney, C. A., & Childs, J. (2023). Improving school attendance data and defining problematic and chronic school absenteeism: the next stage for educational policies and health-based practices. *Preventing School Failure: Alternative Education for Children and Youth*, *67*(4), 265-275.

Kearney, C. A., Gonzálvez, C., Graczyk, P. A., & Fornander, M. J. (2019). Reconciling contemporary approaches to school attendance and school absenteeism: Toward promotion and nimble response, global policy review and implementation, and future adaptability (Part 1). *Frontiers in Psychology*, *10*, 1-16.

Kearney, C. A., Benoit, L., Gonzálvez, C., & Keppens, G. (2022). School attendance and school absenteeism: A primer for the past, present, and theory of change for the future. *Frontiers in Education, 7,* 1-17.

Klein, M., Sosu, E. M., & Dare, S. (2020). Mapping inequalities in school attendance: The relationship between dimensions of socioeconomic status and forms of school absence. *Children and Youth Services Review*, *118*, 1-12.

Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences, 9*(15), 1-14.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Massachusetts Department of Elementary and Secondary Education. (2023). *Early Warning Indicator System (EWIS).* https://www.doe.mass.edu/ccte/ccr/ewis/

Massachusetts Department of Elementary and Secondary Education, & American Institutes for Research. (2013). Technical Descriptions of Risk Model Development: Early and Late Elementary Age Groupings (Grades 1-6). Massachusetts Early Warning Indicator System (EWIS). https://www.doe.mass.edu/ccte/sec-supports/ewis/default.html

Mervosh, S. (2023, November 17). Students are missing school at an alarming rate. *The New York Times.* https://www.nytimes.com/2023/11/17/us/chronic-absenteeism-pandemic-recovery.html

OECD. (2020). Case study: Massachusetts' (United States) Early Warning Indicator System (EWIS). *Strengthening the Governance of Skills Systems : Lessons From Six OECD Countries, OECD iLibrary*. https://doi.org/10.1787/1fbfc1a3-en

Oliver, M. (2023, December 11). How school districts are tackling chronic absenteeism, which has soared since the COVID-19 pandemic. *CBS Evening News.* https://www.cbsnews.com/news/chronic-absenteeism-school-students-covid/

Panorama Education. (2023). https://www.panoramaed.com/products/student-success

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3-14.

Purtell, K. M., & Ansari, A. (2022). Why are children absent from preschool? A nationally representative analysis of Head Start programs. *Frontiers in Education, 7*, 1-13.

Reyes, A. (2020). Compulsory school attendance: The new American crime. *Education Sciences*, *10*(3), 75.

Rogers, T., & Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour*, *2*(5), 335-342.

Romero, M., & Lee, Y. S. (2007). A national portrait of chronic absenteeism in the early grades. *National Center for Children in Poverty, Columbia University,* 1-8.

Sansone, D. (2019). Beyond early warning indicators: High school dropout and machine learning. *Oxford Bulletin of Economics and Statistics*, *81*(2), 456-485.

Shah, A. (2022, July 26). XGBoost (Extreme Gradient Boosting) in Machine Learning. *Medium*. https://medium.com/@jwbtmf/xgboost-extreme-gradient-boosting-in-machine-learning-3427b937b35c

Sheldon, S. B. (2007). Improving student attendance with school, family and community partnerships. *The Journal of Educational Research, 100*(5), 267–275.

Singh, S. (2025). Feeding the algorithm: Legal challenges in AI training data. *Journal of Global Studies, Legal Studies*, *2.*

Swaak, T. (2018, July 31). With Nearly 8 Million Students Chronically Absent From School Each Year, 36 States Set Out to Tackle the Problem in New Federal Education Plans. Will It Make a Difference? *The 74 Million*. https://www.the74million.org/article/chronic-absenteeism-36-states-essa-plans/

Tantithamthavorn, C., Hassan, A. E., & Matsumoto, K. (2018). The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. *IEEE Transactions on Software Engineering*, *46*(11), 1200-1219.

Therriault, S. B., O'Cummings, M., Heppen, J., Yerhot, L., & Scala, J. (2017). Early warning intervention and monitoring system implementation guide. *Michigan Department of Education*.

Tyack, D. (1976). Ways of seeing: An essay on the history of compulsory schooling. *Harvard Educational Review*, *46*(3), 355-389.

U.S. Department of Education. (2016a). *Chronic Absenteeism in the Nation's Schools.* https://www2.ed.gov/datastory/chronicabsenteeism.html#intro

U.S. Department of Education. (2016b). *Issue brief: Early warning systems.* 1-13.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1-67.

Wei, W. (2024). Exploring Patterns of Absenteeism from Prekindergarten Through Early Elementary School and Their Associations With Children's Academic Outcomes. *AERA Open*, *10*.

Weissman, A. (2022). *Friend or foe? The role of machine learning in education policy research.* [Doctoral thesis, University of Michigan].

The White House. (2023, September 13). *Chronic absenteeism and disrupted learning require an all-hands-on-deck approach*. https://www.whitehouse.gov/cea/written-materials/2023/09/13/chronic-absenteeism-and-disrupted-learning-require-an-all-hands-on-deck-approach/

The White House. (2024, January 17). *Fact sheet: Biden-Harris administration announces improving student achievement agenda in 2024.* https://www.whitehouse.gov/briefing-room/statements-releases/2024/01/17/fact-sheet-biden-harris-administration-announces-improving-student-achievement-agenda-in-2024/

Williams, H. D. (1927). Truancy and delinquency. *Journal of Applied Psychology*, *11*(4), 276-288.

Wu, T., Weiland, C., Diemer, M. A., Unterman, R., Shapiro, A., & Staines, T. Measuring "noncognitive" skills at scale: Building longitudinal student behavior composites using administrative data. (EdWorkingPaper: 25-1250). Annenberg Institute at Brown University. https://doi.org/10.26300/7h7f-0j56

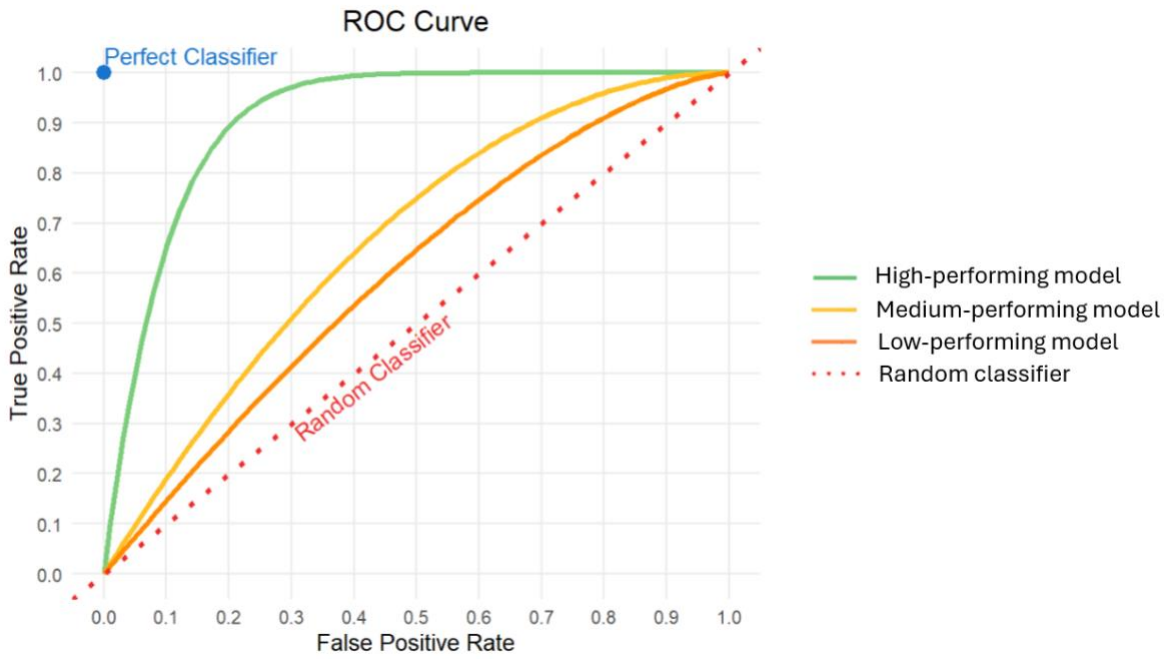
Yu, R., Lee, H., & Kizilcec, R. F. (2021, June). Should college dropout prediction models include protected attributes? In *Proceedings of the Eighth ACM Conference on Learning@ Scale* (pp. 91-100).

