IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Novel Considerations in the ML/AI Modeling of Large-Scale Learning Loss

MIRNA ELIZONDO[1],(Student, IEEE), JUNE YU[1], DANIEL PAYAN[1], LI FENG[2], and JELENA TEŠIĆ[1],(Senior Member, IEEE)

[1]Department of Computer Science, Texas State University, San Marcos, TX 78666, USA
[2]Department of Finance and Economics, Texas State University, San Marcos, TX 78666, USA

Corresponding author: Mirna Elizondo (e-mail: m_e172@txstate.edu).

This work was supported by the NSF #1950292 and NSF #1928696 grants and the Center of Excellence for Community Health and Economic Resilience Research and Center for Analytics and Data Science at Texas State University.

ABSTRACT This study is a path forward for the large-scale, data-driven quantitative analysis of noisy open-source data resources. The goal is to support qualitative findings of smaller studies with extensive open-source data-driven analytics in a new way. The study presented in this research focuses on learning interventions. It uses nine publicly accessible datasets to understand and mitigate factors contributing to learning loss and the practical learning recovery measures in Texas public school districts after the recent school closures. The data came from the Census Bureau 2010, USAFACTS, Texas Department of State Health Services (DSHS), the National Center for Education Statistics (CCD), the US Bureau of Labor Statistics (LAUS), and three sources from the Texas Education Agency (STAAR, TEA, ADA, ESSER). We demonstrate a novel data-driven approach to discover insights from an extensive collection of heterogeneous public data sources. For the pandemic school closure period, the mode of instruction and prior score emerged as the primary resilience factors in the learning recovery intervention method. Grade level and census community income level are the most influential factors in predicting learning loss for both Math and Reading. We demonstrate that data-driven unbiased data analysis at a larger scale can offer policymakers an actionable understanding of how to identify learning-loss tendencies and prevent them in public schools.

INDEX TERMS noisy tabular data, data in the wild, gradient boosting, feature selection, dimensionality reduction

I. INTRODUCTION

Learning loss, within the context of education, can be defined as the depletion or regression of previously attained or expected knowledge and competencies. COVID-19 also had an impact on teacher preparation [1]. As an example, the recent COVID-19 pandemic forced many schools to close, and the global consequences of a five-month closure of schools were equated to less than $10 trillion monetary loss as 43 million students were affected by the school closures [2]. The school closures have led to learning loss among students [3]. The learning loss percentage in some countries was estimated from 0.08 to 0.29 based on the public data [4]. The school closure and subsequent reopening in the United States were uneven as there was no consensus. Thus, the learning loss was not uniform across the states, as documented for Virginia, Maryland, Ohio, and Connecticut in [5]. Individual studies of Rhode Island and North Carolina education data provided estimations of the learning losses and recovery [6]. The data-driven factors contributing to the learning recovery in the abovementioned studies remain elusive and diverse [7]. This complexity posed challenges for policymakers in the other states in determining the learning intervention measures based on their data. 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 [8], [9]. If the apparent factors (census data), location (urban vs. rural), and standardized exam data are good predictors of learning loss. Next, we pinpoint the resilient factors contributing to learning recovery within Texas schools. Our approach is novel in that it integrates data science methodologies with educational policy analysis, offering a data-driven perspective to inform decision-making processes. We identify the factors most important for the schools to experience significant learning loss and recovery using nine open data sources. Next, we introduce the improved automated attribute importance analysis to understand various parameters, including consensus information, demographics of public school districts, instructional modalities, socioeconomic indicators such as income levels, urban or rural settings, student attendance rates, county infection rates, and unemployment statistics, among numerous other factors from 2019 to 2022.

This research uncovers fascinating data-driven indicators: the most resilient factor influencing learning loss in the district is how early or late the students return 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. The remainder of this paper is structured as follows: Section II reviews pertinent literature, Section III-B outlines the research design, Section IV describes data gathering and preparation; Results presents our findings; and Section VI discusses the implications and suggests directions for future research.

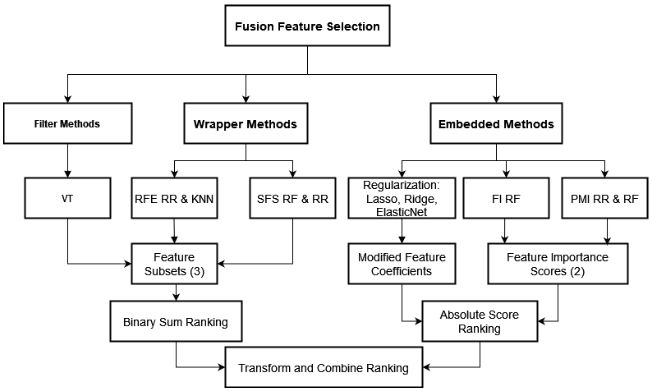


FIGURE 1. Fusion Process Of Aggregating Outcomes From Various Feature Selection Methods For Reading And Math For The Academic Years 2018-2019, 2020-2021, 2021-2022. Refer to Table 1.

II. RELATED WORK

In this section, we focus on (1) quantitative research and machine learning tools to gain insights from data on the relationship with outcomes without overfitting features to the data and (2) 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 networks, decision trees, and neural networks—generally offer excellent classification accuracy above 70% for simple classification tasks [10]. However, excessive reliance on feature engineering may result in less-than-optimal outcomes when translating domain-specific data [11]. ML methods such as deep neural networks (DNN), decision trees, support vector machines (SVM), and K-nearest neighbors (KNN) are widely used for predicting student academic performance [12]. Demographic, educational, familial/personal, and internal assessment factors are standard resources for data-driven evaluations of student performance across classroom metrics, grade levels, and standardized tests [11].

The COVID-19 pandemic significantly disrupted educational systems worldwide, exacerbating inequalities in learning outcomes, particularly for underserved communities. Recent research highlights the unequal effects of these disruptions and the importance of policy interventions in addressing them [13]. The study synthesizes evidence on the achievement gap and emphasizes technology and policy innovations to support equitable recovery and long-term resilience.

