Teacher Attrition Crisis: Predictive Modeling Based on the Historical Data

June Yu
Department of Computer Science
Texas State University
San Marcos, TX, U.S.A.
juneyu@txstate.edu

Li Feng
Finance and Economics
Texas State University
San Marcos, TX, U.S.A.
li.feng@txstate.edu

Jelena Tešić
Department of Computer Science
Texas State University
San Marcos, TX, U.S.A.
jtesic@txstate.edu

***Abstract*—Teacher attrition in U.S. public schools is quickly reaching a crisis point that requires a thorough understanding to fully address the problem. In this paper, we take a novel approach to address this retention crisis by applying a machine learning approach to publicly available survey data that captured attrition information. There have been multiple attempts to address the crisis, but most of them are incidental and lack large-scale data support. There is no large-scale US public teacher attrition dataset available to address the 2022 crisis from a data science perspective. *Historia est vitae magistra*: we utilize the data from the National Center for Education Statistics (NCES), the School and Staffing Survey (SASS 1999-2000), the K-12 educator labor market and the Teacher Follow-Up Survey (TFS 2000-2001) for our data science pipeline to see if historical data can inform the impact factors of attrition, even if 20+ years old. First, we automatically unveil that STEM teachers have a higher turnover rate than non-STEM teachers from this old data, and confirm the qualitative research findings that came years after in the educational research. Our automated attribute selection also revealed that the STEM field was never indicative of the teacher's decision to leave the classroom. Next, we built a novel prediction pipeline if the teacher is more likely to stay or leave. We propose to modify gradient boost and show superiority in terms of accuracy, precision, and recall in the test set. Finally, we predicted that 3402 more teachers were likely to have left in 2000-2001 than the follow-up survey has captured, and we analyzed the predicted outcome in relation to school and principal attributes.**

***Index Terms*—Educational data science, teacher attrition, gradient boosting, predictive modeling.**

I. Introduction

Teacher attrition in K-12 education is prohibitively high in all corners of the world [19]. 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 [34]. In 2016, the attrition rates in public institutions from the K-12 countries surveyed ranged from 3.3% in Israel to 11.7% in Norway [22]. In the United States, the teacher attrition rate was 8% on an annual basis. Incomplete statistical summaries suggest that almost half of new teachers leave the profession in five years or less [32]. Texas has an prohibitively high teacher attrition rate of 10%, much higher than the national average. Historical trends in Texas show that 19% had left teaching after one year and 12% after the second year. Half of newly trained and hired teachers had left the profession in 5 years [28].

The turnover in the teacher population is natural and desirable at a rate between 6% and 8% for public schools around the world [34]. 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 [22]. A 2007 study estimated that the costs of turnover varied 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) [3]. 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 and an additional $8,750 spent to replace each teacher leaving the district [6]. The high teacher attrition rate is expensive and wasteful and also has a poor impact on student academic progress [33]. High teacher turnover reduces the effectiveness and quality of education [33]. COVID-19 has also had an impact, as a recent study indicates how COVID-19 has led many veteran teachers to retire early and novice teachers to consider alternative professions [40].

The OECD estimates that there will be around 94 million teachers worldwide in 2021 [22]. Based on the worldwide teacher attrition rates, the United Nations Education, Science, and Culture Organization (UNESCO) has issued reports stating that close to 69 million new teachers are needed to provide quality primary and secondary education universally by 2030 as part of the UN Sustainable Development Goals (SDGs). [12]. Educational scientists have taken notice and there has been a great deal of research on professional, personal, and social factors that have led teachers to quit their jobs at high rates in Sweden [7], the United Kingdom [31], Finland [14], Canada [15], Malaysia [36], South Korea [23] and several other countries. In recent years, researchers have used data analysis tools to gain insight into the correlation between high teacher attrition and different socioeconomic factors [24]. The current state of the art focuses on simple statistical analysis [35] or a narrow scope of quantifying the correlation between

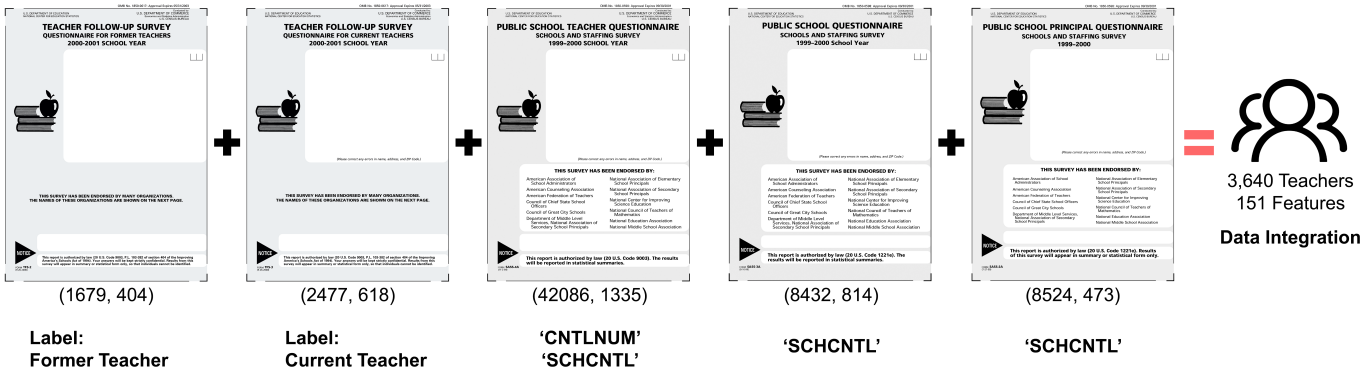

Fig. 1. NCES Data Integration from 5 sources: First, TFS-Former Teacher and TFS-Current Teacher data are concatenated with labeling 1: Current and 0: Former teachers. Then, SASS data are joined with TFS in the order of Public Teachers, Public Schools, then Public Principals by matching control numbers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

social, educational or personal attributes with the teacher attrition rate [21]. These summarized statistics provide a one-dimensional look at the relationship of an outcome with one of the attributes and do not identify the most impactful attributes of all possible candidates [25]. There are no recent teacher attrition data US wide so we utilize 20+ year old big data to demonstrate the value of predictive modeling approach for educational planning and policy making.

