GRADIENT BOOSTING PUBLIC DATA MODELING FOR THE POLICY PLANNING IN EDUCATION

by

June Yu, BA.

A thesis submitted to the Graduate College of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Computer Science
December 2022

Committee Members:

Jelena Tešić, Chair

Li Feng

Lu Wang

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

DEDICATION

In dedication to Lea, my wife, thank you for being only and all of encouragement.

ACKNOWLEDGEMENTS

I thank my advisor, Dr. Jelena Tešić, who has guided and inspired me to find the joy of research throughout my graduate studies from the beginning. I would also like to thank Dr. Li Feng for finding a practical application for my work in the education science, for guiding data pre-processing and integration using domain-specific context, and for and for being in my defense committee. A special thank you to Dr. Lu Wang for agreeing to join my thesis committee. I also thank Daniel Payan, a positive, driven, and hard-working undergraduate student at Texas State University, for his assistance with preprocessing new data sources. And thank you Richard Jason Pinkerton for helping with proofreading my writing. This research has been in part sponsored by the Center of Excellence for Community Health and Economic Resilience Research (CHERR) at Texas State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Teacher retention rate and student learning gain rates in the U.S. public school systems plummeted during 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 the factors of teacher attrition and learning loss. This is the first known study of the 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 teacher attrition and 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 to offer an actionable understanding to policymakers about the (1) recruitment and retention of public teachers, and (2) identifying learning loss tendencies and prevention of them in public schools.

I. Introduction

Teacher attrition in K-12 education is prohibitively high in all corners of the world [1]. Teacher attrition is defined as the number of teachers at a given level of education who leave the profession in a given school year, expressed as the percentage of teachers at that level and in that school year [2]. In 2016, the attrition rates in public institutions in the K-12 countries surveyed ranged from 3.3% in Israel to 11.7% in Norway [3]. In the United States, the teacher attrition rate was 8% on an annual basis. Statistical summaries now suggest that almost half of new teachers leave the profession in five years or less [4]. Texas has a prohibitively high teacher attrition rate of 10%, much higher than the national average. Historical trends in Texas show that 19% leave after one year and 12% after the second year. Half of newly trained and hired teachers leave the profession within 5 years [5].

Change in the teacher population is natural and desirable at a rate between 6% and 8% for public schools around the world [2]. If the teacher attrition rate in a school is less than 5%, it is likely that the school will stagnate. If the teacher attrition rate is greater than 10%, the effect can be detrimental to the effectiveness of a public school. The replacement of teachers has huge financial implications for the public budget [3]. A 2007 study estimated that turnover costs ranged widely from around \$4,000 per teacher (those leaving the New Mexico Public Schools) to almost \$18,000 per teacher (who left Chicago Public Schools) [6]. The study was used as a basis to estimate the total cost of excess teacher turnover in the United States in 2007 at \$7.34 billion annually with costs broken down to \$70,000 per school per year to cover the costs of teachers leaving that school with an additional \$8,750 spent to replace each teacher leaving the district [7]. The high teacher attrition rate is expensive and wasteful and also has a poor impact on student academic progress [8]. A high turnover of teachers reduces the effectiveness and quality of education [8].

COVID-19 also had an impact on teacher preparation [9]. A recent study indicates how COVID-19 has led many veteran teachers to retire early and novice teachers to consider alternative professions [10]. The COVID-19 pandemic also forced many schools to close down across the world [10]. According to the latest UNESCO statistics, there are 43 million students affected by the school shutdowns and country-wide closures [3]. Learning loss even in high-income countries, such as the Netherlands and Belgium, ranged from 0.08 to 0.29 standard deviation [11, 12]. In a recent paper, the global impact of school shutdown of 5 months could generate learning losses that have a present value of $10 trillion [3]. In the US, researchers have not reached a consensus on the impact of school reopening during the spread of COVID-19 [9, 13]. This made it difficult for policy makers to decide when to reopen the school, and these varied between states, counties, and school districts [14]. Thus, the learning losses have not been uniform across the board [15, 16]. The Texas Education Agency published a report documenting the loss of learning in public schools (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 [17]. In this thesis, we propose a unified data science approach to address these two issues: mitigating the U.S. public school teacher attrition crisis, discussed in section IV and identifying resilience factors in Texas public schools in section V.

II. Related Work

In this paper, we propose a novel data-driven approach for public data integration and analysis on a scale, automated attribute importance analysis, and robust prediction modeling. Therefore, we grouped related work into three categories: “Data Science”, “Machine Learning”, and “Economy of Education”. The first group focuses on quantitative research and the use of machine learning tools to gain insight from the data on the relationship with the outcome without overfitting the features to the data. The second group provides directions for selection of machine learning models on predicting teacher attrition and learning loss with given data. Finally, the last group of research products focuses on qualitative research, where the objective is to propose, analyze, and establish the relevance of a single attribute to the teacher attrition rate.

Data Science The application of machine learning (ML) tools for the correlation of attributes with teacher attrition rates has increased from two studies in 2010 to seven studies in 2017 [18]. The most popular ML techniques (logistic regression, support vector machines, Bayesian belief network, decision trees, and neural network) generally offer a good classification accuracy above 70% for simple classification tasks [18]. From a data science perspective, the modeling approaches evaluated are too narrow 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 [19]. An even more elaborate review of 25,771 studies selected 120 quantitative data analyses of teacher turnover in their meta-analysis, and the methods and data sets evaluated suffer from the same drawback as the overfitting attributes used in modeling [20]. 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. [21]. 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 [22]. 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 [23].

Machine Learning State-of-the art gradient boosted decision trees (GBDT) models such as XGBoost [24], LightGBM [25], and CatBoost [26] 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 [27], DNF-Net [28], Neural Oblivious Decision Ensembles (NODE) [29], and TabNN [30]. Although papers have proposed that these deep learning algorithms are outperforming the GBDT models, there is no consensus that deep learning is exceeding GBDT on tabular data because standard benchmarks have been absent and open-source implementations, libraries, and their APIs are lacking [31, 32]. Recent studies provide competitive benchmarks comparing GBDT and deep learning models on multiple tabular data sets [31, 33, 34, 35]; however, all of these benchmarks indicate that there is no dominant winner, and GBDT models still outperform deep learning in general. The studies suggest to develop 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 [33, 35].

Economy of Education Teacher turnover, teacher attrition, teacher retention, and teacher recruitment have been analyzed in worldwide educational literature [3], taking into account specific societal characteristics that influence teachers to quit their jobs in Sweden [36, 37], South Korea [20, 38], the United States [1, 39, 40, 41], Canada [42], Finland [43], Nepal [44] and many other countries. All studies handpicked attributes to explain teacher turnover: teacher characteristics, teacher qualifications, school organizational characteristics, school resources, student body characteristics, relational demography, accountability, and workforce measures.

III. Methodology

The work proposes a unified data science pipeline for tabular data in the wild, and validates the pipeline using teacher attrition data to predict whether a teacher will leave teaching or not (Section IV), and using publicly available education, COVID and census data to predict learning loss in math and reading scores in Texas (Section V). The open multi source data for both cases are heterogeneous tabular data, and we apply the same methodology for attribute selection and prediction modeling.

Attribute Selection and Automated Importance Scoring

The work compares and contracts three major approaches for feature selection: filter methods, embedded methods, and wrapper methods. For each approach, we implement several data-driven automated attribute selection algorithms and offer an interpretable suite of attribute importance analysis approaches and to avoid traps of the 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 linear relationships between two normal distributed variables as $\rho = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y}$. Pearson's coefficient estimate r, also known as a "correlation coefficient," for attribute feature vector $x = (x_1, ...x_n)$ with mean $\bar{x}$ and attribute feature vector $y = (y_1, ...y_n)$ with mean $\bar{y}$ is obtained via a Least-Squares fit as defined in Eq. III.1 as:

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

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

To select and have a glimpse of the features that affect our prediction models, we compare and contrast nine different approaches from the three methods mentioned above—filter methods, embedded methods, and wrapper methods—to evaluate the importance of features. Every approach of selecting minimum redundancy and maximum relevancy feature set yields either a set of features selected or a score of feature importance so that we can reduce the dimensionality of feature space.

- Filter methods:
  - **Variance Threshold**: This is a simple and powerful approach to remove attributes by eliminating all low-variance attributes in the training set [45], since there is no meaningful information in that attribute. We applied the threshold, 0.8*(1-0.8), to the training data set to remove a characteristic containing 80% similar values and select the most relevant attributes of k with the highest variance to the final set.
- Embedded methods:
  - **Lasso Regularization**: This logistic regression with the penalty term L1 shrinks the coefficients by minimizing the loss function during training. As the method reduces the coefficients of the features to be exactly zero, every feature with a nonzero coefficient value is considered and selected in the final set as useful information on prediction, which decreases the variance.
  - **Random Forests Feature Importance**: random forests is a powerful machine learning classification algorithm. The algorithm has a built-in attribute importance measured by the Gini importance or mean decrease impurity. This built-in feature of the algorithm that selects a feature with higher certainty returns the importance of the feature, and we set a threshold of the 50th percentile of the attribute importance to include a relevant attribute to the final set. Interpreting the importance of these selected features in tree-based machine learning models is challenging when features are dependent [46] as tree split can be biased due to the correlation between these dependent features.
- Wrapper methods:

- **Recursive Feature Elimination (RFE)** algorithm starts by fitting the two models to the full set of attributes in our data set, so we can eliminate candidates with the smallest coefficient and remove the importance of the features of the ridge regression and random forest, respectively, which deteriorate the 10-fold cross-validation score of the models in the training data. Attributes that are ranked according to the importance of characteristics in penalized regularization regression modeling on a small scale have been supported in the qualitative research literature [47]. The final set is a set of candidates that does not deteriorate the generalizability of the model [48] with the proper number of features selected by 10-fold cross-validation.
- **Permutation Feature Importance(PFI)** method replaces the values of a feature with redundant noise and measures the difference in the accuracy score or other performance metrics between the baseline and the permuted data set. Although PFI overcomes limitations on the importance of impurity-based characteristics, since the importance of the features drawn from the method does not have a bias toward high-cardinality attributes, it suffers from a bias caused by the correlation between features, as the impurity-based features are [49]. The final set contains any feature with positive importance mean as the method returns positive and higher values for more important features.
- **Sequential Feature Selection (SFS)** model selects an optimal set of features by searching the feature space of all possible combinations in a greedy manner. Each subset of features that add one predictor at a time forward is evaluated based on the 5-fold cross-validation score of the ridge regression and KNN models. As this greedy algorithm adds an attribute one by one, it requires more computational efforts. The method

is set to select a half of provided attributes to the final set.

Prediction Modeling

The two problems, (1) teacher attrition and (2) learning loss, 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 III. 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.

Data preprocessing. 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 labels.

Evaluation metrics. To find the best model, we use performance metrics that are suitable for prediction problems. First, we look at accuracy score for both problems to get a big picture. Then, F1 score is measured to reflect precision and recall harmonically. Additionally, Matthews correlation coefficient (MCC) is regarded to consider 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 deeper understanding of our data set. We have established five state-of-the-art models including the ridge regression as the most common logistic classification model, Support vector machines (SVM) and KNearestNeighbor (KNN) for nonlinear and non-separable data, and two decision-tree-based ensemble methods: random forests and gradient boosting.

Hyperparameter optimization. Each model runs with a 10-fold cross-validation of GridSearch to find optimal hyperparameters, and these hyperparameters for optimization are listed in Table 1:

Table 1: List of hyperparameters optimized for five state-of-the-art models: ridge regression, SVM, KNN, random forests, and gradient boosting

Model	Hyperparameter
Ridge regression	Regularization strength: [0.001, 0.01, 0.1, 1, 10, 100] Solver finding weights minimizing 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 forests	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 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 growing trees: [10, 31, None] Complexity parameter for tree pruning: [0, 0.001, 0.1]

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 [31], 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 [24]. 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 [25].

Handling categorical features. Handling categorical features is a challenge in building a machine learning model for tabular data. While there are several ways to process representing categorical features such as one-hot and ordinal encoding, tree building and and splitting with these methods often result in unbalanced trees and the sparsity of data, especially for high-cardinality categorical features. The four gradient boost models implement and suggest optimal methods for processing categorical features to optimize numerous boost steps for computing time and memory consumption. LightGBM, HistGradientBoosting, and XGBoost use the optimal split method [50] to group the categories of a feature and classify them as continuous partitions according to the target variance to find the best split in the histogram of sorted gradients[51, 52, 53]. CatBoost, however, proposes the method called Ordered Target Statistics (TS), which improves the target encoding method by using the history of all training data to compute TS instead of the target on a test set [54]. All four models accept hyperparameters to handle categorical features, such as categorical feature indices or thresholds to control one-hot encoding or the number of tree split points.

Early stopping rounds. 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 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.

Hyperparameter optimization. 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 [24]. These hyperparameters in Table 2 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 a total of 100 distributions of the parameters; the number of iterations for RandomizedSearch is 20 times.

