Friend or Foe? The Role of Machine Learning in Education Policy Research

by

Amanda Weissman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Educational Studies)
in the University of Michigan
2022

Doctoral Committee:

Associate Professor Christina Weiland, Chair
Professor Susan Dynarski, Harvard University
Associate Professor Ben Hanson
Dr. Kristin Porter, MDRC

Amanda L. Weissman

ketneram@umich.edu

ORCID id: 0000 – 0001 – 5381 – 4516

Dedication

Although I dedicate my dissertation to every person who has supported and influenced me in my career thus far, there are five people without whom this document would not exist. My mom and my dad always supported me emotionally and financially, even when they did not understand what I was doing or necessarily agree with my choices. My undergraduate advisor, Alan Russell, was instrumental in believing in me and my abilities when I did not have faith in myself, and he was the largest source of encouragement when I left my high school teaching career to pursue graduate education. My doctoral advisor, Chris Weiland, has been the best advisor because she held me to high standards while always helping me reach them and supported my life outside of academia. Finally, my husband, Noah Weissman, has been with me for every step of my graduate school experience. He is the one person who has seen me at my worst when I did not think that I was good enough for or could handle my doctoral journey, and he never stopped being my biggest cheerleader. I am so excited to begin our next life phase together as a family of three.

Acknowledgements

The research reported here was conducted as a part of a study funded by Grant R305N160018–17 from the Institute of Education Sciences to MDRC with subcontracts to the University of Michigan, the Boston Public Schools, and the Harvard Graduate School of Education. My work supported by the Institute of Education Sciences Predoctoral Fellowship at the University of Michigan by Grant R305B150012 as well as the University of Michigan Rackham Predoctoral Fellowship. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Thanks to Boston Public Schools, the BPS Department of Early Childhood coaches and staff, the BPS Department of Research, the MDRC research team (particularly Meghan McCormick and Mira Pralica), the Harvard Graduate School of Education research team, the University of Michigan research team (particularly Lillie Moffett and Paola Guerrero Rosada), the Michigan Department of Education Center for Education Performance and Information, the Michigan Education Research Institute, the University of Michigan Education Policy Institute (particularly Kyle Kwaiser, Nicole Wagner Lam, and Jasmina Camo-Biogradlija), and the University of Michigan Consulting for Statistics, Computing, and Analytics Research team. I also want to thank Emanuele Bardelli, Stacey Brockman, Gina Cervetti, Julie Bruch, Katie Lebioda, Paula Clasing, Walter Herring, the participants of the University of Michigan Causal Inference in Education Research Seminar series (particularly Kevin Stange), and my dissertation

committee (Christina Weiland, Susan Dynarski, Ben Hanson, and Kristin Porter) for their time and thoughtful feedback on this work.

Table of Contents

List of Tables

List of Appendices

Abstract

Machine learning, while common in other disciplines, has slowly started to become used more education research. My project seeks to answer the call from researchers for guidance about the nature of machine learning and illustration in how to use machine learning. First, I address the nature of machine learning and provide examples of its capabilities in education policy research. Second, I examine what value machine learning adds over traditional regression in predicting vulnerable student populations for early intervention. Third, I probe how machine learning can aid in identifying students at risk for dropping out of high school. My project will be one of the few projects in the field of education to espouse the potential benefits of machine learning and interrogate its value added to traditional quantitative research methods.

Introduction

Machine learning, while common in other disciplines, is rarely used in education research. Often conflated with buzzwords like “data science,” “artificial intelligence,” and “big data,” researchers without a strong background in computer science can be left out of the conversation. Machine learning is an automated process that discovers patterns of variation within datasets and is a powerful tool that can be used to answer new research questions in education that previous methods could not address. Like other quantitative methods, it has the potential to both address and perpetuate inequity in education and therefore needs to be understood carefully (Athey, 2019; Jacobucci & Grimm, 2020; Mullainathan & Spiess, 2017). My project seeks to answer the call from researchers for more guidance and illustration in *how* to use machine learning by explicitly addressing the nature of machine learning and providing examples of its capabilities in education policy research (Gibson & Ifenthaler, 2017). I will do this by examining what value machine learning adds in predicting different types of vulnerable student populations for early intervention. My project will be one of the few projects in the field of education to espouse the potential benefits of machine learning and interrogate its value added to traditional quantitative research methods. I will also be one of the few authors to discuss its policy implications for use with the type of datasets that are both readily available to administrators and often smaller than datasets historically used with machine learning: administrative data.

My project consists of three related papers aimed at illustrating when machine learning can be a helpful, equity- and justice-oriented tool in the education researcher’s methodological toolbox. The first paper will be conceptual in nature, seeking to explain what machine learning is, how it can be used in education policy research, and ways it can complement causal inference methods. It will serve as a reference for education researchers unfamiliar with machine learning by explaining the basics of machine learning, highlighting relevant examples, and providing resources for further learning. Rather than positing that machine learning can be a silver bullet to solve all methodological issues, as it is sometimes positioned, this paper will examine when machine learning can be beneficial versus when it may be superfluous or even harmful in order to help researchers discern its value for education policy research.

After laying this foundation, my second and third papers will be empirical studies highlighting the value of predictive machine learning and probing the circumstances under which it is beneficial. One empirical paper will focus on how predictive analytics can be helpful in identifying two different vulnerable student populations in the early elementary years: students who were chronically absent in kindergarten, first, and second grades and students who began kindergarten not identified as needing special education but were later identified as needing special services by first and second grade. It will then explore the value added of one researcher-collected measure when combined with the districted-collected measures. Identifying these students earlier could improve their learning trajectories as these are groups who might benefit from early supports and differentiation of instruction, and district administrators expressed interest in working more with these student groups (Coleman et al., 2019; Diamond et al., 2013; Guralnick, 1998; Robinson et al., 2018). This paper will use both administrative and rich demographic and assessment data from the Boston Public Schools, available to me through my involvement in a longstanding research-practice partnership.

My second empirical paper will examine how well machine learning can predict students at risk of dropping out of high school. Using administrative data from Michigan that includes every public-school K-12 student for five years, I will test to what extent different grades' worth of administrative data from fourth through ninth grade predict which students are most at risk of dropping out to identify for early intervention. This study will build upon a handful of existing studies that explore the value added of machine learning for identifying this important vulnerable student population (Lakkaraju et al., 2015; Orooji & Chen, 2019; Sansone, 2019).

Contribution

My goal for my overall project is to be pedagogical in nature by probing the questions of *how* machine learning can be helpful in education policy research, particularly in advancing an equity and justice agenda, and under which conditions its use can be advantageous compared to traditional quantitative methods. I seek to demystify the confusion surrounding machine learning and shed light on when it is beneficial to use in education research by providing a comprehensive overview plus two in-depth applicative examples.

Additionally, to be a helpful tool for education policy research, it is important to study how machine learning fares with the type of data typically available to educators and administrators: administrative data. This project would be one of the first in the field of education that intentionally explores how well machine learning algorithms function under realistic sample sizes available to school personnel. Furthermore, by focusing on vulnerable student populations that would benefit from effective early intervention, my project will promote equity for these students. By examining the performance of machine learning under real world circumstances, the project will demonstrate the policy relevance of a new tool of great potential.

My dissertation will also be one of the few in the field to explicitly compare how supervised machine learning techniques perform versus traditional regression under different types of input data. By building up my models incrementally, I will deepen the conversation around the utility of various predictive analytic approaches by investigating for which data combinations machine learning is more helpful than traditional methods.

Purpose and Research Questions

My dissertation will consist of three related yet standalone papers that I will submit for publication in peer-reviewed journals. My research questions for each paper are:

Paper 1: “Machine learning for education policy research: What is it and why should I care?”

1. How can machine learning be used in education research?
2. How does machine learning work?
3. Can machine learning be used for causal inference?

Paper 2: “What can machine learning offer when predicting special education and chronic absenteeism for early elementary students?”

1. How does machine learning compare to traditional regression methods when identifying students who began kindergarten not identified as requiring special education services but are later identified as needing services by the end of first and second grade? How does this performance vary based on the type and timing of data used?
2. How does machine learning compare to traditional regression methods when identifying students who are chronically absent during kindergarten, first, and second grade? How does this performance vary based on the type and timing of data used?

3. If the district could add one assessment to its standard operations, which researcher-collected assessment would enhance prediction best?

Paper 3: "The role of machine learning in early warning systems for predicting high school dropout in Michigan"

1. How does machine learning compare to traditional regression methods when identifying students who do not graduate from high school? How does this performance vary based using data from different grades as predictors?
2. Do the models work equally well for student racial, socioeconomic, gender, and special education subgroups?

Paper 1: Machine Learning for Education Policy Research:
What is it and Why Should I Care?

Machine learning, while common in other disciplines, is slowly becoming more common in education research. Researchers without a strong background in computer science can be left out of the conversation as machine learning is often conflated with buzzwords like “data science,” “artificial intelligence,” and “big data.” I seek to answer the call from researchers for more guidance and illustration in *how* to use machine learning by explicitly addressing the nature of machine learning and providing examples of its capabilities in education policy research (Gibson & Ifenthaler, 2017). Machine learning is an automated process that discovers patterns of variation within datasets (Hastie et al., 2009). It is a powerful tool that can be used to increase the efficiency of existing methods (such as precisely predicting students at risk of dropping out of high school) and to answer new research questions in education that previous methods could not address (such as identifying complex patterns among students at risk of dropping out) (Athey, 2019; Mullainathan & Spiess, 2017). Like other quantitative methods, it has the potential to both address and perpetuate inequity in education depending on the data used and therefore needs to be understood carefully (Jacobucci & Grimm, 2020).

Other disciplines have adopted machine learning more readily than the broader education research community. For example, machine learning has been used to harness the wealth of electronic medical records data to predict in-hospital mortality, unplanned readmission, longer than expected hospital stays, and final discharge diagnosis (Crown, 2019). Political scientists have used machine learning to study how people discuss politics in online discussion boards and express their political beliefs with social media (Grimmer, 2015). In the field of criminal justice, researchers have used machine learning to predict the risk of arrested people released on bail being charged with committing a crime before the trial in order to study racial discrepancies in the judges' initial decision to allow for release (Kleinberg et al., 2017).

Although education scholars have generally not used machine learning methods to the extent of other fields, those who have are presenting encouraging results. For example, machine learning has been shown to be helpful in predicting which students are most at-risk of failing pivotal exams in order to target for early intervention (Porter, 2019). Researchers have also shown its value by creating an automated process to provide timely formative feedback for teachers about their questioning techniques using audio from classrooms (Donnelly et al., 2017). Machine learning is also helpful for analyzing large amounts of text that would be cumbersome to study manually, such as when researchers reviewed thousands of documents to better understand administrators' response to No Child Left Behind (Sun et al., 2019). These initial applications indicate that machine learning can be a beneficial tool when studying various educational contexts and offer a glimpse into what machine learning can offer the field of education.

My goal is for this manuscript to function as an accessible entry point for education researchers to learn how machine learning can be helpful in both opening up new research questions and improving the efficiency of existing quantitative methods. I particularly focus on readers who use quantitative methods such as regression and econometrics yet are unfamiliar with machine learning. Rather than positioning machine learning as panacea for the pitfalls of traditional research methods, I will discuss the nuances of *when* and *how* machine learning can add value to a study.

As with any decision in research design, there is a tradeoff for using machine learning compared to traditional methods: namely foregoing a level of transparency due to the complexity of the algorithms in order to gain more predictive precision. It may not be beneficial to the broader research study to use a new method that the field at large is unfamiliar with if it does not function better than more simple and traditional methods. The value that machine learning adds needs to be clear in order to justify its use over traditional methods that are broadly accepted in the field. In addition to its methodological contributions, researchers should also be aware of the potential for machine learning algorithms to replicate existing biases in datasets. Therefore, it is important for researchers to understand the conditions under which machine learning is worth pursuing (Athey, 2017; Singer, 2019).

I bring these conditions to light by examining three points:

1: How can machine learning be used in education research?

2: How does machine learning work?

3: Can machine learning be used for causal inference?

Point 1: How Can Machine Learning Be Used in Education Research?

Over the last fifteen years, researchers have begun using machine learning to study different aspects of education, such as predicting students at risk of dropping out of high school (Lakkaraju et al., 2015) and grouping reform efforts in response to policy changes (Sun et al., 2019). From this work, I can see that the two highly useful applications of machine learning for education research are predicting and grouping. Prediction analysis, called supervised machine learning, occurs when researchers use a large set of covariates to predict an outcome, either binary (such as dropping out of high school) or continuous (such as a test score). Grouping analysis, called unsupervised machine learning, occurs when researchers take a large number of either variables or observations and group them into sets that are internally similar yet different from other groups. Both of these can be used with what has traditionally been considered in education research to be quantitative (i.e., numbers) and qualitative (i.e., text) data. By using machine learning on its own as well as in conjunction with other methods, researchers have begun to illustrate the applications of machine learning for education, such as predicting students at risk of dropping out of high school and grouping reform efforts in response to policy changes.

Predicting

Prediction to Identify Students for Intervention

One of the most commonly used applications of machine learning so far in education research has been using supervised machine learning to predict at-risk student populations to target for intervention. Porter (2019) used machine learning to help identify students at risk of failing the end-of-third-grade reading exam (which would then lead to retention) in order to place students into a reading intervention program. By using diagnostic testing data from first, second, and third graders, Porter created three models to determine the value added from machine learning. First, she used the diagnostic reading assessment information from the beginning of third grade to determine which students would likely pass the end-of-year exam using the criteria from the assessment creator. Second, she used traditional regression techniques (such as logistic models) to establish a quantified relationship between the score on the diagnostic exam from the beginning of third grade and whether students passed the end-of-year exam using data from the previous cohort; she then applied this relationship to the current cohort to predict which students were at risk of failing. Third, she used all available data from the diagnostic exams from first and second grades plus the scores from the beginning of third grade with supervised machine learning to predict failing the end of year exam by training her model on the previous cohort.

Across these three models, Porter’s machine learning algorithm correctly identified a higher percentage of students who were likely to fail the end-of-year test (78%) compared to using the publisher’s guidelines from the screening exam (54%) or the traditional regression model (57%). However, the machine learning algorithm also produced a higher false positive rate as well (14% versus 5%) (Porter, 2019). Although a limitation of this study is that the author didn’t strictly compare the performance of machine learning and traditional regression under every combination of data, it is still a positive indication about the predictive capacity of machine learning. In a context where high-stakes third grade reading laws are becoming more common, machine learning is a promising tool because more accurate predictive analytics can be useful for education policy research (Council of Chief State School Officers, 2019).

Similarly, another study from Greece used predictive machine learning[1] with demographic and academic data from 354 students to identify students at risk of dropping out of a distance-learning higher education program. Researchers created two machine learning models: one with baseline demographic factors only and a second with demographic plus academic performance data. They found that the model with demographic information correctly identified 63% of students at risk of dropping out while the second model correctly identified 83%. When they compared this to logistic regression results, they found similar accuracy rates (60% and 83%, respectively), indicating that machine learning may not provide a distinct advantage for this sample (Kotsiantis et al., 2003).

An increasing number of studies have been conducted in the last five years using machine learning to predict which students are at risk of dropping out of high school prior to graduation.

[1] Specifically, researchers used a naïve Bayes, neural network, nearest neighbor, support vector machine, and decision tree algorithms. See Table 2 for more details about the individual algorithms.

Many school systems across the country established an early warning system to target students for intervention based on the established literature (using traditional regression and structural equation modeling) that found factors associated with high school drop-out, such as academic achievement and frequent absences (Battin-Pearson et al., 2000; O'Cummings & Therriault, 2015; Parr & Bonitz, 2015; Rumberger & Larson, 1998). However, new studies have explored how predictive machine learning compared to traditional regression techniques (most commonly, logistic regression) when predicting high school dropout. While they varied in the algorithms, data sources (i.e., administrative versus large scale survey), and geographic location (i.e., United States, Denmark, South Korea, and Mexico), the studies all show high (>90%) accuracy rates with machine learning models (Ara et al., 2015; Chung & Lee, 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016; Orooji & Chen, 2019; Pagani et al., 2008; Sansone, 2019). Furthermore, the study that compared the results of models fit using machine learning techniques versus traditional regression methods (i.e., logistic regression) showed that the most accurate predictions came from the machine learning models[2] when using the same input variables because of its inherent ability to be more flexible and fit the data better than the logistic regression models (Lakkaraju et al., 2015).

These studies show that machine learning is a promising technique for accurately identifying which students to target for intervention, which is particularly helpful when schools want to maximize the efficiency of their limited resources by targeting them towards students who need help the most (Engler, 2020). For school personnel, the question of how much better machine learning techniques are compared to more traditional regression methods when identifying students for intervention largely remains unanswered in the literature. Given how

[2] The most predictive machine learning model used the random forest algorithm.

machine learning algorithms are considered more of a black box than their traditional regression counterparts, they would need to be substantially more accurate to convince schools to use them, an avenue for future research (Bruch et al., 2020).

Prediction to Improve Classroom Instruction

Another area of emerging machine learning research is improving classroom instruction by automating the process of understanding aspects of classroom instruction that would otherwise be cumbersome to study manually. For example, researchers used machine learning algorithms to analyze transcripts from 1,000 fourth and fifth grade language arts classes to determine the amount of time teachers spent in six different areas of instruction. They then combined these elements into constructs, validated them psychometrically, and examined associations between the time spent in constructs and students' value-added assessment scores (Jing & Cohen, 2021). Similarly, researchers from another study recorded the audio from 37 different middle school classes on literature, language arts, and civics classes across 11 teachers. After first coding a portion of the audio transcripts manually to use as a starting point to train the machine learning model, they then used machine learning[3] to parse apart the remaining audio in order to identify when teachers were asking questions and categorize the types of questions they asked (Donnelly et al., 2017). Both studies are first steps towards the ultimate goal of providing personalized formative feedback to teachers about classroom instruction by using an automated process that would be both faster and more efficient than observing classrooms in person.

Machine learning can also be used to detect students' affectation. By using cameras, a machine learning algorithm coded 137 students' facial expressions and body movements in real

[3] Specifically, researchers used the naïve Bayes, logistic, random forest, decision trees, and k-nearest neighbor algorithms.

time while playing a computer physics game in order to determine the students' level of engagement by categorizing students' affect as boredom, confusion, delight, engagement, frustration, or off-task. Part of affect detection studies, this study illustrates how machine learning can be used to provide feedback quickly on student behavior in order to alter the mode of instruction, such as adapting the code of the computer game to change in response to student engagement. The authors found the algorithm to perform similarly for different genders and ethnicities but would have benefitted from a larger sample of students of color to dig deeper into variation across ethnicities (Bosch et al., 2016). By studying both teacher and student behavior, machine learning can provide insight into what is happening in classrooms and has promise to examine discrepancies in instruction among student subgroups in order improve instruction (Petrilli, 2018).

Prediction to Help Students Make Decisions

Researchers have theorized and shown how machine learning can help students make decisions by providing them with data and analyses about which college to attend and courses to take in order to mitigate mismatching (Arndt & Guercio, 2016). One study in China (Ye, 2018) illustrated this by using administrative data from a highly centralized, national higher education application process. In this type of admissions system, an appropriately ranked list was important to increase odds of a successful match because the author noted that previous work had shown that students commonly both under- and overestimated their likelihood of placing into universities, resulting in an ill-ranked list. The researcher used machine learning[4] to determine the predicted probability of acceptance for 5,647 students; they then provided students with their predicted probability to being accepted into various universities in the country. This intervention

[4] Specifically, researchers used a random forest algorithm.

helped students craft their ranked order list of universities that would better match with the universities' ranked list of students, increasing students' likelihood of acceptance. Not only did the predicted probabilities help students make decisions, but the process also mimicked part of a teacher or advisor's role in this crucial process. Because the machine learning algorithm was able to provide help to many more students than teachers could in their limited time, this machine learning-based intervention democratized the valuable knowledge of student acceptance probabilities for a larger number of students who may not otherwise have access to personalized guidance from teachers. Additionally, machine learning relieved teachers of this aspect of advising students and allowed them to focus their efforts elsewhere, increasing the efficiency of teachers as well as assisting students in the application process (Ye, 2018).

Prediction with Modeling Complicated Relationships

The flexibility of machine learning algorithms can more easily allow for complicated modeling techniques – particularly multiple interaction effects and nonlinear relationships within the same model – than traditional regression techniques that can be hampered by collinearity and power concerns. For example, researchers in the United Kingdom were interested in studying the impact of high school curriculum on labor market outcomes (Johnes, 2005). Because students were able to specialize in the courses they took, researchers were particularly interested in modeling the relationship between the various possible course combinations and adult earnings. To explore the added benefit of machine learning, the authors compared results of predictive machine learning models[5] to those with traditional regression methods for 2,970 students and found the machine learning approach predicted earnings more accurately (Johnes, 2005). Although this one study does not represent a referendum on the value added of machine learning

[5] Specifically, researchers used a neural network algorithm.

over traditional regression when modeling complicated relationships, it is a promising indicator of machine learning's utility.

Similarly, machine learning can be helpful for examining heterogeneity when modeling multiple interaction and nonlinear terms. When researchers wanted to better understand the relationship between various student- and school-level inputs and academic achievement for 92,035 students across nine countries, they used machine learning[6] to determine not only which variables were most predictive of achievement but also which types of interactions between those variables were most predictive. Then, they were able to examine geographic differences in these inputs and found that when predicting achievement across countries, the interaction between school size and economic makeup of the student body varied (Masci et al., 2018). By leveraging the functional flexibility inherent in machine learning algorithms, researchers can explore nuanced relationships that linear regression may struggle to appropriately model.

Potential Prediction Pitfall

Although prediction algorithms have been demonstrated to be beneficial in education research, researchers must be careful of potential bias, including but not limited to race, ethnicity, gender, class, and ability (Baker & Hawn, 2021; Broussard, 2018; Criado Perez, 2019; D'Ignazio & Klein, 2020; O'Neil, 2016; Porter et al., 2020). As an example, uses of machine learning algorithms outside of education have had important racial implications. In both healthcare and policing, predictive algorithms have increased racial disparities due to the preexisting racial inequities in the data used to create the machine learning models. When hospitals used machine learning to predict healthcare costs as a proxy for illness, the algorithms under-identified illness among Black patients. This was because the data used in generating the

[6] Specifically, researchers used a random forest algorithm.

machine learning models reflected the fact that Black patients were less likely to seek medical care because they were less likely to have health insurance compared to White patients (Obermeyer et al., 2019).

Similarly, research has shown that police forces that use predictive modeling to predict criminal activity in order to efficiently allocate resources over-estimate the rate of unlawfulness among Black citizens. The machine learning algorithms that created these predictions were constructed using data that reflected the racial bias that officers traditionally exhibited in making arrests, so the predictions replicated these racial biases (Lum & Isaac, 2016; Richardson et al., 2019). Racial problems have also been seen in predictive policing when the police use facial recognition software to identify potential criminals. Depending on the data used to construct the facial recognition software, the algorithms can be less accurate when identifying Black faces compared to White faces, leading to a disproportionate number of misidentification of Black people (Garvie et al., 2016; Hill, 2020).

As predicting machine learning becomes more commonly used to predict education outcomes, researchers must learn from these examples in order not to perpetuate structural racism. We can do this by questioning potential biases in the data used to create the models in order to mitigate what is referred to as “algorithmic bias” (Baker & Hawn, 2021; Gebru, 2021). This is not a problem unique to machine learning algorithms, however. Traditional quantitative methods are also subject to misrepresentative findings when using biased data, such as data that contains measurement error or reflects broader systemic inequities. Rather than disregarding these methods completely, I argue that researchers should think critically about the data used because the machine learning algorithms will reflect and often amplify any existing biases in the data (Engler, 2021; Gillborn et al., 2018; Jacobucci & Grimm, 2020; Lee et al., 2021; Porter et al., 2020). More broadly, I call for researchers to be critical about the results of analyses conducted with any method, including machine learning, to prevent blindly accepting output.

In addition to being aware of existing biases in the data, I also call for researchers to explicitly interrogate their results in two ways. The first is to look for patterns in the identified students. For example, if a machine learning algorithm predicts students at-risk of dropping out of high school with a high degree of precision yet all the students identified are from the same neighborhood, then researchers have an ethical responsibility to question the validity of that model. Is there truly some geographic phenomenon happening, or is the model picking up on underlying characteristics that lead to an overrepresentation of students from that neighborhood while leaving vulnerable students from other neighborhoods to go unnoticed?

The second recommendation is to explicitly examine model performance for important student subgroups. For example, after creating a model to predict high school dropout, researchers should calculate the model performance statistics for students subgroups who drop out at higher rates than their counterparts on average, such as students of color and students from families with low incomes (Hussar et al., 2020). If the model doesn't perform as well for the subgroups as it does for the full sample, then researchers should caution the use of that original model and refine it to work well for subgroups. It is possible that certain model performance statistics – such as the true positive, false positive, true negative, and false negative rates – vary by subgroup. By assuming that a model is measured equally well for every subgroup, it is possible to miss important variation that can lead to biased predictions (Porter et al., 2020). For predictive models that rely on machine learning to be useful in education, they need to work for all types of students (Gebru, 2021; Kantayya, 2020).

Grouping

Grouping to Understand Student Behavior

In addition to prediction, machine learning can also be used to identify groups of students based on similar behaviors that would not necessarily be obvious from the data otherwise. For example, researchers wanted to better understand how preservice teachers responded differentially to writing tasks. They used data from an online learning platform that tracked how long students spent on different types of writing assignments as part of a larger problem-solving study. Harnessing patterns in the data, researchers used a grouping machine learning algorithm[7] to group students into five categories based on the amount of time they spent on each type of writing task, such as students who spent time on deciphering the writing task description versus students who spent time on researching relevant resources. Similarly, researchers applied a similar approach[8] when using data from undergraduate engineering students that tracked their activity on online homework platform to better understand the differences in problem solving techniques between high- and low- performing students. In both studies, researchers used the machine learning algorithm to cluster students together according to how the students behaved similarly (within the groups) and differently (between the groups) and then were able to understand the key identifier of each cluster by applying theories of problem-solving techniques. Furthermore, they were able to do so using multiple variables that would have been cumbersome and difficult to do manually (Antonenko et al., 2012).

Grouping algorithms can also be used to provide insight into student behavior in order for educators and schools to respond more efficiently. When researchers in one study wanted to identify students who had generally performed poorly at the halfway point in a first-year

[7] Specifically, researchers used the Ward’s clustering algorithm.

[8] Specifically, researchers used the k-means clustering algorithm.

undergraduate engineering course, they used machine learning[9] to group students into three different levels of performance (high, medium, and low) using all of the assessment data from the first half of the term. Professors then used these groupings to identify which students to seek out for extra help for the second half of the term (Shovan & Haque, 2012). Another study used grouping to better understand if students' decision to attend a particular vocational high school varied by how far away from the school they lived. They used machine learning[10] with data on their reported reason for attending and measures on how far away they lived from the school to group students into five categories, such students who chose to attend based on their close proximity versus students who attended because of certain academic offerings. The school then used this information to better focus its recruitment efforts for future students (Abadi et al., 2018). These studies illuminate the potential for using machine learning to uncover groupings of students that may not be obvious in order to better understand student behaviors and to respond differentially.

Grouping to Identify Topics in Text

Machine learning can also be helpful when sorting text[11] into previously uncategorized patterns (Sun et al., 2019). When schools in Washington State were tasked with reform initiatives to address student absenteeism and low student achievement under No Child Left Behind, school administrators had leeway to determine how they would address the problems at their institution. Beyond the administrative data used to evaluate the reforms at a macro level, researchers wanted to better understand the decisions at individual schools and how those

[9] Specifically, researchers used the k-means clustering algorithm.

[10] Specifically, researchers used the k-means clustering algorithm.

[11] For a more technical guide on how to scrape educational documents for useful information, see (Anglin, 2019).

reforms were associated with change. Using machine learning,[12] they sorted through the text of planning documents and reports in order to identify twenty different topics of reforms that administrators implemented, such as creating in-house leadership teams to review data for school improvement purposes. After identifying these topics, they were able to determine descriptively the percentage of reform efforts spent on each topic and link the reform themes to the key outcomes, ultimately helping to explain which reform actions helped reduce absenteeism and improve achievement (Sun et al., 2019). By using text as data combined with automated machine learning, researchers were able to more efficiently study documents to learn about patterns in reform efforts.

In one of the few examples of published research that uses machine learning in early childhood education, researchers used machine learning to identify aspects of early childhood education centers that were important to parents. They did this by using a grouping algorithm[13] that read through a large amount of text from over 8,000 online Google and Yelp reviews, identified nine topics that emerged in the text data, such as warmth and communication, and determined the topics discussed in each review. These results provided insight into the issues important to parents, the prevalence of each issue, and how parents' ratings of the overall center corresponded to the topics that they mentioned in their review. Machine learning enabled researchers to use data from an unconventional source to learn more about what parents think about their early childhood education providers in a way that would have been extremely difficult to do manually given how much data was studied (Early & Li, 2020).

[12] Specifically, researchers used a latent dirichlet allocation algorithm.
[13] Researchers used topic modeling.

Overall, machine learning offers a helpful tool when conducting education policy research by automating descriptive statistics via predicting and grouping large amounts of data. By potentially providing more accurate predictions than other methods, machine learning can help researchers identify students who may need interventions the most (Lakkaraju et al., 2015; Porter, 2019). By uncovering groups of subjects, machine learning offers a way to examine profiles of students and themes in text that would otherwise be difficult to detect (Antonenko et al., 2012; Sun et al., 2019). However, machine learning may replicate biases that exist in the datasets, so researchers should use caution when determining which data to use with machine learning algorithms (Gebru, 2021).

Point 2: How Does Machine Learning Work?

Part of the emergence of "big data," computer scientists and statisticians developed machine learning to analyze large datasets more efficiently with potentially thousands of variables and/or observations. Typically considered a subset of artificial intelligence, machine learning is sometimes conflated with the term "data science." Data science generally refers to the process of conducting research with large data sets using multiple methodologies, one of which is machine learning. At its most basic level, machine learning is a way to automate the process of uncovering patterns of variation in large datasets. For education research, machine learning is most helpful for automating descriptive analyses through prediction and grouping. Its value comes from its ability to *learn* from existing patterns in the data while navigating large datasets. There are multiple algorithms for this learning process, and the user specifies which one(s) to use.[14] Whether the goal is predicting or grouping, the most salient research design choices for the researcher are the algorithm and dataset selection.

Machine learning may be best understood by its comparison to traditional regression methods, particularly those drawn from microeconomics, already widely used in education policy research (Table 1). In traditional regression analysis, the researcher chooses the type of regression model to fit (such as a linear or logistic model) and specifies which variables and their functional form to be included in the analysis that produces model coefficients. This choice is driven by theory and evidence from previous literature about the relationship between variables. For example, if a researcher wants to study which students are most at risk of dropping out of high school, they may use a logistic regression to predict a binary indicator of dropping out. They then will choose independent variables according to what the literature has shown to be important predictors of dropping out, such as academic achievement and absenteeism, while also making sure not to choose covariates that would be collinear (Battin-Pearson et al., 2000; Parr & Bonitz, 2015; Rumberger & Larson, 1998). They can then use the results in two ways. First, the fact that regression results display the magnitude and statistical significance of the parameter estimates covariates allows the research to determine which variables are related to dropping out of high school. Second, the researcher can use the predictive capacity of the model to determine how accurately the model identified students who dropped out.

Compared to this deductive process, machine learning presents an *inductive* approach (Singer, 2019). In machine learning, the researcher selects an algorithm (such as a predicting or grouping algorithm) and a dataset. Depending on the algorithm used, the algorithm can

[14] When doing predictive modeling, it is common for researchers to use multiple predictive machine learning algorithms and compare their results. See (Lakkaraju et al., 2015) as an example.

determine the optimal functional form between the predictor and outcome variables and capitalize on the specific input variables most predictive of the outcome. Some algorithms also have an option where, if selected, the algorithm will choose the subset of input variables that most increases model performance to use instead of the full set of predictor variables. For prediction algorithms, algorithms are designed to maximize the predictive capabilities, conceptually similar to trying to reach the highest R^2 value as possible when traditional regression is used for predictive analytics. For grouping, algorithms are designed to create groups that are as internally homogenous as possible while being as different from others; this is similar to the concepts of within- and between-group variance used in traditional ANOVA calculations (Hastie et al., 2009; James et al., 2013).

Due to the nature of the machine learning algorithms, collinearity is not a concern like it is in traditional regression[15] because the models do not present the statistical significance of variables that collinearity can suppress. Instead, machine learning algorithms automatically select the variables that maximize their function. For example, when predicting which students are at risk of dropping out of high school using machine learning, the researcher will specify which prediction algorithm to use and the set of variables that will act as the data frame. Then the algorithm will sort through the full dataset according to the rules stipulated in the particular algorithm to determine which subset of variables to use. Prediction algorithm results do not generally display the magnitude and statistical significance of the input variables like traditional regression does. Instead, the focus in interpreting the algorithm's output how accurate the algorithm was in predicting high school dropout (i.e., true positive, false positive, true negative,

[15] The exception to this is when traditional regression methods, such as logistic regression, are used for predictive analytics.

and false negative rates) along with which variables where most important in determining those predictions (Athey, 2019; Mullainathan & Spiess, 2017).

Types of Machine Learning

While there are several machine learning algorithms already in existence and more that continue to be created, I can broadly condense those that are most applicable to education research into two conceptual groups: predicting (supervised machine learning) and grouping (unsupervised machine learning) (Jordan & Mitchell, 2015). Just as there are different types of traditional regression models to use for different types of outcome variables and the error structure of the model, there are multiple algorithms for both predicting and grouping as discussed more below. For both predicting and grouping machine learning algorithms, there are parametric (i.e., able to be modeled with an equation) and nonparametric (i.e., not able to be modeled by a single equation) algorithm options (Athey, 2019; Mullainathan & Spiess, 2017).

Predicting (Supervised Machine Learning)

Supervised machine learning is commonly known as "predictive analytics" because it is used to predict a known outcome from a set of variables. This is the type of machine learning most conceptually similar to traditional regression in that the goal is to use an algorithm that takes a large set of input variables (**X**) to predict an outcome (Y) for every subject as accurately as possible. Algorithms can be used to predict both continuous and categorical outcomes, and one of the most powerful uses of machine learning for education policy research may be its ability to predict binary outcomes, such as identifying which students are most likely to fail an end of year reading exam (Porter, 2019) or drop out of high school (Lakkaraju et al., 2015; Pagani et al., 2008) in order to target for an intervention. The machine learning process is often referred to as "classification" when predicting a binary outcome coded as a 0 or 1 and called "regression" when the outcome variable is a binary variables treated as continuous to predict likelihood or a traditionally continuous variable (Athey, 2019; Mullainathan & Spiess, 2017).

The overall goal of prediction is to use information from a dataset with known outcomes to estimate, i.e., predict, outcomes for a second dataset with unknown outcomes. Supervised machine learning algorithms are designed to be flexible, meaning that they are designed to maximize their goodness of fit and explanatory power by fitting the data well. This inherent design leads to one of the major concerns of machine learning algorithms: achieving a balance between producing an optimal prediction and overfitting the model. Overfitting the model occurs when an algorithm is so well tailored to the nuances of one dataset that it would not perform well with a similar dataset. This dilemma is referred to as the "variance-bias tradeoff," where optimizing the predictive power leads to minimized variance yet overfitting leads to biased estimates for a similar model (Athey, 2019; Mullainathan & Spiess, 2017). For example, when predicting which students are at risk of dropping out of high school, machine learning algorithms may function well because they can take into account massive amounts of variables available from administrative and assessment data. However, a researcher does not want to adjust the model to the point where it works so exceptionally well for one cohort of students that it does not perform well for the next cohort. Instead, the goal is to create a model that works well for both cohorts.

Researchers attempt to navigate the balance by splitting[16] their dataset into two groups: "training data" and "testing data." Models are first fit using the training data to maximize the

[16] There are multiple ways to split the training and testing data, such as randomly selecting the original data following an 80/20 split, using a previous cohort as the training data and the current data cohort as the testing data, or a process known as *k*-fold cross validation. This occurs when the entire dataset is split into *k* groups, one group is selected as the test data while the other *k-1* are used for training, and then the evaluation indices are averaged across the models creating by cycling through each group as the test data (Hastie et al., 2009; James et al., 2013). Because there is no universally agreed upon best approach for the process of determining the training and testing data, it is up

flexibility, statistically, by minimizing a loss function. Just like the statistical theory underlying traditional regression that seeks to minimize the sum of squared residuals, every supervised machine learning algorithm has a different loss function motivating the algorithm that whose fit statistic measures how accurate the predicted outcome is to the true outcome (Hastie et al., 2009; James et al., 2013). Then, the fit model is applied to the testing data to obtain the performance measures, acting as a check on overfitting to the training data. For example, when predicting high school dropout, researchers may use the first cohort as the training data and the second cohort as the testing data if there is information from two cohorts. If there is only one cohort available, then researchers may randomly select a certain percentage of the cohort as the training data and use the remaining students for the testing data (Athey, 2019; Mullainathan & Spiess, 2017).

Another way that researchers can approach this variance-bias balance is by taking advantage of extra researcher-dictated specifications that are baked into some of the algorithms. Called "hyper parameters" or "tuning parameters," these are ways to limit the scope of the machine learning algorithm by imposing certain constraints in the hopes of making it less likely to overfit to the training data. For example, in a certain predictive algorithm called a decision tree (see more below) that is constructed using recursive partitioning of the dataset, a hyper parameter researchers may opt to use is setting a maximum number of partitions the algorithm can use (Hastie et al., 2009; James et al., 2013).

An important decision for researchers to make for both traditional regression and machine learning is which type of model to choose (Table 2). In traditional regression,

to the researchers' discretion to use the method most appropriate for their dataset. It is becoming increasingly common in the handful of education studies that have employed classification to use multiple algorithms and compare them across multiple fit statistics (Ara et al., 2015; Lakkaraju et al., 2015; Orooji & Chen, 2019; Sansone, 2019).

researchers may opt to use a linear model when estimating test scores and a logistic model when predicting high school dropout. In machine learning, researchers have several options for algorithms when predicting an outcome. For parametric options (i.e., algorithms that are based on an equation with traditional parameters), the most common algorithms for education research include naïve Bayes, linear and quadratic discriminant analysis, neural networks, ridge regression, and LASSO (Least Absolute Shrinkage and Selection Operator). Popular nonparametric options (i.e., algorithms that are built without using traditional equations) include nearest neighbor, decision trees (aggregated to random forests), and vector support machines (Hastie et al., 2009; James et al., 2013). (For an in-depth exploration and example of decision trees and random forests, see the Appendix.)

Model selection for supervised machine learning is often not as clear cut as it is with traditional regression. In traditional regression, there are some clear guides for which type of models to use based on the distribution of the outcome variable (such as using a logistic, probit, or linear probability model when predicting a binary outcome) and researchers often have latitude when determining the error structure (such as using multilevel modeling or clustered standard errors for nesting). While there are a few guiding principles for algorithm selection (such as that neural networks need large datasets to predict with high accuracy), there are generally fewer constraints on which type of situations call for which type of predictive algorithm (Hastie et al., 2009; James et al., 2013). Therefore, it has become increasingly common for education researchers to use a handful of different predictive algorithms and compare their performance. For example, when predicting high school dropout, researchers used four different machine learning algorithms,[17] compared their performance to logistic regression, and found the random forest algorithm to be the most consistently predictive across multiple performance metrics (Lakkaraju et al., 2015).

Similarly, there are multiple ways to evaluate the performance of a predictive machine learning model. In traditional regression, researchers can look at the R^2 to determine how well the model explains variation in the data and can justify model fit based on theory and measures such as the intraclass correlations for multilevel modeling (Snijders & Bosker, 2012). For predictive algorithms, researchers use a variety of performance indicators, including correct and misclassification rates for binary outcomes and Area Under the Curve for continuous likelihoods (Table 3) (Hastie et al., 2009; James et al., 2013). Just as it has become common to use multiple algorithms, it can be helpful to use multiple performance measures to further nuance results, depending on the situation. For example, when predicting the likelihood of an educational outcome to identify students for intervention, researchers can first measure model performance using Area Under the Curve. Then they can pick a cut point in the likelihood to dichotomize the data and examine the true and false positives and negatives to simulate how school personnel would use such a prediction algorithm as part of an early warning system (Bruch et al., 2020).

Grouping (Unsupervised Machine Learning)

Compared to supervised machine learning that takes all of the input variables (**X**) to predict an outcome (Y), unsupervised algorithms only focus on finding patterns among one group of variables. This is typically used to find groups among the covariates but can also be used for the outcome variables. Generally, unsupervised machine learning is helping for taking a

[17] Researchers used decision trees, random forests, AdaBoost, and support vector machines to predict high school dropout.

large number of variables and finding ways to group them that are not initially clear (Athey, 2019). For example, when examining patterns in high school dropout, researchers may want to look at heterogeneity in the types of students that dropout. Although researchers could make theoretically constructed groups defined by variables the literature has shown to be salient to dropout, such as income status and race/ethnicity, a grouping machine learning algorithm would be able to take a large number of variables into account to form groups that may be more complex than those defined by a single variable.

One way to do this is called "clustering," in which the algorithms organize the subjects into non-labeled groups where the subjects within the groups are similar to each other, yet the groups are different from other groups. Graphically speaking, the algorithms accomplish this by minimizing the distance between the data within each cluster while maximizing the distance between clusters. Clustering algorithms include *K*-means, hierarchical, density-based, and parametric model-based clustering. Clustering models can be judged by comparing measures of the inter- and intra-cluster distance, referred to as the Between Sum of Squares Error and Within Sum of Squares Error, respectively, as well as other distance metrics. These within- and between-group variance measurements are similar to the traditional ANOVA. This balance between wanting similarity within clusters and dissimilarity among clusters is a way to quantify the variance-bias tradeoff inherent in every machine learning algorithm (Hastie et al., 2009; James et al., 2013). There is also a modified version of this type of clustering called "fuzzy clustering" that relaxes the stipulation that each group must be separate, allowing the groups to potentially overlap (Kaufman & Rousseeuw, 1990). This type of clustering is often used in education research when conducting text analysis (algorithms include latent dirichlet allocation and topic modeling) to find topics in text documents and then to determine which topics are represented in each document, such as seen in the reform initiative example from Washington discussed above (Sun et al., 2019). Whether grouping subjects or text data, clustering can be helpful for identifying groups in a large dataset that would be difficult to do manually.