The purpose of this paper is to 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 recruitment and retention of public teachers. We propose an end-to-end large-scale educational data modeling pipeline that (1) integrates, cleans, and analyzes educational data; (2) implements automated attribute importance analysis as a way to draw meaningful conclusions; and (3) develops a suite of interpretable teacher retention prediction models utilizing *all* data points and attributes. In the rest of the article, Section III provides background information on US public education data and describes exploratory data analysis; Section IV presents automated approaches to find the most impactful attributes to predict teacher attrition; Section V summarizes the state-of-the-art modeling comparison and the results of the experiment; and Section VI concludes the article with our findings and recommended next steps.

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 two categories of teacher retention analysis: "The Science of Education", and "Machine Learning and Data Science". The first group of research products focuses on qualitative research, where the objective is to propose, analyze, and establish the relevancy of a single attribute to the teacher attrition rate. The second 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 over-engineering the features.

The Science of Education Teacher turnover, teacher attrition, teacher retention, and teacher recruitment have been analyzed in the worldwide educational literature [22], taking into account specific societal characteristics that influence teachers to quit their jobs in Sweden [5, 7], South Korea [21, 23], the United States [19, 20, 26, 18], Canada [15], Finland [14], Nepal [29] 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.

Machine Learning and 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 [4]. 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 [4]. 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 [27]. 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 overengineering attributes used in modeling [21]. Demographic, academic, family / personal and internal assessments were found to be the most frequently used attributes in predicting student performance in class, at grade levels, on standardized tests, etc. [2]. 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 [35]. 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

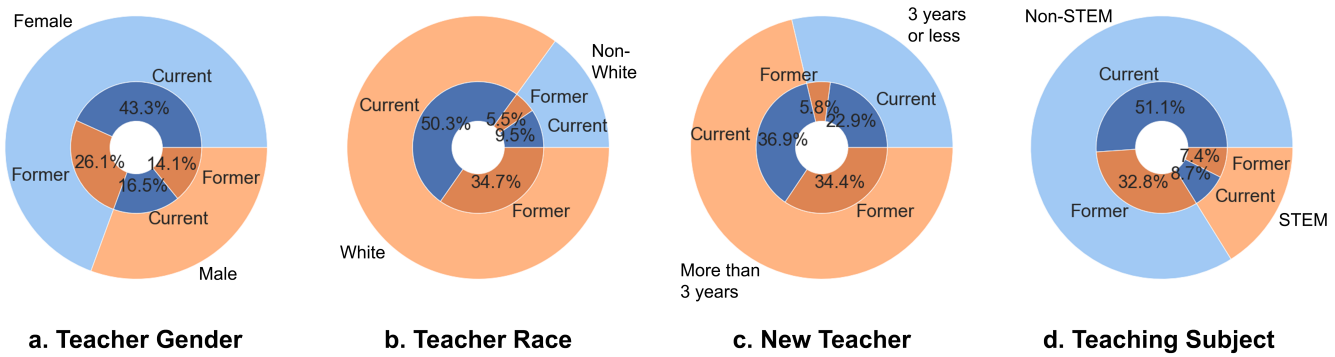

Fig. 2. SASS Exploratory Retention Analysis [10]: (a) female teachers are 2/3 majority while the turnover rate is higher for male teachers; (b) white(Non-Hispanic) teachers are the majority race/ethnicity in U.S. public schools with the higher turnover rate; (c)(d) Teachers working more than 3 years and teachers teaching STEM subjects have higher turnover rate.

noise as DNN [37].

III. United States Educational Public Data Summary

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 [11]. 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 [10]. NCES led SASS seven times between 1987 and 2011 [10]. 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 [10]. The SASS 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 [10]. 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 [10].

In this work, we analyze the data and documents for public use in 1999-2000 SASS and 2000-2001 TFS [10] 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. In our analysis, we included 70 attributes for public teachers, 9 attributes for public principals, and 45 attributes for public schools. The initial set of 124 attributes consists of 107 categorical attributes and 17 numerical attributes [10]. Note that in the integration process, we discarded the data on 301 current teachers and 215 former teachers because the teachers did not provide information about the principal and the school they were associated with. We observed an interesting correlation of known qualitative attributes that affect the teacher attrition rate [26] 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 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)).

IV. Attribute Aggregation and Automated Importance Scoring

In this section, we propose and compare several data-driven automated attribute selection algorithms. Our goal is to offer an interpretable suite of attribute importance analysis approaches and to avoid Garbage In Garbage Out (GIGO) and learning Trivial Models traps. In Section V, we offer policy makers opportunities to draw meaningful conclusions. We utilize unlabeled portions of the data sets and a full set of attributes for the Section IV-A experiment. Section **??** and Section IV-B use a final modeling labeled data set with 3,640 teachers from 2,838 schools, comprised of 53 attributes and labels of 2,176 current teachers and 1,464 former teachers.

