Data Driven Analysis of Intervention Effectiveness for COVID-19 Learning Loss in Texas Public Schools

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

***Abstract*—Student learning gain rates in public school systems in the US 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 public data sources were collected from 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 attributes. We demonstrate a novel data-driven approach to discover insights from a large collection of heterogeneous public data sources and offer an actionable understanding to policymakers to identify learning-loss tendencies and prevent them in public schools.**

I. Introduction

COVID-19 also had an impact on teacher preparation [Choate et al.(2021)]. A study indicates how COVID-19 has led many veteran teachers to retire early and novice teachers to consider alternative professions [Zamarro et al.(2022)]. The COVID-19 pandemic also forced many schools to close across the world [Zamarro et al.(2022)]. According to the latest UNESCO statistics, there are 43 million students affected by school closures and nationwide closures [OECD(2021)]. Even in high-income countries, such as the Netherlands and Belgium, learning loss ranged from 0.08 to 0.29 [Engzell et al.(2021)], [Maldonado and De Witte(2022)]. In a recent article, the global impact of a 5-month school shutdown could generate learning losses with a value of <10 trillion dollars [OECD(2021)].

In a recent paper, the global impact of a school shutdown of 5 months could generate learning losses with a present value of $10 trillion [OECD(2021)]. For the US context, 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 [Courtemanche et al.(2021)]. There is well-documented evidence that learning loss is not uniform across states such as Virginia, Maryland, Ohio, and Connecticut [Halloran et al.(2021)]. Recently, two states Rhode Island and North Carolina published two reports estimating the learning losses in these states ([Betebenner et al.(2021)], [NCD([n. d.])]. Texas Education Agency also published a report documenting the loss of learning [TEA([n. d.])]. There is no clear conclusion on what specifically led to the learning recovery in the aforementioned states, 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 US, researchers have disagreed on the impact of school reopening during the spread of COVID-19 [Choate et al.(2021)], [Courtemanche et al.(2021)]. This made it difficult for policymakers to decide when to reopen the school, and these varied between states, counties, and school districts [Rebai et al.(2020)]. The learning losses have not been uniform across the board [Halloran et al.(2021)], [Betebenner et al.(2021)]. 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 [Agency([n. d.])].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 pre-, during, and post-COVID-19 pandemic. We have collected data from eight public websites and processed data to find what specific factors were most important for the schools to experience a large 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 of influence for learning loss in the district is how early or late the students went back to in-person learning. *Missing: add 2022 findings for resilience*

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 [Cardona et al.(2020)]. From a data science perspective, the modeling approaches evaluated must be narrower in scope, and feature engineering almost guarantees poor domain/data translation results. A more elaborate evaluation of 30 selected articles revealed deep neural networks (DNN), decision trees, support vector machine (SVM), and nearest neighbor k (k-NN) as preferential methods to predict student academic performance [Rao et al.(2019)]. Demographic, academic, family/personal, and internal assessments were found to be the most frequently used attributes to predict student performance in class, at grade levels, on standardized tests, etc. [Baashar et al.(2021)]. A large-scale data science study correlated the Big Fish Little Pond Effect (BFLPE) in 56 countries in fourth grade math and 46 countries in eighth grade math using large data from the Trends in International Mathematics and Science Study (TIMSS) and a simple statistical analysis [Wang(2020)]. Recent findings show that the state of the art in machine learning in tabular data outperforms existing approaches and is not as sensitive to input bias and noise as DNN [Yan(2021)].

State-of-the-art gradient-boosted decision trees (GBDT) models such as XGBoost [Chen and Guestrin(2016)], LightGBM [Guolin Ke(2017)], and CatBoost [Dorogush et al.(2018)] 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 [Arik and Pfister(2021)], DNF-Net [Abutbul et al.(2020)], Neural Oblivious Decision Ensembles (NODE) [Popov et al.(2019)], and TabNN [Ke et al.(2019)]. 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. Open-source implementations, libraries, and their APIs are lacking [Shwartz-Ziv and Armon(2022)], [Joseph(2021)]. Recent studies provide competitive benchmarks comparing GBDT and deep learning models on multiple tabular data sets [Shwartz-Ziv and Armon(2022)], [Borisov et al.(2021)], [Gorishniy et al.(2021)], [Grinsztajn et al.(2022)]; 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 [Borisov et al.(2021)], [Grinsztajn et al.(2022)].

III. Proposed Methodology

Comment: add figure –Jelena The work introduces a unified data science pipeline for handling tabular data. It validates the pipeline from the data science application to educational data by predicting learning loss in math and reading scores in Texas public schools.