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017, April). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics* (pp. 962-970). PMLR.

Zhou, Z. H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press.

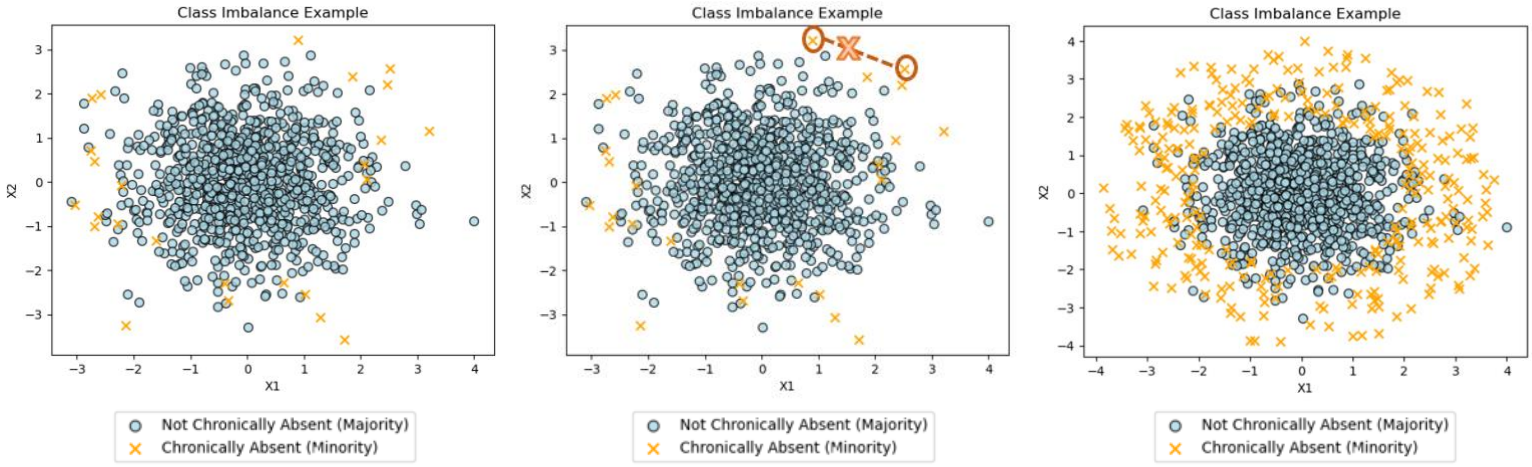
Figures & Tables

Figure 1. Hypothetical ROC Curve Demonstrating Model Performance



Note: This hypothetical ROC curve compares three models with varying predictive performance. The True Positive Rate is plotted against the False Positive Rate. The green curve represents a high-performing model that achieves strong separation between classes. The orange and yellow curves represent progressively weaker models. The red dashed line indicates a random classifier, where the model's predictions are as good as guessing, Curves closer to the upper-left corner reflect better model performance, with the blue dot indicating a perfect model that classifies everything correctly.