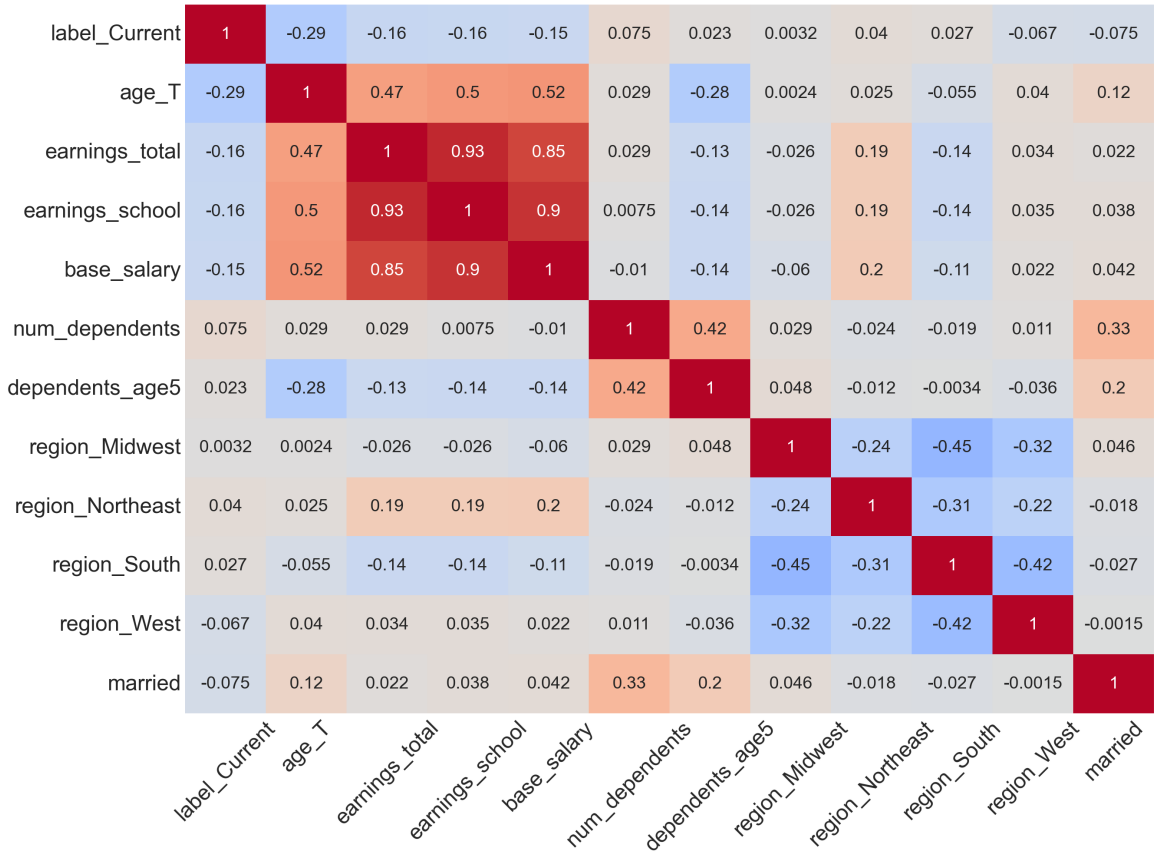

Fig. 3. Attribute Correlation Analysis of the SASS/TFS data. The *base_salary* is highly linearly correlated with the *earnings_school* and *earnings_total*. We use the high correlation coefficient to aggregate linearly dependent attributes into one.

We randomly split the data into a training set (2912 teacher instances, 80%) and a test set (728 teacher instances, 20%).

A. Attribute Filtering by Mutual Correlations

SASS and TFS data provide a lot of overlapping information [10]. In this section, our goal was to build a quazi-orthonormal attribute space. We wanted to avoid artificial weighting of the attributes in the modeling step, so we utilized correlation filtering in this section to aggregate linearly related attributes in the dataset into one attribute. To this end, we have expanded several of categorical attributes to multiple binary attributes as we found that multiple separate categories capture highly overlapping data. Our expanded set contains the 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 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. 1 as:

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

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. The correlation coefficients for the dataset are illustrated in Figure 3, and we use the high correlation coefficient to aggregate linearly dependent attributes into one. The example is illustrated in Figure 3: the base_salary is highly correlated linearly with the earnings_school and earnings_total so we kept only one attribute in our analysis. Table I offer examples of how we combined the highly correlated aggregated attributes into one. We combined all binary dummy-coded variables from related categories as a set in variable selection. This approach is illustrated with new variables in Table I and reduces our attribute dimension to 53. The attributes are listed on x axis in Figure 4.

Recent work in educational data analysis emphasizes the importance of modeling attributes for quantitative research. Statisticians have used the variation of the ridge approaches described in this section and added more penalty terms to account for the correlated attributes and the sampling sensitivity of the forward sampling approaches [36]. In this experiment, we evaluated the importance of 53 selected attributes by fitting a regression model for the importance of features in the training data set. Backward elimination removed attributes that did not have a significant effect on the prediction of teacher attrition. We use the well-known series of global methods that

New Label	From Labels	New Label	From Labels
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 on 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 on salary (1 0)	incen_STEM_comp_sci incen_STEM_math incen_STEM_phys_sci incen_STEM_biology	incen_STEM: Pay recruit incentives on STEM fields (1 0)
urbanicity_LargeCity urbanicity_SmallTown urbanicity_MidCity	urbanicity: Urbanic locale (3 categories)		

TABLE I

EXAMPLES OF AGGREGATED TEACHER ATTRIBUTES AND SCHOOL ATTRIBUTES FILTERED BY CORRELATIONS IN THE SASS DATA SET.