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 DNNs [14]. Gradient-boosted decision trees (GBDT) models such as XGBoost [15], LightGBM [16], and CatBoost [17] are preferred for tabular data due to their superior performance and ability to handle complex feature interactions effectively. Although deep learning models such as TabNet [18], NODE [19], and TabNN [20] show promise, recent benchmarks confirm that GBDT models still outperform deep learning for tabular data in most scenarios [21].

III. METHODOLOGY

The methodology employed in this study aims to systematically uncover factors contributing to both learning loss and learning gain among students. The study utilizes comprehensive datasets from Texas public schools to analyze trends and patterns in Math and Reading scores by leveraging advanced educational data science techniques. This methodology integrates statistical modeling, machine learning algorithms, and data visualization tools to identify critical variables influencing student academic performance over time. By examining multiple academic years, including the 2021-2022 data, the study ensures robustness and reliability in its findings, offering valuable insights into educational outcomes and informing targeted interventions.

A. ATTRIBUTE IMPORTANCE SCORING

First, we introduce an innovative approach to identifying critical features from various potential factors. Heterogeneous data tends to have overlapping information mixed with numerical and categorical data. Logistic Regression coefficients for the actual data often randomly select one out of multiple correlated columns and are not robust enough for the noisy multi-source data analysis [22]. We propose to contrast filter, embedded, and wrapper methods for feature importance and propose a novel aggregation technique for robustness.

Several distinct (ten with variations) algorithms for automated feature selection and interpretative methods for analyzing feature importance are evaluated to assess their effectiveness. These measures aim to mitigate issues associated with "Garbage In, Garbage Out" (GIGO) and trivial modeling. We construct a quasi-orthonormal attribute space by distilling and aggregating highly correlated features.

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:

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

The $\mathrm{cov}(X, Y)$ 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. Here, we modify ten distinct approaches from filter, embedded, and wrapper methods sets to identify and assess the features influencing our prediction models. Each technique aims to select feature sets with minimal redundancy and maximal relevance, resulting in either a chosen set of features or a score indicating feature importance.

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 is biased by the correlation between features [23]. Our ultimate feature set comprises features exhibiting positive mean importance as determined by the PFI, identifying crucial features. We utilize Random Forests *PFI RF* and Logistic Regression with Ridge Regularization *PFI RR*, both of which assign non-zero scores to all features.

Recursive Feature Elimination (RFE) is a method of training a model on a complete set of features in the data set, eliminating the features with the smallest coefficients. This process iterates until the 10-fold cross-validation score of the models with Random Forest *RFE RF* and Logistic Regression with Ridge Regularization *RFE RR* on the training data shows a decrease. The final scores are attribute rankings where 1 indicates the most relevant features [24].

TABLE 1. Resilience Factors and Methods with Abbreviations

Abbreviation	Resilience Factor
LI	Low Income
ATT	Attendance
DEM	Demographics
R/E	Race/Ethnicity
CC	County COVID
DM	District Makeup
MOI	Mode of Instruction
TST	Testing
PS	Prior Score
LOC	Locale
Abbreviation	**Feature Method**
RFE RF	Recursive Feature Elimination - Random Forest
RFE RR	Recursive Feature Elimination - Ridge Regression
VT	Variance Threshold
SFS RR	Sequential Feature Selection - Ridge Regression
SFS KNN	Sequential Feature Selection - K-Nearest Neighbors
FI RF	Feature Importance - Random Forest
PFI RR	Permutation Feature Importance - Ridge
PFI RF	Permutation Feature Importance - Random Forest
Elastic Loss	ElasticNet Logistic Regression Loss
Elastic Expected	ElasticNet Logistic Regression Expected
Elastic Gain	ElasticNet Logistic Regression Gain
Lasso Loss	Logistic Regression L1 (Lasso) Loss
Lasso Expected	Logistic Regression L1 (Lasso) Expected
Lasso Gain	Logistic Regression L1 (Lasso) Gain
Ridge Loss	Logistic Regression L2 (Ridge) Loss
Ridge Expected	Logistic Regression L2 (Ridge) Expected
Ridge Gain	Logistic Regression L2 (Ridge) Gain

Logistic Regression with Filtering and Regularization is a technique that uses L1 *Lasso* or L1 and L2 *Elastic* penalty terms to shrink the coefficients during training. L1 regularization reduces the coefficients of some features to zero for both, and the remaining non-zero coefficients are considered useful information for prediction. On the other hand, L2 regularization, commonly known as *Ridge*, penalizes the square of coefficients, effectively reducing their magnitude without necessarily setting them to zero. This method helps handle multicollinearity and stabilize the model by smoothing out fluctuations in the data, thereby improving generalization performance.

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. The selected set contains features with the top 50% scores.

Variance Threshold (VT) is a straightforward method to eliminate features by removing features with low variance in the training data set [25]. 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 not selected. The final set of features consists of the k features with the highest variance. VT, SFS RR, 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.

The labels are also used in Figures 2 to 4 to illustrate the Section V and comparisons clearly. Figure 1 illustrates the aggregated scoring mechanism detailed in Algorithm 1. This figure emphasizes the innovative approach to combining filter, embedded, and wrapper methods for feature selection, ultimately producing a robust feature importance ranking.

We obtained ten diverse results comprising binary, numerical, and rank scores. We suggest multiple fusion scoring mechanisms for end-users, as detailed in Algorithm 1. First, we look into five approaches that filter out features and rank the features by the binary sum outputs. Next, we take the 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, importance ranking. Figure 1 illustrates the fusion process described above.

B. PREDICTION MODELING

In this study, we address whether the public data collected from web sources is sufficient to predict school district learning performance during the COVID-19 years reliably. Thus, we create five basic baseline models: Logistic Regression with Ridge Regularization, Support Vector Machines (SVM), K-Nearest Neighbor (KNN) suitable for nonlinear and non-separable data, Random Forests, and GBDT. Additionally, we explore four advanced GBDT algorithms: XGBoost, LightGBM, CatBoost, and HistGradientBoosting. Since the data aligns with the features of tabular data, we opt for GBDT methodologies due to their demonstrated robustness in handling diverse tabular datasets [26]. The gradient-boosted decision tree (GBDT) assembles many weak decision trees and grows them sequentially and iteratively based on the residual modeling from the previous trees.