Table 2: List of hyperparameters optimized for four advanced gradient boosting models: XGBoost, LightGBM, CatBoost, HistGradientBoost

Model	Hyperparameter
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]

IV. Mitigating U.S. Public School Teacher Attrition Crisis

United States Educational Public Data Summary

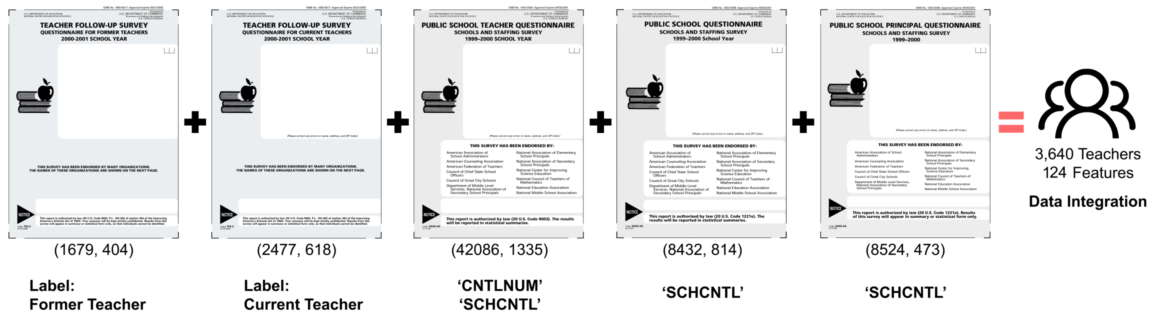


Figure 1: NCES Data Integration from 5 sources. First, TFS-Former Teacher and TFS-Current Teacher data are concatenated with labeling 1: Current and 0: Former teachers. Then, SASS data are joined with TFS in the order of Public Teachers, Public Schools, then Public Principals by matching control numbers such as SCHCNTL and CNTLNUM.

The National Center for Education Statistics (NCES) is the statistical agency that collects all education-related data in the United States of America. NCES collects international assessment data, administrative data on all public schools in the United States, and national survey data and provides them to the research community to inform policy and practice [55]. The Schools and Staffing Survey (SASS) was an integrated multiyear study of public and private school districts, schools, principals, and teachers designed to provide descriptive data on the context of elementary and secondary education [56]. NCES and TFS led SASS seven times between 1987 and 2011, 1987-1988, 1990-1991, 1993-1994, 1999-2000, 2003-2004, 2007-2008, and 2011-2012; however, the last three surveys are restricted-use data [57]. SASS covered a wide range of topics, such as teacher demand, teacher and principal characteristics, general conditions in schools, principals and teachers' perceptions of school climate and problems in their schools, teacher compensation, district hiring and retention practices, and basic characteristics of the student population [56]. The Teacher Follow-Up Survey (TFS) was a survey conducted a year after the SASS survey. TFS surveys K-12 teachers who participated in SASS a year earlier [56]. The collected data consist of a subsample of teachers who left teaching within the year after SASS was administered and a subsample of those who continued teaching, including those who remained in the same school as in the previous year and those who changed schools [56].

In this work, we analyze the latest data and documents available for public-use in 1999-2000 SASS and 2000-2001 TFS [56] in public schools, public school teachers, and public school principals. Raw data include hundreds of attributes on teacher demand, teacher and principal characteristics, general school conditions, principals and teachers perceptions of school climate, teacher compensation, district hiring and retention practices, and student demographics. We use these unchanged attributes in our data science analysis. Furthermore, the TFS data contain binary labels on the decision of teachers to stay teaching (1) or leave teaching (0). The data integration pipeline is illustrated in Figure 1. Of 42,086 public teachers who participated in the School and Staffing Survey (SASS) 1999-2000, only 4,156(<10%) of the teachers participated in the Teacher Follow-Up Survey (TFS) 2000-2001, that is, 2,477 current and 1,679 former teachers. 76.6% of the schools in the dataset have at least 1 teacher who participated in SASS and TFS. 301 current and 215 former teachers did not have the TFS data on the principal and school association, so we excluded them for the labeled data. As a result, the initial set of 124 attributes consists of 107 categorical attributes and 17 numerical attributes for 3,640 teachers, and the 124 attributes include 70 attributes from public teachers, 9 attributes from public principals, and 45 attributes from public schools. We observed an interesting correlation of known qualitative attributes that affect the teacher attrition rate [40] in Fig. 2. Our data shows that female teachers are 2/3 majority, while the turnover rate is higher for male teachers (Fig. 2(a)); white non-Hispanic teachers are the

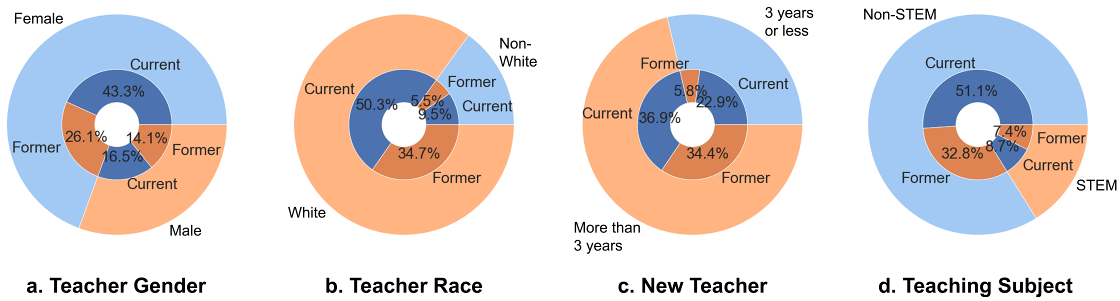


Figure 2: SASS and TFS Exploratory Retention Analysis for gender, race/ethnicity, new teacher, and teaching field. (a) female teachers are 2/3 majority while the turnover rate is higher for male teachers; (b) white(Non-Hispanic) teachers are the majority race/ethnicity in U.S. public schools with the higher turnover rate; (c)(d) Teachers working more than 3 years and teachers teaching STEM subjects have higher turnover rate.

majority race/ethnicity group in public schools in the US, and they have the highest attrition rate (Fig. 2(b)); and the highest attrition yearly rate is for teachers working more than 3 years (Fig. 2(c)) and for the teachers teaching STEM subjects (Fig. 2(d)).

Attribute Aggregation, Selection, and Importance

We use a complete set of attributes for the attribute selection methods introduced in Section III. With the selection of features using these methods, a final modeling labeled data set is prepared for the prediction modeling with 3,640 teachers from 2,838 schools, comprised of 53 attributes and labels of 2,176 current teachers and 1,464 former teachers.

Correlation Filtering for Overlapping Attributes

SASS and TFS data provide a lot of overlap information [56]. Our data set has 124 attributes for the 3,640 teachers, 107 categorical attributes, and 17 numerical attributes, after initial integration. Before calculating correlation of coefficients and examining linearly dependent attributes, we first treated categorical attributes with

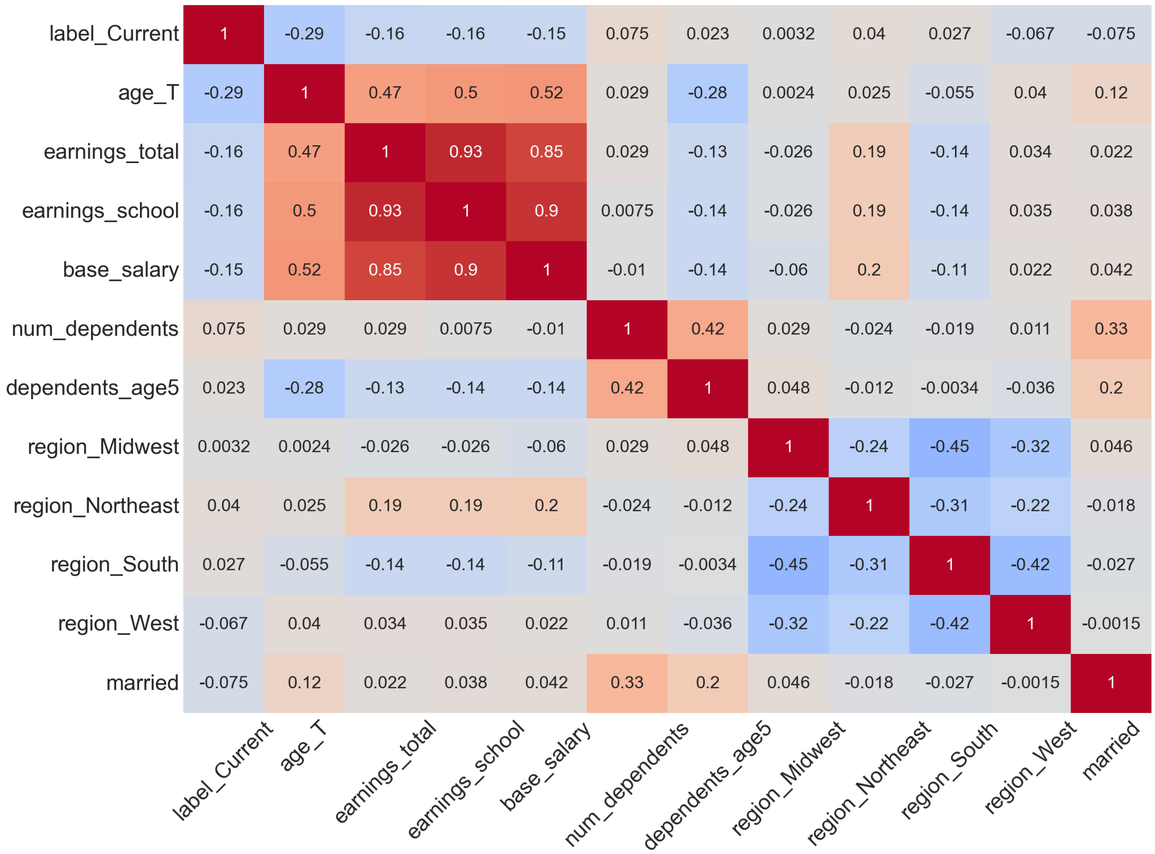


Figure 3: Attribute Correlation Analysis of the SASS/TFS data. The *base_ salary* is highly linearly correlated with the *earnings_ school* and *earnings_ total*. We use the high correlation coefficient to aggregate linearly dependent attributes into one.

high-cardinality of categories, e.g. 80 categories as the major codes for the teachers' BA or MA degrees into two categories, STEM or non-STEM major, and then the rest of the categorical attributes were expanded to multiple binary attributes to detect highly overlapping data. Our expanded set contains a total of 134 categorical and 17 numerical attributes: 78 attributes for public teachers, 17 attributes for public principals, and 56 attributes for public schools. The correlation coefficients of the expanded data are illustrated in Figure 3, e.g. the base_salary is highly correlated with the earnings_school and earnings_total attributes.

With the correlation coefficients of our data, we combined all binary variables if they can be related categories as a set, as in the example of *teaches_ 7to12* that is

Table 3: Aggregated attributes filtered by correlations in SASS and TFS data

From Labels	New Label	From Labels	New Label
teaches_7th teaches_8th teaches_9th teaches_10th teaches_11th teaches_12th	teaches_7to12: Teaching 7 to 12th grades (1 0)	deg_P_Associate deg_P_Bachelors deg_P_Masters deg_P_Edu deg_P_Doctorate	deg_highest_P: Principal's highest degree (5 categories)
pd_stipend pd_tuition_r pd_conference_r pd_travel_r	pd_finance: Professional development pay (1 0)	hrs_tch_math hrs_tch_science	hrs_taught_STEM: Hours of teaching STEM subjects per week
pd_release_t pd_schedule_t	pd_time: Professional development time off(1 0)		
vacnc_gen_elem vacnc_spec_ed vacnc_english vacnc_soc_st vacnc_esl vacnc_foreign_lang vacnc_music_or_art vacnc_vo_tech	vacnc_NonSTEM: Difficulty filling the vacancies in Non-STEM fields (1 0)	incen_gen_elem incen_spec_ed incen_english incen_soc_studies incen_esl incen_foreign_lang incen_music_art incen_voc_ed	incen_NonSTEM: Pay recruit incentives in non-STEM fields (1 0)
type_Alternative type_Elementary type_Regular type_Special type_Voc_Tech	sch_type: School type (5 categories)	vacnc_comp_sci vacnc_math vacnc_biology vacnc_phys_sci	vacnc_STEM: Difficulty of filling vacancies in STEM fields (1 0)
incen_certification incen_excellence incen_prof_dev incen_location	incen_pay: Pay incentives in salary (1 0)	incen_STEM_comp_sci incen_STEM_math incen_STEM_phys_sci incen_STEM_biology	incen_STEM: Pay recruit incentives in STEM fields (1 0)
urbanicity_LargeCity urbanicity_SmallTown urbanicity_MidCity	urbanicity: Urban locale (3 categories)		

an aggregation of the variables from *teaches_ 7th* to *teaches_ 12th*. Second, we combined all dummy-coded variables into a single attribute with multiple categories, e.g. *urbanicity_ LargeCity*, *urbanicity_ SmallTown*, and *urbanicity_ MidCity* became *urbanicity*, as summarized in Table 3. Finally, the dimensionality of the data set has been reduced to 53 attributes with 39 categorical and 14 numerical.

Maximum Relevance Feature Selection

As the dimension space of the attributes in our data set is reduced to the smaller number, 53, we compared the nine different approaches explained in Section III to notice their importance at a glance. Table 4 shows the number of attributes selected from each approach; RFE with ridge regression selected the smallest set of

Table 4: Nine feature selection approaches selected the number of features for Teacher Attrition. The selection is illustrated in Figure 4 with distinguished bar colors marked in the Color column.