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)

TABLE II

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.

use logistic regression modeling and different penalties for regularization [17] throughout the training data set. Regression with *Ridge* regularization uses the penalty term L2 applied to the sum of the squares of all regression coefficients and works

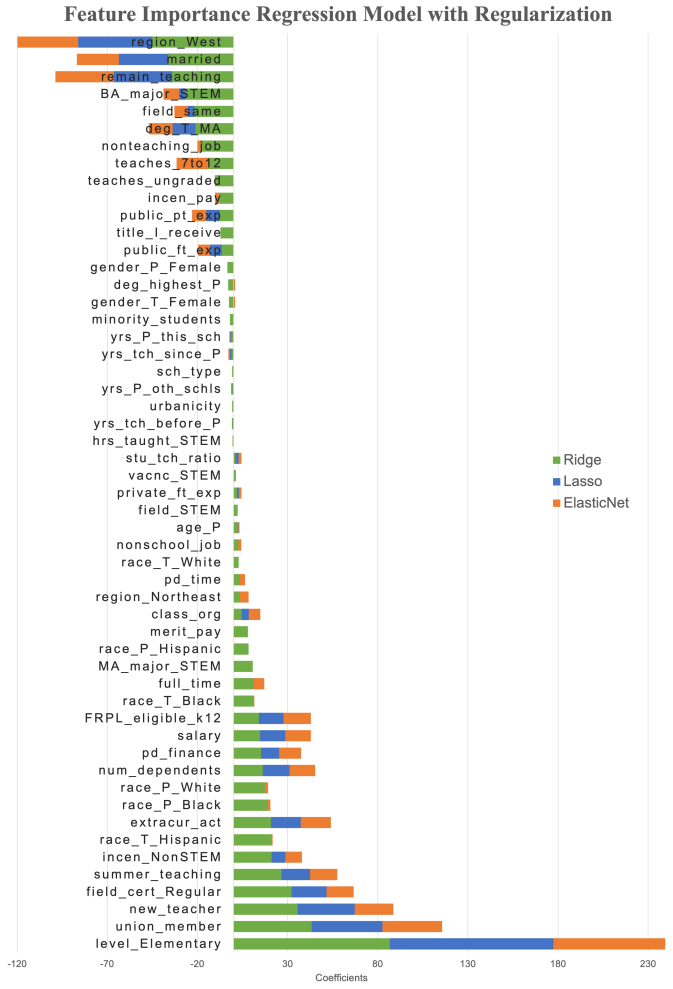

Fig. 4. Automated attribute selection using regression with Ridge (green), Lasso (blue), and ElasticNet (orange) regularization penalty term in loss function, fitted on the training dataset of 2920 instances and 53 attributes. Based on the low importance in all 3 selections, we remove attributes *class_org* and *teaches_ungraded* from the attribute set.

well when many different attributes affect prediction. Lasso regularization uses the L1 penalty term as a way to reduce the number of features in a model by allowing regression coefficients to be zeroed out. Lasso performs well with a sparse data set and if there are few significant predictors that influence the model. If we have two highly correlated attributes in the mix, Lasso will choose one or the other based on its performance in the present data sample. ElasticNet combines the Lasso and Ridge penalty terms for an optimal ranking of the importance of features [9]. Automated attribute weigh-in using regression with the Ridge (green), Lasso (blue), and ElasticNet (orange) regularization penalty term in the loss function, fitted to the training data set of 2920 instances and 53 attributes, is illustrated in Figure 4.

B. Multi-View Relevancy of the Attribute

In this section, we compare and contrast nine different approaches to evaluating the importance of features.

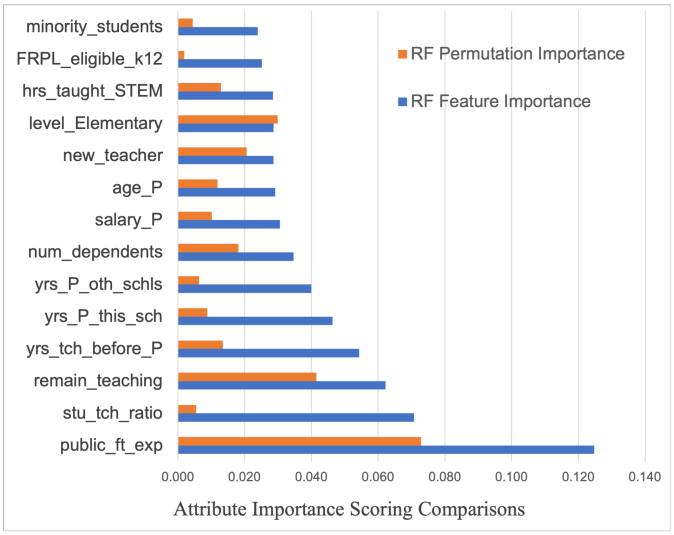

Fig. 5. Random Forest Feature Importance and Permutation Attribute ranking comparison

•**Variance Threshold** is a simple and powerful approach to remove attributes by eliminating all attributes with low variance in the training set [13], since there is no meaningful information in that attribute. We applied a low variance threshold, 0.8*(1-0.8), to the training data set to drop a feature containing 80% of similar values and selected 33 of the most relevant attributes. The selection is illustrated in Figure 6 with the blue bar.

•**Lasso Regularization** This logistic regression with the L1 penalty term shrinks coefficients by minimizing the loss function during the training. As the method reduces the coefficient of features to be exactly zero, it selected 35 of the important features decreasing the variance. The selection is illustrated in Figure 6 with the orange bar.