The GBDT methods handle tricky observations well and are optimized for faster and more efficient fitting using a data sparsity-aware histogram-based algorithm. In contrast to the pointwise split of the traditional GBDT, 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 [15]. The 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 [16]. Our ultimate goal is to assess the predictive power of these nine machine-learning models in this real example.

Algorithm 1 Fusion Scoring Algorithm

```
Input: Feature Selection Importance Scores (binary, numerical)
Output: Final Fusion Importance Ranking
Initialize BinarySumRankings and AbsoluteScoreRankings
for each result in Results do
    if result is binary then
        Filter and sum relevant binary features
        Rank features by binary sums
        Append to BinarySumRankings
    else
        Calculate and rank features by absolute scores
        Append to AbsoluteScoreRankings
    end if
end for
Transform BinarySumRankings
Transform AbsoluteScoreRankings
Combine and merge both rankings for the FinalFeatureImportanceRanking
Return: FinalRanking
```

IV. OPEN SOURCE DATA ACQUISITION AND PROCESSING

A. OPEN DATA SOURCES

The dataset utilized for this analysis integrates information from nine distinct sources, employing both School District I.D. and County FIPS Code to cover a comprehensive range of 1,165 school districts across 253 counties in Texas. The data frames and their respective sources include CCD from the National Center for Education Statistics, providing a matrix of 1189 rows by 66 columns; STAAR and TEA from the Texas Education Agency, with dimensions of 1184x217 and 1182x217 respectively; ADA from the Texas Education Agency, comprising 1226 rows by three columns; ESSER from the Texas Education Agency, with 1208 rows by six columns; Census data from the Census Bureau (2010), with 254 rows by 37 columns; Covid data from USAFacts, featuring 254 rows by eight columns; LAUS from the US Bureau of Labor Statistics, spanning 254 rows by 13 columns; and additional Covid data from DSHS, providing a matrix of 1216 rows by seven columns.

State of Texas Assessments of Academic Readiness (STAAR) data was obtained from the Texas Education Agency (TEA) for the fiscal years 2020, 2021, and 2022 for each school district [27]. 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.

Common Core of Data (CCD) [28] 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, 2020, 2021, and 2022. For example, according to the data acquired, 62.5% of the students attend school in rural areas, 19.8% in town areas, 10.6% in suburban areas, and only 7.1% in the city area.

Texas School COVID-19 campus data comes from the Texas Department of State Health Services (DSHS) [29], including the self-reported student enrollment and on-campus enrollment numbers of the dates 28 September 2020, 30 October 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 [30]. **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 2019, 2020, 2021, and 2022. The **Local Area Unemployment Statistics (LAUS)** data [31] was parsed from the US 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 [32] captures the county's general population characteristics. Upon completing the initial data integration process, merging data from nine sources by matching school district I.D. and county FIPS code, the dataset encompasses 1,165 school districts in Texas, spanning 253 counties with 506 features, one definite, and 505 numerical variables.

For the academic years 2018-2019, 2020-2021, and 2021-2022, the percentage distribution of students by race varied slightly over time. In 2018-2019, the breakdown was as follows: 1.07% Asian or Asian/Pacific Islander, 6.50% Black or African American, 40.41% Hispanic, and 49.09% White. The following year, 2020-2021, saw minor shifts with 1.10% Asian or Asian/Pacific Islander, 6.35% Black or African American, 41.32% Hispanic, and 48.12% White. For the academic year 2021-2022, the proportions were 1.12% Asian or Asian/Pacific Islander, 6.29% Black or African American, 41.67% Hispanic, and 47.70% White.

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 [33], [34]. **CARES ESSER I:** Authorized on 27 March 2020 the Coronavirus Aid Relief and Economic Security (CARES) Act with $13.2 billion for the fiscal year 2020. **CRRSA ESSER II:** Authorized on 27 December 2020, as the Coronavirus Response and Relief Supplemental Appropriations (CRRSA) Act with $54.3 billion for the fiscal year 2021. **ARP ESSER III:** Authorized on 11 March 2021, as the American Rescue Plan (ARP) Act with $122 billion 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 from M13 March 2020 to 30 September 2022. To help policymakers make more informative decisions on learning recovery with localized efforts in each school district, we collected data from nine different sources to determine the qualitative conclusions from small sample datasets. Qualitative findings on the educational impacts of COVID-19 highlight significant disruptions in learning environments and the varied responses of educational systems worldwide. It emphasizes qualitative insights into the socio-emotional and pedagogical challenges students and educators face during the pandemic [35], matching the data-driven findings from significant, heterogeneous, noisy public data sources.

B. DATA CLEANING, AGGREGATION AND FILTERING

Common Core of Data (CCD) [28] 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 [27] from the Texas Education Agency (TEA) for the fiscal years 2019, 2020, 2021, and 2022 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) [29], including the self-reported student enrollment and on-campus enrollment numbers of the dates 28 September 2020, 30 October 2020, and January 29, 2021, at each school district in Texas. Additional COVID-19 data involved confirmed infection and death cases [30] due to Coronavirus at each county from USAFacts. Also, the average daily attendance (ADA) [36], 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 [33] were collected from TEA for school district level. The ADA data for fiscal years 2019 and 2021 capture the impact of district attendance, and the ESSER data reflect the localized efforts of TEA allocating the grant amount at each school district for the fiscal years 2019, 2020, 2021, and 2022. Also, we combined the Local Area Unemployment Statistics (LAUS) data [31] from the US 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 [32] capture the demographic features of a county for the general population. We merged the data from nine sources by matching the school district I.D. and county FIPS code and then integrated based on the district I.D. and county FIPS code.

The analysis aggregates the values from various demographic and educational categories into one consolidated group by calculating the percentage difference between the corresponding pairs of values from 2022-2021, 2020-2021, and 2018-2019. For instance, the percentage difference between the total count of White Students in 2020-2021 and 2018-2019, denoted as '% White Students Diff,' reflects the change in the demographic composition over the two years. Similarly, the percentage difference in total race/ethnicity counts and school-wide Title I program participation provides insights into broader demographic shifts and changes in program enrollment. Moreover, examining the percentage difference in enrollment counts across different grade levels, such as Prek and Kinder, between the two years offers valuable information regarding enrollment trends within specific age groups [37].

