COVID-19 Learning Intervention Impact: Data-Driven Evaluation

Mirna Elizondo *Student Member, IEEE*, June Yu, Daniel Payan, Li Feng and Jelena Tešić *Senior Member, IEEE*

***Abstract*—Student learning gain rates in public school systems in the U.S. plummeted during the COVID-19 pandemic, erasing years of improvements. In this body of research, we collect, integrate, and analyze all available public data in the data science pipeline to see if public data can inform and impact learning loss factors. The general data sources were collected from the Census Bureau 2010, USAFACTS, Texas Department of State Health Services (DSHS), National Center for Education Statistics (CCD), U.S. Bureau of Labor Statistics (LAUS), and three sources from the Texas Education Agency (STAAR, TEA, ADA, ESSER). This is the first known study of public data to address the post-COVID educational policy crisis from a data science perspective. To this end, we have developed an end-to-end large-scale educational data modeling pipeline that i) integrates, cleans, and analyzes educational data, ii) implements automated attribute importance analysis to draw meaningful conclusions, and iii). develops a suite of interpretable learning loss prediction models utilizing all data points and features. We demonstrate a novel data-driven approach to discover insights from an extensive collection of heterogeneous public data sources and offer an actionable understanding to policymakers to identify learning-loss tendencies and prevent them in public schools.**

***Index Terms*—Algorithms, Boosting, Data augmentation, Dimensionality reduction, Random Forest**

I. Introduction

COVID-19 also had an impact on teacher preparation [1]. A study indicates how COVID-19 has led many veteran teachers to retire early and novice teachers to consider alternative professions [2]. The COVID-19 pandemic also forced many schools to close across the world [2]. According to the latest UNESCO statistics, there are 43 million students affected by school closures and nationwide closures [3]. Even in high-income countries like the Netherlands and Belgium, learning loss ranged from 0.08 to 0.29 [4], [5]. In a recent article, the global impact of a 5-month school shutdown could generate learning losses with a value of <10 trillion dollars [3].

In a recent paper, the global impact of a school shutdown for five months generated learning losses with a present value of $10 trillion [3]. In the U.S., school district reopening decisions are difficult for policymakers since there is no consensus on the impact of school reopening on the spread of COVID-19 [6]. The learning loss was not uniform across states, as documented for Virginia, Maryland, Ohio, and Connecticut in [7]. Recently, two states, Rhode Island and North Carolina, published two reports estimating the learning losses in these states ([8], [9]. Texas Education Agency also published a report documenting the Loss of learning [10]. There is no clear conclusion on what specifically led to the learning recovery in the states above, and how to recover these learning losses will be the mounting policy and research questions for the next few years and even decades. In the U.S., researchers have disagreed on the impact of school reopening during the spread of COVID-19 [1], [6]. This made it difficult for policymakers to decide when to reopen the school, and these varied between states, counties, and school districts [11]. The learning losses have not been uniform across the board [7], [8]. The Texas Education Agency published a report documenting the 4% Loss in Reading and 15% Loss in math on the STAAR exam and how the negative impact of COVID-19 erased years of improvement in Reading and math [12]. This paper proposes a novel data-driven approach for public data integration and analysis on a scale, automated attribute importance analysis, and robust prediction modeling. As a proof-of-concept, we fuse and analyze multiple open sources of information on public education in Texas before, during, and after the COVID-19 pandemic. We have collected and processed data from nine public websites to find what specific factors were most important for the schools to experience a significant learning loss. We looked into consensus information, public school district population makeup, mode of instruction, income, urban/rural settings, student attendance, county infection rates, and unemployment rates, among hundreds of other factors in 2019, 2021, and 2022. The data-driven findings show that the most resilient factor influencing learning loss in the district is how early or late the students go back to in-person learning. The size and location of a district, along with the amount of money in the area and the Elementary and Secondary School Emergency Relief Fund received, play critical roles in the recovery process. The results identify the significance of various factors in promoting learning recovery in math and Reading, highlighting the importance of considering a district's economic status, size, locale, demographics, and funding.

II. Related Work

In the introduction, we reviewed the related work from qualitative and reporting perspectives. In this section, we will focus on (1) quantitative research and machine learning tools to gain insight from the data on the relationship with the outcome without overfitting the features to the data or (2) the directions for selecting machine learning models for predicting learning loss with tabular data.

The most popular machine learning (ML) techniques (logistic regression, support vector machines, Bayesian belief network, decision trees, and neural network) for data in the wild generally offer an excellent classification accuracy above 70% for simple classification tasks [13]. From a data science perspective, it's critical to refine the selection of modeling approaches. It has been seen that excessive reliance on feature

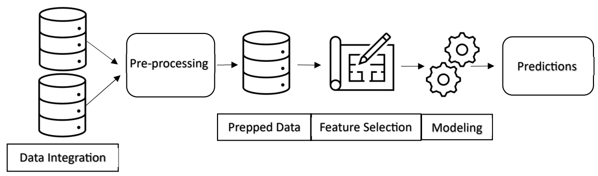

Fig. 1: The proposed noisy tabular data analysis pipeline.

engineering may result in less-than-optimal outcomes when translating domain-specific data. Further analysis of 30 chosen articles indicated that deep neural networks (DNN), decision trees, support vector machines (SVM), and k-nearest neighbor (k-NN) are favored methods for predicting student academic performance [14].

Demographic, academic, familial/personal, and internal assessment factors emerged as the most commonly utilized features for predicting student performance across various metrics such as classroom performance, grade levels, standardized tests performance, etc. [15]. A large-scale data science study examined the Big Fish Little Pond Effect (BFLPE), which describes how individuals often feel better about themselves when they excel in a less competitive environment rather than being average in a highly competitive one, across 56 countries for fourth-grade math and 46 countries for eighth-grade math. This analysis drew on extensive data from the Trends in International Mathematics and Science Study (TIMSS) and employed a straightforward statistical approach [16]. Recent research indicates that state-of-the-art machine learning techniques for tabular data surpass existing methods and exhibit less sensitivity to input bias and noise compared to Deep Neural Networks (DNN) [17].

State-of-the-art gradient-boosted decision trees (GBDT) models such as XGBoost [18], LightGBM [19], and CatBoost [20] are the most popular models of choice when it comes to tabular data. In recent years, deep learning models have emerged as state-of-the-art techniques on heterogeneous tabular data: TabNet [21], DNF-Net [22], Neural Oblivious Decision Ensembles (NODE) [23], and TabNN [24]. Although papers have proposed that these deep learning algorithms outperform the GBDT models, there is no consensus that deep learning exceeds GBDT on tabular data because standard benchmarks have been absent. There is also a shortage of open-source implementations, libraries, and their corresponding APIs for deep learning [25], [26]. Recent studies provide competitive benchmarks comparing GBDT and deep learning models on multiple tabular data sets [25], [27], [28], [29]; however, all of these benchmarks indicate that there is no dominant winner, and GBDT models still outperform deep learning in general. The studies suggest developing tabular-specific deep learning models such that tabular data modalities, spatial and irregular data due to high-cardinality categorical features, missing values, and uninformative features cannot guarantee the same prediction power as deep learning obtains from homogeneous data, including images, audio, or text [27], [29].

III. Proposed Methodology

The study presents a cohesive data science pipeline designed for managing tabular data. It verifies the efficacy of this pipeline by predicting learning deficits in math and reading scores among students in Texas public schools, thus validating its applicability to educational data.

A. Attribute Importance Scoring

This section proposes a novel way to select essential features from the hundreds of features considered. The work compares three different techniques for selecting features in data: filter methods, embedded methods, and wrapper methods. Several algorithms for automated feature selection are tested to evaluate these techniques, and a set of interpretative methods for analyzing feature importance are also provided to avoid the problems of "Garbage In, Garbage Out (GIGO)" and Trivial Modeling.