Figure 2. SMOTE Visualization



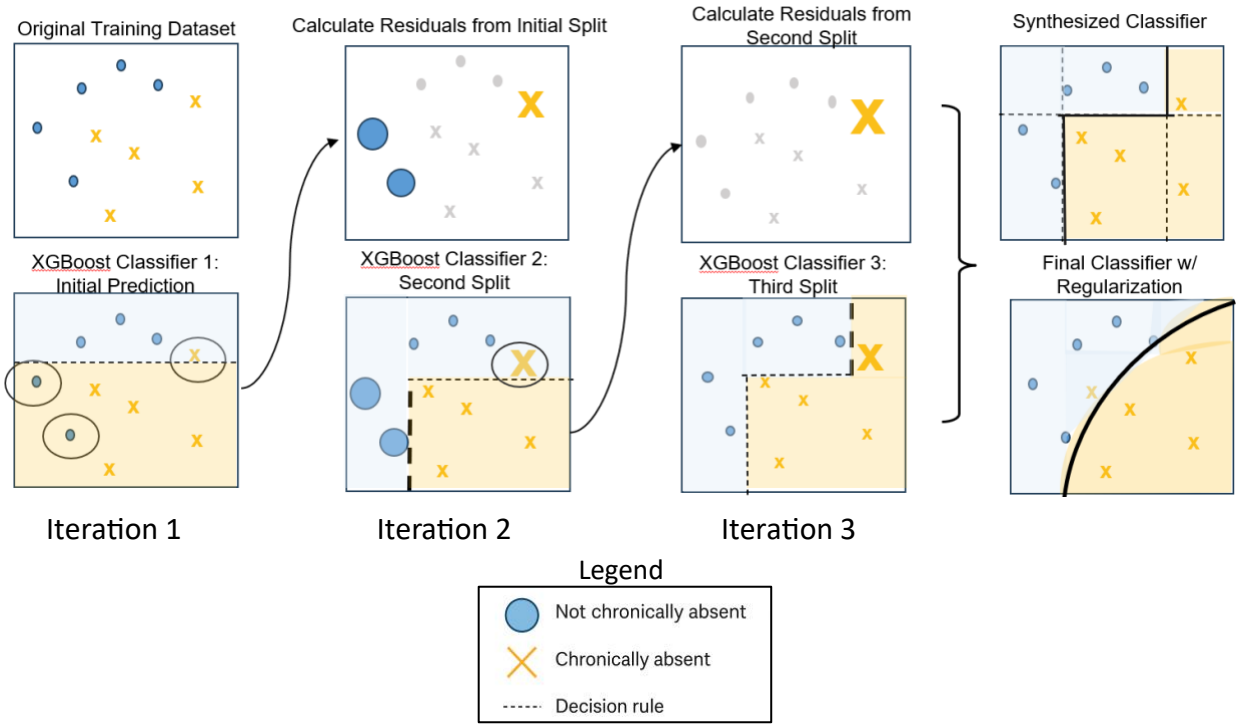
Before SMOTE

Creating 1 Synthetic Instance

After SMOTE

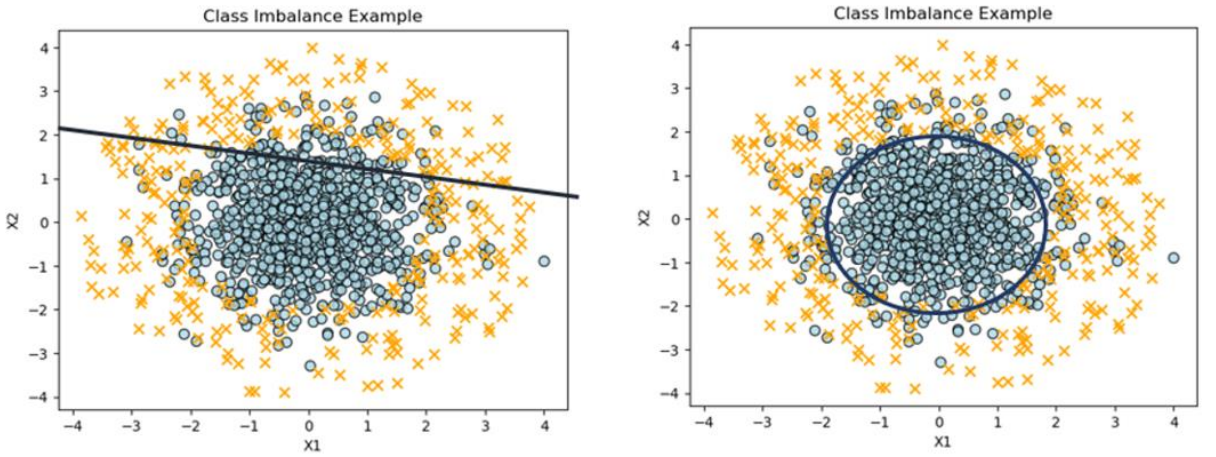
Note: Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al., 2002) is a resampling technique widely employed in machine learning to address class imbalance. This figure illustrates the SMOTE process in a hypothetical two-dimensional feature space. Before SMOTE (left panel), the chronically absent X's are sparse because of the class imbalance. This makes it difficult to classify these points. In the middle panel, the algorithm randomly identifies a chronically absent student (the first circled X) and selects one of its neighboring points using an algorithm called *k* nearest neighbors (the second circled X). Then, it creates a new synthetic instance (the big orange X) along the line segment connecting the two points. After repeating this process for many minority-class examples, the right panel shows how the chronically absent observations are more evenly distributed across the feature space. After this process, the circular boundary line separating the chronically absent from non-chronically absent students becomes clearer.

Figure 3. Graphical Scheme of XGBoost Algorithm (adapted from visualizations by Shah (2020))



Note: While simplified, this figure captures the core logic of XGBoost. At Iteration 1, the model begins with the original training dataset: blue circles represent students who were not chronically absent, and orange X's represent students who were. The model creates a simple decision rule (e.g., a horizontal split), shown as a dotted line. Students above the line are predicted to be blue; those below are predicted to be orange. Misclassified points (i.e., prediction errors) are circled. In each subsequent iteration, the algorithm focuses on these errors. Specifically, XGBoost fits a new decision tree to predict the residuals—the difference between actual and predicted outcomes. For instance, Iteration 2 might introduce a vertical split to better classify the students that were misclassified in the first round. This process continues over several iterations, with each new tree attempting to correct the mistakes of the previous ones. In the final step, XGBoost synthesizes the predictions from each iteration into a final classifier.

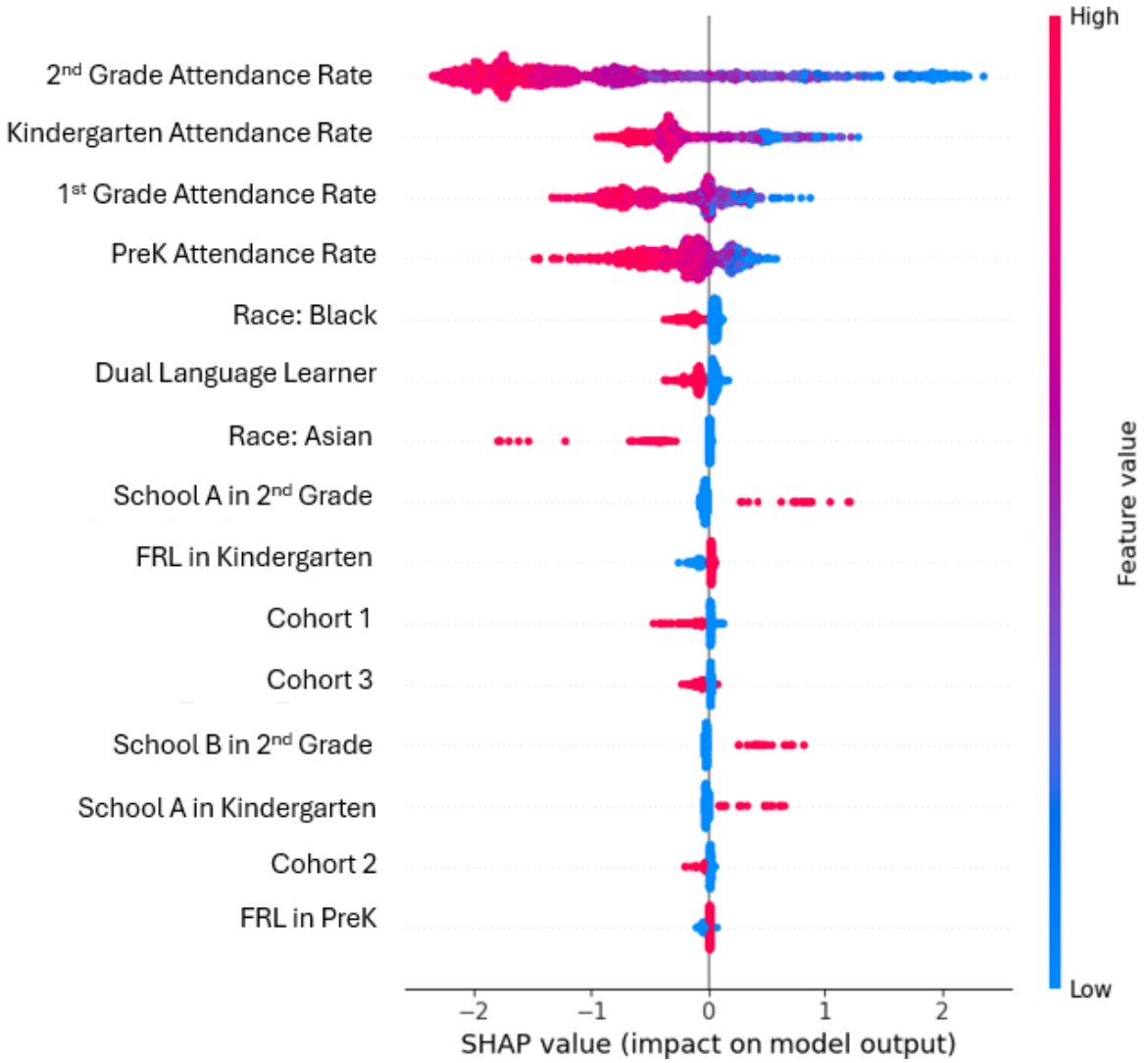
Figure 4. Linear vs. Non-linear Classification Boundary Line



Linear Classification Boundary Line (cannot approximate the necessary non-linear boundary needed)

Non-Linear Classification Boundary Line (needed in order to accurately approximate the boundary)

Figure 5. SHAP Beeswarm Plot for XGBoost + SMOTE Model



Note: The beeswarm plot is designed to display an information-dense summary of how the top features in a dataset impact the model's output. On the y-axis, the top 15 predictors used in the XGBoost + SMOTE model are sorted by feature (predictor) importance, from most to least impactful based on their mean absolute SHAP value. The x-axis shows the SHAP value for each feature, or the impact of each feature on the model prediction output. SHAP values greater than zero means that the feature increases the predicted risk of chronic absenteeism; SHAP values less than zero means the feature decreases the predicted risk. The further from zero, the greater effect the predictor has on the model's overall predictions. In the plot, each dot is one observation in the test data. The color of the dot represents the feature value. Red means a high feature value; blue means a low value.

Table 1. Descriptive Statistics by Chronic Absenteeism Status

Variable	Overall	Not CA in 3rd Grade, N = 6,066	CA in 3rd Grade, N = 632	*p*-value*
Male	50.57%	50.40%	52.40%	0.34
White	17.15%	17.80%	10.80%	<0.001
Black	28.17%	28.00%	29.60%	0.41
Hispanic/Latino	42.25%	41.00%	53.80%	<0.001
Asian	9.32%	10.00%	2.50%	<0.001
Mixed/Other Race	3.11%	3.10%	3.30%	0.74
Free/Reduced Lunch	71.10%	69.30%	88.60%	<0.001
Special Education	17.31%	16.20%	28.30%	<0.001
Dual Language Learner	43.77%	44.60%	35.60%	<0.001
PreK Chronically Absent	26.56%	22.20%	68.00%	<0.001
K Chronically Absent	18.66%	14.50%	59.00%	<0.001
1st Grade Chronically Absent	11.44%	7.60%	48.50%	<0.001
2nd Grade Chronically Absent	10.25%	5.70%	54.00%	<0.001

Note: CA stands for 'chronically absent'. *P*-values are for differences between students who were and were not chronically absent in 3rd grade, calculated using a Pearson's chi-squared test. Time-varying characteristic percentages (free/reduced lunch and special education) are based on students' PreK value. There was a small amount of missing data for students in special education (0.20%), K chronically absent (2.10%), 1st grade chronically absent (3.20%), and 2nd grade chronically absent (0.50%).

