Machine Learning Modeling of Teacher Attrition Data

Educational Researcher
XX(X):1–16
©The Author(s) 2024
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Abstract

Teacher attrition in public schools has reached a critical level, necessitating data-driven strategies to inform educational policies. This study employs a comprehensive analysis of National Center for Education Statistics data to address this issue. We developed an open-source educational data modeling pipeline to generate in-depth insights into attrition dynamics. Utilizing AI/ML techniques, our analysis identified key determinants influencing teacher attrition, including principal demographics, school type, and geographic location. Predictive modeling revealed a projected attrition rate exceeding 10% based on data older than 20 years, aligning with current trends observed in the U.S. education system. These findings underscore the importance of leveraging open-source data for large-scale research and policy development in education.

Keywords

AI/ML, Tabular Data, Missing Value, Correlated attributes

Introduction

Teacher attrition represents a critical challenge in education, profoundly affecting teaching quality and student outcomes. Characterized by the percentage of teachers leaving their positions within a given academic year, attrition rates are pivotal in determining the efficacy of public schools globally. While a turnover rate of 6% to 8% is generally considered normal and beneficial, both excessively low and high attrition rates can severely undermine educational success (UNESCO, 2017). The consequences of teacher attrition are stark: schools experiencing attrition rates below 5% may suffer from stagnation, lacking the influx of fresh perspectives and innovative approaches that new educators provide. On the other hand, when attrition rates exceed 10%, the detrimental effects on a public school's educational effectiveness become increasingly pronounced. The high teacher attrition rate carries substantial costs and detrimental effects on student academic progress. The constant turnover of teachers compromises the continuity and quality of education, hampering students' learning experiences (Sorensen and Ladd, 2020). Moreover, the financial implications of replacing teachers burden public budgets. A study conducted in 2007 estimated the cost of teacher turnover to range from approximately $4,000 to nearly $18,000 per Teacher, with the total annual cost of excess turnover in the United States reaching $7.34 billion (Barnes et al., 2007; Carroll, 2007). Given the urgency and impact of the issue, this research aims to provide data-driven insights into the factors influencing the recruitment and retention of public school teachers in the United States.

The global landscape of teacher attrition is diverse, as evidenced by the wide range of rates observed in different countries. In a survey of K-12 public institutions conducted in 2016, attrition rates varied from 3.3% in Israel to 11.7% in Norway (OECD, 2021). The attrition rate in the United States has traditionally been around 8% per year. However, recent years have seen an alarming increase, with almost half of the new teachers leaving the profession within five years or less (Sims and Jerrim, 2020). The COVID-19 pandemic has exacerbated the problem of teacher attrition in K-12 education worldwide (Madigan and Kim, 2021). The impact has been significant in the United States, with over 300,000 public school teachers and staff leaving their positions between February 2020 and May 2022, resulting in a 3% decrease in the workforce (Dill, 2022). A poll by the National Education Association in 2022 revealed that 55% of teachers desired to leave education earlier than planned, compared to 37% in the previous year (Dill, 2022).

This research embarks on a transformative journey to tackle the pressing issue of teacher attrition through a detailed analysis of National Center for Education Statistics data. By crafting an open-source educational data modeling pipeline, we have unlocked powerful insights into the dynamics of the trends characterizing teacher turnover. Our innovative use of AI/ML techniques has pinpointed critical factors influencing attrition, such as principal demographics, school type, and geographic location. Predictive modeling from over two decades of data forecasts an alarming attrition rate exceeding 10%, mirroring current trends in the U.S. education system.

These compelling findings illuminate the critical need for harnessing open-source data to drive impactful research and shape effective educational policies. By addressing the root causes of teacher attrition and exploring strategies to counteract its detrimental effects, we pave the way for more resilient and dynamic educational environments. Furthermore, our automated attribute importance analysis and interpretable prediction models offer a robust framework for understanding and improving teacher retention. This work not only advances academic knowledge but also provides actionable tools for policymakers and educators dedicated to fostering a thriving educational landscape. Section provides background information on public education data in the United States and outlines the exploratory data analysis conducted. Section introduces automated approaches to identify the most relevant attributes for teacher retention. Section summarizes the state-of-the-art modeling comparison and presents the results of our experiments. Section concludes the article by summarizing our findings and offering insights into future directions for addressing the challenges of teacher hiring and retention.