Method	Approach	Features	Color
Filter	Variance Threshold	34	
Embedded	Lasso Regularization	38	
Embedded	Random Forests Feature Importance	27	
Wrapper	PMI - Random Forests	28	
Wrapper	PMI - Ridge Regression	33	
Wrapper	RFE - Ridge Regression	18	
Wrapper	RFE - Random Forests	49	
Wrapper	SFS - KNN	26	
Wrapper	SFS - Ridge Regression	26	

numbers, 18, while RFE with random forests produced the largest set, which is 49 attibutes, as shown in Figure 4. All nine approaches selected **four** attributes: *remain_ teaching* (teacher responded to the survey question on the likelihood of remaining in teaching), *public_ pt_ exp* (years of part-time teaching experience in public schools), *public_ ft_ exp* (years of full-time teaching experience in public schools) and *level_ Elementary* (level of school in teaching is elementary) as the most important attributes.

Figure 5 indicates the main attributes of random forests and random forests Permutation to predict teacher attrition. If we use a threshold of 0.011, *public_ ft_ exp* (years of full-time teaching experience in public schools), *remain_ teaching* (teacher responded to the survey question on how likely they will remain in teaching), *yrs_ tch_ before_ P* (years of teaching experience prior to becoming a principal), *num_ dependents* (number of dependent teachers), *age_ P* (age of a principal), *new_ teacher* (teachers who teach 3 years or less), *level_ Elementary* (teachers teaching in an elementary school), and *hrs_ taught_ STEM* (hours of teaching STEM subjects per week) are the only eight overlapping highly impactful attributes. Vanilla random forests has 27 features with an impact score greater than 0.011. Both methods select *public_ ft_ exp* as the most

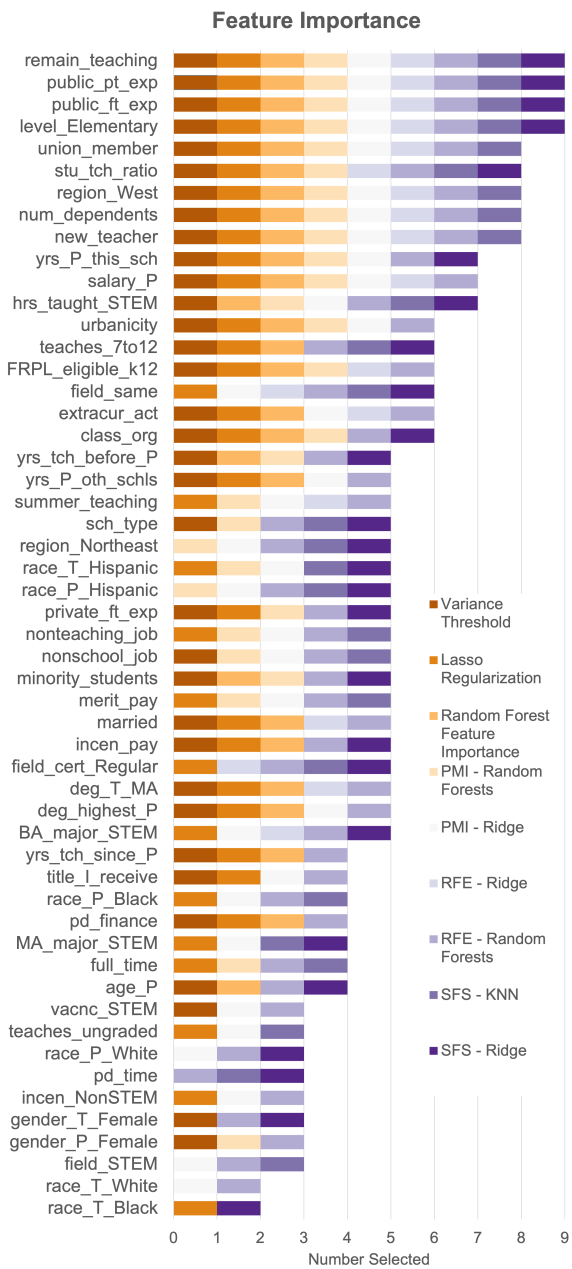


Figure 4: All nine methods select the 4 features *remain_teaching*, *public_pt_exp*, *public_ft_exp*, *level_Elementary* as the most important features.

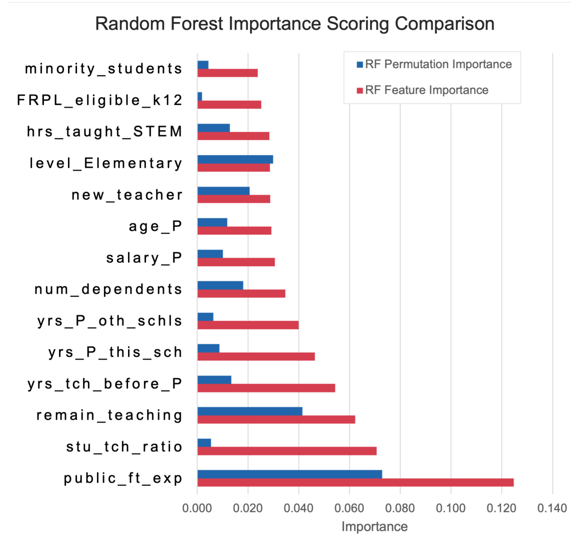


Figure 5: Random Forests Feature Importance and Permutation Attribute ranking comparison

significant characteristic, which is the years of full-time teaching experience in public schools. Specifically, since teachers work longer years as full-time teachers in public schools, we can better predict teacher retention.

In this section, we offer policy makers the opportunity to draw meaningful conclusions. We proposed an elegant and simple way to identify schools with critical attrition personnel in unlabeled data. We analyze and compare state-of-the-art machine learning models for teacher attrition rates. Our main goal was to help educational researchers and policy makers gain insight into data and attrition rates.

Model Evaluation and Comparison

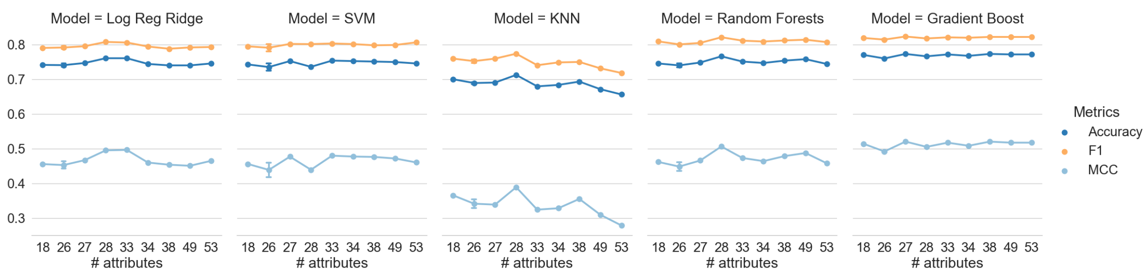


Figure 6: Five machine learning models fitted to the training and test sets with 10-fold cross-validation of GridSearch. Test set accuracy, F1, and MCC results show stable performance for all models except KNN.

Table 5: Best model of the five state-of-the-art machine learning models is gradient boosting training 27 features.

Model	Best Set	Selection Method	Accuracy [0,1]	F1 [0,1]	MCC [-1,+1]
Log Reg Ridge	28	PMI - Random Forests	0.761	0.808	0.496
SVM	33	PMI - Ridge	0.754	0.804	0.48
KNN	28	PMI - Random Forests	0.713	0.774	0.389
Random Forests	28	PMI - Random Forests	0.766	0.821	0.507
Gradient Boost	27	Random Forests Feature Importance	0.773	0.824	0.521

We used the labeled data set with 3,640 teachers from 2,838 schools: 53 attributes and labels of 2,176 current teachers and 1,464 former teachers. We randomly split the data into a training set (2,192 teacher instances, 80%) and a test set (728 teacher instances, 20%) with shuffling and stratification on the label. The feature reduction methods produced a different number of attributes: the full set contains 53 attributes, and 18, 26, 26, 27, 28, 33, 34, 38 attributes are selected by methods such as RFE and SFS with KNN and ridge regression, random forests feature importance, PMI with random forests and ridge regression, Variance threshold, and Lasso regularization, respectively. To evaluate each feature set selected using different dimensionality reduction techniques, the five state-of-the-art machine learning models have been built for the 10 different feature sets: ridge regression, SVM, KNN, random forests, and gradient boosting. Then, the same process was repeated for the advanced gradient boosting models, XGBoost, LightGBM, CatBoost, and HistGradientBoosting, to compare and select the best model on prediction of our label. The specification of the implementation of these models is described in Section III.

The performance of the five state-of-the-art models in the test set, organized by the number of attributes, is illustrated in Figure 6. In general, the metrics, accuracy, F1, and MCC, show steady performance across all models except KNN and feature sets. As the best performance of each model listed in the Table 5, decision-tree based ensemble models, gradient boosting and random forests training 27 and 28 features selected by random forests feature importance and PMI with random forests respectively, are the best performing model with the highest accuracy (77%), and F1 (82%).

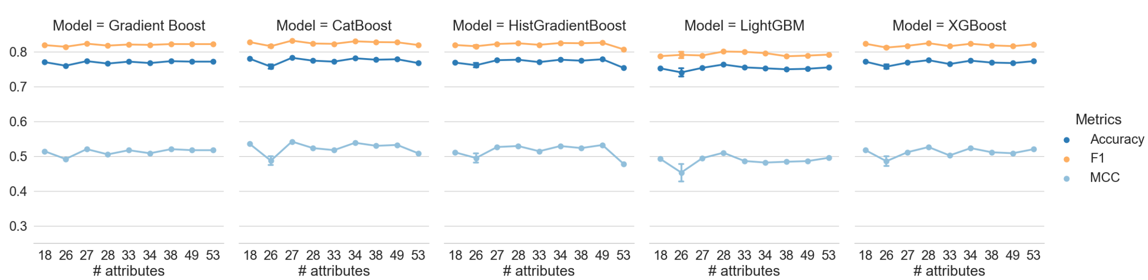


Figure 7: Four gradient boosting models fitted to the training and test sets with 5-fold cross-validation of RandomizedSearch. With accuracy, F1, and MCC for the test set compared with gradient boosting, the dimensionality reduction plays no role for the performance of the models.

Table 6: CatBoost fitting 27 features is the most robust model among Advanced gradient boosting models.

Model	Best Set	Selection Method	Accuracy [0,1]	F1 [0,1]	MCC [-1,+1]
CatBoost	27	Random Forests Feature Importance	0.783	0.832	0.543
HistGradientBoost	49	RFE - Random Forests	0.779	0.826	0.533
LightGBM	28	PMI - Random Forests	0.764	0.801	0.51
XGBoost	28	PMI - Random Forests	0.776	0.825	0.527

Next, we compared the four advanced gradient boosting models with our best performing baseline models, gradient boosting. While the boosting models remain stable for all sets of attributes regarding their test accuracy, F1, and MCC as shown in Figure 7, the most robust performing boosting model is CatBoost trained 27 features selected by random forests feature importance with the best accuracy (78%), F1 (83%), and MCC (54%) as summarized in Table 6. Furthermore, all four gradient boosting algorithms performance is similar to, and not exceeding, the vanilla gradient boost implementation as the difference of accuracy between them is equal to or less than 1%.

Overall, the reduction in dimensionality does not play a role in all nine machine learning models, and gradient boosting algorithms are performing slightly better than the other state-of-the-art models.

Teacher Retention Prediction and Analysis

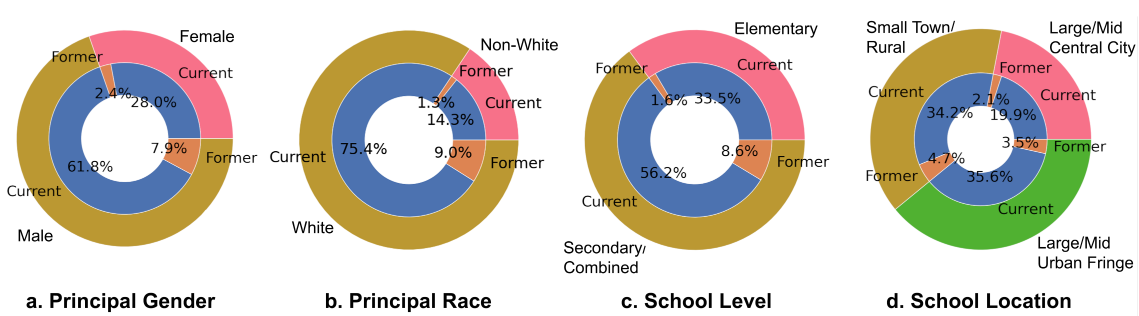


Figure 8: Teacher attrition prediction analysis for school and principal

We proposed using the best gradient boosting model to predict teacher attrition rates per school, and we demonstrated the results for teachers who did not participate in the follow-up survey [56]. The training set contained 2,176 current and 1,464 former teachers and an attrition rate of 40%, much higher than the average United States teacher attrition rate (∼8%). We account for this by tightening the model threshold for model prediction and assigning the label *likely to leave education* only if the confidence in model prediction is greater than 0.8.

