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

Chronic absenteeism is a critical issue that has been linked to many adverse student outcomes. The current study focuses on improving a key system already in place in many school districts—early warning systems (EWSs)—in order to decrease chronic absenteeism in students' earliest schooling years. Using a demographically diverse population 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 improve EWS accuracy in proactively identifying students who are at risk of becoming chronically absent. The best-performing XGBoost model with SMOTE was approximately 52 percentage points more accurate (in terms of recall rate) than the logistic regression model closest to those used in current EWSs in correctly predicting students who would be chronically absent in third grade. Our analyses introduce varying probability thresholds and the incorporation of different years of data, showing the potential of these models to cater to school districts aiming to leverage machine learning predictions while adhering to budgetary or intervention constraints.

VERSION: November 2024

Suggested citation: Wu, Tiffany, and Christina Weiland. (2024). 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

Chronic absenteeism is a critical issue that has been linked to many adverse student outcomes. The current study focuses on improving a key system already in place in many school districts—early warning systems (EWSs)—in order to decrease chronic absenteeism in students' earliest schooling years. Using a demographically diverse population 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 improve EWS accuracy in proactively identifying students who are at risk of becoming chronically absent. The best-performing XGBoost model with SMOTE was approximately 52 percentage points more accurate (in terms of recall rate) than the logistic regression model closest to those used in current EWSs in correctly predicting students who would be chronically absent in third grade. Our analyses introduce varying probability thresholds and the incorporation of different years of data, showing the potential of these models to cater to school districts aiming to leverage machine learning predictions while adhering to budgetary or intervention constraints.

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

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 potential 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)—in order to reduce 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, current EWS models may lack the precision needed for accurate and equitable predictions due to analytical challenges (Sansone, 2019). 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 machine learning holds the potential to overcome these limitations and become a formidable tool capable of improving EWSs. While machine learning 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 explaining why they can be used to enhance the ability of EWSs to provide more proactive and 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. Moreover, our models incorporate varying probability thresholds for prediction to cater to districts aiming to leverage machine learning predictions for early intervention while adhering to budgetary or intervention constraints. Lastly, 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 many individual and family characteristics—such as children's race, health, and socioeconomic status—with higher rates of student absenteeism (Allensworth et al., 2021; Ansari & Gottfriend, 2018; Gottfried & Ansari, 2021; Klein et al., 2020; Purtell & Ansari, 2022). However, although demographic factors play a crucial role in understanding patterns of student absenteeism, many of these factors are not easily malleable, especially in the short term. Schools and policy makers therefore often face limitations to effectively intervening within the constraints of systemic inequity (Gottfried & Hutt, 2019; Kearney et al., 2019).

Early Warning Systems (EWS)

As a result, one strand of attendance research focuses on discovering strategies individual schools and districts can take an active role in diminishing their own rates of absenteeism (Gottfried & Hutt, 2019). Many districts and states have thus 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 failing a course during the semester (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 making the model non-linear, traditional regressions still have poor predictive accuracy compared to more modern machine learning algorithms (Deussen et al., 2017; Sansone, 2019).

These methodological constraints underscore the need for using 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 of students at risk of chronic absenteeism. 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 of absenteeism. This imbalance hampers the regression model's ability to accurately predict the minority of students who are genuinely at risk. Class imbalance is better addressed through non-traditional machine learning algorithms like the synthetic minority oversample technique (SMOTE), 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 students missing more than 10 and 15 percent of school days), behavior (indicators for students who are suspended or expelled at all and for three or more days), and course performance (indicators for failing core courses, failing math courses, and failing English/language arts courses), which have become known as the ABC indicators (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 lack of indicators able to be used in the early years could consequently result in less reliable predictions, but more research is needed examining the predictive accuracy of early EWSs.

Given these limitations, schools that implement EWSs for early grades make utilize available indicators, but they may not focus explicitly on predicting students' risk of chronic absenteeism. For example, the Massachusetts EWS uses student demographics, enrollment, attendance, and suspension indicators to evaluate the risk of students not achieving proficient reading levels by the end of third grade (Massachusetts Department of Elementary and Secondary Education & American Institutes for Research, 2013). While focusing on reading proficiency is important and valuable, it may inadvertently overlook the underlying causes 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). Moreover, attendance rate is one of the few ABC indicators consistently tracked from early childhood. By focusing on identifying absenteeism early alongside reading proficiency levels in third grade, schools can not only enhance the predictive accuracy of their early EWS models but also tackle the foundational barriers to learning, thereby improving overall student engagement and long-term academic success.

Despite this potential, 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.

