Quantifying Texas Public School Learning Loss from Web Sources

Anonymous Author(s)

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. 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.

KEYWORDS

Educational data science, learning loss, gradient boosting, predictive modeling

ACM Reference Format:

Anonymous Author(s). 2024. Quantifying Texas Public School Learning Loss from Web Sources. In *Proceedings of The 17th ACM International Conference on Web Search and Data Mining (WSDM'24).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

COVID-19 had an impact on teacher preparation [10]. A recent study indicates how COVID-19 has led many veteran teachers to retire early and novice teachers to consider alternative professions [35]. The COVID-19 pandemic also forced many schools to close across the world [35]. According to the latest UNESCO statistics, there are 43 million students affected by school closures and nationwide closures [24]. Even in high-income countries, such as the Netherlands and Belgium, learning loss ranged from 0.08 to 0.29 [13, 23]. In a recent article, the global impact of a 5-month school shutdown could generate learning losses with a value of <10 trillion dollars [24]. In the US, researchers have disagreed on the impact of school reopening during the spread of COVID-19 [10, 11]. This made it difficult for policymakers to decide when to reopen the school, and these varied between states, counties, and school districts [29]. The learning losses have not been uniform across the board [5, 19]. 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 [3].

In this paper, we fuse and analyze multiple open sources of information related to public education in Texas on the web and introduce the applied data science pipeline. 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, and 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.

2 RELATED WORK

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. Related work focuses 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 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 [8]. From a data science perspective, the modeling approaches evaluated must be narrower in scope, and feature engineering almost guarantees poor domain/data translation results. State-of-the-art gradient-boosted decision trees (GBDT) models such as XGBoost [9], LightGBM [18], and CatBoost [12] 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 [4], DNF-Net [2], Neural Oblivious Decision Ensembles (NODE) [28], and TabNN [22]. 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 [21, 30]. Recent studies provide competitive benchmarks comparing GBDT and deep learning models on multiple tabular data sets [6, 16, 17, 30]; 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 [6, 17].

WSDM'24, March 04–08, 2024, Mérida, Mexico

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Comment: add figure –Jelena

3 METHODOLOGY

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.

3.1 Attribute Importance Scoring

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. **Variance Threshold** is a straightforward method to eliminate features by removing attributes with low variance in the training data set[15]. 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. **Lasso Regularization** is a technique that uses an L1 penalty term to shrink the coefficients during training. This reduces the coefficients of some features to zero, and the remaining non-zero coefficients are considered useful information for prediction. They are kept in the final feature set. **Random Forests Feature Importance** 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. **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 (ridge regression and random forest) on the training data decreases. The final set of features comprises the candidates that do not negatively impact the generalization performance of the model[1]. **Permutation Feature (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[20]. 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. **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 ridge regression and KNN models. We set to select half of the provided features for the final set. 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 attributes, and RFE ridge regression and RFE Random forest provide attribute ranking. Here, we propose to compare and evaluate the results of several fusion scoring mechanisms. 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.

3.2 Prediction Modeling

We go through two steps of the prediction modeling process to compare and analyze the feature sets selected using the attribute selection methods described in Section 3.1. First, we build baseline models using state-of-the-art machine learning methods. Then, we implement new robust gradient boosting models with gradient boosting to examine the performance and predictability of the models on each feature set. 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.

State-of-the-art Modeling. The choice of State-of-the-art Modeling is rather simple and less complex to train and interpret, as the purpose of having a baseline model is to provide benchmarks of its predictability and a deeper understanding of our data set. We have established five state-of-the-art models including ridge regression as the most common logistic classification model, Support vector machines (SVM) and K-nearest neighbor (KNN) for nonlinear and non-separable data, and two decision-tree-based ensemble methods: random forests and gradient boosting. Each model runs with a 10-fold cross-validation of GridSearch to find optimal hyperparameters.

Gradient Boosting Modeling. Our data fit the description of tabular data. Since gradient boosting approaches showed the most robustness when dealing with heterogeneous tabular data [30], we selected four advanced gradient boosting algorithms: XGBoost, LightGBM, CatBoost, and HistGradientBoosting. 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 [9]. 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 [18]. As the boost 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. 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. 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 [9]. 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.