The entire labeled data (3,640 teachers) then became a training set, and a test set contained 33,198 teachers (entries without principal and school associations were removed). The average number of teachers per 7,428 schools analyzed is 4.47, and only 358 schools have 10+ teachers participating in the SASS survey [56]. We could not produce the teacher attrition rate predictions per 7,428 schools analyzed as the data contains only categorical information on the total number of teachers per school (<24, 25-34, >34) [56]. This new dataset does not contain two attributes available only for the TFS data: marital status and the number of dependents. We fit the XGBoost model with the best hyperparameters, the best gradient boosting model for the Full feature set, with 51 features on the training dataset, and rank predictions on the test set. Our model predicts 3,399 teachers from the unlabeled SASS dataset have also left education (80%+ model confidence). The breakdown of predictions is in Figure 8: (a) female principals have less former teachers(2.4%) than male principals(7.9%); (b) Non-White principals have less Former teachers(1.3%) than the ones for White principals(9%); (c) Secondary or Combined schools have higher Former teacher ratio(8.6%) than Elementary school former teachers(1.6%), and (d) schools located in Small town or rural areas have the highest percentage of former teachers(4.7%) than schools in Large or Mid Urban Fringe (3.5%) following Large or Mid Central City (2.1%).

Full Feature vs. Raw Data for Gradient Boosting Models

Considering that the advanced gradient boosting models handle categorical features that overcome data sparsity and unbalanced tree splits caused by one hot encoding and high-cardinality categorical features, we experimented on categorical feature support by fitting and comparing the models with the full feature set and raw data. The full feature set represents 53 features, including 39 categorical and 14 numerical features that eliminated overlapping information with correlation filtering. As for the raw data, we defined the integrated SASS and TFS data accommodating 124 attributes with 107 categorical attributes and 17 numerical attributes. We do not concern missing values in this experiment, as the SASS and TFS data do not include missing values, as originally provided.

Table 7: Three data sets for experimenting on handling categorical features for XGBoost, LightGBM, CatBoost, and HistGradientBoosting.

Data Set	# of Attributes	# of Categorical	Use Categorical
Full Feature	53	39	Yes
Raw	124	107	Yes
Raw - One-Hot Encoded	596	0	No

We have prepared three data sets for the experiment: Full-Feature, Raw, and Raw One-Hot Encoded are explained in Table 7. The data sets trained in the same gradient boosting modeling process – XGBoost, LightGBM, CatBoost, and HistGradientBoosting – with the choice of providing categorical feature indices as a parameter. As illustrated in Figure 9, the feature engineering of the three data sets, does not significantly affect test accuracy of the four gradient boosting models as the accuracy of the models is steady for all data sets around 76%, except XGBoost, 61%. In fact, XGBoost, which supports categorical features with the recent version 1.6 experimentally, performs the worst with Raw data for F1 as 60% and MCC as 28%, while the same model does not fluctuate for the performance of the rest of the data sets. The key difference in this case is that Raw data have categorical attributes with high-cardinality, since 5 attributes have a large number of categories between 46 to 80; in contrast, Full feature and One-hot encoded Raw data sets have very few categories for an attribute, from 2 to 5.

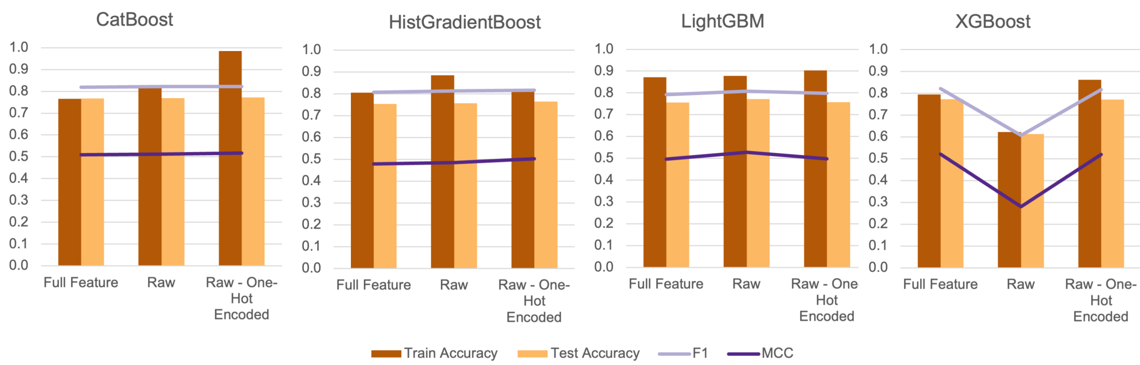


Figure 9: Three data sets fitted four advanced gradient boosting models. Using train and test set split, RandomizedSearch and 5-fold cross validation, feature engineering shows no effects on the performance of the four advanced grandient boosting models except for XGBoost.

V. Identifying Resilience Factors in Texas Public Schools

Texas Public Data Summary

Integrating Public Data

Table 8: Data from eight different sources are integrated by matching school district ID and county FIPS code. After initial integration, data have 1,165 schools districts with 506 attributes in 253 counties in Texas.

Data Abbr	Data	Source	Level	Shape
CCD, NCES	Common Core of Data	National Center for Education Statistics	District	(1189, 66)
STAAR, TEA	State of Texas Assessments of Academic Readiness for 2018-2019 and 2020-2021	Texas Education Agency	District	(1184, 217) (1182, 217)
LAUS, BLS	Local Area Unemployment Statistics	U.S. Bureau of Labor Statistics	County	(254, 13)
Census Bureau	Census Block Group 2010	Census Bureau	County	(254, 37)
Covid, DSHS	Texas Public Schools COVID-19 Data	Texas Department of State Health Services	District	(1216, 7)
Covid, USAFacts	Texas Coronavirus Cases and Deaths	USAFacts	County	(254, 8)
ADA, TEA	Average Daily Attendance	Texas Education Agency	District	(1226, 3)
ESSER, TEA	Elementary and Secondary School Emergency Relief	Texas Education Agency	District	(1208, 6)

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 8 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) [58] 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 [59] 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) [60], 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 [61] due to Coronavirus at each county from USAFacts. Also, the average daily attendance (ADA) [62], 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 [63] – 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 [64] 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 [65] 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.

Labeling into Multi-Class Classification Problem

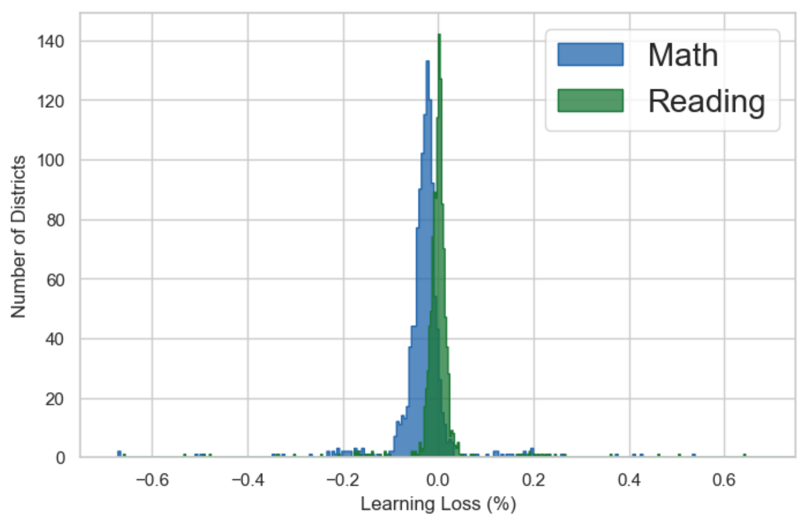


Figure 10: Distribution of learning loss values for math and reading. It shows that more students and school districts experienced learning loss in math and reading subject.

Our data set is unlabeled data; thus we need to create 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 24 attributes indicating the scores for calculating learning loss. We first normalized each cell of the scores by the

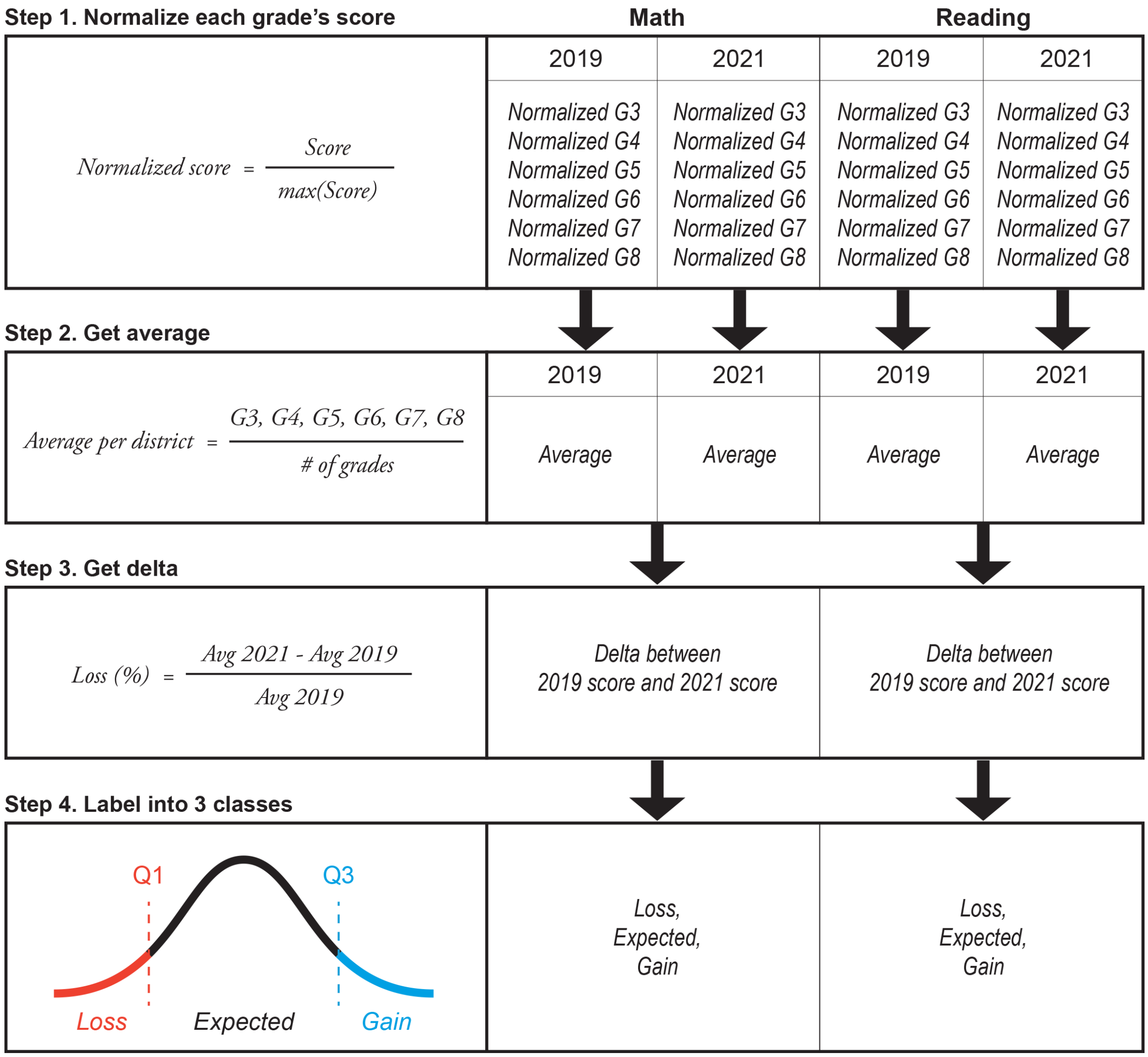
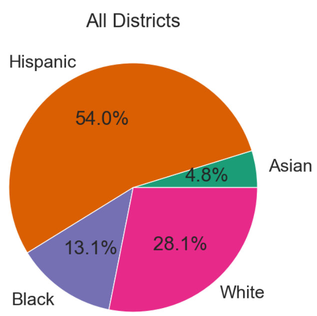
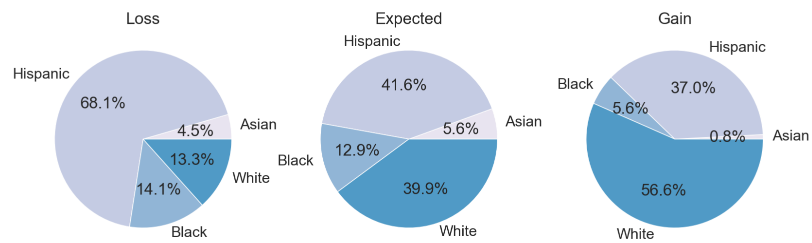


Figure 11: Four steps to label into 3 classes by calculating learning loss using the STAAR scores. First, normalizing each score by dividing by the maximum score, then, getting averages of the normalized scores, and getting delta of the average normalized scores between 2021 and 2019, finally labeling the middle 50% of distribution as "Expected", below Q1 as "Loss", and above Q3 as "Gain".

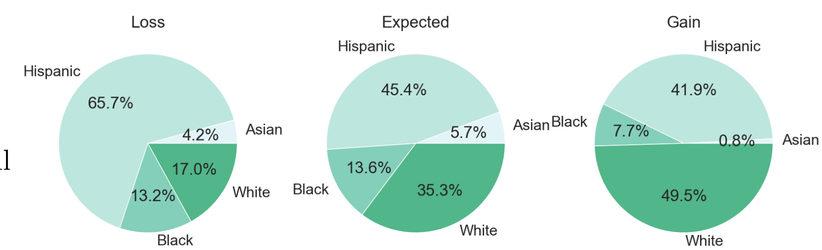
maximum score value of the attribute as described in Figure 11, Step 1. The following Step 2 averaged these normalized scores for each year and subject, and Step 3 calculated the loss as the difference of the scores, between 2019 and 2021 for the perspective of 2019. Consequently, our label – learning loss – is decided depending on the the loss value: if it is positive, there is learning gain, but a negative value corresponds to learning loss. At this point, we plot a distribution of the loss values in Figure 10 to set a threshold determining the loss and gain. The distribution shows that more districts have experienced the loss in math as the median for math (-0.03) is lower than the median for reading (0). We decided to proceed with further analysis and prediction separately for math and reading. Therefore, Step 4 in Figure 11 describes creating 3 label classes; the middle 50% of school districts is labeled as "Expected", and the loss values below 25th percentile are set to be "Loss", and the loss values above 75th percentile become "Gain".