Boston Public Schools Attendance Policy

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. The attendance policy was first revised in 2006 and 2007 to discontinue the use of cutoff times to refuse students' entry into schools and give individual schools more flexibility to promote attendance. 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 the onset of the chronic absenteeism measure (Boston Public Schools, 2022).[1]

Under BPS's attendance policy, a student must be a school for at least half the day in order to be counted as "present." In most schools, a half day means 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

[1] The attendance policy was further revised in 2018 to include cultural and religious holidays as excused absences and in 2021 to discontinue the policy of converting tardies into absences along with issuing "No Credit" grades based on attendance. This period of time is outside our study period.

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. During the academic year, schools use the Aspen system to take attendance and the Panorama Student Success Platform to document absenteeism prevention and intervention plans, the latter of which is used by over 2,000 other school districts in the nation (Panorama Education, 2023). Student attendance data from Aspen is transmitted to Panorama every night. BPS uses this attendance monitoring process to inform its Tiered Attendance System. Tier 1 typically consists of universal attendance-promoting strategies such as reliable attendance reporting, developing a positive school climate, and maintaining positive relationships with parents. Tier 2 consists of more targeted intervention supports such as attendance letters for absent students, attendance contracts, student-family conferences, and mentoring programs for "emerging absenteeism." Tier 3 consists of intensive interventions for severe absenteeism, such as specialized programming or referrals to support services (Boston Public Schools, 2022).

The state of Massachusetts has their own version of an EWS, called the Early Warning Indicator System (EWIS), which was established in 2011 and designed to identify students who may require additional attention to achieve certain academic milestones between 1st and 12th grade (Massachusetts Department of Elementary and Secondary Education, 2023). The EWIS does this by providing risk levels for each student pertaining to meeting the milestone, which it updates annually. The Massachusetts Department of Elementary and Secondary Education then publishes the risk levels on Edwin Analytics, works to make EWIS known to school districts, and provides technical assistance in school districts that use EWIS data. For early elementary students in first to third grade, the milestone focus is meeting or surpassing expectations on the 3rd-grade ELA Massachusetts State Assessment.

The EWIS risk levels are determined through an annually updated multilevel logistic regression model using various indicators with data sourced from existing state-wide collections (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 methodological limitations associated with traditional regression approaches. Furthermore, a 2019 case study found that awareness and 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).

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 EWSs 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. Our hope is that our machine learning models' superior predictive accuracy will enhance the reliability and relevance of EWSs. When predictions are accurate, 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. More broadly, we hope to provide a gentle introduction to SMOTE and XGBoost and enhance understanding of their applicability for decreasing chronic absenteeism. 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 factors contribute to the differences in prediction accuracy between these approaches?
2. How accurately and early can we identify students who will be chronically absent in 3rd grade?
3. How can models be personalized to inform policies regarding chronic absenteeism intervention while taking into account an institution's financial and resource constraints?

Method

Sample

Our sample for this paper is 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. The program began in 2005 with then-Mayor

Thomas Menino's vision of offering free public universal PreK education to all four-year-olds in Boston. It is open to any child in the city, regardless of income, and has garnered a high profile in the past 15 years due to its attention to evidence-based practices (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% was 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. A table of student characteristics for the full sample 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 collect to make our analyses 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[2], 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 and whether the student was female or male.

[2] In the 2014-2015 school year, Massachusetts revised its definition of "low income" and introduced a new income status metric with a slightly different measurement approach However, since we use eligibility for FRL as a predictor variable only from PreK to 2nd grade, and our final cohort was enrolled in 2nd grade during the 2013-2014 school year, this change did not affect our sample.

Analytical Approach

We explain various ML algorithms along with the results from our data 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 S4.

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, we possess a labeled dataset where 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.

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 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 we want 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. In other words, there are not enough instances of the minority class for the model to "learn" how to classify them accurately.

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. SMOTE works by creating synthetic instances of the minority class, as opposed to creating the same minority observation multiple times with replacement. The algorithm does this by first identifying a minority class instance and its *k* nearest neighbors and then generating the synthetic instances along the line segments connecting the original instance with its neighbors. By introducing synthetic minority observations, SMOTE enhances the model's ability to capture patterns separating the minority from the majority class. Figure 2 illustrates a hypothetical example of a 2-D feature space before and after SMOTE. Before SMOTE, the chronically absent x's are sparse because of the class imbalance. This makes it difficult to classify these points. After SMOTE, 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 chose to use 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 regression tends to not be very flexible and make a linear classification boundary line unless the predictor variables contain an interaction.