Table 2. Sample Confusion Matrix

		Actual Label		
		0 (Not CA)	1 (CA)	
Predicted Label	0 (Not CA)	TN	FP	Specificity or TNR = TN/(TN+FP)
	1 (CA)	FN	TP	Recall or TPR = TP/(TP+FN)

Table 3. Class Imbalance Example

		Actual Label	
		0 (Not CA)	1 (CA)
Predicted Label	0 (Not CA)	990	10
	1 (CA)	0	0

Table 4. Performance Metrics for All Models

Row	Algorithm	Inputs				Performance Metrics				
		PreK	K	1st	2nd	Overall Accuracy	Recall/TPR	Specificity/TNR	BER	AUC
1	Logistic	✓				0.899	0.095	0.982	0.461	0.800
2	Logistic		✓			0.905	0.040	0.995	0.483	0.768
3	Logistic			✓		0.904	0.079	0.990	0.465	0.774
4	Logistic				✓	0.910	0.103	0.993	0.452	0.851
5	Logistic w/ Only Attendance Rate				✓	0.908	0.079	0.994	0.463	0.868
6	Logistic	✓	✓	✓	✓	0.905	0.127	0.986	0.444	0.847
7	Logistic w/ Interactions	✓	✓	✓	✓	0.908	0.151	0.987	0.431	0.845
8	Logistic + SMOTE	✓				0.881	0.286	0.943	0.386	0.750
9	Logistic + SMOTE		✓			0.875	0.254	0.939	0.403	0.729
10	Logistic + SMOTE			✓		0.881	0.333	0.938	0.364	0.747
11	Logistic + SMOTE				✓	0.896	0.317	0.956	0.364	0.801
12	Logistic + SMOTE	✓	✓	✓	✓	NC	NC	NC	NC	NC
13	Logistic + SMOTE w/ Interactions	✓	✓	✓	✓	NC	NC	NC	NC	NC
14	XGBoost	✓				0.901	0.040	0.991	0.485	0.825
15	XGBoost		✓			0.910	0.135	0.991	0.437	0.794
16	XGBoost			✓		0.906	0.183	0.982	0.418	0.812
17	XGBoost				✓	0.913	0.325	0.974	0.350	0.864
18	XGBoost	✓	✓	✓	✓	0.916	0.317	0.979	0.352	0.877
19	XGBoost + SMOTE	✓				0.814	0.587	0.838	0.287	0.819
20	XGBoost + SMOTE		✓			0.833	0.556	0.862	0.291	0.808
21	XGBoost + SMOTE			✓		0.851	0.524	0.885	0.296	0.808
22	XGBoost + SMOTE				✓	0.885	0.540	0.921	0.270	0.867
23	**XGBoost + SMOTE**	**✓**	**✓**	**✓**	**✓**	**0.879**	**0.643**	**0.904**	**0.227**	**0.890**
24	XGBoost + SMOTE w/ Only Attendance Rates	✓	✓	✓	✓	0.875	0.627	0.901	0.236	0.874

Note: CA stands for chronically absent. TPR stands for True Positive Rate (recall; proportion of chronically absent students correctly identified; higher is better). TNR stands for True Negative Rate (specificity; proportion of non-chronically absent students correctly identified; higher is better). BER stands for Balanced Error Rate (average of misclassification rate for the group of students who were chronically absent or not; lower is better). AUC stands for Area under the ROC Curve (overall ability of the model to correctly predict students who would be chronically absent or not; higher is better). Unless otherwise specified (i.e., for the models with only attendance rate predictors), models contain all time-varying and non-time-varying predictors. The logistic regression with interactions included interactions for all non-time-varying covariates. NC stands for no convergence, meaning the regression model fit was singular. The outlined logistic regression model in row 4 presents the results closest to using Massachusetts's current multilevel modelling structure for early warning systems. The best performing model (XGBoost with SMOTE and all four years of data) based on recall, BER, and AUC is outlined in row 23.

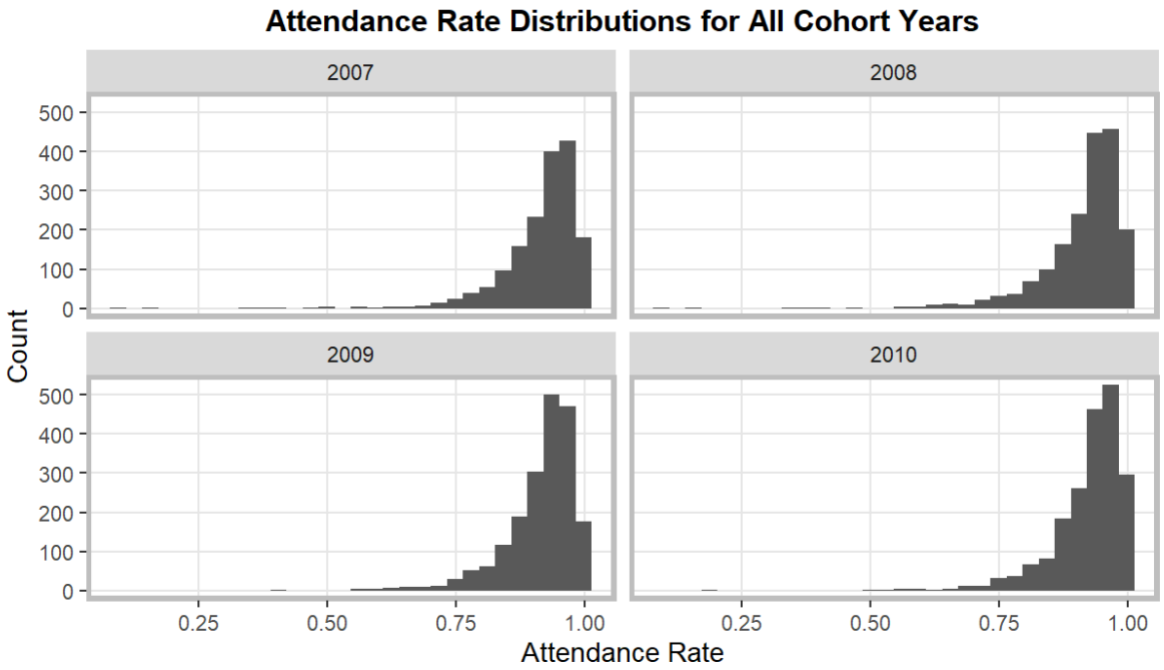
Table 5. Performance Metrics at Varying Probability Thresholds