Another use of unsupervised machine learning is to reduce the dimensionality of a large data set by using a "principal component analysis" algorithm. This algorithm groups the variables into constructs such that each observation has a numerical value for each construct. For example, a principal component analysis may take a dataset with twenty education variables and reduce it to constructs that may conceptually represent classroom characteristics, child demographics, assessment data, and community information. The algorithm does this by creating linear combinations (i.e., weighted combinations) of the variables that maximize the amount of variance captured. It is considered a type of linear projection and therefore has measures of distance between the factors to help evaluate the performance of the algorithm (Hastie et al., 2009; James et al., 2013). Principal component analysis also plays a role in "principal component regression," which is when dimensionality reduction is used for traditional regression with datasets where the number of predictors is larger than the number of observations, leading to concerns about multicollinearity with traditional regression models (Vigneau et al., 1997).

A principal component analysis is very similar to the Exploratory Factor Analyses from the field of psychometrics in the sense that both approaches are used for dimension reduction. There are some slight differences between principal component analysis and Exploratory Factor Analysis, the first one being that Exploratory Factor Analysis relies more heavily on the theory underlying the constructs, places more emphasis on the interpretability of the constructs, and has more guidelines around measuring the fit of the constructs (Furr & Bacharach, 2014; Hu & Bentler, 1999; Worthington & Whittaker, 2006) whereas a principal component analysis is more purely data driven without much regard for the context of the data (Hastie et al., 2009). The second difference is that a principal component analysis automatically produces as many constructs as number of variables (Hastie et al., 2009). Conversely, an Exploratory Factor Analysis undergoes an iterative process where it first fits all of the items onto one construct, checks for model fit, and repeats with an additional construct until it achieves a desired level of fit (Furr & Bacharach, 2014). Given that the Exploratory Factor Analysis was originally designed to measure psychological constructs while the principal component analysis was created purely for dimension reduction, these nuanced differences can help explain the purpose of using each method.

Point 3: Can Machine Learning Be Used for Causal Inference?

For education research, machine learning is generally most helpful for automating descriptive statistics via predicting and grouping. Descriptive statistics provide valuable insight into the nature and variation of the educational landscape, and machine learning may help provide descriptive statistics more efficiently than traditional methods, particularly with large datasets (Loeb et al., 2017). As useful as machine learning is for producing descriptive statistics, it is generally not very helpful for conducting causal inference[18] analysis for education research by determining causality. For most algorithms, there is nothing inherent about the nature of the algorithms that automatically produces causal estimates. Machine learning algorithms are often considered to be a more black box approach to their more transparent traditional regression

[18] The answer to the question of whether machine learning can be used for causal inference largely depends on how causal inference is defined: the Directed Acyclic Graphs (DAG) framework proposed by Pearl or the Potential Outcomes (PO) framework supported by Rubin, Fisher, and Nyman (Angrist & Pischke, 2015; McElreath, 2020; Murnane & Willett, 2010; Pearl & Mackenzie, 2018; Rubin, 1974; Shadish et al., 2002). Because the DAG framework relies heavily on the existence of correlation between variables, machine learning offers more possibilities for causal inference under this framework. For example, supervised machine learning algorithms can model complex relationship among many variables, offering more potential uses under the DAG approach to causality, particularly when examining mediation. While there are few to no examples of this yet for education research, this is an emerging area of research (Imbens, 2019; Zhao & Hastie, 2021).

counterparts, but there is nothing within that box that magically transforms a set of observational data into a source of exogenous variation with a ready counterfactual in order to make a causal claim[19] (Athey, 2019; Imbens, 2019).

However, when considering applications to education policy research in particular – given that the majority of causal inference work conducted has employed the potential outcomes framework – I posit that machine learning algorithms can be most helpful to increase the efficiency of existing causal methods. Just as descriptive and causal methods are generally complementary, machine learning methods can improve the overall richness of a study by combining it with traditional methods (Athey, 2015; Imbens, 2019; Loeb et al., 2017). I provide examples of this for both experimental and quasi-experimental methods.

Experimental Methods

Identifying a Sample Frame

One of the most salient ways that machine learning can complement experimental methods is by identifying a sample frame with supervised machine learning (Crown, 2019). For example, in order to study the effect of a reading intervention for struggling readers at risk of failing a high stakes reading exam, researchers could use a classification algorithm to predict which students are most at risk of failing the reading exam (Porter, 2019). Then researchers could use this list of students as their sample frame, randomizing intervention receipt among these students. This would produce both a treatment and control group of students identified as poor readers, ensuring that the intervention is targeted to students who would most benefit.

[19] There are a few niche extensions of the base machine learning algorithms who developers claim to produce causal estimates under the PO framework. These include an extension of random forests called causal forests (Athey & Imbens, 2016; Wager & Athey, 2018) and a double robust estimator called Targeted Maximum Likelihood Estimator whose inherent flexibility works well when approaching the bias-variance tradeoff in supervised machine learning algorithms (Crown, 2019; Schuler & Rose, 2017).

While it would be ideal to be able to provide help to every student who needs it, researchers are often operating in environments with limited resources. Random assignment is ideal for both identifying the causal impact of an intervention and getting needed buy-in from students if they perceive that they had an equal chance of being selected as their peers. By restricting the sample frame to students most at risk of failing the exam, resources are assured to be delivered to the students who are most likely to have the maximum benefit (Murnane & Willett, 2010).

Creating Blocks

Another way that machine learning can be used to improve the efficiency of randomized control trials is by using unsupervised machine learning to create blocks of participants for stratified random sampling. Sometimes simple randomization can result in undesired imbalances between treatment and control groups. To address this, researchers have found that certain machine learning algorithms can be used to create blocks of participants that are internally homogenous yet are different from other blocks. After randomizing within each group, researchers found that there were fewer imbalances between the overall treatment and control group using the blocks than there were when the counterfactual was created via simple randomization (Grimmer, 2015; Higgins et al., 2014).

Assigning Treatment Status

In a situation with multiple types of treatment as opposed to a binary treatment indicator, machine learning builds on the notion of “multiarmed bandits.” Multiarmed bandits are Bayesian-based algorithms that use information on how previous subjects responded to different types of treatment in order to determine which treatment to assign to new subjects (Scott, 2010). Machine learning algorithms have taken this a step further to incorporate the characteristics of the setting as well as the individual’s characteristics into account in algorithms known as "contextual bandits." The inherent balance with bandits is the need to exploit the previously learned information in order to optimize personalized treatment status for each subject while still wanting to randomly assign treatment in order to identify the causal impact of the treatment (Athey, 2019). In education, this could be helpful when determining the impact of various pedagogical approaches on different types of students in differing classroom contexts, such as low-achieving students in a classroom with high-achieving peers. Bandits have been shown to be most helpful in improving the efficiency of experiments in situations with limited resources, indicating potential future applications to education (Dimakopoulou et al., 2017).

Exploring Heterogeneous Treatment Effects

After a randomized control trial has taken place, it is common to analyze differences in response to the treatment based on certain characteristics with a subgroup analysis (Bloom & Michalopoulos, 2013). While these subgroups are often defined by a pre-intervention characteristic, such as teacher education and experience (Bloom & Michalopoulos, 2013), machine learning offers a way to construct groups who may differentially respond to the treatment based on a combination of large numbers of covariates in a way that traditional regression cannot via grouping algorithms (Chernozhukov et al., 2016, 2018; Grimmer, 2015). Researchers have used different types of supervised machine learning algorithms, such as support vector machines and special types of random forests called "causal forests," to split their treatment group into two groups: one predicted to respond positively to the treatment and the other not predicted to respond positively. Then the control is also split according to these defining subgroup characteristics, allowing researchers to compare the difference in outcomes between the treatment and control groups for each subgroup.

This type of analysis is typically considered exploratory, conducted in a post-hoc manner, and should be specified in a pre-analysis plan to avoid looking for grouping variables that offer statistically significant differences in treatment response (Athey & Imbens, 2016; Davis & Heller, 2017; Friedberg et al., 2019; Imai & Ratkovic, 2013; Wager & Athey, 2018). While this approach is conceptually similar to the process of endogenous stratification (Abadie et al., 2013), the ability of machine learning algorithms to work with high-dimensional datasets allows for more variables to be taken into account. Similarly, work has also been done using the flexible nature of machine learning algorithms, specifically the Bayesian Additive Regression Tree, to automate an examination of a large number of interactions and nonlinear relationships between variables when looking at heterogeneous treatment effects. Researchers note that this broadening of methodological choices frees the researcher from making potentially restrictive decisions about which relationships to include and exclude from the model (Green & Kern, Holger, 2012).

Quasi-Experimental Methods

Creating Synthetic Control Groups

The key to quasi-experimental methods is identifying an appropriate counterfactual to serve as a control group, and one way to do this is with a synthetic control group that has been used as an alternative to researcher-specified control groups (Abadie et al., 2010; Ben-Michael et al., 2021). Machine learning offers an alternative approach to generate control group. One study has shown that a control group created using a grouping machine learning algorithm[20] performed better than the Abadie, Diamond, and Hainmueller synthetic control group across six different simulated data sets on three[21] different performance metrics. The machine learning-generated

[20] Specifically, researchers used principal component analysis.

[21] Researchers evaluated performance with the Mean Square Error, squared bias, and variance.

control group particularly performed better when the underlying data distribution was noisy as opposed to cleanly following a certain distribution because the machine learning algorithm is designed to be more flexible and thus fit the data well (Kinn, 2018).

Improving Matching

Machine learning has also been shown to improve the efficiency of matching techniques that use observed characteristics to create a control group by matching treatment participants to control participants (Cannas & Arpino, 2019; Grimmer, 2015; Hazlett, 2016; Linden & Yarnold, 2016; Sales et al., 2017). For example, one way this has been done is by examining the data that is leftover after the initial matching process was done. Researchers have created an algorithm that uses the unused covariates and unmatched subjects (known as the remnant) to generate more precise estimated treatment effects by decreasing the variance (Sales et al., 2017). Similarly, another study has found a way to use the flexible nature of machine learning algorithms to produce a method for matching the subjects to create treatment and control groups that are more balanced on covariate distributions than traditional matching methods (Grimmer, 2015; Hazlett, 2016). These specific applications of machine learning demonstrate the potential capacity of machine learning to improve estimation methods in the future.

Determining Variable Selection and Functional Form

Researchers have shown how one predictive machine learning algorithm - LASSO (Least Absolute Shrinkage and Selection Operator) – can be helpful for two quasi-experimental designs. One study demonstrated how using the LASSO algorithm for covariate variable selection for regression discontinuity designs improved treatment effect estimates by reducing the standard error estimates. The authors’ simulations showed that using this approach was particularly effective for data sets with fewer than 200 observations, making it appealing for using in education (Anastasopoulos, 2020). Relatedly, researchers of another study used the LASSO algorithm to conduct variable selection and determine functional form for their first stage empirical model for a two-stage least squares regression. Using data to examine the relationship between education and earnings instrumenting on quarter of birth, the researchers found that using the first stage model determined by the LASSO analysis avoided overfitting in the first stage, reduced bias in the second stage, and produced smaller standard error estimates than their traditional first stage model approach (Belloni et al., 2011). Both studies showed how machine learning can be useful for increasing the efficiency of models where the researcher already has an identification strategy for causality.

Conclusion

I have sought to answer the call from researchers to embrace machine learning, understand its strengths and limitations, and interrogate how it can be useful in education research (Gibson & Ifenthaler, 2017; Singer, 2019). Machine learning can be useful in conducting education policy research by helping to automate descriptive statistics by predicting and grouping. It has the potential advantage of over traditional regression when working with large datasets by maximizing its explanatory power to create precise predictions and find seemingly hidden groups of subjects and variables (Athey, 2019; Mullainathan & Spiess, 2017). Predictive machine learning algorithms can sort through a vast number of variables and observations to make accurate predictions while taking multiple interaction and nonlinear relationships into account (Johnes, 2005; Masci et al., 2018). Grouping machine learning algorithms are able to take hundreds of variables into account to create clusters more efficiently than humans alone (Sun et al., 2019). This means that machine learning is helpful for answering research questions such as, “Which students are most at risk of dropping out of high school?” and “What patterns are there among students who drop out of high school?”

Limitations

Limitations of Machine Learning

When incorporated into observational methods, machine learning algorithms have the potential to improve internal validity. However, using machine learning algorithms does not necessarily improve the external validity of an observational study. Unless the datasets used with machine learning algorithms are sampled with the same care as those used in traditional econometric methods, researchers should use caution when making generalized claims about results that extend beyond its dataset (Dede et al., 2016). Similarly, most machine learning algorithms do not inherently produce causal estimates when analyzing observational data. Instead, I argue that machine learning can complement traditional causal methods by increasing their efficiency (Athey, 2019). For example, instead of being used to answer the research question, “What impact does an intervention have on reducing high school drop out?”, machine learning algorithms are generally better suited to answer the question, “Which students should be targeted for an intervention to reduce high school dropout?”

Likewise, at a simplified level, most predictive machine learning algorithms do not usually report the statistical significance of the relationships between covariates and outcomes in the way that traditional regression methods do. For example, if a researcher wants to determine which student characteristics are statistically significantly associated with dropping out of high school, traditional regression is generally more appropriate because machine learning algorithms do not necessarily produce beta coefficient, standard error, and resulting statistical significance estimates. Although some predictive machine learning algorithms can report which predictor variables were most important in generating the predictions, that is a different type of analysis than traditional statistical significance. While predictive machine learning algorithms seek to optimize a function that can be used to create as precise and efficient predictions as possible, they are generally not helpful when exploring theories about *why* those students are likely to be identified (Hastie et al., 2009).

As previously discussed, some of the concerns that plague traditional regression methods can also present problems in machine learning algorithms, such as measurement error and biased data (Jacobucci & Grimm, 2020). If there is inherent bias – such as racial or gender – in the data, the machine learning algorithm may amplify that bias (Gebru, 2021). Therefore, researchers must be just as critical of the dataset they use for machine learning as they are of the data used with traditional quantitative methods in order to produce biases in analytic results (Goldacre, 2008). By being aware of the ways that using machine learning can potentially perpetuate inequity, researchers will be well positioned to avoid that outcome. Furthermore, they could even follow the example of other researchers and use machine learning to explicitly address racial inequity, such as the researchers who used machine learning to analyze police bodycam footage from traffic stops. When they discovered that police spoke to Black residents more disrespectfully than they did to White residents on average, they presented evidence for the department to use in addressing the issue and a method for automating inequality tracking in the future (Voigt et al., 2017).

Limitations of This Paper

One of the limitations of this paper is the need to balance breadth and depth about machine learning. To operate as an entry point for researchers new to machine learning, I did not include several technical details about how to use machine learning or present more advanced machine learning methods, instead focusing on conceptual applications. To learn more about these, I point readers toward my list of recommended readings below. For example, there are ways to combine machine learning methods – called "ensemble learning" – that may hold promise for education research that I did not discuss (Hothorn et al., 2006). I also did not cover every aspect and type of machine learning, rather focusing only on those that I deemed most relevant to education research. Especially as machine learning is a constantly evolving field, there are continuously new innovations in the field that may ultimately be useful for education research.

Policy Implications

Machine learning has the potential to enable researchers to answer new types of research questions and approach analyses more efficiently, such as by automating labor intensive processes (Donnelly et al., 2017). It is better at some tasks (i.e., automating descriptive statistics with predicting and grouping) than others (i.e., causal inference). It has the potential to influence education policies, such as how early warning systems are crafted and how teachers receive feedback on their teaching.

When considering the policy implications of using predictive machine learning algorithms, it is important to consider the contextual implications of true and false positive and negative rates. For example, when predicting which students are most likely to drop out of high school, there are different implications for the true positive rate (i.e., correctly identifying the students who dropped out), true negative rate (i.e., correctly identifying the students who did not drop out), false positive rate (i.e., predicting students to drop out who actually did not), and false negative rate (i.e., predicting students not to drop out who actually did). From a policy perspective in this instance, it is most important to correctly identify the students most likely to drop out to target for intervention. Yet, it still may be acceptable to over-identify students because this ensures that the students who need the intervention receive it and does not necessarily hurt the students who may not need the intervention to receive extra support, assuming that there are enough resources for all these students. Because the desire to maximize and minimize these rates changes based on the context, school personnel should examine these different rates across multiple algorithms as they will not be constant for every model when determining which predictive algorithm to use. Although this is a notable problem for any type of classification strategy, such as with simpler models based on a set of indicators (Cattell & Bruch, 2021), it is important to keep this possibility in mind when using predicting algorithms to determine the most efficiently allocation of resources.

Future Work

Although the research conducted so far is promising, more research is needed to better understand under what conditions machine learning improves upon existing methods (Bird et al., 2021). I see a need to interrogate when it is advantageous to use machine learning for which types of outcomes, data sets, and general contexts. For example, are administrative data sets with only child level information sufficient for accurately predicting high school dropout, or does the addition of family and community level variables enhance the predictive capacity? Furthermore, does this relationship hold for important student subgroups, or are their important nuances to clarify about the performance of machine learning models depending on the students studied? Similarly, when predicting students to target for intervention, more research is needed about the timing of the data used to be useful in a real world – versus research – setting. For example, when predicting high school dropout, while using data from a student's junior year may yield accurate results, it would be helpful to know if data from earlier grades also produced accurate results. By intervening earlier in a student's career, school personnel may be able to make more impact on a student and use fewer resources.

Another important area of focus for the field is how well machine learning techniques work for different sized samples. Given that machine learning was originally designed to work with big data, how do smaller sample sizes generally seen in education research influence the performance of machine learning algorithms? While there have been a few promising studies that use machine learning with small sample sizes (Anastasopoulos, 2020; Bosch et al., 2016; Donnelly et al., 2017), there is more work to be done to convince the field that machine learning works well for small sample sizes. As the education research community better understands the role and applications of machine learning, researchers should consider adding to this technique to their methodological toolbox.

Recommended Texts for Further Reading

There is a vast literature on machine learning and its multiple subspecialities. To help navigate through this multitude, I recommend the following texts based on personal preference for ease of reading while recognizing that there are multiple other excellent texts not mentioned. To learn more about machine learning from an economist perspective, I recommend reading the following texts: Athey (2017), Athey (2019), and Mullainathan & Spiess (2017). To learn more about machine learning from a statistical and computer science perspective, I recommend reading the following text: Hastie, Tibshirani, & Friedman (2009). To learn more about how to measure model performance, I recommend the following texts: Berrar (2019) and Mandrekar (2010). To learn more about how to use machine learning to study education research, I recommend reading the following texts: Bruch et al. (2020), Porter (2019), Donnelly et al. (2017), Sun et al. (2019), Lakkaraju et al. (2019), and Chung & Lee (2019). To learn more about how machine learning can be used with causal inference methods, I recommend reading Athey (2019) and Davis & Heller (2017). To learn more about the ethical implications of using machine learning algorithms, I recommend listening to Gebru's (2021) talk, Kantayya's (2020) documentary, and Lum & Isaac (2016).

Paper 2: What Can Machine Learning Offer when Predicting Special Education and Chronic Absenteeism for Early Elementary Students?

Early warning systems are important for effectively targeting students for intervention. To decrease misidentification of students and use limited resources efficiently, these systems need to be built on accurate prediction models (Engler, 2020; U.S. Department of Education, 2016). In addition to the research on the use of predictive machine learning models in identifying students likely to fail high stakes exams (Porter, 2019), drop out of high school (Ara et al., 2015; Chung & Lee, 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016; Orooji & Chen, 2019; Pagani et al., 2008; Sansone, 2019), and drop out of college (Kotsiantis et al., 2003), there is an emerging field of research that uses these techniques for identifying students who receive special education services (Bone et al., 2016; Duda et al., 2016; Thabtah & Peebles, 2020) and are chronically absent (Bruch et al., 2020) in early elementary school. While this research establishes that machine learning models can accurately predict students who are likely to receive special education services and be chronically absent, none of the research used to identify special education and chronically absent students compares machine learning model performance to traditional regression methods that are more likely to be used in existing early warning systems (O'Cummings & Therriault, 2015).

Our paper speaks to this gap by explicitly comparing how well predictive models perform when based on two traditional regression methods (linear probability and logistic) and three machine learning algorithms (elastic net, decision tree, and random forest). We examine model performance when predicting receipt of special education services in first and second grade for students who were not identified as needing services in kindergarten as well as predicting chronic absenteeism in kindergarten, first grade, and second grade. Our sample consisted of *N* = 1,012 students from Boston Public Schools who were in kindergarten during the 2017-2018 school year who we followed into first and second grades.

In addition to examining five types of models, we also explored model performance with different conceptual blocks of predictor variables. To do so, we draw on an unusually rich data set, spanning the kinds of administrative data common in public schools, district-collected assessments, and additional researcher-collected data on academic achievement, socioemotional measures, family data, and teacher characteristics. Data available on K-2 students typically are less robust than data for older students (Weiland et al., 2021); our rich dataset presents the opportunity to test whether any additional, less-typical data may be particularly useful in early warning systems. Using these data, we compared model performance using the area under the receiver operating characteristic curve (AUC) (Mandrekar, 2010). We then extended our primary analysis to examine model performance for student subgroups, model performance for data only collected by the district, empirically defined likelihood thresholds for identifying students for intervention, true and false positive and negative rates, and the value added of one additional researcher-collected assessment.

Overall, we found that it is possible to create models that are predictive of special education status but not chronically absenteeism that met acceptable performance standards. For the special education models that performed well, the best models came from two of the three machine learning algorithms (elastic net and random forest) and that these models performed well when fit with student demographics and researcher-collected fall of kindergarten academic and executive functioning assessments. Additionally, we found that model performance using only the district-collected data improved when adding one researcher-collected teacher report data source, either the Social Skills Improvement System or the Teacher-Child Rating Scale. We also discuss important policy implications and limitations to our work that we feel shed light on the discussion of how best to identify students to target for intervention.

Review of Literature

Special Education

Importance of Identification for Intervention

Across the wide spectrum of special education identifications, students benefit from high-quality, evidence-based interventions that are tailored to their specific learning needs (Rafferty et al., 2003; Sullivan & Field, 2013). In the short term, interventions have been shown to improve the cognitive development (Guralnick, 1998), literacy and language skills (Diamond et al., 2013; Snowling, 2013), and math achievement (Hanushek et al., 2002) of students with a variety of learning disabilities. Receiving special education services also reduces the likelihood of students being referred for punishment or being suspended due to disruptive behavior (Hurwitz et al., 2021). Research has also shown the importance of effective interventions for longer term impacts, such as decreasing the likelihood of being identified as needing special education services in subsequent school years (Ullery & Katz, 2016), improving academic outcomes by increasing school attendance (Ansari & Gottfried, 2018), and raising the likelihood of completing high school and enrolling in postsecondary education (Ballis & Heath, 2019). Because special education interventions have been shown to be successful for addressing multiple academic and behavioral outcomes, it is imperative that students who need extra support receive it.

Current Process of Identification

Although there are different types of screening tools used to diagnose students in need of special education services based on the type and severity of the disability (Snowling, 2013), the traditional approach begins with a referral. This referral most likely comes from a child's healthcare provider, teacher, or parent (Thabtah & Peebles, 2020). In recent years, an alternative approach to this referral process in which every student participates in a universal screener has become more popular. Students identified as potentially at-risk then complete a second round of more in-depth screenings to determine an appropriate intervention as early as possible in order to mitigate the need for more intensive supports in the future (National Center on Response to Intervention, 2010).

However, the current identification process for special education has opened the door for equity implications. The literature is mixed on whether this process leads to an over- or under-identification of students from minoritized backgrounds, particularly students of color and non-native English speakers when the identification system is based on referrals (Elder et al., 2021; Morgan et al., 2015; Voulgarides et al., 2017). But one thing is clear – identification processes for these student subgroups appear to be different from their majority-culture peers, likely due to multiple factors, including cultural and linguistic barriers between students, families, and schools, implicit racial bias towards students of color, environmental factors, and lower access to resources (Elder et al., 2021; Morgan et al., 2015). Studies also suggest a "frog pond effect" where the same student is less likely to be identified for special education if they are in a classroom with a higher percentage of students in need of special education services compared to if they were one of the few students in a classroom (Elder et al., 2021; Morgan et al., 2015; Voulgarides et al., 2017).

Correct identification for special education students is important for both students who should and should not receive special education services. When students who need special education services are misclassified, they and their general education peers perform worse academically (Ballis & Heath, 2021). Likewise, when students who do not require special education services are incorrectly identified as needing them, they are likely to suffer from lowered self-efficacy that can lead to decreased high school graduation and college enrollment rates. This is particularly true for Black students who are at risk of over-identification (Ballis & Heath, 2021). Just as it is important to receive intervention services as early as possible when they are needed, it is vital to minimize misclassification in order to reduce negative effects on students and to decrease misappropriating limited resources.

Machine Learning for Identification

As a contrasting approach for identifying students in need of special education, machine learning has been shown to be beneficial in special education classification in a few cases, particularly for autism spectrum disorder. Using data from a traditional autism spectrum disorder screening questionnaire, researchers compared the performance of multiple supervised machine learning algorithms and found all models to be between 85%-90% accurate (i.e., true positive rate) for predicting correct identification for adolescents and 90%-95% for adults. Compared to a traditional screening approach that was anecdotally considered subjective and slow, authors found the machine learning approach to be more accurate (Thabtah & Peebles, 2020). A similar study used data from two autism screening questionnaires and found that a machine learning algorithm was able to correctly identify people with autism spectrum disorder at 96.7% accuracy (i.e., true positive rate) for people over ten years old and 89.2% below ten. However, the algorithm did not perform as well when correctly identifying people without autism spectrum disorder; the true negative rate was only 53.4% for those over ten years old and 59% for those below ten (Bone et al., 2016).

In addition to being an accurate predictive tool, machine learning can be helpful for narrowing down the number of items needed on autism spectrum disorder screeners. One study found that compared to the original 65 questions on a traditional screening tool, the multiple predictive machine learning algorithms were highly accurate (over 90%) in correctly discerning (i.e., true positive rate) between autism spectrum disorder and attention deficit hyperactivity disorder for children and adolescents using only five questions (Duda et al., 2016). Although there is limited evidence on using machine learning to identify special education status, preliminary results look promising and invite more exploration.

There are some limitations to these studies when considering their scalability and external validity. Although the Bone et al. (2016) study showed high true positive rates, it used data from a screening questionnaire that took an average of two to three hours per child to administer, which is not scalable for a school to administer to every student (Bone et al., 2016). Another study attempted to address this scalability issue by using simplified data from a screening questionnaire, yet their sample consisted of adults aged 17 to 64, making it an imperfect match for our study that focused on students aged 5 to 6.5 (Thabtah & Peebles, 2020). Although the Duda et al., 2016 study included children in their sample such that the median age was 8.6 years old, their entire sample consisted of participants with either autism spectrum disorder or attention deficit hyperactivity disorder, which may present an inherently different approach to classification compared to determining students who need any type of special education services versus those who do not (Duda et al., 2016). Additionally, all three of these studies did not explicitly compare how well machine learning algorithms compared to traditional linear regression models, meaning the value-added of newer machine learning methods in these early warning systems is unknown.

Chronic Absenteeism

Importance of Identification for Intervention

Although traditionally discussed for older students (Nield & Balfanz, 2006; Rumberger & Thomas, 2000), the prevalence of chronic absenteeism among early elementary students has recently come to light. While 10% of kindergarten through third grade students are chronically absent nationally (Chang & Romero, 2008), this statistic hides the fact that 25% of kindergarten students are chronically absent. Indeed, kindergarten students experience the highest rate of chronic absenteeism until middle school (Ansari & Gottfried, 2018; Chang & Davis, 2015). It is important to identify students who are likely to be chronically absent to target for invention because chronic absenteeism is associated with lower academic achievement and an increased risk of repeating grades and ultimately dropping out of school (Ansari & Purtell, 2018; Chang & Davis, 2015; Morrissey et al., 2014; Nield & Balfanz, 2006; Rumberger & Thomas, 2000). The potential implications of chronic absenteeism do not take years to come to fruition. Due to both missing instruction and out-of-school factors (Pyne et al., 2021), being chronically absent in kindergarten is associated with lower academic performance in first grade (Chang & Romero, 2008). Additionally, chronic absenteeism can have a multiplier effect because students who are chronically absent in early grades are more likely to be chronically absent in later grades (Balfanz & Byrnes, 2012).

There are important equity issues when studying which types of students are most likely to be chronically absent because chronic absenteeism is more common for children living in poverty, students of color (except for Asian Americans), special education students, students in poor health, immigrants, and English language learners (Chang & Davis, 2015; Chang & Romero, 2008). In addition to being more likely to experience chronic absenteeism, many of these groups of students experience the impacts of chronic absenteeism more severely. Although being chronically absent is associated with lower academic achievement for students regardless of gender, race, ethnicity, and income (Balfanz & Byrnes, 2012; Chang & Romero, 2008; Morrissey et al., 2014), this relationship is stronger for students from families with low incomes and special education students as well as students of color (Balfanz & Byrnes, 2012; Chang & Romero, 2008; Ready, 2010). Therefore, it is vital to identify students at risk of chronic absenteeism for all students and especially those from disadvantaged circumstances.

While missing academic instruction is an important effect of being absent, research has shown that absences – particularly unexcused absences – are a signal of broader out-of-school factors that are associated with student outcomes (Pyne et al., 2021). The research that has shown that there are effective interventions for reducing chronic absenteeism for young students targets some of these influential factors. One approach for prekindergarten students in Chicago aimed at changing parental beliefs about the importance of attendance increased the average number of days attended by 2.4 days and reduced chronic absenteeism by 9.3% (Kalil et al., 2019). A second study found that intervention also targeting parental beliefs decreased chronic absenteeism in kindergarten through fifth grade students by 15% (Robinson et al., 2018). Other work has also shown that chronic absenteeism may be reduced by smaller class sizes, race match between students of color and teachers (Tran & Gershenson, 2021), and serving breakfast during school hours (Kirksey & Gottfried, 2021). This recent research indicates that if students can be properly identified, there are successful interventions for reducing chronic absenteeism, particularly in resource limited environments where schools may not be able to implement these interventions universally.

Current Process of Identification

As the potential dangers of chronic absenteeism become more widely known, schools that have traditionally only tracked attendance information for accountability purposes are beginning to use it to identify students at risk for chronic absenteeism (Balfanz & Byrnes, 2012; Chang & Romero, 2008). In particular, schools historically only tracked and reported the average daily attendance for the entire student body, which could mask massive variation in attendance for individual students.

As schools have begun to devote data on student absenteeism, there have been conflicting definitions of chronic absenteeism. While most schools define chronically absent as missing at least 10% of school days, some use a 15-day cutoff. Additionally, some schools also track "severe" or "excessive" chronic absenteeism that refers to missing at least 20% of school days. Depending on the state, some schools only include students who were enrolled for the entire academic year in their chronic absenteeism reporting while others include any student who was enrolled for at least 90 days. Across all of these definitions and nuances, chronic absenteeism is clearly defined as both the sum of excused and unexcused absences as opposed to truancy, which only focuses on unexcused absences and conventionally is associated with misbehavior (Chang & Romero, 2008).

After establishing a method for tracking absences for students, schools are increasingly using early warning systems to identify which students are at risk of being chronically absent to intervene before students miss several days of school (Bruch et al., 2020; U.S. Department of Education, 2016). Similar to the early warning systems used in high schools to target students at risk of dropping out (Battin-Pearson et al., 2000; O'Cummings & Therriault, 2015; Parr & Bonitz, 2015; Rumberger & Larson, 1998), these systems typically use factors known to be associated with chronic absenteeism for their predictions (Bruch et al., 2020; U.S. Department of Education, 2016). At the child and family level, a student is more likely to be chronically absent if they are from an immigrant family, are a non-native English speaker, low income, in poor health, require special education services, are of color, or had parents who had negative experiences with schools themselves (Chang & Davis, 2015; Chang & Romero, 2008). At the school level, a student is more likely to be chronically absent if their school lacks resources and a coordinated system in place to conduct home visits and reach out to families individually (Chang & Romero, 2008).

Overall, the prevalence of such early warning systems for chronic absenteeism for early elementary students is unclear given the somewhat recent attention to the issue and the proprietary nature of the exact predictions (O'Cummings & Therriault, 2015). For example, one organization found that seven states have statewide policies in place requiring the use of early warning systems, but there is little to no data examining how these students operate in practice (National Association of State Boards of Education, n.d.).

Given this paucity of evidence, when considering the structure of early warning systems for chronic absenteeism in the early grades, it may be helpful to consider how early warning systems for high school dropout – a historically more established system – function. As of 2015, 52% of public high schools had an early warning system to identify students likely to drop out. Although larger schools are more likely to use an early warning system than smaller schools are, there is no statistical difference in a school's likelihood of having a system in place based on graduation rate, poverty level, or location. Additionally, at the school level, administrators and guidance counselors are the most common school personnel to monitor the system for updated projects, and they most commonly check the system weekly (U.S. Department of Education, 2016). In terms of effectiveness, one randomized trial study found that high schools in Michigan, Indiana, and Ohio that used early warning systems to monitor chronic absenteeism, among other outcomes, reduced their rate of chronically absent students from 14% to 10% (Faria et al., 2017). Despite not having a more comprehensive look at early warning systems that focus on chronic absenteeism, particularly in early elementary school, it is clear that the accuracy of these systems is crucial in order to identify students who would benefit from early intervention

These identification systems have become more important during the COVID-19 pandemic because school attendance decreased significantly across the country, with one report estimating three million students having little to no access to education during the pandemic (Korman et al., 2021). Students were more likely to miss school if they were from families with low incomes, students of color, English language learners, had a disability, in foster care, in a high poverty district, in a district that predominantly served students of color, or experienced remote instruction (Carminucci et al., 2021; Korman et al., 2021). These absences have important equity implications looking forward, so the use of an accurate system to detect students likely to be absent may be helpful in mitigating the effects of absenteeism.

Machine Learning for Identification

Despite how machine learning has proven effective in identifying other types of students in need of additional supports under certain circumstances with high levels of accuracy (Lakkaraju et al., 2015; Porter, 2019), little work has been done using machine learning to predict students at risk of chronic absenteeism. To our knowledge, only one study has used machine learning to predict students at risk of chronic absenteeism. It used data from two school districts in Pennsylvania, one large (*N* = 28,719) and one small (*N* = 4,614), with three different machine learning algorithms. It found that the best performing model correctly identified 75% elementary students who were chronically absent for the larger district and 74% for the smaller district; likewise, the model for the larger district had a false positive rate of 31% and 34% for the smaller district (Bruch et al., 2020). In this study, the authors used demographic variables, course performance, state test scores, and behavior incidents as predictors. This study indicates that machine learning may be a useful tool for accurately predicting students at risk of chronic absenteeism, but more research is needed.

Current Study

Our exploratory study seeks to build on prior work by examining how – and under what circumstances – machine learning compares to traditional regression methods when predicting special education and chronically absent students in early elementary school. We address three research questions:

1. How does machine learning compare to traditional regression methods when identifying students who began kindergarten not identified as requiring special education services but are later identified as needing services by the end of first and second grade? How does this performance vary based on the type and timing of data used?
2. How does machine learning compare to traditional regression methods when identifying students who are chronically absent during kindergarten, first, and second grade? How does this performance vary based on the type and timing of data used?
3. If the district could add one assessment to its standard operations, which researcher-collected assessment would enhance prediction best?

Our study addresses several gaps in the current literature. Notably, no studies to date have explicitly compared machine learning algorithms to traditional linear or logistic regression that are more likely to be used in early warning systems. The most similar approach that one study (Bruch et al., 2020) used was an elastic net with a logistic base. Although an elastic net is considered a machine learning algorithm, it is more like traditional regression than other algorithms. In this study, we use an elastic net algorithm as a bridge between both a linear and logistic regression and the more functionally free non-parametric machine learning algorithms (i.e., decision tree and random forest). This study will contribute to the gap in the literature by clearly comparing multiple types of machine learning algorithms to traditional regression.

Additionally, our study builds on existing work by exploring under what conditions all the models – both traditional regression and machine learning – operate with different sets of data. In addition to district-collected administrative and academic achievement data, we also have access to rich researcher-collected data on student academic achievement, student intrapersonal and interpersonal skills, family characteristics, and teacher characteristics that school districts typically do not have access to due to limited resources. Our dataset is richer and more nuanced with respect to academic and socioemotional achievement, parental variables, and teacher information than the dataset used in the most similar existing study (Bruch et al., 2020) By capitalizing on this depth data source, we can interrogate whether models respond better to data that school personnel often do not have, leading to potential policy implications if this additional data enhance model performance.

Method

Participants and Setting

The sample for this study consists of 1,012 students who were in enrolled in kindergarten in the Boston Public Schools (BPS) during the 2017-2018 academic year. We recruited students from 130 classrooms and 64 schools as part of a larger longitudinal study examining the effects of the BPS public prekindergarten program (cite Attender, Lottery, and Sustaining Environment papers). We followed these students via administrative data if they stayed in BPS as they progressed to first grade during the 2018-2019 school year (*N* = 894; 88% of study sample) and second grade during the 2019-2020 school year (*N* = 825; 82% of study sample).

Students were diverse with respect to gender, race/ethnicity, income, and language status and were representative of the wider school district (cite Attender paper). On average at the beginning of kindergarten as described in Appendix C Table 1, the study participants were 48% female, 60% eligible for free or reduced-price lunch, 14% Asian, 25% Black, 33% Latinx, 25% White, 3% mixed or other race/ethnicity, 51% Dual Language Learners, and 39% Limited English Proficient. 63% of students reported English as their first language, 74% noted English as their home language, and 75% reported English as their parental language preference. The students were an average of 5.49 years old (SD = 0.29) as of September 1, 2017.

In terms of our two outcomes, our sample was not representative of the broader school district with respect to special education. This was due to how the initial sample was constructed because substantially separate special education classrooms were excluded in the prekindergarten sample, creating a lower percentage of our kindergarten sample identified as special education than the district kindergarten average (sample 7% versus district 14%). However, given our wider recruitment strategy for our kindergarten sample, the percentage of our sample receiving special education services in first (sample 12% versus district 19%) and second (sample 14% versus district 21%) grade is closer to the district average albeit still lower. Our sample was also differed from the broader school district when considering the percentage of students who were chronically absent. Our sample had lower rates of chronic absenteeism than the district in kindergarten (sample 16% versus district 25%), first grade (sample 12% versus district 20%), and second grade (sample 9% versus district 15%) (Table 4). We return to these differences in the limitations section of this paper.

Procedure

This study was conducted with the approval of IRB (Institutional Review Board) at the lead and partner institutions under the approval number was HUM00193769.

School and Classroom Recruitment

The sample for this study is drawn from a larger study sample that initially began its sample with prekindergarten students by randomly selecting 25 public schools offering the BPS public prekindergarten program. Of these, 21 consented to participate and one was chosen as a pilot school, leaving 20 remaining schools for the primary study. Within these 20 schools, all general education and inclusion teachers were invited to participate and 96% (*N* = 51) agreed. As the sample children progressed from prekindergarten to kindergarten, we invited all of the kindergarten teachers assigned to these students to participate and 95% (*N* = 93) agreed (cite Attender, Lottery, and Sustaining Environment papers).

Student Recruitment

As part of the broader study sample, all children in the initial 20 prekindergarten classrooms were invited to participate. Of the 81% of students for whom we received completed consent forms from their parents, we randomly selected 50% (~6-10 students) from each classroom to be a part of the study sample. For this study, we followed these children into kindergarten if they stayed within BPS. In their kindergarten classes, we repeated this same consent process with the participating classrooms in the fall of 2017 to re-consent the students from our prekindergarten sample and to add additional students from their kindergarten classrooms, bring our total baseline sample size of 1,012.

Administrative Data

We used administrative records from BPS from the 2017-2018, 2018-2019, and 2019-2020 school years that provided information on students' demographic information including first, home, and parental language information. We also used this administrative data to confirm and replace when missing which students attended the BPS public prekindergarten enrollment as reported by parents as well as which classrooms and schools students were enrolled in. The administrative data also included our outcomes of interest: special education status and attendance and enrollment records used to create the chronic absenteeism indicators. For the 2019-2020 school year – when our students were in second grade – the attendance and enrollment variables were truncated at March 14, 2020 to account for students' move to at-home schooling as a result of the COVID-19 pandemic.

Direct Assessments

To directly assess children on their math, receptive vocabulary, and executive functioning skills, we trained staff to reliability and conducted the assessments in the fall of 2017 (September 22 – December 18) and again in the spring of 2018 (April 2 – June 11). All of the assessments were conducted on the same day in a quiet place outside of the child's classroom, such as an empty office or classroom. It took an average of 45 minutes per child to administer all assessments at each time point. In order to ensure high quality administration, a master's-level supervisor observed 10% of all field assessments. To determine the language of assessments, we used the Pre-language Assessment Scale (PreLAS) Simon Says and Art Show tests (Duncan &

DeAvila, 1998) as a warm-up to the assessment battery and to determine the administration language for a subset of assessments (Barrueco et al., 2012). 964 (95%) children participated in the fall PreLAS assessment and 923 (91%) participated in the spring PreLAS assessment. Of the 964 children who participated in the fall PreLAS, 23 (2%) did not pass and completed a subset of assessments in Spanish in the fall. Likewise, of the 923 who participated in the spring PreLAS, 6 (1%) did not pass the PreLAS and completed assessments in Spanish in the spring; all 6 students who did not pass the spring screener also did not pass the fall screener. Spanish was the first language for all students who did not pass the PreLAS at either time point. The exceptions to this process were some of the literacy and language assessments – the Dynamic Indicators of Basic Literacy Skills – Next (DIBELS) subscales as discussed more in the measures section below – were administered by the district teachers as opposed to researchers as part of their normal district assessment activities.

Teacher Reports

We asked teachers of participating students to complete a short report on each student assessing children's socio-emotional and self-regulation skills. Teachers completed these reports in the fall of 2017 (September 25 – January 30) and again in the spring of 2018 (April 23 – July 24). Of the 1,012 students in the study, 825 (82%) had completed teacher reports in the fall and 878 (87%) in the spring.

Parent Surveys

In the fall of 2017, we reached out to the parents of the consented study students via email and text message to ask them to complete the 20-minute parent survey. We also sent biweekly reminders to the parents asking them to complete the survey and used backpack mail for parents who did not complete the survey electronically. The survey consisted of parental demographic information along with questions about educational experiences and parental perceptions about the importance of schooling and attendance. Although the majority of parents completed the survey in English, we also translated the survey into Spanish, Vietnamese, and Mandarin. For completing the survey, parents received a $25 gift card. Of the 1,012 students in our study, 437 (43%) of parents completed the survey in kindergarten. When possible, we replaced missing values of the parent survey with data from the parent survey from the first year of the larger study, bringing the total parent survey completion rate to 488 (48%).