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

$$logit\left(P\left(Y_{ij} = 1\right)\right) = \beta_0 + \delta_{ij} + \mu_j + \epsilon_{ij}$$

where *i* denotes student and j denotes school. Y_{ij} represents the third grade chronic absenteeism label for student *i* in school *j.* δ_{ij} represents the vector of student-level predictor variables. μ_j is the random intercept for school j. ϵ_{ij} is the student-level residual error. We included random intercepts for school to account for the nesting of students within schools. We chose this model specifically because it is the closest to what Massachusetts currently uses in their early warning system (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 and a model interacting student demographic characteristics.

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 evaluating and 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:

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

where y_i is the true label value (0 if not chronically absent, 1 if chronically absent), $\widehat{y}_i$ 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 $\widehat{y}_i^{(t-1)} + f_t(x_i)$. A small value for *l* means that the corrected predicted label is closer to y_i, leading to better accuracy. The regularized $\Omega(f_t)$ 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, Figure 3 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 also more robust against overfitting and outliers due to their ability to combine models and adjust hyperparameters. While they lack the interpretability of regression methods, the versatility and accuracy of XGBoosts makes it a valuable tool for education research in instances where accurate predictions are crucial for informed decision-making.

Analytical Process

We trained a total of 22 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. All missing features were imputed using MICE (Multivariate Imputation by Chained Equations; Van Buuren & Groothuis-Oudshoorn, 2011). We tuned each XGBoost model using 5-fold cross-validation for the hyperparameters of the number of trees, depth of trees, learning rate, and gamma, which is a regularization term that penalizes the complexity of the tree. Optimal hyperparameters were chosen by comparing the Area Under the Precision-Recall Curve for each classifier with different hyperparameter values. Optimal hyperparameters for each model are listed in Appendix S3.

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 will pay particular attention to recall/TPR and BER as the main performance metrics when assessing our different models. We emphasize recall/TPR because our focus is on predicting students who will be chronically absent and BER because BER accounts for both false positives and false negatives, making it more robust for evaluating model performance datasets with class imbalance.

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 6 rows are the results for the parsimonious logistic regression model, with row 4 (results from the model using just the second grade data) representing the model closest to that used in

Massachusetts's current EWSs for third grade. This is followed by the set of results for logistic regression with SMOTE, then for XGBoost without SMOTE, and finally for XGBoost with SMOTE.

For accuracy rate, the XGBoost model with predictors from school years PreK-2nd grade performed the best with 91.6% accuracy. However, most of the accuracy rates for all models hovered around 90%. For recall/TPR, the XGBoost models with SMOTE using predictors from all school years PreK-2nd grade had the highest rate at 62.7%. For specificity/TNR, the logistic regression using kindergarten predictors performed the best at 0.995. For BER, the XGBoost model with SMOTE using predictors for all school years performed best at 0.222. Finally, the model with the highest AUC at 0.891 was the XGBoost model with SMOTE using predictors for all school years. The best performing logistic and XGBoost models based on recall and BER are highlighted in darker grey in Table 4.

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 model using SMOTE with predictors from all grades. It had a recall of 0.627 and a BER of 0.222.

Regarding overall patterns in the results, the first six rows of logistic regression models in Table 4 had a high accuracy rate (around 90%) but hovered around 5-10% for recall, even when we add covariates from all school years PreK-2nd grade. The logistic regression with the best recall rate was the one which included all students' PreK-2nd grade data and interaction terms, with a recall rate of 0.151. This means that only a small percentage of students in the test sample who ended up being chronically absent in 3rd grade was identified as having a high risk of chronic absenteeism. The performance of these models, in particular the one using only the second grade predictors, is closest to approximating the performance of existing EWSs. Our results show that current EWSs that rely solely on logistic regression results may not have high predictive accuracy for identifying students who will be chronically absent even if they do include predictors from all students' past school years and interactions.

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 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 in our predictive ability. The tradeoff was 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 an early warning system. 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; Lee & Chung, 2019).

The second reason for the low recall rate is that the logistic regression models without an interaction term make a linear classification boundary which may 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 previously showed, 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 nonlinear logistic regression by including interaction terms. We do this in the sixth row 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 likely non-linear. 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 interactions terms in the logistic regression model with SMOTE also led to non-convergence errors, likely because of multicollinearity issues between the synthetic samples generated and our actual sample. Both these instances demonstrate the limits of using logistic regression for prediction purposes.