Among the 506 features analyzed, 416 display missing values across three data sources, varying from one to 88% within our dataset. Notably, 408 features originate from STAAR and TEA data, six from CCD and NCES, and two from COVID and DSHS data. Within these 416 features, 332 have less than 20% missing values, while 24 exhibits more than 80% missing values. Features exhibiting over 20% missing values primarily originate from the STAAR data, specifically concerning average scores and participation in the STAAR tests. Consequently, we eliminated these features from the STAAR dataset. Additionally, we excluded school districts lacking CCD and COVID data, resulting in 955 public school districts in Texas available for analysis, featuring 119 features devoid of missing values.

Out of 119 features, we aggregate the 58 features that duplicate the 2019, 2021, and 2022 data into 29 differential features. For example, Total Schools 2020-2021 and Total Schools 2018-2019 features are aggregated into Total Schools Diff 2021, reducing the total number of features to 90. For Total Schools 2022-2021 and Total Schools 2020-2021, features are aggregated into Total Schools Diff 2022, reducing the total number of features to 90. The dataset's missing value distribution across columns reveals varying frequencies within distinct count ranges.

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: $NormalizedScore = \frac{\text{grade score}}{\max(\text{grade score})}$. Next, the district average is calculated by summing up the scores of grades G3 to G8 and dividing by the total number of grades. The normalized score provides an overarching view of the academic performance within the district, represented by the equation: $DistrictAverage = \frac{(G3+G4+G5+G6+G7+G8)}{TotalNumberOfGrades}$. The percentage loss in performance is computed over time intervals, reflecting changes in educational outcomes in Eq 4:

$$\text{Score} = \frac{\text{Avg 2021} - \text{Avg 2019}}{\text{Avg 2019}} \tag{3}$$

$$\text{Score} = \frac{\text{Avg 2022} - \text{Avg 2021}}{\text{Avg 2021}} \tag{4}$$

We label the scores as follows: *Gain* if the overall score is more significant than zero, *Expected* if the overall score equals zero, and *Loss* if the overall score is less than zero. This comprehensive process enables the assessment of educational trends, facilitating informed decisions and interventions to enhance learning outcomes.

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. The data provide insights into the educational workforce, encompassing Full-Time Equivalent (FTE) teachers, overall staff counts, and 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 regarding 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 ratioed 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 2010. Lastly, when assessed by date, the normalization process considers the Percentage of students on campus on 28 September 2020, 30 October 2020, and January 29, 2021. The Census block grouped by county [32] 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 time frames. Table 2 illustrates the race, grade, and age groups across counties, highlighting diversity and composition, which are crucial for understanding educational demographics and planning educational resources.

For Math, the average learning gain from 2021 to 2022 was 2.80%, contrasting with the previous average loss of -2.75%. This shift indicates an overall improvement in Math proficiency. The standard deviation increased to 11.93%, suggesting more significant variability in student outcomes. The minimum loss observed was -50.10%, while the maximum gain was 210.09%. In Reading, the average learning gain from 2021 to 2022 was 0.58%, slightly higher than the previous period's gain of 0.32%. The standard deviation remained similar at 9.08%, showing consistent variability. The minimum observed loss in Reading was -50.48%, and the maximum gain was 191.43%. Considering all subjects

TABLE 2. Demographic Proportions of Race/Ethnicity, Gender, and Age Groups Across Counties [32].

Race		Gender		Age (0-24)		Age (25+)	
Category	**Total**	**Category**	**Total**	**Category**	**Total**	**Category**	**Total**
White	17,701,487	Male	12,472,234	0-4	1,928,470	35-44	3,458,373
Black	2,979,598	Female	12,673,245	5-9	1,928,232	45-54	3,435,322
Asian	964,596			10-14	1,881,881	55-64	2,597,668
Hispanic	9,460,903			15-19	1,883,121	65-74	1,472,248
				20-24	1,817,069	75+	1,129,626
				25-34	3,613,469		

combined, the average learning gain from 2021 to 2022 was 1.69%, a significant improvement compared to the previous average loss of -1.22%. The standard deviation increased to 10.22%, indicating more diverse student outcomes. The minimum observed loss across all subjects was -36.35%, while the maximum gain was 193.44%. Figure 5 shows that student learning outcomes improved from 2021 to 2022, with average gains recorded across all subjects. Math showed the most substantial recovery, transitioning from an average loss to a notable gain, while Reading maintained a modest improvement. The increased standard deviations indicate more varied student experiences during this period. Next, we set the threshold to categorize districts based on the STAAR scores into three categories: *Loss*, *Expected*, and *Gain*. The data revealed that more districts experienced a loss in Math, with a median loss value of -0.03, compared to a median of 0 for Reading. We analyzed and predicted outcomes for Math and Reading separately. School districts in the middle 50% of loss values were labeled as *Expected*, those 25% as *Loss*, and those 75% as *Gain*.

This categorization enabled us to explore the correlation between various features and these labels. Our findings indicated that the proportion of White students was higher in districts labeled as *Gain* and decreased in those labeled as *Loss*. Conversely, Hispanic students constituted about two-thirds of the *Loss* category, and their proportion decreased in the *Expected* and *Gain* categories for both Math and Reading. The locale of school districts showed a correlation with learning loss labels as over half of the schools were located in rural areas, with rural locales positively correlated with the *Gain* label. However, an increasing number of losses occurred in schools located in city and suburban areas.

D. DATA PRE-PROCESSING

In **LossA**, we propose a dimensionality reduction dataset to enhance interpretability and pinpoint resilience factors associated with Learning loss. 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. These features are treated independently in the dataset **LossB**. This approach employs the raw integrated data for the GBDT experiment without normalization while considering missing values. *LossB* treats each feature individually, whereas *LossA* utilizes normalized and aggregated features to reduce dimensionality and enhance interpretability. While this approach *LossA* may result in a more prominent feature space and potentially increase computational complexity, it allows for a more detailed analysis of 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. *LossB* provides a complementary perspective to *LossA* and allows for a comprehensive exploration of the dataset, encompassing 506 features across 1,165 school districts.