(a) Race/Ethnicity - All Districts



(b) Race/Ethnicity - Math

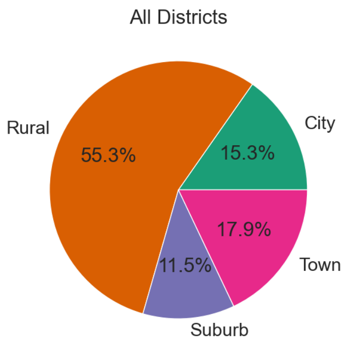


(c) Race/Ethnicity - Reading

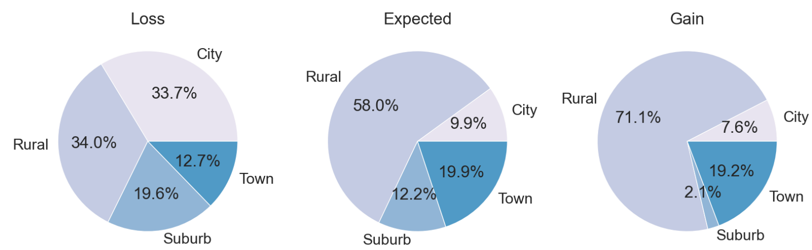
Figure 12: Exploratory data analysis for race/ethnicity of students. (a) the distribution includes four race/ethnicity groups of students, Hispanic, White, Black, and Asian; (b) and (c) show Hispanic and White students correlated to the label as the Hispanic population declines from Loss to Gain in contrast to White population increases in the same direction for both math and reading.

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 12 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

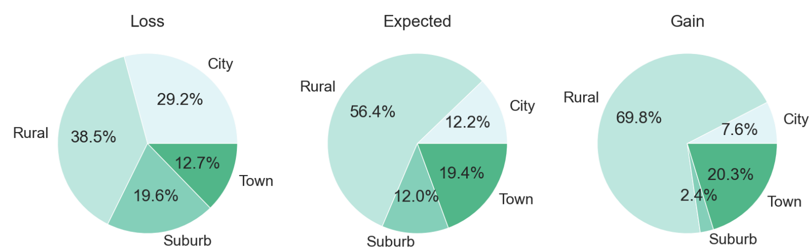
Gain labels for both math and reading. Also, we realized that the locale of school districts is correlated to the label learning loss, as illustrated in Figure 13. Figure 13 (a) confirms that over half the schools are located in rural areas in Texas despite the positive correlation between rural areas and the label from Loss to Gain; however, Loss occurring in schools located in City and Suburb areas increasingly appeared in (b) and (c).



(a) Locale - All Districts



(b) Locale - Math

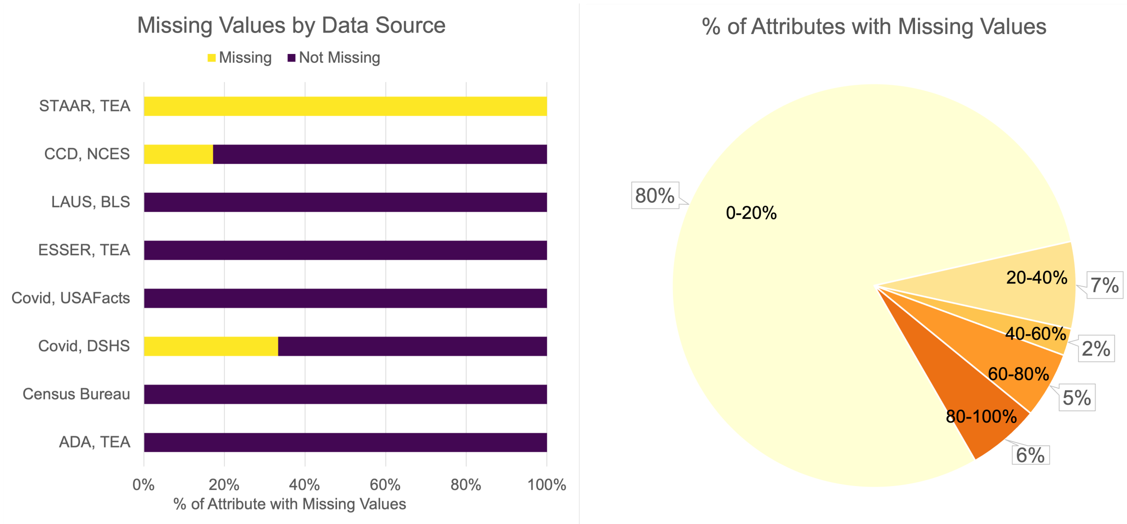


(c) Locale - Reading

Figure 13: Exploratory data analysis for locale of schools: (a) shows that the majority of schools in Texas are located in Rural; (b) and (c) confirm that schools located in City and Suburb experienced severe Loss compared to schools in Rural area growing from Loss to Gain.

Attribute Selection and Importance

Since our data set contains 506 attributes for 1,165 school districts, in this section, we engage in dimensionality reduction to obtain interpretability and identify the resilience factors for learning loss. We first remove noise, missing values, from the data, and then aggregate attributes conveying the same information for each year of 2019 and 2021. In turn, we successfully reduced the number of attributes to 90 to finally adopt the attribute selection methods explained in Section III.



(a) Missing Values by Data Source (b) % of Attributes with Missing Values

Figure 14: 416 attributes out of 506 of the initial data set have missing values from 1 to 88%. 80% of the attributes with missing values has missing values less than 20%.

Handling Missing Values. 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 in Figure 14 (a). Of these 416 attributes, 80% have fewer than 20% missing values, but 11% of them have more than 60% missing values, as shown in Figure 14 (b). While these missing values are predominantly from the STAAR data, related to average scores and participants in the STAAR tests, the rest of the 8 attributes of the CCD and COVID data contain less than 9% missing values. Thus, we decided to simply remove those attributes from the STAAR data and impute the attributes related to the number of students in each grade to 0. Furthermore, dropping rows with missing values from the CCD and COVID data resulted in 119 attributes for 955 school districts.

Aggregating attributes. With the number of attributes reduced to 119 after

Table 9: Examples of aggregating attributes containing the same information for each year 2021 and 2019. By calculating the delta between 2021 and 2019 for the perspective of 2019, 68 of these attributes are reduced by half, 34 attributes. Full list is in Table B.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

handling missing values, we still have 68 attributes containing the same characteristics for each year in 2019 and 2021. To improve readability and interpretability, these attributes were aggregated into a single attribute by obtaining delta values between 2019 and 2021 for the 2019 perspective. For example, the attributes *Total Schools 2020-2021* and *Total Schools 2018-2019* are aggregated into *Total Schools Diff* as provided in the example aggregation list in Table 9. As a result, the dimension is reduced to 90 from 119.

Maximum Relevance Feature Selection

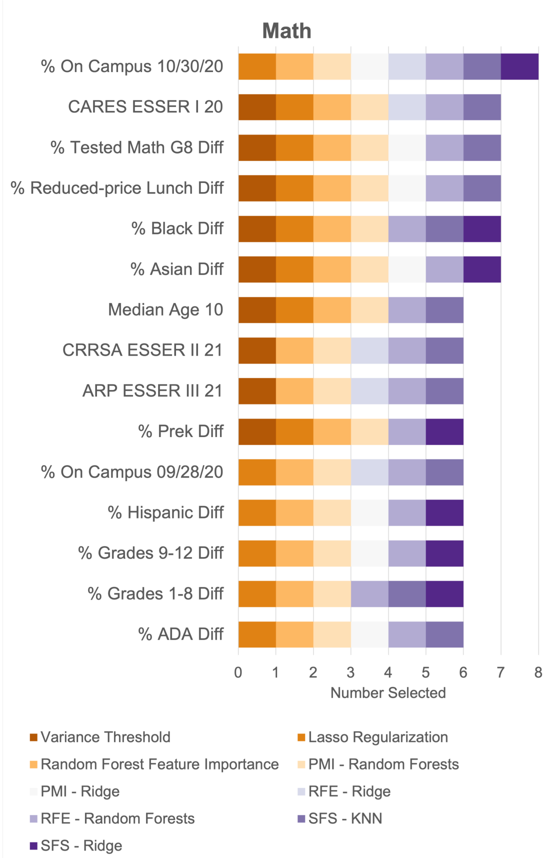
We executed the nine different feature selection approaches described in Section III to detect the resilient factors for the loss of learning due to COVID-10 using the data set with 90 attributes and 955 school districts in Texas. As we discriminate the subjects, math and reading, on predicting learning loss, the feature

Table 10: Nine feature selection approaches selected the number of features for learning loss. The selection is illustrated in Figure 15 with distinguished bar colors marked in the Color column.

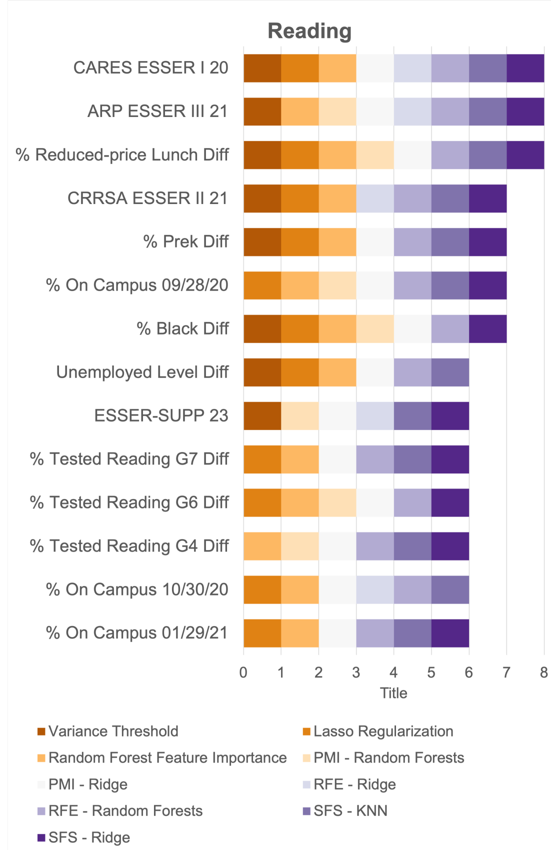
Method	Approach	Features Math	Features Reading	Color
Filter	Variance Threshold	21	20	
Embedded	Lasso Regularization	55	51	
Embedded	Random Forests Feature Importance	45	45	
Wrapper	PMI - Random Forests	70	26	
Wrapper	PMI - Ridge Regression	28	82	
Wrapper	RFE - Ridge Regression	6	5	
Wrapper	RFE - Random Forests	36	36	
Wrapper	SFS - KNN	45	45	
Wrapper	SFS - Ridge Regression	45	45	

selection process has been repeated twice for each subject separately.

Table 10 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, PMI method selected the largest 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 15, (a) Top 15 for math, and (b) Top 14 for reading selected by 6 or more feature selection methods. Note that full selection results are listed in Table B.2. 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, 3 important features are selected, all of which were resilience factors related to 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 6 or more selection methods in Figure 15, we analyzed the



(a) Math - Top 15 features



(b) Reading - Top 14 features

Figure 15: The Most important features for predicting learning loss in math and reading selected by 6 or more feature selection methods. *% On Campus 10/30/20* for Math in (a), *CARES ESSER I 20, ARP ESSER III 21, % Reduced-price Lunch Diff* for Reading in (b). Full list is in Table B.2

resilient factors for seeking the most impactful factor among them. Apparently, Low-income and Grade level are the most influential resilient factors to predict learning loss for both math and reading, as shown in Figure 11. Race/Ethnicity and mode of instruction continued to be powerful resilient factors for both subjects; on the other hand, Attendance and Census demographics are considered as significant factors only by math, and Unemployment is important only for reading.

Although we now realize these important features can identify the resilient factors for Loss or Gain in learning due to COVID-19 pandemic, it is still unknown whether those features are positively impacting the learning or not. As an example,

Table 11: Resilient factors for Top 15 (math) and 14 features (reading). Top 15 and 14 features for math and reading selected by 6 or more dimensionality reduction approaches in Figure 15. 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

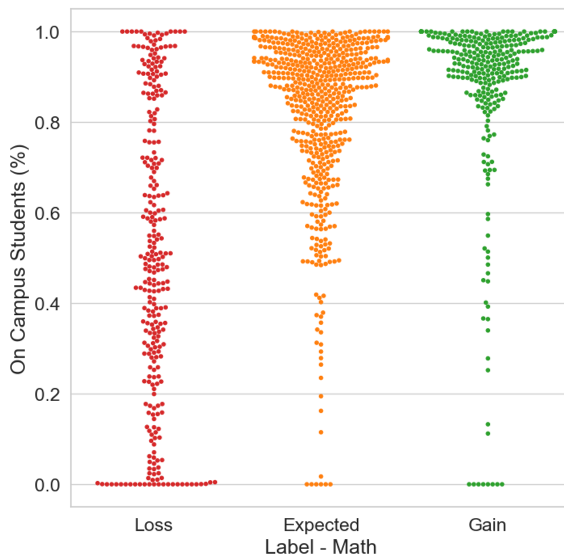


Figure 16: Analysis on the most important feature for predicting learning loss in math: *% On Campus 10/30/20*. School districts in Gain and Expected label have more students went to school on October 30, 2020.