Threshold	Overall Accuracy	Recall/TPR	Specificity/TNR	BER	AUC
Baseline Logistic Model with 2nd Grade Predictors (Table 4, Row 4)					
0.1	0.788	0.778	0.789	0.217	0.851
0.2	0.902	0.397	0.955	0.324	0.851
0.3	0.912	0.222	0.984	0.397	0.851
0.4	0.909	0.119	0.991	0.445	0.851
0.5	0.910	0.103	0.993	0.452	0.851
0.6	0.908	0.079	0.994	0.463	0.851
0.7	0.905	0.032	0.996	0.486	0.851
0.8	0.905	0.032	0.996	0.486	0.851
0.9	0.905	0.016	0.998	0.493	0.851
Best Logistic + SMOTE Model (Table 4, Row 11)					
0.1	0.583	0.849	0.555	0.298	0.801
0.2	0.758	0.651	0.769	0.290	0.801
0.3	0.837	0.540	0.867	0.296	0.801
0.4	0.879	0.429	0.926	0.323	0.801
0.5	0.896	0.317	0.956	0.364	0.801
0.6	0.907	0.254	0.974	0.386	0.801
0.7	0.908	0.175	0.984	0.421	0.801
0.8	0.907	0.127	0.988	0.443	0.801
0.9	0.907	0.071	0.994	0.467	0.801
Best XGBoost + SMOTE Model (Table 4, Row 23)					
0.1	0.674	0.937	0.647	0.208	0.890
0.2	0.775	0.857	0.766	0.188	0.890
0.3	0.822	0.770	0.827	0.202	0.890
0.4	0.860	0.714	0.876	0.205	0.890
0.5	0.879	0.643	0.904	0.227	0.890
0.6	0.895	0.579	0.928	0.247	0.890
0.7	0.913	0.484	0.957	0.279	0.890
0.8	0.915	0.349	0.974	0.339	0.890
0.9	0.917	0.190	0.993	0.408	0.890
XGBoost + SMOTE Model with Only Attendance Rates (Table 4, Row 24)					
0.1	0.622	0.937	0.589	0.237	0.880
0.2	0.739	0.825	0.730	0.222	0.880
0.3	0.810	0.770	0.814	0.208	0.880
0.4	0.843	0.706	0.857	0.218	0.880
0.5	0.875	0.627	0.901	0.236	0.880
0.6	0.901	0.556	0.937	0.254	0.880
0.7	0.912	0.500	0.955	0.273	0.880
0.8	0.915	0.373	0.971	0.328	0.880
0.9	0.913	0.143	0.993	0.432	0.880

Note: CA stands for chronically absent. TPR stands for True Positive Rate (recall; proportion of chronically absent students correctly identified; higher is better). TNR stands for True Negative Rate (specificity; proportion of non-chronically absent students correctly identified; higher is better). BER stands for Balanced Error Rate (average of misclassification rate for the group of students who were chronically absent or not; lower is better). AUC stands for Area under the ROC Curve (overall ability of the model to correctly predict students who would be chronically absent or not; higher is better). In this table, we highlight the Recall/TPR and BER columns because these two metrics are most important for model selection in the context of early warning systems for chronic absenteeism. Our goal is to maximize the identification of chronically absent students (high recall) while minimizing classification error across both groups (low BER). Readers should focus primarily on these columns when comparing model performance across varying probability thresholds. We see that for each probability threshold, both XGBoost + SMOTE models outperform the logistic regression with SMOTE model based on recall rate and BER. The XGBoost + SMOTE models with only attendance rates and with all predictors perform similarly, with the best XGBoost + SMOTE model performing marginally better across all thresholds.

Appendix

Appendix S1. Attendance Rate Distributions for All Cohort Years (2007-2010)



Appendix S2. Descriptive Statistics of Full 12,740 Sample and Broader Boston Public Schools (BPS) Student Population During Study Years

Variable	Study Sample (N=6,698)	Full Sample (N=12,740)	BPS in 2007-08	BPS in 2008-09	BPS in 2009-10	BPS in 2010-11
Male	50.57%	51.72%	50.50%	51.70%	51.80%	51.90%
White	17.15%	17.06%	13.40%	13.40%	13.10%	12.90%
Black	28.17%	28.43%	39.30%	39.30%	36.50%	35.50%
Hispanic/Latino	42.25%	43.87%	36.70%	36.70%	39.60%	40.90%
Asian	9.32%	7.58%	8.50%	8.50%	8.60%	8.40%
Mixed/Other Race	3.11%	3.06%	2.50%	2.50%	2.20%	2.30%
Free/Reduced Lunch	71.10%	65.07%	71.40%	74.30%	75.60%	74.40%
Special Education	17.31%	13.36%	20.10%	20.50%	19.60%	19.40%
Dual Language Learner	43.77%	41.03%	37.70%	38.10%	38.80%	43.40%

Note: There was a small amount of missing data for students from the full sample for sex (0.25%), race/ethnicity (0.25%), free/reduced lunch (1.45%), special education (22.32%), and dual language learner (5.24%).

Appendix S3. Additional Details on Predictor Variables

Variable Name	Definition
Attendance Rate (PK–2)	Proportion of days attended out of total school days enrolled in each year. If a student transferred schools, attendance rate was based on the school the student was enrolled in the longest.
Number of Suspensions (PK–2)	Count of both in-school and out-of-school suspensions in each year from PreK to 2nd grade. We combined these categories due to the structure of the restorative justice intervention program, Succeed Boston, a commonly used program for disciplinary measures in BPS in our study years, where a majority of students in our sample attended during the study period. Succeed Boston operates similarly to an in-school suspension program by providing students with "restorative alternatives to out-of-school suspension" (Succeed Boston, 2025). However, under Massachusetts state reporting requirements, these interventions are classified as out-of-school suspensions because students are not physically present in their school buildings during the program. Consequently, the reported data contain a much higher number of out-of-school suspensions compared to in-school suspensions.
Grade Retention (PK–2)	Binary indicator of whether student was retained (i.e., repeated a grade) in a given year from PreK to 2nd grade.
Free/Reduced Price Lunch (PK–2)	Binary indicator of whether student qualified for free or reduced-price lunch in each year.
Special Education Status (PK–2)	Binary indicator of whether student received special education services in each year.
School Attended (PK–2)	The BPS school where the student was enrolled each year. Students may have attended multiple schools across years. For students who transferred schools in the middle of a school year, we used the school the student was enrolled in the longest. There were a total of 78 unique schools in the sample's PreK year, 222 in kindergarten, 349 in 1st grade, and 421 in 2nd grade.
Race/Ethnicity	Set of binary indicators for whether student identified as Black, Hispanic, Asian, White, or multiracial/other, based on administrative records.
Sex	Binary indicator for student's sex, coded as male or female based on student administrative records during their PreK year.
Dual Language Learner	Binary indicator for whether student was a dual language learner.
Cohort Year	Categorical variable indicating the student's PreK eligibility cohort year. There are 4 cohorts in total.

Appendix S4. Benchmarking Comparisons and Justification for Selecting XGBoost and SMOTE Model

Justification for Model Selection

In the main manuscript, we focused on comparing XGBoost + SMOTE against logistic regression, which is the algorithm currently used in Massachusetts early warning systems (EWSs) and thus represents a policy-relevant baseline for our education research audience. Given space constraints, we have prioritized explaining XGBoost accessibly rather than extensively detailing benchmarking details across a wide range of machine learning algorithms.

However, prior to finalizing XGBoost as our focal method, we conducted a series of benchmarking analyses to compare its performance against several alternative machine learning models. Specifically, we evaluated:

- Random Forest (RF)
- Linear Support Vector Machine (SVM)
- Radial Basis Function (RBF) SVM
- AdaBoost
- LightGBM
- Lasso logistic regression
- A stacked model combining lasso and RF

These models were selected because they are widely regarded as strong-performing classifiers in predictive analytics research (Bertsimas & Dunn, 2017; Cortes & Vapnik, 1995; Freund & Schapire, 1997; Fernández-Delgado et al., 2014; Ke et al., 2017; Singh, 2025).