Attribute Filtering by Mutual Correlations Heterogeneous data tends to have much overlapping information mixed with numerical and categorical data. With this filter method distilling correlated features mutually, we aim to build a quasi-orthonormal attribute space to observe any correlation between two features or a feature and our label. We wanted to avoid artificial weighting of the features in the modeling step, so we utilized this correlation filtering in this section to aggregate linearly related features in our data set into one attribute. To this end, we have expanded several categorical features to multiple binary features as we found that numerous separate categories capture highly overlapping data. The Pearson correlation coefficient ρ measures the linear relationship between two normally distributed variables and is defined in Equation 1:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{1}$$

The $\text{cov}(X, Y)$ represents the covariance between variables X and Y, while σ_X and σ_Y are the standard deviations of X and Y respectively. Pearson's correlation coefficient estimate r, also known as a "correlation coefficient," for attribute feature vectors $x = (x_1, \ldots, x_n)$ with mean $\bar{x}$ and $y = (y_1, \ldots, y_n)$ with mean $\bar{y}$, is obtained via a Least-Squares fit, as defined in Equation 2.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{(y_i - \bar{y})^2}} \tag{2}$$

The $\bar{x}$ and $\bar{y}$ represent the means of vectors x and y respectively. A value of 1 represents a perfect positive relationship, -1 is a perfect negative relationship, and 0 indicates the absence of a relationship between variables. We use features with high correlation coefficients to aggregate them into one attribute, as they are linearly dependent on each other. Eventually, we could keep one attribute, the most highly correlated to our label, of those overlapping features in our analysis. Then, we can combine all binary dummy-coded variables from related categories as a set in variable selection. This approach thus reduces an attribute dimension that provides better interpretability of our attribute set and its importance.

Multi-View Relevancy of the Attribute To select and glimpse the features that affect our prediction models, we compare and contrast ten different approaches from the three methods mentioned above—filter methods, embedded methods, and wrapper methods—to evaluate the importance of features. Every technique of selecting minimum redundancy and maximum relevancy feature sets yields either a set of features chosen or a score of feature importance to reduce the dimensionality of feature space.

Permutation Feature Importance (PFI) is a technique that replaces the values of a feature with noise and measures the change in performance metrics (such as accuracy) between the baseline and permuted data set. This method overcomes some limitations of impurity-based feature importance but can also be biased by the correlation between features[30]. Our final set of features includes any feature with positive mean importance, as the PFI method returns positive values for essential features. We use Random Forests **PFI RF** and Logistic Regression with Ridge Regularization **PFI LR**. All these approaches provide the non-zero scores for all features. **Recursive Feature Elimination (RFE)** is a method training a model on the full set of features in the data set. It then eliminates the features with the smallest coefficients. It continues this process until the 10-fold cross-validation score of the models with Random Forest **RFE RF** and Logistic Regression with Ridge Regularization **RFE LR** on the training data decreases. The final scores are attribute rankings where 1 indicates the most relevant features [31]. **Logistic Regression with Filtering and Regularization** is a technique that uses L1 **LR Lasso** or L1 and L2 **ElasticNet** penalty terms to shrink the coefficients during training. This reduces the coefficients of some features to zero for both, and the remaining non-zero coefficients are considered useful information for prediction. **Feature Importance Random Forest (FI RF)** is a method that leverages the Random Forests machine learning algorithm to determine the importance of each feature. This importance is measured using either the Gini or the mean decrease impurity. A threshold of the 50^{th} percentile of feature importance is used to determine which features should be included in the final set. **Variance Threshold** is a straightforward method to eliminate features by removing features with low variance in the training data set[32]. In this work, the threshold used is 0.8*(1-0.8), meaning that features with 80% similar values in the training data set are removed. The final set of features consists of the k features with the highest variance. Variance Threshold, SFS LR, and SFS KNN provide a binary selection of features.

Sequential Feature Selection (SFS) searches for the optimal set of features by greedily evaluating all possible combinations of features. The method works by adding one feature at a time and assessing each subset based on the 5-fold cross-validation score of logistic regression with ridge regression **SFS RR** and **SFS KNN** models. Overall, we have ten different results: some binary, some numerical, and some rank scores in Alg. 1. We propose several fusion scoring mechanisms for the end user to consider. First, we look into five approaches that filter out features and rank the features by the binary sum outputs. Next, we take five methods that provide scores for all features and rank the attribute importance based on the sum of absolute scores. We transform the scores into rankings and combine them with the filtering and ranking methods to develop the final feature, which is importance ranking.

Algorithm 1: Fusion Scoring Algorithm

Input : Feature Selection Importance Scores(binary, numerical)
Output: Final Fusion Importance Ranking

Initialize BinarySumRankings;
Initialize AbsoluteScoreRankings;
foreach *result in Results* **do**
  if *result is binary* **then**
    Apply filtering mechanism to extract relevant features;
    Calculate the binary sum output for these features;
    Rank the features based on the binary sum outputs;
    Append the ranked features to BinarySumRankings;
  else
    Apply methods to provide scores for all features;
    Calculate the absolute scores for each feature;
    Rank the attribute importance based on the sum of absolute scores;
    Append the ranked attribute importance to AbsoluteScoreRankings;
  end
end
Transform the scores from BinarySumRankings and AbsoluteScoreRankings into rankings;
Combine the rankings derived from both methodologies;
Merge the filtering and ranking methods to generate the FinalFeatureImportanceRanking;
return *FinalRanking*;

B. Prediction Modeling

The second question we are answering in this research is whether the public data we mined from the web is enough to robustly predict school district learning performance during the COVID-19 years.

To this end, we establish five simple baseline models: logistic regression with ridge regularization, Support vector machines (SVM) and K-nearest neighbor (KNN) for nonlinear and non-separable data, random forests, and gradient boosting; and four advanced gradient boosting algorithms: XGBoost, LightGBM, CatBoost, and HistGradientBoosting. Our data fit the description of tabular data. Since gradient boosting approaches showed the most robustness when dealing with heterogeneous tabular data [25], our goal is to assess the predictive power of these nine machine learning models in this real example. Gradient Boosting assembles many weak decision trees, and unlike the random forests, the approach grows trees sequentially and iteratively based on the residuals from the previous trees. Gradient boosting methods handle

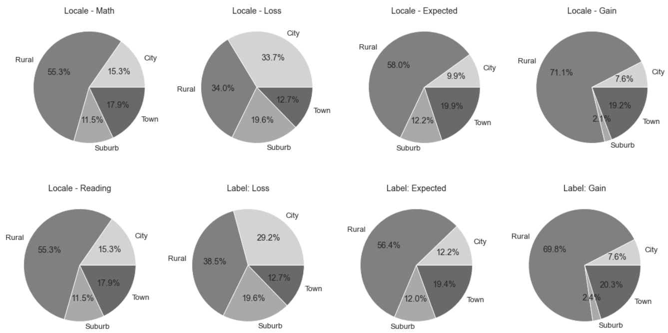

Fig. 2: Exploratory Data Analysis - Locale, Math (top) Reading (bottom)

tricky observations well and are optimized for faster and more efficient fitting using data sparsity-aware histogram-based algorithm.

In contrast to the pointwise split of the traditional Gradient Boosting, which is prone to overfitting, the algorithm's approximate gradient creates estimates by creating a histogram for tree splits. As this histogram algorithm does not handle the sparsity of the data, especially for tabular data with missing values and one-hot encoded categorical features, these algorithms improved tree splits. For example, XGBoost uses Sparsity-aware Split Finding, defining a default direction of tree split in each tree node [18]. Also, LightGBM provides the Gradient-Based One-Side Sampling technique, which filters data instances with a large gradient to adjust the influence of the sparsity, and Exclusive Feature Bundling combining features with non-zero values to reduce the number of columns [19].

IV. Web Data Collection and Processing

A. Data Sources and Collection

We have collected data from nine different public sources as described in Table I. **Common Core of Data (CCD)** [33] is the primary database on public elementary and secondary education supplied by the National Center for Education Statistics (NCES) in the United States. The CCD provided us with public school characteristics, student demographics by grade, and faculty information at the school district in Texas for the fiscal years 2019 and 2021. **State of Texas Assessments of Academic Readiness (STAAR)** data was obtained from the Texas Education Agency (TEA) for the fiscal year 2019 and 2021 for each school district [34]. The STAAR data we collected are the average scores for math and reading tests and the number of students who participated in the grades 3-8 tests. These data also include students' numbers and average scores under various classifications, such as Title 1 participants, economically disadvantaged, free lunch, special education, Hispanic, Black, White, and Asian. **Texas School COVID-19** campus data was provided by the Texas Department of State Health Services (DSHS) [35], including the self-reported student enrollment and on-campus enrollment numbers of the dates September 28, 2020, October 30, 2020, and January 29, 2021, at each school district in Texas. **County COVID-19** data on infection and death cases due to Coronavirus for each Texas County was parsed from USAFacts source[36]. **The average daily attendance (ADA)** is a sum of attendance counts divided by days of instruction per school district and provided by TEA. **Elementary and Secondary School Emergency Relief (ESSER) Grant** data provided by TEA summarizes COVID-19 federal distribution by TEA to school districts for the fiscal years 2020, 2021, 2022, and 2023. The **Local Area Unemployment Statistics (LAUS)** data [37] was parsed from the U.S. Bureau of Labor Statistics (BLS) for the years 2019 and 2021 to examine the workforce impact on learning loss in the counties. **Census block group 2010** data [38] were included to see if the county's general population characteristics make a difference in learning loss. At the end of the initial data integration merging data from nine sources by matching school district I.D. and county FIPS code, the data set represents 1,165 school districts of Texas located in 253 counties with 506 features, consisting of 1 categorical and 505 numerical.