V. RESULTS

In this section, we analyze the results and proposed approaches. For simplicity, Table 1 describes the ten abbreviation labels used for the feature importance scoring. RFE (RF and RR) provide attribute ranking, and SFS (KNN and RR) provide a binary selection of features. RF FI and PMI (RF and RR) offer non-zero scores to all 90 features. Lasso, Ridge, and Elastic fit for the *Gain*, *Expected*, and *Loss* provides scores for a subset of coefficients selected.

We consider Math Learning Loss and Reading Learning Loss separate tasks with separate attribute selections from the same dataset. Table 1 expands on the following abbreviated feature selection methods that separately detect the resilience factors; abbreviations can be found in Table 1, for Learning *Loss* due to COVID-19 using the data set with 90 features and 955 school districts in Texas as a baseline.

A. ATTRIBUTE IMPORTANCE ANALYSIS

Following Algorithm 1 and the process illustrated in Figure 1, we aggregate five filtering method outcomes for Reading and Math: VT, SFS KNN, SFS Ridge, and Elastic Gain and Elastic Loss binarized coefficients. The Initial Importance Values represent raw scores from machine learning methods, which are challenging to compare due to their non-uniformity. The Binary Selection Values are the first output transformation, where VT, SFS KNN, and SFS Ridge outcomes are already binary. RFE methods retain the rank of one feature, assigning it a value of 1, and logistic regression assigns +1 to features with positive coefficients and -1 to those with negative coefficients, ignoring coefficients of 0. Feature importance selects the top 50% of features with positive scores as 1, and permutation feature importance assigns 1 to features with positive scores and zero otherwise. Summing these scores and sorting creates feature importance rankings for each subject out of 9.

The Impact Score Values, the second transformation, normalize scores by dividing each method's scores by the sum of all feature scores to ensure equal contribution to the final ranking. The summation of the absolute value of normalized scores forms feature ranking. The top 20 features with the highest scores are selected for Math and Reading, prioritizing the impact score, which integrates binary and non-zero scores. The binary score is a secondary measure for understanding importance, determining the cutoff point where the impact score drops. Secondary labels are applied to features to categorize their type, enhancing understanding of their significance. This approach facilitates comparing feature importance and identifying the most critical features in educational data analysis.

In this study, we applied several feature selection methods to determine their effectiveness in reducing the dimensionality of features for Math and Reading scores. The methods evaluated include RFE (RF and RR), VT, SFS (RF and KNN), FI RF, PFI (RR and RF), Elastic, Lasso, and Ridge, (*Loss, Gain, Expected*). Lasso, Elastic, PFI RR, PFI RF, and FI RF generate score outputs. Lasso resulted in 51 features for both Math and Reading. Figure 6 shows methods SFS KNN, SFS RR, and VT that focus on selecting subsets of features based on iterative addition or variance criteria. SFS methods iteratively add features to improve model performance, while VT removes low-variance features to enhance efficiency and accuracy. Figure 4 illustrates RFE RF and RFE RR employ recursive techniques to systematically reduce feature sets, facilitating the identification of the most influential features while optimizing computational efficiency. Grouping these methods aids in visualizing each feature's interpretability by each feature selection method. Figures 2 include F1 RF, PFI RR, and PFI RF, emphasizing feature importance assessment through recursive elimination or permutation testing. These methods evaluate how features contribute to model accuracy, identifying crucial predictors for refined model performance.

Figure 3 illustrated the most selected features such as # of *Families 10*, % *On Campus 01/29/21*, and % *On Campus 10/30/20* by feature selection techniques listed in Table 1). The # *of Families 10* showed significant relevance across all methods, indicating its robust influence on educational outcomes during disrupted learning environments. The %*On Campus 01/29/21* and % *On Campus 10/30/20* demonstrated varying impacts depending on the method employed.

The ten 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 RF FI, PMI (RF and RR), RFE (RR and RF) scores, and Elastic Gain and Elastic Loss. Since all of them have an importance ranking per feature (including the sign), we first normalize the scores for each method and then sum them as listed in Table 1.

Elastic reduced the features to 45 for Reading and 41 for Math. PFI RR identified 82 for Reading and 28 for Math, while PFI RF selected 26 for Reading and 70 for Math. FI RF resulted in 45 features for both subjects. Methods producing binary outputs include VT, SFS RR, and SFS KNN. VT reduced the features to 20 for both Reading and Math. Both SFS RR and SFS KNN selected 45 features for each subject. Methods producing rank outputs include RFE RR and RFE RF. RFE RR resulted in the most miniature feature set, with five features for Reading and six for Math, whereas RFE RF identified 36 features for both subjects. The Permutation Feature Importance (PFI) methods identified the most significant number of features, with PFI RF selecting 70 features for Math and PFI RR selecting 82 features for Reading. Table 4 presents the importance ranking of the features and summarizes the top 20 features for Math (a) and for Reading (b) selected by six or more methods scores in 2021 and 2022. The most significant feature predicting learning loss in Math is % *of Campus 10/30/20*, the enrollment of students in the campus district on 30 October 2020, representing the mode of instruction.

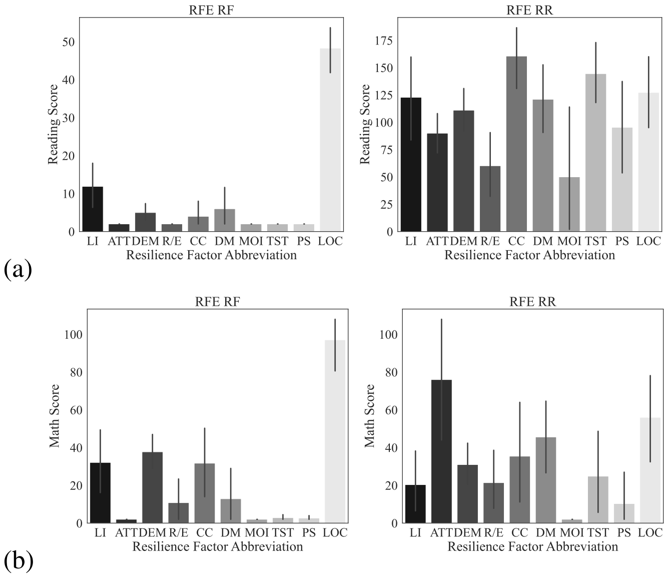


FIGURE 2. Filtering Feature Selection Methods for (a) Reading and (b) Math Comparison of RFE RF and RR.