XGBoost models, on the other hand, do not run into this limitation. The higher recall rates from the XGBoost models provide evidence supporting the need for a more non-linear, non-parametric classification boundary line than a logistic regression is able to produce. The recall for the best XGBoost model was 62.7%, approximately 31 percentage points higher than the best performing logistic regression model with SMOTE (row 10 in Table 4) and 52.4 percentage points higher than the parsimonious logistic regression that best approximates the model used in current EWSs (row 4 in Table 4). This best-performing XGBoost model also had the lowest BER (0.222) and the highest AUC (0.891) out of all the models. 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 and Timeliness in Predicting 3rd Grade Chronic Absenteeism

As aforementioned, the best model (XGBoost using SMOTE with predictors from all grades) could predict third grade chronic absenteeism status with an overall accuracy of 90.1%

and a recall rate of 62.7%. While the recall rate is less than ideal, it is 31 percentage points higher than the recall rate of the top logistic regression model with SMOTE, close to a 50% improvement, and approximately 52 percentage points higher than the recall rate of the logistic regression model 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 will bolster the accuracy of the XGBoost and SMOTE models.

Surprisingly, relying solely on PreK data in the XGBoost and SMOTE model yielded a comparable recall rate of 58.7% (albeit at the cost of decreased specificity, resulting in an overall accuracy of only 81.4% for this configuration). This suggests that even data from children's earliest schooling experience has predictive power for later outcomes. Solely using second grade data led to a recall rate of 54% and accuracy of 88.5%. 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, the XGBoost and SMOTE model can still harness it to identify students at a heightened risk of chronic absenteeism without losing too much precision, especially compared to the accuracy and recall rates obtained by using logistic regression for those same years of data.

RQ3: Personalization of Machine Learning Models

Table 5 presents the performance metrics of the best-performing logistic regression model with SMOTE and the best-performing XGBoost model with SMOTE with varying probability thresholds. The default probability threshold for all models in Table 4 is 0.5 (greater than or equal to 0.5 means that a student is predicted to be chronically absent). 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 may opt to use a lower probability threshold of 0.4 or 0.5 to reach a wider range of students.

Importantly, as seen in Table 5, the XGBoost model performs better than the logistic regression at each given threshold based on recall and BER. Using a probability threshold of 0.8, for instance, the XGBoost model can correctly predict 31% of chronically absent students and 98% of non-chronically absent students. Table 5 demonstrates the ability to personalize machine learning models to inform school and district policies based on their specific needs, including accounting for financial or resource constraints. This adaptability equips educational institutions with a more flexible tool for shaping chronic absenteeism interventions.

Discussion & Conclusion

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. Notably, the top-performing XGBoost model with SMOTE outperformed the logistic regression model closest to that in current EWSs by approximately 52 percentage points and the best logistic regression model with SMOTE by approximately 31 percentage points in accurately forecasting chronic absenteeism among third-grade students. This finding aligns with Sansone's (2019) conclusion that machine learning tools provide more precise predictions compared to the logistic regression models used in many parsimonious early warning systems today. Furthermore, we introduce a table of performance metrics that incorporates flexible probability thresholds, demonstrating how EWSs could be a viable tool for helping decrease chronic absenteeism in early childhood while accounting for school budgetary limitations and intervention severity.

Despite the evidence supporting the use of XGBoost with SMOTE to improve existing early warning systems, there are several important limitations to our approach. First, there are many more hyperparameters we could tune in our XGBoost model that could enhance model performance, such as the maximum delta step and the lambda and alpha regularization terms. Future studies should explore additional hyperparameters that could help reduce overfitting and help the algorithm run faster. Additionally, there are many other supervised learning models such as random forest or neural networks that may yield better recall rates and AUC. Future studies should compare the performance of XGBoost with other modern machine learning models.

Another limitation of our study is that it does not specifically delve into the broader challenges associated with implementing early warning systems within educational institutions. This includes the task of training teams to interpret EWS outputs and the subsequent development of tailored interventions for students once they have been identified (Frazelle & Nagel, 2015). In 2014-15, about 52% of all public high schools nationwide already initiated some form of EWS (U.S. Department of Education, 2016b). A next step could entail the integration of modern machine learning-based algorithms into preexisting EWS frameworks to enhance predictive accuracy and partnering with schools and districts to test the efficacy of this proof of concept for machine learning algorithms. There are also worries that any type of EWS implementation could increase stereotyping of students and bias teachers in a way that will be a self-fulfilling prophecy for students (Brown, 2016). While our findings offer a compelling case for the accuracy of machine learning-enhanced EWSs, we acknowledge that these tools must be implemented with caution to avoid reinforcing existing biases.

Relatedly, it is essential to recognize the potential for bias in machine learning algorithms and how they can perpetuate or even amplify existing biases in EWSs (Feathers, 2023). Given that our predictive models use demographic and disciplinary information, we need to be careful to ensure they do not predict higher or lower risk unfairly for students from marginalized backgrounds. To ensure fairness, machine learning models should be rigorously tested for biases based on race, gender, socioeconomic status, or other demographic factors that may lead to unequal treatment. Therefore, an important next step is to evaluate the bias in our XGBoost and SMOTE models using 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. If unfairness is detected, strategies such as reweighting the training data or applying fairness constraints during model training should be examined to mitigate bias in predictions (Kamiran & Calders, 2012; Zafar et al., 2017).