TABLE I: Data from nine different sources are integrated by matching school district I.D. and county FIPS code for 1,165 school districts with 506 features in 253 Texas counties

Data Frame	Data Source	Level	RowXCol
CCD	National Center for Education Stat [33]	District	1189X66
STAAR	Texas Education Agency [34]	District	1184x217
TEA	Texas Education Agency [10]	District	1182x217
ADA	Texas Education Agency [39]	District	1226X3
ESSER	Texas Education Agency [40]	District	1208X6
Census	Census Bureau 2010 [38]	County	254x37
Covid	USAFacts [36]	County	254X8
LAUS	U.S. Bureau of Labor Statistics [37]	County	254X13
Covid	DSHS [35]	District	1216X7

CARES ESSER I 20, ARP ESSER III 21 features are part of the Elementary and Secondary School Emergency Relief (ESSER) grant programs, which are federal funds granted to State education agencies (SEAs) providing Local education agencies (LEAs) to address the impact due to COVID-19 on elementary and secondary schools across the nation; thus, the funds have been administered by Texas Education Agency (TEA) and allocated in each school district in Texas [40], [41]. **CARES ESSER I:** Authorized on March 27, 2020, as the Coronavirus Aid Relief and Economic Security (CARES) Act with $13.2 billion. Our data shows the allocation amount for the fiscal year 2020. **CRRSA ESSER II:** Authorized on December 27, 2020, as the Coronavirus Response and Relief Supplemental Appropriations (CRRSA) Act with $54.3 billion. Our data show the allocation amount for the fiscal year 2021. **ARP ESSER III:** Authorized on March 11, 2021, as the American Rescue Plan (ARP) Act with $122 billion. The data show the allocation amount for the fiscal year 2021. **ESSER-SUPP:** Authorized by the Texas Legislature to provide additional resources for not reimbursed costs to support students not performing well educationally. The data was collected from March 13, 2020, to September 30, 2022.

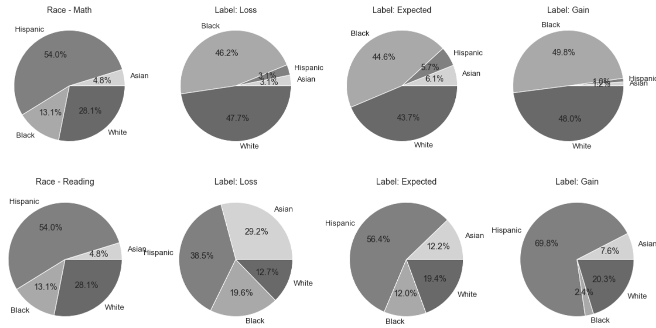

Fig. 3: Math (top) and Reading (bottom) scores broken by race

B. Data Aggregation and Filtering

To help policymakers make more informative decisions on learning recovery with localized efforts in each school district, we collected data from nine different sources as described in Table I to answer our research questions: i) Are students from low-income backgrounds and minority students experiencing more learning loss? ii) Do students of different grade levels experience learning Loss differently? iii) Does the school or school district reopening decision influence learning loss experienced by students? iv) Is the mode of instruction (hybrid, remote, in-person) related to learning loss? v) Is the school or district attendance negatively correlated with learning loss? vi) Does the local or regional infection rate lead to more learning loss? vii) Does the local unemployment rate negatively affect learning losses? If we can answer these questions with our approach, we can also identify resilient factors in learning recovery for Texas public schools.

Primarily, we gathered the Common Core of Data (CCD) [33], which is the primary database on public elementary and secondary education supplied by the National Center for Education Statistics (NCES) in the United States. The CCD provided us with public school characteristics, student demographics by grade, and faculty information at the school district in Texas for the fiscal years 2019 and 2021. Then, we merged the CCD data with the State of Texas Assessments of Academic Readiness (STAAR) data [34] from the Texas Education Agency (TEA) for fiscal years 2019 and 2021 at each school district. The STAAR data we collected are the average scores for math and reading tests and the number of students who participated in the grades 3-8 trials. These data also include students' numbers and average scores under various classifications, such as Title 1 participants, economically disadvantaged, free lunch, special education, Hispanic, Black, White, and Asian. Next, our data merged with COVID-19 campus data from the Texas Department of State Health Services (DSHS) [35], including the self-reported student enrollment and on-campus enrollment numbers of the dates September 28, 2020, October 30, 2020, and January 29, 2021, at each school district in Texas. Additional COVID-19 data involved confirmed infection and death cases [36] due to Coronavirus at each county from USAFacts. Also, the average daily attendance (ADA) [39], which consists of the sum of attendance counts divided by days of instruction, and data from the Elementary and Secondary School Emergency Relief (ESSER) Grant Programs [40] – COVID-19 relief funding – were collected from TEA for school district level. The ADA data for fiscal years 2019 and 2021 were added to our data to see the impact of district attendance, and the ESSER data reflect the localized efforts of TEA allocating the grant amount at each school district in the fiscal years of 2020, 2021, 2022 and 2023. Also, we combined the Local Area Unemployment Statistics (LAUS) data [37] from the U.S. Bureau of Labor Statistics (BLS) for the years 2019 and 2021 to examine the negative impact of the unemployment rate on learning loss at the county level. The census block group 2010 data [38] were included to grasp demographic characteristics at a county for the general population. After completing the initial data integration, data from nine sources were merged by matching the school district I.D. and county FIPS code and then integrated based on the district I.D. and county FIPS code.

TABLE II: Example of 2019 and 2021 attribute aggregates

Attribute	Aggregated Attribute	Data
Total Schools 2020-2021 Total Schools 2018-2019	Total Schools Diff	CCD, NCES
% Title 1 Eligible 2020-2021 % Title 1 Eligible 2018-2019	% Title 1 Eligible Diff	CCD, NCES
% Hispanic 2020-2021 % Hispanic 2018-2019	% Hispanic Diff	CCD, NCES
% Grades 1-8 2020-2021 % Grades 1-8 2018-2019	% Grades 1-8 Diff	CCD, NCES
% Tested Reading G3 2020-2021 % Tested Reading G3 2018-2019	% Tested Reading G3 Diff	STAAR, TEA
Unemployed Rate 2021 Unemployed Rate 2019	Unemployed Rate Diff	LAUS, BLS
% ADA 2020-2021 % ADA 2018-2019	% ADA Diff	ADA, TEA

Out of the 506 features examined, 416 exhibit missing values across three data sources, ranging from one to 88% within our dataset. Specifically, 408 features stem from STAAR and TEA data, six from CCD and NCES, and two from COVID and DSHS data. Among these 416 features, 332 have less than 20% missing values, while 24 have more than 80% missing values. The Distribution is depicted in Figure 4.

The features with over 20% missing values are predominantly from the STAAR data, related to average scores and participants in the STAAR tests, and we have removed those features from the STAAR data. We have also dropped the school districts that do not have the CCD and COVID data and ended up with 955 public school districts in Texas to analyze, with a total of 119 features with no missing values. Out of 119 features, we aggregate the 58 features that duplicate the data for 2019 and 2021 into 29 differential features as illustrated in Table II. For example, Total Schools 2020-2021 and Total Schools 2018-2019 features are aggregated into Total Schools Diff, and the total number of features is reduced to 90.