A. Attribute Importance Scoring

In this section, we propose a novel way to select important attributes from the hundreds of attributes considered. The work compares three different techniques for selecting features in data: filter methods, embedded methods, and wrapper methods. To evaluate these techniques, several algorithms for automated feature selection are tested, and a set of interpretable 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 tend to have a lot of overlapping information mixed with numerical and categorical data. With this filter method distilling correlated attributes mutually, our goal is 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 attributes in the modeling step, so we utilized this correlation filtering in this section to aggregate linearly related attributes in our data set into one attribute. To this end, we first have expanded several categorical attributes to multiple binary attributes as we found that multiple 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:

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

where $\text{cov}(X, Y)$ represents the covariance between variables X and Y, while σ_X and σ_Y are the standard deviations of X and Y respectively. The 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:

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

Here, $\bar{x}$ 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 attributes with high correlation coefficients to aggregate them into one attribute as they are linearly dependent each other. Eventually, we could keep one attribute, the most highly correlated to our label, of those overlapping attributes in our analysis. Then, we can decide to combine all binary dummy-coded variables from related categories as a set in variable selection. This approach thus reduces an attribute dimension that is providing better interpretability of our attribute set and its importance.

Multi-View Relevancy of the Attribute

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[Hooker and Mentch(2019)]. In this work, the final set of features includes any feature with positive mean importance, as the PFI method returns positive values for important 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 attributes. **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 attributes [Abe(2005)]. **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 attributes with low variance in the training data set[Ghojogh et al.(2019)]. 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 attributes 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 evaluating 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, some rank scores. In Alg. *Comment: add algorithm –Jelena* 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 approaches that provide scores for all attributes 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.

B. Prediction Modeling

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

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 [Shwartz-Ziv and Armon(2022)], our goal is to access on this real example the predictive power of these nine machine learning models. 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 approaches handle tricky observations well and are optimized in terms of faster and efficient fitting using data sparsity aware histogram-based algorithm.

In contrast to the pointwise split of the traditional Gradient Boosting that 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 [Chen and Guestrin(2016)]. Also, LightGBM provides the Gradient-Based One-Side Sampling technique, which is filtering data instances with 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 [Guolin Ke(2017)].

IV. Web Data Collection and Processing

A. Data Sources and Collection

We have collected data from eight different public sources as described in Table I. **Common Core of Data (CCD)** [for Education Statistics (NCES)([n. d.])] 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 Texas Education Agency (TEA) for the fiscal year 2019 and 2021 for each school district [(TEA)([n. d.]c)]. The STAAR data we collected are the average scores for math and reading tests and the number of students who participated in the tests for grades 3-8. These data also include the numbers and average scores for students 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) [of State Health Services (DSHS)([n. d.])], 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[USAFacts([n. d.])]. **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 [of Labor Statistics (BLS)([n. d.])] 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 [Bureau([n. d.])] 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 eight sources by matching school district ID and county FIPS code, the data set represents 1,165 school districts of Texas located in 253 counties with 506 attributes, consisting of 1 categorical and 505 numerical.

Data Frame	Data Source	Level	RowXCol
CCD	National Center for Education Stat [for Education Statistics (NCES)([n. d.])]	District	1189X66
STAAR	Texas Education Agency [(TEA)([n. d.]c)]	District	1184x217
TEA	Texas Education Agency [TEA([n. d.])]	District	1182x217
ADA	Texas Education Agency [(TEA)([n. d.]a)]	District	1226X3
ESSER	Texas Education Agency [(TEA)([n. d.]b)]	District	1208X6
census	Census Bureau 2010 [Bureau([n. d.])]	County	254, 37
Covid	USAFacts [USAFacts([n. d.])]	County	254X8
LAUS	U.S. Bureau of Labor Statistics [of Labor Statistics (BLS)([n. d.])]	County	254X13
Covid	DSHS [of State Health Services (DSHS)([n. d.])]	District	1216X7

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

CARES ESSER I 20, ARP ESSER III 21 attributes 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 [(TEA)([n. d.]b)], [(ESE)([n. d.])]. **CARES ESSER I:** Authorized on March 27, 2020, as the Coronavirus Aid Relief and Economic Security (CARES) Act with $13.2 billion. The availability period is from March 13, 2020, to September 30, 2022. Our data have the allocation amount for the fiscal year of 2020. **CRRSA ESSER II:** Authorized on December 27, 2020, as the Coronavirus Response and Relief Supplemental Appropriations (CRRSA) Act with $54.3 billion. The availability period is March 13, 2020, to September 30, 2023. Our data have the allocation amount for the fiscal year of 2021. **ARP ESSER III:** Authorized on March 11, 2021, as the American Rescue Plan (ARP) Act with $122 billion. The availability period is from March 13, 2020, to September 30, 2024. Our data have the allocation amount for the fiscal year of 2021. **ESSER-SUPP:** Authorized by the Texas Legislature to provide additional resources for unreimbursed costs to support students not performing well educationally. The availability period is March 13, 2020, to August 31, 2023. Our data have the allocation amount for the fiscal years 2022 and 2023.