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 and leveraging predictive analytics for outcomes beyond chronic absenteeism. It illuminates the advantages of integrating modern machine learning algorithms into the field of education, and we hope this paper also 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.

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.

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

Brown, E. (2016). Can “early warning systems” keep children from dropping out of school? *The Washington Post.* https://www.washingtonpost.com/local/education/can-early-warning-systems-keep-children-from-dropping-out-of-school/2016/06/21/853c5436-36ef-11e6-a254-2b336e293a3c_story.html

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*.

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*.

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.

Feathers, T. (2023, April 27). False alarm: How Wisconsin uses race and income to label students “high risk”. *The Markup*. https://themarkup.org/machine-learning/2023/04/27/false-alarm-how-wisconsin-uses-race-and-income-to-label-students-high-risk

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.

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.*

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.

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.

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.

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.

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.

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.

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.

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

Figures & Tables

Figure 1. ROC Curve

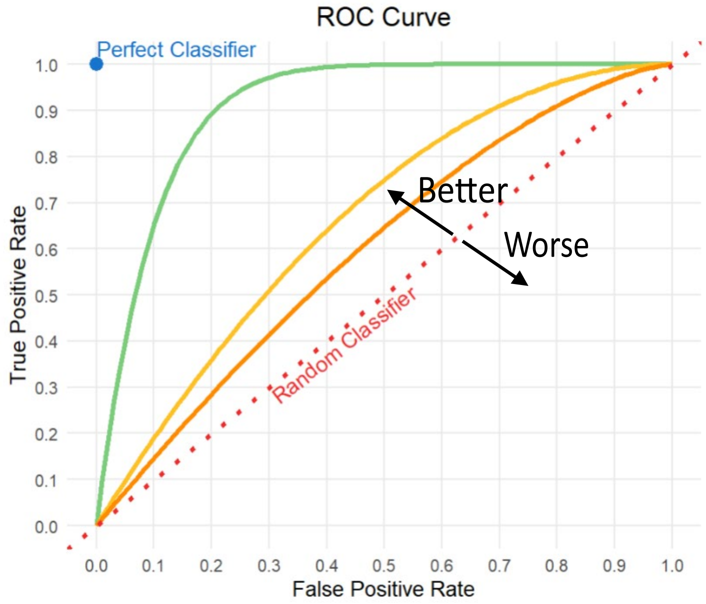
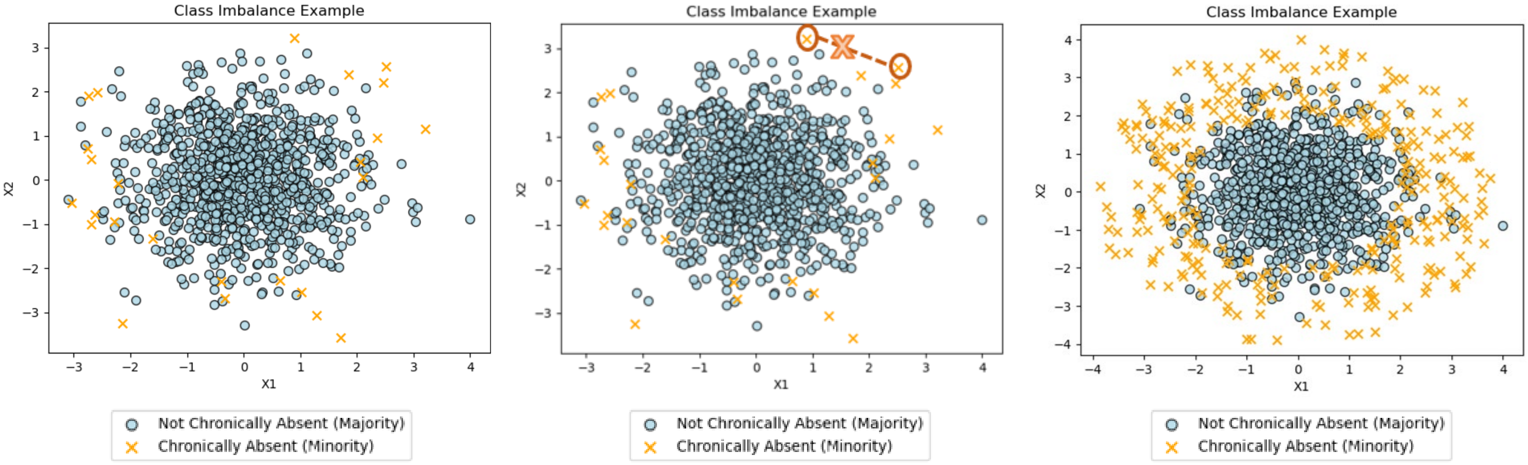