Teacher Surveys

In the spring of 2018, we asked teachers of study participants to complete a survey asking demographic questions as well as questions about their educational background and teaching license. Of the 130 teachers in our study, 128 (98%) completed the teacher survey, representing 915 (90%) of the 1,012 study students. Teachers completed the survey between April 23 and July 24.

Measures

Special Education and Attendance Data from Administrative Data

We used administrative data from BPS to identify our two primary outcome variables: special education and chronic absenteeism status in kindergarten, first, and second grades. We operationalized our binary indicator of special education status if students were flagged as such in the administrative records at any point during the school year. For our main models, we included a student in our special education indicator if they were identified in the administrative data as special education, regardless of the type of special education designation. We discuss this later in the limitation sections of our paper.

In the administrative data, we had records of the number of school days the students were enrolled in BPS and the number of days that they were absent. Although the records included the number of days students had both excused and unexcused absences, we added together these two types of absences for our analysis following the literature in order to examine the number of days those students were absent regardless of the reason. If students switched classrooms and/or schools during the school year, we were able to follow them through these moves and took all attendance and enrollment data into account. We used the attendance records to construct binary indicators of chronic absenteeism. We defined chronic absenteeism as a student being absent – either as excused or unexcused - for 10% or more of school days enrolled for students enrolled for at least 90 days (Chang & Romero, 2008). In Appendix A, we present results using alternative definitions of chronic absenteeism.

Demographic Data from Administrative Data

We used also administrative data from BPS for information on child demographics for the fall of students' kindergarten year. We created binary indicators variables to denote whether students were female, their race/ethnicity (Asian, Black, Latinx, White, or mixed/other race), and whether their first language was English, home language was English, and parental language preference was English. We created indicators for whether students received free or reduced-price lunch (FRPL), were Dual Language Learners, and were classified as Limited English Proficient. We also used a continuous variable for their age as of September 1, 2017. We then used the administrative data to supplement the information about children's prekindergarten experience when possible if it was missing information from the parental surveys as described below. We chose these predictor variables in order to align with previous work done with this sample (McCormick et al., 2020) and because previous literature showed their relationship to child outcomes (Choi et al., 2018; Powell et al., 2010; Reardon & Portilla, 2016).

Achievement Measures from Direct Assessments

We used multiple measures of literacy/language, math, and executive functioning, all of which have been used and discussed in more detail in previous work with our sample (McCormick et al., 2020).

Literacy and Language Measures. To measure receptive vocabulary, we used the raw score from the Peabody Picture Vocabulary Test IV (PPVT) in the fall and spring of kindergarten. The PPVT is a nationally normed exam with strong reliability estimates (Dunn & Dunn, 1997). For this measure only, we assessed all students in English regardless of their PreLAS performance in order to obtain a measure of English vocabulary.

To measure students' literacy skills, we used the raw scores from the teacher-administered[22] Dynamic Indicators of Basic Literacy Skills – Next (DIBELS). The DIBELS has excellent validity and is widely used (Good et al., 2011). We used subtests to measure students' letter knowledge (Letter Naming Fluency: LNF), phonological awareness (First Sound Fluency: FSF; Phoneme Segmentation Fluency: PSF), and alphabetic principle (Nonsense Word Fluency Correct Letter Sounds: CLS; Nonsense Word Fluency Whole Word Read: WWR). In accordance with their curriculum, teachers administered the FSF and LNF subtests in the fall and LNF, PSF, CLS, and WWR subtests in the spring.

Math Measures. To measure math skills, we used the raw scores from both the Woodcock Johnson Applied Problems III (WJAP) and Research-Based Early Mathematics Assessment (REMA). The WJAP assesses numeracy and early mathematics and has strong

[22] The DIBELS was the only direct assessment administered by teachers. For more, see (cite Attender paper).

psychometric properties (Woodcock et al., 2001). For students who did not pass the PreLAS, we assessed them using the Spanish language equivalent called the Batería III Woodcock Muñoz (Schrank et al., 2005). We used the WJAP in both the fall and spring of kindergarten. The REMA assessment goes beyond numeracy to include geometry and measurement skills and has high internal reliability across subscales (Clements & Sarama, 2007). For students who did not pass the PreLAS, we used the Spanish language equivalent (Clements & Sarama, 2007). We only assessed students with the REMA in the spring of kindergarten.

Executive Functioning Measures. To measure short-term memory, we used the categorical score of the Forward Digit Span (FDS) assessment (Gathercole, 1999). This assessment has good test-retest reliability (Lipsey et al., 2017) and is predictive of student achievement (Bull et al., 2008). We used the FDS in both the fall and spring of kindergarten and assessed students in Spanish if they did not pass the PreLAS. To measure working memory, we used the categorical score of the Backward Digit Span (BDS) assessment. This measure is commonly used to measure working memory (Coulacoglou & Saklofske, 2017; Holdnack, 2019). Like the FDS, we assessed students using the BDS in both fall and spring of kindergarten and used the Spanish version if students did not pass the PreLAS.

We also used the percent correct scores of the Hearts and Flowers (H&F) assessment. We used the mixed subscale to measure cognitive flexibility and the incongruent subscale to measure inhibitory control, one of the three components of executive function (Weiland et al., 2014). This measure has strong reliability scores for young children through young teenagers (Davidson et al., 2006). We used both trials of H&F in the fall and spring of kindergarten and administered the assessment in Spanish for students who did not pass the PreLAS. To measure self-regulation, we used the raw scores from the Preschool Self-Regulation Assessment (PSRA). The PSRA is an assessor report with strong reliability evidence, and we used the attention-impulsivity (AI) and positive emotion (PE) subscales (Raver et al., 2011; Smith-Donald et al., 2007).

Additional Assessments from Teacher Reports

We designed the reports that teachers completed on every student in both the fall and spring using questions from two assessments that measure various aspects of students' intrapersonal and interpersonal competencies: the Social Skills Improvement System (SSIS) and the Teacher-Child Rating Scale (TCRS). Both assessments asked teachers to complete a battery of Likert-scale questions that have been previously validated to form constructs. From the SSIS, we used questions to measure six constructs: students' cooperation, engagement, self-control, externalizing behavior, internalizing behavior, and hyperactivity/inattention (Gresham & Elliott, 2008). From the TCRS, we used questions to measure students' academic orientation (Hightower et al., 1986).

Family Data from Parent Surveys

We used the parental surveys to create a set of binary indicator variables describing the parents' education level (high school diploma or less, two-year degree, Bachelor's degree, and advanced degree). We also created a set of continuous variables to measure the age of both the mother and father at first child's birth, the parental age when completing the survey, and the household size. We also created binary indicator variables to denote whether there was at least one person in the household working full time (defined as 35 hours per week), whether parental respondent was married or had a partner, whether the child had received an Early Intervention Services or Individualized Family Service Plan, whether the parent considered both daily prekindergarten and kindergarten attendance very important (a score of 5 on a 1 to 5 Likert scale), and whether the parent was satisfied with their child's school assignment. We created binary indicators to represent whether the child had attended prekindergarten in the BPS public program, a non-BPS center-based prekindergarten program, or did not received center-based care. To measure household income, we created binary indicators to denote whether the household income was less than $25,000, between $25,000 and $59,999, and over $60,000. We also asked parents to what extent they engaged with constrained and unconstrained literacy/language and math activities[23] as well as experiential learning activities with their children on a weekly basis. We also included a continuous measure of how many children's books – including library books – were in the home. We chose these predictor variables in order to align with previous work done with this sample (McCormick et al., 2020) and because previous literature showed their relationship to child outcomes (Bloom & Weiland, 2015; Puma et al., 2005; Weiland et al., 2018).

Teacher Data

From the teacher surveys, we used continuous measures of teachers' years of teaching experience generally, years of teaching experience in kindergarten, and their age. We also created binary indicators of whether teachers were female and their race/ethnicity (White, Black, and Latinx). We also created binary indicators of the type of their highest degree (education specialist/professional diploma, Associate's, Bachelor's, Master's, or Doctorate), the area of their highest education degree (early childhood education, elementary education, special education, child development, reading specialist, curriculum and instruction, bilingual/bicultural education, other type of education, and other non-education), and the area of their current teaching license (early childhood education, elementary education, English Language Learners,

[23] For more detailed information on constrained versus unconstrained literacy/language and math activities, see (McCormick et al., 2020).

teacher of students with moderate disabilities, teacher of students with severe disabilities, other type of teacher, teacher specialist, administrator, professional support personnel, or none). We chose these variables given the previous work done showing their relationship to child outcomes (Early et al., 2007; Landry et al., 2006; Lin & Magnuson, 2018).

Analytic Approach

Missing Data

We had varying rates of missing data depending on the data source. We had no missingness for our outcome or child demographic variables because those came from administrative data. Specifically, we had between 6%-11% missingness for the fall researcher-collected direct assessments; 23% missingness for the fall district-collected academic assessments; 18-19% on the fall teacher report assessments; 9%-24% missingness for the spring researcher-collected direct assessments; 22% missingness for the spring district-collected academic assessments; 13-14% on the spring teacher report assessments; 34-60% on the family data from the parent surveys; and 8-14% on the teacher data from the teacher surveys (Appendix C Table 1).

We used conditional mean imputation, also known as regression imputation, for the variables with missingness by replacing missing values with estimates derived from regression equations fit using non-missing values (Enders, 2010; Harrell Jr., 2015). Although conditional mean imputation can lead to biased parameter estimates and dampened standard error estimates under certain circumstances that multiple imputation may resolve (Enders, 2010), we chose this approach because the focus of our models is on predictive power rather than parameter estimation. By using a simpler missing data strategy, we also hoped to create a model that would be transparent to practitioners. To retain information about which students originally had missing information in case the missingness was informative for the models, we also included a binary indicator for each variable that had missingness that denoted whether a student initially was missing a value for that variable. We also include a version of our analysis conducted using multiple imputation in Appendix A as a robustness check.

RQ1 & 2: Identifying Special Education & Chronic Absenteeism Status

We operationalized our two outcome measures (indicators of special education and chronic absenteeism status) as binary indicators. When coding our predictive models, we forced the software to recognize the binary variables as continuous from zero to one, inclusive, to act as a likelihood risk score to provide more nuanced model interpretation for practitioners and policy makers. This analytical decision could ultimately allow schools to set their own risk threshold when identifying at-risk students (Bruch et al., 2020).

To compare traditional regression to machine learning model performance, we first fit the data with a both linear probability and logistic model. Then we fit the data with three different machine learning algorithms that were common in the machine learning education literature (Bruch et al., 2020; Chung & Lee, 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016): elastic net with a linear base, decision tree, and random forest. In addition to their previous use in the literature, we also choose these algorithms in order to include both parametric (elastic net) and nonparametric (decision tree and random forest) options. While the elastic net model is a machine learning algorithm, we chose it to function as a bridge between the traditional linear probability and logistic models and the more statistically flexible decision tree and random forest models. In contrast to traditional regression models where there is generally a consensus on the main type of model to fit based on the nature of the outcome variable, there are multiple machine learning algorithms designed to be used for the same purpose. Therefore, it is becoming increasingly common to use three to five algorithms and compare predictive performance across models (cite paper 1).

Similarly, there is not a universally agreed upon method of dividing the dataset into the training and testing data for the machine learning algorithms in the education literature. Given the relatively small size and structure of our dataset (i.e., not having multiple cohorts), we randomly selected 80% of the students to be in our training dataset and remaining 20% in our testing dataset. We fit all five models (linear probability, logistic, elastic net, decision tree, and random forest) using the training data in order to establish parameter estimates that define the models. We then used these to predict the outcomes using the testing data. While fitting the model with the training data optimizes its flexibility, we obtained the fit statistics used to measure model performance from the testing data. The use of training and testing data is commonly used in machine learning work as it acts as a check on overfitting the models (Athey, 2019; Mullainathan & Spiess, 2017). To account for our relatively small sample size, especially after splitting into testing and training data, we repeated this 80/20 split four more times to create a total of five model iterations and averaged the fit statistics across the five sets of testing data for our final value of model performance.

To examine the part of the research question that specifies examining how model performance varies based on the input data used, we fit all five models using conceptual blocks of data: Block 1 = child demographics, Block 2 = Block 1 with direct assessments (fall of K), Block 3 = Block 2 with teacher assessments (fall of K), Block 4 = Block 3 with family data, Block 5 = Block 4 with teacher data, Block 6 = Block 5 with direct assessments (spring of K), Block 7 = Block 6 with teacher assessments (spring of K). We chose to fit the conceptual blocks in this order to imitate the timeline of when schools would be able to access similar types of data in practice.

To evaluate model performance given that the outcome will be operationalized as a continuous variable, we used the area under the receiver operating characteristic curve (AUC). AUC is a commonly used measure that ranges from zero to one, with measures closer to one indicating better performance. Generally, an AUC of 0.5 suggests no model discrimination, 0.7-0.8 is acceptable, 0.8-0.9 is excellent, and greater than 0.9 is outstanding model performance (Mandrekar, 2010). The AUC measures the area under the curve created by graphing the sensitivity (i.e., the proportion of true positives out of total actual positive, also known as true positive rate or recall) versus 1 – specificity (i.e., the proportion of true negative out of total actual negative, also known as the true negative rate) at every possible threshold from zero to one for turning the continuous likelihood into a binary classification. Just as using a continuous outcome operationalization provides a more granular and nuanced measure, the AUC provides an analogous measure compared to the performance metrics obtained from using a binary classification (i.e., true/false positive/negative rates) (Bruch et al., 2020).

RQ3: Value Added of One Additional Researcher-Collected Data Point

We examined the potential value that one additional assessment would add if BPS were able to invest in one research-collected assessment to add to their district-collected data. To calculate this, we fit eight models, all using the student demographics and fall of kindergarten DIBELS assessments plus one researcher-collected assessment collected in the fall of kindergarten. The first model included the Peabody Picture Vocabulary Test (PPVT), the second the Woodcock Johnson Applied Problems (WJAP), the third the Research-Based Early Mathematics Assessment (REMA), the fourth the Digit Span Forward (FDS), the fifth both the mixed and incongruent subscales of the Hearts and Flowers (H&F) assessment, the sixth both the attention/impulse control and positive emotion subscales of the Preschool Self-Regulation Assessment (PSRA), the seventh the academic orientation subscale of Teacher-Child Rating Scale (TCRS), and the eighth the cooperation, engagement, self-control, externalizing behavior, internalizing behavior, and hyperattention/inattention subscales of the Social Skills Improvement System (SSIS) assessment. We then repeated these models using data from both the fall and spring of kindergarten, i.e., the student demographics, fall and spring of kindergarten DIBELS assessments, and the fall and spring of kindergarten additional researcher-collected assessment. Similar to the first and second research questions, we randomly selected 80% of students to be in the training data and the remaining 20% in the testing data. We also repeated this process for a total of five times and averaged the fit statistics across the five testing datasets. We used the AUC values to compare model performance to the models fill with the full set of researcher-collected predictors from the first two research questions.

Results

RQ1: Predicting Special Education Status

In our models predicting receipt of special education services, we found overall that model performance improved most with the addition of the fall of kindergarten academic and executive functioning assessments and that the elastic net and random forest models performed best. When predicting first grade special educations status, the linear probability model only had an acceptable model performance when using the fall of kindergarten academic, executive functioning, intra-, and interpersonal assessments (AUC = 0.702, Table 5, Panel A) but had an acceptable fit for every model except for the only with demographics when predicting second grade special education status (Table 5, Panel A). The logistic models did not converge for either grade after the addition of the fall of kindergarten academic and executive functioning assessments. Of the logistic models that did converge, the only one that had acceptable performance was the model predicting second grade special education status using fall of kindergarten academic and executive functioning assessments (AUC = 0.763, Table 5, Panel B).

For the machine learning models, none of the decision tree models performed well (Table 5, Panel D). The elastic net and random forest models performed comparably, with both models reaching an acceptable model performance for both grades for the model that incorporated fall of kindergarten academic and executive functioning assessments and maintaining either an acceptable or excellent fit for every subsequent model (Table 5, Panels C and E). The models with the best performance for predicting first grade special education status was the random forest model with all predictor groups (AUC = 0.770, Table 5, Panel E) and elastic net with all predictor groups for predicting second grade special education status (AUC = 0.844, Table 5, Panel C).

RQ2: Predicting Chronic Absenteeism Status

Overall, every model had a poor performance when predicting chronic absenteeism for every grade. There was a large amount of variability in model performance in that it was unclear how the inclusion of different sets of predictor variables influenced performance, and the linear probability, elastic net, and random forest models had the best performance. For the linear probability models, the best performing model was that predicting first grade chronic absenteeism using fall of kindergarten academic, executive functioning, intra-, and interpersonal assessments (AUC = 0.674, Table 6, Panel A). Like the special education models, the logistic models did not converge for most of the predictor variable groups (Table 6, Panel A).

For the machine learning models, the best performing model predicting kindergarten chronic absenteeism was the elastic net model with fall of kindergarten academic, executive functioning, intra-, and interpersonal assessments (AUC = 0.647, Table 6, Panel C). The best machine learning model for first grade chronic absenteeism was the elastic net model with fall of kindergarten academic and executive functioning assessments (AUC = 0.666, Table 6, Panel C). The best machine learning model for second grade chronic absenteeism was the elastic net model with fall of kindergarten academic, executive functioning, intra-, and interpersonal assessments plus the family data (AUC = 0.666, Table 6, Panel C).

RQ3: Value Added of One Additional Researcher-Collected Data Point

Overall, the results indicated that either the SSIS or TCRS would be the best assessment the district could invest in to improve predictive model performance, particularly if that model was fit with both fall and spring of kindergarten data. When predicting special education services using data from the fall of kindergarten, the best performing model predicting receipt of special education services for first grade was the elastic net model using the SSIS (AUC = 0.733, Table 7, Panel C, Column 1) while the best model for second grade was the random forest model that used the PSRA (AUC = 0.829, Table 7, Panel E, Column 2). When predicting chronic absenteeism using data from the fall of kindergarten, the best kindergarten model was the random forest model with the TCRS (AUC = 0.676, Table 7, Panel E, Column 3), first grade was the linear probability model with the WJAP (AUC = 0.675, Table 7, Panel A, Column 4), and second grade was the elastic net model with the TCRS (AUC = 0.684, Table 7, Panel C, Column 5). All the special education and chronically absent models fit with one additional fall researcher-collected assessment performed either the same or, in most cases, better than the best performing model fit with only fall of kindergarten district-collected data (see extension analysis below).

For the models fit with an additional assessment collected both in the fall and spring of kindergarten, all the models performed best with the addition of either the TCRS or SSIS. The best performing model predicting receipt of special education services in both the first and second grades using data from both the fall and spring of kindergarten was the elastic net model with the SSIS (AUC = 0.779, Table 8, Panel C, Column 1 for first grade; AUC = 0.851, Table 8, Panel C, Column 2 for second grade). When predicting chronic absenteeism using data from the fall and spring of kindergarten, the best kindergarten model was the linear probability model with the SSIS (AUC = 0.683, Table 8, Panel A, Column 3), first grade was the elastic net model with the WJAP (AUC = 0.672, Table 8, Panel C, Column 4), and second grade was the elastic net model with the TCRS (AUC = 0.670, Table 8, Panel C, Column 5). All the special education and chronically absent models fit with one additional fall and spring researcher-collected assessment performed either the same or, in most cases, better than the best performing model fit with fall and spring of kindergarten district-collected data.

The models fit with only the district-collected data plus one extra researcher-collected data point performed similarly to the models fit with all the researcher-collected academic and executive functioning assessments as seen in the first and second research questions. Although the models predicting receipt of special education services improved their performance by adding the fall and spring of kindergarten additional assessment measures, the models predicting chronic absenteeism that used the fall and spring of kindergarten additional assessment performed approximately the same as the models using only the fall of kindergarten additional assessment. Overall, these results indicate that BPS will have more accurate predictive models by investing in one additional assessment, mainly either the TCRS or SSIS.

Robustness Checks

We conducted sensitivity analyses to address five different threats to the model performance and to determine the robustness of our findings. We considered an alternative way to address missing data using multiple imputation, removed students who received special education services during kindergarten, addressed model performance using a different definition of chronic absenteeism that relies on students missing at least 15 school days, refit our models with separate conceptual blocks of predictor variables, and removed the student demographic variables. A full description of these in included in Appendix A. Overall, we found our results to be generally robust to these five threats. While there was some variation in model performance, particularly for the models predicting chronic absenteeism using multiple imputation compared to those fit using conditional mean imputation, the results were not consistent enough to indicate a strong reason to deviate from our primary analytic approach.

Extension Analyses

We extended our main analysis in four ways: subgroup analysis, models fit with only district-collected data, Youden statistic, and confusion matrices (results are in Appendix B). This extension analysis provides a more tangible way of conceptualizing the results by demonstrating how the models presented in the main analysis would be used in an educational setting where school personnel have to set a cut point in the likelihood of dropping out of high school in order to identify students for intervention (Bruch et al., 2020).

Subgroup Analysis

First, to address the question of consequential validity of using predicting models that do not work equally well for important subgroups of students (Brussow, 2018; Gebru, 2021; Kantayya, 2020), we determined the AUC values for our main models when restricting our sample to students identified as Dual Language Learners at baseline and again for students not identified as Dual Language Learners (DLL). In the limitations section, we discuss how we would ideally perform subgroup analysis for other student subgroups but are limited by our sample size.

Overall, the special education models for the DLL students (Appendix C, Table 2) performed worse than the models for the non-DLL students (Appendix C, Table 3), particularly for first grade special education status. We saw a similar trend in the chronically absent models. For all grades, the DLL models (Appendix C, Table 4) again performed worse than the non-DLL models (Appendix C, Table 5). While all of the chronically absent models for the full sample had poor fit, a few of the non-DLL elastic net models for chronic absenteeism in second grade had an acceptable fit. Even though 51% of the sample was identified DLL and the DLL students had approximately similar rates of being chronically absent, these results indicate that the models do not perform equally well for both DLL and non-DLL students. This means that school personnel should use caution when applying these predictive models to DLL students.

District-Collected Data Only

Second, we identified a set of models where all the data was collected by district personnel, i.e., excluding the researcher-collected data that we had access to in this study that the district normally does not have access to – the extra academic assessments, intra- and interpersonal assessments, family data, and teacher characteristics. These models use the student characteristics from the administrative data as seen in Block 1 of the main models plus the assessments (DIBELS) administered by the district teachers as part of their normal assessment activities as the predictor variables. We present two versions of these models. The first one uses the administrative data plus the fall of kindergarten DIBELS assessments. The second version uses all the predictors from the first iteration plus the spring of kindergarten DIBELS assessments. To further understand model performance, we included the distribution of the AUC values across the five model iterations for these two models fit with district-collected data along with the first model from the primary analysis that only used student demographic variables. The distribution of AUC values for all remaining models are available upon request.

Compared to model 1 in the main analysis results that uses only student demographics where none of the models had an acceptable performance, almost every special education model with fall and spring DIBELS information performed better when predicting first grade special education. Particularly, the logistic and random forest models had an acceptable performance. Additionally, every model had either an acceptable or excellent performance when predicting second grade special education with either fall only or fall and spring DIBELS scores (Appendix C, Table 6, Columns 1 and 2). Just like the main models, none of the chronic absenteeism models fit with only the DIBELS had an acceptable performance (Appendix C, Table 6, Columns 3-5). These results indicate the models fit only with the district-collected data perform well only in certain circumstances.

When considering the distribution of AUC values across model iterations, the overall pattern is that the distribution widens as the grade level increases. Within each model type, the distribution of the AUC values from the models predicting first-grade special education status (Appendix B, Figure 1) was narrower than that of the second-grade special education models (Appendix B, Figure 2). Similarly, the distributions of the AUC values within model type from the models predicting chronic absenteeism in kindergarten (Appendix B, Figure 3) and first grade (Appendix B, Figure 4) were narrower than that of second grade (Appendix B, Figure 5). This indicates that the amount of sampling variability arising from the partitioning of the full dataset into the testing and training datasets increased as the models predicted higher grades.

Youden Statistic

Third, following the example of Bruch et al. (2020), we then extended our analysis of performance metrics by finding the optimal point on the receiver operating characteristic curves that optimizes the balance between the sensitivity and specificity. This point on the curve is known as the Youden statistic (also known as Youden's J statistic) and is defined as $J = max_t\{sensitivity(t) + specificity(t) - 1\}$ (Berrar, 2019; Youden, 1950). Sensitivity is true positive rate while specificity is the true negative rate (cite paper 1). While the Youden statistic is the empirically-defined optimal risk score in balancing true positive and negatives, we use it as an example of how schools can choose a risk score that reflects their preferences in which students to prioritize due to resource constraints and policy directives (Bruch et al., 2020). We present the Youden statistic for the models that only use district-collected data available to district personnel under normal circumstances (i.e., the models with only the student characteristics from administrative data, the models fit with the administrative data plus the fall of kindergarten DIBELS assessments, and the models with the administrative data and fall and spring of kindergarten DIBELS assessments). Youden statistics for all remaining models are available upon request. Additionally, we included the distribution of the Youden statistic values across the five model iterations for these two models fit with district-collected data along with the first model from the primary analysis that only used student demographic variables. The distributions of Youden statistic values for all remaining models are available upon request.

Overall, for the best performing special education models (i.e., elastic net and random forest), the Youden statistic increased as more predictor variables were added. For example, the elastic net model using only student demographics predicting second grade special education status had a cut point of 0.062, 0.098 with the fall of kindergarten DIBELS, and 0.127 with the fall and spring of kindergarten DIBELS (Appendix C, Table 7, Column 2, Panel C). This means that the empirically defined optimal balance point between the true positive and negatives for the elastic net models ranged from a 6.2% to 12.7% likelihood of being identified as needing special education services based on the predictor variables used. Compared to the special education models, the models predicting chronic absenteeism generally tended to have smaller Youden statistics as the grade increased across model type. For example, based on the predictor variables used, the Youden statistic for the elastic net models ranged from 0.148 to 0.166 for kindergarten, 0.130 to 0.136 for first grade, and 0.085 to 0.093 for second grade (Appendix C, Table 7, Columns 3-5, Panel C). This means that if the district followed the empirically based risk score for elastic net models to identify students likely to be chronically absent, they will need to use between 8.5% and 16.6% as their cut point, depending on their exact model used.

Compared to the AUC value distributions, there is no clear pattern in the distribution of Youden statistic values across the five model iterations other than the distributions for the machine learning models predicting chronic absenteeism across all grades tend to be slightly wider than the traditional regression models predicting chronic absenteeism (Appendix B, Figures 6-10). Overall, this indicates that there is variability in the Youden statistic based on the random sampling done to obtain the testing and training sets. As we are presenting the Youden statistic values as an example of an empirically defined cut score that schools could use to identify students, we also want to stress that schools should investigate alternative cut points for their specific context.

Confusion Matrices

Fourth, we present a confusion matrix that is derived by dichotomizing the dataset based on the Youden statistic for one sampling dataset. This extends the analysis done with the Youden statistic to provide an example for how schools can use a cutoff point in the predicted likelihood of receiving special education services or being chronically absent to determine the predicted true positive, false positive, true negative, and false negative rates (Berrar, 2019). We did this to examine model performance in a more tangible way than the more fine-grained AUC value provides. We present confusion matrices for the models that only use the district-collected data (i.e., the models with only the student characteristics from administrative data, the models fit with the administrative data plus the fall of kindergarten DIBELS assessments, and the models with the administrative data and fall and spring of kindergarten DIBELS assessments). Confusion matrices for all remaining models are available upon request.

Even though the confusion matrices were created using the Youden statistic as a cut point that should optimize both the true positive and negative rates, there is great variability in these rates across models. For example, the decision tree model predicting first grade special education status using only student demographic data showed a very high specificity (99.5%) yet a low sensitivity (10%), meaning that while it was very good at correctly identifying students who were not receiving special education services, the model did poorly at correctly identifying students who did receive special education services (Appendix C, Table 8, Model 1, Panel D). Conversely, the logistic model predicting second grade chronic absenteeism also using only student demographic data showed both moderately high specificity (72.3%) and sensitivity

(76.9%), implying the model did well both in correctly identifying students who were and were not chronically absent (Appendix C, Table 12, Model 1, Panel B). Similar trends were seen across the models predicting special education in second grade (Appendix C, Table 9) and chronic absenteeism in kindergarten (Appendix C, Table 10) and first grade (Appendix C, Table 11). Overall, these results indicate that there is a wide variation in how well the models correctly identify students who do and do not receive special education services and are chronically absent; school administrators must consider which group they would like to prioritize when selecting a predictive model to use. Additionally, these results were obtained when using the Youden statistic as the cut point and may change based on the use of a different cut point based on district priorities and resources.

Discussion

We sought to understand the ability of machine learning versus traditional regression models to predict two outcomes in early elementary education that have important equity implications: receipt of special education services and chronic absenteeism. Students from families with low incomes, students of color, and English language learners suffer disproportionately from being misidentified for special education services and being chronically absent (Chang & Davis, 2015; Chang & Romero, 2008; Elder et al., 2021; Morgan et al., 2015). Not receiving special education services when needed and being chronically absent are both associated with lower academic achievement (Balfanz & Byrnes, 2012; Chang & Romero, 2008; Diamond et al., 2013; Guralnick, 1998; Hanushek et al., 2002; Morrissey et al., 2014; Snowling, 2013), future school attendance (Ansari & Gottfried, 2018), behavior (Hurwitz et al., 2021), and completing high school and enrolling in postsecondary education (Ballis & Heath, 2019). Thus, it is important to identify students who need special education services and who are likely to be chronically absent as early as possible in order to mediate with proven interventions (Kalil et al., 2019; Kirksey & Gottfried, 2021; Rafferty et al., 2003; Robinson et al., 2018; Sullivan & Field, 2013; Tran & Gershenson, 2021).

While there are a few studies showing the promising value of machine learning algorithms in predicting need for special education services (Bone et al., 2016; Thabtah & Peebles, 2020), particularly for students on the autism spectrum, and chronic absenteeism (Bruch et al., 2020) in early elementary students, this is the first study to both explore how machine learning algorithms compare to traditional regression models as well as how all of these models perform with different types of predictor variables. In addition to administrative data, we had access to rich data on student academic achievement and socioemotional skills and family and teacher characteristics that extends previous studies and provided the opportunity to examine model performance using data that is not usually available to school administrators. Analyses of these data point to measures not typically collected that may be worth consideration for districts interested in early warning systems.

Special Education Results

When looking at our models predicting receipt of special education services, we found that none of our models had an acceptable performance when using student demographics alone. Adding the fall of kindergarten academic achievement and executive functioning measures increased the elastic net and random forest models to an acceptable performance level for first grade and excellent for second grade. Although adding the additional conceptual blocks of predictor variables did not consistently improve model performance, every elastic net and random forest model with additional predictor variables had either an acceptable or excellent performance. Additionally, the linear probability models predicting second grade special education status also had an acceptable performance. This indicates that it is possible for schools to construct well-performing models that predict special education receipt in first and second grade for students who did receive services in kindergarten using certain model types and predictor variables. We discuss further nuanced implications of this below.

Although we cannot directly compare our main results to the previous studies predicting special education status because we compared model performance using AUC values instead of true and false positive rates, the confusion matrices that we presented in our additional extended analysis allow us a better opportunity to put our results in context. Compared to previous studies that reported a 85-90% correct identification rate for adolescents (Thabtah & Peebles, 2020) and 89.2% for participants under the age of ten (Bone et al., 2016) when predicting the presence of autism spectrum disorder, our models predicting receipt of special education services had true positive rates ranging from 8.3% (decision tree for second grade) to 100% (linear probability for second grade) with an average of 54.3% across grades when using only child demographic information. This average increases to 56.1% when including the DIBELS fall of kindergarten district-collected data and 60.2% when including both the fall and spring DIBELS information. Although our results look less promising than those previously found in the literature, we also had faced issues due to sample size, as discussed in limitations.

Chronic Absenteeism Results

We found that when predicting chronic absenteeism, none of our models had an acceptable performance. Unlike the special education models, the chronic absenteeism models did not improve in performance with the addition of the fall of kindergarten academic achievement and executive functioning measures nor any subsequent sets of predictor variables. Even though none of the models performed well when predicting the likelihood of being chronically absent, the models with the best performance were the linear probability, elastic net, and random forest models. Of these, there was no clear pattern for which grade the models performed best.

To put our results in context with the prior literature, the one study predicting chronic absenteeism for elementary school also used AUC values to compare model performance (Bruch et al., 2020). The authors found AUC values of between 0.75 and 0.77 depending on model type. None of our models predicting chronic absenteeism had AUC values as strong as those in Bruch et al. (2020). This study also presented true positive rates of 74% and 75% for its elementary chronic absenteeism models that is created using a Youden statistic. Comparatively, we had an average true positive rate of 65.1% when predicting kindergarten chronic absenteeism using child demographic data (64.2% when including fall of kindergarten DIBELS and 62.4% when including fall and spring of kindergarten DIBELS), 62.9% when predicting first-grade chronic absenteeism using child demographic data (57.2% when including fall of kindergarten DIBELS and 52.8% when including fall and spring of kindergarten DIBELS), and 87.8% when predicting first-grade chronic absenteeism using child demographic data (86.5% when including fall of kindergarten DIBELS and 72.5% when including fall and spring of kindergarten DIBELS). This indicates that our models predicting chronic absenteeism did not perform as well as those in Bruch et al., 2020. Similarly, a follow up study to Bruch et al. (2020) found that models predicting chronic absenteeism using machine learning performed similarly to models based on a simple set of indicators (Cattell & Bruch, 2021). Although this methodological approach is not exactly the same as ours, our findings were in line with theirs since we also found that the machine learning models performed similarly to our traditional regression models.

Policy Implications

Because the school district had access to student characteristics from the administrative data at baseline and administered the DIBELS literacy and language assessments in the fall and spring of kindergarten to every student in the district, we consider the first model from the main results plus the models conducted using the child demographics plus the DIBELS in Appendix B to be the most policy relevant in terms of limited resources if the district wanted to construct predictive models without collecting any more data. Overall, the model using the child demographics plus fall of kindergarten DIBELS data performed better than the model with only child demographics, and the model with child demographics and fall and spring of kindergarten DIBELS data performed better than the model with child demographics and only fall of kindergarten DIBELS data. Yet all three of these models generally performed comparably or slightly worse than the main text models with child demographic data plus the full range of academic assessments in the fall of kindergarten. Additionally, while almost all of the models with child demographics and fall and spring of kindergarten DIBELS data had an acceptable performance when predicting receipt of special education services in both grades, none of the models predicting chronic absenteeism had an acceptable model performance. We interpret this to mean that the district may fit acceptable models predicting both special education using district-collected data, but they would be restricted in the model type they use (i.e., any type for predicting special education service receipt but only linear probability, logistic, and elastic net for chronic absenteeism). They would also forgo better fitting models by not collecting more data, particularly either the SSIS or TCRS, which are quick teacher reports compared to more time-intensive direct assessments.

Similarly, it is important for school administrators to consider the trade-offs of forgoing an early warning system built on a traditional regression technique versus one built on a machine learning algorithm. Although machine learning algorithms show promise for accurate predictions due to their statistically flexible nature (Ara et al., 2015; Bone et al., 2016; Bruch et al., 2020; Chung & Lee, 2019; Duda et al., 2016; Kotsiantis et al., 2003; Lakkaraju et al., 2015; Márquez-Vera et al., 2016; Orooji & Chen, 2019; Pagani et al., 2008; Porter, 2019; Sansone, 2019; Thabtah & Peebles, 2020), they are also inherently less transparent than a linear or logistic model. It may be more difficult to get buy-in from stakeholders who would work with the students these prediction models identify because the predictions come from a system that is often viewed as a black box (cite paper 1). Therefore, we suggest that school officials consider how much better prediction models built on machine learning algorithms need to work compared to traditional regression models in order to justify their use.

Another important aspect for school administrators to consider when deciding to construct predictive models is the timeline to accessing data. As researchers, we are privileged to use data that may take multiple months to collect, validate, and clean, such as the fall of kindergarten academic assessments. It took the research staff the entire fall semester (from September 22 to December 18) to collect this data, and it is very possible that it would take longer to conduct these assessments on every student rather than those in the study sample in addition to using more resources given how relatively labor-intensive they are. Meanwhile, teachers conducted the DIBELS assessments on their students in 1.5 months (from September 14 to October 31). Although the special education models using these assessments generally had acceptable model performance, schools may be able to get access to the DIBELS data faster than other academic achievement assessments and thus able to identify students sooner. This is particularly important when considering that the research shows that successful interventions to reduce chronic absenteeism include targeting parental beliefs (Kalil et al., 2019; Robinson et al.,

2018) and smaller classes, race match between students and teachers (Tran & Gershenson, 2021), and serving breakfast during school hours (Kirksey & Gottfried, 2021), all of which take time to implement. When building these models in real time, administrators should consider how quickly they can access data in order to intervene in a timely and effective manner.

Other policy considerations for administrators to keep in mind are the consequences of true and false positive and negative rates because there are differential impacts for these groups of students (Brussow, 2018). Administrators must pick a threshold in the likelihood of students receiving special education services and being chronically absent, and that threshold can vary based on resources and the desire to optimize the type of students identified. In this context for the models predicting special education receipt, the false negatives are students who the model does not predict to need special education services yet actually receive them while the false positives are students who the model predicts to need services yet actually do not receive them. While it is ideal to minimize the numbers of students in both groups, we argue that there are greater implications for the false negative students because those are students who would not receive important services that they need (Ansari & Gottfried, 2018; Ballis & Heath, 2019; Diamond et al., 2013; Guralnick, 1998; Hanushek et al., 2002; Hurwitz et al., 2021; Snowling, 2013; Ullery & Katz, 2016) compared to the false positive students who would go through extra screening only to later be determined that they do not need special education services (National Center on Response to Intervention, 2010). Even though the confusion matrices we present in Appendix B are based on the Youden statistic that is designed to empirically balance the true positive and negative rates (Berrar, 2019; Youden, 1950), our analysis showed that certain models vastly favored one of the other. Therefore, administrators should be cautious of this and not assume an equality of rates even when using the data-based cut point.

Similarly, it is important to consider what a context-specific threshold would be for a model to qualify as having an acceptable performance. Although we based our decision of an AUC value of 0.7 as this threshold following the statistical literature (Mandrekar, 2010) and previous work done in this field (Bruch et al., 2020), we recognize that the application of the AUC in the field of education early warning systems is novel enough to not have a large enough precedence to have definitively established a threshold agreed upon throughout both the research and practitioner communities. Furthermore, this threshold may vary based on context, including geography, school and district size, and outcome studied. Thus, administrators should consider what an appropriate cutoff would be for their particular context when deciding which statistical model to use for an early warning system.

With these implications in mind, and considering the limitations discussed below, we stress that this work is preliminary. Given the scant amount of existing work of this nature that we are building on, this exploratory research indicates that schools may be well served in investing resources into more research on the feasibility and performance of these predictor models under practical circumstances. While we have made every attempt to replicate what we believe to be operating school conditions, school personnel would be better equipped to make decisions such as how to simulate the types of data readily available to them, which types of performance metrics to focus on, and how logistically reasonable an early warning system would be given staffing capabilities. Our results indicate that such an early warning system based on predictor models shows promise, particularly for special education in first and second grades, yet we present our results as a first step towards putting them into practice.

Limitations

We note several important limitations to our work that should be taken into consideration when interpreting our results. The first is that our sample size is smaller compared to other similar studies. The other studies examining special education status had larger sample sizes (*N* = 1726; Bone et al., 2016) and/or larger percentages of their samples identified as special education than our study did (26.8%, Thabtah & Peebles, 2020; 73.2%, Bone et al., 2016). Similarly, the study that used machine learning to predict chronic absenteeism had a much larger sample size than our study (*N* = 28, 719 for one model and *N* = 4,614 for another model; Bruch et al., 2020). Because machine learning models were designed to capitalize on large data sets, it is possible that our three machine learning models would have performed better with a larger sample size (cite paper 1). However, given that the average school district in the United States enrolls approximately 3,583 K-12 students, our sample size is not unrealistic for the average number of elementary students in a school district (U.S. Department of Education National Center for Education Statisics, 2019). Similarly, predictive models may be most helpful in execution when using data from previous cohorts of students to predict the performance of future cohorts of students. The fact that our sample consisted of only one cohort indicates that future research ideally should be done using not only larger sample sizes but students across multiple cohorts.

Our sample size also means that we were not able to differentitate between excused and unexecused absences in our chronic absenteeism models. Although we followed the literature on the most common approach of defining chronic absenteeism in our main text (missing 10% or more of school days while being enrolled for at least 90 days) and then another common definition on our robustness check (missing at least 15 school days without a minimum number of days enrolled) that count combined both excused and unexcused absences together (Chang & Romero, 2008), research also indicates an important difference in out-of-school factors that contribute to long-term effects between students who have a higher number of excused absences versus unexcused absences (Pyne et al., 2021). Ideally, a model predicting chronic absenteeism for a school would be able to distinguish between these two types of absences.

Likewise, our relatively low sample size meant that we were not able to look at model performance based on two subgroups that we know from research are important for both special education and chronic absenteeism identification: students of color and students from families with low incomes (Chang & Davis, 2015; Chang & Romero, 2008; Elder et al., 2021; Morgan et al., 2015). Although we were able to have a large enough sample size to look at model performance based on Dual Language Learner (DLL) status, we were only able to do this because 51% of our sample identified as DLL; even for those models, it is very possible that they would have performed better with a larger sample size. Similarly, we had to aggregate all of the special education students in our sample together for our analysis despite knowing that there are important differences between types of special education identification and that they therefore require different types of intervention (Rafferty et al., 2003; Snowling, 2013; Sullivan & Field, 2013). For example, one study that used machine learning techniques to differentiate between autism spectrum disorder and attention deficit hyperactivity disorder had a sample size of N = 2925 that was solely comprised of two diagnosis options (Duda et al., 2016).