C. Data Labeling

Our data set is unlabeled; thus, the process begins by normalizing the individual grade scores, ensuring consistency across different scales, through the equation Normalized Score $= \frac{\text{grade score}}{\max(\text{grade score})}$. Following this, the district average is calculated by summing up the scores of grades G3 to G8 and dividing by the total number of

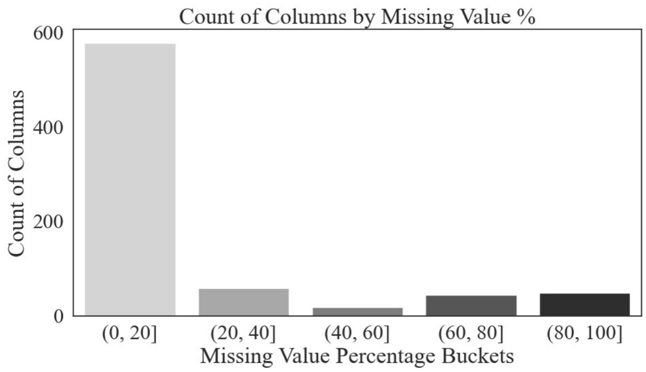

Fig. 4: Percentage of missing values for 416 features in the aggregated data. 2022-2023

grades. This provides an overarching view of the academic performance within the district, represented by the equation District Average $= \frac{G3+G4+G5+G6+G7+G8}{\text{Total number of grades}}$. Subsequently, the percentage loss in performance is computed over time intervals, reflecting changes in educational outcomes. This is expressed through the equations $\%$ Loss $= \frac{\text{Avg 2021}-\text{Avg 2019}}{\text{Avg 2019}}$ and $\%$ Loss $= \frac{\text{Avg 2022}-\text{Avg 2021}}{\text{Avg 2021}}$. Finally, the obtained loss percentages are interpreted to categorize the observed trends. These interpretations are encapsulated in the labeling criteria: Learning Gain if the Loss is greater than zero, Expected if the Loss equals zero, and Learning Loss if the Loss is less than zero. This comprehensive process enables the assessment of educational trends, facilitating informed decisions and interventions to enhance learning outcomes. *Comment: Should the labeling be moved up to explain what we mean by Loss early on?? –Jelena*

When analyzed by year, the normalization process encompasses various facets of educational institutions, such as the count of operational public schools, identification of Schoolwide Title 1 designations, and Title 1 eligibility. Additionally, it includes insights into the educational workforce, encompassing Full-Time Equivalent (FTE) teachers and overall staff counts, along with lunch program statistics like free and reduced-price lunch participants. Race and ethnicity distributions among Asian, Hispanic, Black, and White demographics, delineated by grade groups from Prekindergarten to Grade 12, are normalized for accurate assessment. Attendance metrics undergo normalization in terms of average daily attendance (ADA) and as a percentage of total students per district. By grade, the standardization involves the Percentage of students taking the STAAR reading and math tests, with average scores ratio-ed to the 100th percentile in each grade, regarding population metrics, normalization factors in confirmed COVID-19 cases, and deaths as percentages of the county population. It also encompasses race/ethnicity and age group distributions as a percentage of the county population in 2010. Lastly, when assessed by date, the normalization process considers the Percentage of students on campus on September 28, 2020, October 30, 2020, and January 29, 2021. Additionally, it categorizes different household types and housing units as percentages of the total number of households and housing units in 2010, respectively. This comprehensive standardization methodology ensures a consistent and comparable analysis across diverse data points and timeframes.

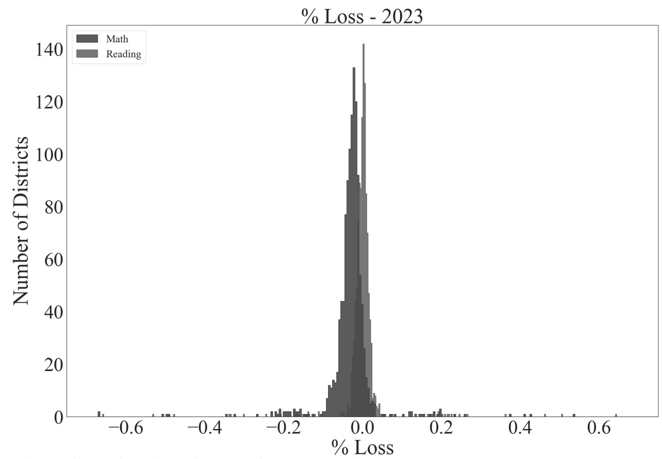

Fig. 5: Distribution of normalized STAAR scores between 2019-2023. More school districts in Texas faced learning loss in math than in reading

The Distribution of the loss values in Figure 5 informed us to set a threshold determining the Loss and Gain. The Distribution shows that more districts have experienced Loss in math as the median for math (-0.03) is lower than for Reading (0). We proceeded with further analysis and prediction separately for math and Reading. The middle 50% of school districts are labeled as "Expected," the loss values below the 25th percentile are set to be "Loss," and the loss values above the 75th percentile become "Gain."

With the data labeled as learning loss, Expected, and Gain, we analyzed each in-depth concerning a correlation between features and the label. Figure 3 illustrates (a) White students are correlated to our label as they are the majority population for Gain and decreased towards Loss label; (b) Hispanic students are 2/3 of Loss students then reduced as for Expected and Gain labels for both math and Reading. Also, we realized that the locale of school districts is correlated to the label learning loss, as illustrated in Figure 2 (a), confirms that over half the schools are located in rural areas in Texas despite the positive correlation between rural areas and the label from Loss to Gain; however, Loss occurring in schools located in City and Suburb areas increasingly appeared in (b) and (c).

D. Data Pre-Processing

In the dataset "LossA," we condense 58 duplicated features representing data for 2019 and 2022 into 29 differential features, as demonstrated in Table II. For instance, features such as "Total Schools 2020-2021" and "Total Schools 2018-2019" are combined into a single feature, "Total Schools Diff," resulting in a total reduction to 90 features. Conversely, in the dataset "LossB," these features are treated independently. The comparison of their importance in modeling is detailed in Section III-A. This dataset encompasses 506 attributes across 1,165 school districts.

We propose dimensionality reduction to obtain interpretability and identify the resilience factors for

TABLE III: Resilient factors for Top 15 (math) and 14 features (Reading) Low income and grade level are both subjects' most impactful resilience factors

Resilient Factor	Math	Reading
Low-income	4	5
Grade Level	4	4
Race/Ethnicity	3	1
Mode of instruction	2	3
Attendance	1	0
Census demographics	1	0
Unemployment	0	1

learning Loss as follows:

LossA We remove noise and missing values from the data and then aggregate attributes conveying the same information for 2019 and 2021. In turn, we successfully reduced the number of attributes to 90 to finally adopt the attribute selection methods Section III-A.

LossB In this approach, raw integrated data is utilized for the Gradient Boosting experiment without normalization but with missing values. Unlike LossA, where attributes are aggregated and normalized to reduce dimensionality and enhance interpretability, LossB treats each feature individually. This means that each feature is considered independently without any aggregation or normalization across different attributes. While this approach may result in a more prominent feature space and potentially increased computational complexity, it allows for a more detailed analysis of individual features and their impact on learning loss. By examining each feature in isolation, we aim to gain insights into the specific factors contributing to learning loss without the influence of normalization or aggregation techniques. This approach provides a complementary perspective to LossA and allows for a comprehensive exploration of the dataset.

V. RESULTS

A. Attribute Importance Analysis

We executed the ten different feature selection approaches described in Section III-A to detect the resilient factors for learning Loss due to COVID-19 using the data set with 90 features and 955 school districts in Texas as a baseline.

As we discriminate the subjects, math and Reading, on predicting learning loss, the feature selection process has been repeated for each subject separately. Variance Threshold, SFS Ridge, and SFS KNN provide a binary selection of features. ElastiNet Logistic Regression fit for the Gain and Loss provides scores for a subset of coefficients that are not zeroed out. R.F. feature importance, R.F. permutation, and Ridge permutation importance offer non-zero scores to all 90 features, and RFE ridge regression and RFE Random forest provide attribute ranking. Figure 7 sums up the filtering results. The five methods ranked 18 features as top importance and agreed to exclude 33 descriptors, mainly from the workforce, Census, and COVID data sources. The difference between free lunch and the COVID deaths in the county had little impact on learning loss. Next, we sort the remaining 57 features using Random Forest feature Importance, Random Forest permutation, Ridge permutation importance, RFE Ridge and R.F. scores, and ElastiNet Gain and ElastinNet Loss. Since all of them have importance ranking per feature (including the sign), we first normalize the scores for each method and then sum them.