•**Recursive Feature Elimination (RFE)** We propose Recursive Feature Elimination (RFE) as a baseline model. Attributes ranked according to the importance of characteristics in penalized regularization regression modeling on a small scale have been supported in the qualitative research literature [38]. Here, we employ the RFE cross-validation score of the regression model with a ridge penalty and random forest for the selection of backward features. The algorithm starts by fitting ridge regression and random forest models to the full set of 51 attributes, so we can eliminate candidates with the smallest coefficient and feature importance from the ridge regression and random forest respectively that deteriorate the 10-fold cross-validation score of the models on the training data. The final set is a set of candidates that do not deteriorate the generalizability of the model [1]. This approach selected 18 and 46 of the most relevant attributes from each model, and the selection is illustrated in Figure 6 with beige and pink bars.

•**Random Forests** Random Forests is a powerful machine learning classification algorithm. The Random Forest algorithm has a built-in attribute importance measured by the Gini importance or mean decrease impurity. This is a built-

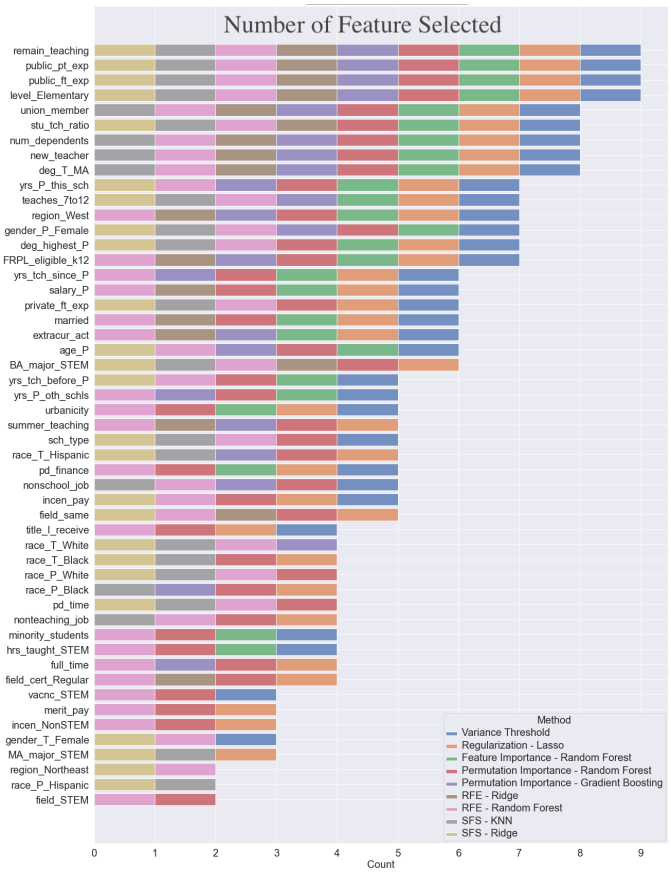

Fig. 6. All nine methods select same 4 features *remain_teaching, public_pt_exp, public_ft_exp, level_Elementary* as the most important features.

in feature of the algorithm, and we set the threshold for the importance of the feature at 0.011 that is 50th percentile of the feature importance. This approach selected 26 of the most relevant attributes, and the selection is illustrated in Figure 6 with green bars. Interpreting the importance of features in machine learning models is challenging when features are dependent [41].

•**Permutation feature importance** (PMI) overcomes limitations on the importance of impurity-based characteristics, as the method combined with Random Forests and Gradient Boosting does not have a bias toward high-cardinality attributes. This approach selected 44 and 23 of the most relevant attributes from Random Forests and Gradient Boosting respectively, and the selection is illustrated in Figure 6 with red and purple bars. •**Sequential Feature Selection** (SFS) selects an optimal set of features by searching the feature space of all possible combinations in a greedy fashion. Each subset of features adding one predictor at a time forward is evaluated based on the 5-fold cross-validation score of ridge regression and KNN models, and both models selected 25 of the most important features that are shown in Figure 6 gold and silver bars.

In Figure 6, all nine approaches selected **four** attributes: *remain_teaching* (teacher responded to the survey question on how likely they will remain 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* (is the elementary school of the school) as the most important attributes.

Figure 5 indicates the top important attributes from Random Forest and Random Forest 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 dependents of teachers), *age_P* (age of a principal), *new_teacher* (teachers who teaching 3 years or less), *level_Elementary* (teachers teaching at a elementary school), and *hrs_taught_STEM* (hours of teaching STEM subjects per week) are the only eight overlapping highly impactful attributes. Vanilla Random Forest has 26 features with an impact score greater than 0.011. Both methods select *public_ft_exp* as the most significant characteristic, which is the years of full-time teaching experience in public schools. Specifically, as teachers work longer years as full-time teachers in public schools, we can better predict the teacher retention.

V. Analysis and Prediction Modeling of Teacher Attrition

We proposed an elegant and simple way to identify schools with critical attrition personnel in unlabeled data. We analyzed, compared, and contrasted 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.