Another limitation relates to the generalizability of our study. As we noted in our Participants and Setting section, our study sample is not perfectly representative of the larger BPS school district, particularly with respect to our two outcomes. Our sample has a lower percentage of students identified as receiving special education services in kindergarten, first grade, and second grade; our sample also had lower rates of chronic absenteeism in all three grades. Because these lower rates could be related to lower model performance, it is possible that our models could have performed differently if our sample had larger percentages of special education and chronically absent students that were more representative of the district.

When interpreting our results, it is important to keep in mind two additional cautions. The first is that our special education models were predicting receipt of special education services, not the inherent need for such services. Our data only shows which students have been identified by the district as needing services. Although the district employs a universal screener that hopefully mitigates the inequity that is more common with a referral system (National Center on Response to Intervention, 2010; Thabtah & Peebles, 2020), it is possible that there is existing bias in the data that we are replicating with our models due to structural and implicit bias affecting students with color, students from families with low incomes, and students from non-majority cultural and linguistic backgrounds (Elder et al., 2021; Morgan et al., 2015; Voulgarides et al., 2017).

The second additional caution is that our paper should not be interpreted as an ultimate referendum on the value of traditional regression versus machine learning models when predicting receipt of special education and chronic absenteeism in early elementary grades. As discussed in (cite paper 1), it is common to use multiple supervised machine learning algorithms when conducting predictive analytics. We chose three such algorithms based on their previous use in the literature and their methodological relevance to this study (Bruch et al., 2020; Chung & Lee, 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016), but there are several more predictive machine learning algorithms that exist that we could have chosen (Athey, 2019; Hastie et al., 2009; James et al., 2013). Additionally, we could have adjusted our existing models using their internal hyperparameters that could have resulted in a different model performance (Hastie et al., 2009; James et al., 2013). Given the possibility for a different outcome based on these decisions, we want to clarify that our results should be interpreted only in the context of the analysis we conducted.

Conclusion

Despite these limitations, our findings offer insight into how traditional regression models (linear probability and logistic) compared to machine learning models (elastic net, decision tree, and random forest) when predicting two important outcomes for early elementary students with important equity implications: receipt of special education services and chronic absenteeism. We build on the existing literature that shows that it is possible to construct predictive machine learning models with acceptable performance for those two outcomes in this age group (Bone et al., 2016; Bruch et al., 2020; Duda et al., 2016; Thabtah & Peebles, 2020) and extend it by comparing model performance to traditional regression methods. We also examine under what circumstances – i.e., choice of predictor variables – model performance changes given our access to rich administrative, assessment, parent, and teacher data. Our work demonstrates that there are ways to approach building early warning systems that incorporate both traditional and new methods that can achieve the goal of predicting students as early as possible for intervention.

Paper 3: The Role of Machine Learning in Early Warning Systems for Predicting High School Dropout in Michigan

Dropping out of high school is associated with worse financial, familial, and societal outcomes later in life (Alliance for Excellent Education, 2011; Bridgeland et al., 2006). Accordingly, a great deal of attention and resources have been directed to creating effective interventions to reduce dropout (Dynarski et al., 2008). One such intervention is the use of early warning systems that alert school personnel to which students are at high risk of dropping. These systems are designed to more effectively target intervention resources by identifying which students do and do not need intervention (O'Cummings & Therriault, 2015; U.S. Department of Education, 2016).

To be useful and efficient given limited resources, the predictive models underlying these systems need to be as accurate as possible (Engler, 2020; U.S. Department of Education, 2016). Recent research has shown that machine learning offers a promising avenue for constructing highly predictive models when predicting students likely to drop out of high school (Ara et al., 2015; Chung & Lee, 2019; Coleman et al., 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016; Orooji & Chen, 2019; Pagani et al., 2008; Sansone, 2019). However, machine learning techniques are somewhat opaque and complex for localities to operate, particularly compared to traditional regression models. None of the studies to date examining the accuracy of machine learning techniques in predicting dropout have compared the performance of newer ML models to traditional regression models. Prior studies also have not explicitly explored model performance for student subgroups that have differential dropout rates. This is an important consideration given prior work in other fields have showed that predictive machine learning models do not necessarily work equally well for subgroups, leading to bias in the predictions (Garvie et al., 2016; Hill, 2020; Obermeyer et al., 2019).

Using administrative data from the entire state of Michigan for five cohorts of students (*N* = 416,105) expected to graduate high school between academic years 2012-2013 and 2016-2017, I seek to help address these two critical gaps in the literature. I compare traditional regression models to machine learning models, exploring how early in a student's career the models can accurately predict their likelihood of dropping out, and how the models perform for important student subgroups. Leveraging Michigan's longitudinal database, I was able to collect data on students spanning back to fourth grade. Using information on student demographics, attendance, behavior, and academic performance, I fit two types of traditional regression models (linear probability and logistic) and three types of machine learning models (elastic net, decision tree, and random forest) predicting the likelihood of dropping out of high school using multiple conceptual blocks of predictors beginning with data in fourth grade. I used the area under the receiver operating characteristic curve (AUC) to evaluate model performance (Mandrekar, 2010). Then I example model performance for four types of student subgroups with differential dropout rates: gender, race and ethnicity, socioeconomic, and receipt of special education services.

Overall, I found that the random forest models were most accurate, yet the linear probability, logistic, and elastic net models performed almost as well, to the point where it was not convincing that the machine learning models performed better than the traditional regression models. I also found that the models fit with fourth and fifth grade data had a strong model performance and that this performance increased with each additional grade's worth of data. On average across the four best performing model types, the models had approximately a 71.5% chance of correctly identifying students who dropped out of high school using data from fourth and fifth grade. This increased by an average of between 1-2% for each additional grade of data added to the model. When analyzing model performance across subgroups, I found that while it was possible to construct well performing models for each subgroup, it should not be assumed that models fit for the full sample work equally well for every subgroup. I then discuss implications for policy and limitations to consider when interpreting the results on how to build predictive models for early warning systems to identify students likely to drop out of high school.

Review of Literature

Importance of Identification for Intervention

It is crucial to accurately identify students who are at risk of dropping out of high school because students who do not finish high school are more likely to experience worse financial, familial, and social outcomes later in life. For example, students who do not finish high school earn on average 71% of the earnings of their peers who hold only a high school diploma, 54% for an Associate's degree, and 42% for a Bachelor's degree (Alliance for Excellent Education, 2011). Students who drop out of high school are also more likely to be unemployed, receive social assistance benefits, live in poverty, commit crimes, take drugs, be in prison, be unhealthy, have a shorter life span, vote, volunteer, be a teenage parent, be either divorced or a single parent, have unhealthy children, and raise a child who does not graduate from high school (Alliance for Excellent Education, 2011; Bridgeland et al., 2006). Given these negative impacts of dropping out of high school, early warning systems can help schools use their resources more efficiently to better identify students with a high likelihood of dropping out of high school.

Although there are multiple ways of defining high school graduation and dropout rate, it is clear across definitions that not every student is equally likely to be at risk. Nationally, the adjusted cohort graduation rate, which considers students who transfer in and out of the school, indicates that 85% of high school students graduate with a traditional diploma within four years of entering high school. This average masks important variation across race and ethnic lines; although 92% of Asian and Pacific Islander and 89% of White students graduate within four years, only 81% of Hispanic, 79% of Black, and 74% of American Indian or Alaska Native students do (Hussar et al., 2020). Similarly, another measure is the status dropout rate metric, which is calculated as the percentage of 16 to 24-year-olds who are not enrolled in school and have not earned either a traditional high school diploma or a GED (General Educational Development) equivalent. Nationally, the average status dropout rate is 5.3%, yet the Pacific Islander rate is 8.1%, Hispanic 8.0%, Black 6.4%, multiracial 5.2%, White 4.2%, and Asian 1.9% (Hussar et al., 2020). Additionally, in almost every racial/ethnic category, males are more likely to drop out then females, students in institutionalized settings (such as adult and juvenile jails and prison and health care facilities) are more likely to drop out than those who are not, and foreign-born students are more likely drop out compared to U.S.-born students. Furthermore, students with a disability are 2.3 times more likely to drop out than students without a disability. All of these statistics indicate important equity concerns when studying which students are most are risk of dropping out of high school (Hussar et al., 2020).

Fortunately, there are evidence-based interventions that have been shown to be effective in mitigating high school dropout, including proper student identification, student-level, and school-level interventions. Researchers recommend that schools track key student data longitudinally to better identify which students are most at risk of dropping out, often via an early warning system as discussed more below. Effective student-level interventions include assigning students an individual adult advocate who are appropriately trained with low enough caseloads such that they can meet regularly with students; providing students with extra academic support such as individual or small group tutoring, extra study time, and credit recovery; and implementing support programs to improve students' classroom behavior, often coordinating with social services and mental health programs as needed. At the school-wide level, interventions that lead to a higher quality of learning for all students have been shown to help reduce dropout. These include fostering a personalized learning environment via small classes or learning communities, team teaching, and student participation in extracurricular activities that promote a sense of belonging among students as well as providing high-quality, engaging instruction that prepares students for life after high school, such as professional development for teachers, integrating academic and career-ready content, and work internships (Dynarski et al., 2008). Because every intervention is not appropriate for every student, schools show discretion in which interventions to use for their students, such as higher-poverty schools using adult advocates and credit recovery more often than lower-poverty schools (U.S. Department of Education, 2016). As schools are often limited in the resources they can allocate to dropout prevention, it is important to be as precise as possible when identifying which students need intervention (Engler, 2020). Indeed, to help control intervention costs, research recommends using an early warning system to identify students who are likely to drop out of high school (Dynarski et al., 2008)

Current Process of Identification

Over half of U.S. states[24] use an early warning system for school improvement, including to flag students who are at risk of dropping out of high school (Data Quality Campaign, 2013; O'Cummings & Therriault, 2015; U.S. Department of Education, 2016). Part of an initiative to reduce dropout, early warning systems are used to identify students to determine an appropriate intervention based on the circumstances indicating why the individual students are deemed at-risk. In some districts, early warning systems also use data to detect students who are at risk of not being reading proficient by the end of third grade and who are not ready for the transition from elementary to middle school and high school to college (Faria et al., 2017; O'Cummings & Therriault, 2015).

Early warning systems offer a data-driven approach to identifying at-risk students as opposed to solely relying on educators' intuition about which students should receive interventions. While some districts use their own historical student data to create a tailored set of indicators for flagging students, other districts rely on the established literature about the factors commonly associated with dropping out of high school (Faria et al., 2017; O'Cummings & Therriault, 2015; Therriault et al., 2017). For example, previous research has shown that students are more likely to drop out of high school if they move frequently (Rumberger & Larson, 1998), have poor academic achievement (Battin-Pearson et al., 2000; Parr & Bonitz, 2015), are frequently absent, have low parental involvement (Parr & Bonitz, 2015), engage in deviant behavior, bond with antisocial peers, and are from families with low-incomes (Battin-Pearson et al., 2000). This research has been distilled into three main categories of commonly-used indicators for early warning systems: attendance, behavior, and course performance (Therriault et al., 2017; U.S. Department of Education, 2016). Although the exact indicators used for early

[24] For a full list of states, see Data Quality Campaign (2013).

warning systems are typically considered proprietary information, known flags include whether a student missed 20% of instructional time in middle school or 10% in high school, failed either an English or math course in middle school or any course in high school, had a GPA of 2.0 or lower in high school, and lacked enough course credits to advance to the next grade in high school (O'Cummings & Therriault, 2015; Therriault et al., 2017). Additionally, high-poverty schools and schools with particularly low graduation rates tend to also use data from out-of-school factors, such as homelessness status and experience with the juvenile justice system (U.S. Department of Education, 2016). Regardless of the precise indicators that schools use, early warning systems rely on data to identify students who are likely to drop out of high school before they do so to keep them in school.

There is variation in whether early warning systems are administered and monitored at the district or school level. Although larger schools are more likely to use an early warning system than smaller schools are, there is no statistically significant difference in a school's likelihood of having a system in place based on graduation rate, poverty level, or location. Within a school, researchers found that school administrators and guidance counselors are the most common personnel who monitor the early warning systems, followed by student support teams, teachers, case managers, district administrators, and mentors. With respect to how often educators monitor the early warning system, 44% of schools reported checking the data weekly, citing capacity constraints for not being able to check it more often. Research has also shown that school educators and administrators generally like early warning systems because they find them accessible, easy to interpret, and accurate (U.S. Department of Education, 2016). Also, in large schools where teachers are siloed in departments, early warning systems alert teachers to the fact that students struggling in their class may also be do poorly in other classes, broadening the focus from a single course to the whole student (O'Cummings & Therriault, 2015). The widespread adoption and benefits of early warning systems indicate that they are expected to continue to be used in the future.

Machine Learning for Identification

Traditionally, the research investigating the factors associated with dropping out of high school that were used to create the indicators in early warning systems was conducted using traditional regression (Rumberger & Larson, 1998) and structural equation modeling (Battin-Pearson et al., 2000; Parr & Bonitz, 2015). However, recent research has shown that a relatively new technique – machine learning – holds promise for predicting which students are at-risk of dropping out of high school very accurately, i.e., potentially over 90% correctly predicted (Ara et al., 2015; Chung & Lee, 2019; Coleman et al., 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016; Orooji & Chen, 2019; Pagani et al., 2008; Sansone, 2019).

Among this research, there is a wide variety of the geographic areas studied and the type of data used in the models. For example, machine learning has been used to predict high school drop out in Louisiana (Orooji & Chen, 2019), large U.S. unnamed cities (Coleman et al., 2019; Lakkaraju et al., 2015), across the entire United States (Sansone, 2019), Denmark (Ara et al., 2015), South Korea (Chung & Lee, 2019), Quebec (Pagani et al., 2008), and Mexico (Márquez-Vera et al., 2016). While most of the studies used only administrative data (Ara et al., 2015; Chung & Lee, 2019; Lakkaraju et al., 2015; Orooji & Chen, 2019), two studies used administrative data plus extra assessment data (Coleman et al., 2019; Márquez-Vera et al., 2016) while two other studies used nationally-representative survey data instead (Pagani et al., 2008; Sansone, 2019).

Consistent with the use of machine learning studying other educational outcomes (cite dissertation paper 1), researchers conducted their analyses with different algorithms and performance metrics, making it difficult to directly compare the performance of the new machine learning models both to each other and to than traditional regression or binary indicator methods. However, trends indicate that machine learning is most predictive with large sample sizes (Ara et al., 2015; Chung & Lee, 2019; Coleman et al., 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016; Orooji & Chen, 2019; Pagani et al., 2008; Sansone, 2019). While the majority of these studies did not explicitly compare the performance of machine learning models to traditional regression methods, the one study that did showed that machine learning models were more accurate than logistic regression (Lakkaraju et al., 2015). This is most likely due to the flexible nature of machine learning models compared to traditional regression; without the constraints of regression models, most machine learning algorithms can fit the nuances of individual datasets well. Additionally, machine learning models are able to consider more input variables than traditional methods are because collinearity is not a concern (Hastie et al., 2009; James et al., 2013).

Although these studies demonstrate the capacity of machine learning for identifying which students are likely to drop out of high school, I build on this research by addressing three gaps. First, I address the call for an explicit examination of the potential added of machine learning over the traditional regression approach widely used already. Second, I examine model performance when using predictor variables that are measured prior to entering high school, something that only half of the existing studies did (Coleman et al., 2019; Lakkaraju et al., 2015; Orooji & Chen, 2019; Pagani et al., 2008). If it is possible to accurately predict students who are likely to drop out of high school as early as possible, then that may give school personnel more time to intervene. Third, I study model performance for student subgroups, something that no previous study did. Because the literature demonstrates that students are differentially likely to drop out based on their gender, race/ethnicity, special education, and socioeconomic status (Hussar et al., 2020) and given the consequential validity of using a tool that should be attuned to these nuances in the data (Brussow, 2018), it is important not to assume that a model fit on an entire student sample would work equally well for these important student subgroups. If this assumption is not explored, it could lead to algorithmic bias and biased predictions (Gebru, 2021; Kantayya, 2020). This study will further nuance the conversation about how to construct predictive models for early warning systems addressing high school dropout.

Current Study

This exploratory study seeks to build on prior research by examining how – and under what circumstances – machine learning compares to traditional regression methods when predicting high school dropout. I answer two research questions:

1. How does machine learning compare to traditional regression methods when identifying students who do not graduate from high school? How does this performance vary based using data from different grades as predictors?
2. Do the models work equally well for student racial, socioeconomic, gender, and special education subgroups?

Method

Participants and Setting

The sample consists of 416,105 students in five cohorts who were expected[25] to graduate high school across the entire state of Michigan between academic years 2012-2013 and 2016-2017. I used administrative data to retroactively collect data on these students from fourth through twelfth grade. Because I used demographic, attendance, behavior, and assessment data from fourth through tenth grade as predictor variables, I only included students in my sample for whom I had a value for at least one predictor variable. For example, this means that my sample includes students who transferred into Michigan beginning in tenth grade and reached their expected four-year high school graduation yet excludes students who transferred into Michigan beginning in eleventh grade and reached their expected four-year high school graduation. Of the 416,105 students for whom I had graduation data, I was able to get fourth grade data on 342,533 students (82%); 346,344 students in fifth grade (83%); 351,743 students in sixth grade (85%); 356,883 students in seventh grade (86%); 360,103 students in eighth grade (87%); 387,161 students in ninth grade (93%); and 386,608 students in tenth grade (93%) across the five cohorts.

Students were diverse with respect to gender, race and ethnicity, and socioeconomic status. 48% of my sample was female, 68% White, 20% Black, 6% Latinx, 3% Asian, 1% Native American or Alaskan Native, 2% multiracial, and less than 1% Hawaiian or Pacific Islander. 39% of my sample was classified as being economically disadvantaged, defined by the state of Michigan as a student being either classified as free/reduced price lunch, receiving SNAP or TANF benefits, migrant, homeless, or in foster care. 11% of my sample received special education services while 4% of my sample was identified as being Limited English Proficient. 11% of students in our sample dropped out of high school, and virtually 100% (99.92%) of

[25] The state of Michigan defines the expected graduation year as four years after a student's first year in ninth grade (Center for Educational Performance and Information, 2021).

students in my sample were enrolled in a public school when they were expected to graduate (Table 9).

Procedure

This study was conducted with the approval of IRB (Institutional Review Board) at the lead institution under the approval number was HUM00192810.

Administrative Data

I used administrative records spanning every student enrolled in the state of Michigan for five cohorts of students who were expected to graduate high school between 2012-2013 and 2016-2017. All the variables came from this administrative data as it was comprehensive in including graduation records, demographic variables, and information on attendance, behavior, and assessments. I restricted the study to students who were expected to graduate between the 2012-2013 and 2016-2017 school years due to the state consistently giving assessments across grades during those years.

Measures

Drop Out Status

I used administrative data to measure our outcome of interest: dropping out of high school. According to the state of Michigan, students are only classified as dropping out if they did not earn a traditional high school diploma in four years or less (71% of our sample), did not earn a traditional high school diploma in five years or more (4%), did not earn a GED (1%), did not earn another type of certificate (1%), did not graduate in four years but are still continuing in school (2%), are exempt for reasons such as reaching the maximum special education age or moving out of state (5%), or have a missing expected record (5%). Thus, although the four year graduation rate is 71%, the dropout rate is only 11% instead of 100% – 71% = 29% (Center for

Educational Performance and Information, 2021). I coded the outcome as a binary indicator where 1 = dropped out and 0 = any other outcome.

Demographics

For the student demographic variables from the administrative data, I used baseline measures. These consisted of gender, and mutually exclusive race and ethnicity indicator variables (White, Black, Latinx, Asian, Native American or Alaskan Native, Hawaiian or Pacific Islander, or multiracial), and indicators of being identified as Limited English Proficient, receiving special education services, and being economically disadvantaged (i.e., defined as an indicator for if a student received SNAP or TANF benefits, was migrant, was homeless, or was in foster care).

Attendance

Following research showing that attendance is an important variable to use in early warning systems predicting students likely to drop out of high school (Therriault et al., 2017; U.S. Department of Education, 2016), I used a continuous variable tracking attendance in fourth through tenth grades. Because schools had a different number of required school days, I used the percentage of school days each student attended in a given school year from administrative data.

Behavior

Similarly to attendance, research shows that early warning systems should include measures of student behavior (Therriault et al., 2017; U.S. Department of Education, 2016). Using administrative data, I included indicator variables that noted whether students were suspended in-school, suspended out-of-school, or expelled at least once during the school year. I also used four binary variables that indicated whether a student was involved in four types of incidents during the school year. I included whether students were involved in a violent incident, an incident involving a weapon, an incident involving substance abuse, or another type of incident. A student received a value of one for each variable if they were involved in at least incident in a given school year.

Although I had this data for fourth through tenth grade, I dropped some of the variables from the earlier grades because no students in my sample were suspended, expelled, or involved in any incidents. Specially, I excluded both in- and out-of-school suspension and expulsion variables from fourth, fifth, and sixth grade. There were no students who had incidents that were violent or involved weapons or another type of incident in fourth or fifth grade. I also dropped the substance abuse incident variable in fourth, fifth, and sixth grade.

Assessments

The third category of variables that research indicates is important to use for early warning systems addressing high school dropout is academic performance (Therriault et al., 2017; U.S. Department of Education, 2016). From the administrative data, I used values from the state-wide standardized exams (Michigan Education Assessment Program, *MEAP*) in various subjects across years. I only included subjects that were tested in every grade across the five cohorts in my sample. Therefore, I included values from the math assessment in fourth through eighth grade, reading in fourth through eighth grade, social studies in sixth and ninth grades, and science in eighth grade. All assessments were measured in scaled scores.

Analytic Approach

Missing Data

I had varying rates of missing across grades data due to both natural missingness and given the changing number of students present in each grade across all the cohorts (Table 9). I had no missingness in the dropout rate, by design, as well as no missingness in the gender and race/ethnicity variables. There was 18% missingness for economically disadvantaged, receiving special education services, and Limited English Proficient. The missingness for percent of school days attended ranged from 8% to 19% across years. There was between 7% and 14% missingness for both suspension variables and the expulsion indicator. For the indicators of being involved in a violent, weapons, or other type of incident, missingness ranged from 7% to 15% and from 7% to 14% for a substance abuse incident. Math assessment scores missingness ranged from 19% to 22%, reading 18% to 23%, social studies 13% to 20%, and science 18%.

I used conditional mean imputation, also known as regression imputation, for the variables with missingness by replacing missing values with estimates derived from regression equations fit using non-missing values (Enders, 2010; Harrell Jr., 2015). Although conditional mean imputation can lead to biased parameter estimates and dampened standard error estimates under certain circumstances that multiple imputation may resolve (Enders, 2010), I chose this approach because the focus of our models is on predictive power rather than parameter estimation. By using a simpler missing data strategy, I also hoped to create a model that would be transparent to practitioners. To retain information about which students originally had missing information in case the missingness was informative for the models, I also included a binary indicator for each variable that had missingness that denoted whether a student initially was missing a value for that variable. To account for the fact that our missingness was a combination of naturally occurring missingness and different numbers of students across grades, I included a version of my analysis conducted using only the students who were present in every grade in Appendix D.

RQ1: Machine Learning and Traditional Regression Analyses

I operationalized the outcome measure of dropping out of high school as a binary indicator for the descriptive statistics. Then, when coding the predictive models, I forced the software to recognize the binary variable as continuous from zero to one, inclusive, to act as a risk score to provide more nuanced model interpretation for practitioners and policy makers. By doing this, it would allow schools to set their own risk threshold when identifying at-risk students (Bruch et al., 2020).

To compare traditional regression to machine learning model performance, I first fit the data with a both linear probability and logistic model. Next, I fit the data with three different machine learning algorithms that were common in the machine learning education literature (Bruch et al., 2020; Chung & Lee, 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016): elastic net with a linear base, decision tree, and random forest. As well as their previous use in the literature, I also choose these algorithms to include both parametric (elastic net) and nonparametric (decision tree and random forest) options. While the elastic net model is technically a machine learning algorithm, I selected it to function as a bridge between the traditional linear probability and logistic models and the more statistically flexible decision tree and random forest models. Compared to traditional regression models where there is generally a consensus on the main type of model to use based on the nature of the outcome variable and the error structure based on how the dataset is structured, there are multiple machine learning algorithms designed to be used for the same purpose. To that end, it is becoming increasingly common to use multiple algorithms and compare predictive performance across models (cite paper 1).

Similarly, in the education literature, there is not a universally agreed upon method of dividing the dataset into the training and testing data. To simulate how such predictive models could be implemented in early warning systems where data from previous students was used to predict the performance of future students, I used the first four cohorts (2012-2013 through 2015-2016) of students as my training dataset and the most recent cohort (2016-2017) as my testing dataset. I fit all five models (linear probability, logistic, elastic net, decision tree, and random forest) using the training data to establish parameter estimates that define the models. I then used these model values to predict the outcome using the testing data. While fitting the model with the training data optimizes its flexibility, I used the model applied to the testing data to obtain the statistics used to measure model performance. It is common to use training and testing data because it acts as a check on overfitting the models (Athey, 2019; Mullainathan & Spiess, 2017).

To examine the aspect of the research question that specifies looking at how model performance varies based on year of data used, I fit all five models using conceptual blocks of data: Block 1 = fourth grade measures, Block 2 = Block 1 with all fifth grade measures, Block 3 = Block 2 with all sixth grade measures, Block 4 = Block 3 with all seventh grade measures, Block 5 = Block 4 with all eighth grade measures, Block 6 = Block 5 with all ninth grade measures, Block 7 = Block 6 with all tenth grade measures, and Block 8 = Block 7 with demographics. I chose to fit the conceptual blocks in this order to explore how early schools could set up early warning systems to make accurate predictions of students likely to drop out of high school for intervention and to align with the literature on what information is typically included in an early warning system (Therriault et al., 2017; U.S. Department of Education, 2016).

Given that the outcome was operationalized as continuous variable, I used the area under the receiver operating characteristic curve (AUC) to evaluate model performance. AUC is a commonly used measure that ranges from zero to one, with measures closer to one indicating a better performance. Generally, an AUC of 0.5 suggests no model discrimination, 0.7-0.8 is acceptable, 0.8-0.9 is excellent, and greater than 0.9 is outstanding model performance (Mandrekar, 2010). The AUC measures the area under the curve created by graphing the sensitivity (i.e., the true positive rate - the proportion of true positives out of total actual positive) versus 1 – specificity (i.e., the true negative rate - the proportion of true negative out of total actual negative) at every possible threshold from zero to one for turning the continuous likelihood into a binary classification. In context, the AUC can be interpreted as the percent chance a model has of correctly identifying students to drop out of high school (Bruch et al., 2020). Just as using a continuous outcome operationalization provides a more granular and nuanced measure, the AUC provides an analogous measure compared to the performance statistics obtained from using a binary classification (i.e., true/false positive/negative rates) and is commonly used in the literature evaluating predictive models (Ara et al., 2015; Bruch et al., 2020; Chung & Lee, 2019; Coleman et al., 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016; Sansone, 2019).

RQ2: Comparing Model Performance Across Student Subgroups

To address the important assumption that early warning systems should work equally well for student subgroups (Brussow, 2018; Gebru, 2021; Kantayya, 2020), particularly for subgroups that the literature shows are more likely to drop out (Hussar et al., 2020), I fit the models from the first research question restricted to the subgroup of interest. I did this for four types of student subgroups based on gender, race and ethnicity, socioeconomic status, and receipt of special education services. For gender, I fit the models first only with females and again only with males. Then, for race and ethnicity, I fit the models separately for each category: White, Black, Latinx, Asian, Native American or Alaskan Native, Hawaiian or Pacific Islander, or multiracial. Lastly, for the economically disadvantaged and special education subgroups, I fit the models separately for students who were identified as being economically disadvantaged or receiving special education services and again for the students who were not identified as such. Because these two subgroups were allowed to vary across grades compared to gender and race/ethnicity, I used the imputed baseline (i.e., fourth grade) value to determine subgroup status for this analysis.

Results

RQ1: Machine Learning and Traditional Regression Analyses

Overall, we found that almost every model type performed well beginning with the fourth and fifth grade data and with performance improving as each additional year's worth of data was added (Table 10). Specifically, the linear probability (AUC = 0.715), logistic (AUC = 0.715), elastic net (AUC = 0.715), and random forest (AUC = 0.716) models all had acceptable model performance with the second block of predictors. This means that it was possible to construct well performing models predicting high school dropout using data on attendance and academic performance from fourth and fifth grade. Unsurprisingly, these four models improved their performance monotonically as each year of data was added, and their performance went from acceptable to excellent (AUC = 0.806 for linear probability, AUC = 0.800 for logistic, AUC = 0.806 for elastic net, AUC = 0.818 for random forest) in the sixth model iteration (i.e., the model fit with fourth through ninth grade data). In context, these results imply that, on average across the four best performing model types, the models had approximately a 71.5% chance of correctly identifying students who dropped out of high school using data from only fourth and fifth grade.

This increased by an average of between 1-2% for each additional grade of data added to the model.

Although the random forest model performed slightly better than the other mode types (i.e., 0.5% to 1.5% more accurate when identifying students who dropped out), the linear probability, logistic, and elastic net models performed almost as well, to the point where my results were not convincing that machine learning models produced more accurate predictions. The slight outlier was the decision tree model that performed well once it reached the fourth model iteration (i.e., fourth through seventh grade data), but it consistently underperformed compared to the other four types.

RQ2: Comparing Model Performance Across Student Subgroups

Descriptively, there were differential graduation rates based on the four types of subgroups. Compared to the full sample dropout rate of 11%, females had a dropout rate of 8% compared to 13% for males. Based on race and ethnicity, White students saw a dropout rate of 8%, Black students 19%, Latinx students 15%, Asian students 4%, Native American or Alaskan Native students 16%, Hawaiian or Pacific Islander students 10%, and multiracial students 11%. Students who were identified as being economically disadvantaged had a 18% dropout rate compared to their peers who were not economically disadvantaged who had a dropout rate of 5%. Lastly, students who received special education services had a dropout rate of 18% compared to the 10% of students who did not receive services.

Overall, the predictive models performed differentially based on subgroups, favoring the subgroup with the lower dropout rate. In terms of gender, the model performed slightly better for females than it did for the full sample with an average AUC value of 0.07 higher (Table 11, Panel A). The model performed approximately the same for males as it did for the full sample

(Table 11, Panel B). For the model performance based on race and ethnicity, the model performed slightly better for White subgroup compared to the full sample with an average ACU of 0.006 higher (Table 12, Panel A). It also performed better for the Asian subgroup compared to the full sample models with an average AUC value of 0.04 higher (Table 12, Panel D). However, the remaining subgroups performed worse than the full sample. The model performed an average AUC value of 0.04 lower for Black students compared than the full sample (Table 12, Panel B), Latinx 0.06 lower (Table 12, Panel C), Native American/Alaskan Native 0.03 lower (Table 12, Panel E), Hawaiian/Pacific Islander between 0.06 lower (Table 12, Panel F), and multiracial 0.02 lower (Table 12, Panel G). In context, these results indicate that the model has between a 2% and 6% lower chance of correctly identifying Black, Latinx, Native American/Alaskan Native, Hawaiian/Pacific Islander, and multiracial students compared to the full sample.

For the students that were economically disadvantaged (Table 13, Panel A) or received special education services (Table 14, Panel A), the model performed approximately 7-9% worse compared to the full sample. For these subgroups, the model did not have an acceptable performance until eighth grade data was incorporated. The model performed as well as or slightly better for students who were not economically disadvantaged (Table 13, Panel B) or did not receive special education services (Table 14, Panel B) compared to the full sample.

For all subgroups, the random forest models performed slightly better than the linear probability, logistic, and elastic net models on average but not to the point where its advantages were convincing. Additionally, like the full sample, almost every model type improved its performance with each additional grade’s worth of data. These results indicate that although it is possible to construct well performing models for every subgroup of interest, it should not be assumed that the models designed for the full sample will automatically perform equally well for the subgroups.

Robustness Checks

I conducted sensitivity analyses to address four different threats to model performance and to determine the robustness of my findings. I fit models to address the possibility that results were driven by student demographics, a saturation of predictor variables, an unstable number of students across grades, and different model performance based on urbanicity. A full description of these in included in Appendix D. Overall, I found the results to be robust to the first two of these threats. I found that model performance improved when the model was fit with students present in all grades, indicating that our results may be slightly dampened by error introduced with imputation. Additionally, I found that the model performed slightly worse for students in a city or town compared to a suburb or rural area, implying that it cannot be assumed that a model fit for an entire state automatically works equally well for students from different geographic areas.

Extension Analyses

I extended the main analysis by examining the Youden statistic and resulting specificity and sensitivity rates and confusion matrices (results are in Appendix E). This extension analysis provides a more tangible way of conceptualizing the results by demonstrating how the models presented in the main analysis would be used in an educational setting where school personnel have to set a cut point in the likelihood of dropping out of high school in order to identify students for intervention (Bruch et al., 2020). For brevity, I only included the extension analyses for the first research question; similar analyses for the second research question are available upon request.

Youden Statistic

First, following the example of Bruch et al. (2020) and (cite paper 2), I examined what it would look like to take the nuanced continuously measured AUC measure and apply a cut point in the risk score of dropping out of high school. I did this by finding the optimal point on the receiver operating characteristic curve that optimizes the balance between the sensitivity and specificity. This point on the curve is known as the Youden statistic (also known as Youden's J statistic) and is defined as $J = max_t\{sensitivity(t) + specificity(t) - 1\}$ (Berrar, 2019; Youden, 1950). Sensitivity is true positive rate (i.e., the likelihood of correctly identifying students who dropped out) while specificity is the true negative rate (i.e., the likelihood of correctly identifying students who did not drop out) (cite paper 1). While the Youden statistic is the empirically defined optimal risk score in balancing true positive and negatives, I presented it as an example of how schools can choose a risk score that reflects their preferences in which students to prioritize due to resource constraints and policy directives (Bruch et al., 2020), as I mention later in the discussion.

Across all model types, the Youden statistic was approximately similar, ranging from 0.075 for the logistic models 2-4 (Appendix E, Table 1, Panel A, Column 2) to 0.137 for the decision tree models 7-8 (Appendix E, Table 1, Panel A, Column 4). In practice, this means that the empirically defined cut point for identifying students for intervention to prevent high school dropout would be between 7.5% and 13.7%, depending on the predictive model used. The linear probability, elastic net, and random forest models consistently had the highest Youden statistics on average although the decision tree models had the absolute highest scores. On average, it appears that the Youden statistic increases in response to using models with more years' worth of data (Appendix E, Table 1, Panel A).

Specificity, Sensitivity, and Confusion Matrices

Because the Youden statistic is defined to maximize both the specificity and sensitivity across all options derived from the possible cut points, Appendix E, Table 1, Panels B and C shows the resulting specificity and sensitivity rates, respectively, when calculated using the Youden statistics presented in Appendix E, Table 1, Panel A. These results can also be derived using the confusion matrices presented for the linear probability (Appendix E, Table 2), logistic (Appendix E, Table 3), elastic net (Appendix E, Table 4), decision tree (Appendix E, Table 5), and random forest (Appendix E, Table 6) models. Overall, the models had similar specificity rates within model type, ranging from 0.556 for the linear probability model 2 (Appendix E, Table 1, Panel B, Column 1) to 0.830 for the decision tree models 7-8 (Appendix E, Table 1, Panel B, Column 4). This means that, across all model types, the percentage of students who did not drop out who were correctly identified as not dropping out ranged from 55.6% to 83%. While the decision tree models were the most unstable in terms of specificity, the other four model types had somewhat consistent specificity values. On average, the specificity values increased as data from more years was added to the models (Appendix E, Table 1, Panel B).

Across all models, the sensitivity rates ranged from 0.611 for the decision tree models 7-8 (Appendix E, Table 1, Panel C, Column 4) to 0.824 for the random tree model 8 (Appendix E, Table 1, Panel C, Column 5), meaning that the percentage of students who dropped out who were correctly identified to drop out ranged from 61.1% to 82.4%. Similar to the specificity results, the sensitivity rates were most stable among the linear probability, logistic, elastic net, and random forest models. Additionally, the sensitivity rates increased on average as the models added more years' worth of data for the linear probability, logistic, elastic net, and random forest models (Appendix E, Table 1, Panel C).

Discussion

I interrogated how traditional regression models compared to machine learning models in predicting how likely students were to drop out of high school. Additionally, I wanted to explore how early in a student's educational career these predictive models could be built and if the models would work equally well for significant student subgroups. It is important to identify students who are likely to drop out of high school because dropping out of high school is associated with worse financial and social outcomes later in life, including lower lifetime earnings, a shorter life span, and a higher likelihood of living in poverty, committing crimes, taking drugs, and being incarcerated (Alliance for Excellent Education, 2011; Bridgeland et al., 2006). Additionally, students are more likely to drop out of high school if they are male, Pacific Islander, Latinx, Black, multiracial, live in an institutionalized setting, or have a disability (Hussar et al., 2020), meaning that preventing high school dropout is also an issue with deep equity implications. Given that there are proven effective interventions to reduce high school dropout (Dynarski et al., 2008; U.S. Department of Education, 2016) and that schools are often limited in the amount of resources they can allocate for intervention (Engler, 2020), schools need a way to accurately detect students who are likely to drop out.

Overall, I found that it is possible to build predictive models with good performance for the full sample using data from fourth and fifth grade and that model performance improves with each addition of a grade's worth of data. Of the five types of model types that I used, the random forest models performed slightly better than the linear probability, logistic, and elastic net models while the decision tree models had the worst performance. I then examined model performance across four student subgroups that have differential dropout rates: gender, race and ethnicity, economically disadvantaged, and receipt of special education services. These results generally showed that the model did not perform well for most subgroups with higher dropout rates (i.e., male, Black, Latinx, Asian, Native American/Alaskan Native, Hawaiian/Pacific Islander, multiracial, economically disadvantaged, and special education students) compared to either their counterparts nor the full sample. Although it was possible to construct predictive models that perform well for every subgroup using multiple years' worth of data, this indicates that it should not be assumed that models fit with the full sample will automatically work equally well for every subgroup.

Comparison to Prior Literature

There is a growing literature that uses machine learning to develop models to predict students likely to drop out of high school (Ara et al., 2015; Chung & Lee, 2019; Coleman et al., 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016; Orooji & Chen, 2019; Pagani et al., 2008; Sansone, 2019). This research spans multiple geographic boundaries, including individual school districts (Coleman et al., 2019; Lakkaraju et al., 2015) and states (Orooji & Chen, 2019) in the United States plus nationally representative samples (Sansone, 2019) as well as Denmark (Ara et al., 2015), South Korea (Chung & Lee, 2019), Quebec (Pagani et al., 2008), and Mexico (Márquez-Vera et al., 2016). However, only one of these papers looks at data from an entire state – Louisiana – (Orooji & Chen, 2019), so this research adds to the literature about how predictive models can inform early warning systems at a state-wide level. Additionally, this is the first paper to study machine learning models predicting high school drop out in for students across Michigan.

Similarly, administrative data is the type of data that school personnel would have readily available to use to construct an early warning system without having to collect additional data. This paper builds on the work of other research that exclusively uses administrative data to build predictive models (Ara et al., 2015; Chung & Lee, 2019; Lakkaraju et al., 2015; Orooji & Chen, 2019) as compared to using additional assessment data (Coleman et al., 2019; Márquez-Vera et al., 2016) or survey data (Pagani et al., 2008; Sansone, 2019). Accordingly, this approach is particularly scalable and cost effective.

From all of the work exploring how machine learning can predict high school dropout, only one other study explicitly compared the performance of machine learning based models to traditional regression models (Lakkaraju et al., 2015), something that will be crucial to convince school personnel to buy into using new methods that seem more opaque than traditional methods with which they are more familiar (cite paper 1). While this study found that random forest models performed 9% better than logistic regression when identifying students who dropped out (Lakkaraju et al., 2015), my results showed a much closer margin to the point where my results were not substantially convincing that machine learning models were better.

This study is the first to explicitly examine predictive model performance for subgroups when examining the likelihood of dropping out of high school. Given that my sample has differential high school dropout rates for the four subgroups that I examined in my second research question, i.e., subgroups based on gender, race and ethnicity, economic background, and receipt of special education services, and that work from other disciplines showed that machine learning models did not work equally well for subgroups from traditionally disadvantaged backgrounds (Garvie et al., 2016; Gebru, 2021; Kantayya, 2020; Lum & Isaac, 2016; Obermeyer et al., 2019; Richardson et al., 2019), it is important not to assume that a model fit for the full sample works equally well for these subgroups. To explicitly address this, I followed suit of other predictive modeling work in education to examine model performance for each subgroup (Bruch et al., 2020; Cattell & Bruch, 2021). A similar concern that is not so easily empirically tested is the idea of replicating inherent bias in the data, a concept that I discuss further in the limitation section. Although no previous work on predicting high school dropout discusses this, it should be noted that after using predictive modeling to identify students likely to drop out of high school, one study used another machine learning method to group these students into internally homogenous groups to more effectively target for intervention (Sansone, 2019).

Policy Implications

When constructing an early warning system to identify students likely to drop out of high school, one important consideration for school personnel is how early they can make accurate predictions. My models showed that it is possible to fit models that predict high school dropout accurately (i.e., an approximate 71.5% chance of correctly identifying students who dropped out of high school) using data from only fourth and fifth grade. This is encouraging if schools have resources to intervene as early as fifth grade as it may be more cost effective to intervene earlier and more impactful for the student. There is a breadth of research discussing effective interventions to prevent high school drop out across a student's P-12 educational experience, including with high-quality early childhood education and family engagement in early grades (National Dropout Prevention Center, 2022).

However, schools should also consider how much more accurate their early warning system would be if they waited one or more years to use more data. On average, my results indicated that the models improved their accuracy of correctly identifying students who dropped out of high school between 1-2% for every grade of data added after fifth grade. Additionally, both the sensitivity and specificity of the models increased with each additional year's worth of data, implying that models become more accurate in identifying students who do and do not drop out of high school with more data. Schools should balance the desire for more accurate models that come from more years' worth of data, the value of early intervention, and their ability to intervene early due to limited resources (Engler, 2020).

When considering the potential consequences of using a model that does not work equally well for subgroups in producing biased predictions (Brussow, 2018), schools should consider waiting until a model has enough years' worth of data that is has an acceptable performance for all subgroups of interest. For example, the model that incorporated fourth through seventh grade was the first model iteration that had an acceptable model performance (i.e., AUC value of 0.7 or greater) for all racial and ethnic subgroups. Given that it may be difficult to construct a model that works equally well (i.e., the same AUC and true and false positive and negative values) for every subgroup, a potential first step toward addressing this equity issue would be to make sure that a model achieves a specified performance threshold for each subgroup. Although this would not eliminate the possibility of identifying students at differential rates, it would be preferrable to using a model that has an acceptable model performance only for the full sample.

