Deeper Roots Before the Storm

Utilizing Machine Learning to Alert School Districts of Permanent School Closures

Michael L. Chrzan, Francis A. Pearman, Benjamin W. Domingue

Stanford Graduate School of Education

Abstract

The increasing rate of permanent school closures in U.S. public school districts presents unprecedented challenges for administrators and communities alike. This study develops an early-warning indicator model to predict mass closure events - defined as a district closing at least 10% of its schools - five years in advance. Leveraging administrative data from the National Center for Education Statistics from 2000-2018, we evaluated a suite of supervised machine learning models - including elastic-net regularized logistic regression, random forests, XGBoost, LSTM neural networks, and SuperLearner ensembles - to determine the degree to which they could predict mass closures using enrollment, financial, and demographic predictors. Comparative analysis based on Area Under the Precision–Recall Curve (AUC-PR), and Recall revealed that XGBoost provided predictive accuracy while effectively handling class imbalance. Our findings demonstrate the technical feasibility of using advanced analytics in educational settings and also offer a glimpse into their potential for generating actionable insights for policymakers to proactively manage resources and support equitable decision-making in the face of systemic challenges. To this end, we include a case study of The School District of Philadelphia's mass school closures between 2012 and 2013 which this model predicts in 2007 which includes recommendations districts could use based on our predictions.

1. Introduction

School districts across the United States are facing a crisis. With rises in chronic absenteeism and declining enrollment since the pandemic, declining birth rates nationally, and the proliferation of school choice policies, public school districts across the nation – particularly districts in urban areas – are finding themselves in need of reorganizing their resources, namely their school portfolios (Peetz 2024; Bingamon 2024; Alejo 2024).

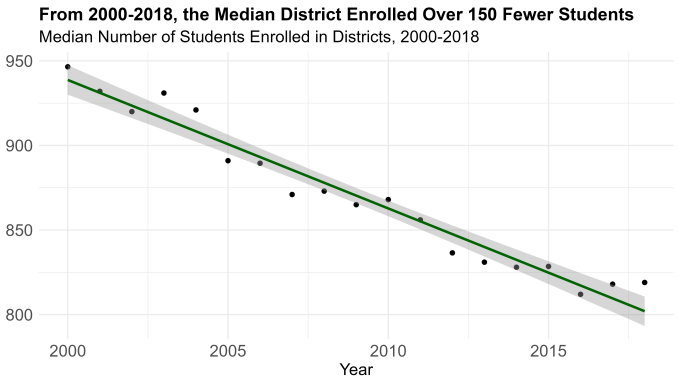

Figure 1: **Median number of students enrolled in all U.S. school districts from 2000-2018** according to NCES agency data. A linear regression line is fit to the data and displayed to further demonstrate the clear, decreasing pattern and the strength of the pattern over the 18 years between 2000 and 2018.

Many districts have schools that are underenrolled, buildings in dire need of repair, and students who need resources that are too dispersed to be effective, such as social work and mental health services, particularly after the COVID-19 pandemic (Ewing and Green 2022). Using National Center for Education Statistics data, the median U.S. School District lost between 150 and 200 students between 2000 and 2018 (see **Figure 1**), which - by certain estimates of the funding per student across states (Lieberman 2023) - would result in a loss of around $2.9 million per district. As each of these crises continues to worsen, more and more districts across the nation will be faced with the difficult and complex task of deciding how to reorganize and the consequences that will stem from deciding which schools to close, making it imperative to develop methods to accurately predict risks of closures. Districts need guidance in making these crucial and contentious decisions. This research attempts to support that guidance effort by creating

an early-warning model to help districts better prepare for a potential reorganization.

1.1 Prior Work

Research thus far has attempted to help districts understand the impacts of school closure to inform their decisions. This research has provided mixed findings into the impact of school closures on the students and communities impacted by them. With regard to student impacts, some studies show no significant effects (or even positive effects) on student achievement, while others suggest negative consequences for students, particularly students displaced by their school closing (Torre and Gwynne 2009; Brummet 2014; Billger 2010; Kirshner, Gaertner, and Pozzoboni 2010). For example, Engberg et. al. has shown that displaced students often experience a temporary decline in achievement but typically return to their pre-closure trajectory (Engberg et al. 2012). Other research by Steinberg and MacDonald has shown displaced students may face increased absences and suspensions, particularly if they have to travel longer distances to new schools (Steinberg and MacDonald 2019).

Research has also examined the effects of closure beyond students to try to understand how to make these choices. In studying the impact of school closures on teachers in Chicago, Lee and Sartain found that school closures can lead to teacher layoffs and rehires, which can potentially affect teacher retention and quality within a district (Lee and Sartain 2020). Pearman and Green found that closures of schools in historically marginalized communities can exacerbate gentrification and spatial inequality (Pearman and Greene 2022) and Gordon discussed how the opposite directionality can also be true - where once occupied school buildings can become vacant and underutilized, further contributing to neighborhood decline rather than gentrification (Gordon 2014).

Furthermore, literature in education, economics, and other fields have examined the process of closing schools, as opposed to the impact after the fact. This is of particular importance considering how contentious the process can become. This research finds that school closure decisions often stem from stated rationale – such as declining enrollment, financial constraints, low academic performance – and unstated rationale – such as racialized policies and charter school expansion – or, most commonly, combinations of these factors (Engberg et al. 2012; Weber, Farmer, and Donoghue 2020). The process typically involves district-wide planning, data analysis to identify target schools, and the development of a closure plan. In certain cases, these plans also incorporate stakeholder engagement strategies and will detail practices to lessen the potential harmful effects of closures on students (Sunderman and Payne 2009). This literature has been instrumental in identifying best practices districts can implement during the closure process. It emphasizes stakeholder engagement throughout the process, providing clear information to affected communities, and offering support to displaced students and staff. It also names the crucialness of considering long-term implications, including impacts on the neighborhood and overall community cohesion (Ewing 2018; Hahnel and Marchitello 2023).

1.2 Research Question

It's clear from this research on closure that American school districts need supports that are evidence-based and designed for practitioners that can help districts anticipate oncoming storms of school closures so they can better navigate those storms. Districts need to be able to anticipate closures, proactively develop plans to engage the communities they serve, understand how to best make the decision in equitable ways, and prepare to ease the potential negative effects associated with having to make this type of decision. This research project aims to address this need by answering the following question:

> To what extent can supervised machine learning models accurately predict mass school closures in U.S. school districts to enable proactive resource reallocation?

We define a "mass" closure as a district closing at least 10% of the schools they operate over a segment of time. We consider here a temporal window of five years. In other words, we consider how district and community characteristics predict instances of mass closure as far as five years in advance. We then train supervised machine learning models on historical administrative data to predict a binary indicator of mass closure occurring over the next five years and compare their performance.

Advanced data analysis methods, such as supervised machine learning models, are relatively new in the landscape of social science research, particularly in education (Rahal, Verhagen, and Kirk 2024). Much of the literature involving using machine learning in education thus far has focused on student level outcomes, such as dropout risk and academic performance, and much of that work has taken place outside of the United States context (Kucak, Juricic, and Dambic 2018). There is much more work to be done on utilizing the power and flexibility of machine learning models in more systemic topics in education beyond even individual student outcomes, such as school closures as this project proposes.

That novelty in methodology and lack of direct examples in prior research also engenders understandable concern. Contexts that have employed machine learning algorithms in social contexts have been shown to reproduce problematic and unjust biases (O'Neil 2016; Buolamwini 2023). Given research discussed above that has shown our processes for

school closures thus far to be impacted by at least racialized disparities, we plan to take particular care in selecting predictors and studying the potential bias present in our model to be sure districts who use the model do not reproduce traditionally biased harms on the communities they serve.

1.3 Contributions

Our study offers contributions to both educational administrative praxis and the broader applied machine learning literature in education. We present here a robust early-warning model that forecasts mass school closures - defined as districts closing at least 10% of their schools - five years in advance, thereby equipping school districts with actionable insights to strategically reallocate resources and mitigate adverse impacts. Leveraging nearly two decades of comprehensive administrative data from the National Center for Education Statistics, we present a replicable research design that rigorously evaluates a suite of machine learning techniques, including ensemble methods and deep learning approaches, which adds to a growing body of research establishing effective predictive paradigms in education. By addressing challenges such as class imbalance and temporal dynamics while judiciously incorporating demographic predictors, our approach not only enhances predictive accuracy but also helps us evaluate our models' ability to promote equitable outcomes and mitigate historical biases. We incorporate key methodological innovations, such as grouped cross-validation to prevent data leakage to further bolster the model's generalizability and ethical integrity, underscoring its potential utility for policymakers and educational leaders confronting systemic challenges.

This research has significant implications for policymakers and school administrators as well as the communities served by them. By providing a tool to predict a district's risk of experiencing a mass school closure, we can allow leaders to make more informed strategic decisions ahead of time to address the challenges associated with school closures, such as the displacement of students and staff, the disruption of educational services, and the ever clearer negative impacts on the communities' schools are located within, particularly communities that have historically served our country's most vulnerable populations.

2. Data

The dataset used for this study is comprised of data from the National Center for Education Statistics Institute of Education Sciences (NCES) which contains annual data from every public school and district in the United States. For the purpose of this project, we focus on data since the turn of the twenty-first century.

The NCES data is gathered from the Common Core of Data (CCD) in their publicly accessible Elementary and Secondary Information System database (ElSi). This resource is the Department of Education's primary database on public elementary and secondary education in the United States ("Common Core of Data (CCD)" 2024). For the purposes of this study, the data have been pulled for each year and then joined together based on the unique agency (i.e. district) and school IDs provided by NCES. This dataset includes a binary indicator in each year which is 1 in the year when a school closes, which is a key variable for deriving the outcome of interest.

Table 1: District Demographics

Characteristic	Mean (SD) or Percent	Min	Median	Max
Mass Closure Prevalence	12.6% (33%)			
Number of Schools	5.59 (15.39)	1	3	1111
Total Number of Students	2646.51 (8201.04)	1	868	747009
Elementary Schools	3.31 (10.33)	0	1	697
Middle Schools	1.02 (2.55)	0	1	212
High Schools	1.01 (2.26)	0	1	183
Schools of Other Levels	0.26 (1.17)	0	0	91
Magnet Schools	0.16 (2.64)	0	0	333
Title 1 Schools	3.09 (11.74)	0	1	855
Schools More Diverse Than the District	2.51 (10.27)	0	1	873
Percent White	70% (31%)	0%	83%	100%
Percent Hispanic	12% (20%)	0%	4%	100%
Percent Black	11% (22%)	0%	1%	100%
Percent 'Other'	5% (12%)	0%	1%	100%
Percent Asian	2% (5%)	0%	1%	100%
Percent Free-and-Reduced Lunch	43% (25%)	0%	42%	100%
Rural	50.1%			
Suburban	22.1%			
Town	16.2%			
Urban	11.5%			
1-Regular local school district that is NOT a component of a supervisory union	83.8%			
7-Independent Charter District	11.9%			
2-Local school district that is a component of a supervisory union	3.2%			
4-Regional Education Service Agency (RESA)	0.4%			
5-State agency providing elementary and/or secondary level instruction	0.4%			
8-Other education agencies	0.1%			
9-Specialized public school district	0.1%			
3-Supervisory union administrative center (or county superintendent's office serving the same purpose)	0%			
6-Federal agency providing elementary and/or secondary level instruction	0%			
Per-Pupil Expenditure	$12157.5 (8483.7)	$0	$10429	$1408600
Student:Teacher Ratio	14.5 (4.22)	1	14.22	33

2.1 Data Processing

To prepare the data for analysis, several critical data manipulations were applied at the school level for each year. We first cleaned the data by correcting the following inconsistencies and presumed clerical errors (Cornman, n.d.).

Student-teacher ratios were winsorized to mitigate the impact of extreme values, enhancing the stability of our analyses. We winsorized the values to the 5th and 95th percentiles, capping all values below the 5th percentile and above the 95th percentile to these respective limits. This approach mitigates the influence of outliers while keeping the size of the dataset constant. Demographic percentage values were adjusted to ensure they did not exceed 100%, providing accurate representations of school populations. We also removed records for schools that persisted in the dataset post-closure. Additionally, we assigned each district the most recent locale type (e.g. Urban/Suburban) and agency type (e.g. Regular/Charter) to make these classification variables consistent. To that same end, we adjusted Title-1 classifications to match the previous "1-Yes"/"2-No" classification scheme for Title-1 schools. A clerical error in district enrollment figures was also corrected, where some district's enrollment totals did not align with the sum of the enrollments of its corresponding schools. We corrected this to ensure accurate data at the district level due to the known importance of enrollment in closure considerations (Gilblom and Sang 2021; DeAngelis and Flanders 2019; Hahnel and Pearman 2023). Together, these transformations prepared the data for robust and reliable analysis.

Once these steps were completed, the data was then aggregated to the district level. To capture potentially relevant and useful information at the school level in the aggregated dataset, we generate 4 features for each school-level feature to capture the distribution: the mean, standard deviation, minimum, and the maximum. This approach ensures that for every underlying school-level measure we retain information about both the typical district experience (via the mean), its variability (via the standard deviation), and the range of conditions that exist across schools (via the minimum and maximum). By summarizing each feature in this way, we preserve intra-district heterogeneity and equip downstream models with a richer, more nuanced representation of district-level circumstances.

This aggregation was done for two reasons. First, aggregation to the district level aided in generating our outcome variable of interest, which is an indicator variable regarding the percentage of schools that a district closes over a given number of years. Second, and more importantly, this is done to match the scope of the resulting model to the intended use cases. It is our intent that this model be used to alert districts of the need to prepare to reorganize their resources, not to inform which specific school closure decisions to make within that process. Our hope is that this early warning system allows districts to have more time to implement effective practices for ensuring equity in their school closure processes, as described by Hahnel and Marchitello (2023) .