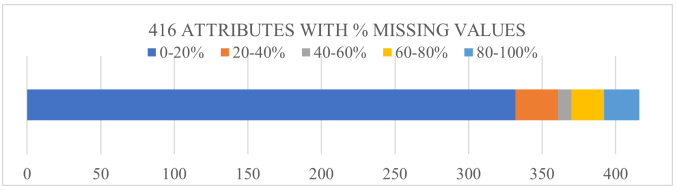


Fig. 1: Percentage of missing values for 416 attributes in the aggregated data.*Comment: add hiResimage –Jelena*

B. Data Aggregation and Filtering

We wanted to help policy makers make more informative decisions on learning recovery with localized efforts on each school district. Therefore, we collected data from eight difference 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 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) [for Education Statistics (NCES)([n. d.])] 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 schools characteristics, student demographics by grade, and faculty information at the school district in the state of Texas for the fiscal year 2019 and 2021. Then, we merged the CCD data with the State of Texas Assessments of Academic Readiness (STAAR) data [(TEA)([n. d.]c)] from Texas Education Agency (TEA) for fiscal year 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 tests for grade 3-8. These data also include the numbers and average scores for students 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) [of State Health Services (DSHS)([n. d.])], 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 [USAFacts([n. d.])] due to Coronavirus at each county from USAFacts. Also, the average daily attendance (ADA) [(TEA)([n. d.]a)], 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 [(TEA)([n. d.]b)] – COVID-19 relief funding – were collected from TEA for school district level. The ADA data for fiscal year 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 year of 2020, 2021, 2022 and 2023. Also, we combined the Local Area Unemployment Statistics (LAUS) data [of Labor Statistics (BLS)([n. d.])] from U.S. Bureau of Labor Statistics (BLS) for the year 2019 and 2021 to examine the negative impact of unemployment rate to learning loss at the county level. Additionally, Census block group 2010 data [Bureau([n. d.])] were included to grasp demographic characteristics at a county for general population. At the end of the initial data integration merging data from eight sources by matching school district ID and county FIPS code, the data set represents 1,165 school districts of Texas located in 253 counties with 506 attributes, consisting of 1 categorical and 505 numerical.

All eight sources were integrated by the district ID and county FIPS code, and the aggregated dataset covers 1,165 school districts of Texas located in 253 counties with 506 attributes, 1 categorical and 505 numerical.

The aggregated data set contains 506 attributes for 1,165 school districts in Texas. Among the 506 attributes, 416 attributes contain missing values from 3 data sources ranging from 1 to 88% in our data set: 408 attributes from STAAR, TEA, 6 attributes from CCD, NCES, and 2 attributes from COVID, DSHS data. Of these 416 attributes, 332 attributes have fewer than 20% missing values and 24 attributes have more than 80% of missing values, and the distribution is illustrated in Figure 1.

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

TABLE II: Example of 2019 and 2021 attribute aggregation

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

C. Data Labeling

Our data set is unlabeled; thus we need to create a ground truth label for further prediction processes. The data set contains average scale scores of the STARR for math and reading between grades 3 and 8 for the fiscal years of 2019 and 2021. This means that each school district has total of 24 attributes indicating the scores for calculating learning loss. We first normalized each cell of the scores by the maximum score value of the attribute as described in Figure 2, Step 1. Step 2 averaged these normalized scores for each year and subject, and Step 3 calculated the loss as the difference between the scores between 2019 and 2021 for the perspective of 2019. Consequently, our label – learning loss – is decided depending on the loss value: if it is positive, there is learning gain, but a negative value corresponds to learning loss. The distribution of the loss values in Figure **??** 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. Step 4 in Figure 2 describes creating 3 label classes; the middle 50% of school districts is labeled as "Expected", and the loss values below the 25th percentile are set to be "Loss", and the loss values above 75th percentile become "Gain".

With the data labeled as learning loss, Expected, and Gain, we analyzed each of them in depth with respect to a correlation between attributes and the label. For instance, Figure **??** reveals that White students are correlated to our label as they are the majority population for Gain and decreased towards Loss label; on the other hand, 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 **??**. Figure **??** (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

Comment: describe two different approaches here and compare –Jelena In the dataset *LossA*, we aggregate aggregate the 58 attributes that duplicate the data for 2019 and 2021 in 29 differential attributes as illustrated in Table II. For example, the attributes *Total Schools 2020-2021* and *Total Schools 2018-2019* are aggregated into *Total Schools Diff*, and the total number of attributes is reduced to 90. In the dataset *LossB*,