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

Figure 16 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 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, was the key to avoid Loss in math.

The most important features for reading, *CARES ESSER I 20, ARP ESSER III 21*, provide a perspective on the federal funding. The features are part of the Elementary and Secondary School Emergency Relief (ESSER) grant programs which are federal funds granted to State education agencies (SEAs) providing Local education agencies (LEAs) to address the impact due to COVID-19 on elementary and secondary schools across the nation; thus, the funds have been administered by Texas Education Agency (TEA) and allocated in each school district in Texas [63, 66]. The ESSER funds have four programs as below:

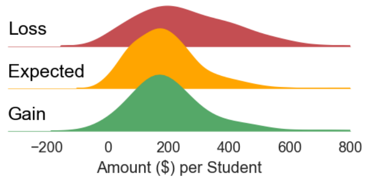
CARES ESSER I: Authorized on March 27, 2020, as the Coronavirus Aid Relief, and Economic Security (CARES) Act with $13.2 billion. Period of availability is March 13, 2020 to September 30, 2022. Our data have the allocation amount for the fiscal year of 2020.

CRRSA ESSER II: Authorized on December 27, 2020, as the Coronavirus Response and Relief Supplemental Appropriations (CRRSA) Act with $54.3 billion. Period of availability is March 13, 2020 to September 30, 2023. Our data have the allocation amount for the fiscal year of 2021.

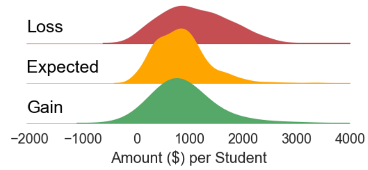
ARP ESSER III: Authorized on March 11, 2021, as the American Rescue Plan (ARP) Act with $122 billion. Period of availability is March 13, 2020 to September 30, 2024. Our data have the allocation amount for the fiscal year of 2021.

ESSER-SUPP: Authorized by the Texas Legislature to provide additional resources for unreimbursed costs to support students not performing well educationally. Period of availability is March 13, 2020 to August 31, 2023. Our data have the allocation amount for the fiscal year of 2022 and 2023.

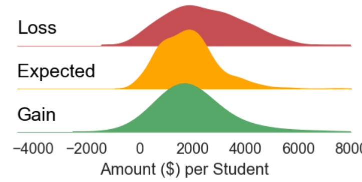
As shown in Figure 17 displaying the distribution of each fund amount converted to the amount per student, the students experienced Loss in reading received larger amount of funding for all funding programs on average than the



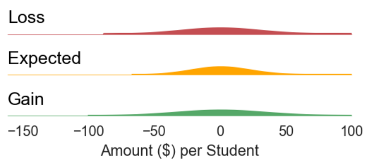
(a) CARES ESSER I (2020)



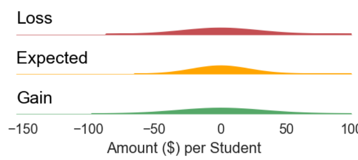
(b) CRRSA ESSER II (2021)



(c) ARP ESSER III (2021)



(d) ESSER-SUPP (2022)



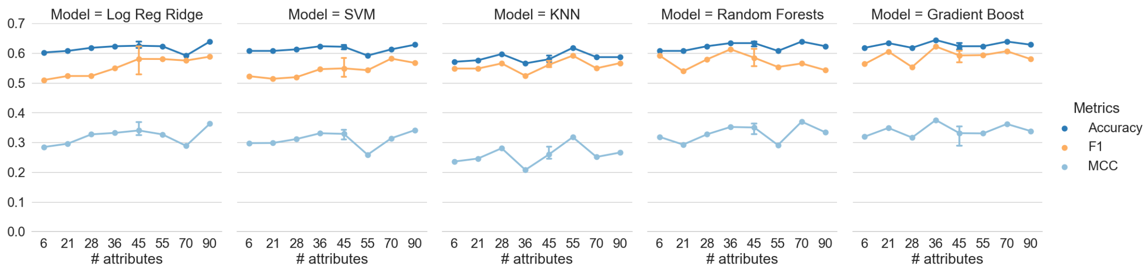
(e) ESSER-SUPP (2023)

Figure 17: Analysis on the most important feature for predicting learning loss in reading: *CARES ESSER I 20, ARP ESSER III 21*. Five ESSER funding programs allocation for school district per student indicate that school districts allocated higher amount to the districts accommodating more students in Loss.

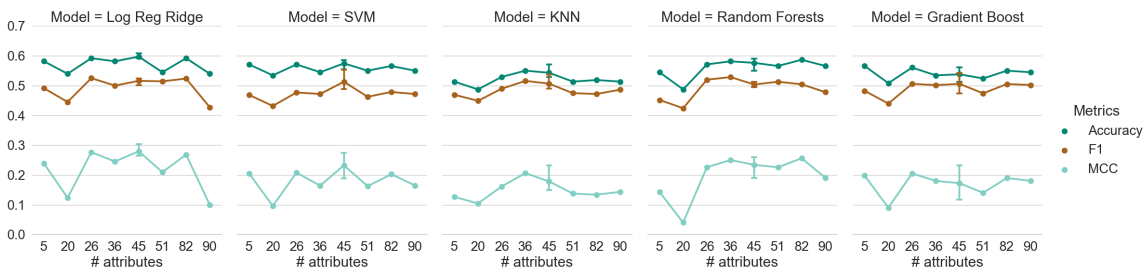
students experienced Gain or Expected in the same subject. Meaning that the ESSER amounts have been distributed to proper districts in need of financial help for adapting and preparing learning loss due to COVID-19 as the ESSER fund amounts are calculated by formula based on the Title I, Part A grant that is considered as a poverty proxy [63, 66].

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



(a) Accuracy, F1, and MCC for Math



(b) Accuracy, F1, and MCC for Reading

Figure 18: Five state-of-the-art models fitted to 10 different feature sets for predicting learning loss. With 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 performs similarly.

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 Figure 15: 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 III.

Table 12: (a) CatBoost fitting 36 features for math and (b) CatBoost fitting 82 are the most robust models predicting learning loss. These are resulted from the nine machine learning models training the 10 features sets.

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

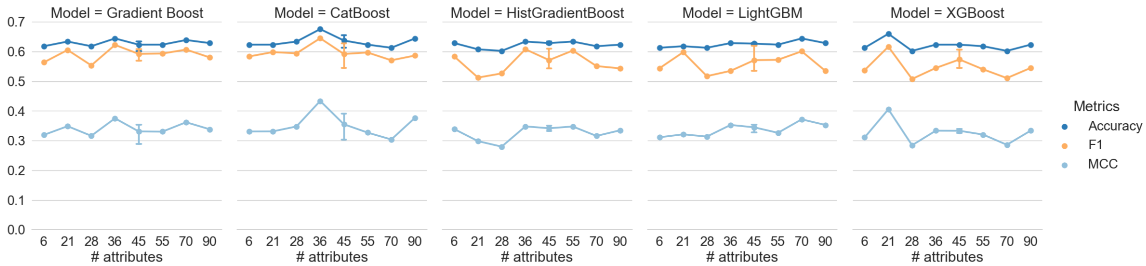
(a) Math

Model	Best Set	Selection Method	Accuracy [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

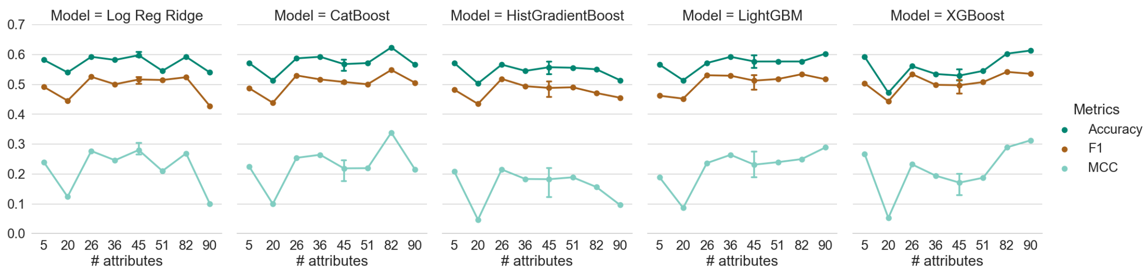
(a) Reading

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 18 (a) for math and (b) for reading; predicting learning loss for reading shows weak performance compared to math generally. While no clear 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%)



(a) Accuracy, F1, and MCC for Math



(b) Accuracy, F1, and MCC for Reading

Figure 19: 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).

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 19, (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.

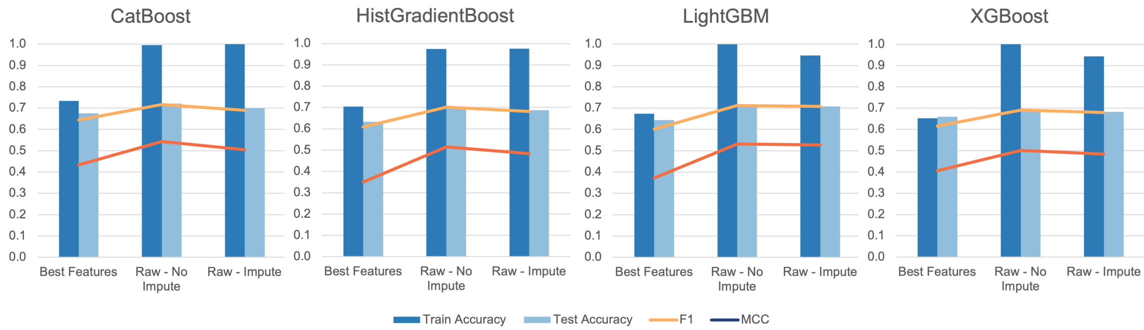
For all nine models, the best feature set for each model is described in Table 12 (a) for math and (b) for reading; both subjects suggest CatBoost as the most robust models: 36 features selected by RFE with random forests with accuracy (68%), F1 (65%) and MCC (43%) for math and 82 features selected by PMI with ridge regression with accuracy (62%), F1 (55%) and MCC (34%) for reading.

Overall, the gradient boosting algorithms, CatBoost and XGBoost, are the best choice of all machine learning models we have experimented to predict learning loss for both subjects. While these models performed better for predicting the loss 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 accuracy difference is around 3%. Also, no clear indication emerged when it comes to the best dimensionality reduction technique that performs across all models.

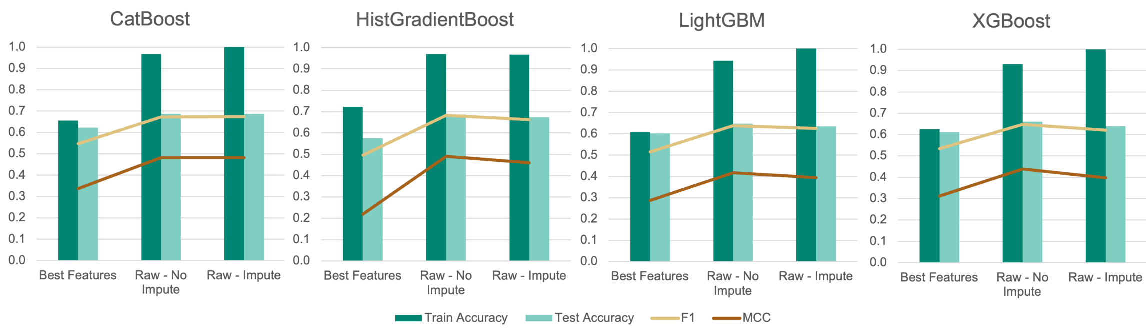
Best Features vs. Raw Data for Gradient Boosting Models

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 1 categorical) for 1,165 school districts, includes 416 attributes with missing values as small as 1% and as large as 88% of each attribute as shown in Figure 14. In this experiment, we executed the pipeline of building the advanced gradient boosting models for Raw data and compared with the models trained the data processed the feature engineering techniques in terms of prediction power on learning loss. The classification task completed for the respective subjects, math and reading. All attributes with missing values except for 8 attributes 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 attributes for this experiment for each subject. 212 out of 302 attributes contain missing values.

We have three data sets for comparison: (1) The best sets of features in Table 12 from the four gradient boosting models performance result in Figure 19, (2) Raw data with no imputation for missing values, and (3) Raw data imputed missing values with mean values. Note that our data has only 1 categorical attribute



(a) Train & Test Accuracy, MCC for Math



(b) Train & Test Accuracy, MCC for Reading

Figure 20: 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 appeared in Table 12.

including no missing values, so the imputation method is limited to average. As for 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 20; HistGradientBoost increased MCC the most as 47% following LightGBM (43%), CatBoost (25%) and XGBoost (24%) for math, and the improved MCC for reading is even steeper with 124% for HistGradientBoost and 45%, 43%, and 41% for LightGBM, CatBoost, and XGBoost, respectively. For a closer look, we also observed that Raw data set without imputation perform slightly better compared to Raw data set with imputation for all models and subjects; MCC for math rose the most, over 6%, in CatBoost and HistGradientBoost; in contrast, XGBoost showed the largest growing for MCC in reading with 10%.