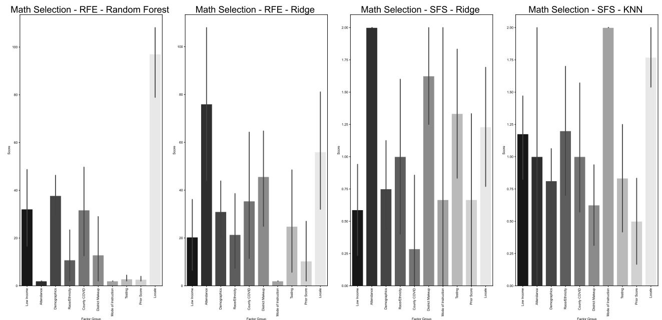

Fig. 6: Analysis of Resilience Factors: Random Forest Recursive Feature Elimination, Ridge Recursive Feature Elimination, Sequential Feature Selection with Ridge, and K-Nearest Neighbors to identify significant math resilience factors

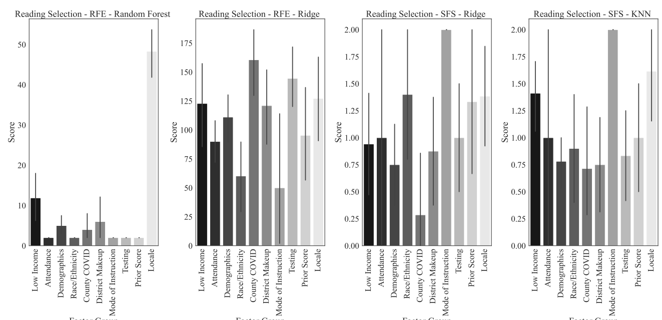

Fig. 7: Analysis of Resilience Factors: Random Forest Recursive Feature Elimination, Ridge Recursive Feature Elimination, Sequential Feature Selection with Ridge, and K-Nearest Neighbors to identify significant reading resilience factors

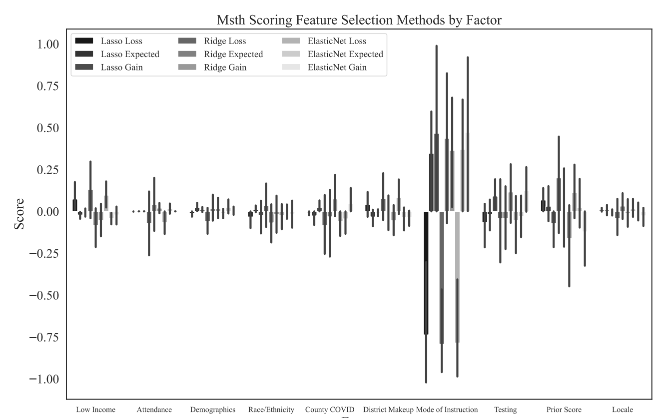

Fig. 8: Comparison of Loss, Expected, and Gain Metrics for Lasso, Ridge, and ElasticNet Regularization Techniques to identify significant math resilience factors

First, we aggregate five filtering method outcomes for Reading and math: Variance Threshold, SFS KNN, SFS Ridge, and ElastiNet Gain and ElastiNet Loss binarized coefficients. The Initial Importance Values are the raw scores from the machine learning methods and are initially tricky to compare due to their non-uniformity. The Binary Selection Values are

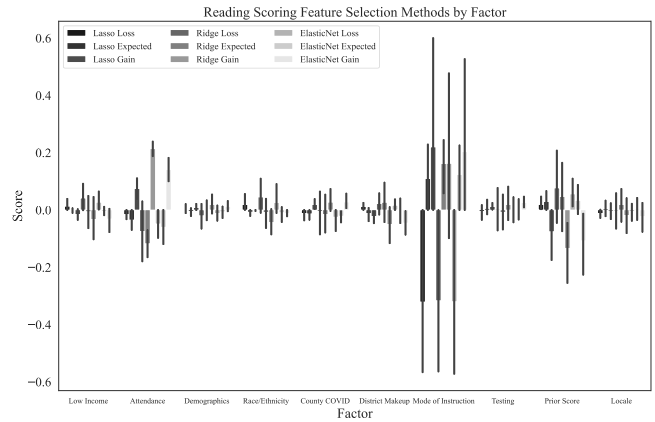

Fig. 9: Comparison of Loss, Expected, and Gain Metrics for Lasso, Ridge, and ElasticNet Regularization Techniques to identify significant reading resilience factors

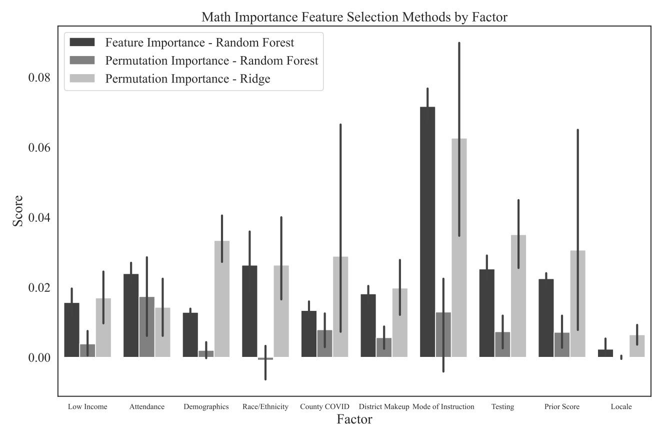

Fig. 10: Analysis of Resilience Factors: Feature Importance and Permutation Importance - Random Forest; Permutation Importance - Ridge comparison to identify the significant math resilience factors

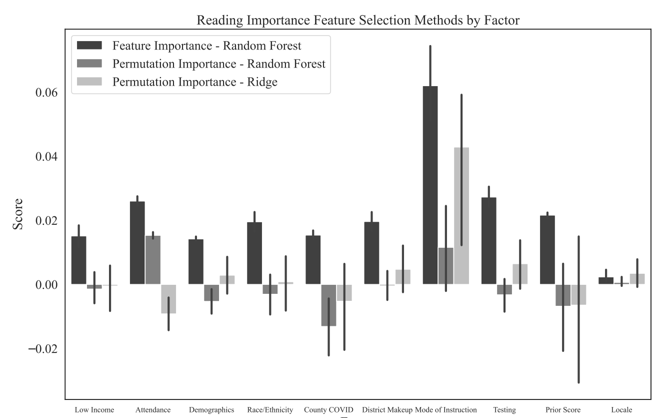

Fig. 11: Analysis of Resilience Factors: Feature Importance and Permutation Importance - Random Forest; Permutation Importance - Ridge comparison to identify the significant reading resilience factors

the first output transformation, where we binarize all scores as SFS KNN, SFS RR, and Variance Threshold, which are already binary. To transform the features into a binary format,

TABLE IV: Method index and corresponding full name and output format. Feature dimensions after the method is applied are in the last two columns for math and reading

Index	Method	Output	*Math*	*Reading*
LR Lasso	Logistic Regression with L1 Reg.	score	51	51
LR ElasticNet	Logistic Regression with L1+L2 Reg.	score	41	45
PFI LR	Permutation Feature importance for LR L2	score	28	82
PFI RF	Permutation Feature importance for Random Forest	score	70	26
FI RF	Feature Importance Random Forest	score	45	45
VR	Variance Threshold	binary	20	20
SFS LR	Sequential Feature Search with Ridge Regression	binary	45	45
SFS KNN	Sequential Feature Search KNN	binary	45	45
RFE LR	Recursive Feature Elimination with Ridge Regression	rank	6	5
RFE RF	Recursive Feature Elimination Random Forest	rank	36	36

we use the following approach: For RFE methods, we retain only the rank of one feature and assign a value of 1 to it while the others get a value of 0. For logistic regression, we give a +1 score to features with a positive coefficient and -1 to those with a negative coefficient, while the coefficients with a value of 0 are ignored. For feature importance, we select the top 50% of features with positive scores and assign a value of 1 to them, while the others get a value of 0. For the importance of permutation features, we give 1 to features with positive scores and 0 to those with negative or zero scores. Finally, we sum the scores and sort the feature importance for each subject out of 9. The Impact Score Values are the second transformation of the output. They are obtained by normalizing the scores of each method by dividing them by their sum of overall features. This normalization ensures that each feature contributes equally to the final ranking. Next, we calculate the absolute value of the normalized score for each attribute and sum them up to create a feature ranking. The top 20 features with the highest scores are selected for math and reading by prioritizing the impact score, as it combines both binary and non-zero scores. In contrast, the binary score is used as a secondary measure to understand the importance. The number of features selected is based on a drop in impact score after the top 20 features, labeled the cutoff point. Secondary labels were also applied to the features to understand what "type" of the feature was most significant. Overall, this approach allows us to compare the relative importance of each feature and identify the most important ones.