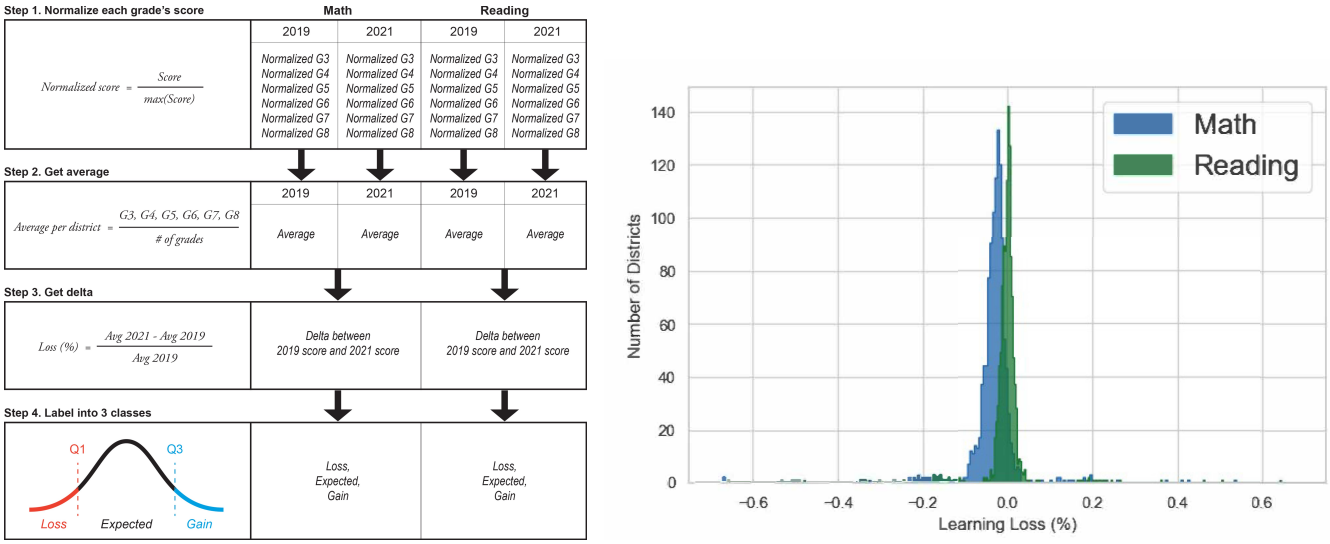


Fig. 2: Distribution of normalized STAAR scores between 2019 and 2021. More school districts in Texas faced learning loss in math than in the reading subject.Four steps to label learning loss with "Expected", "Loss", and "Gain" using the STAAR scores. First, normalizing each score, then getting averages and delta of the scores between 2021 and 2019.

we treat them as independent attributes. The experiment in comparing importance modeling of the two is illustrated in Section III-A.

V. Results

A. Setup

B. Attribute Importance Analysis

Attribute	Math	Attribute	Reading
% On Campus 10/30/20	5	CARES ESSER I 20	5
% Black Diff	5	CRRSA ESSER II 21	5
CARES ESSER I 20	5	ARP ESSER III 21	5
% Tested Math G8 Diff	4	% Prek Diff	4
% Reduced-price Lunch Diff	4	Unemployed Level Diff	4
% Asian Diff	4	ESSER-SUPP 23	4
% Grades 1-8 Diff	4	% Black Diff	4
CRRSA ESSER II 21	4	% Reduced-price Lunch Diff	4
ARP ESSER III 21	4	% Tested Reading G7 Diff	4
% PreK Diff	4	# of Families 10	4
Median Age 10	4	% Tested Reading G4 Diff	4
% On Campus 09/28/20	4	Avg Household Size 10	4
% White Diff	4	ESSER-SUPP 22	4
Rural: Distant	4	Median Age Female 10	4
ESSER-SUPP 22	4	% Asian Diff	4
Unemployed Level Diff	4	% County Deaths 10/30/20	4
% ADA Diff	3	% Free Lunch Diff	3
% Grades 9-12 Diff	3	% Grades 1-8 Diff	3

TABLE III: Top 18 attributes selected by ranking filtering outcomes of five approaches for math: 3 modes of instruction, 1 district attendance, 4 district race/ethnicity, 2 district poverty levels, 2 school population, and 3 census location.

We executed the ten different feature selection approaches described in Section III-A to detect the resilient factors for learning loss due to COVID-10 using the data set with 90 attributes 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. RF feature importance, RF permutation, and Ridge permutation importance provide non-zero scores for all 90 attributes, and RFE ridge regression and RFE Random forest provide attribute ranking. Table III sums up the filtering results. The five methods ranked 18 features as top importance and agreed on excluding 33 descriptors, mostly 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 attributes using Random Forest feature Importance, Random Forest permutation, Ridge permutation importance, RFE Ridge and RF scores, and ElastiNet Gain and ElastinNet Loss. Since all of them have importance ranking per feature (including the sign), first we normalize the scores for each method, and then we sum them.

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.