Directly counter to this goal would be to create a model that predicts which exact schools will close. Given the known bias in the school closure process (Hahnel and Pearman 2023), any model built using these demographic predictors to predict closures at the school level could be repeating the same inequitable processes used for decades to decide on which schools to close, which would further entrench already inequitable outcomes (O'Neil 2016). However, for the purposes of a district-level early-warning system, we feel it is important to include these features in our model to best capture the historical patterns that have led to closure and create the most accurate predictions for districts. Thus, we aggregate to the district level in order to appropriately scope our model. Given the decision-making power for closure choices is at the district leadership level, while our model could exhibit the historical biases present in the data at the school level, we believe the early alert of oncoming closures would allow district and state leaders to make decisions that change the trajectory towards equitable outcomes and even has the potential to allow them to avoid closures altogether.

Finally, we derive our outcome variable. From our analysis, we find that the average district in an given year will close 1-3% of their schools, with schools in Urban locales having higher rates. After calculating the percentages of schools a district closed in a given year, we then create binary indicators over the 5-year time frame with the goal of predicting mass closure events 5 years in advance using the methodology described in **Figure 3** and find that this outcome has a prevalence of 13%.

After removing rows with no outcome due to using a lag to derive our outcome, the resulting dataset contains $207,254$ observations where the unit of analysis is a year for a given district. **Table 1** gives a descriptive analysis of the districts included in the data, with the mean and percentages calculated over all years present in the dataset.

2.2 Features

In our predictive framework, each district–year observation in our final dataset is characterized by a rich set of structural, demographic, and equity-oriented covariates hypothesized to underlie mass-closure risk.

We begin with temporal and size metrics - including the calendar year, the number of schools closed in that year, total schools, and counts of recently opened schools - to capture historical closure trends and district growth. School-level composition is summarized via counts of elementary, middle, high-school, no-level, charter, magnet, and Title-I campuses, as well as the number of "more diverse" schools (i.e., those exhibiting lower Theil-index segregation than the district average, see Methods).

Diversity and resource equity are further quantified through the Theil index and student–teacher ratio within schools (each summarized by mean, standard deviation, minimum, and maximum), and by distributions of free-reduced-price

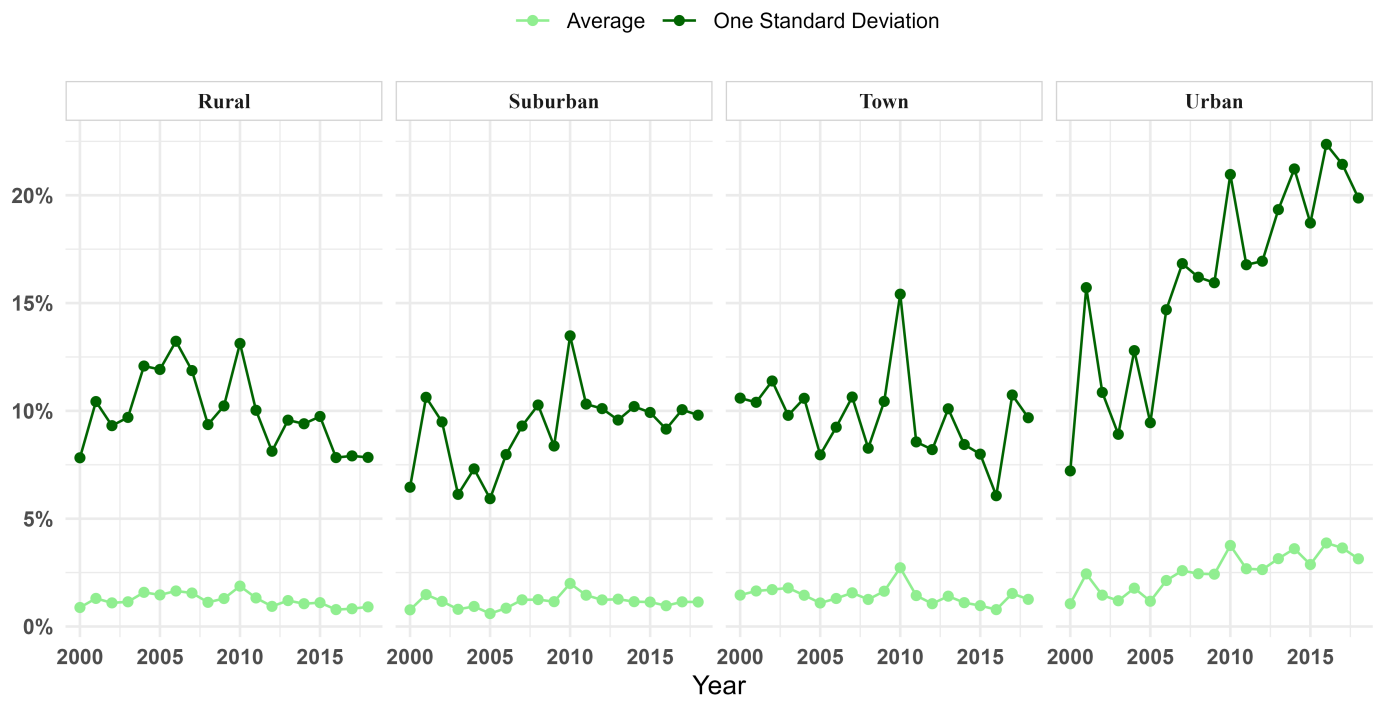

Figure 2: Figure displays the average percent of schools closed each year, disaggregated by locale. On average, in a given year, districts close 1-2.5% of their schools, depending on the locale (with urban districts having around twice the average of other locales). The light green line represents the average percentage of closures, while the dark green line indicates one standard deviation from the mean. Districts with no reliable locale data not included.

lunch enrollment. We also include both counts and proportions of students by race (American Indian, Asian, Black, Hispanic, White, and Other) and overall enrollment concentration (the percentage of total district enrollment residing in each school). These demographic and socioeconomic indicators are crucial, as shifts in student populations and concentrations of need are often linked to enrollment trends and resource allocation decisions that may precede school closures.

Finally, district-level predictors - state, NCES-reported district type and locale, total enrollment, per-pupil expenditure, student–teacher ratio, proportions of free-reduced-price lunch and each racial subgroup, and the district-level Theil index - provide contextual information on resource allocation, financial situation, and segregation patterns that may presage whether a district will engage in a mass school closure event.

3. Methods

This study evaluated the effectiveness of five machine learning algorithms - elastic-net regularized logistic regression, random forests, XGBoost, LSTM neural networks, and SuperLearner ensembles - in predicting the binary classification task of mass closures five years in advance for U.S. school districts, an outcome that is rare and thus leads to an imbalanced outcome variable (14% prevalence in the dataset). Each algorithm was chosen for its approach to handling relationships within data, from linear associations to complex non-linear patterns. These methods are also ones commonly seen and demonstrated to be effective in education prediction tasks, in particular Random Forests, XGBoost, and Neural Networks (Oppong 2023). The following sections describe each method, including underlying mathematical formulations, key parameters, and the hyperparameter tuning techniques employed to handle the class imbalance and we discuss the approach to model selection.

The data is split into training and testing sets using an 80/20 split where we assign all the years for a given district to the same set such that no district appears in both the training and test sets. This splitting methodology avoids the issue of data leakage due to within-district correlation, in which the model embeds the patterns of characteristics of a specific district during training that it then uses in testing, inflating performance metrics. This strategy is also useful for making our models more generalizable to new districts. Consider that districts change over time, eventually every district included in our training would also be considered "out of distribution" at some point in the future (Sriliana,

Budiantara, and Ratnasari 2022; Admojo and Nurul Rismayanti 2024; Baldo, Manthos, and Miani 2019), thus we consider it important to have a model that is more robust to these changes. After examining key distributions, we also identified a need to stratify this split to ensure a balance among the agency types present in the data and incorporate that stratification into our splitting process.

In the process of preparing the data for the split, we examined the distributions of features to examine a need for stratification in the split to ensure representation across features to ensure the split datasets represent similar distributions. In this examination, we determined that given their low prevalence in the dataset, it was not possible to ensure districts with certain characteristics could meaningfully be split into our training and testing sets. Those characteristics were districts within the state of Hawaii as well as particular types of districts (under the NCES classification of type), namely keeping only types 1 (Regular local school district that is NOT a component of a supervisory union), 2 (Local school district that is a component of a supervisory union), and 7 (Independent charter district). Due to their inability to be meaningfully represented within the splits, these districts were removed from the dataset.

From here, we chose to implement a stratified, grouped 5-fold cross-validation approach due to the longitudinal nature of the data. This approach ensures that all the samples for a district are in the same fold for each cross-validation step to prevent data leakage during the training process and generate unbiased estimates of model performance. We also ensure that the data is standardized within the cross-validation process, where each fold used in training is standardized together to address potential data leakage concerns. While the current literature has mixed results on the necessity of standardized data for certain algorithms used here, there is some evidence that standardized data either makes no difference or does help with estimation in prediction tasks (Shanker, Hu, and Hung 1996; Ozsahin et al. 2022).

To address the issue of class imbalance, we used class weights to penalize the loss functions in each method for incorrect predictions of the minority class within the training set. Class weights were calculated directly from the outcome in the training data by first tabulating the counts of each class, identifying the majority and minority classes, and then assigning a weight of 1 to the majority class and a weight equal to the ratio of majority-to-minority counts to the minority class. This weighting scheme was incorporated directly into the optimization routines of logistic regression, random forest, and gradient boosting models as well as into the cross-entropy loss for neural networks so that misclassification of the less frequent mass closure cases incurred a proportionally larger penalty.

Further, the models were evaluated during cross-validation using Area Under the Precision-Recall Curve (or AUC-PR) for model evaluation and selection. AUC-PR evaluates the trade-off between **Precision** (how many predicted positives are actually positive, also called *specificity*) and **Recall** (how many actual positives are correctly predicted, also called *sensitivity*) across various thresholds. Note that here a positive outcome refers to a district closing 10% of their schools over the next 5-years from the year of the prediction.

We choose to focus on AUC-PR due to the inherent class imbalance in our dataset, where the positive class represents a small minority of the total instances. The AUC-PR metric was selected over the more commonly used Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for several reasons:

1. **Sensitivity to class imbalance**: AUC-PR is more sensitive to improvements in the minority class prediction, which is crucial when dealing with imbalanced datasets (Dabek et al. 2022).
2. **Focus on positive class**: In our context, accurately identifying the minority (positive) class is of primary importance. AUC-PR places more emphasis on the correct classification of positive instances (Subramaniam, Kane, and Krishnamurti 2023).
3. **Robustness to changes in class distribution**: AUC-PR provides a more informative and discriminative measure when the class distribution may vary between the training set and real-world applications (Pérez Míguez et al. 2024).

The AUC-PR score ranges from 0 to 1, with higher values indicating better model performance. It represents the model's ability to correctly identify positive instances while minimizing false positives across various classification thresholds. By using AUC-PR, we aim to provide a more accurate and meaningful evaluation of our model's performance, particularly in its ability to correctly classify the instances of closures while balancing false negatives and false positives. In the case that AUC-PR is not particularly discerning between models, we also will examine the cross-validated Recall of the models.

In our consideration of the potential impact of misclassifications, we consider false negatives to be particularly problematic to the goals of the early-warning nature of the task. While false positive results could lead to unnecessary distress for districts and the communities they serve, we can imagine scenarios where the due diligence to prepare for a mass closure event could lead to positive outcomes for the district, for example by auditing budgets more closely to better allocate resources. However, false negative results could lead to much more dire results where districts who otherwise would have been able to prepare for and possibly avoid the mass closure event would now lose the early-warning to enable a more equitable and reasonable process to take place. Our intent is that by evaluating performance using AUC-PR, we are able to balance these two outcomes

and also focus on reducing false negatives if necessary as well. Thus, in our analysis, we will report the AUC-PR score and compare it among models to demonstrate the effectiveness of our model in handling the imbalanced nature of the data.

3.1 Feature Engineering

Once the dataset was finalized, we engineered domain-specific indicators to measure aspects of overcrowding and diversity. To assess overcrowding, we calculated the average percentage of district enrollment across all constituent schools, thereby providing a direct measure of the spatial and resource pressures within a district. Following Pearman and Greene (2022), we adopted the Theil index to quantify the amount of racial diversity in the district based on the demographic enrollment data available (White 1986). The Theil index offers a nuanced view of the distribution of various demographic groups across schools, reflecting intra-district heterogeneity. Building on this, we derived a secondary feature by counting the number of schools within each district that exhibit diversity levels exceeding the district-wide average. This dual-layer approach to capturing diversity not only informs the overall demographic balance but also highlights localized variations that may have distinct implications for educational outcomes. The full list of predictors can be found in **Appendix A3**.

3.2 Models

3.2.1 Elastic Net Logistic Regression

Logistic regression is a linear model for binary classification that estimates the probability of class membership using the logistic function. For a binary outcome $y \in 0, 1$, logistic regression models the log-odds of $y = 1$ as a linear combination of the predictors:

$$\text{logit}(P(y = 1|X)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}$$

The parameters β are estimated by minimizing the binary cross-entropy (or log-loss) cost function:

$$J(\beta) = -\frac{1}{N} \sum_{i=1}^{N} (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i))$$

where $\hat{y}_i$ is the predicted probability for the i-th observation.

In this analysis, logistic regression was implemented with Elastic Net Regularization, which combines both $L1$ (lasso) and $L2$ (ridge) penalties on top of this loss function. Elastic Net regularization is particularly advantageous when dealing with datasets that may have collinear or highly correlated features, as it can perform variable selection (through the lasso penalty) while also controlling the overall magnitude of coefficients (through the ridge penalty).