VI. Conclusion

Our intentional data science pipeline can automatically uncover important attributes using public-use data and the nine different feature selection methods to model two projects in this paper: teacher attrition and learning loss. While reduction in dimensionality of data plays no role for 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 for both of projects. In fact, the gradient boosting models such as XGBoost and CatBoost are superior for handling categorical features and missing values as we experimented with raw data for each project; 107 out of 124 attributes are categorical for teacher attrition data set, and over 2/3 of attributes of learning loss data set contain missing values. Reproducible experiments and datasets are published on [67].

Implications and Future Work

Policy makers can use our predictive models and analysis to focus resources on the public school system including schools, students, and teachers to keep teachers in public schools, and mitigate learning loss with possible interventions.

Teacher Attrition. Utilizing public-use SASS and TFS data in other years, 1987-1988, 1990-1991, 1993-1994, as test sets can determine the best performing model among the current models as teacher samples are drawn from the different years and population. In addition, we can build prediction models with restricted-use data such as SASS and TFS in 2003-2004, 2007-2008, and 2011-2012 [57], or National Teacher and Principal Survey (NTPS) data, redesigned survey of SASS conducted in 2015-2016 and 2017-2018 [68], to discover unknown important factors for teacher attrition as those data contain unseen attributes and up-to-date information. After affirming the current pipeline by experimenting with the new data, we can apply the pipeline to segmented problems, for example, STEM teacher retention.

Learning Loss. To alleviate learning loss, the findings of the project can shape interventions in terms of policy planning. First of all, the most prominent resilient factor of learning loss identified by both subjects, math and reading, is Low-income background of students; so, Low-income factor can be considered for policy making as in the example of the ESSER programs. Grade level – the important resilient factor – and the different distributions of learning loss in each subject in Figure 10 suggest that personalize interventions with grade-specific and subject-specific efforts might be required. Moreover, policy makers can also contemplate specific and cross-sectional factors for the personalized interventions; for instance, Hispanic and City resident students are the largest population suffering from learning loss as appeared in the EDA on Race/Ethnicity of students and Locale of schools in Figure 12 and 13. In the future, the key factors recognized and suggested in this paper will be examined and tracked by running the current pipeline for the up-to-date STARR 2021-2022 scores as learning recovery label and analyzing whether these factors contribute to achieving the recovery.

Gradient Boosting Modeling. We confirmed that the gradient boosting models such as CatBoost, XGBoost, and LightGBM offer enhanced quality of classification when training raw data without preprocessing categorical features or missing values. However, the models are "black-Box" as it is challenging to interpret feature importance. Future work could include explaining the gradient boosting models fitting raw data using the tree explainers such as SHAP and LIME, so that we can obtain global feature importance for all districts in Texas and local feature importance for each district or county level. These results will be compared with the current feature selection methods to evaluate the key factors.

APPENDIX SECTION

APPENDIX A

Table A.1: Selected Teacher, Principal, and School Attributes in the SASS dataset. Value (1 0): If the statement is true, the attribute value is 1, otherwise it is 0.

Teacher Label	Description	**Teacher Label**	Description
num_dependents	Number of dependents of teachers	deg_T_MA	Master's degree (1 0)
married	Married teacher (1 0)	pd_time	Professional development time off(1 0)
race_T_White	Teacher's race (1 White 0 Others)	pd_finance	Professional development pay (1 0)
race_T_Black	Teacher's race (1 Black 0 Others)	remain_teaching	Likely to remain in teaching (5-pt scale)
race_T_Hispanic	Teacher's Ethnicity (1 Hispanic 0 Others)	field_STEM	STEM is main teaching job (1 0)
gender_T_Female	Teacher's gender (1 F 0 M)	hrs_taught_STEM	Hours of teaching STEM subjects per week
summer_teaching	Teaching summer school (1 0)	public_ft_exp	Years of full-time teaching in public schools
nonteaching_job	Has a nonteaching summer job (1 0)	public_pt_exp	Years of part-time teaching in public schools
nonschool_job	Has a nonschool summer job (1 0)	private_ft_exp	Years of full-time teaching in private schools
extracur_act	Extracurricular Pay(1-T 0-F)	field_same	Same teaching field as 1yo (1 0)
merit_pay	Income from merit pay (1 0)	full_time	Teaching full-time (1 0)
union_member	Union member (1 0)	teaches_7to12	Teaching 7 to 12th grades (1 0)
BA_major_STEM	STEM major for BA (1 0)	new_teacher	Teaching 3 years or less (1 0)
MA_major_STEM	STEM major for MA (1 0)	stu_tch_ratio	Student-Teacher ratio
field_cert_Regular	Certificate type (1 Regular 0 Others)		
Principal Label	Description	**School Label**	Description
age_P	Age of principal	vacnc_STEM	Difficulty of filling vacancies in STEM fields (1 0)
salary_P	Annual salary of principal	region_Northeast	School Location (1 Northeast 0 Others)
yrs_P_this_sch	Years at current job	region_West	School Location (1 West 0 Others)
yrs_P_oth_schls	Years as principal elsewhere	minority_students	Minority students percent
yrs_tch_before_P	Years teaching prior to principal	FRPL_eligible_k12	Free or reduced-price lunch eligible students percent
yrs_tch_since_P	Years teaching since principal	sch_type	School type (5 categories)
deg_highest_P	Principal's highest degree (5 categories)	level_Elementary	School level (1 Elementary 2 Others)
race_P_Black	Principal's race/Ethnicity (1 Black 0 Others)	urbanicity	Urbanic locale (3 categories)
race_P_White	Principal's race/Ethnicity (1 White 0 Others)	title_I_receive	Students receive Title I (1 0)
race_P_Hispanic	Principal's race/Ethnicity (1 Hispanic 0 Others)	incen_pay	Pay incentives on salary (1 0)
gender_P_Female	Principal's gender (1 F 0 M)	incen_NonSTEM	Pay recruit incentives on non-STEM fields (1 0)

APPENDIX B

Table B.1: Aggregated attributes containing the same information for each year 2021 and 2019. By calculating the delta between 2021 and 2019 for the perspective of 2019, 68 of these attributes are reduced by half, 34 attributes.

Attribute	Aggregated Attribute	Data
Total Students 2020-2021 Total Students 2018-2019	Total Students Diff	CCD, NCES
Total Schools 2020-2021 Total Schools 2018-2019	Total Schools Diff	CCD, NCES
Teachers:Students 2020-2021 Teachers:Students 2018-2019	Teachers:Students Diff	CCD, NCES
Staff:Students 2020-2021 Staff:Students 2018-2019	Staff:Students Diff	CCD, NCES
% White 2020-2021 % White 2018-2019	% White Diff	CCD, NCES
% Operational Schools 2020-2021 % Operational Schools 2018-2019	% Operational Schools Diff	CCD, NCES
% Title 1 Eligible 2020-2021 % Title 1 Eligible 2018-2019	% Title 1 Eligible Diff	CCD, NCES
% Title 1 School-wide 2020-2021 % Title 1 School-wide 2018-2019	% Title 1 School-wide Diff	CCD, NCES
% Reduced-price Lunch 2020-2021 % Reduced-price Lunch 2018-2019	% Reduced-price Lunch Diff	CCD, NCES
% Prek 2020-2021 % Prek 2018-2019	% Prek Diff	CCD, NCES
% Kinder 2020-2021 % Kinder 2018-2019	% Kinder Diff	CCD, NCES
% Hispanic 2020-2021 % Hispanic 2018-2019	% Hispanic Diff	CCD, NCES
% Grades 9-12 2020-2021 % Grades 9-12 2018-2019	% Grades 9-12 Diff	CCD, NCES
% Grades 1-8 2020-2021	% Grades 1-8 Diff	CCD, NCES

% Grades 1-8 2018-2019		
% Free Lunch 2020-2021 % Free Lunch 2018-2019	% Free Lunch Diff	CCD, NCES
% Black 2020-2021 % Black 2018-2019	% Black Diff	CCD, NCES
% Asian 2020-2021 % Asian 2018-2019	% Asian Diff	CCD, NCES
% Tested Reading G8 2020-2021 % Tested Reading G8 2018-2019	% Tested Reading G8 Diff	STAAR, TEA
% Tested Reading G7 2020-2021 % Tested Reading G7 2018-2019	% Tested Reading G7 Diff	STAAR, TEA
% Tested Reading G6 2020-2021 % Tested Reading G6 2018-2019	% Tested Reading G6 Diff	STAAR, TEA
% Tested Reading G5 2020-2021 % Tested Reading G5 2018-2019	% Tested Reading G5 Diff	STAAR, TEA
% Tested Reading G4 2020-2021 % Tested Reading G4 2018-2019	% Tested Reading G4 Diff	STAAR, TEA
% Tested Reading G3 2020-2021 % Tested Reading G3 2018-2019	% Tested Reading G3 Diff	STAAR, TEA
% Tested Math G8 2020-2021 % Tested Math G8 2018-2019	% Tested Math G8 Diff	STAAR, TEA
% Tested Math G7 2020-2021 % Tested Math G7 2018-2019	% Tested Math G7 Diff	STAAR, TEA
% Tested Math G6 2020-2021 % Tested Math G6 2018-2019	% Tested Math G6 Diff	STAAR, TEA
% Tested Math G5 2020-2021 % Tested Math G5 2018-2019	% Tested Math G5 Diff	STAAR, TEA
% Tested Math G4 2020-2021 % Tested Math G4 2018-2019	% Tested Math G4 Diff	STAAR, TEA
% Tested Math G3 2020-2021 % Tested Math G3 2018-2019	% Tested Math G3 Diff	STAAR, TEA
Unemployed Rate 2021 Unemployed Rate 2019	Unemployed Rate Diff	LAUS, BLS
Unemployed Level 2021	Unemployed Level Diff	LAUS, BLS

Unemployed Level 2019		
Labor Force 2021 Labor Force 2019	Labor Force Diff	LAUS, BLS
Employed 2021 Employed 2019	Employed Diff	LAUS, BLS
% ADA 2020-2021 % ADA 2018-2019	% ADA Diff	ADA, TEA

Table B.2: Full list of the dimensionality reduction approaches for predicting learning loss for math and reading. It is organized by the number selected by the approaches for math.

Feature	Selected Math	Selected Reading	Data Source	Resilient Factor	Description
% On Campus 10/30/20	8	6	Covid, DSHS	District oncampus enrollmen of students (%) on 10/30/20	Mode of instruction
CARES ESSER I 20	7	8	ESSER, TEA	CARES ESSER I amount($) (2020)	Lowincome
% Asian Diff	7	5	CCD, NCES	Hispanic students difference (%) (20212019)	Race/ethnicity
% Black Diff	7	7	CCD, NCES	Black students difference (%) (20212019)	Race/ethnicity
% Reducedprice Lunch Diff	7	8	CCD, NCES	Reducedprice lunch eligible students difference (%) (20212019)	Lowincome
% Tested Math G8 Diff	7	4	STAAR, TEA	Grade 8 students tested math exam difference (20212019)	Grade level
% ADA Diff	6	4	ADA, TEA	Average daily attendance difference (%) (20212019)	Attendance
Median Age 10	6	4	Census Bureau	Median age of county population (2010)	Census demographics
% On Campus 09/28/20	6	7	Covid, DSHS	District oncampus enrollmen of students (%) on 09/28/20	Mode of instruction
ARP ESSER III 21	6	8	ESSER, TEA	ARP ESSER I amount($) (2021)	Lowincome
CRRSA ESSER II 21	6	7	ESSER, TEA	CRRSA ESSER II amount($) (2021)	Lowincome
% Grades 18 Diff	6	5	CCD, NCES	Grades 18 students difference (%) (2021-2019)	Grade level
% Grades 912 Diff	6	5	CCD, NCES	Grades 912 students difference (%) (2021-2019)	Grade level
% Hispanic Diff	6	4	CCD, NCES	Hispanic students difference (%) (20212019)	Race/ethnicity
% Prek Diff	6	7	CCD, NCES	Prek students difference (%) (20212019)	Grade level

# of Households 10	5	4	Census Bureau	Number of households (2010)	Census demographics
% HH MarriednoChild 10	5	4	Census Bureau	Married couple households with no children (%) (2010)	Census demographics
Avg Family Size 10	5	4	Census Bureau	Average family size (2010)	Census demographics
Avg Household Size 10	5	5	Census Bureau	Average household size (2010)	Census demographics
Median Age Female 10	5	5	Census Bureau	Median age of female population (2010)	Census demographics
Median Age Male 10	5	4	Census Bureau	Median age of male population (2010)	Census demographics
% County Infected 09/28/20	5	4	Covid, USAFacts	Covid infection cases (%) on 09/28/20	COVID cases
% On Campus 01/29/21	5	6	Covid, DSHS	District oncampus enrollmen of students (%) on 01/29/21	Mode of instruction
ESSERSUPP 22	5	4	ESSER, TEA	ESSERSUPP amount($) (2022)	Lowincome
Unemployed Level Diff	5	6	LAUS, BLS	Unemployment level difference (20212019)	Unemployment
% Kinder Diff	5	5	CCD, NCES	Kindergarten students difference (%) (2021-2019)	Grade level
% White Diff	5	5	CCD, NCES	White students difference (%) (20212019)	Race/ethnicity
% Tested Math G3 Diff	5	4	STAAR, TEA	Grade 3 students tested math exam difference (20212019)	Grade level
% Tested Math G5 Diff	5	4	STAAR, TEA	Grade 5 students tested math exam difference (20212019)	Grade level
% Tested Math G6 Diff	5	6	STAAR, TEA	Grade 6 students tested math exam difference (20212019)	Grade level
% Tested Math G7 Diff	5	6	STAAR, TEA	Grade 7 students tested math exam difference (20212019)	Grade level
Locale_42Rural: Distant	5	3	CCD, NCES	School locale Rural: Distant	School Information
# of Families 10	4	4	Census Bureau	Number of families (2010)	Census demographics
# of Housing Units 10	4	3	Census Bureau	Number of housing units (2010)	Census demographics
% Age 1519 Pop 10	4	3	Census Bureau	Age 1519 Population (%) (2010)	Census demographics
% Age 2534 Pop 10	4	2	Census Bureau	Age 2534 Population (%) (2010)	Census demographics
% Age 59 Pop 10	4	2	Census Bureau	Age 59 Population (%) (2010)	Census demographics
% Age 7584 Pop 10	4	1	Census Bureau	Age 7584 Population (%) (2010)	Census demographics
% Asian Pop 10	4	4	Census Bureau	Asian Population (%) (2010)	Census demographics