A. Prediction Leave Decision Modeling

We built a robust prediction model to determine whether the teacher will leave or not, which was a simple binary classification challenge. We established a baseline model using state-of-the-art machine learning methods using grid model search and 10-fold cross-validation to find the optimal parameters for Logistic Regression, KNN, linear and RBF kernel SVM, random forests, gradient boosting and stochastic gradient descent classifiers [38]. We used a 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 training set (2,192 teacher instances, 80%) and test set (728 teacher instances, 20%). Each model was evaluated with 10-fold cross-validation and GridSearch to find the best hyperparameters with shuffling and stratification on the label. Each of the methods is evaluated for the full set of attributes and the nine-dimensionality reduction methods explained in Section IV-B. The feature reduction methods produce a different number of data points: the full set is the 51 attributes, and 48, 43, 32, 22, 15, 9, and 4 attributes are selected by the 3, 4, 5, 6, 7, and 8 methods respectively out of the nine-dimensionality reduction methods. The performance of the method in the training set by the number of features is

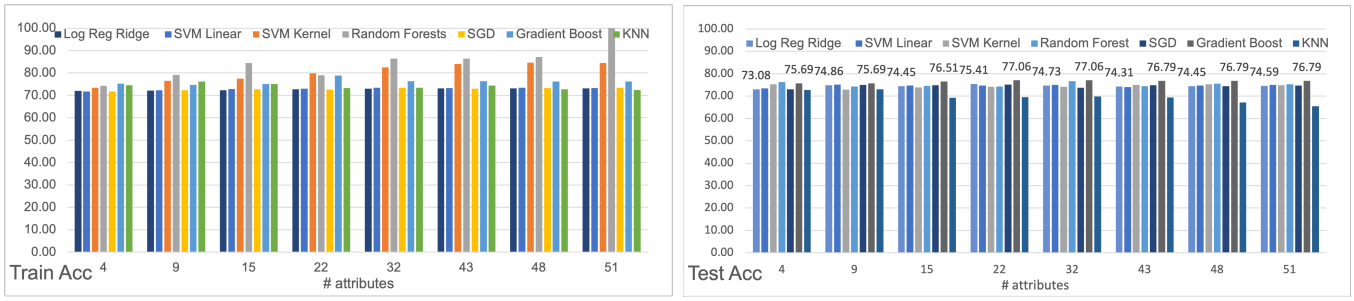

Fig. 7. Seven Machine Learning Models fitted to the training dataset, compared through training and test accuracy.

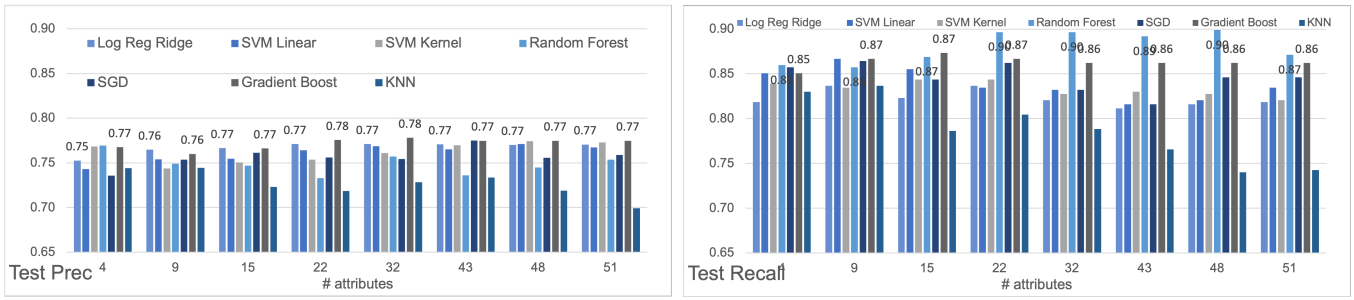

Fig. 8. Seven Machine Learning Models compared on the test set through precision and recall scores demonstrate that Gradient Boost has the most robust and realizable modeling performance.

illustrated in Figure IV-B and Figure 8. The best performing feature set in the train set includes 48 features that are selected by the 3 dimensionality reduction methods and trained with Random Forests with 87% accuracy, followed by SVM with RBF kernel 85% and Gradient Boosting 79% accuracy trained with 48 and 22 features selected by the 3 and 6 dimensionality reduction methods.

Although the accuracy and precision of the logistic regression test are comparable to the best performing methods in Figure IV-B and Figure 8, the recall score is much lower. Gradient Boosting followed by Random Forests are the baseline models of choice with 77% accuracy and 78% precision, and more than 87% recall in the test set. Our data fit the description of tabular data, since gradient-boosting approaches showed the most robustness when dealing with heterogeneous tabular data. Gradient Boosting assembles many weak decision trees, but unlike random forests, it grows trees sequentially and iteratively, growing the trees based on the residuals of the previous trees, thus Gradient Boosting approaches handle tricky observations well. We selected four state-of-the-art gradient-boosting algorithms: XGBoost, HistGradientBoosting, LightGBM, and CatBoost [30]. These algorithms are optimized compared to the traditional Gradient Boosting in terms of faster and efficient fitting: histogram-based algorithm and data sparsity. In contrast to pointwise split of the traditional Gradient Boosting that is prone to be overfitting, the algorithms approximate gradient estimates by creating a histogram for tree splits. As this histogram algorithm does not handle sparsity of 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 [8]. Also, LightGBM provide the techniques Gradient-based One-Side Sampling that 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 [16]. The four boosting algorithms are improved by tuning regularization hyperparameters evaluated with 10 iterations of 10-fold cross-validation RandomizedSearch. To constrain tree structures curbing complex and longer trees growing, the parameters such as the number of trees, the depth of trees, the number of leaves per tree are adjusted. In addition, setting smaller learning rate, less than 0.5, allows weighting trees slowing down the learning by a small amount at each iteration to reduce errors. Next, we compared them with our best-performing baseline models, Gradient Boosting and Random Forests, in Figure 9(left) and showed that the reduction in dimensionality does not play a role in these boosting algorithms as their train accuracy scores remain stable. Next, we compared accuracy, precision and recall in the test set in Figure 9(right). All four state-of-the-art gradient-boosting algorithms perform well similar to the vanilla Gradient Boost implementation for our data set. In this comparison, we have included a simple neural network model with 1 dense layer, 130 neurons, ReLU activation function, and binary cross-entropy loss function. Note that we did not optimize this DNN as recent data show that gradient boosting is superior to any DNN in tabular data [30]. Although the training data measures were greater than 95% for accuracy, precision and recall, our model was overfitted, resulting in the lowest precision, recall, and accuracy in the test set in Figure 9(right).