Thus, the final objective function minimized is given by:

$$\min_{\beta} [J(\beta) + \lambda(\alpha \sum_{j=1}^{p} |\beta_j| + (1 - \alpha) \sum_{j=1}^{p} \beta_j^2)]$$

where:

- $J(\beta)$ is the binary cross-entropy loss function described above
- λ is the regularization parameter controlling the overall strength of the penalty term
- α controls the balance between the $L1$ and $L2$ penalties such that when $\alpha = 1$ the model reduces to lasso regularization and when $\alpha = 0$ the model reduces to ridge regularization. Values between 0 and 1 represent a mix of both penalties.

In this study, the regularization parameters λ and α were optimized using grid search and 5-fold cross-validation to find the optimal balance that minimizes overfitting while improving model AUC-PR performance using the "glmnet" R package (Friedman et al. 2023), with training control handled through the "caret" R package (Kuhn [aut et al. 2024). Elastic Net's dual penalty approach provided flexibility in handling the high-dimensional nature of the dataset, allowing the model to maintain interpretability by potentially shrinking some coefficients to zero (like lasso) and reducing the impact of correlated variables. Grid search details of the 550 unique parameter combinations used for this model can be found in **Appendix A2**.

3.2.2 Random Forests

Random forests are an ensemble learning method based on the concept of bagging (Bootstrap Aggregating), which builds multiple decision trees and averages their predictions to reduce variance and improve accuracy. Each tree in the forest is trained on a bootstrap sample, and at each split, a random subset of features is considered, making the trees decorrelated.

For classification tasks, such as in this study, the prediction for a given observation is the majority vote across all trees:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_M(x)\}$$

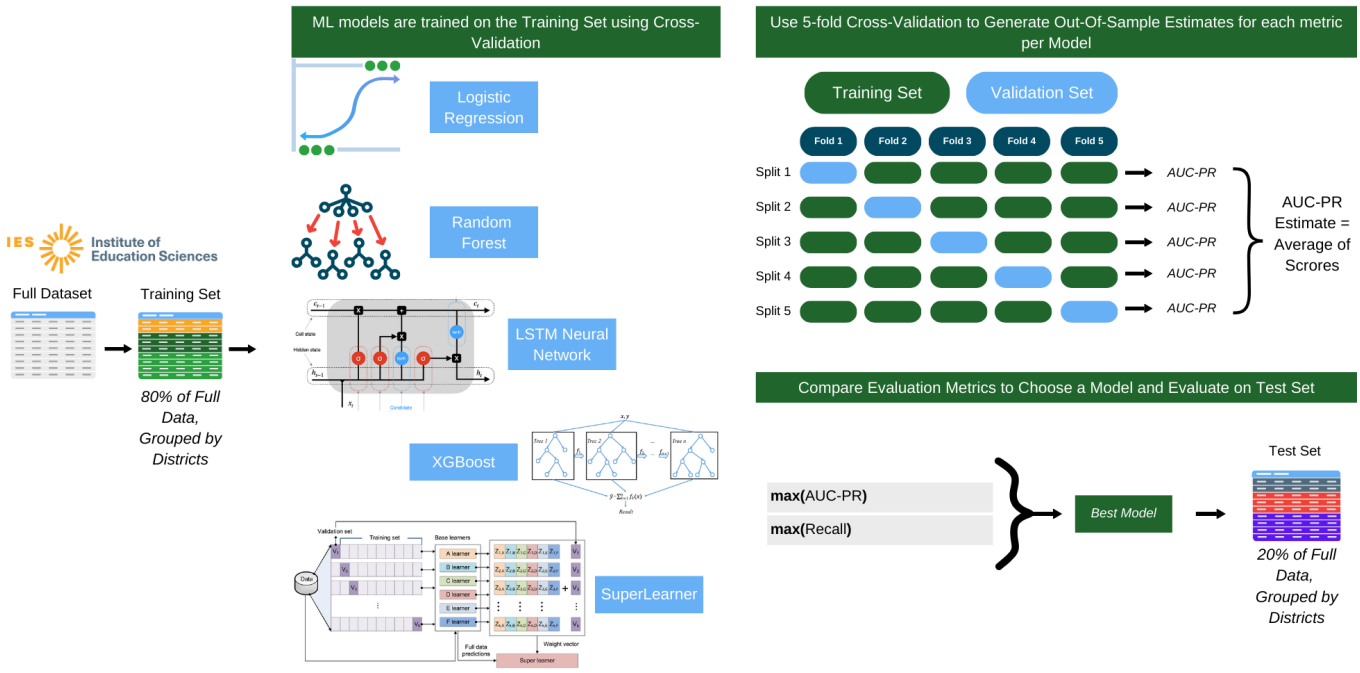

Figure 3: Research design. We begin with taking the data and splitting it into a training set (80%) and a testing set (20%). We then train 5 models on the training data, using 5-fold cross-validation to estimate AUC-PR and Recall estimates. Once trained, we examine these estimates to determine model generalizability. Evaluating model performance based on these two measures as well as fold performance during cross-validation, we choose a 'Best Model' to examine on the test data for a generalizable performance estimate of the model. Note: the splits and cross-validation are done using a grouped design, such that all the years for a district follow the district's fold assignment.

where $T_m(x)$ is the prediction from the m-th tree. In this study, to improve model sensitivity to the rare positive cases, we explored several splitting criteria that are better suited for imbalanced data:

- **Gini impurity**: Measures the likelihood of an incorrect classification by a node; tuned to prioritize splits that enhance detection of the minority class.
- **Hellinger distance**: Specifically designed for imbalanced datasets, this criterion measures class separation more effectively by reducing the influence of the majority class.
- **Extremely randomized trees (ExtraTrees)**: Creates splits by randomly selecting thresholds, increasing variance among trees and potentially capturing more rare patterns.

Each splitting criterion has specific advantages. Gini impurity is robust and standard. Hellinger distance is especially suitable for imbalanced datasets due to its focus on maximizing class separation. ExtraTrees introduces randomness, which increases tree diversity and helps detect rare patterns in the minority class through varied splits.

Using a combination of these criteria within cross-validation allows us to better balance model performance across the majority and minority classes, improving the model's ability to detect the rare positive cases effectively. Each tree's parameters, including the number of trees M, maximum depth, and minimum samples per leaf, were optimized through grid search and 5-fold cross-validation with emphasis on AUC-PR and trained using the "ranger" R package implementation (Wright, Wager, and Probst 2024), with training control handled through the "caret" R package (Kuhn [aut et al. 2024). More information on these splitting methods can be found in the appendix (**Appendix A1**) along with details on the full grid of 105 unique parameter combinations used (**Appendix A2**).

3.2.3 XGBoost

XGBoost, or Extreme Gradient Boosting, is a highly efficient implementation of gradient boosting, which iteratively improves predictions by fitting additional trees to the residual errors of the previous trees. For a set of training data $(x_i, y_i)_{i=1}^N$, XGBoost builds an ensemble of decision trees $f_t(x)$ in an additive manner:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i)$$

where each f_t minimizes an objective function $\mathcal{L}$ composed of a loss function l and a regularization term Ω:

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t)$$

Here, l is the binary cross-entropy, while the regularization term $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_j w_j^2$ penalizes the complexity of each tree f_t based on the number of leaves T and the leaf weights w_j.

A comprehensive grid search of 1,080 unique parameter combinations was implemented to optimize the hyperparameters of an XGBoost classifier (see **Appendix A2**). The custom tuning grid spans a range of values for key hyperparameters such as the number of boosting iterations (nrounds), maximum tree depth (max_depth), learning rate (eta), regularization parameter (gamma), column subsampling ratio (colsample_bytree), minimum child weight (min_child_weight), and the subsample ratio for training instances. This grid was designed to explore a broad spectrum of model complexities and learning behaviors, thereby aiding in the selection of a configuration that maximizes predictive performance. The model was trained using the "xgboost" package in R (Chen et al. 2024), using the "caret" package to manage the training process described here.

3.2.4 Long Short-Term Memory (LSTM) Neural Networks

LSTM neural networks, a type of recurrent neural network (RNN), are designed to handle sequential data by retaining long-term dependencies through memory cells. LSTMs mitigate the vanishing gradient problem common in RNNs, allowing them to capture temporal relationships in time series data. For each time step t, the LSTM cell computes the hidden state h_t and cell state C_t using gates that regulate information flow:

- **Input gate** $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$
- **Forget gate** $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$
- **Output gate** $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$

The cell state C_t and hidden state h_t are updated as follows:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \quad h_t = o_t * \tanh(C_t)$$

where $\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$.

Model training employed a two-layer LSTM backbone: the first layer comprised N units (e.g., 64) with L2 kernel regularization and a subsequent dropout layer for preventing overfitting. A second LSTM layer with twice as many units further distilled temporal patterns, followed again by dropout. To capture long-range dependencies beyond what recurrent units alone can learn, a multi-head self-attention block (four heads) was applied to the LSTM outputs; its output was summed residually with the LSTM activations and normalized via layer normalization. Dense processing was performed in a time-distributed manner, using a ReLU-activated layer of 32 units and an additional dropout layer before projecting to a sigmoid output at each time step. The final output was reshaped to yield a sequence of probability estimates for the positive class. Training was also controlled by two callback mechanisms: early stopping on validation PR-AUC with a patience of 15 epochs (restoring the best weights), and a ReduceLROnPlateau callback to halve the learning rate after seven stagnant validation steps, down to a floor of 1e-5.

The LSTM neural network was tuned across 216 parameter combinations (**Appendix A2**). We explored variations in network architecture (units), regularization (dropout rate), optimization parameters (learning rate, batch size), and training duration (epochs) to identify the optimal configuration. The binary cross-entropy (weighted to account for the class imbalance) was minimized using the Adam optimizer, and a grouped cross-validation method was applied. The model was trained using the TensorFlow package in python with a custom cross-validation method (Abadi et al., n.d.).

3.2.5 SuperLearner Ensemble

To predict the likelihood of a district experiencing a mass school-closure event, we implemented a weighted ensemble using the R SuperLearner framework (Van Der Laan, Polley, and Hubbard 2007; Polley and Laan 2010). The binary outcome variable was coded as 1 for districts meeting the closure threshold and 0 otherwise. All remaining district–year features - including time-fixed effects, financial indicators, and demographic covariates - were automatically selected as predictors, excluding only the outcome and the clustering identifier, here each district's unique ID.

Our ensemble library comprised four algorithms: the simple mean predictor (SL.mean), penalized logistic regression (SL.glmnet), random forest, and gradient-boosted trees. To incorporate class weights in the training process, we developed custom wrapper functions - SL.ranger.wt and SL.xgboost.wt - that pass the computed weights to the underlying `ranger` and `xgboost` calls, respectively and also used the weights in training the ensemble. Hyperparameters for the Random Forest and XGBoost were chosen based on the best hyperparameters identified in each models separate training prior to training the SuperLearner. This was done to adjust for some computation constraints in the nested cross-validation scheme described below.

We employed a nested cross-validation scheme to both select and evaluate ensemble weights, preserving the clustering structure by specifying the it in SuperLearner's V-fold CV. Pre-defined outer folds (five in total, identical to the other methods described previously) were used for unbiased performance estimation of the SuperLearner, with

each holdout fold comprising all observations from a distinct subset of districts. For hyperparameter tuning and base-learner combination within each outer training partition, we ran an inner five-fold CV. This nested design mitigates overfitting in weight estimation and allows a robust assessment of generalization error.

3.3 Missing Data

The impact of missing data is expected to vary across the five machine learning methods:

1. **Logistic Regression with Elastic Net Regularization**: Logistic regression is particularly sensitive to missing data due to its reliance on linear relationships and the assumption of independent predictors. Imputation may improve its stability, while exclusion could reduce statistical power and weaken variable selection.
2. **Random Forests**: Random forests are naturally resilient to missing data, as splits at individual trees can be determined based on non-missing features. However, imputation might still enhance performance by providing a more consistent dataset across all trees, whereas exclusion could reduce model diversity by removing rows.
3. **XGBoost**: XGBoost has a built-in mechanism to handle missing values by learning optimal split directions for missing data during tree construction. Thus, the model may perform well even without imputation. However, imputing missing values could still enhance predictive power by reducing potential noise introduced by the missingness itself.
4. **LSTM Neural Networks**: LSTMs are sensitive to missing sequential data, as gaps can disrupt temporal dependencies. Imputation is typically considered to be critical for LSTMs, where missing data can cause significant drops in performance. Excluding rows, especially in smaller datasets, may lead to poor generalization by limiting the availability of training sequences.
5. **SuperLearner**: Missing data can significantly affect a SuperLearner ensemble by disrupting the training and prediction processes of its constituent models. Since SuperLearner relies on combining predictions from multiple algorithms, any model within the ensemble that cannot handle missing values may either fail to train or produce biased estimates. This inconsistency can weaken the overall performance of the ensemble and reduce its predictive accuracy.

Missing data in the study is anticipated to arise from incomplete records on demographic, temporal, or school-level variables, which could introduce bias if handled improperly as described above. The missing data in our dataset accounts for only 5% of the total data points but is distributed across many rows, suggesting that missingness is sparse but widespread rather than concentrated. This study addresses missing data by training on a dataset where rows with missing data are excluded. While this approach results in removal of approximately 49% of the rows, the large number of remaining rows (over 200,000) ensures sufficient data remains for robust model training.

4. Results

The increasing prevalence of permanent school closures as a budgetary and reform strategy underscores the urgency of developing predictive tools that can help districts navigate these difficult decisions proactively. This study sought to determine whether machine learning models could provide an early-warning system capable of forecasting the threat of mass school closures - defined as a district closing at least 10% of its schools five years in advance.

In this section, we evaluate the predictive performance of five machine learning models: elastic-net regularized logistic regression, random forests, XGBoost, LSTM neural networks, and SuperLearner ensembles. We assess their ability to detect mass closure events using key performance metrics, namely Area Under the Precision-Recall Curve (AUC-PR) and Recall. We begin by discussing model performance.