Table IV indicates the dimension reduced to the various numbers by each approach. 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 importance ranking of the features resulting from the ten approaches is shown in Figure IV, (a) Top 15 for math, and (b) Top 14 for reading selected by six or more feature selection methods. The most significant feature predicting learning loss in math is *% of Campus 10/30/20*, the enrollment of students in the campus district on October 30, 2020, representing the mode of instruction. For reading subject, three critical features

Method Index	Full name	Output	*Math*	*Reading*
LR Lasso	Logistic Regression with L1 Reg.	score	51	51
LR ElasticNet	Logistic Regression with L1+L2 Reg. score	score	Z	
PFI LR	Permutation Feature importance for LR L2 model	score	28	82
PFI RF	Permutation Feature importance for Random Forest model	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 model	rank	36	36

TABLE IV: Feature dimension is X. After method Y is applied the feature dimension is Z. *TODO: Need to find Elasticnet Z – Mirna*

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). Based on the characteristics of the top 15 (math) and 18 (reading) important features selected by six or more selection methods in Figure IV, we analyzed the resilient factors for seeking the most impactful factor among them. Low-income and Grade level is the most influential resilient factors to predict learning loss for math and reading, as shown in Figure VI. 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.

TABLE V: Resilient factors for Top 15 (math) and 14 features (reading). Low income and Grade level are the most impactful resilient factors for both subjects.

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

TABLE VI: *TODO: Here we will introduce another level of aggregation and present the aggregated impact score that way. We also need to determine if the overall impact was positive or negative. Please update the approximate labels in the Excel sheet, column A – Mirna*

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, in math and reading, we analyzed positive or negative correlations between the most critical features and our label, Loss, Expected, or Gain.

Figure 3 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 populated the most where the enrollment

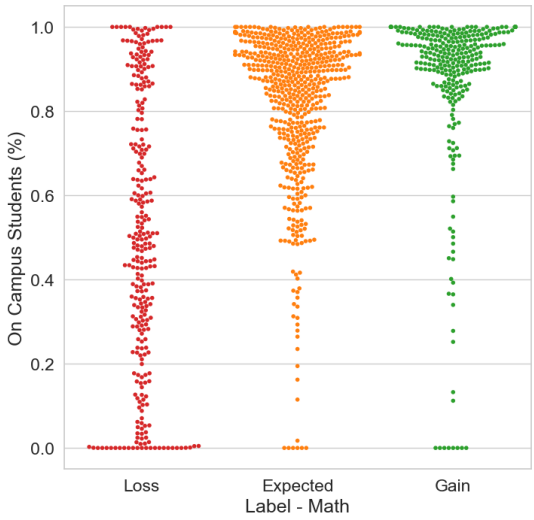


Fig. 3: Analysis on the most important feature for predicting learning loss in math: *% On Campus 10/30/20*. School districts in Gain and Expected label have more students who went to school on October 30, 2020.*TODO: Keeper. – Mirna*

is 0%. It is clear that in-person classes, the mode of instruction, were the key to avoiding Loss in math.

Figure*Missing: fig* **??** shows the distribution of each ESSER fund amount converted to the amount per student, 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. Meaning that the ESSER amounts 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 [(TEA)([n. d.]b)], [(ESE)([n. d.])].

C. Modeling Learning Loss from Public Data

Primarily, 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. To find the best model, we use performance metrics suitable for prediction problems. First, we look at the accuracy score for both problems to get a big picture. Then, F1 score is measured to reflect the precision and recall harmonically. Additionally, Matthews correlation coefficient (MCC) considers true negatives, class imbalance,

and multi-class of data. Each model runs with a 10-fold cross-validation of GridSearch to find optimal hyperparameters, see TableVIII. As the boosting algorithm trains weak learners iteratively, early stopping is used to reduce training time and avoid overfitting. At every round of the boost, 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 is set to 10% of the maximum number of boosting iterations are provided.

Five state-of-the-art machine learning models – ridge regression, SVM, KNN, random forests, and gradient boosting – fit our full set of 90 attributes and another ten different sets of selected features from RFE with ridge regression and random forests, Variance Threshold, SFS with ridge regression and KNN, random forests feature importance, Lasso regularization, and PMI with ridge regression and random forests as shown in *TODO: Table 4: 6, 21, 28, 45, 45, 45, 55, and 70 features for math, and 5, 20, 26, 36, 45, 45, 45, 51, and 82 features for reading.Needs to be confirmed – Mirna* The performance, accuracy, F1, and MCC of these models are plotted on bar graphs in Figure 6(a) for math and (b) for reading; predicting learning loss for reading shows weak performance compared to math generally. While no apparent differences between the performance of all models, except KNN, and the number of attributes have been observed for both subjects, gradient boosting for math and ridge regression for reading indicate the best accuracy, F1, and MCC on average.