Similarly, administrators should keep in mind the consequences of true and false positive and negative rates in terms of differential impacts for these groups of students (Brussow, 2018). Like how I used the Youden statistic to dichotomize the continuous likelihood, administrators must pick a threshold in the likelihood of students dropping out of high school. That threshold can vary based on resources and the desire to optimize the type of students identified. Because these models predicted students dropping out of high school, the false negatives were students who the model did not predict to drop out yet actually did while the false positives were students who the model predicted to drop out yet actually did not. While it is ideal to minimize the numbers of students in both groups, I argue that there are greater implications for the false negative students because those are students who would not receive the intervention that they need. Given that the type of interventions shown to reduce high school dropout include actions that may be generally helpful to all students, such as extra tutoring, a more personalized learning environment, and wraparound services (Dynarski et al., 2008), it would most likely not be as harmful to provide intervention to students who end up not needing it compared to withholding intervention from those who do. All of this is tempered by the resources that schools have to allocate for intervention, stressing the need for the most accurate models possible.

Similarly, it is important to consider what an appropriate threshold would be for a model to qualify as having an acceptable performance for this specific context. Although I based my decision of an AUC value of 0.7 as this threshold following the statistical literature (Mandrekar, 2010) and previous work done in this field (Bruch et al., 2020), I recognize that there has not been enough work in the field of education early warning systems using AUC values to have definitively established a threshold agreed upon throughout both the research and practitioner communities. Furthermore, this threshold may vary based on context, including geography, school and district size, and outcome studied. Thus, administrators should consider what an appropriate cutoff would be for their particular context when deciding which statistical model to use for an early warning system.

Another point for school personnel to consider when choosing the method underlying their early warning system is whether to choose a traditional regression method that is often more familiar versus a machine learning method that is likely newer and more opaque. Generally speaking, predictive machine learning algorithms hold promise for this type of application given their statistically flexible nature to fit the data better than traditional regression models do (Hastie et al., 2009; James et al., 2013). However, the trade-off for this highly predictive nature is a less transparent model, potentially making it difficult to garner support from stakeholders to agree to an approach that is often considered a black box (cite paper 1). Although the results for the whole sample indicate that the random forest models performed slightly better than the linear probability and logistic models (i.e., 0.5% to 1.5% more accurate when identifying students who dropped out), their AUC values may not be different enough to convince stakeholders to use a less traditional approach.

With these implications in mind, and considering the limitations discussed below, it is important to keep in mind that this work is preliminary. This exploratory research indicates that schools may be well served by investing resources into more research on the feasibility and performance of these predictor models under practical circumstances. While I have attempted to replicate what I believe to be operating school conditions, school personnel would be better equipped to make decisions such as how to simulate the types of data readily available to them, which types of performance metrics to focus on, and how logistically reasonable an early warning system would be given staffing capabilities. Although my results indicate that such an early warning system based on predictor models shows promise, I present my results as a first step towards putting them into practice.

Limitations

There are several limitations to my study that should be taken into consideration when interpreting the results. First, the results of the predictive models are only as good as the data used to construct them. If there is any inherent bias in the data reflecting structural and/or systemic issues that lead to certain types of students being more likely to drop out of high school, then the models will replicate that bias by design (Goldacre, 2008). For example, my data descriptively showed that students of color, except for Asian students, were more likely to drop out of high school compared to White students, very likely reflecting the presence of such barriers that would be reflected in the predictive models. Although making sure that the models work equally well for these racial and ethnic subgroups is a helpful check on this, it is still a limitation to consider.

Similarly, it is possible that the models would be more predictive if schools used different data as predictive variables. Despite using variables from each of the three categories used to build early warning systems (i.e., attendance, behavior, and performance) (Therriault et al., 2017; U.S. Department of Education, 2016), it is possible that there are other variables in these categories that would improve model performance. For example, whereas I used student performance on standardized assessments, other studies used grade point average, grades in individual courses, and credits earned as proxies for academic performance (Coleman et al., 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016). Similarly, my models may have performed differently if I had used variables measured at the classroom, school, or district level. The literature on situational context beyond an individual student is vast, and just as it is possible that a single student may perform differently in alternative contexts, the models predicting outcomes for individual students may perform better taking these variables into account (Elder et al., 2021; Masci et al., 2018; Rumberger & Thomas, 2000). Before committing to a set model for an early warning system, schools should consider alternate variables that they have access to in case those variables improve model performance.

Another limitation is the size of my subgroups for the second research question. Compared to the full sample of 416,105 students, only 429 (<1%) students identified as Hawaiian or Pacific Islander. While the other student race and ethnicity subgroups were large enough to not be concerning, this subgroup was small enough to potentially have underpowered results. It is possible that I would have found a different model performance – perhaps one closer to that of the full sample – if this subgroup had been larger. Therefore, the differential subgroup performance for this group in particular should not automatically be assumed to be fully attributable to bias.

In terms of generalizability, the models based on students in Michigan may not directly translate to other students in other states. Michigan has an excellent longitudinal data collection system that allowed me to track students as far back as the 2004-2005 school year that other states may not have. Michigan may also be different from other states based on its dropout rate. While my sample had a dropout rate of 11%, the national status dropout rate is 5.3% (Hussar et al., 2020). Likewise, the demographic composition of Michigan may not apply to other states. Compared to the rest of the country, my sample from Michigan had a higher proportion of White students and a lower proportion of Black, Latinx, Native American/Alaskan Native, Asian, Hawaiian/Pacific Islander, and multiracial students compared to the rest of the country (Irwin et al., 2021). Therefore, caution should be applied when extrapolating these results to other states.

Lastly, this paper should be not interpreted as the conclusive answer to how traditional regression models compare to machine learning models when predicting high school dropout. Although I chose commonly used algorithms that have been used in previous studies (Ara et al., 2015; Chung & Lee, 2019; Coleman et al., 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016; Orooji & Chen, 2019; Pagani et al., 2008; Sansone, 2019), there are several more algorithms in existence that could provide different results (Athey, 2019; Hastie et al., 2009; James et al., 2013). Additionally, I could have used alternative ways of splitting my sample into training and testing data, such as randomly selecting 80% of students to be in my training dataset and the remaining 20% to be in my testing dataset or using the first three cohort as the training data with later two cohorts as the testing data. Similarly, I could have used hyperparameters specific to each algorithm to adjust model performance (Hastie et al., 2009; James et al., 2013). Given different model choices could produce alternative results, I want to impress that these results should be interpreted only in the context of the analysis I conducted.

Conclusion

Despite these limitations, these results offer important insight into how machine learning models compare to traditional regression models when predicting the likelihood of students dropping out of high school. Additionally, I explored how accurate these models were using multiple years' worth of data and model performance for specific student subgroups. Similar to previous studies (Ara et al., 2015; Chung & Lee, 2019; Coleman et al., 2019; Lakkaraju et al., 2015; Márquez-Vera et al., 2016; Orooji & Chen, 2019; Pagani et al., 2008; Sansone, 2019), I found that the best performing model was based on a machine learning algorithm, specifically the random forest model. However, I extended beyond the previous literature by finding that the linear probability, logistic, and elastic net models performed almost as well, to the point where the random forest model's superiority was not convincing. I also found that it is possible to construct well performing models using data from fourth and fifth grade for the full sample. My results also indicated that although it is possible to create well performing models for each subgroup, it should not be assumed that the full models fit for the full sample automatically work equally well for each subgroup. These results provide helpful insight for schools looking to build robust early warning systems to identify students likely to drop out of high school to identify for intervention.

Conclusion

My dissertation spoke to the need in the field for a pedagogical paper that explained the role of machine learning in education policy research to scholars unfamiliar with the relatively new methodology. In this first paper, I identified the ability to predict students for intervention as one of the most promising uses of machine learning but noted that while most of the research conducted on this showed highly accurate results, very few explicitly compared how machine learning algorithms compared to traditional regression methods.

I used this as the basis for both of my empirical papers, investigating the value added of machine learning over linear probability and logistic models in early warning systems. My second paper used researcher- and district-collected data from Boston Public Schools to predict chronic absenteeism and receipt of special education services for kindergarten through second grade students. My third paper used administrative data from Michigan to predict high school dropout across five cohorts of students. In both empirical papers, I compared the performance of three machine learning algorithms (elastic net, decision tree, and random forest) to two traditional regression models (linear probability and logistic). I also used sample sizes that were realistic for educational settings and placed an emphasis on the feasibility of timing for practitioners using data in real world applications that was absent in some of the previous work done on this topic.

For both empirical papers, I found that the machine learning models performed well, reinforcing the fact that it is possible to build accurate predictive models for all my outcomes.

However, in the third paper, I also found that the machine learning models did not perform better enough than the traditional regression methods to be convincing for school personnel to embrace the new methodology as part of an early warning system that is based on traditional methods. Furthermore, both the machine learning and traditional regression models did not perform equally well for important student subgroups as they did for the full sample, a caution for schools looking for build early warning systems that are unbiased and work equally well for all students.

Although my dissertation addressed an important gap in the field, it should not be considered the ultimate referendum on either the role of machine learning in education policy research or how predictive machine learning algorithms compare to traditional regression models. This is a vast field of study that is constantly evolving, and future research may contradict my findings as methods evolve. Looking forward, I hope to build on these three papers to further nuance and better understand how researchers and practitioners can leverage machine learning to improve educational experience and outcomes for all students.

Table 1. Main conceptual differences between traditional regression and machine learning.

	Traditional regression	Machine learning
Key focus area	Estimating parameters	Precise prediction of outcome (supervised)/groups variables or subjects (unsupervised)
Attention on variables	Emphasis on magnitude and statistical significance of coefficients	Rarely consider point estimates, sometimes do not know which variables are used, can look at which variables are most important
Role of researcher	Specifies variables and model structure to use	Specifies dataset and algorithm to use
Type of approach	Theory-driven (deductive)	Data-driven (inductive)
Major concerns	Collinearity, spurious variables	Balance between predicting and overfitting (variance-bias trade-off)

Table 2. Sample of commonly used algorithms in education research.

Algorithm	Function	Approach	Example R code	Example Stata code*	Example article using method
Decision tree	Predicting	Recursive partitioning	“tree”	“crtrees”	Lakkaraju et al. (2015)
Random forest	Predicting	Combining multiple decision trees	“randomForest”	“crtees” with “rforests” option	Chung & Lee (2019)
Neural network	Predicting	Two-step prediction through hidden layer	“neuralnet”	-	Johnes (2005)
Naïve Bayes	Predicting	Uses Bayes rule	“naivebayes”	-	Márquez-Vera et al. (2016)
Support vector machine	Predicting	Generate line that graphically separates groups	“e1071”	“svmachines”	Lakkaraju et al. (2015)
K-means	Grouping	Grouping closest subjects iteratively	“kmeans”	“cluster kmeans”	Antonenko et al. (2012)
Latent dirichlet allocation	Grouping	Parses about individual words then groups	“topicmodels”	“ldagibbs”	Sun et al. (2019)

* Beginning with Stata 16, Stata officially recommends implementing machine learning algorithms with integrated Python code. We present sample user-written packages for algorithms.

Table 3. Sample of commonly used performance metrics for predictive algorithms.

Metric	Definition	Equation	Also known as
True positive (TP)	Predicted positive subjects who are actually positive	-	-
False positive (FP)	Predicted positive subjects who are actually negative	-	-
True negative (TN)	Predicted negative subjects who are actually negative	-	-
False negative (FN)	Predict negative subjects who are actually positive	-	-
Accuracy	Proportions of correct predictions out of total predictions	(TP+TN)/(TP+FP+TN+FN)	-
Sensitivity	Proportion of true positives out of total actual positive	TP/(TP+FN)	True positive rate, recall
Specificity	Proportion of true negative out of total actual negative	TN/(TN+FP)	True negative rate
Precision	Proportion of true positive out of total predicted positive	TP/(TP+FP)	-
Area Under the Curve (AUC)*	Area under the curve when plotting (1-specificity) versus sensitivity	-	Area under ROC (receiver operating characteristic) curve

Note: Often, a 2x2 grid called a "confusion table" or "confusion matrix" displays the true positive, false positive, true negative, and false negative rates (Berrar, 2019). See Chung & Lee (2019) page 349 as an example.

*Typically, AUC values of 0.7 or greater indicate acceptable model performance (Mandrekar, 2010).

Table 4. Descriptive statistics for special education and attendance variables for paper two.

Variable	Mean/Percent	Standard deviation	*N*
Special education			
Identified as SPED in kindergarten	0.07	-	66
Identified as SPED in first grade	0.12	-	103
Identified as SPED in first grade but not in kindergarten	0.05	-	44
Identified as SPED in second grade	0.14	-	112
Identified as SPED in second grade but not in kindergarten	0.08	-	63
Attendance & enrollment			
Days enrolled in kindergarten	177.51	5.75	978
Days enrolled in first grade	178.97	7.59	878
Days enrolled in second grade	115.73	1.67	804
Days absent in kindergarten	10.55	9.59	978
Days absent in first grade	9.14	7.85	878
Days absent in second grade	4.81	4.82	804
Percent of days absent in kindergarten	0.06	0.05	978
Percent of days absent in first grade	0.05	0.04	878
Percent of days absent in second grade	0.04	0.04	804
Chronically absent in kindergarten	0.16	-	153
Chronically absent in first grade	0.12	-	103
Chronically absent in second grade	0.09	-	69

Notes: *N* = 1,012 in kindergarten, *N* = 894 in first grade, *N* = 825 in second grade. There is no missing data for any of these variables. For attendance and chronic absenteeism variables, samples are restricted to students who are enrolled at least 90 days. Chronically absent defined as missing ten percent or more of days enrolled. For the 2019-2020 school year, the attendance and enrollment data stopped at March 14, 2020 due to the COVID-19 pandemic.

Table 5. AUC values for models predicting special education status for paper two.

Model type	First grade (1)	Second grade (2)
Panel A. Linear Probability Model		
Demographics	0.522	0.641
+ Fall of K academic/executive functioning	0.683	0.781
+ Fall of K intra/interpersonal	0.702	0.774
+ Family data	0.680	0.702
+ Teacher data	0.680	0.719
+ Spring of K academic/executive functioning	0.697	0.754
+ Spring of K intra/interpersonal	0.698	0.745
Panel B. Logistic Model		
Demographics	0.501	0.614
+ Fall of K academic/executive functioning	0.657	0.763
+ Fall of K intra/interpersonal	-	-
+ Family data	-	-
+ Teacher data	-	-
+ Spring of K academic/executive functioning	-	-
+ Spring of K intra/interpersonal	-	-
Panel C. Elastic Net		
Demographics	0.543	0.631
+ Fall of K academic/executive functioning	0.713	0.813
+ Fall of K intra/interpersonal	0.737	0.820
+ Family data	0.733	0.809
+ Teacher data	0.734	0.818
+ Spring of K academic/executive functioning	0.720	0.828
+ Spring of K intra/interpersonal	0.769	0.844
Panel D. Decision Tree		
Demographics	0.517	0.579
+ Fall of K academic/executive functioning	0.599	0.709
+ Fall of K intra/interpersonal	0.511	0.701
+ Family data	0.626	0.582
+ Teacher data	0.579	0.627
+ Spring of K academic/executive functioning	0.568	0.594
+ Spring of K intra/interpersonal	0.637	0.634
Panel E. Random Forest		
Demographics	0.563	0.649
+ Fall of K academic/executive functioning	0.735	0.813
+ Fall of K intra/interpersonal	0.750	0.814
+ Family data	0.696	0.794

+ Teacher data	0.709	0.805
+ Spring of K academic/executive functioning	0.731	0.812
+ Spring of K intra/interpersonal	0.770	0.843

Table 6. AUC values for models predicting chronic absenteeism for paper two.

Model type	Kindergarten (1)	First grade (2)	Second grade (3)
Panel A. Linear Probability Model			
Demographics	0.618	0.634	0.616
+ Fall of K academic/executive functioning	0.609	0.691	0.580
+ Fall of K intra/interpersonal	0.625	0.674	0.560
+ Family data	0.615	0.620	0.618
+ Teacher data	0.593	0.642	0.589
+ Spring of K academic/executive functioning	0.570	0.649	0.614
+ Spring of K intra/interpersonal	0.610	0.650	0.635
Panel B. Logistic Model			
Demographics	0.617	0.628	0.619
+ Fall of K academic/executive functioning	0.608	0.676	0.577
+ Fall of K intra/interpersonal	0.625	0.650	-
+ Family data	-	-	-
+ Teacher data	-	-	-
+ Spring of K academic/executive functioning	-	-	-
+ Spring of K intra/interpersonal	-	-	-
Panel C. Elastic Net			
Demographics	0.649	0.637	0.604
+ Fall of K academic/executive functioning	0.630	0.666	0.612
+ Fall of K intra/interpersonal	0.647	0.662	0.654
+ Family data	0.643	0.649	0.666
+ Teacher data	0.633	0.627	0.656
+ Spring of K academic/executive functioning	0.633	0.643	0.652
+ Spring of K intra/interpersonal	0.653	0.643	0.654
Panel D. Decision Tree			
Demographics	0.608	0.591	0.564
+ Fall of K academic/executive functioning	0.576	0.579	0.472
+ Fall of K intra/interpersonal	0.550	0.573	0.556
+ Family data	0.572	0.524	0.520
+ Teacher data	0.571	0.542	0.550
+ Spring of K academic/executive functioning	0.586	0.554	0.563
+ Spring of K intra/interpersonal	0.563	0.545	0.559
Panel E. Random Forest			
Demographics	0.661	0.647	0.554
+ Fall of K academic/executive functioning	0.613	0.627	0.578
+ Fall of K intra/interpersonal	0.617	0.627	0.629

+ Family data	0.618	0.615	0.601
+ Teacher data	0.616	0.602	0.618
+ Spring of K academic/executive functioning	0.613	0.616	0.611
+ Spring of K intra/interpersonal	0.592	0.620	0.625

Table 7. AUC values for models predicting receipt of special education services and chronic absenteeism for with student demographics, fall of Kindergarten DIBELS assessments, and one extra fall of Kindergarten assessment for paper two.

	Special education		Chronic absenteeism		
Model type	First grade	Second grade	Kindergarten	First grade	Second grade
	(1)	(2)	(3)	(4)	(5)
Panel A. Linear Probability Model					
PPVT	0.660	0.809	0.640	0.661	0.587
WJAP	0.667	0.805	0.630	0.675	0.595
REMA	0.643	0.801	0.627	0.650	0.608
FDS	0.688	0.818	0.635	0.662	0.593
H&F	0.653	0.784	0.626	0.664	0.532
PSRA	0.688	0.809	0.638	0.668	0.619
TCRS	0.692	0.809	0.658	0.670	0.593
SSIS	0.692	0.803	0.645	0.656	0.577
Panel B. Logistic Model					
PPVT	0.667	0.812	0.639	0.653	0.588
WJAP	0.676	0.804	0.631	0.663	0.570
REMA	0.671	0.817	0.629	0.649	0.586
FDS	0.706	0.827	0.634	0.660	0.590
H&F	0.637	0.801	0.626	0.655	0.582
PSRA	0.690	0.815	0.637	0.663	0.624
TCRS	0.691	0.821	0.654	0.664	0.616
SSIS	0.686	0.802	0.644	0.633	0.564
Panel C. Elastic Net					
PPVT	0.675	0.822	0.637	0.647	0.578
WJAP	0.692	0.814	0.636	0.648	0.604
REMA	0.678	0.820	0.640	0.651	0.615
FDS	0.700	0.826	0.635	0.660	0.590
H&F	0.673	0.804	0.627	0.638	0.571
PSRA	0.712	0.820	0.637	0.667	0.612
TCRS	0.728	0.825	0.665	0.651	0.684
SSIS	0.733	0.826	0.648	0.647	0.633
Panel D. Decision Tree					
PPVT	0.639	0.740	0.542	0.581	0.539
WJAP	0.651	0.691	0.576	0.578	0.549
REMA	0.577	0.746	0.554	0.587	0.558
FDS	0.641	0.718	0.540	0.564	0.515
H&F	0.652	0.746	0.551	0.545	0.560

PSRA	0.615	0.725	0.558	0.598	0.589
TCRS	0.618	0.746	0.602	0.619	0.534
SSIS	0.523	0.681	0.543	0.556	0.538
Panel E. Random Forest					
PPVT	0.660	0.810	0.634	0.626	0.594
WJAP	0.682	0.793	0.641	0.625	0.585
REMA	0.695	0.812	0.650	0.647	0.601
FDS	0.688	0.826	0.643	0.636	0.592
H&F	0.685	0.801	0.636	0.617	0.579
PSRA	0.711	0.829	0.638	0.655	0.600
TCRS	0.668	0.805	0.676	0.602	0.627
SSIS	0.697	0.790	0.632	0.601	0.594

Notes: All models fit with student demographics, fall of K DIBELS measures, and one extra fall of K assessment as noted. Model 1 = Peabody Picture Vocabulary Test (PPVT), Model 2 = Woodcock Johnson Applied Problems (WJAP), Model 3 = Research-Based Early Mathematics Assessment (REMA), Model 4 = Digit Span Forward (DSF), Model 5 = Hearts and Flowers (H&F) both mixed and incongruent subscales, Model 6 = Preschool Self-Regulation Assessment (PSRA) both attention/impulse control and positive emotion subscales, Model 7 = Teacher-Child Rating Scale (TCRS) academic orientation subscale, Model 8 = Social Skills Improvement System (SSIS) cooperation, engagement, self-control, externalizing behavior, internalizing behavior, and hyperattention/inattention subscales.

Table 8. AUC values for models predicting receipt of special education services and chronic absenteeism for with student demographics, fall and spring of Kindergarten DIBELS assessments, and one extra fall and spring of Kindergarten assessment for paper two.

	Special education		Chronic absenteeism		
Model type	First grade	Second grade	Kindergarten	First grade	Second grade
	(1)	(2)	(3)	(4)	(5)
Panel A. Linear Probability Model					
PPVT	0.694	0.814	0.654	0.638	0.606
WJAP	0.690	0.815	0.652	0.660	0.612
REMA	0.674	0.810	0.650	0.654	0.609
FDS	0.708	0.814	0.654	0.646	0.582
H&F	0.696	0.810	0.643	0.647	0.606
PSRA	0.718	0.824	0.650	0.656	0.613
TCRS	0.740	0.825	0.681	0.669	0.609
SSIS	0.729	0.816	0.683	0.652	0.590
Panel B. Logistic Model					
PPVT	0.693	0.814	0.646	0.635	0.607
WJAP	0.683	0.801	0.641	0.655	0.610
REMA	0.694	0.819	0.638	0.651	0.608
FDS	0.719	0.820	0.647	0.643	0.606
H&F	0.702	0.819	0.635	0.641	0.572
PSRA	0.713	0.817	0.642	0.655	0.616
TCRS	0.748	0.824	0.670	0.664	0.606
SSIS	0.727	0.808	0.673	0.632	0.591
Panel C. Elastic Net					
PPVT	0.702	0.830	0.633	0.650	0.617
WJAP	0.707	0.832	0.645	0.672	0.618
REMA	0.701	0.834	0.644	0.659	0.619
FDS	0.707	0.830	0.627	0.665	0.614
H&F	0.697	0.834	0.627	0.657	0.607
PSRA	0.722	0.831	0.635	0.671	0.631
TCRS	0.772	0.845	0.661	0.661	0.670
SSIS	0.779	0.851	0.665	0.654	0.624
Panel D. Decision Tree					
PPVT	0.650	0.689	0.596	0.501	0.492
WJAP	0.616	0.660	0.587	0.585	0.486
REMA	0.596	0.699	0.550	0.573	0.516
FDS	0.650	0.703	0.573	0.570	0.511
H&F	0.572	0.616	0.558	0.558	0.435

PSRA	0.646	0.653	0.588	0.543	0.484
TCRS	0.628	0.642	0.570	0.561	0.514
SSIS	0.566	0.673	0.544	0.459	0.523
Panel E. Random Forest					
PPVT	0.700	0.821	0.623	0.601	0.575
WJAP	0.721	0.810	0.628	0.623	0.574
REMA	0.721	0.816	0.633	0.632	0.562
FDS	0.736	0.821	0.621	0.610	0.572
H&F	0.728	0.820	0.603	0.618	0.600
PSRA	0.764	0.850	0.634	0.638	0.566
TCRS	0.749	0.847	0.647	0.617	0.592
SSIS	0.777	0.839	0.622	0.593	0.583

Notes: All models fit with student demographics, fall and spring of K DIBELS measures, and one extra fall and spring of K assessment as noted. Model 1 = Peabody Picture Vocabulary Test (PPVT), Model 2 = Woodcock Johnson Applied Problems (WJAP), Model 3 = Research-Based Early Mathematics Assessment (REMA), Model 4 = Digit Span Forward (DSF), Model 5 = Hearts and Flowers (H&F) both mixed and incongruent subscales, Model 6 = Preschool Self-Regulation Assessment (PSRA) both attention/impulse control and positive emotion subscales, Model 7 = Teacher-Child Rating Scale (TCRS) academic orientation subscale, Model 8 = Social Skills Improvement System (SSIS) cooperation, engagement, self-control, externalizing behavior, internalizing behavior, and hyperattention/inattention subscales.

Table 9. Descriptive statistics for variables for paper three.

Variable	Mean or percent	Standard deviation	Percent missing
<u>*Outcome*</u>			
Dropout of high school	0.11	-	0%
<u>*Predictors*</u>			
Female	0.48	-	0%
White	0.68	-	0%
Black	0.20	-	0%
Latinx	0.06	-	0%
Asian	0.03	-	0%
Native American/Alaskan Native	0.01	-	0%
Hawaiian/Pacific Islander	<0.01	-	0%
Two or more races	0.02	-	0%
Economically disadvantaged	0.39	-	18%
Special education	0.11	-	18%
Limited English Proficient	0.04	-	18%
Percent of school days attended			
G4	0.96	0.06	19%
G5	0.96	0.06	18%
G6	0.95	0.07	17%
G7	0.95	0.08	15%
G8	0.95	0.08	14%
G9	0.94	0.11	8%
G10	0.93	0.12	8%
Suspended: In-school			
G7	<0.01	-	14%
G8	<0.01	-	13%
G9	0.01	-	7%
G10	0.01	-	7%
Suspended: Out-of-school			
G7	<0.01	-	14%
G8	<0.01	-	13%
G9	0.01	-	7%
G10	0.02	-	7%
Expelled			
G7	<0.01	-	14%
G8	<0.01	-	13%
G9	<0.01	-	7%
G10	<0.01	-	7%
Incident: Violent			
G6	<0.01	-	15%
G7	<0.01	-	14%
G8	<0.01	-	13%
G9	0.01	-	7%

G10	0.01	-	7%
Incident: Weapon			
G6	<0.01	-	15%
G7	<0.01	-	14%
G8	<0.01	-	13%
G9	<0.01	-	7%
G10	<0.01	-	7%
Incident: Substance abuse			
G7	<0.01	-	14%
G8	<0.01	-	13%
G9	<0.01	-	7%
G10	<0.01	-	7%
Incident: Other			
G6	<0.01	-	15%
G7	<0.01	-	14%
G8	0.01	-	13%
G9	0.03	-	7%
G10	0.03	-	7%
Assessments: Math			
G4			
G5	422.55	24.03	22%
G6	518.32	27.05	21%
G7	619.31	28.62	20%
G8	722.28	27.23	19%
Assessments: Reading			
G4	425.37	26.40	23%
G5	525.84	29.24	22%
G6	624.32	27.49	20%
G7	723.61	29.96	19%
G8	821.84	25.79	18%
Assessments: Social studies			
G6	614.68	22.75	20%
G9	916.99	25.12	13%
Assessments: Science			
G8	820.13	25.65	18%

Notes: *N* = 416,105. Percent missing includes differential number of students across grades. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade. Economically disadvantaged indicates that a student was either classified as free/reduced price lunch, receiving SNAP or TANF benefits, migrant, homeless, or in foster care. Suspended, expelled, and incident variables indicate event occurred at least one per school year. There were no students who were suspended in- or out-of-school or expelled in G4, G5, or G6. There were no students who had incidents that were violent or involved weapons or another type of incident in G4 or G5. There were no students who had a substance abuse incident in G4, G5, or G6. Assessments are in scale scores.

Table 10. AUC values for models with full sample for paper three.

Model type	Linear probability (1)	Logistic (2)	Elastic net (3)	Decision tree (4)	Random forest (5)
G4 data	0.696	0.696	0.696	0.609	0.698
+ G5 data	0.715	0.715	0.715	0.616	0.716
+ G6 data	0.740	0.740	0.740	0.642	0.746
+ G7 data	0.759	0.756	0.759	0.704	0.768
+ G8 data	0.782	0.777	0.783	0.724	0.790
+ G9 data	0.806	0.800	0.806	0.730	0.818
+ G10 data	0.825	0.819	0.826	0.728	0.843
+ Demographics	0.837	0.833	0.838	0.728	0.850

Notes: *N* = 416,105. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Table 11. AUC values for models with based on gender for paper three.

Model type	Linear probability (1)	Logistic (2)	Elastic net (3)	Decision tree (4)	Random forest (5)
Panel A. Females					
G4 data	0.705	0.705	0.705	0.613	0.706
+ G5 data	0.723	0.723	0.723	0.628	0.723
+ G6 data	0.747	0.747	0.748	0.646	0.754
+ G7 data	0.766	0.763	0.766	0.710	0.776
+ G8 data	0.789	0.784	0.790	0.734	0.798
+ G9 data	0.813	0.807	0.814	0.741	0.827
+ G10 data	0.833	0.828	0.834	0.736	0.852
+ Demographics	0.842	0.838	0.843	0.736	0.856
Panel B. Males					
G4 data	0.686	0.686	0.686	0.600	0.690
+ G5 data	0.707	0.707	0.707	0.610	0.709
+ G6 data	0.733	0.733	0.733	0.639	0.739
+ G7 data	0.751	0.750	0.752	0.699	0.760
+ G8 data	0.775	0.769	0.775	0.717	0.782
+ G9 data	0.798	0.794	0.799	0.724	0.810
+ G10 data	0.818	0.812	0.819	0.723	0.834
+ Demographics	0.826	0.822	0.827	0.723	0.840

Notes: *N* = 201,533 for females, *N* = 214,572 for males. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Table 12. AUC values for models with based on race and ethnicity for paper three.

Model type	Linear	Logistic	Elastic net	Decision	Random
	probability			tree	forest
	(1)	(2)	(3)	(4)	(5)
Panel A. White					
G4 data	0.699	0.699	0.699	0.608	0.702
+ G5 data	0.719	0.718	0.719	0.613	0.718
+ G6 data	0.745	0.743	0.745	0.630	0.750
+ G7 data	0.765	0.761	0.765	0.699	0.774
+ G8 data	0.790	0.783	0.791	0.724	0.799
+ G9 data	0.815	0.808	0.816	0.722	0.829
+ G10 data	0.834	0.827	0.835	0.714	0.853
+ Demographics	0.848	0.843	0.849	0.714	0.861
Panel B. Black					
G4 data	0.652	0.651	0.652	0.587	0.653
+ G5 data	0.674	0.671	0.674	0.583	0.665
+ G6 data	0.698	0.697	0.698	0.631	0.700
+ G7 data	0.718	0.715	0.719	0.671	0.725
+ G8 data	0.739	0.735	0.739	0.675	0.744
+ G9 data	0.755	0.753	0.756	0.704	0.770
+ G10 data	0.780	0.778	0.781	0.721	0.799
+ Demographics	0.788	0.786	0.788	0.721	0.803
Panel C. Latinx					
G4 data	0.636	0.635	0.636	0.568	0.642
+ G5 data	0.651	0.650	0.651	0.574	0.664
+ G6 data	0.686	0.684	0.686	0.613	0.700
+ G7 data	0.701	0.697	0.701	0.669	0.717
+ G8 data	0.725	0.720	0.725	0.682	0.738
+ G9 data	0.754	0.748	0.755	0.685	0.772
+ G10 data	0.782	0.774	0.783	0.697	0.804
+ Demographics	0.792	0.785	0.793	0.697	0.809
Panel D. Asian					
G4 data	0.748	0.751	0.748	0.670	0.737
+ G5 data	0.782	0.785	0.781	0.691	0.788
+ G6 data	0.795	0.797	0.795	0.678	0.792
+ G7 data	0.803	0.803	0.802	0.714	0.800
+ G8 data	0.821	0.821	0.822	0.729	0.823
+ G9 data	0.853	0.847	0.853	0.755	0.865
+ G10 data	0.870	0.860	0.871	0.733	0.882
+ Demographics	0.871	0.863	0.872	0.733	0.881

Panel E. Native American or Alaskan Native					
G4 data	0.667	0.666	0.667	0.591	0.666
+ G5 data	0.668	0.672	0.668	0.605	0.682
+ G6 data	0.684	0.688	0.684	0.590	0.690
+ G7 data	0.706	0.712	0.706	0.649	0.728
+ G8 data	0.718	0.719	0.718	0.632	0.732
+ G9 data	0.766	0.766	0.766	0.700	0.777
+ G10 data	0.781	0.783	0.782	0.708	0.807
+ Demographics	0.791	0.791	0.790	0.708	0.817
Panel F. Hawaiian or Pacific Islander					
G4 data	0.737	0.742	0.737	0.598	0.692
+ G5 data	0.693	0.703	0.692	0.646	0.690
+ G6 data	0.760	0.773	0.759	0.725	0.772
+ G7 data	0.776	0.770	0.776	0.734	0.770
+ G8 data	0.749	0.763	0.755	0.710	0.759
+ G9 data	0.743	0.767	0.756	0.661	0.790
+ G10 data	0.753	0.768	0.766	0.619	0.759
+ Demographics	0.779	0.811	0.795	0.619	0.789
Panel G. Two or more races					
G4 data	0.659	0.659	0.658	0.562	0.662
+ G5 data	0.682	0.683	0.682	0.588	0.692
+ G6 data	0.716	0.713	0.716	0.614	0.726
+ G7 data	0.748	0.744	0.748	0.697	0.763
+ G8 data	0.773	0.770	0.774	0.722	0.793
+ G9 data	0.818	0.811	0.818	0.736	0.828
+ G10 data	0.829	0.823	0.829	0.736	0.850
+ Demographics	0.836	0.830	0.837	0.736	0.855

Notes: *N* = 283,553 for White students, *N* = 85,830 for Black students, *N* = 24,048 for Latinx students, *N* = 10,883 for Asian students, *N* = 3,500 for Native American or Alaskan Native students, *N* = 429 for Hawaiian or Pacific Islander students, *N* = 7,862 for students of two or more races. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Table 13. AUC values for models with based on economically disadvantaged status for paper three.

Model type	Linear probability (1)	Logistic (2)	Elastic net (3)	Decision tree (4)	Random forest (5)
Panel A. Economically disadvantaged					
G4 data	0.621	0.619	0.620	0.563	0.622
+ G5 data	0.638	0.636	0.638	0.568	0.636
+ G6 data	0.659	0.656	0.659	0.600	0.666
+ G7 data	0.680	0.675	0.680	0.649	0.689
+ G8 data	0.707	0.699	0.708	0.665	0.713
+ G9 data	0.734	0.725	0.734	0.676	0.747
+ G10 data	0.760	0.751	0.761	0.698	0.780
+ Demographics	0.770	0.765	0.770	0.698	0.786
Panel B. Not economically disadvantaged					
G4 data	0.706	0.706	0.706	0.616	0.709
+ G5 data	0.723	0.723	0.723	0.622	0.726
+ G6 data	0.751	0.750	0.751	0.635	0.753
+ G7 data	0.770	0.767	0.770	0.699	0.779
+ G8 data	0.796	0.789	0.796	0.724	0.807
+ G9 data	0.822	0.815	0.822	0.728	0.837
+ G10 data	0.842	0.835	0.843	0.707	0.859
+ Demographics	0.847	0.840	0.848	0.707	0.864

Notes: *N* = 180,091 for students who are economically disadvantaged, *N* = 236,014 for students who are not economically disadvantaged. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Table 14. AUC values for models with based on receipt of special education services for paper three.

Model type	Linear probability (1)	Logistic (2)	Elastic net (3)	Decision tree (4)	Random forest (5)
Panel A. Received services					
G4 data	0.588	0.583	0.587	0.520	0.597
+ G5 data	0.610	0.604	0.609	0.540	0.608
+ G6 data	0.638	0.631	0.639	0.590	0.646
+ G7 data	0.655	0.645	0.656	0.629	0.667
+ G8 data	0.700	0.677	0.701	0.654	0.711
+ G9 data	0.731	0.714	0.733	0.663	0.750
+ G10 data	0.758	0.742	0.761	0.695	0.781
+ Demographics	0.768	0.753	0.770	0.695	0.787
Panel B. Did not receive services					
G4 data	0.699	0.699	0.699	0.606	0.701
+ G5 data	0.719	0.719	0.719	0.612	0.720
+ G6 data	0.745	0.744	0.745	0.634	0.750
+ G7 data	0.765	0.763	0.765	0.704	0.773
+ G8 data	0.787	0.783	0.788	0.724	0.794
+ G9 data	0.810	0.806	0.811	0.730	0.823
+ G10 data	0.830	0.825	0.831	0.730	0.847
+ Demographics	0.843	0.840	0.843	0.730	0.855

Notes: *N* = 49,003 for students who received special education services, *N* = 367,102 for students who did not receive special education services. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

References

Abadi, S., Mat The, K. S., Nasir, B. M., Huda, M., Ivanova, N. L., Sari, T. I., Maseleno, A., Satria, F., & Muslihudin, M. (2018). Application model of k-means clustering: Insights into promotion strategy of vocational high school. *International Journal of Engineering and Technology*, *7*(2.27), 182–187. https://doi.org/10.14419/ijet.v7i2.11491

Abadie, A., Chingos, M., & West, M. (2013). Endogenous Stratification in Randomized Experiments. In *NBER Working Paper Series* (No. 19742). https://doi.org/10.3386/w19742

Abadie, A., Diamond, A., & Hainmueller, A. J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Alliance for Excellent Education. (2011). *The high cost of high school dropouts: What the nation pays for inadequate high schools*. http://www.all4ed.org/files/HighCost.pdf

Anastasopoulos, L. J. (2020). *Principled estimation of regression discontinuity designs*. http://arxiv.org/abs/1910.06381

Anglin, K. L. (2019). Gather-narrow-extract: A framework for studying local policy variation using web-scraping and natural language processing. *Journal of Research on Educational Effectiveness*, *12*(4), 685–706. https://doi.org/10.1080/19345747.2019.1654576

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering metrics: The path from cause to effect*. Princeton University Press.

Ansari, A., & Gottfried, M. A. (2018). Early childhood educational settings and school absenteeism for children with disabilities. *AERA Open*, *4*(2), 1–15. https://doi.org/10.1177/2332858418785576

Ansari, A., & Purtell, K. M. (2018). School absenteeism through the transition to kindergarten. *Journal of Education for Students Placed at Risk*, *23*(1–2), 24–38. https://doi.org/10.1080/10824669.2018.1438202

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, *60*(3), 383–398. https://doi.org/10.1007/s11423-012-9235-8

Ara, N.-B., Halland, R., Igel, C., & Alstrup, S. (2015). High-school dropout prediction using machine learning: A Danish large-scale study. In M. Verleysen (Ed.), *ESANN 2015: 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (pp. 319–324). i6doc.com

Arndt, T., & Guercio, A. (2016). A formalism for PLAN: A big data personal learning assistant for university students. *Journal of E-Learning and Knowledge Society*, *12*(2), 13–25.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*(February), 483–485.

Athey, S. (2019). The impact of machine learning on economics. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence* (pp. 507–547). The University of Chicago Press.

Athey, S. (2015). Machine learning and causal inference for policy evaluation. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 5–6.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects.

Proceedings of the National Academy of Sciences of the United States of America, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Baker, R. S., & Hawn, A. (2021). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00285-9

Balfanz, R., & Byrnes, V. (2012). *The importance of being In school: A report on absenteeism in the nation's public schools*. http://new.every1graduates.org/wp-content/uploads/2012/05/FINALChronicAbsenteeismReport_May16.pdf

Ballis, B., & Heath, K. (2019). *The long-run impacts of special education* (No. 19–151; EdWorking Paper).

Ballis, B., & Heath, K. (2021). *Direct and spillover effects of limiting minority student access to special education* (No. 21–364; EdWorkingPaper). https://www.edworkingpapers.com/ai21-364

Barrueco, S., Lopez, M., Ong, C., & Lozano, P. (2012). *Assessing Spanish-English bilignual preschoolers: A guide to best approaches and measures*. Paul H. Brookes Publishing.

Battin-Pearson, S., Newcomb, M. D., Abbott, R. D., Hill, K. G., Catalano, R. F., & Hawkins, J. D. (2000). Predictors of early high school dropout: A test of five theories. *Journal of Educational Psychology*, *92*(3), 568–582. https://doi.org/10.1037/0022-0663.92.3.568

Belloni, A., Chernozhukov, V., & Hansen, C. (2011). LASSO methods for Gaussian instrumental variables models. In *MIT Department of Economics Working Paper Series* (No. 11–14). https://doi.org/10.2139/ssrn.1908409

Ben-Michael, E., Feller, A., & Rothstein, J. (2021). Synthetic controls with staggered adoption. In *NBER Working Paper Series* (No. 28886). https://doi.org/10.2139/ssrn.3861415

Berrar, D. (2019). Performance measures for binary classification. *Encyclopedia of*

Bioinformatics and Computational Biology, *1*, 546–560. https://doi.org/10.1016/B978-0-12-809633-8.20351-8

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. In *EdWorkingPaper* (No. 21–438). https://doi.org/10.26300/hd2e-7e02

Bloom, H. S., & Michalopoulos, C. (2013). When is the story in the subgroups?: Strategies for interpreting and reporting intervention effects for subgroups. *Prevention Science*, *14*, 179–188. https://doi.org/10.1007/s11121-010-0198-x

Bloom, H. S., & Weiland, C. (2015). *Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study*. http://www.mdrc.org/sites/default/files/quantifying_variation_in_head_start.pdf

Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., Lord, C., & Narayanan, S. S. (2016). Use of machine learning to improve autism screening and diagnostic instruments: Effectiveness, efficiency, and multi-instrument fusion. *Journal of Child Psychology and Psychiatry*, *57*(8), 927–937. https://doi.org/10.1111/jcpp.12559

Bosch, N., D'Mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems*, *6*(2), 1–26. https://doi.org/10.1145/2946837

Bridgeland, J. M., DiIulio Jr., J. J., & Morison, K. B. (2006). *The silent epidemic: Perspectives of high school dropouts*. https://docs.gatesfoundation.org/documents/thesilentepidemic3-06final.pdf

Broussard, M. (2018). *Artificial unintelligence: How computer misunderstand the world*. MIT

Press.