4.1 Model Performance

Effective predictive modeling relies on selecting optimal hyperparameters that balance model complexity, generalization, and computational efficiency. Through cross-validation and hyperparameter tuning, we identified the best-performing hyperparameters for each model that required hyperparameter choices. These final hyperparameters and their implications can be found in **Appendix A3**.

The performance metrics of five machine learning models are presented in **Table 2**, which reports five-fold cross–validated performance (mean $\pm$ SD) of our candidate classifiers on the target prediction task, using area under the precision–recall curve (AUC-PR) and recall as primary metrics:

- **AUC-PR:** XGBoost achieved the highest mean AUC-PR (0.396 ± 0.017), closely followed by the SuperLearner (0.393 ± 0.023) and the LSTM (0.391 ± 0.028). Random Forest (0.374 ± 0.011) and Elastic Net (0.369 ± 0.017) trailed behind, with Random Forest showing the most consistent precision-recall performance (lowest SD).
- **Recall:** The LSTM model delivered the best sensitivity, with a mean recall of 0.776 ± 0.046. Elastic Net also demonstrated high recall (0.724 ± 0.013), whereas

tree-based methods (Random Forest: 0.684 ± 0.017; XGBoost: 0.678 ± 0.015) were more conservative. Notably, the SuperLearner ensemble exhibited poor recall (0.454 ± 0.102) and high variability, suggesting that its blending strategy favored precision over sensitivity in this setting.

Table 2 - Model Performance

Model	AUC-PR	AUC-PR sd	Recall	Recall sd
Elastic Net	0.369	0.017	0.724	0.013
LSTM	0.391	0.028	0.776	0.046
Random Forest	0.374	0.011	0.684	0.017
SuperLearner	0.393	0.023	0.454	0.102
XGBoost	0.396	0.017	0.678	0.015

Taken together, these results suggest a trade-off between optimizing precision (as captured by AUC-PR) and maximizing case detection (recall): XGBoost and the SuperLearner excel at ranking positives higher overall, whereas the LSTM provides superior coverage of true positives.

Beyond mean performance, the fold-by-fold shown in **Figure 4** results reveal important differences in model stability. XGBoost not only attains the highest average AUC-PR but also shows relatively tight clustering of its five fold points - suggesting consistently strong ranking of positives across resamples. In contrast, the LSTM exhibits the widest spread in both AUC-PR and recall: one fold dips below 0.35 in precision–recall performance while another peaks above 0.42, and its recall spans roughly 0.70–0.80. Elastic Net and Random Forest each produce more compact fold clouds, indicating dependable - albeit lower - overall performance. The SuperLearner's AUC-PR variability is moderate, yet its recall points span nearly 0.35 to 0.65, underscoring that the ensemble's sensitivity is especially inconsistent. Taken together, these patterns highlight XGBoost as both a high-performer and low-variance option, whereas the LSTM and SuperLearner may deliver less predictable results post-training.

Given these results, XGBoost represents the best balance of high discriminatory power and robust generalizability among our candidates, making it the clear choice for final test-set evaluation. It achieved the top mean AUC-PR (0.396) while maintaining one of the smallest fold-to-fold variances, indicating that its ranking of true positives is both strong and reproducible. Although the LSTM slightly outperformed XGBoost on recall, its much wider variability across folds raises concerns about unstable sensitivity in new data. In contrast, XGBoost delivers near-leading recall (0.678 ± 0.015) with minimal fluctuation, providing confidence that its sensitivity will hold up outside the training environment. Moreover, XGBoost's fast training time, built-in regularization, and native feature-importance measures further support its deployment: it not only maximizes precision–recall performance but also offers interpretability and scalability for large, real-world datasets. We therefore proceed with XGBoost for our final test-set evaluation and subsequent application.

4.2 Test Set Performance

Given its balanced performance on the evaluation metrics, we selected the XGBoost model for our predictive model on our test set of districts. **Table 3** shows the confusion matrix for the model, which shows how well the predictions matched the ground truth performance within the test set.

As seen in **Table 4**, on the held-out test set XGBoost achieved an overall accuracy of 0.7213 (95% CI: 0.7156–0.7270), meaning it correctly classified roughly 72% of examples. However, this figure should be interpreted in light of the No Information Rate (NIR) of 0.8463 - the accuracy that would be obtained by always predicting the majority class (e.g. never predicting closure). Here, our model's accuracy is significantly below the NIR (p-value = 1 for Acc > NIR), indicating that raw accuracy alone overstates performance when the data are heavily imbalanced. Cohen's Kappa of 0.277 reflects only fair agreement beyond chance, again underscoring the challenge posed by class imbalance, even when managed via class weighting as done here.

Turning to class-specific metrics, the model's Sensitivity is 0.6816, meaning it correctly identifies about 68% of actual positives. Specificity (true negative rate) is 0.7285, so roughly 73% of actual negatives are correctly flagged. These results yield a balanced accuracy of 0.7051, indicating the model does somewhat better than chance on each class, but still leaves substantial room for improvement.

Precision for the positive class is low: the Positive Predictive Value is only 0.3133, i.e. fewer than one in three of the observations predicted as mass closure is truly a mass closure. In contrast, the Negative Predictive Value is high (0.9264), meaning that when the model predicts no mass closure, it is almost always correct. Because the prevalence of mass closure in the test set is just 15.37%, the model produces a detection prevalence of 33.45% - it over-calls the positive class relative to its true frequency, generating a large number of false positives (5,551) compared to true positives (2,532). Finally, McNemar's test p-value (<2e-16) indicates a significant asymmetry in the pattern of false positives versus false negatives, suggesting that our classifier errs more heavily in one direction.

Table 3 - XGBoost Confusion Matrix

	Reference	
Prediction	Non-Extreme Closure	Extreme Closure
Non-Extreme Closure	14898	5551
Extreme Closure	1183	2532

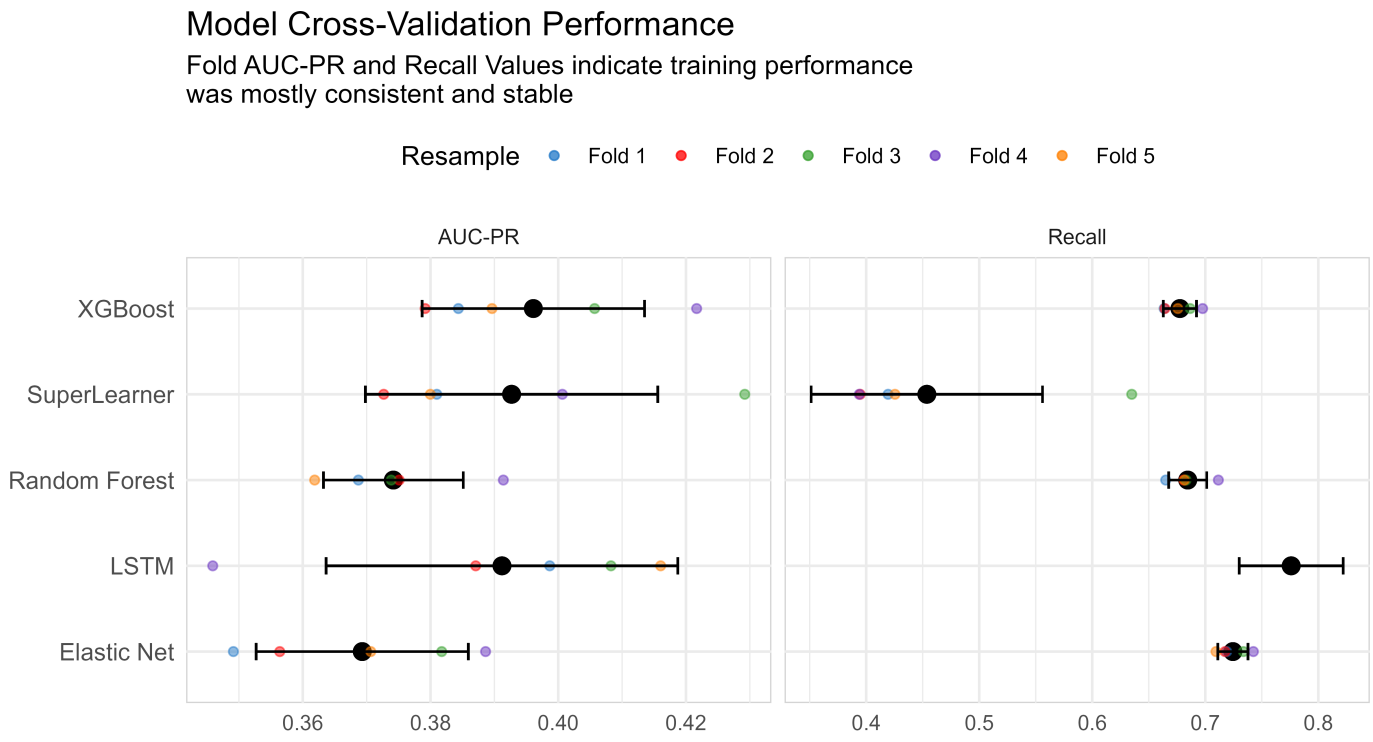

Figure 4: Cross-validated Area Under the Precision–Recall curve (AUC–PR) and Recall for our five different classification models - Random Forest, XGBoost, Elastic Net Logistic Regression, LSTM, and SuperLearner - valuated via five-fold cross-validation. The horizontal axis depicts the values for each metric, ranging from 0.0 to 1.0, while the vertical axis lists the respective models. For each model, five distinct points correspond to the performance in each of the five cross-validation folds while the mean and standard deviation across the folds is depicted in the black dot and error bars.

Table 4 - XGBoost Model Performance Metrics

Metric	Value
Model Accuracy Metrics	
Accuracy	0.7213
95% CI	(0.7156, 0.7270)
No Information Rate	0.8463
P-Value [Acc > NIR]	1
Kappa	0.2769
Mcnemar's Test P-Value	< 2e-16
Classification Metrics	
Sensitivity	0.6816
Specificity	0.7285
Pos Pred Value	0.3133
Neg Pred Value	0.9264
Prevalence	0.1537
Detection Rate	0.1048
Detection Prevalence	0.3345
Balanced Accuracy	0.7051

4.3 Error Analysis

Understanding the errors made in predictive modeling is essential for evaluating model performance and ensuring its practical applicability. In this study, we conduct a detailed error analysis to assess the nature of misclassifications in our predictions of mass closure. Here, we extend our previous analysis of the accuracy of our model to an examination of specific error patterns of the false positives and false negatives, to identify potential biases or systematic issues in our most performant model and to potentially identify areas for improvement. We analyze the distribution of errors across key subgroups to determine whether model performance varies by demographic or contextual factors. This analysis provides critical insights into the model's reliability and areas for potential improvement.

Figure 5 (and the detailed breakdown for each numeric predictor in **Appendix)** illustrate the differences in numeric feature distributions between correctly classified cases and misclassified cases (false positives and false negatives). The plot highlights key discrepancies in variables. For instance, observing the distributions for "District Number of Students," false negatives tend to occur in districts with a seemingly lower number of students compared to correctly identified closures or false positives, while false positives sometimes appear in districts with a higher student population than correctly identified non-closures. Similarly, variables such as "Number of Schools in District" show that false negatives are more common in districts with

fewer schools, whereas false positives can occur across a wider range, sometimes overlapping with districts that did experience closures.

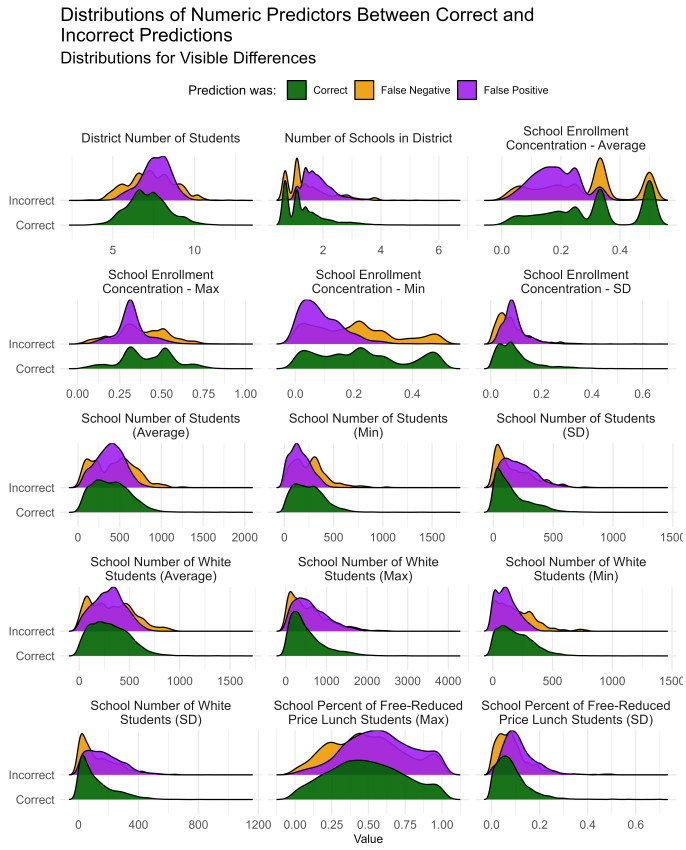

Figure 5: **Distributions of numeric predictors for correct and incorrect predictions.** This figure displays the distribution of key numeric predictors stratified by prediction outcome: correct predictions (green), false negatives (orange), and false positives (purple). Each subplot corresponds to a different variable used in the classification model to predict school closures, such as district and school-level student counts, enrollment concentration metrics, racial composition, and socioeconomic indicators (e.g., percent of students eligible for free/reduced-price lunch). Distributions are faceted by variable and prediction correctness to visually assess which features are associated with systematic model errors. Notable patterns include higher concentrations of false positives at low values of "Number of Schools in District" and higher variability in false negatives across "School Enrollment Concentration" measures, suggesting model sensitivity to contextual school system characteristics.