Data Frame	Data Source	Level	RowXCol
census	Census Bureau 2010	County	254, 37
Covid	USAFacts	County	254X8
Covid	DSHS	District	1216X7
CCD	National Center for Education Stat.	District	1189X66
LAUS	U.S. Bureau of Labor Statistics	County	254X13
STAAR	Texas Education Agency	District	1184x217
TEA	Texas Education Agency	District	1182x217
ADA	Texas Education Agency	District	1226X3
ESSER	Texas Education Agency	District	1208X6

Table 1: 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.*Comment: add web links –Jelena*

4 WEB DATA AGGREGATION AND FILTERING

We have collected data from eight different public sources as described in Table 1. **Common Core of Data (CCD)** [14] 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 [32]. 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) [27], 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[33]. **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 [26] 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 [7] were included to see if the county's general population characteristics make a difference in learning loss. 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.

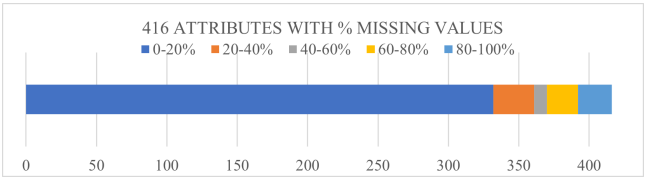


Figure 1: Percentage of missing values for 416 attributes in the aggregated data.*Comment: add hiResimage –Jelena*

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 [25, 31]. **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.

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 2: 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 in 29 differential attributes as illustrated in Table 2. 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.

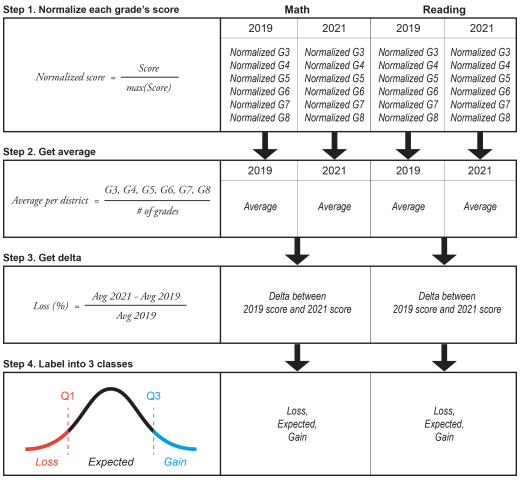


Figure 2: Four steps to label learning loss with "Expected", "Loss", and "Gain" using the STAAR scores. First, normalizing each score, then getting averages and delta of the scores between 2021 and 2019.

5 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 3 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".

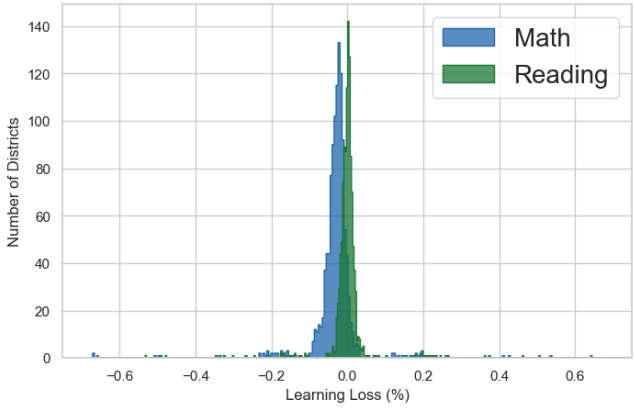


Figure 3: Distribution of normalized STAAR scores between 2019 and 2021. More school districts in Texas faced learning loss in math than in the reading subject.

6 ATTRIBUTE IMPORTANCE ANALYSIS

We executed the nine different feature selection approaches described in Section 3.1 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 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 3 sums up the filtering results. The five methods ranked 18 features as of 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

Attribute	Math Filter	Attribute	Reading Filter
% On Campus 10/30/20	4		
CARES ESSER I 20	4		
% PreK Diff	4		
% Asian Diff	4		
% Black Diff	4		
City: Mid-size	4		
Rural: Distant	4		
% On Campus 01/29/21	3		
% On Campus 09/28/20	3		
% ADA Diff	3		
% Reduced-price Lunch Diff	3		
% Title 1 Diff	3		
% Grades 1-8 Diff	3		
% Hispanic Diff	3		
ARP ESSER III 21	3		
CRRSA ESSER II 21	3		
Town: Remote	3		
Teachers:Students Diff	3		

Table 3: 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. ***Comment: finish –Jelena***

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.