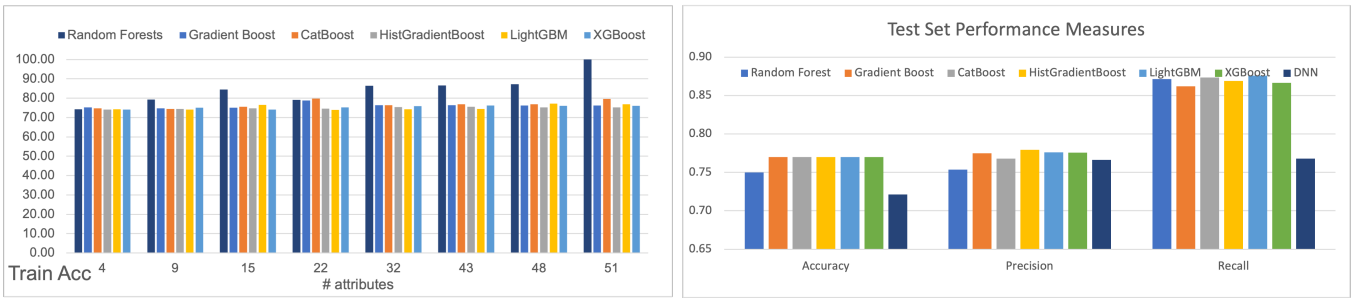

Fig. 9. Advanced gradient boosting method accuracy on the train set shows that dimensionality reduction plays no role in the model performance (left); Full attribute space test set metric (right).

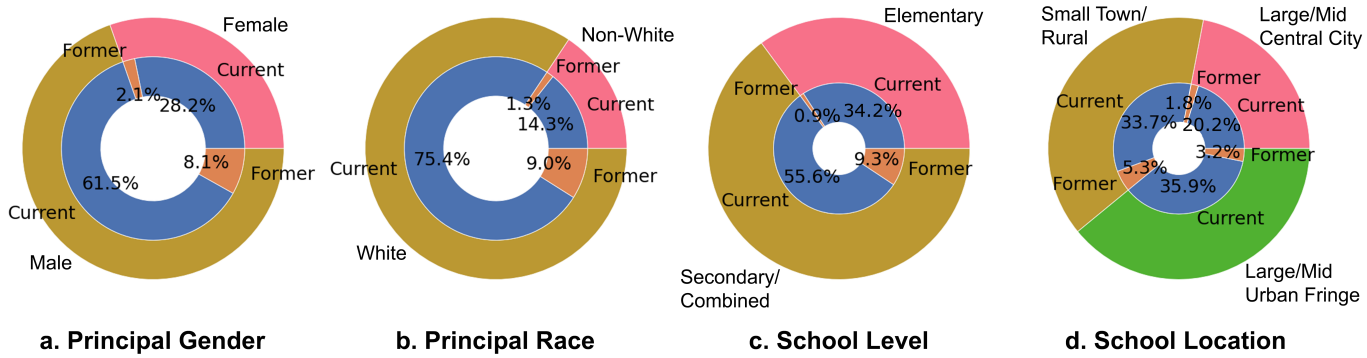

Fig. 10. Teacher attrition prediction analysis per school and principal attributes.

B. Teacher Retention Prediction and Analysis

We proposed using the best gradient boost model (Sect. V-A) to predict teacher attrition rates per school, and we demonstrated the results for teachers who did not participate in the follow-up survey [10]. 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 356 schools have 10+ teachers participating in the SASS survey [10]. 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) [10]. This new dataset does not contain two attributes available only for the TFS data: marital status and the number of dependents. We fit a new CatBoost model with 49 features on the training dataset, and rank predictions on the test set. Our model predicts 3402 teachers from the unlabeled SASS dataset have also left education (80%+ model confidence). The breakdown of predictions is in Figure 10: (a) female principals have less former teachers(7%) than male principals(11.7%); (b) Non-White principals have less Former teachers(8.2%) than the ones for White principals(10.6%); (c) Secondary/Combined schools have significantly higher Former teacher ratio(14.3%) than Elementary school former teachers(2.7%), and (d) schools located in rural areas have higher ratio of former teachers(13.5%) than schools in urban areas (8.1%).

VI. Conclusion and Future Work

Our intentional data science pipeline can automatically uncover important attributes for modeling teacher attrition. In addition, gradient-boosting models are superior for this challenge and predict teacher attrition aggregated per school for unlabeled data. Policy makers can use our predictive model to focus resources on schools and teachers that are highly likely to leave the system and personalize the effort to keep teachers in public schools. The next steps are to expand our attribute base and automate attribute aggregation, and expand the analysis to world-wide rteacher attrition data. Reproducible experiments will be published on [39] at the presentation time.

References

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

[2] Yahia Baashar, Gamal Alkawsi, Nor'ashikin Ali, Hitham Alhussian, and Hussein T Bahbouh. Predicting student's performance using machine learning methods: A systematic literature review. In *2021 International Conference on Computer & Information Sciences (ICCOINS)*, pages 357–362. IEEE, 2021.