Benchmarking Results

Table S3.1 summarizes the performances of XGBoost and the benchmarking models using predictors from PreK to 2nd grade and SMOTE for oversampling:

Table S4.1: Model Performance for XGBoost+SMOTE and Benchmarking Algorithms

Algorithm	**Performances**					**Hyperparameters**
	Accuracy	**Recall/TPR**	**Specificity/TNR**	**BER**	**AUC**	
Logistic Regression (no SMOTE, baseline model comparison)	0.91	0.103	0.993	0.452	0.851	None
Logistic Regression + SMOTE	0.896	0.317	0.956	0.364	0.801	None
XGBoost + SMOTE	0.879	0.643	0.904	0.227	0.890	max_depth': [5, **7**], learning_rate': **[0.1,** 0.2], n_estimators': [**100,** 500], 'gamma': [**10**]
Linear SVM + SMOTE	0.861	0.579	0.890	0.265	0.833	'C': [**0.01,** 0.1, 1, 10, 100]
RBF SVM + SMOTE	0.878	0.413	0.927	0.330	0.821	C': [0.01, 0.1, 1, **10,** 100], 'gamma': [**'scale',** 0.01, 0.1, 1, 10]
AdaBoost + SMOTE	0.871	0.643	0.895	0.231	0.865	n_estimators': [50, 100, **200**], 'learning_rate': [0.5, **1.0**]

LightGBM + SMOTE	0.906	0.492	0.949	0.280	0.872	n_estimators': [100, **200**], 'learning_rate': [0.05, **0.1],** 'num_leaves': [31, **63]**
Lasso + SMOTE	0.870	0.548	0.904	0.274	0.834	{'C': [0.001, 0.01, **0.1,** 1, 10]}
RF + SMOTE	0.895	0.574	0.928	0.249	0.879	n_estimators': [100, **200**], max_depth': [**None,** 10, 20], min_samples_split': [**2,** 5]
Stacked Lasso/RF + SMOTE	0.896	0.452	0.942	0.303	0.873	lasso__C': [0.01, **0.1,** 1, 10], rf__n_estimators': [100, **200**], rf__max_depth': [**None,** 10], rf__min_samples_split': [**2,** 5], final_estimator__C': [0.01, 0.1, **1]**

Note: Given that AdaBoost and random forest showed similar performance metrics in initial runs, we conducted additional analyses to assess the stability of these results across different random seeds. Specifically, we reran each model using 10 different random state initializations, and averaged their performance metrics, which we indicate as "(avg)" in the table for applicable algorithms.

We selected XGBoost with SMOTE as our final predictive model based on a combination of balanced performance, robustness, and scalability across multiple metrics. While both XGBoost and AdaBoost performed competitively and had the two highest recall rates (0.643), XGBoost ultimately had a higher overall accuracy, specificity, and AUC, along with a slightly lower BER. In addition to its predictive strength, XGBoost provides practical advantages. It includes built-in regularization (L1 and L2), which helps prevent overfitting and improves generalization across different data contexts. It is also known to be more robust to noisy data and outliers, unlike AdaBoost, which emphasizes misclassified observations in each iteration and can become distorted by mislabeled or difficult-to-classify students (Hastie, Tibshirani, & Friedman, 2009). Finally, XGBoost was built for speed and scalability, with optimized C++ back-end code and native support for parallel processing, making it well-suited for use in real-time or district-wide predictive systems. This means XGBoost is less likely to overfit the training data and can scale up more efficiently when deployed on larger student datasets, such as those used by district-level student information systems.

That said, we acknowledge that this paper could just as easily have focused on AdaBoost since the performance gains of XGBoost over AdaBoost was so slight. Our choice to highlight XGBoost was driven not by a belief that it will always outperform other algorithms, but by a desire to provide a detailed, pedagogically clear example of a robust, widely used boosting model that could perform well across other education use cases as well. In fact, with additional tuning of AdaBoost, it is very well possible it could beat XGBoost (or vice versa). Taken together, our results should be interpreted not only as support for XGBoost specifically, but as an endorsement of the predictive power and practical promise of boosting models more broadly in early warning system development.

Robustness to Training-Test Data Split Method

In the original analysis, we used an 80/20 random split for the purposes of our proof-of-concept of XGBoost and SMOTE. This design choice was motivated by two factors: First, given the relatively small sample size (6,698 students, ~600 chronically absent cases), an 80/20 split also allowed us to maximize training data while maintaining a sizable test set. Second, the historical nature of the dataset (2007-08 to 2010-11) made current deployment impractical without

retraining on post-pandemic cohorts. Given the substantial structural changes to schooling post-pandemic (e.g., remote learning, shifts in attendance patterns post-pandemic), the models trained on pre-pandemic data should be viewed as illustrative rather than directly deployable.

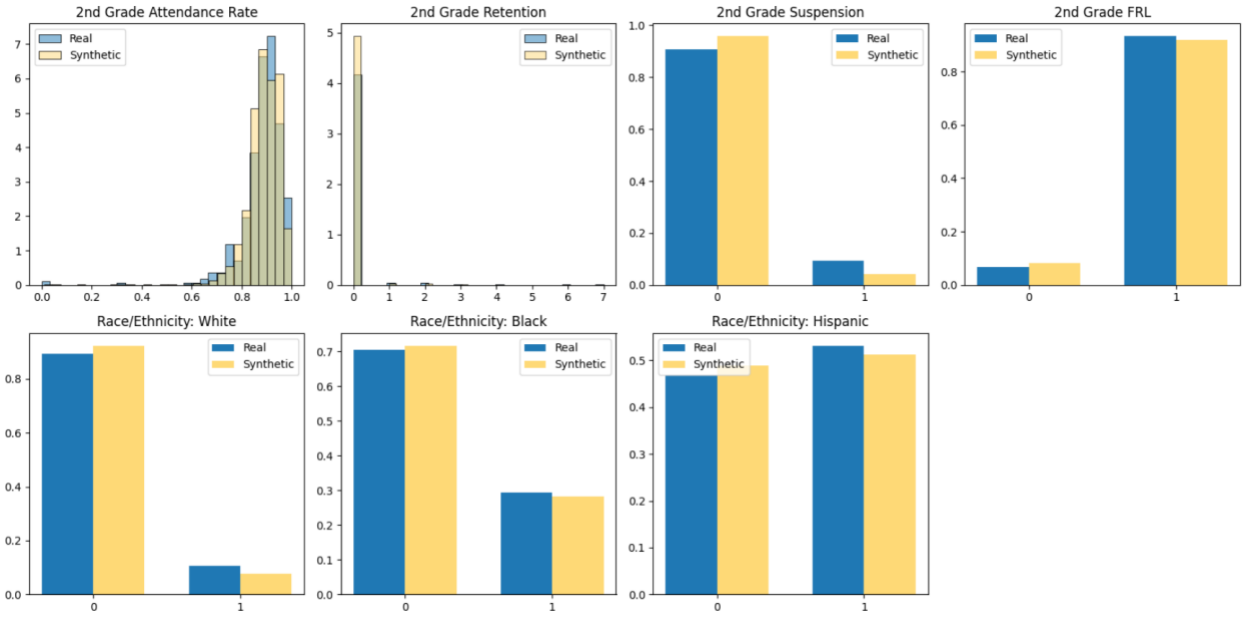
To fully address concerns about potential over-optimism, we implemented a rolling-origin validation approach, training models on earlier cohorts (2008-2010) and testing on the future cohort (2011). Below, the results show that the XGBoost and SMOTE model remained the most balanced model, offering the strongest tradeoff between recall and specificity and the lowest BER. While the traditional logistic regression with SMOTE achieved the highest recall (0.969), its extremely low specificity (0.099) makes it impractical for EWSs where both false positives and false negatives matter. The Lasso regression performance, while better than the regular logistic regression, fell into the same category of high recall at the cost of specificity. The Random Forest and SMOTE model had the best overall accuracy and second highest AUC, but it came at the cost of decreased recall. AdaBoost and the stacked model performed competitively compared to XGBoost. Both had better recall compared to the XGBoost model. The stacked model, which combined Lasso and Random Forests, achieved the lowest BER (0.218) and strong recall (0.704). However, this came at a moderate cost to specificity (0.859 for the stacked model and 0.861 for AdaBoost) and increased model complexity for the stacked model. In contrast, XGBoost + SMOTE offered higher specificity (0.919), a higher AUC, a higher overall accuracy, and nearly equivalent BER (0.239), while maintaining a simpler and more interpretable structure.