Distributions for school enrollment concentration metrics (average, max, min, SD) also reveal patterns. For example, false negatives for "School Enrollment Concentration - Average" appear more frequently at lower concentration values, while false positives might show higher concentration averages than true negatives. The "School Percent of Free-Reduced Price Lunch Students (Max)" shows that false positives tend to have a higher maximum percentage of students eligible for free-reduced price lunch compared to correctly identified non-closure instances. These patterns suggest that the model might be struggling with specific profiles of districts, potentially misinterpreting signals from these numeric predictors, leading to specific types of errors. For example, the model may associate very high percentages of students on free-reduced price lunch with impending closure, leading to false positives if other mitigating factors are not captured or weighted appropriately. Conversely, districts with fewer students or schools might be underestimated for closure risk, resulting in false negatives.

An analysis of the XGBoost model's performance in predicting mass school closures across different geographical locales was also examined and is presented in **Figure 6**. The results detail the distribution of correct predictions, false negatives, and false positives as a percentage of total predictions within each NCES locale category: Rural, Suburban, Town, and Urban. Overall, the model's predictive accuracy varied across locales. The highest percentage of correct predictions was observed in Rural districts, where the model accurately predicted the outcome for 75.8% of cases. Performance was comparable in Suburban districts with 72.3% correct predictions. The model exhibited lower accuracy in Town and Urban districts, with 64.0% and 65.3% correct predictions, respectively.

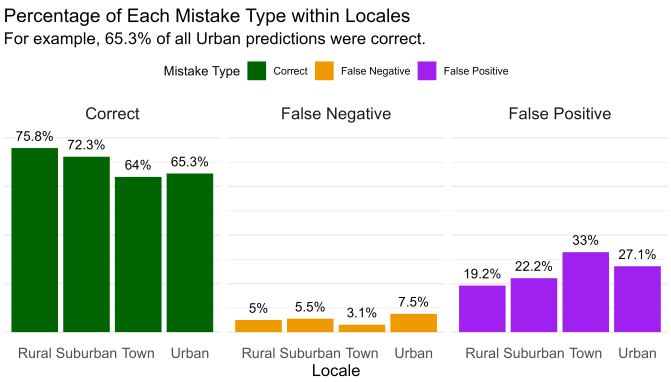

Figure 6: **Distribution of Prediction Outcomes Across Locales.** This figure illustrates the percentage of correct predictions, false negatives, and false positives for a predictive model, categorized by four NCES locales: Rural, Suburban, Town, and Urban. The analysis reveals that correct predictions were highest in the Rural locale (75.8%), followed by Suburban (72.3%), Urban (65.3%), and Town (64%). False negatives were most prevalent in the Urban setting (7.5%), while Rural areas showed the lowest incidence of this error type (5%). Conversely, false positives were most frequent in the Town locale (33%), with Rural areas exhibiting the lowest rate of false positives (19.2%). Overall, the figure highlights variations in model performance across different geographical settings and suggests that the model particularly struggled as locales became more urban.

An examination of error types reveals distinct patterns across the locales. False negatives, which represent instances where the model failed to predict impending mass school closures that subsequently occurred, were lowest in Town districts at 3.1% of all predictions for that locale. Rural and Suburban districts showed slightly higher false

negative rates of 5.0% and 5.5%, respectively. The highest rate of false negatives was observed in Urban districts, where 7.5% of predictions were false negatives. Such errors could have significant implications.

Conversely, false positives, where the model incorrectly predicted mass closures that did not materialize, showed a different distribution. Rural districts had the lowest false positive rate at 19.2%. This was followed by Suburban districts with 22.2% and Urban districts with 27.1%. Town districts exhibited the highest false positive rate, with 33.0% of all predictions for this locale being false positives. A high incidence of false positives could lead to unnecessary alarm, misallocation of support resources, or potentially damaging reputational effects for districts erroneously identified as facing mass closures. While for the expected use case of the model we argue these concerns are more minimal, the high incidence of false positives in the model makes them more likely and - thus - potentially more problematic when considering model deployment.

Further, **Figure 7** shows that the model also struggled to predict Charter districts. For regular local school districts not part of a supervisory union (Type 1 in the NCES categorization scheme), the model correctly predicted outcomes in 72.21% of cases, with false negatives occurring in 4.89% and false positives in 22.89% of predictions. Local school districts that are components of a supervisory union (Type 2) showed a similar pattern, with 71.53% correct predictions, 3.65% false negatives, and 24.82% false positives. In contrast, independent charter districts (Type 7) exhibited a lower percentage of correct predictions at 60.29%, with false negatives at 8.82% and a higher percentage of false positives at 30.88%. This suggests that the model's performance varies depending on the type of school district, with independent charter districts showing the highest rate of incorrect predictions.

4.4 Synthesizing Results

Taken together, the analysis of the model's predictive performance reveals an accuracy that, while demonstrating predictive potential, varies considerably across different district typologies and locales. A predominant characteristic observed across multiple segments is the model's propensity towards False Positive errors, where mass closures are predicted but do not subsequently occur. This tendency was particularly notable, suggesting a potential over-sensitivity of the model to certain closure indicators within specific contexts. For instance, Town locales exhibited a 33% False Positive rate, and Independent Charter Districts showed a 30.88% False Positive rate.

The model's efficacy is significantly diminished when applied to Independent Charter Districts. This category not only registered the lowest overall predictive accuracy (60.29% correct) but also concurrently displayed the highest rates for both False Negative (8.82%) and False Positive (30.88%) predictions. This suggests that the factors driving closures or the characteristic data profiles for charter districts are not as effectively captured by the current model compared to other district types.

Performance also diverged notably by geographic locale. While Rural districts were predicted with the highest accuracy (75.8%), Town districts demonstrated markedly lower accuracy (64%), primarily attributable to their elevated False Positive rate. In contrast, Urban districts were characterized by a higher False Negative rate (7.5%), indicating a greater likelihood of the model failing to anticipate impending closures in these settings, potentially hindering proactive interventions or resource allocation for districts at growing risk of substantial closures given **Figure 2**. Regular local school districts, whether or not part of a supervisory union, showed more moderate and comparable performance (71-72% accuracy), though districts within a supervisory union had a notably lower False Negative rate (3.65%).

Further examination of the distributions of numeric predictors for correct versus incorrect predictions provides insights into these performance variations. The model consistently demonstrated reduced reliability for smaller districts. Specifically, a higher incidence of incorrect predictions was observed for districts characterized by a low total number of students, few constituent schools, and low average or minimum student numbers per school. Moreover, districts exhibiting low internal variability (i.e., low standard deviation) on several key metrics - such as school enrollment concentration, the standard deviation of student numbers across schools, and the standard deviation in the percentage of students eligible for free or reduced-price lunch - were more frequently associated with prediction errors. This implies that the model faces challenges in accurately distinguishing outcomes for districts that are either very small in scale or display a high degree of homogeneity across their constituent schools on these specific structural and demographic indicators - characteristics particularly (though not uniquely) fitting of many small charter districts within urban environments.

5. Discussion

5.1 Findings Overview

The primary goal of this study was to develop an early-warning indicator for mass school closures using supervised machine learning models. Our findings underscore both the potential and challenges of applying advanced predictive methods to administrative education data. The increasing prevalence of permanent school closures necessitates proactive tools, and this research sought to determine if machine

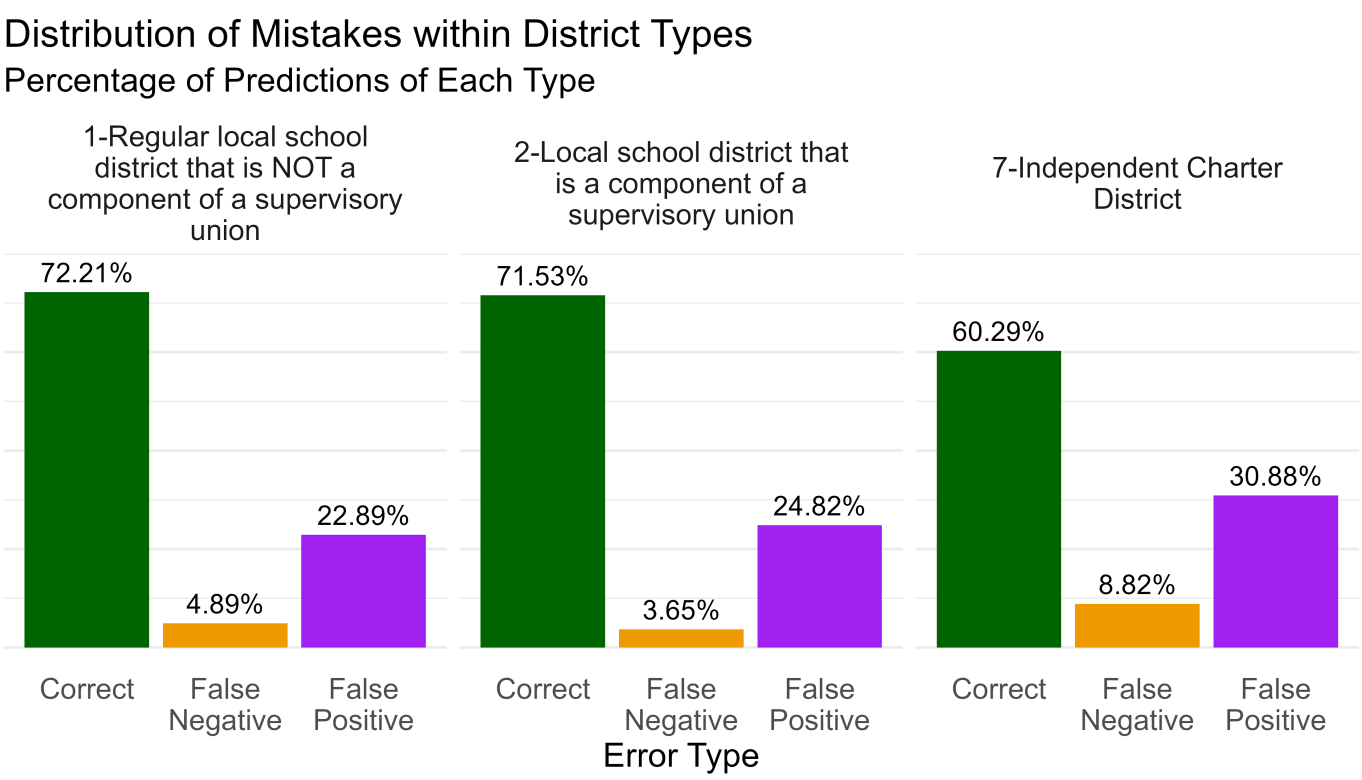

Figure 7: Prediction accuracy across three distinct types of school districts. This figure illustrates the percentage of predictions classified as Correct (green), False Negative (orange), and False Positive (purple) for each district category. The district types analyzed are: (1) Regular local school districts not part of a supervisory union, (2) Local school districts that are components of a supervisory union, and (7) Independent Charter Districts. For regular local school districts not part of a supervisory union (Type 1), correct predictions accounted for 72.21% of the outcomes. False negatives constituted 4.89%, and false positives represented 22.89% of predictions. In the case of local school districts that are components of a supervisory union (Type 2), a similar distribution was observed, with correct predictions at 71.53%. This district type exhibited the lowest percentage of false negatives at 3.65%, while false positives were slightly higher at 24.82%. Independent Charter Districts (Type 7) demonstrated a notably different profile. Correct predictions were lower, at 60.29%. Conversely, this category showed the highest rates for both false negatives (8.82%) and false positives (30.88%).

learning could provide a reliable system for forecasting the threat of mass school closures. In the following discussion, we analyze model performance, address practical implications for deployment, consider methodological choices regarding the unit of analysis and use of demographic predictors, demonstrate appropriate use of the model and what its use enables via a case study of The School District of Philadelphia's 2012-2013 closures, and situate our work within the broader literature on machine learning in education.

Our evaluation across five machine learning models - elastic-net regularized logistic regression, random forests, XGBoost, LSTM neural networks, and SuperLearner ensembles - focused on AUC-PR and Recall. XGBoost emerged as the strongest candidate, achieving the highest mean AUC-PR (0.396 ± 0.017) during cross-validation, indicating superior precision in identifying true positives, and demonstrating robust, consistent performance across folds. While the LSTM model showed the highest mean recall (0.776 ± 0.046), its wider variability across folds raised concerns about its stability on new data. The SuperLearner also showed good AUC-PR but had poor and inconsistent recall. Ultimately, XGBoost presented the best balance of high discriminatory power and generalizable performance, leading to its selection for test set evaluation. On the test set, the XGBoost model achieved an overall accuracy of 0.7213 and a balanced accuracy of 0.7051. While its accuracy was below the No Information Rate, highlighting the challenge of the imbalanced outcome, its recall of 0.6816 indicates it correctly identified about 68% of actual mass closure events.

Our error analysis also revealed specific challenges the model faced, despite its passable performance. A notable tendency was the prevalence of False Positive predictions across various district segments, indicating the model may be overly sensitive to certain closure signals. This issue was particularly pronounced for Independent Charter Districts, which not only exhibited the lowest overall predictive accuracy but also the highest rates of both False Positives and False Negatives, suggesting their unique characteristics are not fully captured. Though disappointing, this performance on charter districts is unsurprising given that closure

processes for charter districts are not governed by the same logic and pressures faced by traditional school districts and thus it makes sense that the model would struggle to pick up the patterns leading to their outcomes. Performance also varied significantly by locale; for instance, Town districts were disproportionately affected by False Positives, while Urban districts showed a higher propensity for False Negatives, meaning actual closures were more frequently missed.

Underlying many of these discrepancies, the model demonstrated reduced reliability when applied to very small districts and those characterized by high internal homogeneity on key structural and demographic indicators, highlighting these profiles as distinct predictive hurdles. These findings suggest that the current feature set might not fully capture the unique operational and financial dynamics of these district types. This highlights areas for future model refinement.