Table IV indicates the dimension each approach reduces to the various numbers. RFE with random forests only selected 6 and 5 features for math and Reading, respectively; however, the PMI method selected the most significant number of features for both subjects: 70 features for math using random forests and 82 features for reading using ridge regression. The 2022-2023 importance ranking of the features resulting from the ten approaches is shown in Table V, (a) Top 20 for math, and (b) Top 20 for reading selected by six or more feature selection methods selection results are listed in Table VI.

The most significant feature predicting learning loss in math is *% of Campus 10/30/20*, the enrollment of students

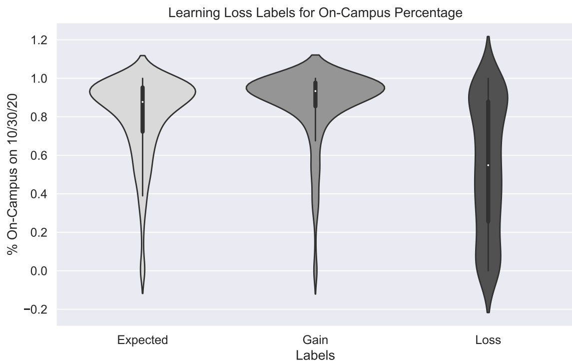

Fig. 12: Analysis on the most important feature for predicting learning loss: % On Campus 10/30/20. Gain and Expected label school districts have more students who went to school on October 30, 2020,

in the campus district on October 30, 2020, representing the mode of instruction. For Reading, three critical features were selected, all of which were resilience factors related to the Low-income backgrounds of students: *CARES ESSER I 20* (Coronavirus Aid, Relief and Economic Security (CARES) grant amount in 2020), *ARP ESSER III 21* (American Rescue Plan Act (ARP) grant amount in 2021), *% Reduced-price Lunch Diff* (Reduced-price Lunch Eligible Students Difference in percent between 2019 and 2021). The top 20 in math in Figure 6 and the top 20 in Reading in Figure 7 were important features selected by six or more selection methods. Low income and Grade level are the most influential resilient factors to predict learning loss for Math and Reading, as shown in Table III. Race/Ethnicity and mode of instruction continued to be decisive, resilient factors for both subjects; on the other hand, Attendance and Census demographics are considered significant factors only in math, and Unemployment is essential only for Reading.

Although we now realize these essential features can identify the resilient factors for Loss or Gain in learning due to the COVID-19 pandemic, it is still unknown whether those features positively impact learning. For example, we analyzed positive or negative correlations between the most critical features and our label, Loss, Expected, or Gain, in math and Reading.

The students who experienced Loss in Reading received more significant funding for all funding programs on average than the students who participated, gained, or Expected in the same subject. The ESSER funds have been distributed to proper districts in need of financial help for adapting and preparing for learning Loss due to COVID-19 as the ESSER fund amounts are calculated by a formula based on Title I, Part A grant that is considered as a poverty proxy [40], [41].

Figure 12 indicates that *% of Campus 10/30/20* is positively correlated with Gain as the Distribution of school districts with the highest proportion of students on a campus populated more for Gain and Expected in math; however, the students experienced Loss are inhabited the most where the enrollment is 0%. It is clear that in-person classes, the mode of instruction, were the key to avoiding Loss in math.

TABLE V: Top 20 Attributes for 2022 data

Math		
Feature	**Impact Score**	**Binary Score**
Median Household Income	6.6214	5
Total Students 2018-2019	6.2266	7
Total Students 2020-2021	6.1418	6
Total Students 2021-2022	6.1089	7
Rural: Distant	6.0521	3
# of Families 10	5.8406	4
Average Annual Pay	5.8252	2
ARP ESSER III 21 NORM	5.7606	3
CARES ESSER I 20 NORM	5.7590	4
Rural: Remote	5.7405	3
# of Housing Units 10	5.7040	3
# of Households 10	5.7005	3
Per Capita Income	5.6966	3
% of Pop Under 18 in Poverty	5.6840	3
Median Age Male 10	5.6834	3
County Population	5.6794	2
% of Pop in Poverty	5.6740	2
CRRSA ESSER II 21 NORM	5.6704	2
Median Age 10	5.6540	2
Median Age Female 10	5.5848	1
Reading		
Feature	**Impact Score**	**Binary Score**
Average Annual Pay	6.4049	3
Per Capita Income	6.2658	4
Total Students 2021-2022	6.0159	6
County Population	5.9212	5
# of Families 10	5.9140	6
Total Students 2018-2019	5.8932	5
Total Students 2020-2021	5.8707	5
# of Households 10	5.8357	5
% of Pop Under 18 in Poverty	5.8048	4
CRRSA ESSER II 21 NORM	5.8094	4
Median Household Income	5.7823	5
# of Housing Units 10	5.7847	4
Median Age Female 10	5.7612	3
% of Pop in Poverty	5.7661	4
Rural: Distant	5.7042	3
CARES ESSER I 20 NORM	5.7132	4
ARP ESSER III 21 NORM	5.6947	4
Median Age Male 10	5.6591	3
Median Age 10	5.5855	2
Rural: Remote	5.5603	2

B. Modeling Learning Loss from Public Data

The data sets have been randomly split into 80% of the training set and 20% of the test set with shuffling and stratification on the label. We use performance metrics suitable for prediction problems to find the best model. The accuracy score for both Gain and Loss is used to get a big picture, and the F1 score is used for an in-depth measure as it harmonically includes the precision and the recall scores. Matthews' correlation coefficient (MCC) considers true negatives, class imbalance, and multi-class data. Each model runs with a 10-fold cross-validation of GridSearch to find optimal hyperparameters. As the boosting algorithm trains weak learners iteratively, early stopping reduces training time and avoids overfitting. At every boost round, the model evaluates and decides whether to stop or continue the training when the model shows no more improvement for a certain number of consecutive rounds in terms of the evaluation metric specified as the fit parameter. For early stopping, a validation set, the split test set at the beginning of the modeling process, and the number of early stopping rounds that are set to 10% of the

TABLE VI: Top 20 Attributes - 2023

Math		
Feature	**Impact Score**	**Binary Score**
Total Students 2018-2022	6.2266	7
% On Campus 10/30/20	1.4300	5
% White Students 2020-2021	0.6324	5
% Tested Math - G3 2020-2021	0.7360	5
Median Household Income	6.6214	5
% On Campus 09/28/20	0.6603	4
% White Students 2018-2019	0.6082	4
% On Campus 01/29/21	0.6892	4
Total Staff 2020-2021	0.6071	4
Total Teachers 2020-2021	1.2075	4
# of Families 10	5.8406	4
CARES ESSER I 20 NORM	5.7590	4
% Tested Math - G5 2018-2019	0.8683	4
% Asian Students 2018-2019	0.5235	4
City: Small	0.4131	4
Suburb: Mid-size	0.3970	4
% White Students 2021-2022	0.5163	3
% Hispanic Pop 10	0.1885	3
% Tested Math - G5 2020-2021	1.0852	3
% Tested Math - G6 2020-2021	0.2561	3
Reading		
Feature	**Impact Score**	**Binary Score**
# of Families 10	5.9140	5
Total Students 2021-2022	6.0159	5
County Population	5.9212	5
# of Households 10	5.8357	4
Total Students 2018-2019	5.8932	4
Total Students 2020-2021	5.8707	4
# of Housing Units 10	5.7847	4
CRRSA ESSER II 21 NORM	5.8094	4
% Asian Pop 10	0.3936	3
% Prek 2018-2019	0.3689	3
% Tested Reading - G7 2021-2022	0.3498	3
Median Household Income	5.7823	3
Total Teachers 2020-2021	0.8628	3
% On Campus 10/30/20	0.4773	3
% Tested Reading - G8 2018-2019	0.2727	3
% White Students 2020-2021	0.5280	3
% White Students 2021-2022	0.3913	3
% of Pop Under 18 in Poverty	5.8048	3
% of Pop in Poverty	5.7661	3
City: Small	0.1889	3

maximum number of boosting iterations are provided.