Taken together, we believe this shows that XGBoost + SMOTE demonstrated a stronger trade-off between recall, specificity, and practicality. However, we also recognize that there may be use cases where the stacked model or AdaBoost's higher recall or slightly lower BER could be prioritized, particularly if the goal is to cast a wider net and ensure that fewer at-risk students are missed. Ultimately, model selection should align with the operational priorities and resource constraints of the local educational context. For our purposes, balancing interpretability, specificity, and overall performance, XGBoost + SMOTE represented the most generalizable modeling strategy.

Table S4.2: Rolling-Origin Validation Results

Algorithm	Performances					Hyperparameters
	Accuracy	Recall/TPR	Specificity/TNR	BER	AUC	
Logistic without SMOTE	0.906	0.000	1.000	0.500	0.835	None
Logistic + SMOTE	0.181	0.969	0.099	0.466	0.728	None
XGBoost + SMOTE	0.889	0.604	0.919	0.239	0.889	max_depth': [**5,** 7], learning_rate': [**0.2**], n_estimators': [**100,** 500], 'gamma': [**10**],
Lasso + SMOTE	0.338	0.931	0.276	0.396	0.817	{'C': [0.001, 0.01, **0.1**, 1, 10]}
RF + SMOTE	0.898	0.553	0.934	0.256	0.869	n_estimators': [100, **200**], max_depth': [**None,** 10, 20], min_samples_split': [**2**, 5],
Stacked Lasso/RF + SMOTE	0.845	0.704	0.859	0.218	0.838	lasso__C': [0.01, **0.1,** 1, 10], rf__n_estimators': [100, **200**], rf__max_depth': [**None,** 10], rf__min_samples_split': [**2,** 5], final_estimator__C': [0.01, 0.1, **1**]
AdaBoost + SMOTE	0.834	0.667	0.851	0.241	0.806	n_estimators': [50, 100, **200**], 'learning_rate': [0.5, **1.0**]

Appendix S5. Empirical Verification of SMOTE Samples



To empirically verify the quality of the synthetic samples generated via SMOTE, we compared the distributions of key predictors between real and synthetic minority class instances. For continuous variables (e.g., 2nd grade attendance rate), we used kernal density estimation (KDE) plots to examine whether synthetic values fall within plausible ranges of the real data. For binary predictors (e.g., race/ethnicity), we used side-by-side bar plots to compare the proportions of 0s and 1s. Our findings suggest that SMOTE-generated samples broadly reflect the characteristics of the original minority class, supporting its use in our modeling pipeline.

Appendix S6. Optimal Hyperparameters & Tuning Process for XGBoost Models

Algorithm	Inputs				Optimal Hyperparameters
	PreK	K	1st	2nd	
XGBoost	✓				Depth of trees = 7, Learning rate = 0.2, # trees = 100, Gamma = 5
XGBoost		✓			Depth of trees = 7, Learning rate = 0.01, # trees = 100, Gamma = 5
XGBoost			✓		Depth of trees = 5, Learning rate = 0.01, # trees = 100, Gamma = 5
XGBoost				✓	Depth of trees = 5, Learning rate = 0.01, # trees = 100, Gamma = 5
XGBoost	✓	✓	✓	✓	Depth of trees = 5, Learning rate = 0.01, # trees = 100, Gamma = 5
XGBoost + SMOTE	✓				Depth of trees = 7, Learning rate = 0.2, # trees = 100, Gamma = 5
XGBoost + SMOTE		✓			Depth of trees = 7, Learning rate = 0.2, # trees = 100, Gamma = 5
XGBoost + SMOTE			✓		Depth of trees = 7, Learning rate = 0.2, # trees = 100, Gamma = 5
XGBoost + SMOTE				✓	Depth of trees = 7, Learning rate = 0.1, # trees = 100, Gamma = 5
XGBoost + SMOTE	✓	✓	✓	✓	Depth of trees = 7, Learning rate = 0.1, # trees = 100, Gamma = 10
XGBoost + SMOTE w/ Only Attendance Rate	✓	✓	✓	✓	Depth of trees = 7, Learning rate = 0.1, # trees = 100, Gamma = 10

We tuned each XGBoost model via a grid search within 5-fold cross-validation for four core hyperparameters described by Chen & Guestrin (2016). The hyperparameters were the maximum depth of trees (max_depth), learning rate (eta), number of trees (n_estimators), and the minimum loss reduction required to make further splits (gamma). Initially, we set the hyperparameters to default values recommended by Chen & Guestrin (2016): max_depth = 6, eta = 0.3, n_estimators = 100, and gamma = 0. We systematically explored alternative values around these defaults using a grid search procedure. The evaluation metric guiding model selection was the Area Under the Precision-Recall Curve.

We inspected the cross-validation performance for signs of under- and over-fitting to guide other hyperparameter values to try. For example, while we had initially started with a gamma of 0 in our first grid search, we eventually chose to initiate our tuning procedure from gamma = 5 based on preliminary analyses indicating the potential for overfitting to the training data when gamma was lower. Starting with gamma = 5 provided a more conservative baseline, effectively limiting unnecessary splits from the outset. This strategy streamlined our grid search and facilitated identifying models that were both parsimonious and robust across grades.

We believe the best XGBoost model hyperparameter configuration strikes a balance between capturing non-linear patterns in multi-year attendance data (max depth of 7), avoiding excessive ensemble size (100 trees), learning quickly enough to converge within those 100 rounds (learning rate of 0.1), and pruning weak splits (gamma of 5).

Appendix S7. Top 15 SHAP Interaction Values

Feature 1	Feature 2	Interaction Value
PreK Attendance Rate	2nd Grade Attendance Rate	0.1095
Kindergarten Attendance Rate	2nd Grade Attendance Rate	0.1083
1st Grade Attendance Rate	2nd Grade Attendance Rate	0.0943
PreK Attendance Rate	1st Grade Attendance Rate	0.0545
PreK Attendance Rate	Kindergarten Attendance Rate	0.0419
Kindergarten Attendance Rate	1st Grade Attendance Rate	0.0320
Race: Asian	2nd Grade Attendance Rate	0.0303
2nd Grade Attendance Rate	Cohort 1	0.0214
Race: Black	Dual Language Learner	0.0196
Special Education in 2nd Grade	1st Grade Attendance Rate	0.0115
Cohort 4	School A in 2nd Grade	0.0115
2nd Grade Attendance Rate	School A in Kindergarten	0.0106
Kindergarten Attendance Rate	School A in 2nd Grade	0.0093
Special Education in 2nd Grade	2nd Grade Attendance Rate	0.0083
Dual Language Learner	PreK Attendance Rate	0.0082

Since SHAP values reflect a feature's marginal contribution across all possible subsets of input features, they incorporate both the feature's individual contribution and interaction effects with other features (Lundberg & Lee, 2017). However, in tree-based models like XGBoost, the SHAP values used to calculate feature importance can sometimes favor continuous variables, which have more potential split points than binary variables. To assess whether the high importance of attendance rate predictors reflected meaningful contributions rather than this modeling artifact, we computed SHAP interaction values as a robustness check here. SHAP interaction values isolate the true interaction strength between features, rather than attributing all predictive gain to the feature that simply splits more often (which tends to be continuous). Even if a continuous feature splits more often in the tree, the pairwise interaction matrix reveals whether those splits are interacting meaningfully with other features or just coincidentally fitting noise.

In the above table, each interaction value shows the average absolute interaction effect between two features, capturing how much the combined presence of both features shifts model predictions, beyond what each would contribute on its own. High values indicate that the model is capturing meaningful interactions between those two features in its predictions. For instance, the interaction with the highest interaction value is PreK Attendance Rate x 2nd Grade Attendance Rate. That means that the impact of 2nd grade attendance on predicted chronic absenteeism really depends on how the student attended in PreK and suggests that the XGBoost model learns trajectory patterns; a dip in attendance in 2nd grade is more concerning if PreK attendance was also low and multi-year attendance decline would be more predictive of risk than a single-year anomaly.