For comparison purposes, four advanced gradient boost models, XGBoost, LightGBM, CatBoost, and HistGradient-Boosting, train the same sets of features. To improve the gradient boosting models, we can penalize and regularize the algorithm by hyperparameter tuning so that we aim at increasing accuracy and avoiding overfitting, see TableVIII. 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 search 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 longer 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, normally 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 [Chen and Guestrin(2016)]. 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 the 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.

DistrictA Math 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

DistrictA Reading 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

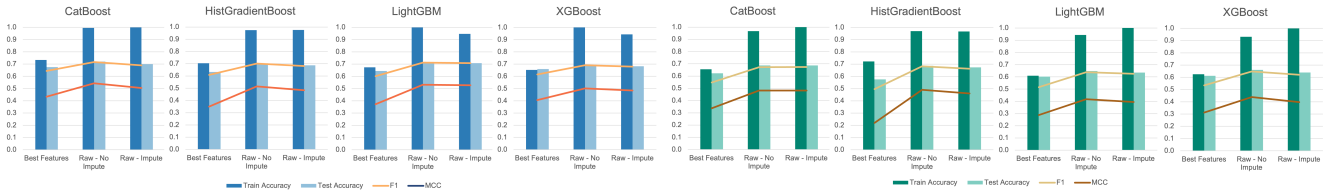
(b) Reading

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

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 is the best choice 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.

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 attributes (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, as shown in Figure ??*Missing: fig* In this experiment, we executed the pipeline of building the advanced gradient boosting models for raw data. We compared it with the models



(a) Train & Test Accuracy, MCC for Math; (b) Train & Test Accuracy, MCC for Reading Four advanced gradient boosting models training Raw data, including missing values with or without imputation. MCC improved compared to the results using the data with the best features selected through feature engineering in Table VII. *Comment: address this experiment. –Jelena*

Model	Hyperparameter(s)
Ridge Regression	Regularization strength: [0.001, 0.01, 0.1, 1, 10, 100] Solver for finding weights minimizing the cost function: ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
SVM	Regularization strength: [0.001, 0.01, 0.1, 1, 10, 100] Kernel type: ['linear', 'rbf']
KNN	Number of neighbors: [1, 3, 5, 7, 9, 11] Algorithm computing nearest neighbors: ['ball_tree', 'kd_tree', 'brute'] Leaf size passed to 'ball_tree' or 'kd_tree': [10, 30, 50]
Random Forest	Maximum depth of the tree: [1, 6, None] Number of trees in the forest: [50, 100, 200] Function measuring the quality of a split: ['gini', 'entropy'] Minimum number of samples at a leaf node: [1, 5, 10] Ratio of samples to draw from X: [0.1, 0.5, None] Maximum number of leaf nodes when growing trees: [10, 31, None] Complexity parameter for tree pruning: [0, 0.001, 0.1]
Gradient Boosting	Learning rate: [0.1, 0.2, 0.3] Number of boost iterations: [50, 100, 200] Minimum number of samples at a leaf node: [1, 5, 10] Minimum weighted fraction of total sample weights at a leaf: [0.0, 0.1, 0.5] Maximum depth of tree: [1, 3, 6] Maximum number of leaf nodes when growing trees: [10, 31, None] Complexity parameter for tree pruning: [0, 0.001, 0.1]
XGBoost	Number of boosting iterations: [50, 100, 200] Maximum depth of the tree: [1, 6, 0] Minimum sum hessian in one leaf: [0, 0.001, 0.1, 1] Learning/shrinkage rate: [0.01, 0.1, 0.2, 0.3] L1 regularization term (alpha): [0, 0.1, 10] L2 regularization term (lambda): [0, 0.1, 10] Minimum loss reduction (gamma): [0, 0.1, 10]
LightGBM	Number of boosting iterations: [50, 100, 200] Maximum depth of tree: [1, 6, -1] Minimum sum hessian in one leaf: [0, 0.001, 0.1, 1] Learning/shrinkage rate: [0.01, 0.1, 0.2, 0.3] L1 regularization term (alpha): [0, 0.1, 10] L2 regularization term (lambda): [0, 0.1, 10] Minimal gain to perform split: [0, 0.1, 10]
CatBoost	Number of boosting iterations: [50, 100, 200] Maximum depth of the tree: [3, 6, 9] Minimum number of samples per leaf: [1, 5, 10] Learning/shrinkage rate: [0.01, 0.1, 0.2, 0.3] L2 regularization term (lambda): [0, 0.01, 0.1, 1, 10] Amount of randomness for scoring splits: [0, 5, 10, 15]
HistGradientBoosting	Number of boosting iterations: [50, 100, 200] Maximum depth of tree: [1, 6, None] Maximum number of leaves for each tree: [10, 31, 50, 64] Minimum number of samples per leaf: [10, 20, 30] Learning/shrinkage rate: [0.01, 0.1, 0.2, 0.3] L2 regularization term (lambda): [0, 0.01, 0.1, 1, 10]