Bruch, J., Gellar, J., Cattell, L., Hotchkiss, J., & Killewald, P. (2020). *Using data from schools and child welfare agencies to predict near-term academic risks.* https://files.eric.ed.gov/fulltext/ED606230.pdf

Brussow, J. A. (2018). Consequential validity evidence. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 372–374). SAGE Publications, Inc. https://doi.org/10.4135/9781506326139.n142

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology*, *33*(3), 205–228. https://doi.org/10.1080/87565640801982312

Cannas, M., & Arpino, B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal*, *61*(4), 1049–1072. https://doi.org/10.1002/bimj.201800132

Carminucci, J., Hodgman, S., Rickles, J., & Garet, M. (2021). *Student attendance and enrollment Loss in 2020–21.* https://www.air.org/sites/default/files/2021-07/research-brief-covid-survey-student-attendance-june-2021_0.pdf

Cattell, L., & Bruch, J. (2021). *Identifying students at risk using prior performance versus a machine learning algorithm.* https://ies.ed.gov/ncee/rel/regions/midatlantic/pdf/REL_2021126.pdf

Center for Educational Performance and Information. (2021). *Understanding Michigan's cohort gradution and dropout rates.* https://www.michigan.gov/documents/cepi/Understanding_Michigans_Cohort_Grad-

Drop_Rates_599718_7.pdf

Chang, H. ., & Davis, R. (2015). *Mapping the early attendance gap: Charting a course for school success*. https://www.attendanceworks.org/mapping-the-early-attendance-gap/

Chang, H. ., & Romero, M. (2008). *Present, engaged, and accounted for: The critical importance of addressing chronic absence in the early grades*. http://www.nccp.org/wp-content/uploads/2008/09/text_837.pdf

Chernozhukov, V., Demirer, M., Duflo, E., & Fernández-Val, I. (2016). *Generic machine learning inference on heterogeneous treatment effects in randomized experiments with an application to immunization in India* (No. 24678; NBER Working Paper Series).

Chernozhukov, V., Fernández-Val, I., & Luo, Y. (2018). The sorted effecs method: Discovering heterogeneous effects beyond their averages. *Econometrica*, *86*(6), 1911–1938.

Choi, J. Y., Jeon, S., & Lippard, C. (2018). Dual language learning, inhibitory control, and math achievement in Head Start and kindergarten. *Early Childhood Research Quarterly*, *42*(September 2017), 66–78. https://doi.org/10.1016/j.ecresq.2017.09.001

Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, *96*, 346–353. https://doi.org/10.1016/j.childyouth.2018.11.030

Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, *38*(2), 136–163.

Coleman, C., Baker, R. S., & Stephenson, S. (2019). A better cold-start for early prediction of student at-risk status in new school districts. *Proceedings of the 12th International Conference on Educational Data Mining*.

Coulacoglou, C., & Saklofske, D. H. (2017). Executive function, theory of mind, and adaptive behavior. In *Psychometrics and psychological assessment: Principles and applications* (Issue 1, pp. 91–130). Academic Press.

Council of Chief State School Officers. (2019). *Third grade reading laws: Implementation and impact*. https://ccsso.org/sites/default/files/2019-08/CCSSO CEELO third grade reading.pdf

Criado Perez, C. (2019). *Invisible women: Data bias in a world designated for men*. Abrams Press.

Crown, W. H. (2019). Real-World Evidence, Causal Inference, and Machine Learning. *Value in Health*, *22*(5), 587–592. https://doi.org/10.1016/j.jval.2019.03.001

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

Data Quality Campaign. (2013). *Supporting early warning systems: Using data to keep students on track to success*. https://dataqualitycampaign.org/wp-content/uploads/2016/03/Supporting-Early-Warning-Systems.pdf

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037–2078. https://doi.org/10.1016/j.neuropsychologia.2006.02.006

Davis, J. M. V., & Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review: Papers & Proceedings*, *107*(5), 546–550. https://doi.org/10.1257/aer.p20171000

Dede, C., Ho, A., & Mitros, P. (2016). Big Data Analysis in Higher Education: Promises and Pitfalls. *Educause*, *51*(5). https://doi.org/http://nrs.harvard.edu/urn-3:HUL.InstRepos:34785368

Diamond, K. E., Justice, L. M., Siegler, R. S., & Snyder, P. A. (2013). Synthesis of IES research on early intervention and early childhood education. In *National Center for Special Education Research* (Issue July). https://doi.org/NCESR 2013-3001

Dimakopoulou, M., Zhou, Z., Athey, S., & Imbens, G. (2017). Estimation considerations in contextual bandits. *ArXiv*, 1–46. http://arxiv.org/abs/1711.07077

Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. K. (2017). Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. *International Learning Analytics & Knowledge Conference*. https://doi.org/10.1097/01.NUMA.0000547263.07462.95

Duda, M., Ma, R., Haber, N., & Wall, D. P. (2016). Use of machine learning for behavioral distinction of autism and ADHD. *Translational Psychiatry*, *6*(2), e732. https://doi.org/10.1038/tp.2015.221

Duncan, S. E., & DeAvila, E. A. (1998). *PreLAS*. CBT McGraw Hill.

Dunn, L. M., & Dunn, L. M. (1997). *PPVT-III: Peabody picture vocabulary test*.

Dynarski, M., Clarke, L., Cobb, B., Finn, J., Rumberger, R. W., & Smink, J. (2008). *Dropout prevention: A practice guide (NCEE 2008-4025)*. http://ies.ed.gov/ncee/wwc

Early, D. M., & Li, W. (2020). *Google and Yelp reviews as a window into public perceptions of early care and education in Georgia*. http://www.decal.ga.gov/documents/attachments/GoogleYelpBrief.pdf

Early, D. M., Maxwell, K. L., Burchinal, M., Bender, R. H., Ebanks, C., Henry, G. T., Iriondo-Perez, J., Mashburn, A. J., Pianta, R. C., Alva, S., Bryant, D., Cai, K., Clifford, R. M., Griffin, J. A., Howes, C., Jeon, H.-J., Peisner-Feinberg, E., & Vandergrift, N. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development*, *78*(2), 558–580.

Elder, T. E., Figlio, D., Imberman, S. A., & Persico, C. (2021). School segregation and racial gaps in special education identification. *Journal of Labor Economics*, *39*(S1), S151–S197. https://doi.org/10.1017/CBO9781107415324.004

Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press.

Engler, A. (2020). *What all policy analysts need to know about data science*. https://www.brookings.edu/research/what-all-policy-analysts-need-to-know-about-data-science/

Engler, A. (2021). *Enrollment algorithms are contributing to the crises of higher education*. https://www.brookings.edu/research/enrollment-algorithms-are-contributing-to-the-crises-of-higher-education/

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*. https://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/REL_2017272.pdf

Friedberg, R., Tibshirani, J., Wager, S., & Athey, S. (2019). *Local linear forests*.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An Introduction* (Second). SAGE Publishing.

Garvie, C., Bedoya, A., & Frankle, J. (2016). *The perpetual line-up: Unregulated police face recognition in America*. https://www.perpetuallineup.org/

Gathercole, S. E. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Sciences*, *3*(11), 410–419. https://doi.org/10.1016/S1364-6613(99)01388-1

Gebru, T. (2021). *Computer vision: Who is helped and who is harmed?* https://midas.umich.edu/event/midas-webinar-series-presents-timnit-gebru/

Geverdt, D. (2018). *Education Demographic and Geographic Estimates (EDGE) geocodes: Public schools and location education agencies*. https://nces.ed.gov/programs/edge/docs/EDGE_GEOCODE_PUBLIC_FILEDOC.pdf

Gibson, D. C., & Ifenthaler, D. (2017). Preparing the next generation of education researchers for big data in higher education. In *Big Data and Learning Analytics in Higher Education: Current Theory and Practice* (pp. 29–42). https://doi.org/10.1007/978-3-319-06520-5

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: education, policy, 'Big Data' and principles for a critical race theory of statistics. *Race Ethnicity and Education*, *21*(2), 158–179. https://doi.org/10.1080/13613324.2017.1377417

Goldacre, B. (2008). *Bad science: Quacks, hacks, and big pharma flacks*. Farrar, Straus and Giroux.

Good, R. H., Kaminsky, R. A., Cummings, K. D., Dufour-Martel, C., Peterson, K., Powell-Smith, K., Stollar, S., & Wallin, J. (2011). *DIBELS next assessment manual*.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Green, D. P., & Kern, Holger, L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian Additive Regression Trees. *The Public Opinion Quarterly*, *76*(3), 491–511.

Gresham, F. M., & Elliott, S. N. (2008). *Social Skills Improvement System Rating Scales manual*. NCS Pearson.

Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *Political Science and Politics*, *48*(1), 80–83. https://doi.org/10.1017/S1049096514001784

Guralnick, M. J. (1998). Effectiveness of early intervention for vulnerable chidren: A developmental perspective. *American Journal on Mental Retardation*, *102*(4), 319–345. https://doi.org/https://doi.org/https://doi.org/10.1352/0895-8017

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2002). Inferring program effects for special populations: Does special education raise achievement for students with disabilities? *The Review of Economics and Statistics*, *84*(4), 584–599.

Harrell Jr., F. E. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. In *Technometrics* (Second, Vol. 45, Issue 2). Springer. https://doi.org/10.1198/tech.2003.s158

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer-Verlag. https://doi.org/10.1007/978-1-4419-9863-7_941

Hazlett, C. (2016). *Kernel Balancing: A flexible non-parametric weighting procedure for estimating causal effects*. http://arxiv.org/abs/1605.00155

Higgins, M. J., Savje, F., & Sekhon, J. S. (2014). *Improving experiments by pptimal blocking: Minimizing the maximum within-block distance*. http://sekhon.berkeley.edu/papers/HigginsSekhon.pdf

Hightower, A. D., Work, W. C., Cowen, E. L., Lotyczewski, B. S., Spinell, A. P., Guare, J. C., & Rohrbeck, C. A. (1986). The Teacher-Child Rating Scale: A brief objective measure of elementary children’s school problem behaviors and competencies. *School Psychology Review*, *15*(3), 393–409. https://doi.org/10.1080/02796015.1986.12085242

Hill, K. (2020, June 24). Wrongfully accused by an algorithm. *New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Holdnack, J. A. (2019). The development, expansion, and future of the WAIS-IV as a cornerstone in comprehensive cognitive assessments. In G. Goldstein, D. N. Allen, & J. DeLuca (Eds.), *Handbook of psychological assessment* (Fourth, pp. 103–139). Academic Press. https://doi.org/https://doi.org/10.1016/C2014-0-01970-3

Hothorn, T., Buhlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355–373. https://doi.org/10.1093/biostatistics/kxj011

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hurwitz, S., Cohen, E. D., & Perry, B. L. (2021). Special education is associated with reduced odds of school discipline among students with disabilities. *Educational Researcher*, *50*(2), 86–96. https://doi.org/10.3102/0013189X20982589

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., Smith, M., Bullock Mann, F., Barmer, A., & Dilig, R. (2020). *The condition of Education 2020 (NCES 2020-144)*. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020144

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, *7*(1), 443–470. https://doi.org/10.1214/12-AOAS593

Imbens, G. W. (2019). *Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics*. http://arxiv.org/abs/1907.07271

Irwin, V., Zhang, J., Hein, S., Wang, K., Roberts, A., York, C., Barmer, A., Bullock Mann, F., Dilig, R., & Parker, S. (2021). *Report on the condition of education 2021 (NCES 2021-144)*. https://nces.ed.gov/pubs2021/2021144.pdf

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer. https://doi.org/10.2200/S00899ED1V01Y201902MAS024

Jing, L., & Cohen, J. (2021). *Measuring teaching practices at scale: A novel application of text-as-data methods* (No. 21–369; EdWorkingPaper). https://doi.org/10.26300/6bqj-vh81

Johnes, G. (2005). Don't know much about history: Revisiting the impact of curriculum on subsequent labour market outcomes. *Bulletin of Economic Research*, *57*(3), 249–272.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Kalil, A., Mayer, S. E., & Gallegos, S. (2019). Using behavioral insights to increase attendance at subsidized preschool programs: The Show Up to Grow Up intervention. *Organizational Behavior and Human Decision Processes*. https://doi.org/10.1016/j.obhdp.2019.11.002

Kantayya, S. (2020). *Coded bias*. 7th Empire Media.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316801

Kinn, D. (2018). *Synthetic control methods and big data*. http://arxiv.org/abs/1803.00096

Kirksey, J. J., & Gottfried, M. A. (2021). The effect of serving "Breakfast After-the-Bell" meals on school absenteeism: Comparing results from regression discontinuity designs. *Educational Evaluation and Policy Analysis*, *43*(2), 305–328.

https://doi.org/10.3102/0162373721991572

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ledwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. In *NBER Working Paper Series* (No. 23180). https://doi.org/10.1142/s1793557119500840

Korman, H. T. N., O'Keefe, B., & Repka, M. (2021). *Missing in the margins 2021: Revisiting the COVID-19 attendance crisis*.

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In V. Palade, R. J. Howlett, & L. C. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems. KES 2003. Lecture Notes in Computer Science, vol 2774* (pp. 267–274). Springer-Verlag. https://doi.org/10.1007/978-3-540-45226-3_37

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1909–1918. https://doi.org/10.1145/2783258.2788620

Landry, S. H., Swank, P. R., Smith, K. E., Assel, M. A., & Gunnewig, S. B. (2006). Enhancing early literacy skills for preschool children: Bringing a professional development model to scale. *Journal of Learning Disabilities*, *39*(4), 306–324. https://doi.org/10.1177/00222194060390040501

Lee, V. R., Wilkerson, M. H., & Lanouette, K. (2021). A call for a humanistic stance toward K–12 data science education. *Educational Researcher*, 1–9. https://doi.org/10.3102/0013189X211048810

Lin, Y., & Magnuson, K. A. (2018). Classroom quality and children's academic skills in child care centers: Understanding the role of teacher qualifications. *Early Childhood Research Quarterly*, *42*, 215–227. https://doi.org/10.1016/j.ecresq.2017.10.003

Linden, A., & Yarnold, P. R. (2016). Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, *22*(6), 864–870. https://doi.org/10.1111/jep.12592

Lipsey, M. W., Nesbitt, K. T., Farran, D. C., Dong, N., Fuhs, M. W., & Wilson, S. J. (2017). Learning-related cognitive self-regulation measures for prekindergarten children: A comparative evaluation of the educational relevance of selected measures. *Journal of Educational Psychology*, *109*(8), 1084–1102. https://doi.org/10.1037/edu0000203

Loeb, S., Dynarski, S., Mcfarland, D., Morris, P., Reardon, S., & Reber, S. (2017). *Descriptive analysis in education: A guide for researchers*.

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Fardoun, H. M., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Masci, C., Johnes, G., & Agasisti, T. (2018). Student and school performance across countries: A machine learning approach. *European Journal of Operational Research*, *269*(3), 1072–1085. https://doi.org/10.1016/j.ejor.2018.02.031

McCormick, M. P., Weissman, A. K., Weiland, C., Hsueh, J., Sachs, J., & Snow, C. (2020).

Time well spent: Home learning activities and gains in children's academic skills in the prekindergarten year. *Developmental Psychology*, *56*(4), 710–726. https://doi.org/10.1037/dev0000891.supp

McElreath, R. (2020). The Haunted DAG &amp; The Causal Terror. In *Statistical Rethinking* (2nd ed., pp. 161–190). Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608-6

Morgan, P. L., Farkas, G., Hillemeier, M. M., Mattison, R., Maczuga, S., Li, H., & Cook, M. (2015). Minorities are disproportionately underrepresented in special education: Longitudinal evidence across five disability conditions. *Educational Researcher*, *44*(5), 278–292. https://doi.org/10.3102/0013189X15591157

Morrissey, T. W., Hutchison, L., & Winsler, A. (2014). Family income, school attendance, and academic achievement in elementary school. *Developmental Psychology*, *50*(3), 741–753. https://doi.org/10.1037/a0033848

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press, USA.

National Association of State Boards of Education. (n.d.). *Chronic absenteeism early warning systems*. https://statepolicies.nasbe.org/health/categories/social-emotional-climate/chronic-absenteeism-early-warning-systems

National Center on Response to Intervention. (2010). Essential components of RTI: A closer look at response to intervention. In *U.S. Office of Special Education Programs*. http://www.rti4success.org

National Dropout Prevention Center. (2022). *Effective Strategies*.

https://dropoutprevention.org/effective-strategies/

Nield, R. C., & Balfanz, R. (2006). An extreme degree of difficulty: The educational demographics of urban neighborhood high schools. *Journal of Education for Students Placed at Risk*, *11*(2), 123–141. https://doi.org/10.1207/s15327671espr1102

O'Cummings, M., & Therriault, S. B. (2015). *From accountability to prevention: Early warning systems put data to work for struggling students* (Issue May).

O'Neil, C. (2016). *Weapons of math destruction*. Broadway Books.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Orooji, M., & Chen, J. (2019). *Predicting Louisiana public high school dropout through imbalanced learning techniques*. https://arxiv.org/pdf/1910.13018.pdf

Pagani, L. S., Japel, C., Vitaro, F., Tremblay, R. E., Larose, S., & McDuff, P. (2008). When predictions fail: The case of unexpected pathways toward high school dropout. *Journal of Social Issues*, *64*(1), 175–193. https://doi.org/10.1111/j.1540-4560.2008.00570.x

Parr, A. K., & Bonitz, V. S. (2015). Role of family background, student behaviors, and school-related beliefs in predicting high school dropout. *Journal of Educational Research*, *108*(6), 504–514. https://doi.org/10.1080/00220671.2014.917256

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Petrilli, M. J. (2018). Big data transforms education research: Can machine learning unlock the keys to great teaching? *Education Next*, 86–87.

Porter, K. E. (2019). *Using predictive analytics to combine indicators of third-grade reading* *proficiency*. https://www.mdrc.org/publication/using-predictive-analytics-combine-indicators-third-grade-reading-proficiency

Porter, K. E., Redcross, C., & Miratrix, L. (2020). *Balancing promise and caution in pretrial risk assessments*. https://www.mdrc.org/publication/balancing-promise-and-caution-pretrial-risk-assessments

Powell, D. R., Son, S.-H., File, N., & San Juan, R. R. (2010). Parent–school relationships and children's academic and social outcomes in public school pre-kindergarten. *Journal of School Psychology*, *48*(4), 269–292. https://doi.org/10.1016/j.jsp.2010.03.002

Puma, M., Bell, S., Cook, R., Heid, C., & Lopez, M. (2005). *Head Start impact study: First year findings*.

Pyne, J., Grodsky, E., Vaade, E., McCready, B., Camburn, E., & Bradley, D. (2021). The Signaling Power of Unexcused Absence from School. *Educational Policy*, 089590482110494. https://doi.org/10.1177/08959048211049428

Rafferty, Y., Piscitelli, V., & Boettcher, C. (2003). The impact of inclusion on language development and social competence among preschoolers with disabilities. *Exceptional Children*, *69*(4), 467–479. https://doi.org/10.1177/001440290306900405

Raver, C. C., Li-Grining, C., Bub, K., Jones, S. M., Zhai, F., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development*, *82*(1), 362–378.

Ready, D. D. (2010). Socioeconomic disadvantage, school attendance, and early cognitive development: The differential effects of school exposure. *Sociology of Education*, *83*(4), 271–286. https://doi.org/10.1177/0038040710383520

Reardon, S. F., & Portilla, X. A. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open*, *2*(3), 1–18. https://doi.org/10.1177/2332858416657343

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review*, *94*(2), 192–233. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423#

Robinson, C. D., Lee, M. G., Dearing, E., & Rogers, T. (2018). Reducing student absenteeism in the early grades by targeting parental beliefs. *American Educational Research Journal*, *55*(6), 1163–1192. https://doi.org/10.3102/0002831218772274

Rubin, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. http://www.fsb.muohio.edu/lij14/420_paper_Rubin74.pdf

Rumberger, R. W., & Larson, K. A. (1998). Student mobility and the increased risk of high school dropout. *American Journal of Education*, *107*(1), 1–35. https://doi.org/10.1086/444201

Rumberger, R. W., & Thomas, S. L. (2000). The distribution of dropout and turnover rates among urban and suburban high schools. *Sociology of Education*, *73*(1), 39–67. https://doi.org/10.2307/2673198

Sales, A. C., Hansen, B. B., & Rowan, B. (2017). Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates. *Journal of Educational and Behavioral Statistics*, *43*(1), 3–31. https://doi.org/10.3102/1076998617731518

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

https://doi.org/10.1111/obes.12277

Schrank, F. A., McGrew, K. S., Ruef, M. L., Alvarado, C. G., Muñoz-Sandoval, A. F., & Woodcock, R. W. (2005). *Overview and technical supplement*.

Schuler, M. S., & Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, *185*(1), 65–73. https://doi.org/10.1093/aje/kww165

Scott, S. L. (2010). A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, *26*, 639–658. https://doi.org/10.1002/asmb

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs*. Houghton Mifflin Company.

Shovan, H. I., & Haque, M. (2012). An approach of improving student's academic performance by using K-means clustering algorithm and decision tree. *International Journal of Advanced Computer Science and Applications*, *3*(8). https://doi.org/10.14569/ijacsa.2012.030824

Singer, J. (2019). *Shaping the arc of education*. https://www.sree.org/video/index.php?s=2019SBall1

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly*, *22*(2), 173–187. https://doi.org/10.1016/j.ecresq.2007.01.002

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Snowling, M. J. (2013). Early identification and interventions for dyslexia: A contemporary view. *Journal of Research in Special Educational Needs*, *13*(1), 7–14.

https://doi.org/10.1111/j.1471-3802.2012.01262.x

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Sullivan, A. L., & Field, S. (2013). Do preschool special education services make a difference in kindergarten reading and mathematics skills?: A propensity score weighting analysis. *Journal of School Psychology*, *51*(2), 243–260. https://doi.org/10.1016/j.jsp.2012.12.004

Sun, M., Liu, J., Zhu, J., & LeClair, Z. (2019). Using a text-as-data approach to understand reform processes: A deep exploration of school improvement strategies. *Educational Evaluation and Policy Analysis*, *41*(4), 510–536. https://doi.org/10.3102/0162373719869318

Thabtah, F., & Peebles, D. (2020). A new machine learning model based on induction of rules for autism detection. *Health Informatics Journal*, *26*(1), 264–286. https://doi.org/10.1177/1460458218824711

Therriault, S. B., O'cummings, M., Heppen, J., Yerhot, L., Scala, J., Bailey, B., Dailey, D., Schanfield, M., & Zuber, T. (2017). *Early warning intervention and monitoring system implementation guide*. https://www.michigan.gov/documents/mde/Michigan_EWIMS_Implementation_Guide_606186_7.pdf

Tran, L., & Gershenson, S. (2021). Experimental estimates of the student attendance production function. *Educational Evaluation and Policy Analysis*, *43*(2), 183–199. https://doi.org/10.3102/0162373720984463

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

http://ies.ed.gov/ncee/edlabs/projects/ews.asp.

U.S. Department of Education National Center for Education Statisics. (2019). Table 214.20.: Number and percentage distribution of regular public school districts and students, by enrollment size of district: Selected years, 1979-80 through 2017-18. In *Digest of Education Statistics* (2019th ed.).

Ullery, M. A., & Katz, L. (2016). Beyond Part C: Reducing middle school special education for early intervention children with developmental delays. *Exceptionality*, *24*(1), 1–17. https://doi.org/10.1080/09362835.2014.986601

Vigneau, E., Devaux, M. F., Qannari, E. M., & Robert, P. (1997). Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. *Journal of Chemometrics*, *11*(3), 239–249. https://doi.org/10.1002/(SICI)1099-128X(199705)11:3<239::AID-CEM470>3.0.CO;2-A

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., Jurgens, D., Jurafsky, D., & Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(25), 6521–6526. https://doi.org/10.1073/pnas.1702413114

Voulgarides, C. K., Fergus, E., & Thorius, K. A. K. (2017). Pursuing equity: Disproportionality in special education and the reframing of technical solutions to address systemic inequities. *Review of Research in Education*, *41*(March), 61–87. https://doi.org/10.3102/0091732X16686947

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–

1242. https://doi.org/10.1080/01621459.2017.1319839

Weiland, C., Barata, M. C., & Yoshikawa, H. (2014). To co-occuring development of executive function skills and receptive vocabulary in preschool-aged children: A look at the direction of the developmental pathways. *Infant and Child Development*, *23*, 4–21. https://doi.org/10.1002/icd

Weiland, C., McCormick, M., Mattera, S., Maier, M., & Morris, P. (2018). Preschool curricula and professional development features for getting to high-quality implementation at scale: A comparative review across five trials. *AERA Open*, *4*(1), 1–16. https://doi.org/10.1177/2332858418757735

Weiland, C., Sachs, J., McCormick, M., Hsueh, J., & Snow, C. (2021). Fast-response research to answer practice and policy questions. *Future of Children*, *31*(1), 75–96. www.futureofchildren.org

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of cognitive abilities*.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838. https://doi.org/10.1177/0011000006288127

Ye, X. (2018). Personalized advising for college math: Experimental evidence on the use of human expertise and machine learning to improve college choice. In *Working Paper*. http://www-personal.umich.edu/~yxy/Ye_jmp.pdf

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35.

Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business and Economic Statistics*, *39*(1), 272–281. https://doi.org/10.1080/07350015.2019.1624293

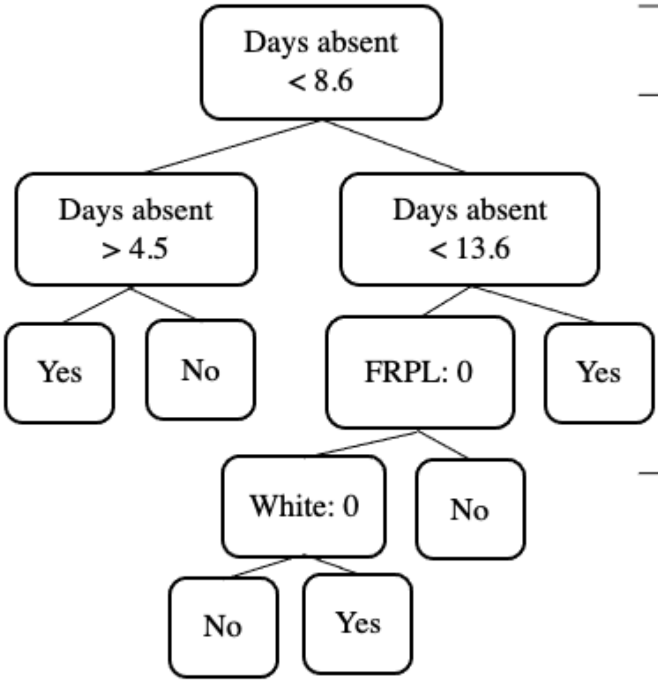
Appendix A

Decision trees are one of the easiest supervised machine learning algorithms to conceptually understand and also have the added benefit of often having lower error rates than more complicated algorithms in education research conducted so far, such as support vector machines and naïve Bayes (Ara et al., 2015; Lakkaraju et al., 2015). This nonparametric approach classifies data points using "recursive partitioning." The algorithm sorts through the entire dataset and determines which variable (at which value) divides the dataset into the desired outcome groups with the least amount of error. After this first partition, it repeats this process for each partition until the data is sufficiently portioned or until it reaches the maximum number of variables or partitions used if specified by the researcher ahead of time. This iterative process can also be thought of graphically by plotting each data point in a k-dimensional space with k number of covariates. Each partition slices through the plane corresponding to the variable used to create the partition, physically separating data points into classified groups (Hastie et al., 2009; James et al., 2013; Strobl et al., 2009).

Decision trees are aptly named because their output looks like a tree. Starting at the top, the researcher can trace the classification decision for individual data point by moving to the left at each partition of the decision is true and to the right if the decision is false. The example seen in Appendix Figure 1 was constructed with a mock dataset consisting of $N = 298$ students to predict which students would drop out of high school during their junior year. The covariate variables consisted of the number of days absent that school year and demographic variables (gender, bilingual status, free-reduced price lunch (FRPL) status, and race/ethnicity indicators). Of the 298 students in the dataset, *N* = 249 (84%) did not drop out of high school. This decision tree did not use all of the possible covariates and repeated one variable (days absent) twice (Appendix A Figure 1). The option to repeat variables is one of many possible the user-dictated hyper parameters, along with the number of partitions to create.

Despite their benefits, decisions have some drawbacks, namely that they can be become unstable (i.e., difficult to replicate due to overfitting) with large amounts of input variables because a slight change in the covariates could potentially change the entire tree structure depending on that variable's classification utility. Therefore, instead of using a single decision tree, most researchers use a "random forest," which is a large number (such as *N* = 1,000) of decision trees. Each decision tree within the forest is generated using a random sample of the input variables, and the data used to create each tree is comprised of bootstrapped (i.e., random sampling with replacement) samples from the training data in a process called "bagging." The ultimate classification decision is taken from an average across all of the trees in the forest. It is also possible to determine which variables were most important in the classification process on average across the forest. Random forests are more reliable than single decision trees and are not as complicated as some other algorithms (Hastie et al., 2009; James et al., 2013; Strobl et al., 2009).

Appendix A Figure 1
Sample decision tree with corresponding data points.



Days absent	**FRPL**	**White**	**Predicted dropout**
9.0	Yes	Yes	No
13.5	No	Yes	Yes
12.7	No	No	No
8.5	Yes	Yes	Yes
4.0	Yes	No	No

Appendix B

Overview of Robustness Checks for Paper 2

Threats to the robustness of our results include: 1) missing data, 2) inclusion of students identified as receiving special education services in kindergarten, 3) sensitivity of using alternative definitions of chronic absenteeism, 4) saturation of predictor variables, and 5) sensitivity of using student demographics. Overall, we found our results to be robust to these five threats. In all models, just as in our primary analytic strategy, we used five model types (linear probability, logistic, elastic net, decision tree, and random forest) with seven conceptual blocks of predictor variables (Model 1 = child demographics, Model 2 = Model 1 with academic and executive functioning assessments (fall of K), Model 3 = Model 2 with inter- and intrapersonal assessments (fall of K), Model 4 = Model 3 with family data, Model 5 = Model 4 with teacher data, Model 6 = Model 5 with academic and executive functioning assessments (spring of K), Model 7 = Model 6 with inter- and intrapersonal assessments (spring of K)). The exception to this was the models addressing threat 4, which fit models with separate conceptual blocks of predictor variables, and threat 5, which removed the student demographics from all conceptual blocks. For threats 2-5, we used the conditional mean imputation as used in the primary analysis. We examined the stability of the AUC values across models and compared them to the primary results.

Threat 1: Missing data. To address the missing data in our sample, we refit our models using multiple imputation (Graham, 2009). We imputed 100 datasets using a multivariate normal distribution and only imputed predictor variables. Overall, we found our results to be robust.

When predicting special education status, the linear probability and logistic results performed better than the primary results to the point where they mirrored the elastic net and random forest primary analysis results (Appendix B, Table 1, Panels A and B). For the chronic absenteeism models, the results were very similar to the primary analysis results (Appendix B, Table 2). Although a few models had an acceptable fit whereas none of the primary analysis models did, there was not a clear pattern to this. Because most of the multiple imputation models were robust to the mean conditional imputation models, we feel confident focusing on our main primary analytic approach.

Threat 2: Inclusion of students identified as receiving special education services in kindergarten. In the primary analytic approach for the models predicting receipt of special education services, we included all students who we had data for at baseline. This meant that students who were identified as receiving special education services during kindergarten were not included in the group receiving special education services in either first or second grade, i.e., they were counted as zero in the outcome variable. To address the possibility that the inclusion of these students was hindering the ability of the models to pick up the variation among the remaining students, we refit our special education models excluding the students (*N* = 66) who were identified as receiving special education services during kindergarten.

Overall, the refit models performed the same as the conditional mean imputation models except for the linear probability models, which performed better. Even though several of the linear probability models had an acceptable performance, the elastic net and random forest models performed the best, consistent with the main results. On average, the second-grade models performed better than the first-grade models, also consistent with the main results (Appendix B, Table 3). These results indicate that including the students who were identified as receiving special education services during kindergarten was an appropriate choice.

Threat 3: Sensitivity of using alternative definitions of chronic absenteeism. As Chang and Romero (2008) discuss, schools and districts use varying definitions of chronic absenteeism. While our primary analytic approach uses a definition of students missing at least 10% of school days, while enrolled a minimum of 90 days, refit our chronic absentee models using a definition of chronic absenteeism where students miss at least 15 school days with no minimum days of enrollment (Chang & Romero, 2008). We only did this for models predicting chronic absenteeism in kindergarten and first grade because the sample size of students missing at least 15 school days in second grade was too small due to the attendance data being truncated on March 14, 2020 to account for students' move to at-home schooling as a result of the COVID-19 pandemic. Using this definition of chronic absenteeism, *N* = 218 (21.5%) students were identified as being chronically absent in kindergarten while *N* = 159 (17.8%) students were identified as being chronically absent in first grade.

Overall, the models refit using this new definition of chronic absenteeism performed approximately the same in that all but one model had an unacceptable fit. The AUC values ranged between 0.543 and 0.690, and the only model that had an acceptable fit was the elastic model with every predictor group (AUC = 0.700, Appendix B, Table 4). Overall, these results indicate that model performance was not sensitive to the definition of chronic absenteeism used.

Threat 4: Saturation of predictor variables. Even though we built up our main analysis models using conceptual blocks of predictor variables in the chronological order that we thought district would have access to throughout the kindergarten school year, we also wanted to address the possibility that some blocks of predictor variables may be more predictive on their own instead of combined with other blocks. Overall, we found that most of the models fit with separate conceptual blocks of predictor variables performed the same or worse than the main analytic models. This was true for models predicting both receipt of special education services (Appendix B, Table 5) and chronic absenteeism (Appendix B, Table 6). Also, for both types of outcomes, there were different logistic models that converged compared to the main text models, such as the spring of kindergarten academic achievement and intra- and interpersonal skills measures. Interestingly, there were a few linear probability and logistic models that had an acceptable fit, including the spring of kindergarten assessment and teacher report measures. With these few well-fitting models as exceptions, we take these results as indication that our primary analytic models were an appropriate choice.

Threat 5: Sensitivity of using demographic variables. To address the possibility that our results were being driven by the demographic variables, we refit all our models removing the demographic variables. This is consistent with previous work on early warning systems that focus on attendance, behavior, and course performance variables instead of demographic variables (Therriault et al., 2017; U.S. Department of Education, 2016). Overall, we found our results to be robust to the exclusion of demographic variables with the exception that several of the special education linear probability models and the second-grade special education logistic models had an acceptable performance (Appendix B, Table 7). Consistent with the main results, none of the chronic absenteeism models had an acceptable performance (Appendix B, Table 8). On average, we take these results as indication that our primary analytic results were not driven by the inclusion of demographic variables.

Appendix B Table 1 (Addresses Threat 1)

AUC Values for Models Predicting Special Education Status using Multiple Imputation

Model type	First grade (1)	Second grade (2)
Panel A. Linear Probability Model		
Demographics	0.425	0.696
+ Fall of K academic/executive functioning	0.725	0.893
+ Fall of K intra/interpersonal	0.752	0.887
+ Family data	0.726	0.771
+ Teacher data	0.765	0.816
+ Spring of K academic/executive functioning	0.779	0.851
+ Spring of K intra/interpersonal	0.787	0.858
Panel B. Logistic Model		
Demographics	0.447	0.674
+ Fall of K academic/executive functioning	0.717	0.874
+ Fall of K intra/interpersonal	0.743	0.871
+ Family data	-	0.736
+ Teacher data	-	-
+ Spring of K academic/executive functioning	-	-
+ Spring of K intra/interpersonal	-	-
Panel C. Elastic Net		
Demographics	0.500	0.582
+ Fall of K academic/executive functioning	0.923	0.889
+ Fall of K intra/interpersonal	0.910	0.884
+ Family data	0.894	0.806
+ Teacher data	0.884	0.749
+ Spring of K academic/executive functioning	0.668	0.546
+ Spring of K intra/interpersonal	0.739	0.614
Panel D. Decision Tree		
Demographics	0.547	0.532
+ Fall of K academic/executive functioning	0.609	0.680
+ Fall of K intra/interpersonal	0.645	0.716
+ Family data	0.518	0.578
+ Teacher data	0.564	0.531
+ Spring of K academic/executive functioning	0.646	0.674
+ Spring of K intra/interpersonal	0.674	0.672
Panel E. Random Forest		
Demographics	0.503	0.683
+ Fall of K academic/executive functioning	0.774	0.827
+ Fall of K intra/interpersonal	0.805	0.867
+ Family data	0.759	0.846

+ Teacher data	0.771	0.889
+ Spring of K academic/executive functioning	0.793	0.935
+ Spring of K intra/interpersonal	0.793	0.934

Appendix B Table 2 (Addresses Threat 1)

AUC Values for Models Predicting Chronic Absenteeism using Multiple Imputation

Model type	Kindergarten	First grade	Second grade
	(1)	(2)	(3)
Panel A. Linear Probability Model			
Demographics	0.624	0.688	0.721
+ Fall of K academic/executive functioning	0.601	0.733	0.720
+ Fall of K intra/interpersonal	0.628	0.743	0.723
+ Family data	0.632	0.637	0.530
+ Teacher data	0.571	0.622	0.572
+ Spring of K academic/executive functioning	0.571	0.615	0.621
+ Spring of K intra/interpersonal	0.584	0.578	0.636
Panel B. Logistic Model			
Demographics	0.617	0.680	0.722
+ Fall of K academic/executive functioning	0.592	0.733	0.696
+ Fall of K intra/interpersonal	0.617	0.739	0.701
+ Family data	0.587	-	-
+ Teacher data	0.546	-	-
+ Spring of K academic/executive functioning	-	-	-
+ Spring of K intra/interpersonal	-	-	-
Panel C. Elastic Net			
Demographics	0.634	0.690	0.510
+ Fall of K academic/executive functioning	0.668	0.603	0.492
+ Fall of K intra/interpersonal	0.689	0.554	0.474
+ Family data	0.760	0.490	0.487
+ Teacher data	0.665	0.514	0.474
+ Spring of K academic/executive functioning	0.694	0.519	0.537
+ Spring of K intra/interpersonal	0.725	0.605	0.535
Panel D. Decision Tree			
Demographics	0.601	0.611	0.630
+ Fall of K academic/executive functioning	0.601	0.576	0.519
+ Fall of K intra/interpersonal	0.601	0.578	0.552
+ Family data	0.553	0.585	0.421
+ Teacher data	0.571	0.580	0.465
+ Spring of K academic/executive functioning	0.570	0.577	0.388
+ Spring of K intra/interpersonal	0.549	0.578	0.501
Panel E. Random Forest			
Demographics	0.622	0.637	0.630
+ Fall of K academic/executive functioning	0.675	0.701	0.644
+ Fall of K intra/interpersonal	0.621	0.671	0.700
+ Family data	0.731	0.774	0.764

+ Teacher data	0.714	0.699	0.803
+ Spring of K academic/executive functioning	0.747	0.723	0.863
+ Spring of K intra/interpersonal	0.739	0.701	0.842

Appendix B Table 3 (Addresses Threat 2)

AUC Values for Models Predicting Special Education Status without Students who Received Special Education Services in Kindergarten

Model type	First grade (1)	Second grade (2)
Panel A. Linear Probability Model		
Demographics	0.593	0.728
+ Fall of K academic/executive functioning	0.727	0.818
+ Fall of K intra/interpersonal	0.744	0.810
+ Family data	0.704	0.794
+ Teacher data	0.689	0.769
+ Spring of K academic/executive functioning	0.678	0.775
+ Spring of K intra/interpersonal	0.700	0.775
Panel B. Logistic Model		
Demographics	0.589	0.737
+ Fall of K academic/executive functioning	-	-
+ Fall of K intra/interpersonal	-	-
+ Family data	-	-
+ Teacher data	-	-
+ Spring of K academic/executive functioning	-	-
+ Spring of K intra/interpersonal	-	-
Panel C. Elastic Net		
Demographics	0.578	0.709
+ Fall of K academic/executive functioning	0.769	0.844
+ Fall of K intra/interpersonal	0.789	0.846
+ Family data	0.773	0.843
+ Teacher data	0.768	0.846
+ Spring of K academic/executive functioning	0.777	0.855
+ Spring of K intra/interpersonal	0.810	0.863
Panel D. Decision Tree		
Demographics	0.513	0.555
+ Fall of K academic/executive functioning	0.650	0.662
+ Fall of K intra/interpersonal	0.621	0.725
+ Family data	0.601	0.700
+ Teacher data	0.506	0.683
+ Spring of K academic/executive functioning	0.610	0.646
+ Spring of K intra/interpersonal	0.632	0.609
Panel E. Random Forest		
Demographics	0.626	0.671
+ Fall of K academic/executive functioning	0.824	0.844
+ Fall of K intra/interpersonal	0.807	0.858

+ Family data	0.787	0.839
+ Teacher data	0.779	0.828
+ Spring of K academic/executive functioning	0.804	0.844
+ Spring of K intra/interpersonal	0.820	0.862

Appendix B Table 4 (Addresses Threat 3)

AUC Values for Models Predicting Chronic Absenteeism with Alternative Definition of Chronic Absenteeism