In contrast, interactions involving fixed demographic characteristics had substantially lower interaction values. While the interactions still suggest that certain student groups could experience differential effects in how their chronic absenteeism risk is computed, their more limited interaction values suggest that the model's risk assessments are driven primarily by malleable, longitudinal attendance trajectory indicators. This finding supports the substantive importance of attendance predictors over static demographic factors in absenteeism prediction and helps alleviate concerns about structural favoritism toward continuous features while using SHAP values.

Appendix S8. Sample Code for XGBoost with SMOTE

```
# Import necessary libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
import seaborn as sns
import matplotlib as mpl
import matplotlib.pyplot as plt
import xgboost as xgb
from sklearn.impute import SimpleImputer
pd.set_option('display.max_rows', None)

# Read in datset
training = pd.read_csv("dataset.csv")

# Splitting data into training (80%) and test (20%) set.
from sklearn.model_selection import train_test_split
training_train, training_test = train_test_split(training,
                                                 test_size=0.2,
                                                 stratify=training['chronic_absence_fy4'], #fy4 is 3rd grade
                                                 random_state=28)

# Split into X and y
X_train = training_train.drop(['chronic_absence_fy4'], axis=1)
X_test = training_test.drop(['chronic_absence_fy4'], axis=1)
y_train = training_train.chronic_absence_fy4
y_train_df = training_train.loc[:, ['chronic_absence_fy4']]

y_test = training_test.chronic_absence_fy4
y_test_df = training_test.loc[:, ['chronic_absence_fy4']]

# SMOTE-NC
from imblearn.over_sampling import SMOTE
from imblearn.over_sampling import SMOTENC
from sklearn.impute import SimpleImputer

# Apply SMOTE-NC to training data
smotenc = SMOTENC(categorical_features=categorical_feature_indices, random_state=28, sampling_strategy = 0.7)
X_train_resampled, y_train_resampled = smotenc.fit_resample(X_train, y_train)

# Convert the resampled numeric data back to a DataFrame
X_train_resampled_df = pd.DataFrame(X_train_resampled, columns=X_train.columns)

# XGBoost cannot take categorical vars, so we need to one-hot encode
# Identify categorical variables
categorical_vars = X_train_resampled_df.select_dtypes(include=['object', 'category'])

# Perform one-hot encoding
X_train_encoded = pd.get_dummies(X_train_resampled_df, columns=categorical_vars.columns)

# View the encoded dataframe
print(X_train_encoded.head())

# Do the same for X_test
# XGBoost cannot take categorical vars, so we need to one-hot encode
# Identify categorical variables
categorical_vars = X_test.select_dtypes(include=['object', 'category'])

# Perform one-hot encoding
X_test_encoded = pd.get_dummies(X_test, columns=categorical_vars.columns)

```

```
# Compare column sets
train_columns_set = set(X_train_encoded.columns)
test_columns_set = set(X_test_encoded.columns)

# Check if the column sets are equal
if train_columns_set == test_columns_set:
    print("X_train_encoded and X_test_encoded have the same columns.")
else:
    print("X_train_encoded and X_test_encoded do not have the same columns.")

# Make both df's have the same column order or else XGBoost won't run
# Get the column order from X_train_encoded
column_order = X_train_encoded.columns

# Reorder the columns in X_test_encoded
X_test_encoded = X_test_encoded[column_order]

# Disable warnings
import warnings
warnings.filterwarnings('ignore')

import xgboost as xgb
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import make_scorer, roc_auc_score

# Define XGBoost model
xgb_model = xgb.XGBClassifier(
    objective='binary:logistic',
    seed=28,
    eval_metric='aucpr',
    use_label_encoder=False  # suppresses a warning message
)

# Set up the hyperparameter grid
param_grid = {
    'max_depth': [3, 5, 7],
    'learning_rate': [0.01, 0.1, 0.2],
    'n_estimators': [100, 500, 1000],
    'gamma': [5, 10, 20],
  #  'early_stopping_rounds': [10, 20],
  #  'missing': ['nan'],
  #  'reg_alpha': [0, 0.1],
  #  'reg_lambda': [0, 0.1, 0.5, 1],
  #  'subsample': [0.6, 0.8, 1.0],
  #  'colsample_bytree': [0.6, 0.8, 1.0]
}

# Set up the scorer for GridSearchCV
scorer = make_scorer(roc_auc_score)

# Perform GridSearchCV
grid_search = GridSearchCV(estimator=xgb_model, param_grid=param_grid, scoring=scorer, cv=5)
grid_search.fit(X_train_encoded, y_train_resampled)

# Print the best hyperparameters and the corresponding ROC-AUC score
print("Best Hyperparameters: ", grid_search.best_params_)
print("Best ROC-AUC Score: ", grid_search.best_score_)

# Make predictions on the test data using the best model
best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test_encoded)

# Calculate accuracy
accuracy = (y_test == y_pred).mean()
print('Accuracy:', accuracy)
```

```
126.
127. # Predict class probabilities for test data
128. y_prob = best_model.predict_proba(X_test_encoded)[:, 1]
129.
130. from sklearn.metrics import roc_auc_score, balanced_accuracy_score, classification_report
131.
132. # Calculate AUC
133. auc = roc_auc_score(y_test, y_prob)
134. print('AUC:', auc)
135.
136. # Calculate BER
137. y_pred = best_model.predict(X_test_encoded)
138. ber = 1 - balanced_accuracy_score(y_test, y_pred)
139. print('BER:', ber)
140.
141. # Calculate recall
142. report = classification_report(y_test, y_pred, target_names=['Negative', 'Positive'], digits =
3)
143. print('Recall:\n', report)
144.
145. # Calculate accuracy
146. accuracy = (y_test == y_pred).mean()
147. print('Accuracy:', accuracy)
148.
149. # Define the probability thresholds to test
150. thresholds_to_test = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
151.
152. # Initialize lists to store evaluation metrics for each threshold
153. accuracy_scores = []
154. auc_scores = []
155. ber_scores = []
156. recall_positive_scores = []
157. recall_negative_scores = []
158.
159. # Loop through each threshold and calculate metrics
160. for threshold in thresholds_to_test:
161.     # Apply the threshold to the predicted probabilities
162.     y_pred_custom_threshold = (y_prob > threshold).astype(int)
163.
164.     # Calculate metrics
165.     accuracy_custom_threshold = (y_test == y_pred_custom_threshold).mean()
166.     auc_custom_threshold = roc_auc_score(y_test, y_prob)
167.     ber_custom_threshold = 1 - balanced_accuracy_score(y_test, y_pred_custom_threshold)
168.     report_custom_threshold = classification_report(y_test, y_pred_custom_threshold,
target_names=['Negative', 'Positive'], output_dict=True)
169.
170.     # Append metrics to lists
171.     accuracy_scores.append(accuracy_custom_threshold)
172.     auc_scores.append(auc_custom_threshold)
173.     ber_scores.append(ber_custom_threshold)
174.     recall_positive_scores.append(report_custom_threshold['Positive']['recall'])
175.     recall_negative_scores.append(report_custom_threshold['Negative']['recall'])
176.
177. # Create a DataFrame to store the metrics for each threshold
178. results_df = pd.DataFrame({
179.     'Probability Threshold': thresholds_to_test,
180.     'Accuracy': accuracy_scores,
181.     'AUC': auc_scores,
182.     'BER': ber_scores,
183.     'Recall (Positive)': recall_positive_scores,
184.     'Recall (Negative)': recall_negative_scores
185. })
186.
187. # Display the results DataFrame
188. print(results_df)
```

```
189.
190. # Export the results DataFrame to an Excel file
191. results_df.to_excel('XGBoost_probthresholds.xlsx', index=False)
192.
```