Method	Selected Attributes	Method	Selected Attributes
Variance Threshold	Z	Random Forest	50%
PFI Ridge	Z	PFI Random Forests	z
LR w Lasso	Z	LR w ElasticNet	z
RFE Ridge	Z	RFE Random Forests	Z
SFS Ridge	Z	SFS KNN	Z

Table 4: Feature dimension is X. After method Y is applied the feature dimension is Z. ***Comment: finish –Jelena***

Table 4 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 nine approaches is shown in Figure 4, (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 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 14 (reading) important features selected by six or more selection methods in Figure 4, 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 6. 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 5: Resilient factors for Top 15 (math) and 14 features (reading). Low-income and Grade level is 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 6: ***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 – June***

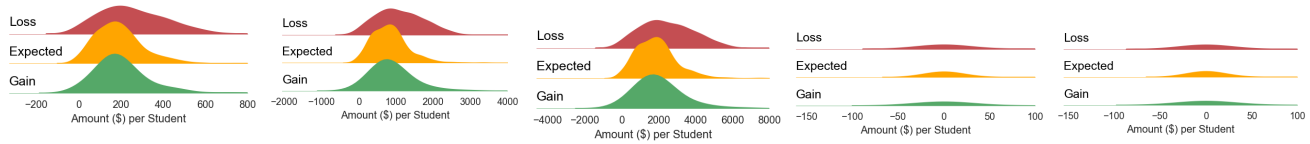
Although we now realize these essential features can identify the resilient factors for Loss or Gain in learning due to the COVID-19 pandemic, it is still unknown whether those features positively impact learning. For example, in math and reading, we analyzed positive or negative correlations between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

Figure 5 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 is 0%. It is clear that in-person classes, the mode of instruction, were the key to avoiding Loss in math.

Figure 4 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 [25, 31].

7 ANALYSIS AND PREDICTION MODELING OF LEARNING LOSS

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 3.2. Then, our initial data set was also experimented with gradient boosting models in terms of missing values and their imputation.



(a) CARES ESSER I (2020) (b) CRRSA ESSER II (2021) (c) ARP ESSER III (2021) (d) ESSER-SUPP (2022) (e) ESSER-SUPP (2023)

Figure 4: Analysis on the most important feature for predicting learning loss in reading: *CARES ESSER I 20, ARP ESSER III 21.* Five ESSER funding allocation for school district per student confirms that the funds have been distributed to the districts needing help as those districts have more students who experienced learning loss in reading.

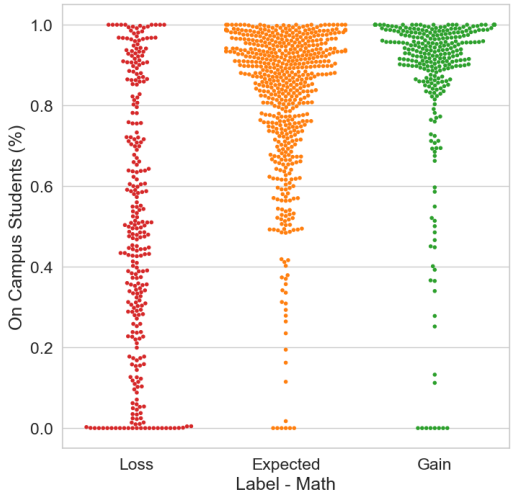


Figure 5: Analysis on the most important feature for predicting learning loss in math: *% On Campus 10/30/20.* School districts in Gain and Expected label have more students who went to school on October 30, 2020.*TODO: Keeper. – June*

Model	Best Set	Method	Acc [0,1]	F1 [0,1]	MCC [-1,+1]
Log Reg Ridge	45	Random Forests	0.639	0.622	0.368
SVM	45	SFS Ridge	0.628	0.584	0.343
KNN	55	Lasso Reg	0.618	0.591	0.318
Random Forests	45	Att scores	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	Variance Threshold	0.66	0.616	0.405

(a) Math

Model	Best Set	Selection Method	Acc [0,1]	F1 [0,1]	MCC [-1,+1]
Log Reg Ridge	45	SFS - Ridge	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 - Ridge	0.592	0.513	0.26
Gradient Boost	45	SFS - Ridge	0.56	0.542	0.231
CatBoost	82	PMI - Ridge	**0.623**	0.548	**0.338**
HistGB	45	SFS - Ridge	0.576	0.495	0.219
LightGBM	90	No Reduction	0.602	0.516	0.288
XGBoost	90	No Reduction	0.613	0.535	0.312

(b) Reading

Table 7: Performance of the nine machine learning models are trained for the ten features for (a) Math and (b) Reading.