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). Table 3 summarizes the top 20 impact features of the learning loss. Table 4 summarizes the top 20 features for learning recovery. The attribute is important if selected by six or more selection methods summarized in Table 1. Figure 4 illustrate that income and Grade level are the most influential resilient factors to predict learning loss for Math and Reading. The 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 primary features can identify the resilient factors for *Loss* or *Gain* in learning due to

TABLE 3. Top 14 Math Features for 2021 and 2022 Datasets

Math Scores 2021			Math Scores 2022		
Feature	Impact	Binary	Feature	Impact	Binary
Median Household Income	**6.62**	**5**	**Average Annual Pay**	**6.40**	**3**
Total Students 2018-2019	**6.23**	**7**	**Per Capita Income**	**6.27**	**4**
Total Students 2020-2021	6.14	6	Total Students 2021-2022	6.02	6
Total Students 2021-2022	6.11	7	County Population	5.92	5
Rural: Distant	6.05	3	# of Families 10	5.91	6
# of Families 10	5.84	4	Total Students 2018-2019	5.89	5
Average Annual Pay	5.83	2	Total Students 2020-2021	5.87	5
ARP ESSER III 21 NORM	5.76	3	# of Households 10	5.84	5
CARES ESSER I 20 NORM	5.76	4	% of Population Under 18 in Poverty	5.80	4
Rural: Remote	5.74	3	CRRSA ESSER II 21 NORM	5.81	4
# of Housing Units 10	5.70	3	Median Household Income	5.78	5
# of Households 10	5.70	3	# of Housing Units 10	5.78	4
Per Capita Income	5.70	3	Median Age Female 10	5.76	3
% of Population Under 18 in Poverty	5.68	3	% of Population in Poverty	5.77	4

TABLE 4. Top 14 Reading Features for 2021 and 2022 Datasets

Reading Scores 2021			Reading Scores 2022		
Feature	Impact	Binary	Feature	Impact	Binary
Median Household Income	**5.78**	**5**	**Total Students 2018-2019**	**5.89**	**4**
Total Students 2018-2019	**5.89**	**4**	**Total Students 2020-2021**	**5.87**	**5**
Total Students 2020-2021	5.87	5	# of Households 10	5.84	5
Total Students 2021-2022	5.85	5	% of Population Under 18 in Poverty	5.80	4
# of Households 10	5.84	5	CRRSA ESSER II 21 NORM	5.81	4
# of Housing Units 10	5.78	4	Median Household Income	5.78	5
Median Age Female 10	5.76	3	# of Housing Units 10	5.78	4
% of Population in Poverty	5.77	4	Median Age Female 10	5.76	3
Rural: Distant	5.70	3	% of Population in Poverty	5.77	4
CARES ESSER I 20 NORM	5.71	4	Rural: Distant	5.70	3
ARP ESSER III 21 NORM	5.69	4	CARES ESSER I 20 NORM	5.71	4
Median Age Male 10	5.66	3	ARP ESSER III 21 NORM	5.69	4
Median Age 10	5.59	2	Median Age Male 10	5.66	3
Rural: Remote	5.56	2	Median Age 10	5.59	2

the COVID-19 pandemic, whether those features positively impact learning is still unknown. 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, showed *Gain* or *Expected* in the same subject. The districts in need of financial help for adapting and preparing for learning *Loss* due to COVID-19 received the ESSER funds amounts calculated by a formula based on Title I and Part A grant [33], [34].

Figure 4 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%. t is clear that in-person classes, the mode of instruction, were the key to avoiding *Loss* in Math.

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. e 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. The boosting algorithm trains weak learners iteratively, and 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 it shows no more improvement for a certain number of consecutive rounds in terms of the evaluation metric specified as the fit parameter. The number of early stopping rounds is set to 10% of the maximum number of boosting iterations.

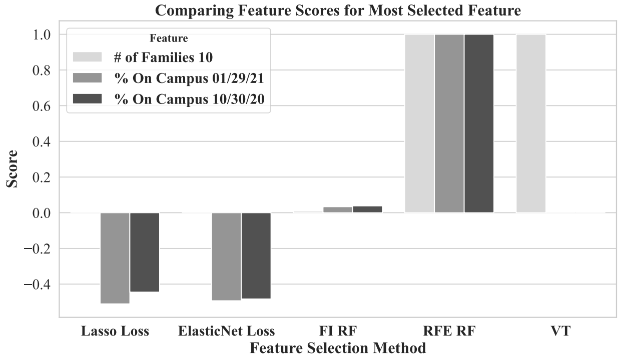


FIGURE 3. Most Selected Features: *# of Families 10*, *% of Campus 10/30/20*,and *% On Campus 01/29/21*.

We employed five state-of-the-art machine learning mod-

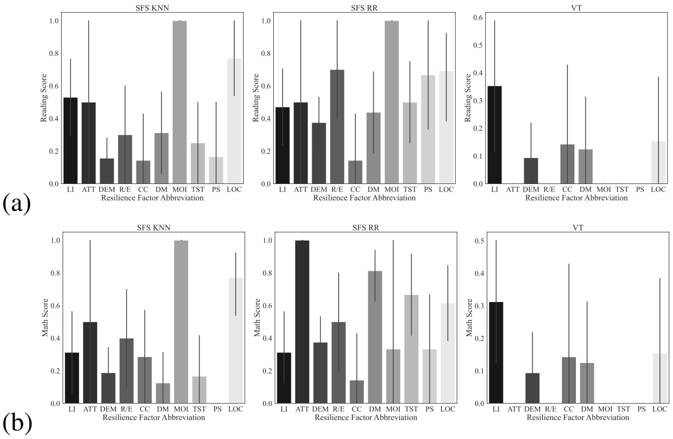


FIGURE 4. Importance Feature Selection Methods for (a) Reading and (b) Math: Comparison of SFS KNN, SFS RR, and VT for identifying significant resilience factors.