% Black Pop 10	4	4	Census Bureau	Black Population (%) (2010)	Census demographics
% HH FemaleChild 10	4	2	Census Bureau	Femaleheaded households with children (%) (2010)	Census demographics
% Housing Renter Occup 10	4	3	Census Bureau	Renter Occupied Housing (%) (2010)	Census demographics
% County Infected 10/30/20	4	3	Covid, USAFacts	Covid infection cases (%) on 10/30/20	COVID cases
County Population	4	4	Covid, USAFacts	County Population	Census demographics
Labor Force Diff	4	5	LAUS, BLS	Labor force difference (20212019)	Unemployment
Staff:Students Diff	4	4	CCD, NCES	Staffs and students ratio difference (2021-2019)	School Information
Teachers:Students Diff	4	4	CCD, NCES	Fulltime teachers and students ratio difference (20212019)	School Information
Total Schools Diff	4	5	CCD, NCES	Number of schools difference (20212019)	School Information
Total Students Diff	4	5	CCD, NCES	Number of students difference (20212019)	School Information
% Tested Math G4 Diff	4	6	STAAR, TEA	Grade 4 students tested math exam difference (20212019)	Grade level
Locale_12City: Midsize	4	3	CCD, NCES	School locale City: Midsize	School Information
% Age 04 Pop 10	3	2	Census Bureau	Age 04 Population (%) (2010)	Census demographics
% Age 2024 Pop 10	3	5	Census Bureau	Age 2024 Population (%) (2010)	Census demographics
% Age 3544 Pop 10	3	4	Census Bureau	Age 3544 Population (%) (2010)	Census demographics
% Age 6574 Pop 10	3	2	Census Bureau	Age 6574 Population (%) (2010)	Census demographics
% Female Pop 10	3	4	Census Bureau	Female Population (%) (2010)	Census demographics
% HH MaleChild 10	3	4	Census Bureau	Maleheaded households with children (%) (2010)	Census demographics
% Hispanic Pop 10	3	2	Census Bureau	Hispanic Population (%) (2010)	Census demographics
% Male Pop 10	3	4	Census Bureau	Male Population (%) (2010)	Census demographics
% White Pop 10	3	4	Census Bureau	White Population (%) (2010)	Census demographics
% County Infected 01/29/21	3	2	Covid, USAFacts	Covid infection cases (%) on 01/29/21	COVID cases
ESSERSUPP 23	3	6	ESSER, TEA	ESSERSUPP amount($) (2023)	Lowincome
% Free Lunch Diff	3	5	CCD, NCES	Free lunch eligible students difference (%) (20212019)	Lowincome

% Operational Schools Diff	3	3	CCD, NCES	Operational schools difference (%) (2021-2019)	School Information
% Title 1 Eligible Diff	3	3	CCD, NCES	Title 1 targeted assistance eligible schools difference (%) (20212019)	Lowincome
% Title 1 Schoolwide Diff	3	4	CCD, NCES	Title 1 schoolwide eligible schools difference (%) (20212019)	Lowincome
Locale_21Suburb: Large	3	3	CCD, NCES	School locale Suburb: Large	School Information
Locale_22Suburb: Midsize	3	3	CCD, NCES	School locale Suburb: Midsize	School Information
Locale_23Suburb: Small	3	3	CCD, NCES	School locale Suburb: Small	School Information
Locale_32Town: Distant	3	4	CCD, NCES	School locale Town: Distant	School Information
Locale_33Town: Remote	3	3	CCD, NCES	School locale Town: Remote	School Information
% Age 1014 Pop 10	2	2	Census Bureau	Age 1014 Population (%) (2010)	Census demographics
% Age 4554 Pop 10	2	1	Census Bureau	Age 4554 Population (%) (2010)	Census demographics
% Age 5564 Pop 10	2	2	Census Bureau	Age 5564 Population (%) (2010)	Census demographics
% Age 85Up Pop 10	2	4	Census Bureau	Age 85Up Population (%) (2010)	Census demographics
% HH 1 Female 10	2	0	Census Bureau	1person female households (%) (2010)	Census demographics
% HH 1 Male 10	2	2	Census Bureau	1person male households (%) (2010)	Census demographics
% County Deaths 09/28/20	2	2	Covid, USAFacts	Covid death cases (%) on 09/28/20	COVID cases
% County Deaths 10/30/20	2	1	Covid, USAFacts	Covid death cases (%) on 10/30/20	COVID cases
Employed Diff	2	5	LAUS, BLS	Employed difference (20212019)	Unemployment
Unemployed Rate Diff	2	4	LAUS, BLS	Unemployment rate difference (20212019)	Unemployment
Locale_11City: Large	2	3	CCD, NCES	School locale City: Large	School Information
Locale_13City: Small	2	4	CCD, NCES	School locale City: Small	School Information
Locale_31Town: Fringe	2	4	CCD, NCES	School locale Town: Fringe	School Information
Locale_41Rural: Fringe	2	3	CCD, NCES	School locale Rural: Fringe	School Information
% HH MarriedChild 10	1	2	Census Bureau	Married couple households with children (%) (2010)	Census demographics
% Housing Owner Occup 10	1	3	Census Bureau	Owner Occupied Housing (%) (2010)	Census demographics

% Housing Vacant 10	1	3	Census Bureau	Vacant housing units (%) (2010)	Census demographics
% County Deaths 01/29/21	1	1	Covid, USAFacts	Covid death cases (%) on 01/29/21	COVID cases
Locale_43Rural: Remote	1	4	CCD, NCES	School locale Rural: Remote	School Information

REFERENCES

[1] D. J. Madigan and L. E. Kim, "Towards an understanding of teacher attrition: A meta-analysis of burnout, job satisfaction, and teachers' intentions to quit," *Teaching and teacher education*, vol. 105, p. 103425, 2021.

[2] UNESCO, "Global education monitoring report 2017/8: Accountability in education–meeting our commitments," 2017.

[3] OECD, *Education at a Glance 2021.* Organisation for Economic Co-operation and Development, 2021.

[4] S. Sims and J. Jerrim, *TALIS 2018: Teacher Working Conditions, Turnover and Attrition. Statistical Working Paper.* ERIC, 2020.

[5] P. Reyes and C. Alexander, "Policy brief: A summary of texas teacher attrition," *For the Education Research Center, The University of Texas At Austin: Austin, TX*, pp. 1–8, 2017.

[6] G. Barnes, E. Crowe, and B. Schaefer, "The cost of teacher turnover in five school districts: A pilot study.," *National Commission on Teaching and America's Future*, 2007.

[7] T. G. Carroll, "Policy brief: The high cost of teacher turnover," *National Commission on Teaching and America's Future*, 2007.

[8] L. C. Sorensen and H. F. Ladd, "The hidden costs of teacher turnover," *AERA Open*, vol. 6, no. 1, p. 2332858420905812, 2020.

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

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

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

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

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

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

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

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

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

[18] T. Cardona, E. A. Cudney, R. Hoerl, and J. Snyder, "Data mining and machine learning retention models in higher education," *Journal of College Student Retention: Research, Theory & Practice*, p. 1521025120964920, 2020.

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

[20] T. D. Nguyen, L. D. Pham, M. Crouch, and M. G. Springer, "The correlates of teacher turnover: An updated and expanded meta-analysis of the literature," *Educational Research Review*, vol. 31, p. 100355, 2020.

[21] Y. Baashar, G. Alkawsi, N. Ali, H. Alhussian, and H. T. Bahbouh, "Predicting student's performance using machine learning methods: A systematic literature review," in *2021 International Conference on Computer & Information Sciences (ICCOINS)*, pp. 357–362, IEEE, 2021.

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

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

[24] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Information Fusion*, pp. 785–794, 2016.

[25] T. F. e. a. Guolin Ke, Qi Meng, "Lightgbm: A highly efficient gradient boosting decision tree," *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3149–3157, 2017.

[26] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: gradient boosting with categorical features support," *arXiv preprint arXiv:1810.11363*, 2018.

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

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

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

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

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

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

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

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

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

[36] R. Carlsson, P. Lindqvist, and U. K. Nordänger, "Is teacher attrition a poor estimate of the value of teacher education? a swedish case," *European Journal of Teacher Education*, vol. 42, no. 2, pp. 243–257, 2019.

[37] J. Casely-Hayford, C. Björklund, G. Bergström, P. Lindqvist, and L. Kwak, "What makes teachers stay? a cross-sectional exploration of the individual and contextual factors associated with teacher retention in sweden.," *Teaching and Teacher Education*, vol. 113, p. 103664, 2022.

[38] L. D. Pham, T. D. Nguyen, and M. G. Springer, "Teacher merit pay: A meta-analysis," *American Educational Research Journal*, vol. 58, no. 3, pp. 527–566, 2021.

[39] V. Marz and G. Kelchtermans, "The networking teacher in action: A qualitative analysis of early career teachers' induction process," *Teaching and Teacher Education*, vol. 87, 2020.

[40] R. R. Raab, "A statistic's five years: A story of teacher attrition," *Qualitative Inquiry*, vol. 24, no. 8, pp. 583–591, 2018.

[41] E. Han, "The gendered effects of teachers' unions on teacher attrition: Evidence from district-teacher matched data in the us," *Feminist Economics*, vol. 0, no. 0, pp. 1–33, 2022.

[42] T. M. Gunn and P. A. McRae, "Better understanding the professional and personal factors that influence beginning teacher retention in one canadian province," *International Journal of Educational Research Open*, vol. 2, p. 100073, 2021.

[43] M. Greufe, "Evaluating teacher turnover rates in america, canada, and finland," 2020.

[44] R. K. Shrestha, "Teacher retention in private schools of nepal: A case from bhaktapur district," *KMC Journal*, vol. 4, pp. 167–183, Aug. 2022.

[45] B. Ghojogh, M. N. Samad, S. A. Mashhadi, T. Kapoor, W. Ali, F. Karray, and M. Crowley, "Feature selection and feature extraction in pattern analysis: A literature review," *arXiv preprint arXiv:1905.02845*, 2019.

[46] Z. Zhou and G. Hooker, "Unbiased measurement of feature importance in tree-based methods," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 2, pp. 1–21, 2021.

[47] J. E. Yoo and M. Rho, "Exploration of predictors for korean teacher job satisfaction via a machine learning technique, group mnet," *Frontiers in Psychology*, vol. 11, p. 441, 2020.

[48] S. Abe, "Modified backward feature selection by cross validation.," in *ESANN*, pp. 163–168, Citeseer, 2005.

[49] G. Hooker and L. Mentch, "Please stop permuting features: An explanation and alternatives," *arXiv e-prints*, pp. arXiv–1905, 2019.

[50] W. D. Fisher, "On grouping for maximum homogeneity," *Journal of the American Statistical Association*, vol. 53, no. 284, pp. 789–798, 1958.

[51] "LightGBM optimal split for categorical features." `https://lightgbm.readthedocs.io/en/v3.3.2/Features.html#optimal-split-for-categorical-features`.

[52] "HistGradientBoostingClassifier categorical features support." `https://scikit-learn.org/stable/modules/ensemble.html#categorical-support-gbdt`.

[53] "XGBoost optimal partitioning." `https://xgboost.readthedocs.io/en/stable/tutorials/categorical.html#optimal-partitioning`.

[54] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," 2017.

[55] N. C. for Education Statistics, "The national center for education statistics (nces) is the primary federal entity for collecting and analyzing data related to education.." https://nces.ed.gov.

[56] N. C. for Education Statistics, "1999-2000 sass public-use data and documentation & 2000-01 tfs public-use data and documentation." https://nces.ed.gov/surveys/sass/dataprod9901.asp.

[57] N. C. for Education Statistics, "Schools and staffing survey (sass)." https://nces.ed.gov/surveys/sass/dataproducts.asp.

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

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

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

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

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

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

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

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

[66] O. of Elementary & Secondary Education, "Elementary and secondary school emergency relief fund." https://oese.ed.gov/offices/education-stabilization-fund/elementary-secondary-school-emergency-relief-fund/.

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

[68] N. C. for Education Statistics, "Snational teacher and principal survey (ntps)." https://nces.ed.gov/surveys/ntps/ntps-sass-data.asp.