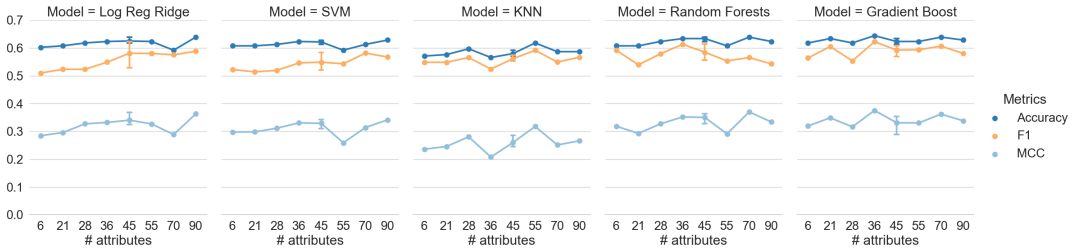
7.1 Model Evaluation and Comparison

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 nine 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 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. For comparison purposes, four advanced gradient boost models, XGBoost, LightGBM, CatBoost, and HistGradientBoosting, train the same sets of features. Including hyparameter optimization, details of these models establishments are described in Section 3.2.

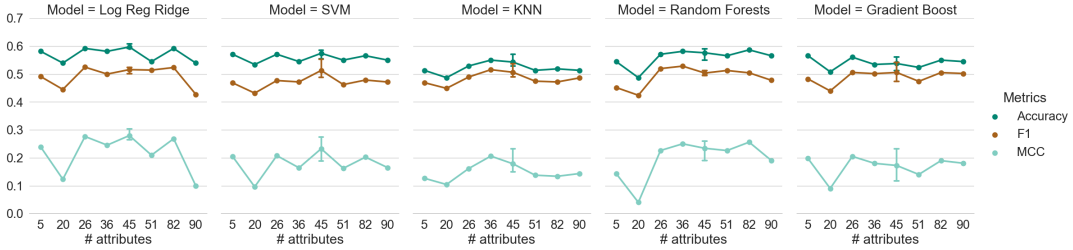
After training the five state-of-the-art models using 10-fold cross-validation of GridSearch for training (80%) testing (20%) split sets, the performance, accuracy, F1, and MCC of these models are plotted on bar graphs in Figure 6(a) for math and (b) for reading; predicting learning loss for reading 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.

We also train the four gradient boosting models for the same sets of features used above with 5-fold cross-validation of RandomizedSearch and train(80%) test(20%) split sets, and the performance comparison with the best state-of-the-art models, gradient boosting for math and ridge regression for reading, are shown in Figure 7, (a) math, and (b) reading. 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.

For the nine models, the best set of features for each model is described in Table 7 (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.