[3] Gary Barnes, Edward Crowe, and Benjamin Schaefer. The cost of teacher turnover in five school districts: A pilot study. *National Commission on Teaching and America's Future*, 2007.

[4] Tatiana Cardona, Elizabeth A Cudney, Roger Hoerl, and Jennifer Snyder. Data mining and machine learning retention models in higher education. *Journal of College Student Retention: Research, Theory & Practice*, page 1521025120964920, 2020.

[5] Rickard Carlsson, Per Lindqvist, and Ulla Karin Nordänger. Is teacher attrition a poor estimate of the value of teacher education? a swedish case. *European Journal of Teacher Education*, 42(2):243–257, 2019.

[6] Thomas G Carroll. Policy brief: The high cost of teacher turnover. *National Commission on Teaching and America's Future*, 2007.

[7] Jeffrey Casely-Hayford, Christina Björklund, Gunnar Bergström, Per Lindqvist, and Lydia 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*, 113:103664, 2022.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Information Fusion*, page 785–794, 2016.

[9] Lixin Cui, Lu Bai, Yanchao Wang, Xin Jin, and Edwin R Hancock. Internet financing credit risk evaluation using multiple structural interacting elastic net feature selection. *Pattern Recognition*, 114:107835, 2021.

[10] National Center 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.

[11] National Center 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.

[12] UNESCO Institute for Statistics. The world needs almost 69 million new teachers to reach the 2030 education goals. *(Fact Sheet No. 39)*, 2016.

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

[14] Mary Greufe. Evaluating teacher turnover rates in america, canada, and finland. 2020.

[15] Thelma M Gunn and Philip A McRae. Better understanding the professional and personal factors that influence beginning teacher retention in one canadian province. *International Journal of Educational Research Open*, 2:100073, 2021.

[16] Thomas Finley et al. 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*, pages 3149–3157, 2017.

[17] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[18] Eunice Han. The gendered effects of teachers' unions on teacher attrition: Evidence from district–teacher matched data in the us. *Feminist Economics*, 0(0):1–33, 2022.

[19] Daniel J Madigan and Lisa E Kim. Towards an understanding of teacher attrition: A meta-analysis of burnout, job satisfaction, and teachers' intentions to quit. *Teaching and teacher education*, 105:103425, 2021.

[20] Virginie Marz and Geert Kelchtermans. The networking teacher in action: A qualitative analysis of early career teachers' induction process. *Teaching and Teacher Education*, 87, 2020.

[21] Tuan D Nguyen, Lam D Pham, Michael Crouch, and Matthew G Springer. The correlates of teacher turnover: An updated and expanded meta-analysis of the literature. *Educational Research Review*, 31:100355, 2020.

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

[23] Lam D Pham, Tuan D Nguyen, and Matthew G Springer. Teacher merit pay: A meta-analysis. *American Educational Research Journal*, 58(3):527–566, 2021.

[24] Xin Qiao and Hong Jiao. Data mining techniques in analyzing process data: A didactic. *Frontiers in psychology*, page 2231, 2018.

[25] Lixia Qin. Why they want to leave? a three-level hierarchical linear modeling analysis of teacher turnover intention. In *Methodology for Multilevel Modeling in Educational Research*, pages 311–337. Springer, 2022.

[26] Rebecca Raine Raab. A statistic's five years: A story of teacher attrition. *Qualitative Inquiry*, 24(8):583–591, 2018.

[27] A. Ravishankar Rao, Yashvi Desai, and Kavita Mishra. Data science education through education data: an end-to-end perspective. In *2019 IEEE Integrated STEM Education Conference (ISEC)*, pages 300–307, 2019.

[28] Pedro Reyes and Celeste Alexander. Policy brief: A summary of texas teacher attrition. *For the Education Research Center, The University of Texas At Austin: Austin, TX*, pages 1–8, 2017.

[29] Rajan Kumar Shrestha. Teacher retention in private schools of nepal: A case from bhaktapur district. *KMC Journal*, 4(2):167–183, Aug. 2022.

[30] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[31] Sam Sims. Modelling the relationships between teacher working conditions, job satisfaction and workplace mobility. *British Educational Research Journal*, 46(2):301–320, 2020.

[32] Sam Sims and John Jerrim. *TALIS 2018: Teacher Working Conditions, Turnover and Attrition. Statistical Working Paper.* ERIC, 2020.

[33] Lucy C. Sorensen and Helen F. Ladd. The hidden costs of teacher turnover. *AERA Open*, 6(1):2332858420905812, 2020.

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

[35] Ze 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*, 11, 2020.

[36] Azizi Yahaya, Ismail Maakip, Peter Voo, Sharon Kwan Sam Mee, Balan Rathakrishnan, and Sonny Anak Jumpo. An exploratory study on predictors associated with teachers' job satisfaction in malaysian sports schools. *Hong Kong Journal of Social Sciences*, 2021.

[37] Kuan Yan. Student performance prediction using xgboost method from a macro perspective. In *2021 2nd International Conference on Computing and Data Science (CDS)*, pages 453–459, 2021.

[38] Jin Eun Yoo and Minjeong Rho. Exploration of predictors for korean teacher job satisfaction via a machine learning technique, group mnet. *Frontiers in Psychology*, 11:441, 2020.

[39] June Yu and Jelena Tešić. NCESS SASS and TFS teacher attrition modeling code. https://github.com/DataLab12/NCESanalysis.

[40] Gema Zamarro, Andrew Camp, Dillon Fuchsman, and Josh 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.

[41] Zhengze Zhou and Giles Hooker. Unbiased measurement of feature importance in tree-based methods. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2):1–21, 2021.