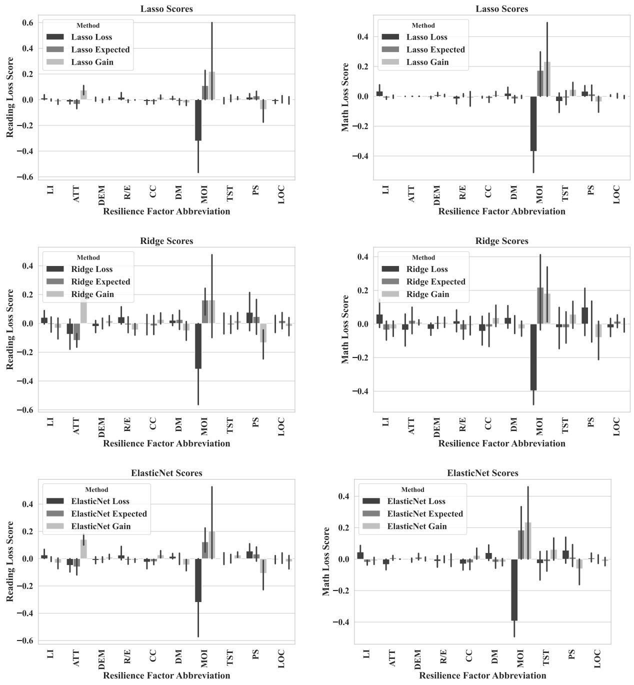


FIGURE 5. Comparison of Loss, Expected, and Gain Metrics for Lasso, Ridge, and Elastic Regularization Techniques to identify significant Reading (left) and Math (right) resilience factors

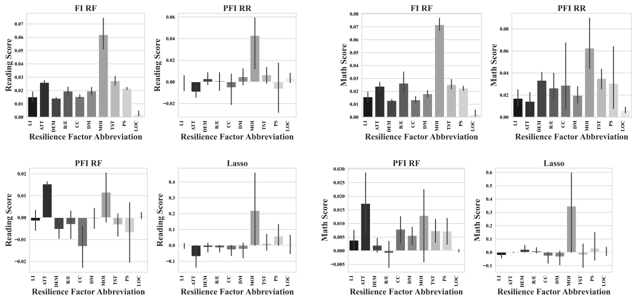


FIGURE 6. Scoring Feature Selection Methods for Reading (left) and Math (right) comparison of F1 RF, PFI RR, PFI RF, and Lasso for identifying significant resilience factors

els: Ridge Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forests, and GBDT and their variations, as summarized in Table 1. We trained the models using our complete set of 90 features and ten additional feature groups derived from various feature selection techniques, and Table 1 summarizes the model characteristics. Table 5 outlines the performance of the machine learning methods. Performance metrics, including accuracy, F1 score, and Matthews correlation coefficient (MCC), for these models, are presented in bar graphs in Figure 5 for baseline models and in Figure 5 for GBDT models. The prediction of learning loss for Reading exhibits weaker performance than for Math. Although most models perform similarly across both subjects, except KNN, GBDT for Math, and ridge regression for Reading, they demonstrate the highest average accuracy, F1 score, and MCC.

We penalize and regularize the algorithm by hyperparameter tuning so that we aim to increase accuracy and avoid overfitting to improve the gradient boosting modeling for XGBoost, LightGBM, CatBoost, and HistGradientBoosting. 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. XGBoost is supposed to explore 100 distributions of the parameters; the number of iterations for RandomizedSearch is 20 times. The constraints on tree structures aid in curbing the growth of overly complex and extensive trees, limiting the number of trees, tree depth, and the number of leaves per tree in the model. A lower learning rate (below 0.5) allows for gradually adjusting tree weights during each iteration, thereby minimizing errors. Ridge and Lasso regularization terms further the models by simplifying the complexity and size of the model [15]. The GBDT 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.

TABLE 5. Performance Metrics of Machine Learning Models Evaluated on the Test Set (20% of the Data). The table includes state-of-the-art models and advanced GBDT models.

Model	Accuracy	Precision	Recall	F1	MCC
GB	0.6329	0.6131	0.6329	0.5877	0.3513
KNN	0.5911	0.5731	0.5911	0.5548	0.2698
RF	**0.6339**	**0.6355**	**0.6339**	0.5651	**0.3602**
Ridge	0.6273	0.5974	0.6273	**0.5662**	0.3453
SVM	0.6268	0.5982	0.6268	0.5588	0.3441
CatBoost	**0.6337**	**0.6310**	**0.6337**	0.5778	**0.3543**
HistGB	0.6330.3	0.6157	0.6304	**0.5805**	0.3464
LightGBM	0.6281	0.6117	0.6281	0.5735	0.3415
XGBoost	0.6247	0.6047	0.6247	0.5635	0.3363

For Math, the best-performing model was CatBoost, achieving an accuracy of 67.5%, an F1 score of 64.5%, and an MCC of 43.4% using 36 features selected by RFE RF. thTherotable performances included Gradient Boost with an accuracy of 64.4%, an F1 score of 62.2%, and an MCC of 37.5% using the same feature selection method, and XG-Boost with an accuracy of 66.0%, an F1 score of 61.6%, and an MCC of 40.5% using 21 features selected by Variance Threshold (VT).

For Reading, CatBoost also emerged as the top performer, achieving an accuracy of 62.3%, an F1 score of 54.8%, and an MCC of 33.8% using 82 features selected by PMI Ridge. The second-best performance was from XGBoost, with an accuracy of 61.3%, an F1 score of 53.5%, and an MCC of 31.2% using 90 features from all feature sets. Third is the Ridge approach, with an accuracy of 60.7%, an F1 score of 52.2%, and an MCC of 30.3% using 45 features selected by SFS Ridge.

Overall, the GBDT algorithms, CatBoost and XGBoost, were the best choices among all the machine learning models tested for predicting learning loss in both subjects. Despite better predicting Math scores than Reading scores, the performance gap between the four GBDT models and the five other state-of-the-art models, except KNN, was negligible, with a difference in accuracy of around 3%.

Additionally, no single dimensionality reduction technique consistently outperformed others across all models. 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.