Figure 2. SMOTE Visualization



Before SMOTE

Creating 1 Synthetic Instance

After SMOTE

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

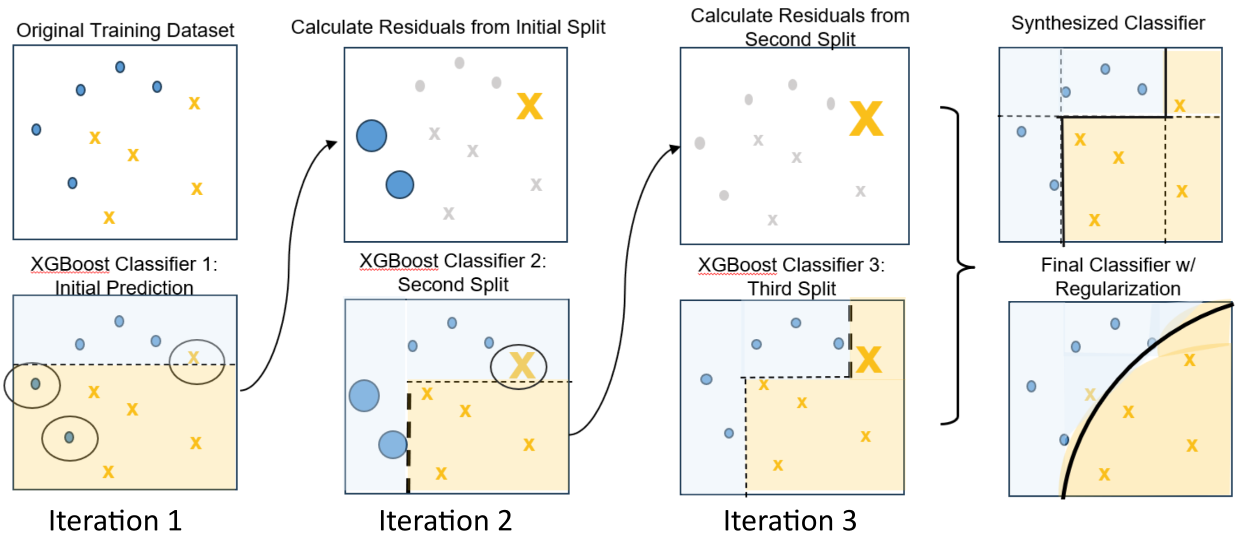
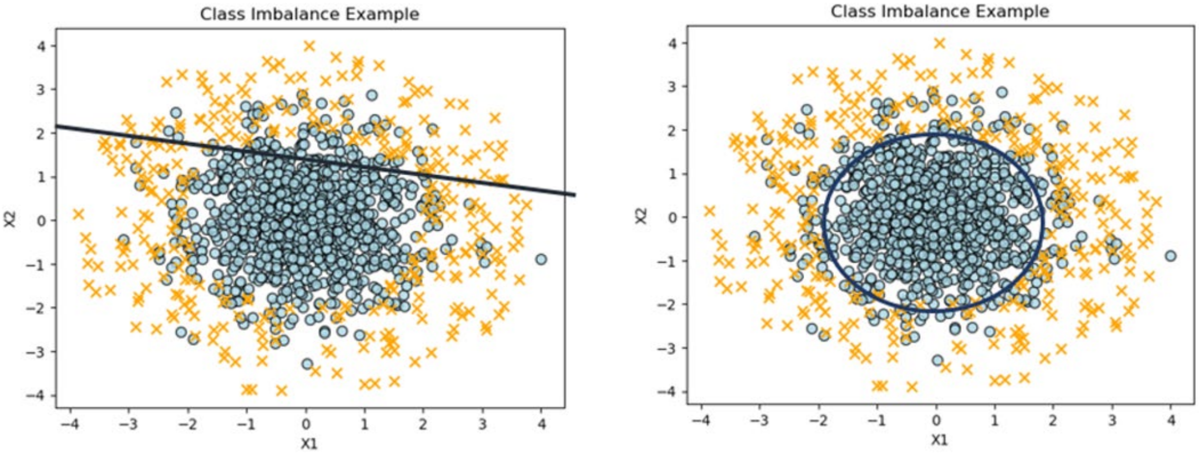

Original Training Dataset
Calculate Residuals from Initial Split
Calculate Residuals from Second Split
Synthesized Classifier
XGBoost Classifier 1: Initial Prediction
XGBoost Classifier 2: Second Split
XGBoost Classifier 3: Third Split
Final Classifier w/ Regularization
Iteration 1
Iteration 2
Iteration 3