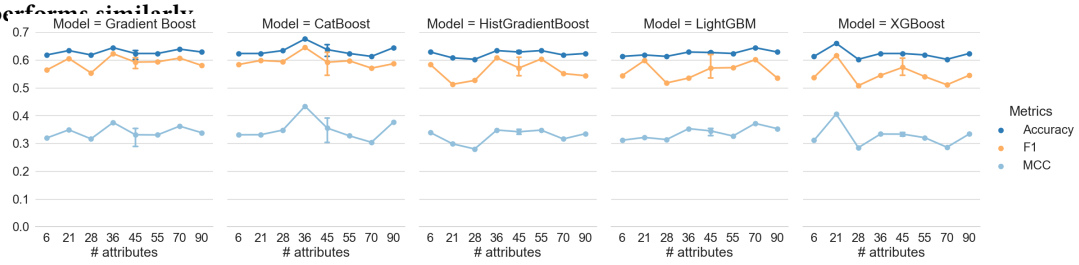


(a) Accuracy, F1, and MCC for Math

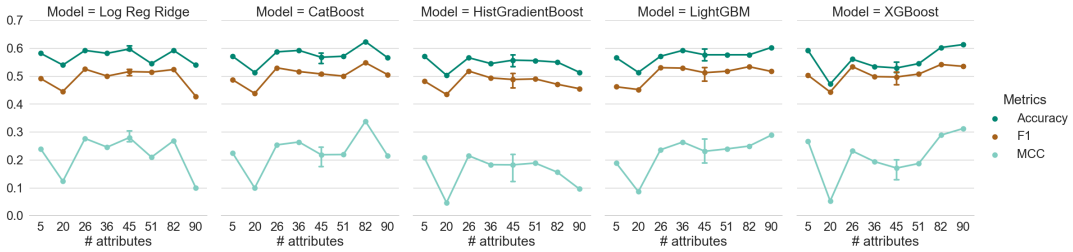


(b) Accuracy, F1, and MCC for Reading

Figure 6: Five state-of-the-art models fitted to 10 feature sets for predicting learning loss. With the train-test split, GridSearch, and 10-fold cross-validation, (a) gradient boosting for math and (b) ridge regression perform the best, while the rest, except KNN, also perform similarly.



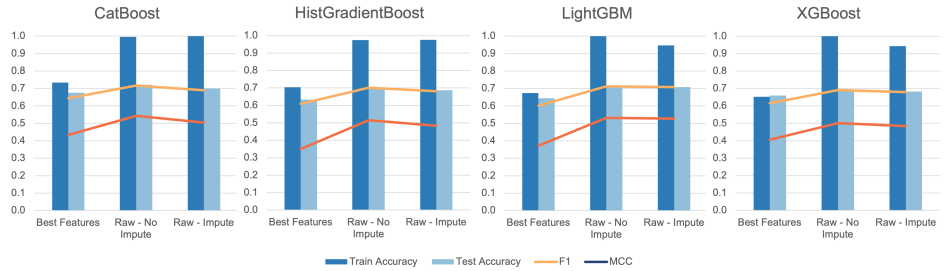
(a) Accuracy, F1, and MCC for Math



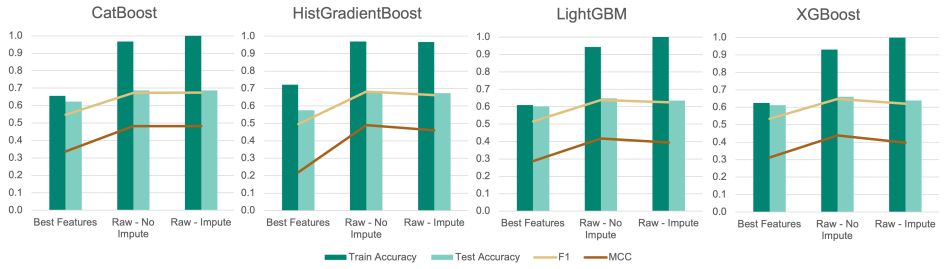
(b) Accuracy, F1, and MCC for Reading

Figure 7: Four advanced gradient boosting models fitted to 10 different feature sets for predicting learning loss. With train-test split, RandomizedSearch, and 5-fold cross-validation, the best state-of-the-art models, gradient boosting, and ridge regression are compared with for math in (a) and reading in (b).

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.



(a) Train & Test Accuracy, MCC for Math



(b) Train & Test Accuracy, MCC for Reading

Figure 8: 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 7.

7.2 Best Features vs. Raw Data for Gradient Boosting Models

All four gradient boosting models we 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 **??**. *Comment: fix this! –Jelena* In this experiment, we executed the pipeline of building the advanced gradient boosting models for raw data. We compared it with the models 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 7 from the performance results of the four gradient boosting models in Figure 7, (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 8; 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%.

8 CONCLUSION AND FUTURE WORK

The intentional data science pipeline can automatically uncover important attributes using public-use data and the nine 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 nine 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 [34]. 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

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

[2] 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

[3] 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.

[4] 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

[5] 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).

[6] 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

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

[8] 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.

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

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

[11] 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.

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

[13] 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).

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

[15] 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).

[16] 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

[17] LÃ©o Grinsztajn, Edouard Oyallon, and GaÃ«l Varoquaux. 2022. Why do tree-based models still outperform deep learning on tabular data? https://doi.org/10.48550/ARXIV.2207.08815

[18] 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

[19] 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.

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

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

[22] 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

[23] 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.

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

[25] OFFICE of Elementary & Secondary Education. [n. d.]. Elementary and Secondary School Emergency Relief Fund. https://oese.ed.gov/offices/education-stabilization-fund/elementary-secondary-school-emergency-relief-fund/.

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

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

[28] 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

[29] 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.

[30] 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

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

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

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

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

[35] 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).