5.2 Policy Implications

Importantly, our prediction task was deliberately executed at the district level rather than the school level. This strategic choice aims to mitigate the risk of reinforcing historical closure biases - an issue extensively discussed in prior work by Pearman and Greene (2022) - often linked to individual school characteristics. The intent is to inform broader resource allocation, strategic planning, and community engagement efforts at a systems level, rather than directly dictating individual school closures.

A key methodological facet of our study is the incorporation of demographic data, which has generated considerable debate in the literature. On the one hand, including demographic predictors has been shown to improve the accuracy of predictions by capturing underlying structural inequalities that are often unobserved in non-demographic variables (Baker et al. 2022; Rahal, Verhagen, and Kirk 2024; Hofman, Sharma, and Watts 2017). This approach addresses concerns related to "color-blind" racism by acknowledging the role of protected attributes in shaping educational outcomes. On the other hand, the use of such data raises potential issues regarding reduced actionability and the risk of reinforcing preexisting biases. In our application, however, the predictive task is strictly aimed at forecasting mass closures at the district level. The results suggest that while demographic data serve as powerful predictors, the model's design - focusing on actionable, system-level predictions - helps counteract the risk of discriminatory decision-making. Our findings thus contribute to a growing body of evidence that carefully integrated demographic data can enhance predictive performance while supporting equitable policy decisions (Hofman, Sharma, and Watts 2017).

From a policy and operational standpoint, the deployment of an early-warning system based on these models, particularly the XGBoost model, could provide critical lead time for school districts facing potential closures. The observed trade-off between optimizing precision (AUC-PR) and maximizing case detection (Recall) is a crucial consideration. Decision-makers should consider models like XGBoost that offer a good balance, or potentially calibrate models towards higher Recall if the priority is to minimize missed detections, even at the cost of more false alarms. The relatively low Positive Predictive Value (0.3133) of the XGBoost model indicates that many districts flagged would not actually experience mass closure, emphasizing the need for careful interpretation and follow-up investigation rather than direct action based solely on the prediction, a reasonable best practice even with a more performant model. However, the high Negative Predictive Value (0.9264) suggests that when the model predicts no mass closure, it is highly reliable.

Importantly, these predictive insights could be used not just for forecasting, but for proactive planning. Districts flagged as at-risk could begin reallocating resources - such as staff, student support services, or capital improvements - to strengthen vulnerable schools. Leaders might also develop community engagement strategies or targeted academic interventions in advance, creating contingency plans that soften the impact or even prevent closures altogether. In this way, predictive modeling becomes not just a reactive tool, but a strategic asset for equity-focused, long-term planning. At the state level, policymakers and political leaders could use these early-warning signals to identify systemic patterns across districts and allocate supplemental funding, technical assistance, or policy flexibility where it is most needed. They might also use model outputs to inform legislation or program design aimed at stabilizing enrollment, enhancing school quality, or supporting districts undergoing demographic shifts. Embedding predictive analytics into governance structures would enable more anticipatory and coordinated responses, helping to avert crisis conditions before they materialize.

5.3 Case Study on Potential Model Use

To exhibit the potential for the model to aid state and district leaders in the process, here we examine The School District of Philadelphia's mass school closures between 2012 and 2013, an instance from our test set which our model correctly predicts.

The closures were prompted by a $1.35 billion deficit, declining enrollment, and underutilized facilities-serve as a cautionary tale for urban educational restructuring. Using quantitative metrics such as facility condition, utilization rates, and academic performance, the district closed 24 schools (10% of its inventory), disproportionately impacting Black students (81% of those displaced versus 58% of the district population). Protests leading up to the decision drew thousands, leading to 19 arrests, including the head

of a national teacher's union ("The Devastating Impact of School Closures on Students and Communities," n.d.; Jack and Sludden, n.d.). While the closures were projected to save $24.5 million annually, the outcomes were mixed: displaced students experienced increased absences and suspensions, receiving schools were strained by sudden influxes, and vacant buildings contributed to neighborhood blight. Academic gains were limited to the small subset of students placed in higher-performing schools, and community opposition highlighted how the process reinforced existing inequities and disrupted neighborhood stability ("What Happened When Philly Closed 30 Schools? New Study Offers Answers," n.d.; "The Impact of the 2013 School Closures" 2023; "The Devastating Impact of School Closures on Students and Communities," n.d.).

A five-year head start on this process could have enabled a more strategic, equitable, and community-centered approach (Hahnel and Marchitello 2023). With extended planning, the district could have implemented tiered intervention systems, allowing struggling schools time and resources to improve or transition gradually, while protecting high-performing or historically significant schools. Early, collaborative planning with communities could have facilitated the re-purposing of closed buildings into mixed-use developments, job-training centers, or charter/community schools, thereby preventing blight and preserving neighborhood vibrancy (Admin 2022). Longer timelines would also have enabled robust student transition supports, such as guaranteed placements in higher-performing schools, teacher cohort transfers to maintain continuity, and curriculum alignment to ease academic disruption (Admin 2022).

Predictive analytics could have been used to anticipate enrollment trends, model building degradation, and proactively address equity concerns through race-conscious closure algorithms and mobility audits. Financially, phased closures would allow for creative solutions like energy performance contracting and targeted reinvestment, potentially tripling savings while offsetting transition costs. Finally, extended engagement would foster trust, allowing for deeper community input, cultural preservation efforts, and the development of impact-weighted closure criteria to safeguard the most vulnerable populations.

We also note that this work need not be done by districts alone. State leaders in the state of Pennsylvania could use this model to monitor for districts that need to implement practices like those mentioned here, require districts to engage in necessary audits, and offer support in preparing or avoiding closures entirely. While our model only has 64% overall accuracy for the state, when we focus on districts in the state with more than 3 schools (given our models struggles mentioned previously), the model is correct on 89% of the mass closure events district-year's for Pennsylvania within the test set. Thus, this model could give states like Pennsylvania support for requiring districts to proactively incorporate equity into their leadership decisions.

Philadelphia's experience demonstrates that the timeline and process of school closure decision-making are critical determinants of equity and effectiveness. Rushed, metrics-driven closures risk amplifying social costs and community disruption, while predictive, participatory, and phased approaches offer opportunities to balance fiscal realities with educational and social stability. For districts facing similar pressures, predictive models that integrate facility use, demographic trends, and community impact-implemented with sufficient lead time-are essential for minimizing harm and maximizing long-term benefits.

6. Conclusion

This study illustrated that machine learning methodologies can play a pivotal role in anticipating mass school closures and thereby inform strategic resource reallocation within U.S. public school districts. Through the use of extensive administrative data and a range of predictive models, our analysis confirms that machine learning can aid district and state leadership by helping forecasting closure events five years ahead. By incorporating demographic predictors in a thoughtful manner, we have shown that it is possible to enhance model performance while addressing concerns of reinforcing historical biases - a critical consideration in educational policy.

This successful application of machine learning methods in this study also joins a larger movement within educational research aimed at leveraging big data for policy innovation. Our work not only demonstrates the feasibility of using administrative data to anticipate mass school closures but also highlights the importance of addressing methodological biases inherent in historical datasets. Given the predictive accuracy achieved the tool developed here could be used to inform proactive strategies. For example, state and district leaders might use high-risk predictions as a prompt to engage in resource auditing, targeted community outreach, or preemptive policy adjustments that could forestall the need for closures entirely. Sans preventative measures, it also allows districts to plan ahead effectively to ensure equitable school closure processes ensure the democratization of precarity in these situations.

However, this study is also subject to several limitations. First, the analysis was confined to data collected prior to the COVID-19 pandemic, which was done to avoid potential data quality issues but may limit the generalizability of findings to the post-pandemic context. Second, the absence of community-level predictors in the current dataset may have constrained model performance, evidenced by the particular inability for the model to distinguish among signal and noise in particular Urban contexts, which research has

shown are particularly heterogeneous (Welsh and Swain, n.d.). Lastly, the handling of missing data through listwise deletion, although a common approach, might have introduced bias, as some models could have benefited from more sophisticated imputation techniques. Addressing these limitations in future research will be critical for advancing the predictive accuracy and practical utility of administrative early-warning systems in education.

Future work could include a follow-up examination that allows for interpretation of model outcomes (especially after improving this final predictive model by address the limitations described previously). That work is warranted to uncover evidence of the underlying mechanisms that drive these predictions and could yield valuable insights and refine recommendations for educational leaders. For example, techniques like SHapley Additive exPlanations - or SHAP - values can generate insights for each individual district about which model features are driving the prediction towards their mass closure probability (Lundberg and Lee 2017). Further, future work should sincerely focus on enhancing administrative data collection to improve the reliability of district and school level data and the ability incorporate community-level covariates. Such improvements would enable researchers to explore the complex interactions between local socioeconomic factors and district-level predictors, thereby facilitating a deeper understanding of the causes behind school closures, improve the predictive accuracy of models, and can provide further levers for change. Expanding the dataset - both in terms of quantity and the temporal window - could also further enhance the robustness of these models. Finally, exploring other applications of administrative education data, such as developing predictive models for teacher support and retention, represents an important and underexplored avenue in the current literature.

In summary, this study illustrates the promise of machine learning in anticipating system-level challenges in education, while simultaneously acknowledging the ethical and practical complexities of deploying such models. Our findings advocate for a balanced approach that leverages advanced analytics to support, rather than dictate, nuanced decision-making processes within school districts in carrying out their responsibilities towards the students, staff, and communities they serve.

Acknowledgements

The authors would like to thank the 2025 cohort of the Stanford Education Data Science program at the Stanford Graduate School of Education for their continued feedback throughout the process of designing and implementing this project, including the Program Director Sanne Smith and program Teaching Assistants Radhika Kapoor and Mridul Joshi. Additionally, we thank the dedicated team of Ritu Khanna, Mele Lau-Smith, Moonhawk Kim from San Francisco Unified School District for their work in trying to implement equitable closure practices that informed the need for this work.

As some of the computing for this project was performed on the Sherlock cluster, we would also like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results (Cavalotti, n.d.).

Reflexivity Statements

Michael L. Chrzan

As the primary researcher in this project examining school closures, I acknowledge that my personal, professional, and academic experiences shape my approach to the research and the interpretation of the data. I have worked closely in and with urban school districts over my career as a K-12 teacher and as a data scientist, which provides me with a deep understanding of the challenges these institutions face at multiple levels – particularly in terms of resource allocation, community impact, and equity. However, I know that these experiences also engender potential biases.

I identify as mixed-race, African-American, cisgender, heterosexual, Christian man living an American middle-class lifestyle after growing up in poverty. I am aware that my background may influence how I understand and interpret the lived experiences of students, parents, and teachers from communities different from my own and that my experience *within* these identities does not make me a spokesperson for the multiple ways these identities exist. For instance, as someone who has experienced the impacts of school closures firsthand, I must be careful not to project my experiences onto the data and presume that my experiences are generalizable, let alone that they generalize well enough to be examined at the scale quantitative research enables.

My positionality as a researcher also intersects with my role as an advocate for equitable education. While this advocacy aligns with the goals of ensuring fairness and equity in school closure decisions, it may introduce a bias toward solutions that prioritize racial and socioeconomic equity over other important factors worth consideration.

Throughout this project, I committed to ongoing self-reflection, particularly in moments when I was required to make interpretive choices. I continue to engage with literature that challenges my assumptions and biases and expands my knowledge of the problem space I am working in, and I regularly seek feedback from peers, mentors, and stakeholders to ensure that my work remains grounded in both academic rigor and community accountability to the communities my work aims to serve.

Francis A. Pearman

I am an assistant professor in the Stanford Graduate School of Education. My research on poverty, inequality, and urban education is both professionally and personally motivated by my experiences as a Black man in the United States. I approach topics like gentrification and school closure with an awareness of how my own identity and institutional affiliation shape the questions I ask, the methods I use, and the ways my findings may be interpreted or applied. While I leverage quantitative tools and big data to surface structural patterns often overlooked in policy debates, I remain critically aware that data is never neutral. My aim is to conduct research that is not only methodologically rigorous but also ethically grounded - attuned to the lived experiences of marginalized communities and accountable to those whose lives and schools are too often treated as sites of experimentation rather than partnership.

Benjamin W. Domingue

I am an associate professor in the Stanford Graduate School of Education, where I apply quantitative methods to a variety of research questions, with a particular focus on issues in educational and psychological measurement. I identify as a cisgender white man who has benefited from structural advantages, including access to elite educational spaces. I recognize that this privilege may create blind spots that limit my understanding of certain lived experiences. I am committed to engaging critically with those limitations by learning from scholars with different perspectives and being reflective about how my identity shapes my research choices and interpretations.

Code

All code used in this study is available at the corresponding author's GitHub repository: https://github.com/mlchrzan/Deeper-Roots. Data is available on request by emailing the author at mlchrzan1@gmail.com.

AI Usage Statement

In the preparation of this research paper, generative artificial intelligence (AI) tools were used to assist with specific aspects of coding challenges and writing refinement. AI was leveraged to troubleshoot programming issues, optimize code efficiency, and explore alternative approaches to data analysis. Additionally, AI-based writing tools were used for grammar and clarity improvements, but all substantive arguments, interpretations, and conclusions were independently developed by the authors. Any outputs generated by AI were critically reviewed and edited to ensure accuracy and alignment with the research objectives.

Appendix

Appendix A1 - Splitting Rules for Random Forest

For handling imbalanced datasets, we used a variety of splitting criteria in the random forests model: **Gini Impurity**, **Hellinger Distance**, and **Extremely Randomized Trees (ExtraTrees)**. Each criterion evaluates different aspects of information gain or separation to determine the best split at each node, improving model sensitivity to the rare positive cases. Below is the mathematical formulation for each criterion.

1. Gini Impurity

Gini impurity measures the likelihood of an incorrect classification by randomly selecting an instance from a dataset. It is calculated as:

$$\text{Gini}(D) = 1 - \sum_{k=1}^{K} p_k^2$$

where:

- D is the dataset at a given node,
- K is the number of classes (in this binary case, $K = 2$),
- p_k is the proportion of instances in class k within D.