We employed five state-of-the-art machine learning models, including ridge regression, support vector machine (SVM), k-nearest neighbors (KNN), random forests, and gradient boosting, on our dataset. These models were trained using our complete set of 90 features and ten additional feature groups derived from various feature selection techniques. These techniques encompassed Recursive Feature Elimination (RFE) with ridge regression and random forests, Variance Threshold, Sequential Forward Selection (SFS) with ridge regression and KNN, random forests feature importance, Lasso regularization, and Pointwise Mutual Information (PMI) with ridge regression and random forests. The application of these techniques is detailed in Figure 14. Performance metrics, including accuracy, F1 score, and Matthews correlation coefficient (MCC), for these models, are presented in bar graphs in Figure 13 for baseline models and in Figure 14 for gradient boosting models. Overall, the prediction of learning loss for Reading exhibits weaker performance than for math. Although most models perform similarly across both subjects, with the exception of KNN, gradient boosting for math and ridge regression for Reading demonstrate the highest average accuracy, F1 score, and MCC.

TABLE VII: Best Performance of the ten machine learning models that are trained for (a) Math and (b) Reading for *DistrictA* dataset. CatBoost is the overall winner

Model	Best Set	Feature Selection	Acc [0,1]	F1 [0,1]	MCC [-1,+1]
LR Ridge	45	FI RF	0.639	0.622	0.368
SVM	45	SFS LR	0.628	0.584	0.343
KNN	55	LR Lasso	0.618	0.591	0.318
Random Forests	45	SFS LR	0.639	0.582	0.363
Gradient Boost	36	RFE RF	0.644	0.622	0.375
CatBoost	36	RFE RF	**0.675**	**0.645**	**0.434**
HistGB	45	SFS KNN	0.634	0.609	0.35
LightGBM	70	PMI RF	0.644	0.601	0.372
XGBoost	21	VR	0.66	0.616	0.405

(a) Math

Model	Best Set	Feature Selection	Acc [0,1]	F1 [0,1]	MCC [-1,+1]
LR Ridge	45	SFS LR	0.607	0.522	0.303
SVM	45	SFS KNN	0.586	**0.553**	0.274
KNN	45	SFS KNN	0.571	0.536	0.232
Random Forests	45	SFS LR	0.592	0.513	0.26
Gradient Boost	45	SFS LR	0.56	0.542	0.231
CatBoost	82	PMI - Ridge	**0.623**	0.548	**0.338**
HistGB	45	SFS LR	0.576	0.495	0.219
LightGBM	90	All	0.602	0.516	0.288
XGBoost	90	All	0.613	0.535	0.312

(b) Reading

Four advanced gradient boost models, XGBoost, LightGBM, CatBoost, and HistGradientBoosting, train the same sets of features for comparison purposes. To improve the gradient boosting models, we can penalize and regularize the algorithm by hyperparameter tuning so that we aim to increase accuracy and avoid overfitting. These hyperparameters are searched with a 5-fold cross-validation RandomizedSearch with the number of iterations that is 20% of parameter distributions of each model. For example, XGBoost is supposed to explore 100 distributions of the parameters; the number of iterations for RandomizedSearch is 20 times.

To begin with, constraining tree structures reduces the growth of complex and more extended trees by optimizing parameters such as the number of trees, the depth of trees, and the number of leaves per tree. In addition, setting a smaller learning rate, usually less than 0.5, allows weighting trees to slow the learning by a small amount at each iteration to reduce errors. Furthermore, setting the optimal L1 and L2 regularization terms penalizing the sum of the leave weights improves the models by simplifying the complexity and size of the model [18]. The gradient boosting algorithms also show higher prediction power for math than reading and indicate no significant model exceeding other models, including the best state-of-the-art models, in terms of performance.

The various dimensions of the selected features were experimented with to examine the effects of dimensionality reduction methods and the best set of the features by predicting learning loss with the machine learning models introduced in Section III-B. Then, our initial data set was also experimented with gradient boosting models in terms of missing values and their imputation.

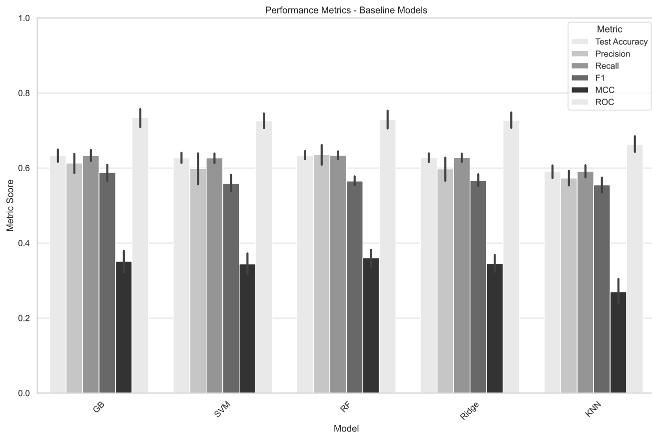

Fig. 13: Five state-of-the-art machine learning models fitted to 10 feature sets for predicting learning loss. With the train-test split, GridSearch, and 10-fold cross-validation.

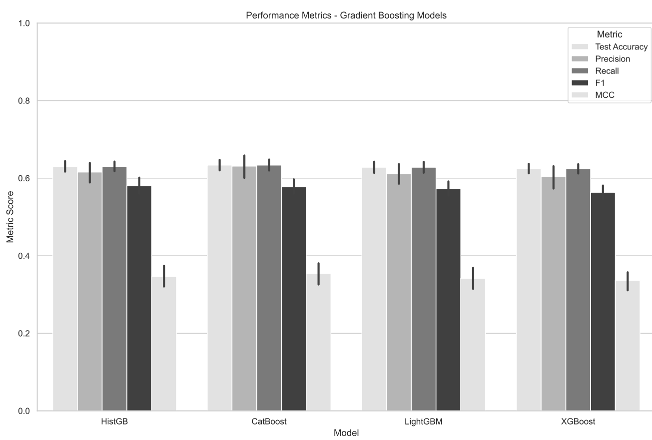

Fig. 14: Four advanced gradient boost models fitted to 10 feature sets for predicting learning loss. With the train-test split, GridSearch, and 10-fold cross-validation.

For the ten models, the best set of features for each model is described in Table VII (a) for math and (b) for Reading; both subjects suggest CatBoost as the most robust models: 36 features selected by RFE with random forests with precision (68%), F1 (65%) and MCC (43%) for math and 82 features selected by PMI with ridge regression with precision (62%), F1 (55%) and MCC (34%) for Reading.

Overall, the gradient boosting algorithms CatBoost and XGBoost are the best choices of all the machine learning models we have experimented with to predict learning loss for both subjects. Although these models performed better in predicting failure in math rather than reading, in general, the performance gap between the four gradient boosting models and the five state-of-the-art models, except KNN, is negligible, as their difference in accuracy is around 3%. Furthermore, no clear indication of the best dimensionality reduction technique that performs across all models emerged.

TODO: For most figures: Increase the text size, right??? – Mirna

C. Best Features vs. Raw Data for Gradient Boosting Models

All four gradient boosting models built – XGBoost, LightGBM, CatBoost, and HistGrandientBoosing – are aware of the sparsity of data, such as missing values, by finding optimal tree split. Recall that the initial data set, also known as Raw data, containing 506 features (505 numerical and one categorical) for 1,165 school districts, includes 416 details with missing values as small as 1% and as large as 88% of each point. In this experiment, we executed the pipeline of building the advanced gradient boosting models for raw data. We conducted a comparison with models trained on data processing using various feature engineering techniques to assess their predictive power concerning learning loss. The classification tasks for math and Reading were completed. All features with missing values except for eight details are subject-specific, e.g., the number of grade 3 students tested in math. After dropping the subject-specific math features for Reading and vice versa, 302 was the dimension of characteristics for this experiment for each subject. 212 of 302 details contain missing values. We have three data sets for comparison: (1) the best sets of features in Table VII from the performance results of the four gradient boosting models in Figure VII, (2) raw data without imputation for missing values, and (3) raw data impute missing values with mean values. Our data has only one categorical attribute, including no missing values, so the imputation method is limited to average. Regarding the performance of Best Features vs. Raw data, all models improved with Raw data throughout all performance metrics, especially MCC, for both subjects; HistGradientBoost increased MCC the most by 47% following LightGBM (43%), CatBoost (25%) and XGBoost (24%) for math, and the improved MCC for Reading is even higher with 124% for HistGradientBoost and 45%, 43%, and 41% for LightGBM, CatBoost, and XGBoost, respectively. For a closer look, we also observed that the Raw data set without imputation performed slightly better compared to the Raw data set with imputation for all models and subjects; MCC for math rose the most, over 6%, in CatBoost and HistGradientBoost; on the contrary, XGBoost showed the most significant growth for MCC in reading with 10%.