TABLE VIII: List of hyperparameters optimized for five state-of-the-art models & four advanced gradient boosting models *TODO: make it nicer ... I think I can find the optimal ones, just need to check all the gits again – Mirna*

trained the data processed the feature engineering techniques regarding prediction power on learning loss. The classification task was completed for the respective subjects, math, and reading. All attributes 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 attributes 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, as appeared in Figure *Missing: fig*; 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 can automatically uncover important attributes using public-use data and the ten feature selection methods to model learning loss due to COVID-19 in this paper. While the reduction in the dimensionality of data plays no role in the prediction power, as the ten machine learning models training the feature sets selected by the feature selection method did not exhibit significant improvement for the performance, the gradient boosting algorithms are generally performing better in both projects. The gradient boosting models such as XGBoost and CatBoost are superior for handling missing values as we experimented with raw data for the project; over 2/3 of attributes of the learning loss data sets contain missing values. Reproducible experiments and datasets are published on [Yu and Tešić([n. d.])]. Policymakers can use our predictive models and analysis to focus resources on the public school system, including schools, students, and teachers, to mitigate and recover learning loss with possible interventions in public schools.

References

[NCD([n. d.])] [n. d.]. https://www.dpi.nc.gov/about-dpi/state-board-education

[TEA([n. d.])] [n. d.]. https://tea.texas.gov/texas-schools/accountability/academic-accountability/performance-reporting/2021-tac-accountability-presentation-final.pdf

[Abe(2005)] Shigeo Abe. 2005. Modified backward feature selection by cross validation.. In *ESANN*. Citeseer, 163–168.

[Abutbul et al.(2020)] Ami Abutbul, Gal Elidan, Liran Katzir, and Ran El-Yaniv. 2020. DNF-Net: A Neural Architecture for Tabular Data. *CoRR* abs/2006.06465 (2020). arXiv:2006.06465 https://arxiv.org/abs/2006.06465

[Agency([n. d.])] Texas Educational Agency. [n. d.]. Impacts of COVID-19 and Accountability Updates for 2022 and Beyond. https://tea.texas.gov/sites/default/files/2021-tac-accountability-presentation-final.pdf.

[Arik and Pfister(2021)] Sercan Arik and Tomas Pfister. 2021. TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 8 (May 2021), 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826

[Baashar et al.(2021)] Yahia Baashar, Gamal Alkawsi, Nor'ashikin Ali, Hitham Alhussian, and Hussein T Bahbouh. 2021. Predicting student's performance using Machine Learning Methods: A Systematic Literature Review. *2021 International Conference on Computer and Information Sciences (ICCOINS)* (2021). https://doi.org/10.1109/iccoins49721.2021.9497185

[Betebenner et al.(2021)] Damian Betebenner, A Van Iwaarden, A Cooperman, M Boyer, and N Dadey. 2021. Assessing the academic impact of COVID-19 in summer 2021. *Center for Assessment* (2021).

[Borisov et al.(2021)] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2021. Deep Neural Networks and Tabular Data: A Survey. https://doi.org/10.48550/ARXIV.2110.01889

[Bureau([n. d.])] Census Bureau. [n. d.]. Census Block Group 2010. https://schoolsdata2-93b5c-tea-texas.opendata.arcgis.com/datasets/census-block-group-2010-tx/.

[Cardona et al.(2020)] Tatiana Cardona, Elizabeth A Cudney, Roger Hoerl, and Jennifer Snyder. 2020. Data Mining and Machine Learning Retention Models in Higher Education. *Journal of College Student Retention: Research, Theory & Practice* (2020), 1521025120964920.

[Chen and Guestrin(2016)] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *Information Fusion* (2016), 785–794. https://doi.org/abs/1603.02754

[Choate et al.(2021)] Kathryn Choate, Dan Goldhaber, and Roddy Theobald. 2021. The effects of COVID-19 on teacher preparation. *Phi Delta Kappan* 102, 7 (2021), 52–57.

[Courtemanche et al.(2021)] Charles J Courtemanche, Anh H Le, Aaron Yelowitz, and Ron Zimmer. 2021. *School reopenings, mobility, and COVID-19 spread: Evidence from Texas*. Technical Report. National Bureau of Economic Research.

[Dorogush et al.(2018)] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).

[Engzell et al.(2021)] Per Engzell, Arun Frey, and Mark D Verhagen. 2021. Learning loss due to school closures during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences* 118, 17 (2021).

[(ESE)([n. d.])] (ESE). [n. d.]. Elementary and Secondary School Emergency Relief Fund. https://oese.ed.gov/offices/education-stabilization-fund/elementary-secondary-school-emergency-relief-fund/.

[for Education Statistics (NCES)([n. d.])] National Center for Education Statistics (NCES). [n. d.]. Common Core of Data (CCD). https://nces.ed.gov/ccd/elsi/tableGenerator.aspx.