For a binary classification, this becomes:

$$\text{Gini}(D) = 2 \cdot p_0 \cdot p_1$$

where p_0 and p_1 represent the proportions of the two classes, typically the majority (negative) and minority (positive) classes. Lower Gini impurity values indicate purer nodes, i.e., nodes with a dominant class distribution. In imbalanced datasets, tuning thresholds for Gini-based splits may help amplify the rare class during tree construction.

2. Hellinger Distance

Hellinger distance is a measure specifically designed to be robust for imbalanced datasets, as it separates classes based on probabilistic distributions. It's used as an alternative to Gini impurity or information gain for calculating node purity. For binary classification, Hellinger distance at a node split is computed as follows:

$$\text{Hellinger}(D_{\text{left}}, D_{\text{right}}) = \sqrt{2 \left(1 - \sum_{k=1}^{K} \sqrt{p_k(D_{\text{left}}) \cdot p_k(D_{\text{right}})} \right)}$$

where:

- D_{left} and D_{right} are the two subsets created by the split,
- $p_k(D_{\text{left}})$ and $p_k(D_{\text{right}})$ are the class probabilities for each subset for class k.

For binary classification, this simplifies to:

$$\text{Hellinger}(D_{\text{left}}, D_{\text{right}}) = \sqrt{2\left(1 - \left(\sqrt{p_0(D_{\text{left}}) \cdot p_0(D_{\text{right}})} + \sqrt{p_1(D_{\text{left}}) \cdot p_1(D_{\text{right}})}\right)\right)}$$

This distance measure favors splits that maximize the separation between classes, making it particularly effective in skewed datasets by focusing on preserving the representation of the minority class.

3. Extremely Randomized Trees (ExtraTrees)

Extremely Randomized Trees (ExtraTrees) employ a different approach by introducing randomness in both the feature selection and the split threshold. Instead of calculating the best possible split based on Gini impurity or entropy, ExtraTrees randomly selects a threshold within the range of feature values. The selected feature and threshold for each split are chosen randomly, without considering any measure of impurity at this point.

The mathematical process of ExtraTrees is as follows:

1. Randomly select a subset of features at each node.
2. For each selected feature j, generate a random threshold t within the range $[\min(x_j), \max(x_j)]$ of the feature values in D
3. Split the node using the feature-threshold pair that yields the greatest separation based on randomness rather than an impurity measure.

This process can increase variance and diversity among the trees, which is beneficial in ensemble methods like random forests, as it allows some trees to capture the minority class even in imbalanced datasets. The randomization reduces the likelihood of overfitting while capturing patterns in the minority class due to increased variance across trees.

Appendix A2 - Hyperparameter Grid Search Details

To optimize four of the models presented, we conducted grid searches over key hyperparameters. This comprehensive grid search enabled us to identify the combination of parameters that maximized model performance.

For the Elastic Net Model, we varied:

- The mixing parameter (alpha) between 0 (ridge regression) and 1 (lasso regression) in increments of 0.1,
- The regularization strength (lambda) was explored over 50 logarithmically spaced values from 10^{-3} to 10

For the Random Forest model, we varied:

- The number of predictors sampled at each split (mtry) over values 2, 3, 4, and 5;
- Three different split criteria - Gini impurity, extratrees, and Hellinger
- Explored minimum node sizes of 1, 3, and 5.

For the XGBoost model, we varied:

- The number of boosting rounds (nrounds: 50, 100)
- Tree depth (max_depth: 3, 5, 9)
- Learning rate (eta: 0.01, 0.05, 0.1)
- Minimum loss reduction (gamma: 0, 0.1, 0.5)
- Subsample ratio of columns (colsample_bytree: 0.6, 0.8, 1)
- Minimum child weight (min_child_weight: 1, 5, 10),
- Subsample ratio of training instances (subsample: 0.8, 1)

And lastly, for the LSTM model, we varied:

- The number of LSTM units (32, 64, 128) to assess model capacity
- Explored dropout rates of 0.2, 0.3, and 0.4 to mitigate overfitting
- The learning rate across two orders of magnitude (0.0001 and 0.00001).
- Batch sizes of 32 and 64
- Training over 50 and 100 epochs
- Sequence lengths of 3, 4, and 5

Appendix A3 - Hyperparameter Outcomes

For elastic-net logistic regression, the optimal values were $\alpha = 0.1$ and $\lambda = 0.001206793$. The α parameter, at 0.1, indicates a stronger inclination towards L2 (ridge) regularization, although L1 (lasso) still plays a role. This suggests that while some feature selection (L1) is occurring to reduce unnecessary predictors and help stabilize coefficient estimates, the primary regularization effect comes from shrinking the coefficient estimates (L2) to mitigate multicollinearity and prevent overfitting. The smaller λ value (0.001206793) indicates that a less aggressive penalty was needed to achieve optimal generalization with the adjusted dataset. This implies the model was able to effectively capture patterns without requiring as much regularization as before.

For the random forest model, the best-performing hyperparameters were mtry = 16, splitrule = extratrees, and min.node.size = 30. These have meaningful implications for the model's behavior and performance. A larger mtry value of 16 suggests that the model considers a large subset of features at each split. This can potentially increase model variance slightly but may also reduce bias if there are more informative predictors that need to be considered together. The use of the "extratrees" splitrule introduces randomness by selecting split points at random rather than optimizing them, which is beneficial for enhancing generalization and reducing overfitting, especially with noisy data or highly correlated features. Lastly, min.node.size at 30 encourages the model to create shallower trees, limiting the model's ability to overfit on smaller patterns in the training data and promoting smoother, more generalizable predictions. Taken together, these hyperparameters reflect a model that is encouraged to explore a wider range of feature interactions while still maintaining control over overfitting through randomization and tree depth.

For XGBoost, the optimal hyperparameters were nrounds = 300, max_depth = 5, eta = 0.05, gamma = 0, colsample_bytree = 1, min_child_weight = 10, and subsample = 0.8. These indicate a model designed for careful, conservative learning with a focus on generalization. A moderate max_depth of 5 limits the complexity of individual trees, helping to prevent overfitting while still allowing the model to capture non-linear relationships. The low learning rate (eta = 0.05) slows down the boosting process, requiring more boosting rounds (nrounds = 300) to converge, which typically leads to more stable and accurate models while gamma = 0) means that there is no minimum loss reduction required to make a split beyond the default behavior of the algorithm. This could allow the model to make splits that are less immediately impactful but might contribute to the overall model performance in later boosting rounds, potentially capturing more nuanced patterns. A min_child_weight of 10 restricts tree growth by requiring a higher minimum sum of instance weight in a child node, promoting simpler and more robust splits. Meanwhile, colsample_bytree = 1 ensures all features are considered for each tree, and subsample = 0.8 introduces row-level randomness to each boosting round, which helps reduce variance. Overall, this configuration prioritizes stability and generalizability while being slightly less restrictive on individual split conditions.

For LSTM, the optimal hyperparameters were batch_size = 32, dropout_rate = 0.2, learning_rate = 0.001, lstm_units = 32, and num_epochs = 30. The batch size of 32 represents a common trade-off between computational efficiency and model stability during training. The dropout rate of 0.2 introduces regularization by randomly deactivating 20% of neurons during training, which is a key technique to reduce the risk of overfitting in neural networks. The learning rate of 0.001 ensures gradual convergence during backpropagation, preventing large, destabilizing weight updates and helping the model find a good minimum in the loss landscape. The LSTM unit count of 32 means that each sequence was processed through 32 memory cells, striking a balance between the model's capacity to capture long-term dependencies in sequential data and avoiding excessive computational costs or overfitting. Finally, training for 30 epochs suggests that the model converged within a reasonable time frame without excessive training cycles that could lead to overfitting on the training data.

Appendix A4 - Error Distributions for Numeric Features

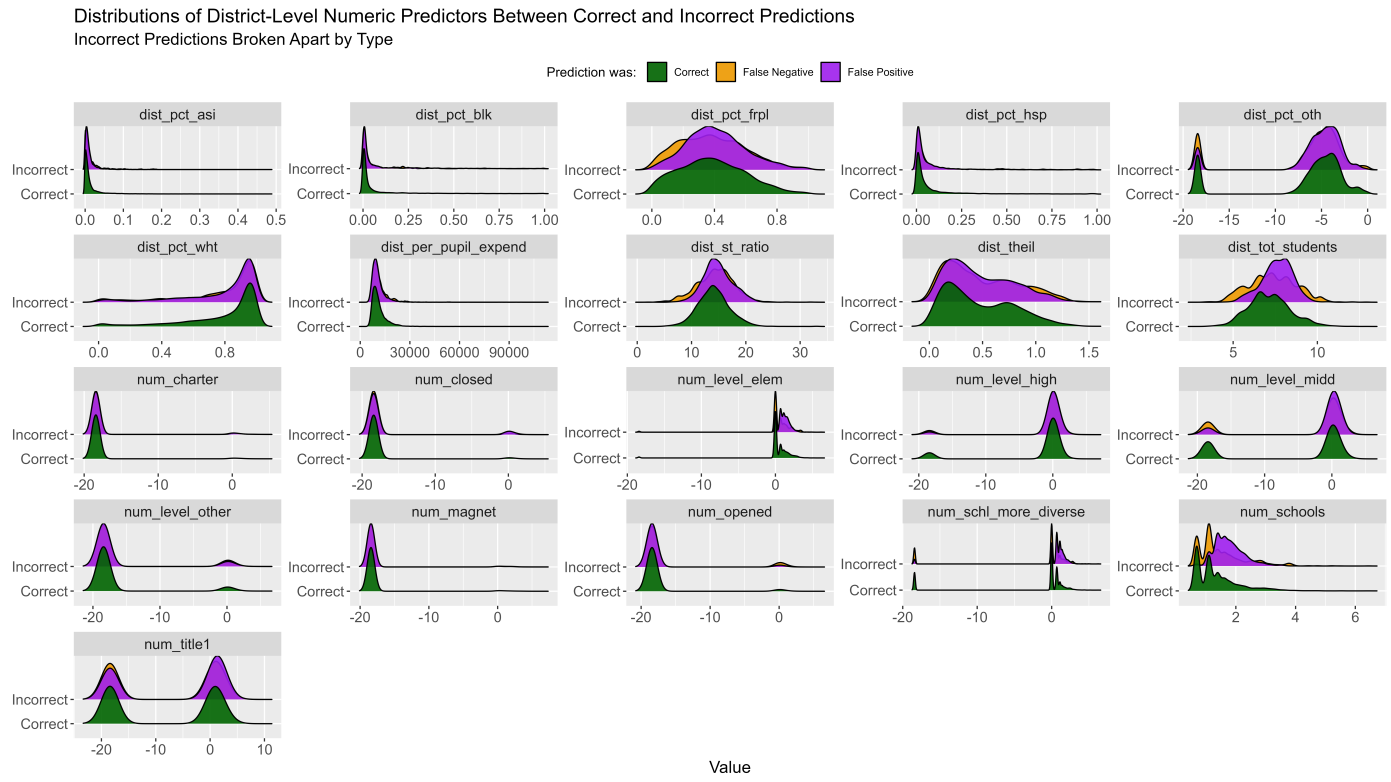

Figure 8: Distributions of District-Level Numeric Predictors Between Correct and Incorrect Predictions, with Incorrect Predictions Broken Down by Type. This figure illustrates the distribution of numeric predictors across correct predictions (blue) and incorrect predictions, further categorized into false negatives (orange) and false positives (purple). Each panel represents a specific predictor variable, comparing the distributions for correct and incorrect classifications. The visualizations highlight differences in predictor values that may contribute to classification errors, offering insights into model behavior and potential areas for improvement. Note: Some predictors have been log-transformed in order to be more carefully examined.