Model type	Kindergarten (1)	First grade (2)
Panel A. Linear Probability Model		
Demographics	0.646	0.591
+ Fall of K academic/executive functioning	0.664	0.675
+ Fall of K intra/interpersonal	0.672	0.667
+ Family data	0.689	0.652
+ Teacher data	0.683	0.655
+ Spring of K academic/executive functioning	0.679	0.640
+ Spring of K intra/interpersonal	0.675	0.647
Panel B. Logistic Model		
Demographics	0.643	0.618
+ Fall of K academic/executive functioning	0.661	0.671
+ Fall of K intra/interpersonal	-	0.654
+ Family data	-	-
+ Teacher data	-	-
+ Spring of K academic/executive functioning	-	-
+ Spring of K intra/interpersonal	-	-
Panel C. Elastic Net		
Demographics	0.647	0.637
+ Fall of K academic/executive functioning	0.668	0.666
+ Fall of K intra/interpersonal	0.673	0.664
+ Family data	0.692	0.647
+ Teacher data	0.685	0.632
+ Spring of K academic/executive functioning	0.690	0.637
+ Spring of K intra/interpersonal	0.700	0.640
Panel D. Decision Tree		
Demographics	0.608	0.584
+ Fall of K academic/executive functioning	0.575	0.601
+ Fall of K intra/interpersonal	0.601	0.578
+ Family data	0.590	0.580
+ Teacher data	0.620	0.555
+ Spring of K academic/executive functioning	0.598	0.570
+ Spring of K intra/interpersonal	0.588	0.543
Panel E. Random Forest		
Demographics	0.626	0.647
+ Fall of K academic/executive functioning	0.645	0.638
+ Fall of K intra/interpersonal	0.643	0.618

+ Family data	0.664	0.586
+ Teacher data	0.676	0.605
+ Spring of K academic/executive functioning	0.681	0.622
+ Spring of K intra/interpersonal	0.686	0.617

Appendix B Table 5 (Addresses Threat 4)

AUC Values for Models Predicting Special Education Status with Separate Predictor Variable Blocks

Model type	First grade (1)	Second grade (2)
Panel A. Linear Probability Model		
Demographics	0.522	0.641
Fall of K academic/executive functioning	0.706	0.784
Fall of K intra/interpersonal	0.698	0.723
Family data	0.577	0.607
Teacher data	0.533	0.590
Spring of K academic/executive functioning	0.716	0.821
Spring of K intra/interpersonal	0.784	0.806
Panel B. Logistic Model		
Demographics	0.501	0.614
Fall of K academic/executive functioning	0.665	0.757
Fall of K intra/interpersonal	0.691	0.711
Family data	-	-
Teacher data	0.525	0.583
Spring of K academic/executive functioning	0.721	0.821
Spring of K intra/interpersonal	0.797	0.803
Panel C. Elastic Net		
Demographics	0.543	0.631
Fall of K academic/executive functioning	0.720	0.808
Fall of K intra/interpersonal	0.735	0.758
Family data	0.519	0.568
Teacher data	0.500	0.492
Spring of K academic/executive functioning	0.719	0.820
Spring of K intra/interpersonal	0.802	0.810
Panel D. Decision Tree		
Demographics	0.517	0.579
Fall of K academic/executive functioning	0.626	0.675
Fall of K intra/interpersonal	0.576	0.656
Family data	0.548	0.514
Teacher data	0.505	0.508
Spring of K academic/executive functioning	0.617	0.727
Spring of K intra/interpersonal	0.674	0.675
Panel E. Random Forest		
Demographics	0.563	0.649
Fall of K academic/executive functioning	0.720	0.800
Fall of K intra/interpersonal	0.658	0.648

Family data	0.527	0.545
Teacher data	0.553	0.584
Spring of K academic/executive functioning	0.762	0.841
Spring of K intra/interpersonal	0.765	0.798

Appendix B Table 6 (Addresses Threat 4)

AUC Values for Models Predicting Chronic Absenteeism with Separate Predictor Variable Blocks

Model type	Kindergarten (1)	First grade (2)	Second grade (3)
Panel A. Linear Probability Model			
Demographics	0.618	0.634	0.616
Fall of K academic/executive functioning	0.577	0.690	0.552
Fall of K intra/interpersonal	0.618	0.612	0.622
Family data	0.589	0.594	0.620
Teacher data	0.577	0.507	0.577
Spring of K academic/executive functioning	0.648	0.660	0.603
Spring of K intra/interpersonal	0.687	0.689	0.650
Panel B. Logistic Model			
Demographics	0.617	0.628	0.619
Fall of K academic/executive functioning	0.569	0.685	0.548
Fall of K intra/interpersonal	0.618	0.609	0.623
Family data	-	-	-
Teacher data	0.577	0.532	0.578
Spring of K academic/executive functioning	0.635	0.665	0.588
Spring of K intra/interpersonal	0.688	0.689	0.653
Panel C. Elastic Net			
Demographics	0.649	0.637	0.604
Fall of K academic/executive functioning	0.588	0.633	0.582
Fall of K intra/interpersonal	0.621	0.592	0.674
Family data	0.618	0.577	0.498
Teacher data	0.501	0.500	0.488
Spring of K academic/executive functioning	0.607	0.658	0.624
Spring of K intra/interpersonal	0.652	0.615	0.628
Panel D. Decision Tree			
Demographics	0.608	0.591	0.564
Fall of K academic/executive functioning	0.549	0.587	0.514
Fall of K intra/interpersonal	0.597	0.591	0.541
Family data	0.538	0.530	0.502
Teacher data	0.507	0.490	0.527
Spring of K academic/executive functioning	0.515	0.597	0.552
Spring of K intra/interpersonal	0.637	0.661	0.557
Panel E. Random Forest			
Demographics	0.661	0.647	0.554
Fall of K academic/executive functioning	0.571	0.575	0.558
Fall of K intra/interpersonal	0.583	0.576	0.581

Family data	0.577	0.587	0.553
Teacher data	0.649	0.537	0.567
Spring of K academic/executive functioning	0.600	0.605	0.593
Spring of K intra/interpersonal	0.619	0.595	0.632

Appendix B Table 7 (Addresses Threat 5)

AUC Values for Models Predicting Special Education Status without Demographic Variables

Model type	First grade (1)	Second grade (2)
Panel A. Linear Probability Model		
Fall of K academic/executive functioning	0.706	0.784
+ Fall of K intra/interpersonal	0.702	0.773
+ Family data	0.672	0.713
+ Teacher data	0.675	0.721
+ Spring of K academic/executive functioning	0.700	0.760
+ Spring of K intra/interpersonal	0.706	0.758
Panel B. Logistic Model		
Fall of K academic/executive functioning	0.665	0.757
+ Fall of K intra/interpersonal	0.639	0.732
+ Family data	-	-
+ Teacher data	-	-
+ Spring of K academic/executive functioning	-	-
+ Spring of K intra/interpersonal	-	-
Panel C. Elastic Net		
Fall of K academic/executive functioning	0.720	0.808
+ Fall of K intra/interpersonal	0.742	0.814
+ Family data	0.736	0.812
+ Teacher data	0.734	0.818
+ Spring of K academic/executive functioning	0.722	0.822
+ Spring of K intra/interpersonal	0.769	0.844
Panel D. Decision Tree		
Fall of K academic/executive functioning	0.626	0.675
+ Fall of K intra/interpersonal	0.525	0.716
+ Family data	0.589	0.606
+ Teacher data	0.515	0.627
+ Spring of K academic/executive functioning	0.585	0.603
+ Spring of K intra/interpersonal	0.634	0.667
Panel E. Random Forest		
Fall of K academic/executive functioning	0.720	0.800
+ Fall of K intra/interpersonal	0.743	0.815
+ Family data	0.687	0.797
+ Teacher data	0.722	0.797
+ Spring of K academic/executive functioning	0.739	0.826
+ Spring of K intra/interpersonal	0.770	0.844

Appendix B Table 8 (Addresses Threat 5)
AUC Values for Models Predicting Chronic Absenteeism without Demographic Variables

Model type	Kindergarten (1)	First grade (2)	Second grade (3)
Panel A. Linear Probability Model			
Fall of K academic/executive functioning	0.577	0.690	0.552
+ Fall of K intra/interpersonal	0.589	0.661	0.574
+ Family data	0.614	0.625	0.601
+ Teacher data	0.599	0.640	0.592
+ Spring of K academic/executive functioning	0.610	0.648	0.597
+ Spring of K intra/interpersonal	0.629	0.655	0.605
Panel B. Logistic Model			
Fall of K academic/executive functioning	0.569	0.685	0.548
+ Fall of K intra/interpersonal	0.589	0.640	0.559
+ Family data	-	-	-
+ Teacher data	-	-	-
+ Spring of K academic/executive functioning	-	-	-
+ Spring of K intra/interpersonal	-	-	-
Panel C. Elastic Net			
Fall of K academic/executive functioning	0.588	0.633	0.582
+ Fall of K intra/interpersonal	0.616	0.627	0.658
+ Family data	0.630	0.613	0.676
+ Teacher data	0.626	0.606	0.661
+ Spring of K academic/executive functioning	0.631	0.626	0.659
+ Spring of K intra/interpersonal	0.650	0.634	0.656
Panel D. Decision Tree			
Fall of K academic/executive functioning	0.549	0.587	0.514
+ Fall of K intra/interpersonal	0.529	0.563	0.582
+ Family data	0.622	0.523	0.522
+ Teacher data	0.588	0.538	0.537
+ Spring of K academic/executive functioning	0.581	0.547	0.566
+ Spring of K intra/interpersonal	0.559	0.533	0.545
Panel E. Random Forest			
Fall of K academic/executive functioning	0.571	0.575	0.558
+ Fall of K intra/interpersonal	0.566	0.609	0.632
+ Family data	0.592	0.589	0.602
+ Teacher data	0.557	0.596	0.630
+ Spring of K academic/executive functioning	0.607	0.614	0.616
+ Spring of K intra/interpersonal	0.614	0.618	0.616

Appendix C

Supplementary Tables and Figures for Paper 2

Appendix C Table 1

Descriptive Statistics for Predictor Variables

Variable	Mean/Percent	Standard deviation	Percent missing
Child demographics (Fall of K)			
Female	0.48	-	0%
Asian	0.14	-	0%
Black	0.25	-	0%
Latinx	0.33	-	0%
White	0.25	-	0%
Mixed/other race	0.03	-	0%
First language: English	0.63	-	0%
Home language: English	0.74	-	0%
Parent language preference: English	0.75	-	0%
Free or reduced-priced lunch	0.60	-	0%
Dual language learner	0.51	-	0%
Limited English proficient	0.39	-	0%
Age of Sept 1, 2017	5.49	0.29	0%
Direct assessments (Fall of K)			
PPVT (raw score)	91.05	26.80	6%
DIBELS FSF	16.47	12.66	23%
DIBELS LNF	25.52	17.17	23%
WJAP (raw score)	16.41	5.13	5%
REMA (raw score)	11.96	5.94	7%
FDS (categorical score)	3.58	0.97	7%
H&F mixed	0.71	0.21	10%
H&F incongruent	0.85	0.25	11%
PSRA AI	2.65	0.46	10%
PSRA PE	2.40	0.46	10%
Direct assessments (Spring of K)			
PPVT (raw score)	103.84	25.87	9%
DIBELS LNF	52.83	18.54	22%
DIBELS PSF	43.53	17.64	22%
DIBELS CLS	41.02	26.47	22%

DIBELS WWR	7.64	11.07	22%
WJAP (raw score)	19.53	4.77	9%
REMA (raw score)	16.51	8.23	9%
FDS (categorical score)	3.80	0.89	9%
BDS (categorical score)	2.53	0.76	24%
H&F mixed	0.77	0.21	11%
H&F incongruent	0.91	0.20	11%
PSRA AI	2.65	0.46	10%
PSRA PE	2.40	0.46	10%
Teacher report assessments (Fall of K)			
Academic orientation	3.42	1.08	19%
Cooperation	3.15	0.68	18%
Engagement	3.23	0.56	18%
Self-control	3.07	0.68	19%
Externalizing behavior	1.50	0.51	19%
Internalizing behavior	1.43	0.44	19%
Hyperactivity/inattention	1.79	0.65	19%
Teacher report assessments (Spring of K)			
Academic orientation	3.57	1.10	14%
Cooperation	3.22	0.65	13%
Engagement	3.34	0.53	13%
Self-control	3.13	0.67	13%
Externalizing behavior	1.49	0.51	13%
Internalizing behavior	1.44	0.43	13%
Hyperactivity/inattention	1.75	0.64	13%
Family data			
Parent ed: High school diploma or less	0.31	-	51%
Parent ed: Two-year degree	0.31	-	51%
Parent ed: Bachelor's degree	0.17	-	51%
Parent ed: Advanced degree	0.21	-	51%
Age of mother at first child's birth	26.19	6.65	53%
Age of father at first child's birth	24.50	13.36	57%
Parent age when completing survey	36.27	7.09	
Household size	4.28	1.37	51%
At least one adult in household working full time	0.89	-	51%
Married/partner	0.55	-	58%
Child had Early Intervention Services or Individualized Family Service Plan	0.16	-	58%
PreK experience: BPS PreK	0.64	-	34%
PreK experience: Something other than BPS PreK	0.24	-	34%

PreK experience: No center-based care	0.12	-	34%
Parental perception that PreK attendance is important	0.82	-	52%
Parental perception that daily school attendance is important	0.83	-	59%
Parent satisfaction with school assignment	3.58	0.67	60%
Parental engagement: Literacy constrained	2.98	0.73	52%
Parental engagement: Language unconstrained	3.02	0.66	51%
Parental engagement: Math constrained	2.79	0.70	52%
Parental engagement: Math unconstrained	2.47	0.76	52%
Household income: Less than $25000	0.31	-	55%
Household income: Between $2500-$59999	0.32	-	55%
Household income: $60000 or more	0.37	-	55%
Experiential learning activities	0.24	0.17	50%
Number of children’s books at home (including library books)	5.53	3.67	52%
Teacher data			
Years of teaching experience	13.11	7.51	9%
Years of Kindergarten teaching experience	9.10	0.25	14%
Female	0.93	-	14%
White	0.59	-	14%
Black	0.14	-	14%
Asian	0.04	-	11%
Latinx	0.23	-	8%
Age	39.05	9.19	11%
Highest education: Education specialist/professional diploma	0.03	-	8%
Highest education: Associate’s or other	0.01	-	8%
Highest education: Bachelor’s	0.11	-	8%
Highest education: Master’s	0.84	-	8%
Highest education: Doctorate	0.01	-	8%
Major of highest degree: Early childhood education	0.43	-	8%
Major of highest degree: Elementary education	0.25	-	8%
Major of highest degree: Special education	0.32	-	10%
Major of highest degree: Child development	0.04	-	8%
Major of highest degree: Reading specialist	0.07	-	10%
Major of highest degree: Curriculum & instruction	0.07	-	10%
Major of highest degree: Bilingual/bicultural education	0.04	-	10%
Major of highest degree: Other education	0.13	-	8%
Major of highest degree: Other non-education	0.02	-	10%
Current teaching license: Early childhood education	0.94	-	8%
Current teaching license: Elementary education	0.42	-	8%
Current teaching license: English Language Learners	0.54	-	8%

Current teaching license: Speech	0.00	-	10%
Current teaching license: Teacher of students with moderate disabilities	0.49	-	8%
Current teaching license: Teacher of students with severe disabilities	0.03	-	10%
Current teaching license: Other teacher	0.03	-	10%
Current teaching license: Teacher specialist	0.03	-	10%
Current teaching license: Administrator	0.04	-	10%

Notes: *N* = 1,012. PPVT = Peabody Picture Vocabulary Test, FSF = First sound fluency score, LNF = Letter naming raw score, PSF = Phoneme segmentation fluency score, CLS = Nonsense word fluency correct letter sounds score, WWR = Nonsense word fluency whole word read score, WJAP = Woodcock Johnson Applied Problems, REMA = Research-Based Early Mathematics Assessment, FDS = Digit Span Forward, BDS = Digit Span Backward, H&F = Hearts and Flowers, PSRA = Preschool Self-Regulation Assessment, AI = Attention/Impulse Control, PE = Positive Emotion. Major of highest degree and currently teaching license for teachers are not mutually exclusive to allow for dual degree and license.

Appendix C Table 2

AUC Values for Models Predicting Special Education Status for Dual Language Learners

Model type	First grade (1)	Second grade (2)
Panel A. Linear Probability Model		
Demographics	0.629	0.655
+ Fall of K academic/executive functioning	0.632	0.795
+ Fall of K intra/interpersonal	0.652	0.773
+ Family data	0.729	0.663
+ Teacher data	0.689	0.695
+ Spring of K academic/executive functioning	0.731	0.701
+ Spring of K intra/interpersonal	0.727	0.709
Panel B. Logistic Model		
Demographics	0.593	0.656
+ Fall of K academic/executive functioning	0.614	0.782
+ Fall of K intra/interpersonal	-	-
+ Family data	-	-
+ Teacher data	-	-
+ Spring of K academic/executive functioning	-	-
+ Spring of K intra/interpersonal	-	-
Panel C. Elastic Net		
Demographics	0.576	0.642
+ Fall of K academic/executive functioning	0.667	0.838
+ Fall of K intra/interpersonal	0.701	0.832
+ Family data	0.690	0.827
+ Teacher data	0.692	0.831
+ Spring of K academic/executive functioning	0.686	0.847
+ Spring of K intra/interpersonal	0.731	0.851
Panel D. Decision Tree		
Demographics	0.513	0.563
+ Fall of K academic/executive functioning	0.543	0.743
+ Fall of K intra/interpersonal	0.492	0.670
+ Family data	0.555	0.648
+ Teacher data	0.533	0.704
+ Spring of K academic/executive functioning	0.517	0.645
+ Spring of K intra/interpersonal	0.599	0.682
Panel E. Random Forest		
Demographics	0.628	0.677
+ Fall of K academic/executive functioning	0.657	0.868
+ Fall of K intra/interpersonal	0.711	0.843
+ Family data	0.631	0.815

+ Teacher data	0.625	0.822
+ Spring of K academic/executive functioning	0.679	0.841
+ Spring of K intra/interpersonal	0.648	0.875

Appendix C Table 3

AUC Values for Models Predicting Special Education Status for non-Dual Language Learners

Model type	First grade (1)	Second grade (2)
Panel A. Linear Probability Model		
Demographics	0.562	0.659
+ Fall of K academic/executive functioning	0.711	0.779
+ Fall of K intra/interpersonal	0.734	0.794
+ Family data	0.654	0.765
+ Teacher data	0.669	0.789
+ Spring of K academic/executive functioning	0.672	0.798
+ Spring of K intra/interpersonal	0.670	0.773
Panel B. Logistic Model		
Demographics	0.578	0.646
+ Fall of K academic/executive functioning	0.674	0.753
+ Fall of K intra/interpersonal	-	-
+ Family data	-	-
+ Teacher data	-	-
+ Spring of K academic/executive functioning	-	-
+ Spring of K intra/interpersonal	-	-
Panel C. Elastic Net		
Demographics	0.513	0.619
+ Fall of K academic/executive functioning	0.742	0.792
+ Fall of K intra/interpersonal	0.771	0.812
+ Family data	0.765	0.795
+ Teacher data	0.769	0.808
+ Spring of K academic/executive functioning	0.729	0.812
+ Spring of K intra/interpersonal	0.793	0.834
Panel D. Decision Tree		
Demographics	0.516	0.624
+ Fall of K academic/executive functioning	0.574	0.717
+ Fall of K intra/interpersonal	0.568	0.603
+ Family data	0.456	0.517
+ Teacher data	0.524	0.519
+ Spring of K academic/executive functioning	0.544	0.727
+ Spring of K intra/interpersonal	0.503	0.582
Panel E. Random Forest		
Demographics	0.606	0.621
+ Fall of K academic/executive functioning	0.753	0.753
+ Fall of K intra/interpersonal	0.754	0.776
+ Family data	0.727	0.767

+ Teacher data	0.747	0.786
+ Spring of K academic/executive functioning	0.746	0.778
+ Spring of K intra/interpersonal	0.801	0.804

Appendix C Table 4

AUC Values for Models Predicting Chronic Absenteeism for Dual Language Learners

Model type	Kindergarten (1)	First grade (2)	Second grade (3)
Panel A. Linear Probability Model			
Demographics	0.546	0.595	0.628
+ Fall of K academic/executive functioning	0.566	0.688	0.560
+ Fall of K intra/interpersonal	0.583	0.677	0.647
+ Family data	0.600	0.652	0.634
+ Teacher data	0.596	0.675	0.662
+ Spring of K academic/executive functioning	0.577	0.694	0.639
+ Spring of K intra/interpersonal	0.581	0.688	0.642
Panel B. Logistic Model			
Demographics	0.543	0.597	0.627
+ Fall of K academic/executive functioning	0.568	0.696	0.548
+ Fall of K intra/interpersonal	0.587	0.691	-
+ Family data	-	-	-
+ Teacher data	-	-	-
+ Spring of K academic/executive functioning	-	-	-
+ Spring of K intra/interpersonal	-	-	-
Panel C. Elastic Net			
Demographics	0.637	0.600	0.678
+ Fall of K academic/executive functioning	0.619	0.658	0.611
+ Fall of K intra/interpersonal	0.647	0.648	0.636
+ Family data	0.633	0.619	0.628
+ Teacher data	0.614	0.606	0.600
+ Spring of K academic/executive functioning	0.615	0.602	0.603
+ Spring of K intra/interpersonal	0.647	0.580	0.609
Panel D. Decision Tree			
Demographics	0.556	0.531	0.561
+ Fall of K academic/executive functioning	0.543	0.614	0.446
+ Fall of K intra/interpersonal	0.548	0.601	0.567
+ Family data	0.604	0.595	0.568
+ Teacher data	0.605	0.662	0.644
+ Spring of K academic/executive functioning	0.641	0.598	0.663
+ Spring of K intra/interpersonal	0.593	0.568	0.602
Panel E. Random Forest			
Demographics	0.670	0.698	0.592
+ Fall of K academic/executive functioning	0.616	0.561	0.634
+ Fall of K intra/interpersonal	0.628	0.570	0.657
+ Family data	0.594	0.556	0.632

+ Teacher data	0.592	0.597	0.632
+ Spring of K academic/executive functioning	0.577	0.561	0.625
+ Spring of K intra/interpersonal	0.550	0.535	0.618

Appendix C Table 5

AUC Values for Models Predicting Chronic Absenteeism for non-Dual Language Learners

Model type	Kindergarten (1)	First grade (2)	Second grade (3)
Panel A. Linear Probability Model			
Demographics	0.680	0.660	0.635
+ Fall of K academic/executive functioning	0.668	0.693	0.581
+ Fall of K intra/interpersonal	0.676	0.670	0.587
+ Family data	0.643	0.634	0.629
+ Teacher data	0.618	0.619	0.638
+ Spring of K academic/executive functioning	0.616	0.614	0.632
+ Spring of K intra/interpersonal	0.660	0.594	0.654
Panel B. Logistic Model			
Demographics	0.681	0.645	0.633
+ Fall of K academic/executive functioning	0.658	0.661	0.582
+ Fall of K intra/interpersonal	0.665	0.613	-
+ Family data	-	-	-
+ Teacher data	-	-	-
+ Spring of K academic/executive functioning	-	-	-
+ Spring of K intra/interpersonal	-	-	-
Panel C. Elastic Net			
Demographics	0.674	0.639	0.585
+ Fall of K academic/executive functioning	0.650	0.674	0.626
+ Fall of K intra/interpersonal	0.656	0.675	0.681
+ Family data	0.663	0.671	0.726
+ Teacher data	0.663	0.653	0.718
+ Spring of K academic/executive functioning	0.657	0.675	0.702
+ Spring of K intra/interpersonal	0.655	0.687	0.727
Panel D. Decision Tree			
Demographics	0.659	0.608	0.577
+ Fall of K academic/executive functioning	0.611	0.551	0.500
+ Fall of K intra/interpersonal	0.571	0.536	0.546
+ Family data	0.538	0.546	0.562
+ Teacher data	0.539	0.512	0.580
+ Spring of K academic/executive functioning	0.517	0.537	0.589
+ Spring of K intra/interpersonal	0.518	0.523	0.527
Panel E. Random Forest			
Demographics	0.648	0.603	0.599
+ Fall of K academic/executive functioning	0.614	0.663	0.562
+ Fall of K intra/interpersonal	0.609	0.689	0.573
+ Family data	0.636	0.658	0.592

+ Teacher data	0.620	0.618	0.580
+ Spring of K academic/executive functioning	0.625	0.634	0.611
+ Spring of K intra/interpersonal	0.639	0.660	0.615

Appendix C Table 6

AUC Values for Models Predicting Special Education and Chronic Absenteeism Status with District-Collected Data

Model type	Special education		Chronically absent		
	First grade	Second grade	Kindergarten	First grade	Second grade
	(1)	(2)	(3)	(4)	(5)
Panel A. Linear Probability Model					
Demographics + Fall of K DIBELS	0.651	0.802	0.641	0.660	0.597
+ Spring of K DIBELS	0.694	0.811	0.669	0.651	0.604
Panel B. Logistic Model					
Demographics + Fall of K DIBELS	0.665	0.814	0.637	0.653	0.604
+ Spring of K DIBELS	0.701	0.819	0.656	0.649	0.604
Panel C. Elastic Net					
Demographics + Fall of K DIBELS	0.673	0.821	0.635	0.650	0.621
+ Spring of K DIBELS	0.698	0.828	0.633	0.661	0.624
Panel D. Decision Tree					
Demographics + Fall of K DIBELS	0.657	0.733	0.542	0.560	0.542
+ Spring of K DIBELS	0.620	0.702	0.607	0.534	0.475
Panel E. Random Forest					
Demographics + Fall of K DIBELS	0.645	0.798	0.650	0.632	0.594
+ Spring of K DIBELS	0.704	0.806	0.593	0.617	0.578

Appendix C Table 7

Youden Statistic Values for Models Predicting Special Education and Chronic Absenteeism Status with District-Collected Data

Model type	Special education		Chronically absent		
	First grade	Second grade	Kindergarten	First grade	Second grade
	(1)	(2)	(3)	(4)	(5)
Panel A. Linear Probability Model					
Demographics	0.029	0.058	0.173	0.139	0.078
+ Fall of K DIBELS	0.065	0.104	0.164	0.096	0.076
+ Spring of K DIBELS	0.075	0.124	0.156	0.117	0.100
Panel B. Logistic Model					
Demographics	0.040	0.079	0.160	0.119	0.074
+ Fall of K DIBELS	0.052	0.062	0.143	0.106	0.081
+ Spring of K DIBELS	0.024	0.060	0.129	0.089	0.087
Panel C. Elastic Net					
Demographics	0.043	0.062	0.166	0.136	0.085
+ Fall of K DIBELS	0.058	0.098	0.148	0.133	0.093
+ Spring of K DIBELS	0.068	0.127	0.151	0.130	0.089
Panel D. Decision Tree					
Demographics	0.176	0.110	0.134	0.142	0.069
+ Fall of K DIBELS	0.036	0.041	0.137	0.091	0.140
+ Spring of K DIBELS	0.048	0.064	0.128	0.240	0.220
Panel E. Random Forest					
Demographics	0.060	0.054	0.140	0.125	0.068
+ Fall of K DIBELS	0.047	0.114	0.189	0.091	0.087
+ Spring of K DIBELS	0.075	0.136	0.220	0.111	0.061

Appendix C Table 8

Confusion Matrices Based on Youden Statistic Values for Models Predicting Special Education Status in First Grade with District-Collected Data

			Model 1		Model 2		Model 3	
Model type			Actual		Actual		Actual	
			Not SPED	SPED	Not SPED	SPED	Not SPED	SPED
Panel A. Linear Probability Model								
	Predicted	Not SPED	80	5	111	3	179	5
		SPED	112	5	81	7	13	5
Panel B. Logistic Model								
	Predicted	Not SPED	87	6	186	7	167	5
		SPED	105	4	6	3	25	5
Panel C. Elastic Net								
	Predicted	Not SPED	45	1	125	4	177	6
		SPED	146	10	66	7	14	5
Panel D. Decision Tree								
	Predicted	Not SPED	191	9	164	6	182	5
		SPED	1	1	28	4	10	5
Panel E. Random Forest								
	Predicted	Not SPED	105	7	147	4	171	4
		SPED	87	3	45	6	21	6

Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments, SPED = received special education services.

Appendix C Table 9

Confusion Matrices Based on Youden Statistic Values for Models Predicting Special Education Status in Second Grade with District-Collected Data

			Model 1		Model 2		Model 3	
Model type			Actual		Actual		Actual	
			Not SPED	SPED	Not SPED	SPED	Not SPED	SPED
Panel A. Linear Probability Model								
	Predicted	Not SPED	79	0	170	5	181	4
		SPED	111	12	20	7	9	8
Panel B. Logistic Model								
	Predicted	Not SPED	178	7	163	4	166	3
		SPED	12	5	27	8	24	9
Panel C. Elastic Net								
	Predicted	Not SPED	96	2	116	2	113	5
		SPED	88	16	68	16	71	13
Panel D. Decision Tree								
	Predicted	Not SPED	189	11	173	9	180	5
		SPED	1	1	17	3	10	7
Panel E. Random Forest								
	Predicted	Not SPED	71	1	173	5	171	3
		SPED	119	11	17	7	19	9

Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments, SPED = received special education services.

Appendix C Table 10

Confusion Matrices Based on Youden Statistic Values for Models Predicting Chronic Absenteeism in Kindergarten with District-Collected Data

			Model 1		Model 2		Model 3	
Model type			Actual		Actual		Actual	
			Not absent	Absent	Not absent	Absent	Not absent	Absent
Panel A. Linear Probability Model								
	Predicted	Not absent	99	10	122	11	130	12
		Absent	63	21	40	20	32	19
Panel B. Logistic Model								
	Predicted	Not absent	100	10	123	12	135	14
		Absent	62	21	39	19	27	17
Panel C. Elastic Net								
	Predicted	Not absent	110	10	114	11	102	9
		Absent	56	20	52	19	64	21
Panel D. Decision Tree								
	Predicted	Not absent	98	13	98	13	99	13
		Absent	64	18	64	18	63	18
Panel E. Random Forest								
	Predicted	Not absent	101	11	98	8	123	10
		Absent	61	20	64	23	39	21

Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Table 11

Confusion Matrices Based on Youden Statistic Values for Models Predicting Chronic Absenteeism in First Grade with District-Collected Data

Model type			Model 1		Model 2		Model 3	
			Actual		Actual		Actual	
			Not absent	Absent	Not absent	Absent	Not absent	Absent
Panel A. Linear Probability Model								
	Predicted	Not absent	112	12	125	13	117	11
		Absent	34	14	21	13	29	15
Panel B. Logistic Model								
	Predicted	Not absent	87	8	105	9	104	9
		Absent	59	18	41	17	42	17
Panel C. Elastic Net								
	Predicted	Not absent	70	3	95	5	91	4
		Absent	82	21	57	19	61	20
Panel D. Decision Tree								
	Predicted	Not absent	109	15	119	16	142	22
		Absent	37	11	27	19	4	4
Panel E. Random Forest								
	Predicted	Not absent	100	10	126	16	128	15
		Absent	46	16	20	10	18	11

Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Table 12

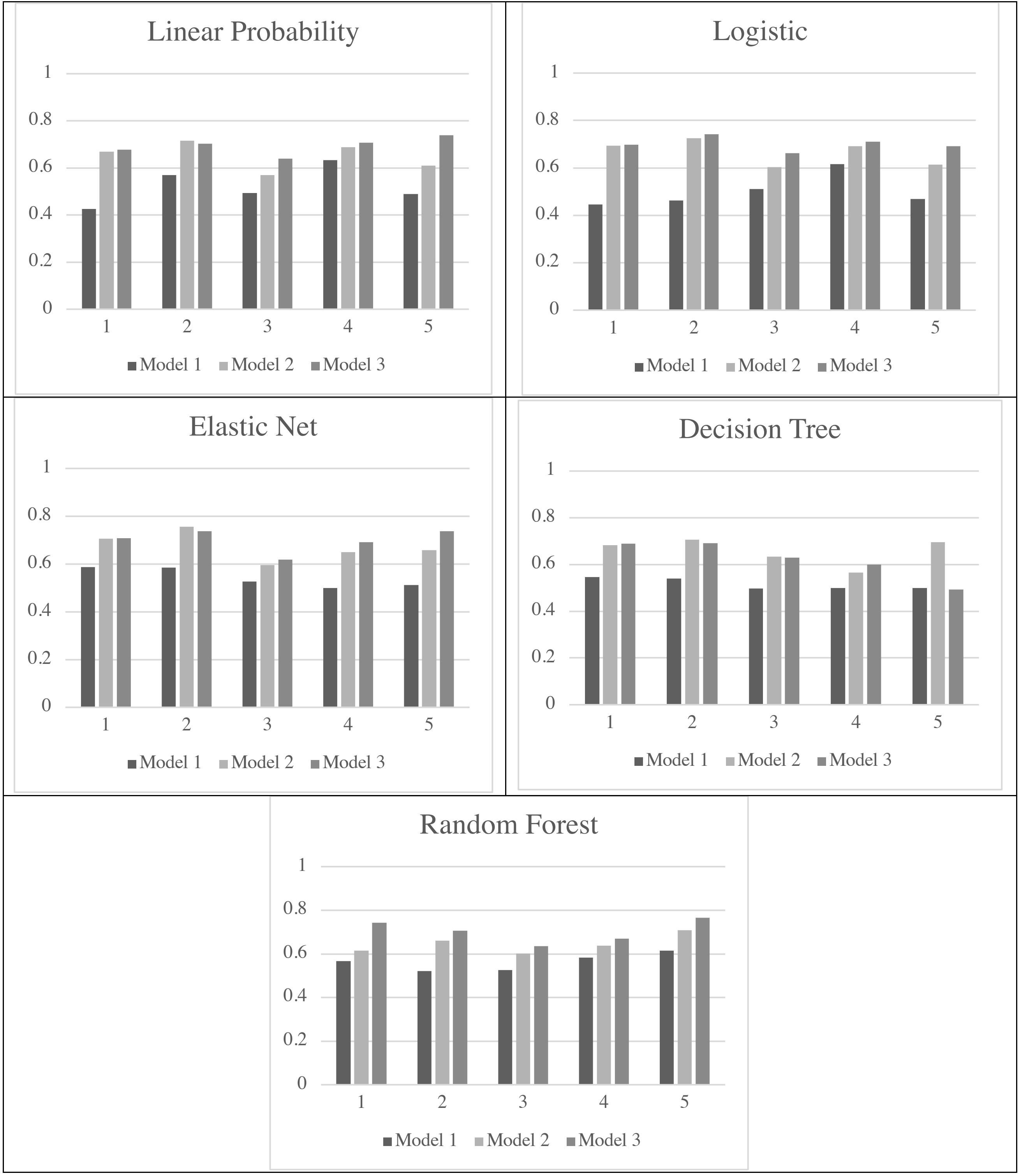
Confusion Matrices Based on Youden Statistic Values for Models Predicting Chronic Absenteeism in Second Grade with District-Collected Data

			Model 1		Model 2		Model 3	
Model type			Actual		Actual		Actual	
			Not absent	Absent	Not absent	Absent	Not absent	Absent
Panel A. Linear Probability Model								
	Predicted	Not absent	89	2	95	2	108	4
		Absent	52	11	46	11	33	9
Panel B. Logistic Model								
	Predicted	Not absent	102	3	93	2	104	4
		Absent	39	10	48	11	37	9
Panel C. Elastic Net								
	Predicted	Not absent	66	0	99	3	79	2
		Absent	81	14	48	11	68	12
Panel D. Decision Tree								
	Predicted	Not absent	61	2	61	2	122	7
		Absent	80	11	80	11	19	6
Panel E. Random Forest								
	Predicted	Not absent	54	1	39	0	52	1
		Absent	87	12	102	13	89	12

Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Figure 1

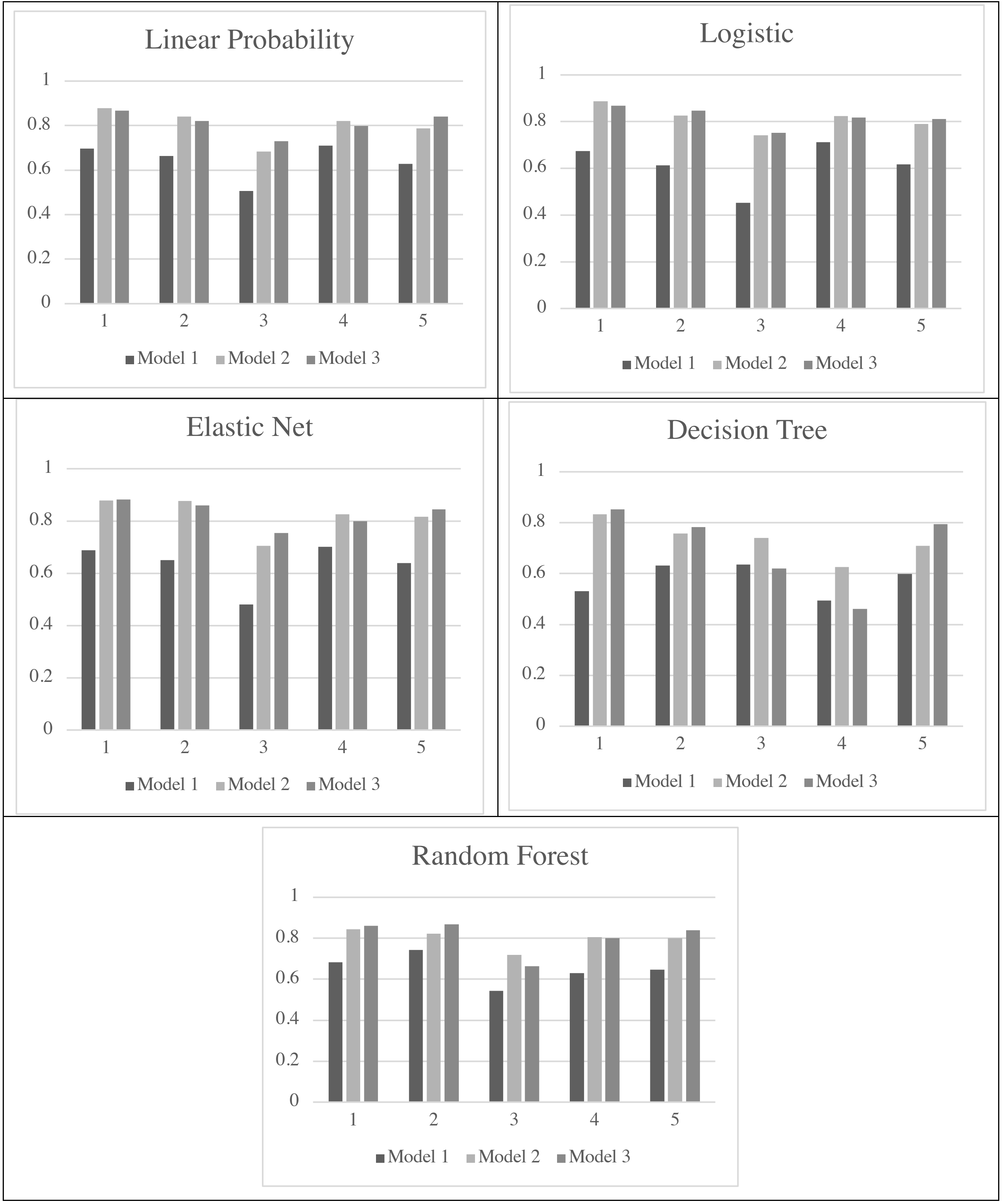
Distribution of AUC Values for Models Predicting First Grade Special Education Status



Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Figure 2

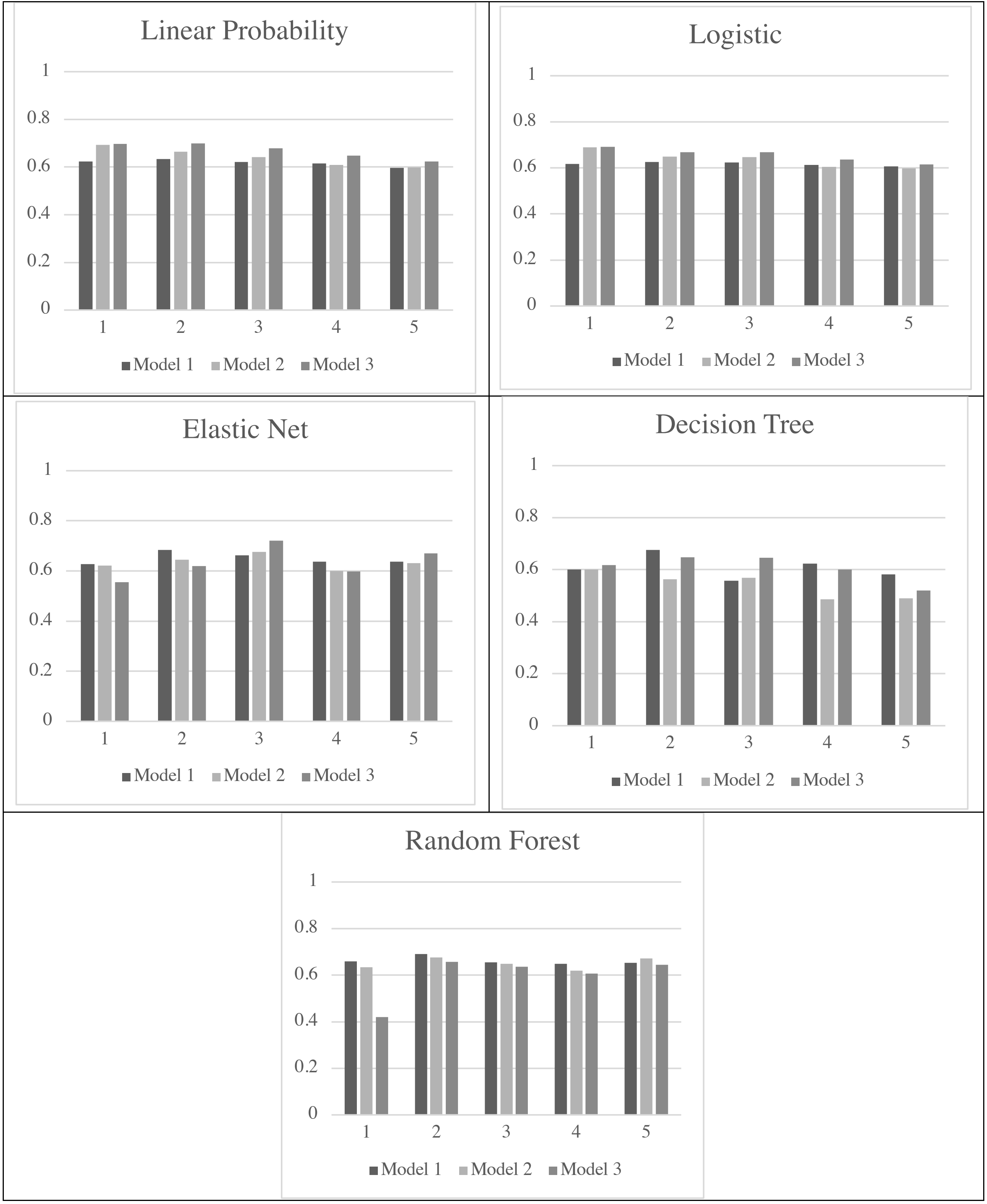
Distribution of AUC Values for Models Predicting Second Grade Special Education Status



Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Figure 3

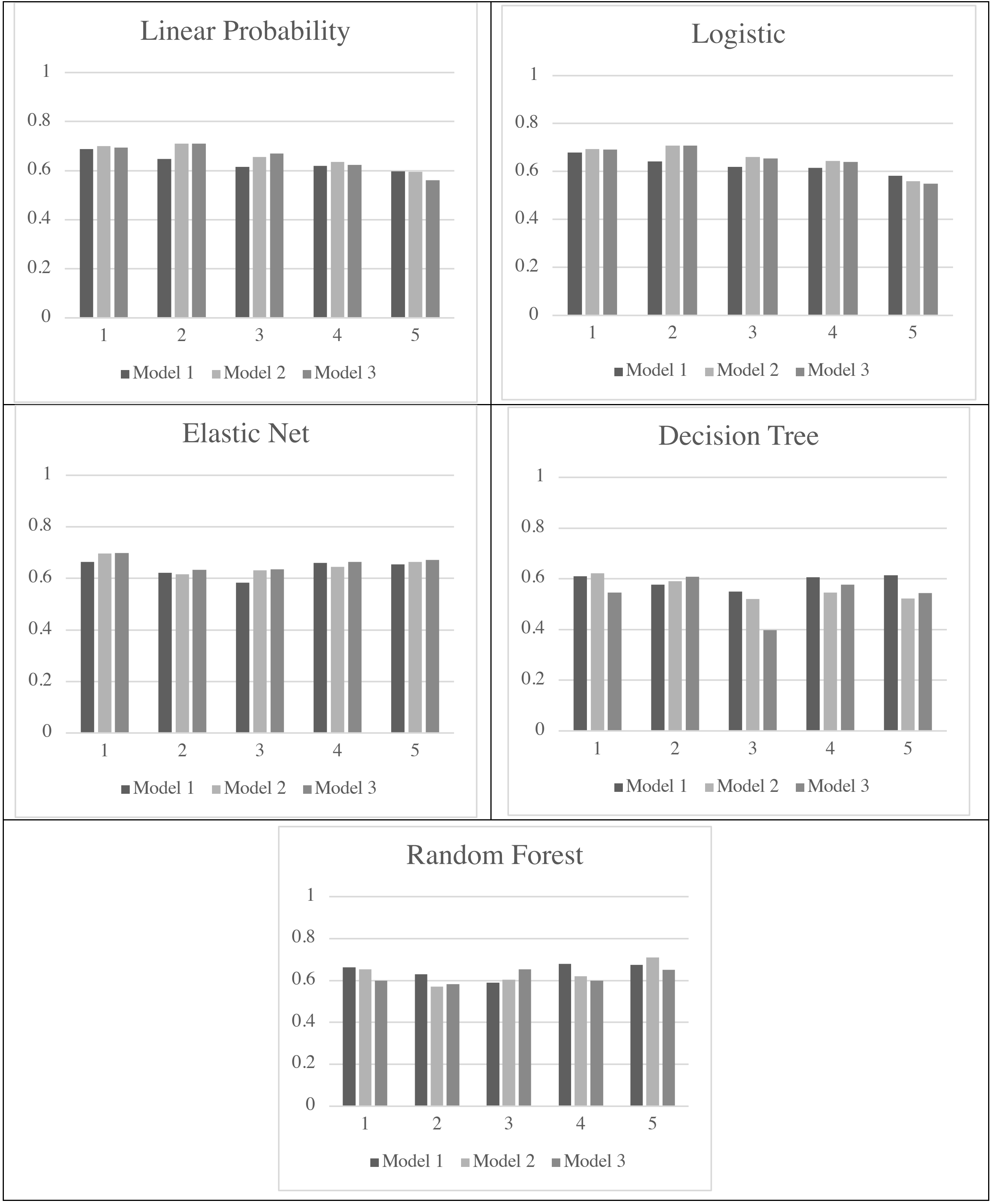
Distribution of AUC Values for Models Predicting Kindergarten Chronic Absenteeism



Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Figure 4

Distribution of AUC Values for Models Predicting First Grade Chronic Absenteeism



Linear Probability

Logistic

Elastic Net

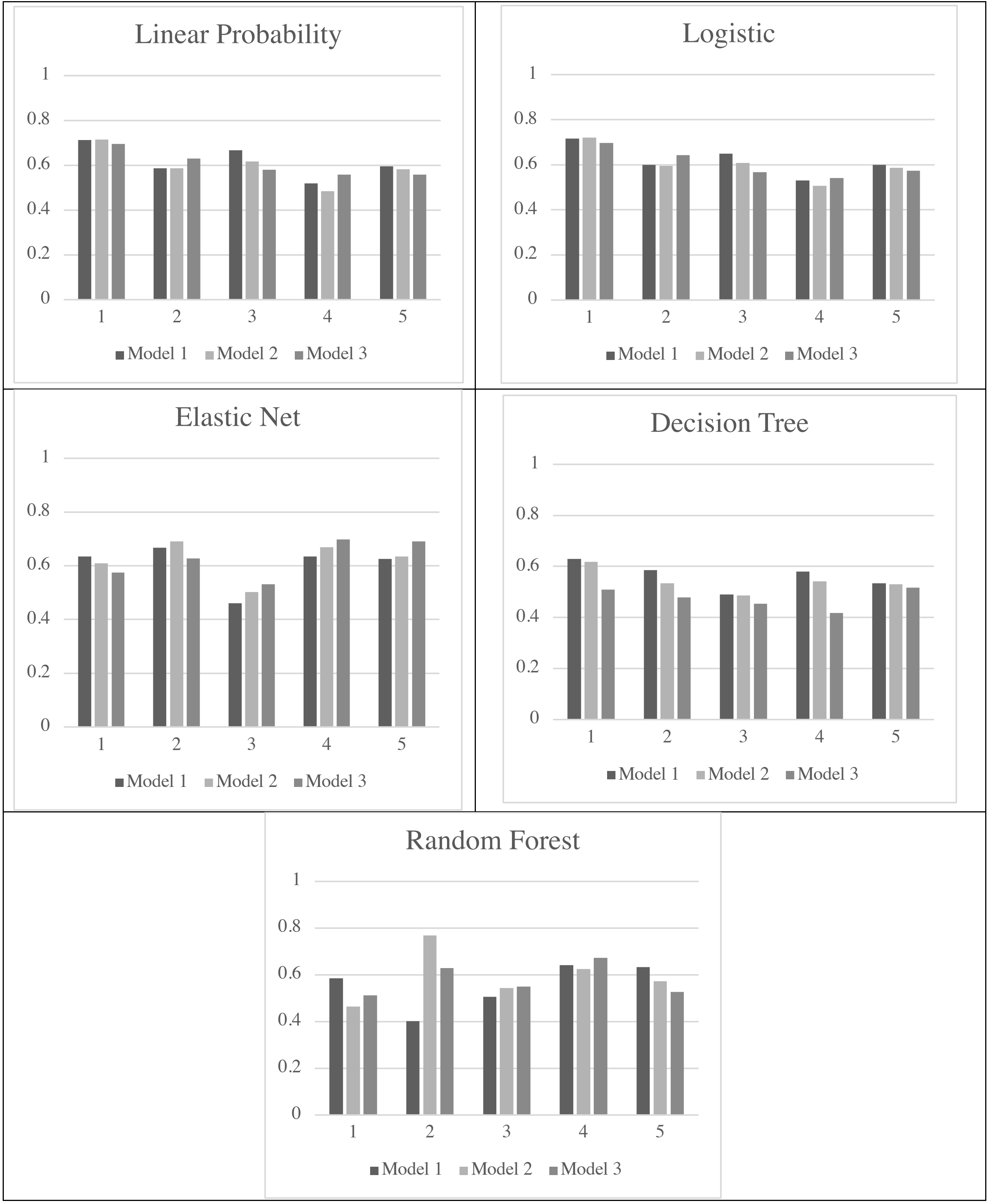
Decision Tree

Random Forest

Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Figure 5

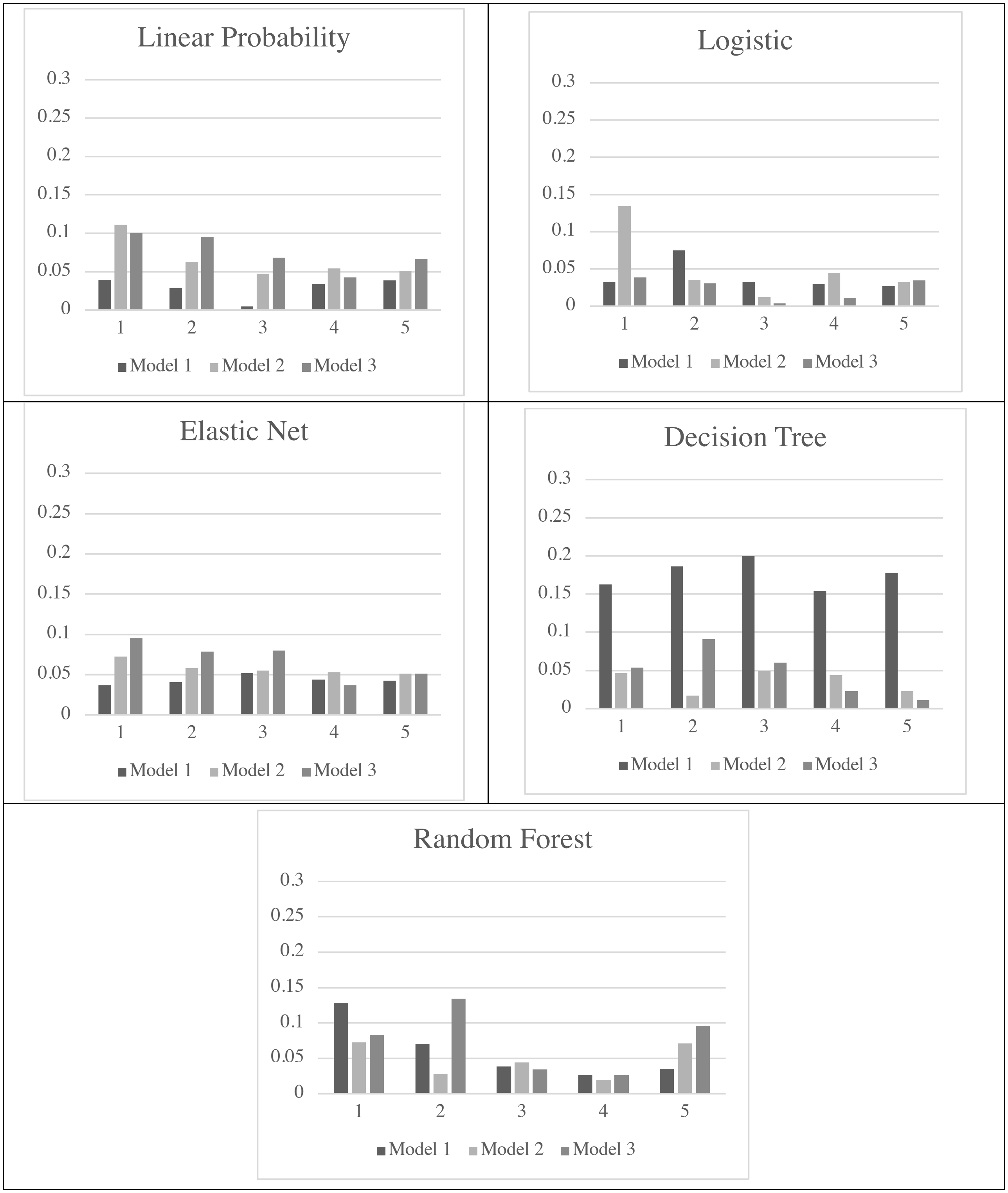
Distribution of AUC Values for Models Predicting Second Grade Chronic Absenteeism



Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Figure 6

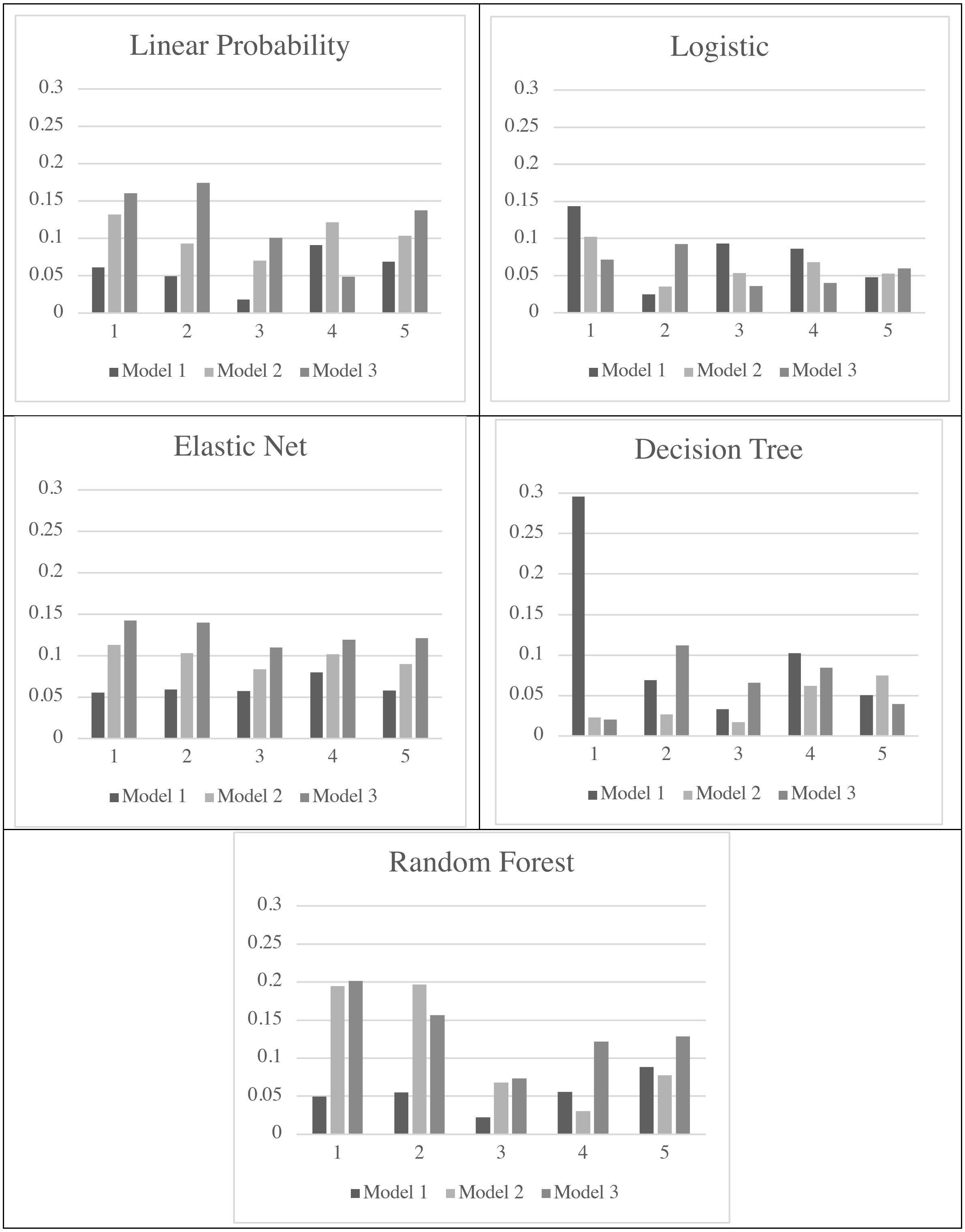
Distribution of Youden Statistic Values for Models Predicting First Grade Special Education Status



Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Figure 7

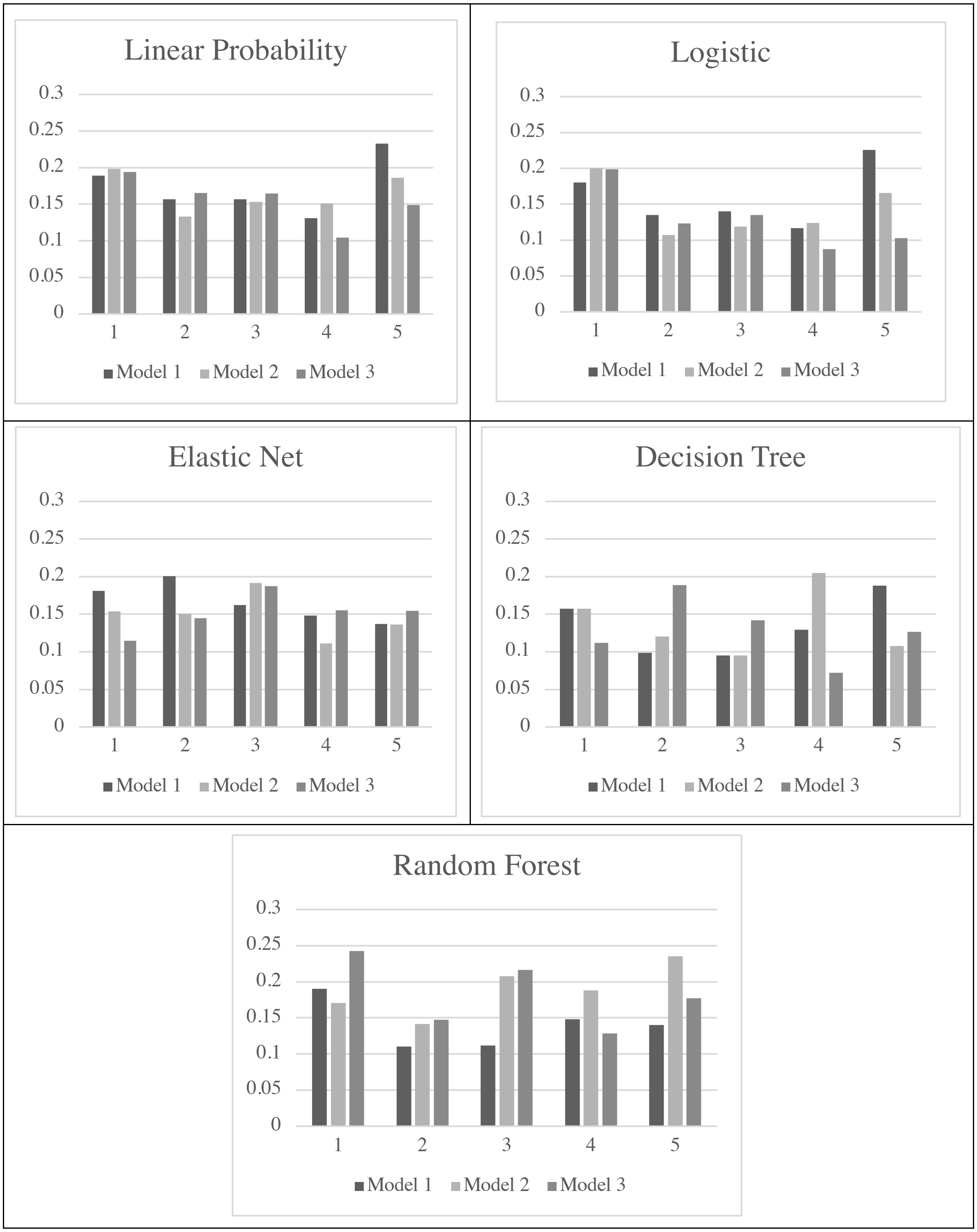
Distribution of Youden Statistic Values for Models Predicting Second Grade Special Education Status



Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Figure 8

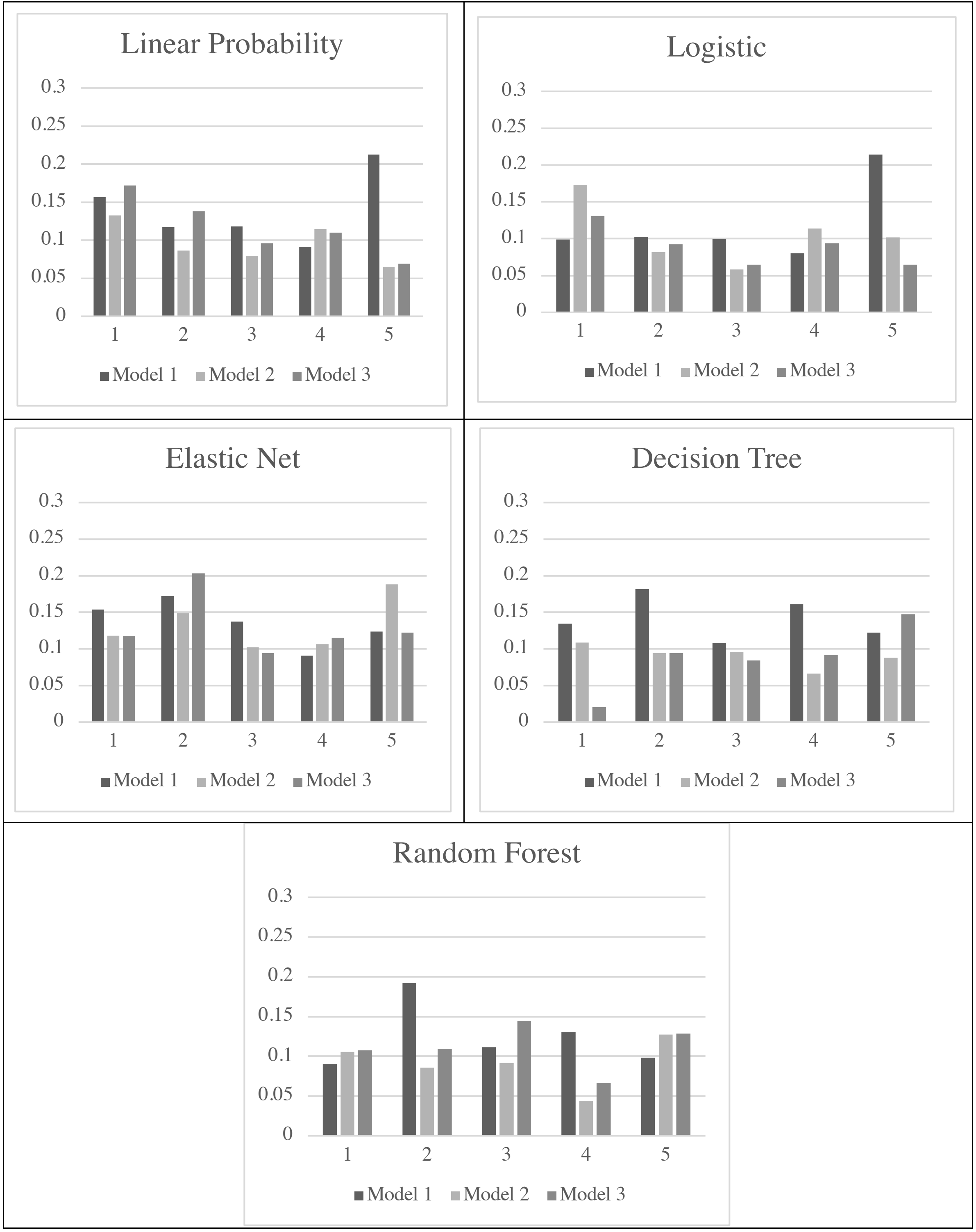
Distribution of Youden Statistic Values for Models Predicting Kindergarten Chronic Absenteeism



Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Figure 9

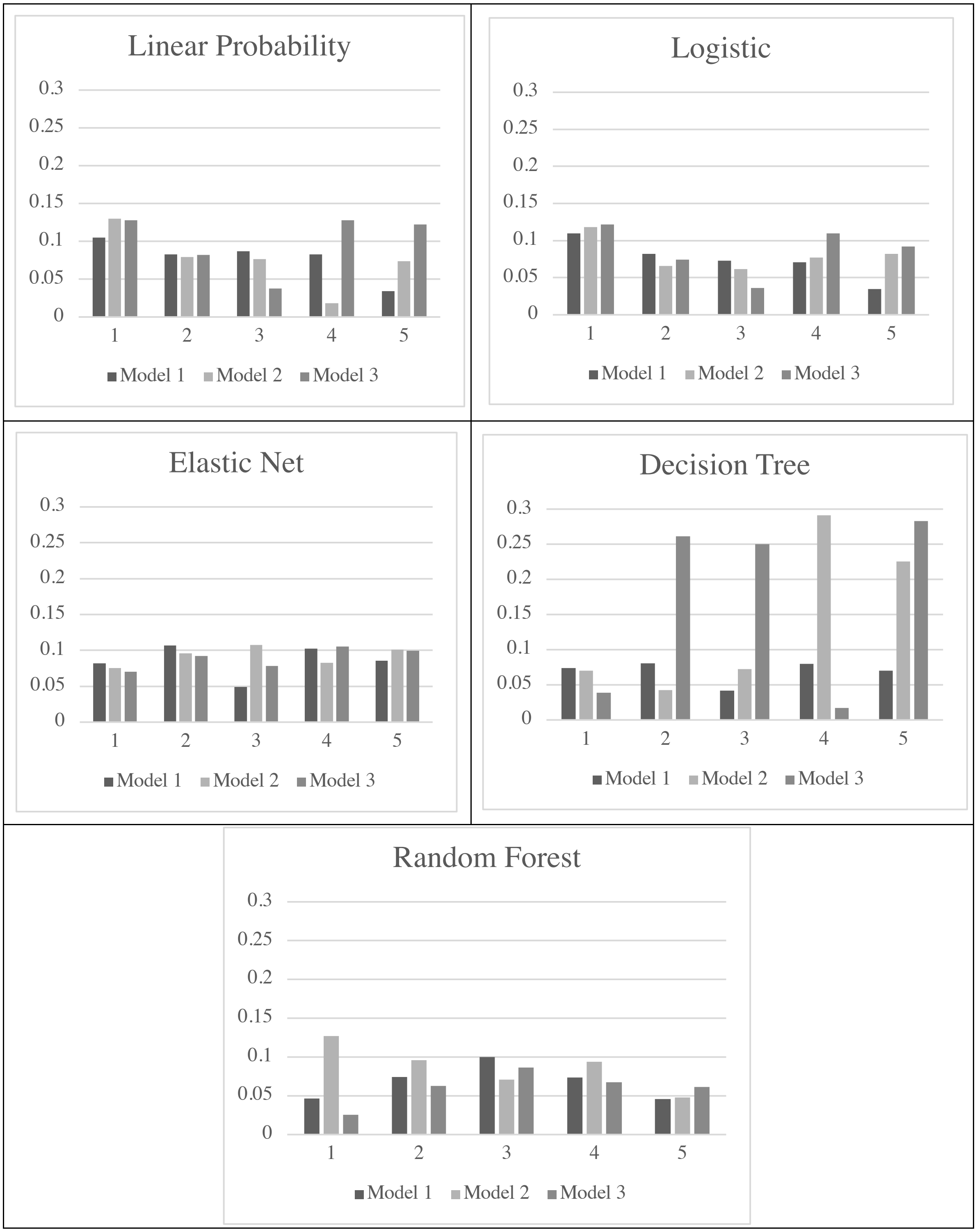
Distribution of Youden Statistic Values for Models Predicting First Grade Chronic Absenteeism



Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix C Figure 10

Distribution of Youden Statistic Values for Models Predicting Second Grade Chronic Absenteeism



Notes: Model 1 = child demographics from administrative data, Model 2 = Model 1 with fall of K DIBELS assessments, Model 3 = Model 2 with spring of K DIBELS assessments.

Appendix D

Overview of Robustness Checks for Paper 3

Threats to the robustness of my results include: 1) results being driven by student demographics, 2) saturation of predictor variables, 3) unstable numbers of students across grades, and 4) differential model performance based on urbanicity. Overall, I found the main results to be robust to these threats. In all models, just as in the primary analytic strategy, I used five model types (linear probability, logistic, elastic net, decision tree, and random forest) with seven conceptual blocks of predictor variables (Block 1 = fourth grade measures, Block 2 = Block 1 with all fifth grade measures, Block 3 = Block 2 with all sixth grade measures, Block 4 = Block 3 with all seventh grade measures, Block 5 = Block 4 with all eighth grade measures, Block 6 = Block 5 with all ninth grade measures, Block 7 = Block 6 with all tenth grade measures, and Block 8 = Block 7 with demographics). The exceptions to this were the models addressing threat 1 that used only demographic information and threat 2 that were fit with separate blocks of predictors. I used conditional mean imputation for all models to address missingness. I examined the stability of the AUC values across models and compared them to the primary results.

Threat 1: Results being driven by student demographics. To address the possibility that the results were driven by student demographics instead of the attendance, behavior, and academic performance variables, I refit the main models using only the demographic variables. I found that only one model had an acceptable performance – the logistic model – but this model did not perform as well as the main model fit with the fourth and fifth grade data (Appendix D, Table 1). These results indicate that the main models with the full range of predictor variables represent the best approach.

Threat 2: Saturation of predictor variables. Even though I built up the main analysis models using conceptual blocks of predictor variables in the chronological order that school personnel would have access to that data, I also wanted to address the possibility that some blocks of predictor variables may be more predictive on their own instead of combined with other blocks. To examine this, I refit the models using separate conceptual blocks of predictors (i.e., one grade per model). Overall, I found that although the models beginning with the fifth grade data had an acceptable performance, they performed worse than the main models fit with multiple years' worth of data (Appendix D, Table 2). This indicates that while the main model specification is preferrable for optimal model performance, it is possible to build models with acceptable performance using a single grade's worth of data if that is all that schools have access to in practice.

Threat 3: Unstable numbers of students across grades. In the primary analysis, I choose to include any student who had at least one value of one predictor variable if they had a graduation record to simulate the circumstances under which school personnel would be using an early warning system. However, from a methodological standpoint, this instability and the resulting imputation that had to take place to address missingness could introduce error into the models. To address this, I refit the main models using only the 312,194 students who were present in the data for every grade for every year, i.e., fourth through tenth grade. The dropout rate for this subset of students was 7%, lower than the 11% in the full sample. I found that when restricting my sample to only students who were present in each year, the models performed better by an average AUC value of 0.02 compared to the models fit with the full sample

(Appendix D, Table 3). These results indicate that the results may be partially being driven by error introduced during imputation to account for different numbers of students across grades.

Threat 4: Differential model performance based on urbanicity. Michigan is a large and diverse state with differential graduation rates based on urbanicity. Using definitions of urbanicity as defined by the National Center for Education Statistics, I used administrative data to determine the urbanicity of the school from where students were expected to graduate for 99% of my sample and found that 23% of the sample attended school in a city, 44% in a suburb, 11% in a town, and 22% in a rural area (Geverdt, 2018). The students who lived attended a school in a city had a 14% dropout rate compared to 10% for a suburb, 10% for a town, and 8% for a rural area. To address the possibility that the model may not perform well for every geographic area (Gebru, 2021; Kantayya, 2020), I examined model performance for students in each of the four geographic areas. The model for the students graduating from a school in a city and town performed slightly worse than the full sample while the model for those graduating from a school in a suburb or rural area performed the same as or slightly better than the full sample model (Appendix D, Table 4). This indicates that care should be taken when constructing a predictive model that covers students from a wide range of geographic areas.

Appendix D Table 1 (Addresses Threat 1)
AUC Values for Models Fit with Demographic Data Only

Linear probability	Logistic	Elastic net	Decision tree	Random forest
(1)	(2)	(3)	(4)	(5)
0.692	0.701	0.692	0.659	0.696

Notes: $N = 416{,}105$.

Appendix D Table 2 (Addresses Threat 2)
AUC Values for Models Fit with Separate Data by Grade

Model type	Linear probability (1)	Logistic (2)	Elastic net (3)	Decision tree (4)	Random forest (5)
G4 data	0.696	0.696	0.696	0.609	0.699
+ G5 data	0.707	0.707	0.707	0.616	0.712
+ G6 data	0.734	0.734	0.734	0.642	0.737
+ G7 data	0.749	0.746	0.749	0.704	0.750
+ G8 data	0.772	0.767	0.773	0.724	0.780
+ G9 data	0.787	0.777	0.787	0.739	0.790
+ G10 data	0.739	0.739	0.739	0.695	0.781

Notes: *N* = 416,105. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Appendix D Table 3 (Addresses Threat 3)
AUC Values for Models with Students Present in Every Year

Model type	Linear probability (1)	Logistic (2)	Elastic net (3)	Decision tree (4)	Random forest (5)
G4 data	0.723	0.718	0.723	0.629	0.716
+ G5 data	0.741	0.734	0.741	0.637	0.729
+ G6 data	0.759	0.753	0.759	0.643	0.755
+ G7 data	0.776	0.769	0.776	0.666	0.777
+ G8 data	0.797	0.791	0.798	0.724	0.800
+ G9 data	0.821	0.814	0.822	0.745	0.825
+ G10 data	0.838	0.832	0.839	0.721	0.849
+ Demographics	0.851	0.850	0.851	0.721	0.858

Notes: *N* = 312,194. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Appendix D Table 4 (Addresses Threat 4)
AUC Values for Models with Based on Urbanicity

Model type	Linear probability	Logistic	Elastic net	Decision tree	Random forest
	(1)	(2)	(3)	(4)	(5)
Panel A. City					
G4 data	0.681	0.681	0.681	0.605	0.686
+ G5 data	0.697	0.697	0.697	0.609	0.695
+ G6 data	0.726	0.726	0.726	0.646	0.729
+ G7 data	0.745	0.743	0.746	0.700	0.750
+ G8 data	0.768	0.764	0.768	0.709	0.771
+ G9 data	0.793	0.789	0.794	0.729	0.803
+ G10 data	0.820	0.815	0.821	0.734	0.836
+ Demographics	0.831	0.828	0.831	0.734	0.842
Panel B. Suburb					
G4 data	0.708	0.708	0.708	0.612	0.710
+ G5 data	0.728	0.726	0.727	0.620	0.730
+ G6 data	0.752	0.749	0.752	0.646	0.760
+ G7 data	0.772	0.767	0.772	0.713	0.783
+ G8 data	0.794	0.787	0.795	0.737	0.804
+ G9 data	0.815	0.808	0.815	0.736	0.829
+ G10 data	0.831	0.825	0.831	0.728	0.848
+ Demographics	0.841	0.837	0.841	0.728	0.855
Panel C. Town					
G4 data	0.663	0.665	0.663	0.592	0.665
+ G5 data	0.682	0.684	0.682	0.596	0.684
+ G6 data	0.706	0.708	0.706	0.615	0.718
+ G7 data	0.723	0.724	0.724	0.670	0.735
+ G8 data	0.753	0.748	0.754	0.696	0.767
+ G9 data	0.783	0.776	0.784	0.701	0.800
+ G10 data	0.815	0.804	0.817	0.705	0.836
+ Demographics	0.831	0.823	0.833	0.705	0.844
Panel D. Rural					
G4 data	0.685	0.686	0.685	0.600	0.687
+ G5 data	0.709	0.709	0.709	0.607	0.710
+ G6 data	0.732	0.732	0.732	0.626	0.735
+ G7 data	0.753	0.751	0.753	0.690	0.761
+ G8 data	0.777	0.772	0.777	0.713	0.782
+ G9 data	0.804	0.798	0.805	0.717	0.816
+ G10 data	0.817	0.812	0.818	0.715	0.837
+ Demographics	0.830	0.827	0.831	0.715	0.845

Notes: *N* = 93,178 for students who graduated from a school in a city, *N* = 183,222 for students who graduated from a school in a suburb, *N* = 47,273 for students who graduated from a school in a town, *N* = 89,268 for students who graduated from a rural school. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Appendix E

Supplementary Tables for Paper 3

Appendix E Table 1

Youden Statistic, Specificity, and Sensitivity Values for Models with Full Sample

Model type	Linear probability (1)	Logistic (2)	Elastic net (3)	Decision tree (4)	Random forest (5)
Panel A. Youden Statistic					
G4 data	0.091	0.078	0.091	0.096	0.087
+ G5 data	0.086	0.075	0.085	0.103	0.083
+ G6 data	0.096	0.075	0.095	0.087	0.107
+ G7 data	0.093	0.075	0.093	0.080	0.086
+ G8 data	0.101	0.077	0.100	0.075	0.093
+ G9 data	0.104	0.076	0.101	0.080	0.113
+ G10 data	0.099	0.079	0.099	0.137	0.112
+ Demographics	0.108	0.087	0.107	0.137	0.105
Panel B. Specificity					
G4 data	0.570	0.563	0.570	0.597	0.587
+ G5 data	0.556	0.567	0.551	0.633	0.568
+ G6 data	0.614	0.595	0.612	0.710	0.655
+ G7 data	0.617	0.613	0.617	0.609	0.610
+ G8 data	0.658	0.631	0.653	0.598	0.641
+ G9 data	0.704	0.659	0.693	0.645	0.711
+ G10 data	0.718	0.727	0.716	0.830	0.731
+ Demographics	0.718	0.742	0.716	0.830	0.717
Panel C. Sensitivity					
G4 data	0.735	0.741	0.735	0.646	0.715
+ G5 data	0.779	0.765	0.784	0.622	0.760
+ G6 data	0.761	0.773	0.764	0.556	0.714
+ G7 data	0.784	0.780	0.784	0.755	0.793
+ G8 data	0.775	0.792	0.781	0.788	0.799
+ G9 data	0.764	0.798	0.776	0.753	0.777
+ G10 data	0.784	0.764	0.787	0.611	0.801
+ Demographics	0.808	0.771	0.812	0.611	0.824

Notes: *N* = 416,105. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Appendix E Table 2

Confusion Matrices Based on Youden Statistic Values for Models with Full Sample for Linear Probability Models

Model type			Actual	
			Not Dropout	Dropout
G4 data	Predicted	Not Dropout	66551	2664
		Dropout	50191	7401
+ G5 data	Predicted	Not Dropout	64860	2223
		Dropout	51882	7842
+ G6 data	Predicted	Not Dropout	71712	2406
		Dropout	45030	7659
+ G7 data	Predicted	Not Dropout	71994	2174
		Dropout	44748	7891
+ G8 data	Predicted	Not Dropout	76792	2262
		Dropout	39950	7803
+ G9 data	Predicted	Not Dropout	82237	2374
		Dropout	34505	7691
+ G10 data	Predicted	Not Dropout	83844	2176
		Dropout	32898	7889
+ Demographics	Predicted	Not Dropout	83865	1930
		Dropout	32877	8135

Notes: *N* = 416,105. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Appendix E Table 3

Confusion Matrices Based on Youden Statistic Values for Models with Full Sample for Logistic Models

Model type			Actual	
			Not Dropout	Dropout
G4 data	Predicted	Not Dropout	65752	2610
		Dropout	50990	7455
+ G5 data	Predicted	Not Dropout	66228	2364
		Dropout	50514	7701
+ G6 data	Predicted	Not Dropout	69412	2281
		Dropout	47330	7784
+ G7 data	Predicted	Not Dropout	71512	2210
		Dropout	45230	7855
+ G8 data	Predicted	Not Dropout	73654	2090
		Dropout	43088	7975
+ G9 data	Predicted	Not Dropout	76929	2037
		Dropout	39813	8028
+ G10 data	Predicted	Not Dropout	84820	2371
		Dropout	31922	7694
+ Demographics	Predicted	Not Dropout	86677	2302
		Dropout	30065	7763

Notes: *N* = 416,105. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Appendix E Table 4

Confusion Matrices Based on Youden Statistic Values for Models with Full Sample for Elastic Net Models

Model type			Actual	
			Not Dropout	Dropout
G4 data	Predicted	Not Dropout	66556	2664
		Dropout	50186	7401
+ G5 data	Predicted	Not Dropout	64280	2173
		Dropout	52462	7892
+ G6 data	Predicted	Not Dropout	71402	2375
		Dropout	45340	7690
+ G7 data	Predicted	Not Dropout	72015	2171
		Dropout	44727	7894
+ G8 data	Predicted	Not Dropout	76246	2207
		Dropout	40496	7858
+ G9 data	Predicted	Not Dropout	80908	2250
		Dropout	35834	7815
+ G10 data	Predicted	Not Dropout	83629	2141
		Dropout	33113	7924
+ Demographics	Predicted	Not Dropout	83560	1895
		Dropout	33182	8170

Notes: *N* = 416,105. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Appendix E Table 5

Confusion Matrices Based on Youden Statistic Values for Models with Full Sample for Decision Tree Models

Model type			Actual	
			Not Dropout	Dropout
G4 data	Predicted	Not Dropout	69699	3576
		Dropout	47043	6498
+ G5 data	Predicted	Not Dropout	73891	3800
		Dropout	42851	6265
+ G6 data	Predicted	Not Dropout	82930	4469
		Dropout	33812	5596
+ G7 data	Predicted	Not Dropout	71078	2462
		Dropout	45664	7603
+ G8 data	Predicted	Not Dropout	69848	2129
		Dropout	46894	7936
+ G9 data	Predicted	Not Dropout	75244	2487
		Dropout	41498	7578
+ G10 data	Predicted	Not Dropout	96949	3913
		Dropout	19793	6152
+ Demographics	Predicted	Not Dropout	96949	3913
		Dropout	19793	6152

Notes: *N* = 416,105. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.

Appendix E Table 6

Confusion Matrices Based on Youden Statistic Values for Models with Full Sample for Random Forest Models

Model type			Actual	
			Not Dropout	Dropout
G4 data	Predicted	Not Dropout	68550	2869
		Dropout	48192	7196
+ G5 data	Predicted	Not Dropout	66362	2417
		Dropout	50380	7648
+ G6 data	Predicted	Not Dropout	76446	2880
		Dropout	40296	7185
+ G7 data	Predicted	Not Dropout	71197	2088
		Dropout	45545	7977
+ G8 data	Predicted	Not Dropout	74781	2022
		Dropout	41961	8043
+ G9 data	Predicted	Not Dropout	83036	2249
		Dropout	33706	7816
+ G10 data	Predicted	Not Dropout	85390	2007
		Dropout	31352	8058
+ Demographics	Predicted	Not Dropout	83721	1774
		Dropout	33021	8291

Notes: *N* = 416,105. G4 = fourth grade, G5 = fifth grade, G6 = sixth grade, G7 = seventh grade, G8 = eighth grade, G9 = ninth grade, G10 = tenth grade.