[Ghojogh et al.(2019)] Benyamin Ghojogh, Maria N Samad, Sayema Asif Mashhadi, Tania Kapoor, Wahab Ali, Fakhri Karray, and Mark Crowley. 2019. Feature selection and feature extraction in pattern analysis: A literature review. *arXiv preprint arXiv:1905.02845* (2019).

[Gorishniy et al.(2021)] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 18932–18943. https://proceedings.neurips.cc/paper/2021/file/9d86d83f925f2149e9edb0ac3b49229c-Paper.pdf

[Grinsztajn et al.(2022)] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on tabular data? https://doi.org/10.48550/ARXIV.2207.08815

[Guolin Ke(2017)] Thomas Finley et al. Guolin Ke, Qi Meng. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017), 3149–3157. https://doi.org/doi/10.5555/3294996.3295074

[Halloran et al.(2021)] Clare Halloran, Rebecca Jack, James C Okun, and Emily Oster. 2021. *Pandemic schooling mode and student test scores: Evidence from us states*. Technical Report. National Bureau of Economic Research.

[Hooker and Mentch(2019)] Giles Hooker and Lucas Mentch. 2019. Please stop permuting features: An explanation and alternatives. *arXiv e-prints* (2019), arXiv–1905.

[Joseph(2021)] Manu Joseph. 2021. PyTorch Tabular: A Framework for Deep Learning with Tabular Data. https://doi.org/10.48550/ARXIV.2104.13638

[Ke et al.(2019)] Guolin Ke, Jia Zhang, Zhenhui Xu, Jiang Bian, and Tie-Yan Liu. 2019. TabNN: A Universal Neural Network Solution for Tabular Data. https://openreview.net/forum?id=r1eJssCqY7

[Maldonado and De Witte(2022)] Joana Elisa Maldonado and Kristof De Witte. 2022. The effect of school closures on standardised student test outcomes. *British Educational Research Journal* 48, 1 (2022), 49–94.

[OECD(2021)] OECD. 2021. *Education at a Glance 2021*. Organisation for Economic Co-operation and Development. 474 pages. https://doi.org/10.1787/b35a14e5-en

[of Labor Statistics (BLS)([n. d.])] U.S. Bureau of Labor Statistics (BLS). [n. d.]. Local Area Unemployment Statistics (LAUS). https://www.bls.gov/lau.

[of State Health Services (DSHS)([n. d.])] Texas Department of State Health Services (DSHS). [n. d.]. Texas Public Schools COVID-19 Data. https://dshs.texas.gov/coronavirus/schools/texas-education-agency/.

[Popov et al.(2019)] Sergei Popov, Stanislav Morozov, and Artem Babenko. 2019. Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. *CoRR* abs/1909.06312 (2019). arXiv:1909.06312 http://arxiv.org/abs/1909.06312

[Rao et al.(2019)] A. Ravishankar Rao, Yashvi Desai, and Kavita Mishra. 2019. Data Science Education Through Education Data: An end-to-end perspective. *2019 IEEE Integrated STEM Education Conference (ISEC)* (2019). https://doi.org/10.1109/isecon.2019.8881970

[Rebai et al.(2020)] Sonia Rebai, Fatma Ben Yahia, and Hédi Essid. 2020. A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences* 70 (2020), 100724.

[Shwartz-Ziv and Armon(2022)] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81 (2022), 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

[(TEA)([n. d.]a)] Texas Education Agency (TEA). [n. d.]a. 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.

[(TEA)([n. d.]b)] Texas Education Agency (TEA). [n. d.]b. 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.

[(TEA)([n. d.]c)] Texas Education Agency (TEA). [n. d.]c. 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.

[USAFacts([n. d.])] USAFacts. [n. d.]. Texas Coronavirus Cases and Deaths. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/texas.

[Wang(2020)] Ze Wang. 2020. When large-scale assessments meet data science: The big-fish-little-pond effect in fourth- and eighth-grade mathematics across nations. *Frontiers in Psychology* 11 (2020). https://doi.org/10.3389/fpsyg.2020.579545

[Yan(2021)] Kuan Yan. 2021. Student performance prediction using XGBOOST method from a macro perspective. *2021 2nd International Conference on Computing and Data Science (CDS)* (2021). https://doi.org/10.1109/cds52072.2021.00084

[Yu and Tešić([n. d.])] June Yu and Jelena Tešić. [n. d.]. Tabular Data in the Wild: Gradient Boosting Modeling Improvement. https://github.com/DataLab12/educationDataScience.

[Zamarro et al.(2022)] Gema Zamarro, Andrew Camp, Dillon Fuchsman, and Josh B McGee. 2022. Understanding how Covid-19 has changed teachers' chances of remaining in the classroom. *Sinquefield Center for Applied Economic Research Working Paper No. Forthcoming* (2022).