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)

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

Algorithm	**Inputs**				**Performances**				
	PreK	**K**	**1st**	**2nd**	**Accuracy**	**Recall/TPR (Predicting CA)**	**Specificity/TNR (Predicting non-CA)**	**BER (want low)**	**AUC (want high)**
Logistic	✓				0.899	0.095	0.982	0.461	0.8
Logistic		✓			0.905	0.04	0.995	0.483	0.768
Logistic			✓		0.904	0.079	0.99	0.465	0.774
Logistic				✓	0.91	0.103	0.993	0.452	0.851
Logistic	✓	✓	✓	✓	0.905	0.127	0.986	0.444	0.847
Logistic w/ Interactions	✓	✓	✓	✓	0.908	0.151	0.987	0.431	0.845
Logistic + SMOTE	✓				0.881	0.286	0.943	0.386	0.75
Logistic + SMOTE		✓			0.875	0.254	0.939	0.403	0.729
Logistic + SMOTE			✓		0.881	0.333	0.938	0.364	0.747
Logistic + SMOTE				✓	0.896	0.317	0.956	0.364	0.801
Logistic + SMOTE	✓	✓	✓	✓	NC	NC	NC	NC	NC
Logistic + SMOTE w/ Interactions	✓	✓	✓	✓	NC	NC	NC	NC	NC
XGBoost	✓				0.901	0.04	0.991	0.485	0.825
XGBoost		✓			0.91	0.135	0.991	0.437	0.794
XGBoost			✓		0.906	0.183	0.982	0.418	0.812
XGBoost				✓	0.913	0.325	0.974	0.35	0.864
XGBoost	✓	✓	✓	✓	0.916	0.317	0.979	0.352	0.877
XGBoost + SMOTE	✓				0.814	0.587	0.838	0.287	0.819
XGBoost + SMOTE		✓			0.833	0.556	0.862	0.291	0.808
XGBoost + SMOTE			✓		0.851	0.524	0.885	0.296	0.808
XGBoost + SMOTE				✓	0.885	0.54	0.921	0.27	0.867
XGBoost + SMOTE	✓	✓	✓	✓	0.901	0.627	0.929	0.222	0.891

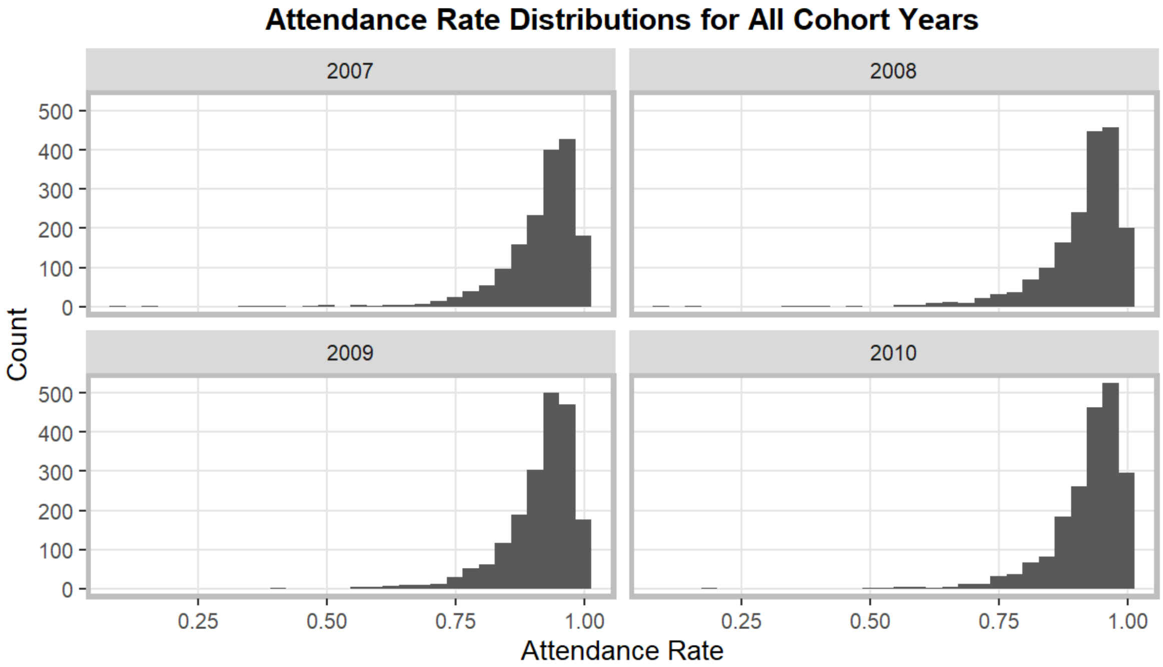
Note: CA stands for chronically absent. TPR stands for True Positive Rate. TNR stands for True Negative Rate. BER stands for Balanced Error Rate. AUC stands for Area under the ROC Curve. The logistic regression with interactions included interactions for all non-time-varying covariates: sex, race, and whether a student was a dual language learner. NC stands for no convergence, meaning the regression model fit was singular.