VI. Conclusion and Future Work

The intentional data science pipeline employed in this study automatically unearthed crucial features using publicly available data, leveraging ten distinct feature selection methods to model the impact of COVID-19 on learning loss. However, despite the reduction in data dimensionality facilitated by these methods, they exhibited limited influence on prediction accuracy. Surprisingly, the performance of the ten machine learning models trained on the feature-selected sets showcased negligible improvements. Notably, gradient-boosting algorithms, such as XGBoost and CatBoost, emerged as consistently superior in both projects. These models demonstrated remarkable efficacy in managing missing values, a prevalent issue as more than two-thirds of the features in the learning loss datasets contained missing values. Our reproducible experiments and datasets are accessible at [42], offering valuable tools for policymakers.

Upon deeper analysis of the most influential features, it became evident that shifts in significance were predominantly observed at the individual feature level rather than through changes in resilience factor importance from 2022 to 2023. Notably, across this period, the mode of instruction and prior score stood out as the most significant resilience factors in the realm of education. This finding underscores the centrality of these factors in understanding and addressing learning loss dynamics. Policymakers can leverage our predictive models and analytical insights to strategically allocate resources and interventions within the public school system, targeting schools, students, and educators to mitigate and counteract the effects of learning loss.

Acknowledgments This research was partly supported by the Community Health and Economic Resilience Research at Translational Health Research Center at Texas State University.

REFERENCES

[1] K. Choate, D. Goldhaber, and R. Theobald, "The effects of covid-19 on teacher preparation," *Phi Delta Kappan*, vol. 102, no. 7, pp. 52–57, 2021.

[2] G. Zamarro, A. Camp, D. Fuchsman, and J. B. McGee, "Understanding how covid-19 has changed teachers' chances of remaining in the classroom," *Sinquefield Center for Applied Economic Research Working Paper*, vol. 22, no. 01, 2022.

[3] OECD, *Education at a Glance 2021*. https://doi.org/10.1787/b35a14e5-en: Organisation for Economic Co-operation and Development, 2021.

[4] P. Engzell, A. Frey, and M. D. Verhagen, "Learning loss due to school closures during the covid-19 pandemic," *Proceedings of the National Academy of Sciences*, vol. 118, no. 17, 2021.

[5] J. E. Maldonado and K. De Witte, "The effect of school closures on standardised student test outcomes," *British Educational Research Journal*, vol. 48, no. 1, pp. 49–94, 2022.

[6] C. J. Courtemanche, A. H. Le, A. Yelowitz, and R. Zimmer, "School reopenings, mobility, and covid-19 spread: Evidence from texas," tech. rep., National Bureau of Economic Research, 2021.

[7] C. Halloran, R. Jack, J. C. Okun, and E. Oster, "Pandemic schooling mode and student test scores: Evidence from us states," tech. rep., National Bureau of Economic Research, 2021.

[8] D. Betebenner, A. Van Iwaarden, A. Cooperman, M. Boyer, and N. Dadey, "Assessing the academic impact of covid-19 in summer 2021," 2021.

[9] N. C. D. of Public Instruction, 2022.

[10] T. E. A. (TEA), "Impacts of covid-19 and accountability updates for 2022 and beyond," 2022.

[11] S. Rebai, F. B. Yahia, and H. Essid, "A graphically based machine learning approach to predict secondary schools performance in tunisia," *Socio-Economic Planning Sciences*, vol. 70, p. 100724, 2020.

[12] T. E. Agency, "Impacts of covid-19 and accountability updates for 2022 and beyond." https://tea.texas.gov/sites/default/files/2021-tac-accountability-presentation-final.pdf, 2021.

[13] T. Cardona, E. A. Cudney, R. Hoerl, and J. Snyder, "Data mining and machine learning retention models in higher education," *Journal of College Student Retention: Research, Theory & Practice*, vol. 25, no. 1, p. 1521025120964920, 2020.

[14] A. R. Rao, Y. Desai, and K. Mishra, "Data science education through education data: an end-to-end perspective," in *2019 IEEE Integrated STEM Education Conference (ISEC)*, (U.S.), pp. 300–307, IEEE, 2019.

[15] Y. Baashar, G. Alkawsi, N. Ali, H. Alhussian, and H. Bahbouh, "Predicting student's performance using machine learning methods: A systematic literature review," in *2021 International Conference on Computer & Information Sciences (ICCOINS)*, (U.S.), pp. 357–362, IEEE, 2021.

[16] Z. Wang, "When large-scale assessments meet data science: The big-fish-little-pond effect in fourth-and eighth-grade mathematics across nations," *Frontiers in Psychology*, vol. 11, p. 579545, 2020.

[17] K. Yan, "Student performance prediction using xgboost method from a macro perspective," in *2021 2nd International Conference on Computing and Data Science (CDS)*, (U.S.), pp. 453–459, IEEE, 2021.

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), p. 785–794, Association for Computing Machinery, 2016.

[19] G. Ke, Q. Meng, T. Finely, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30 (NIP 2017)*, (https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/), pp. 1–9, Advances in Neural Information Processing Systems 30 (NIP 2017), December 2017.

[20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[21] S. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 6679–6687, May 2021.

[22] A. Abutbul, G. Elidan, L. Katzir, and R. El-Yaniv, "Dnf-net: A neural architecture for tabular data," *CoRR*, vol. abs/2006.06465, 2020.

[23] S. Popov, S. Morozov, and A. Babenko, "Neural oblivious decision ensembles for deep learning on tabular data," *CoRR*, vol. abs/1909.06312, pp. 1–12, 2019.

[24] G. Ke, J. Zhang, Z. Xu, J. Bian, and T.-Y. Liu, "TabNN: A universal neural network solution for tabular data," 2019.

[25] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[26] M. Joseph, "Pytorch tabular: A framework for deep learning with tabular data," 2021.

[27] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," 2021.

[28] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, (-), pp. 18932–18943, Curran Associates, Inc., 2021.

[29] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," 2022.

[30] G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance," *Statistics and Computing*, vol. 31, pp. 1–16, 2021.

[31] S. Abe, "Modified backward feature selection by cross validation.," in *ESANN*, (U.S.), pp. 163–168, Springer, 2005.

[32] B. Ghojogh, M. Samad, S. Mashhadi, T. Kapoor, W. Ali, F. Karray, and M. Crowley, "Feature selection and feature extraction in pattern analysis: A literature review," *arXiv:1905.02845v1*, vol. 1, pp. 1–14, 05 2019.

[33] N. C. for Education Statistics (NCES), "Common core of data (ccd)." https://nces.ed.gov/ccd/elsi/tableGenerator.aspx, 2022.

[34] T. E. A. (TEA), "State of texas assessments of academic readiness (staar) for 2018-2019 and 2020-2021." https://tea.texas.gov/student-assessment/testing/staar/staar-aggregate-data, 2022.

[35] T. D. of State Health Services (DSHS), "Texas public schools covid-19 data." https://dshs.texas.gov/coronavirus/schools/texas-education-agency/, 2022.

[36] USAFacts, "Texas coronavirus cases and deaths." https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/texas, 2022.

[37] U. B. of Labor Statistics (BLS), "Local area unemployment statistics (laus)." https://www.bls.gov/lau, 2022.

[38] C. Bureau, "Census block group 2010." https://schoolsdata2-93b5c-tea-texas.opendata.arcgis.com/datasets/census-block-group-2010-tx/, 2010.

[39] T. E. A. (TEA), "Average daily attendance (ada)." https://tea.texas.gov/finance-and-grants/state-funding/state-funding-reports-and-data/average-daily-attendance-and-wealth-per-average-daily-attendance, 2022.

[40] T. E. A. (TEA), "Elementary and secondary school emergency relief (esser) grant programs." https://tea.texas.gov/finance-and-grants/grants/elementary-and-secondary-school-emergency-relief-esser-grant-programs, 2021.

[41] (ESE), "Elementary and secondary school emergency relief fund." https://oese.ed.gov/offices/education-stabilization-fund/elementary-secondary-school-emergency-relief-fund/, 2022.

[42] J. Yu and J. Tešić, "Tabular data in the wild: Gradient boosting modeling improvement." https://github.com/DataLab12/educationDataScience, 2022.