C. BEST FEATURES VS. RAW DATA FOR GBDT MODELS

In this section, we analyze the performance of four GBDT models—XGBoost, LightGBM, CatBoost, and HistGradientBoosting—on different datasets to evaluate their predictive power regarding learning loss in Math and Reading. These models handle data sparsity, including missing values, by finding optimal tree splits. The initial dataset, referred to as *LossB*, consists of 506 features (505 numerical and one categorical) across 1,165 school districts, with 416 features containing missing values ranging from 1% to 88%.

We compared the performance of models trained on three datasets: (1) the best feature sets identified through various feature engineering techniques, (2) raw data without imputation for missing values, and (3) raw data with missing values imputed using mean values. The subject-specific features differ for Math and Reading. Each subject had 302 features, with 212 features containing missing values.

The comparison in Table 5 showed that all models improved their performance metrics, especially the MCC, when using the best feature sets compared to raw data. istGradientBoosting exhibited the most significant improvement in MCC for Math, increasing by 47%, followed by LightGBM 43%, CatBoost 25%, and XGBoost 24%. In Reading, the improvement in MCC was even more pronounced, with HistGradientBoosting showing a 124% increase and LightGBM, CatBoost, and XGBoost improving by 45%, 43%, and 41%, respectively. Additionally, models trained on raw data without imputation performed slightly better than those with imputed data across all subjects and models. MCC for Math increased by over 6% for CatBoost and HistGradientBoosting, while XGBoost showed the most significant MCC growth for Reading, with an increase of 10%. In conclusion, gradient-boosted decision tree (GBDT) models trained on carefully selected feature sets significantly outperformed those trained on raw data, highlighting the importance of feature engineering in predictive modeling. Moreover, avoiding the imputation of missing values yielded better performance than mean imputation, emphasizing the models' capability to handle raw data effectively. Table 5 also illustrates that over ten feature selection methods, the GBDT models are robust against changes in feature subsets as the standard deviation of the results (in brackets is usually 1%). The models maintain similar performance levels regardless of the specific features used for training, which is beneficial in ensuring reliable predictions across different datasets or real-world applications.

VI. CONCLUSION AND FUTURE WORK

In this study, we employ a data-driven approach to investigate the impact of the COVID-19 pandemic on learning loss, utilizing an intentional data science pipeline. Despite employing ten distinct feature selection methods to facilitate the automatic extraction of crucial features from publicly available datasets, our findings reveal a limited influence on prediction accuracy across the nine machine learning models trained on feature-selected sets and raw data. Notably, GBDT algorithms, particularly XGBoost and CatBoost, consistently outperform other models, demonstrating remarkable efficacy in managing missing values in the raw datasets. Your reproducible experiments and datasets are accessible at [38], providing valuable tools for policymakers to strategically allocate resources and interventions to mitigate the effects of learning loss. A deeper analysis of 2021 to 2022 data revealed that shifts in feature significance primarily occurred at the individual feature level rather than through changes in resilience factor importance. Significantly, the mode of instruction and prior score emerged as the primary resilience factors during this period. Overall, low income and grade level proved to be the most influential factors in predicting learning loss in both Math and Reading. Noteworthy contributors to Math performance include attendance and census demographics, particularly the *% of Campus 10/30/20*. Additionally, students from low-income backgrounds and regions with higher unemployment rates were particularly impactful in predicting Reading learning loss. In future research, we aim to broaden the temporal scope of our analysis and incorporate more granular data sources to deepen our understanding of the enduring repercussions of the COVID-19 pandemic on education. Additionally, exploring novel feature engineering techniques or enhancing existing ones could bolster prediction accuracy across various datasets.

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

[3] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] A. Moubayed124, M. Injadat, N. Alhindawi, G. Samara, S. Abuasal, and R. Alazaidah, "A deep learning approach towards student performance prediction in online courses: Challenges based on a global perspective," Jan 2024.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] Y. Kim, Y.-K. Choi, and S. Emery, "Logistic regression with multiple random effects: A simulation study of estimation methods and statistical packages," *The American Statistician*, vol. 67, no. 3, pp. 171–182, 2013.

[23] 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.

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

[25] 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.

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

[27] 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.

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

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

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

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

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

[33] 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.

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

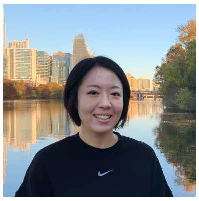
[35] R. Donnelly and H. A. Patrinos, "Learning loss during covid-19: An early systematic review," *Prospects*, vol. 51, pp. 601–609, 2022.

[36] 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.

[37] National Institute for Early Education Research, "Impacts of COVID-19 on Preschool enrollment and spending," 2021.

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

MIRNA ELIZONDO is a Ph.D. student in Computer Science at Texas State University and a Run2R1 Fellow. She holds a B.A. in Computer Science, awarded in 2020. Her research interests encompass network science, machine learning, and data science, focusing on developing predictive models and analyzing complex and noisy data structures in healthcare applications. She is a student member of the IEEE.



JUNE YU received her Bachelor's degree in Public Relations, Advertising, and Applied Communication from Kookmin University in 2020 and her Master's degree in Computer Science from Texas State University in 2022. Her research focuses on machine learning, data analysis, and data visualization, with a particular interest in data management and applications in marketing.

DANIEL PAYAN received a dual bachelor's degree in Computer Science and Business Management. His research interests are AI/ML modeling of large unstructured multi-source data, open-source dashboard development for promoting equity in science and education, and practical data science implementations. He is a Data Scientist at Love's Travel Stops and Country Stores.

LI FENG, PH.D. is a Professor of Economics in the Department of Finance and Economics at Texas State University, her research, funded by the NSF, IES, AERA, and NAS, focuses on education, labor, and health economics. She holds a B.A. in International Economic Cooperation from Xi'an Foreign Language University and a Ph.D. in Economics and Ed.S. from Florida State University.

JELENA TEŠIĆ, PH.D. is an Associate Professor at the Department of Computer Science, Texas State University. She received her Ph.D. from the University of California Santa Barbara, CA, USA. Dr. Tešić has authored over 80 peer-reviewed scientific papers and holds seven US patents. Her research, funded by NSF, DoD, DoE, and industry, focuses on unstructured data representation, analysis, AI/ML data modeling, and graph network science at scale. She is a senior IEEE member.