Table 5. Performance Metrics at Varying Probability Thresholds

Threshold	Accuracy	Recall/TPR	Specificity/TNR	BER	AUC
Best Logistic Regression (with SMOTE) Model					
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 (with SMOTE) Model					
0.1	0.699	0.905	0.678	0.209	0.891
0.2	0.804	0.833	0.801	0.183	0.891
0.3	0.848	0.770	0.856	0.187	0.891
0.4	0.879	0.714	0.896	0.195	0.891
0.5	0.901	0.627	0.929	0.222	0.891
0.6	0.904	0.508	0.945	0.274	0.891
0.7	0.912	0.405	0.965	0.315	0.891
0.8	0.914	0.294	0.979	0.364	0.891
0.9	0.916	0.183	0.992	0.413	0.891

Online-Only Appendix

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



Appendix S2. Descriptive Statistics of Full 12,740 Sample

Variable	Full Sample Missingness	Full Sample (*N*=12,740)	Study Sample (*N*=6,098)
Male	0.25%	51.72%	50.57%
White	0.25%	17.06%	17.15%
Black	0.25%	28.43%	28.17%
Hispanic/Latino	0.25%	43.87%	42.25%
Asian	0.25%	7.58%	9.32%
Mixed/Other Race	0.25%	3.06%	3.11%
Free/Reduced Lunch	1.45%	65.07%	71.10%
Special Education	22.32%	13.36%	17.31%
Dual Language Learner	5.24%	41.03%	43.77%

Note: CA stands for 'chronically absent'. Time-varying characteristics (free/reduced lunch and special education) are based on students' PreK value. *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. There was a small amount of missing data for students from the Study Sample in special education (0.20%), K chronically absent (2.10%), 1st grade chronically absent (3.20%), and 2nd grade chronically absent (0.50%).

Appendix S3. Optimal Hyperparameters of XGBoost Models

Algorithm	Inputs				
	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 = 5

Appendix S4. Sample code for XGBoost with SMOTE

```python
# 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
```

```python

# 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)

# Predict class probabilities for test data
y_prob = best_model.predict_proba(X_test_encoded)[:, 1]

from sklearn.metrics import roc_auc_score, balanced_accuracy_score, classification_report

# Calculate AUC
auc = roc_auc_score(y_test, y_prob)
print('AUC:', auc)

# Calculate BER
y_pred = best_model.predict(X_test_encoded)
ber = 1 - balanced_accuracy_score(y_test, y_pred)
print('BER:', ber)

# Calculate recall
report = classification_report(y_test, y_pred, target_names=['Negative', 'Positive'], digits = 3)
print('Recall:\n', report)

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

# Define the probability thresholds to test
thresholds_to_test = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]

# Initialize lists to store evaluation metrics for each threshold
accuracy_scores = []
auc_scores = []
ber_scores = []
recall_positive_scores = []
recall_negative_scores = []

# Loop through each threshold and calculate metrics
for threshold in thresholds_to_test:
    # Apply the threshold to the predicted probabilities
    y_pred_custom_threshold = (y_prob > threshold).astype(int)

    # Calculate metrics
    accuracy_custom_threshold = (y_test == y_pred_custom_threshold).mean()
    auc_custom_threshold = roc_auc_score(y_test, y_prob)
    ber_custom_threshold = 1 - balanced_accuracy_score(y_test, y_pred_custom_threshold)
    report_custom_threshold = classification_report(y_test, y_pred_custom_threshold, target_names=['Negative', 'Positive'], output_dict=True)

    # Append metrics to lists
    accuracy_scores.append(accuracy_custom_threshold)
    auc_scores.append(auc_custom_threshold)
    ber_scores.append(ber_custom_threshold)
    recall_positive_scores.append(report_custom_threshold['Positive']['recall'])
    recall_negative_scores.append(report_custom_threshold['Negative']['recall'])

# Create a DataFrame to store the metrics for each threshold
results_df = pd.DataFrame({
    'Probability Threshold': thresholds_to_test,
    'Accuracy': accuracy_scores,
    'AUC': auc_scores,
    'BER': ber_scores,
    'Recall (Positive)': recall_positive_scores,
    'Recall (Negative)': recall_negative_scores
})

# Display the results DataFrame
print(results_df)

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

```