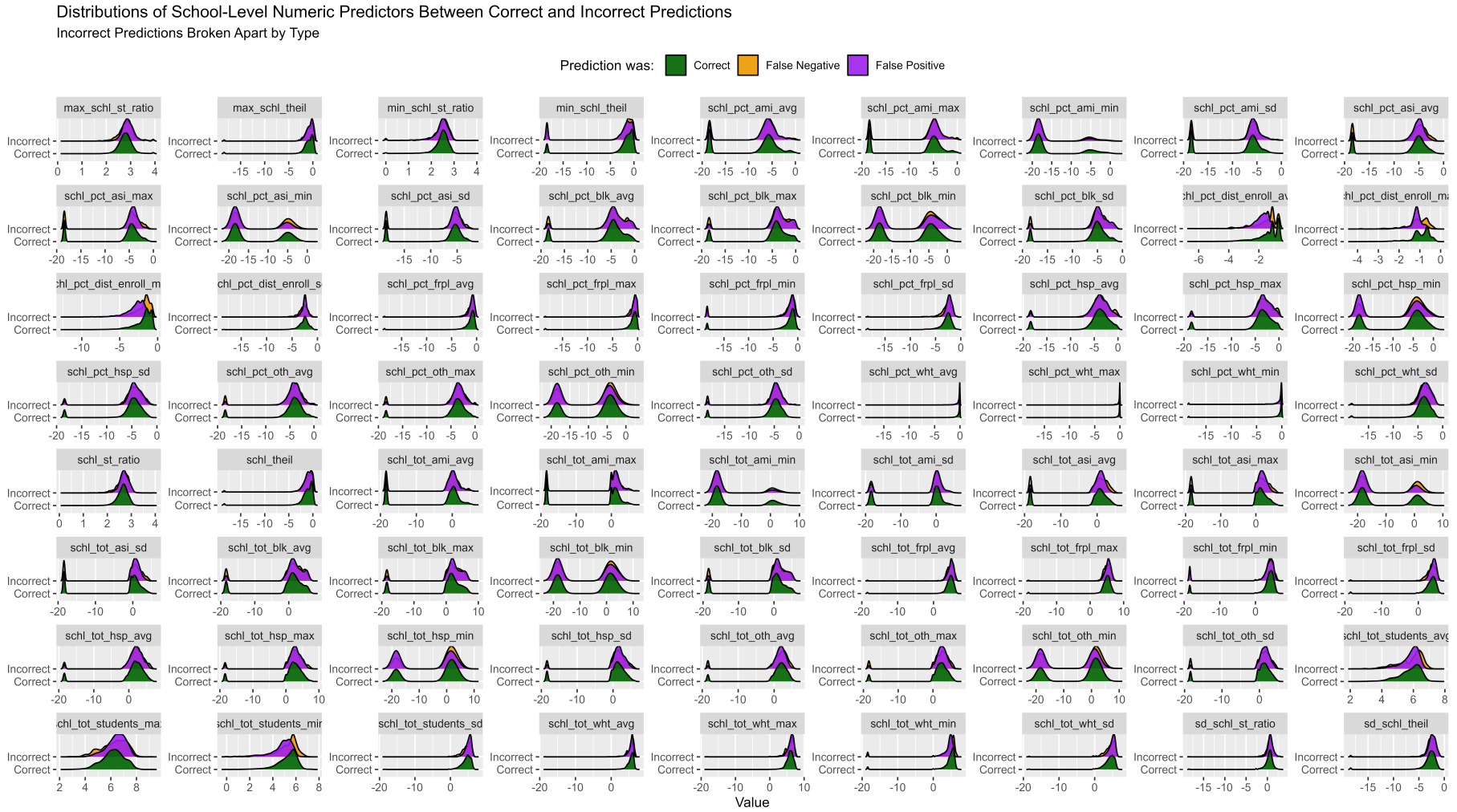

Figure A4: Distributions of Numeric Predictors Between Correct and Incorrect Predictions, with Incorrect Predictions Broken Down by Type. This figure illustrates the distribution of numeric predictors across correct predictions (blue) and incorrect predictions, further categorized into false negatives (orange) and false positives (purple). Each panel represents a specific predictor variable, comparing the distributions for correct and incorrect classifications. The visualizations highlight differences in predictor values that may contribute to classification errors, offering insights into model behavior and potential areas for improvement. Note: Some predictors have been log-transformed in order to be more carefully examined.

References

Abadi, Martın, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, et al. n.d. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems."

Admin, CRPE. 2022. "Closing Schools in a Time of Enrollment Decline – Center on Reinventing Public Education." https://crpe.org/closing-schools-in-a-time-of-enrollment-decline/.

Admojo, Fadhila Tangguh, and Nurul Rismayanti. 2024. "Estimating Obesity Levels Using Decision Trees and k-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions." *Indonesian Journal of Data and Science* 5 (1): 37–44. https://doi.org/10.56705/ijodas.v5i1.126.

Alejo, Anna. 2024. "Denver Public Schools Seeks Community Input Ahead of Possible School Closures - CBS Colorado." https://www.cbsnews.com/colorado/news/denver-public-schools-seeks-community-input-ahead-of-possible-school-closures/.

Baker, Ryan Shaun, Lief Esbenshade, Jon Vitale, and Shamya Karumbaiah. 2022. "Using Demographic Data as Predictor Variables: A Questionable Choice." EdArXiv. https://doi.org/10.35542/osf.io/y4wvj.

Baldo, Nicola, Evangelos Manthos, and Matteo Miani. 2019. "Stiffness Modulus and Marshall Parameters of Hot Mix Asphalts: Laboratory Data Modeling by Artificial Neural Networks Characterized by Cross-Validation." *Applied Sciences* 9 (17): 3502. https://doi.org/10.3390/app9173502.

Billger, Sherrilyn M. 2010. "Demographics, Fiscal Health, and School Quality: Shedding Light on School Closure Decisions." 2010. https://www.iza.org/publications/dp/4739/demographics-fiscal-health-and-school-quality-shedding-light-on-school-closure-decisions.

Bingamon, Brant. 2024. "Austin ISD Budget Shortfall Could Put School Closures on the Table. Austin Chronicle." June 7, 2024. https://www.austinchronicle.com/news/2024-06-07/austin-isd-budget-shortfall-could-put-school-closures-on-the-table/.

Brummet, Quentin. 2014. "The Effect of School Closings on Student Achievement." *Journal of Public Economics* 119 (November): 108–24. https://doi.org/10.1016/j.jpubeco.2014.06.010.

Buolamwini, Joy. 2023. *Unmasking AI: A Story of Hope and Justice in a World of Machines*. First edition. New York: Random House.

Cavalotti, Kilian. n.d. "Sherlock Documentation - Sherlock." https://www.sherlock.stanford.edu/docs/.

Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. 2024. "Xgboost: Extreme Gradient Boosting." https://cran.r-project.org/web/packages/xgboost/index.html.

"Common Core of Data (CCD)." 2024. https://nces.ed.gov/ccd/aboutccd.asp.

Cornman, Stephen Q. n.d. "Documentation for the NCES Common Core of Data National Public Education Financial Survey (NPEFS), School Year 2020–21 (Fiscal Year 2021): Provisional File Version 1a."

Dabek, Filip, Peter Hoover, Kendra Jorgensen-Wagers, Tim Wu, and Jesus J. Caban. 2022. "Evaluation of Machine Learning Techniques to Predict the Likelihood of Mental Health Conditions Following a First mTBI." *Frontiers in Neurology* 12 (February): 769819. https://doi.org/10.3389/fneur.2021.769819.

DeAngelis, Corey A., and Will Flanders. 2019. "The Education Marketplace: The Predictors of School Growth and Closures in Milwaukee." *Journal of School Choice* 13 (3): 355–79. https://doi.org/10.1080/15582159.2019.1595949.

Engberg, John, Brian Gill, Gema Zamarro, and Ron Zimmer. 2012. "Closing Schools in a Shrinking District: Do Student Outcomes Depend on Which Schools Are Closed?" *Journal of Urban Economics* 71 (2): 189–203. https://doi.org/10.1016/j.jue.2011.10.001.

Ewing, Eve L. 2018. *Ghosts in the Schoolyard: Racism and School Closings on Chicago's South Side*. Chicago: The University of Chicago Press.

Ewing, Eve L., and Terrance L. Green. 2022. "Beyond the Headlines: Trends and Future Directions in the School Closure Literature." *Educational Researcher* 51 (1): 58–65. https://doi.org/10.3102/0013189X211050944.

Friedman, Jerome, Trevor Hastie, Rob Tibshirani, Balasubramanian Narasimhan, Kenneth Tay, Noah Simon, Junyang Qian, and James Yang. 2023. *Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. https://cran.r-project.org/web/packages/glmnet/index.html.

Gilblom, Elizabeth, and Hilla Sang. 2021. "Charter School Closure in Ohio's Largest Urban Districts: The Effects of Management Organizations, Enrollment Characteristics and Community Demographics on Closure Risk." *Journal of Education and Learning* 10 (3): p1. https://doi.org/10.5539/jel.v10n3p1.

Gordon, Emily. 2014. "Pittsburgh School Closures: The Impact on Physical and Social Neighborhood Dynamics." Master of Science in Urban Planning, Graduate School of Architecture, Planning; Preservation, Columbia University.

Hahnel, Carrie, and Max Marchitello. 2023. "Centering Equity in the School-Closure Process in California."

Hahnel, Carrie, and Francis A. Pearman. 2023. "Declining Enrollment, School Closures, and Equity Considerations." *Policy Analysis for California Education*, September.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. "Prediction and Explanation in Social Systems." *Science* 355 (6324): 486–88. https://doi.org/10.1126/science.aal3856.

Jack, James, and John Sludden. n.d. "School Closings in

Philadelphia."

Kirshner, Ben, Matthew Gaertner, and Kristen Pozzoboni. 2010. "Tracing Transitions: The Effect of High School Closure on Displaced Students." *Educational Evaluation and Policy Analysis* 32 (3): 407–29. https://doi.org/10.3102/0162373710376823.

Kucak, Danijel, Vedran Juricic, and Goran Dambic. 2018. "Machine Learning in Education - a Survey of Current Research Trends." In *DAAAM Proceedings*, edited by Branko Katalinic, 1st ed., 1:0406–10. DAAAM International Vienna. https://doi.org/10.2507/29th.daaam.proceedings.059.

Kuhn [aut, Max, cre, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, et al. 2024. "Caret: Classification and Regression Training." https://cran.r-project.org/web/packages/caret/index.html.

Lee, Helen, and Lauren Sartain. 2020. "School Closures in Chicago: What Happened to the Teachers?" *Educational Evaluation and Policy Analysis* 42 (3): 331–53. https://doi.org/10.3102/0162373720922218.

Lieberman, Mark. 2023. "How Much Money Do Public Schools Get? The Latest Numbers." *Education Week*, June. https://www.edweek.org/leadership/how-much-money-do-public-schools-get-the-latest-numbers/2023/06.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown.

Oppong, Stephen Opoku. 2023. "Predicting Students' Performance Using Machine Learning Algorithms: A Review." *Asian Journal of Research in Computer Science* 16 (3): 128–48. https://doi.org/10.9734/ajrcos/2023/v16i3351.

Ozsahin, Dilber Uzun, Mubarak Taiwo Mustapha, Auwalu Saleh Mubarak, Zubaida Said Ameen, and Berna Uzun. 2022. "Impact of Feature Scaling on Machine Learning Models for the Diagnosis of Diabetes." In *2022 International Conference on Artificial Intelligence in Everything (AIE)*, 87–94. https://doi.org/10.1109/AIE57029.2022.00024.

Pearman, Francis A., and Danielle Marie Greene. 2022. "School Closures and the Gentrification of the Black Metropolis." *Sociology of Education* 95 (3): 233–53. https://doi.org/10.1177/00380407221095205.

Peetz, Caitlynn. 2024. "As Enrollment Declines, Districts Consider Closing Schools." *Education Week*, January. https://www.edweek.org/leadership/as-enrollment-declines-districts-consider-closing-schools/2024/01.

Pérez Míguez, Carlos, Jose Angel Diaz Arias, Davide Crucitti, Jesús Gómez Fernández, Manuel Piñeiro Fiel, Lucrecia Yanez San Segundo, and Adrian Mosquera Orgueira. 2024. "Smartcytoflow: A Machine Learning Decision Support System for Flow Cytometry Analysis in b Cell Acute Lymphoblastic Leukemia Diagnosis and Monitoring." *Blood* 144 (November): 7483. https://doi.org/10.1182/blood-2024-201439.

Polley, Eric, and Mark van deer Laan. 2010. "Super Learner in Prediction." *U.C. Berkeley Division of Biostatistics Working Paper Series*, May. https://biostats.bepress.com/ucbbiostat/paper266.

Rahal, Charles, Mark Verhagen, and David Kirk. 2024. "The Rise of Machine Learning in the Academic Social Sciences." *AI & SOCIETY* 39 (2): 799–801. https://doi.org/10.1007/s00146-022-01540-w.

Shanker, M., M. Y. Hu, and M. S. Hung. 1996. "Effect of Data Standardization on Neural Network Training." *Omega* 24 (4): 385–97. https://doi.org/10.1016/0305-0483(96)00010-2.

Sriliana, Idhia, I Nyoman Budiantara, and Vita Ratnasari. 2022. "A Truncated Spline and Local Linear Mixed Estimator in Nonparametric Regression for Longitudinal Data and Its Application." *Symmetry* 14 (12): 2687. https://doi.org/10.3390/sym14122687.

Steinberg, Matthew P., and John M. MacDonald. 2019. "The Effects of Closing Urban Schools on Students' Academic and Behavioral Outcomes: Evidence from Philadelphia." *Economics of Education Review* 69 (April): 25–60. https://doi.org/10.1016/j.econedurev.2018.12.005.

Subramaniam, Rajagopalan, Michael Kane, and Lakshmanan Krishnamurti. 2023. "Individualized Prediction of Outcomes of Hematopoietic Cell Transplantation for Sickle Cell Disease: A Machine Learning Approach." *Blood* 142 (November): 1058. https://doi.org/10.1182/blood-2023-178352.

Sunderman, Gail L, and Alexander Payne. 2009. "Does Closing Schools Cause Educational Harm? A Review of the Research."

"The Devastating Impact of School Closures on Students and Communities." n.d. https://populardemocracy.org/news-article/news-and-publications-devastating-impact-school-closures-students-and-communities/.

"The Impact of the 2013 School Closures." 2023. https://philadelphianeighborhoods.com/2023/05/11/education-the-lasting-effects-of-the-2013-school-closures/.

Torre, Marisa de la, and Julia Gwynne. 2009. "When Schools Close Effects on Displaced Students in Chicago Public Schools."

Van Der Laan, Mark J., Eric C Polley, and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1). https://doi.org/10.2202/1544-6115.1309.

Weber, Rachel, Stephanie Farmer, and Mary Donoghue. 2020. "Predicting School Closures in an Era of Austerity: The Case of Chicago." *Urban Affairs Review* 56 (2): 415–50. https://doi.org/10.1177/1078087418802359.

Welsh, Richard O., and Walker A. Swain. n.d.

"(Re)defining Urban Education: A Conceptual Review and Empirical Exploration of the Definition of Urban Education - Richard o. Welsh, Walker a. Swain, 2020." https://journals.sagepub.com/doi/full/10.3102/0013189X20902822.

"What Happened When Philly Closed 30 Schools? New Study Offers Answers." n.d. https://whyy.org/articles/what-happened-when-philly-closed-30-schools-new-study-offers-answers/.

White, Michael J. 1986. "Segregation and Diversity Measures in Population Distribution." *Population Index* 52 (2): 198–221. https://doi.org/10.2307/3644339.

Wright, Marvin N., Stefan Wager, and Philipp Probst. 2024. "Ranger: A Fast Implementation of Random Forests." https://cran.r-project.org/web/packages/ranger/index.html.