Related Work

The field of data science has witnessed an increase in the application of machine learning (ML) tools to correlate attributes with teacher attrition rates. From just two studies in 2010, the number rose to seven in 2017 (Cardona et al., 2020). These studies employed popular ML techniques such as logistic regression, support vector machines, Bayesian belief networks, decision trees, and neural networks. While these techniques achieved accuracy above 70% for simple classification tasks, their narrow scope and limited feature engineering often resulted in poorly translating domain-specific knowledge into effective models (Cardona et al., 2020). A more comprehensive evaluation of 30 selected articles revealed that deep neural networks (DNN), decision trees, support vector machines (SVM), and nearest neighbor (kNN) methods were preferred for predicting student academic performance (Rao et al., 2019). Additionally, a detailed review of 25,771 studies, incorporating 120 quantitative data analyses of teacher turnover, highlighted the overfitting of attributes in the evaluated methods (Nguyen et al., 2020). Demographic, academic, family/personal, and internal assessment attributes usually predict student performance across various contexts (Baashar et al., 2021). In the realm of data science for education application, a large-scale study analyzed the Big Fish Little Pond Effect (BFLPE) across 56 countries in fourth-grade math and 46 countries in eighth-grade math, utilizing extensive data from the Trends in International Mathematics and Science Study (TIMSS) (Wang, 2020). This study employed simple statistical analysis to establish correlations. Furthermore, recent findings indicate that state-of-the-art machine learning techniques in tabular data outperform existing approaches, demonstrating robustness to input bias and noise (Yan, 2021). In the domain of machine learning, gradient-boosted decision trees (GBDT) models, such as XGBoost, LightGBM, and CatBoost, have gained popularity for analyzing tabular data (Chen and Guestrin, 2016; Dorogush et al., 2018; Guolin Ke, 2017). Deep learning models, including TabNet, DNF-Net, and Neural Oblivious Decision Ensembles (NODE), have emerged as state-of-the-art techniques for tabular data analysis (Abutbul et al., 2020; Arik and Pfister, 2021; Popov et al., 2019). However, there is no consensus on whether deep learning surpasses GBDT in tabular data, as standard benchmarks and open-source implementations have been limited (Joseph, 2021; Shwartz-Ziv and Armon, 2022). Recent studies have provided competitive benchmarks comparing GBDT and deep learning models across multiple tabular datasets, revealing that GBDT models still generally outperform deep learning models (Borisov et al., 2021; Grinsztajn et al., 2022; Shwartz-Ziv and Armon, 2022). The field of education economics has extensively analyzed teacher turnover, attrition, retention, and recruitment on a global scale (OECD, 2021). Various studies have investigated these issues in specific contexts, including Sweden, South Korea, the United States, Canada, Finland, Nepal, and many other countries, considering factors such as teacher characteristics, qualifications, school organizational characteristics, resources, student body characteristics, relational demography, accountability, and workforce measures (Carlsson et al., 2019; Casely-Hayford et al., 2022; Greufe, 2020; Gunn and McRae, 2021; Han, 2022; Marz and Kelchtermans, 2020; Pham et al., 2021; Raab, 2018; Shrestha, 2022). However, no principal data-driven study has comprehensively identified the attributes that explain teacher attrition.

Open Source Data for Education

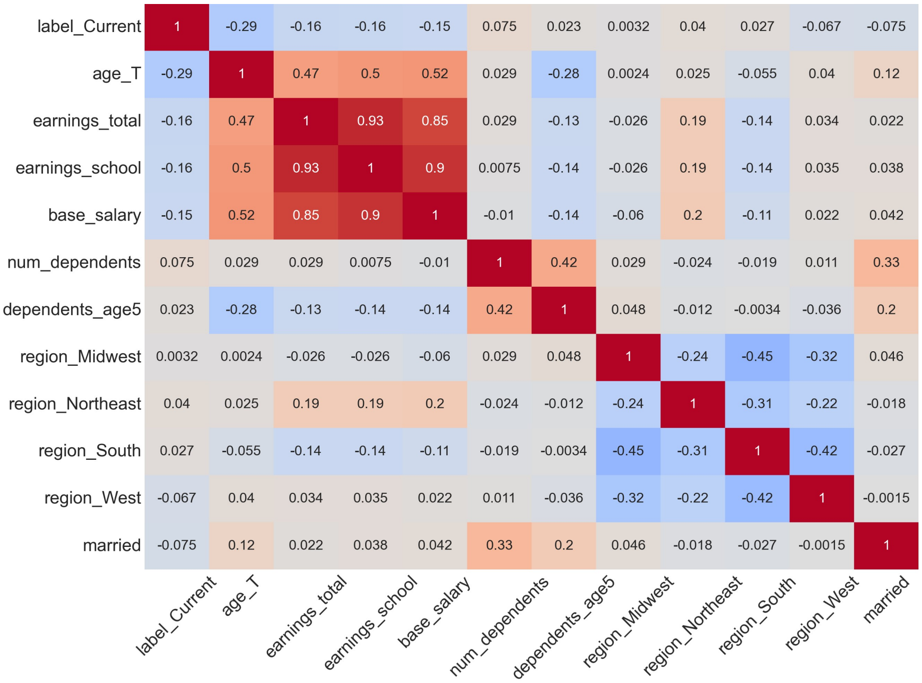
The National Center for Education Statistics (NCES) in the United States is the primary statistical agency responsible for collecting education-related data. NCES gathers international assessment data, administrative data from all public schools in the country, and national survey data, available to the research community to inform policy and practice (National Center, 2022). One of the significant studies conducted by NCES is the Schools and Staffing Survey (SASS), a multiyear study encompassing public and private school districts, schools, principals, and teachers. SASS aims to provide descriptive data on various aspects of elementary and secondary education. It covers teacher demand, characteristics of teachers and principals, school conditions, perceptions of school climate, teacher compensation, district hiring, retention practices, and essential student characteristics within the school (National Center, 2001). The Teacher Follow-Up Survey (TFS) is conducted a year later in conjunction with SASS. TFS focuses on K-12 teachers who participated in the previous SASS survey. The data collected includes a sub-sample of teachers who left teaching within the year and a sub-sample of those who continued teaching, whether in the same school or a different one (National Center, 2001).

Here, we analyze the open source data from the SASS and TFS, as they focus on public schools, teachers, and principal retention in the U.S. (National Center, 2001). The TFS data provide binary labels indicating whether teachers decided to continue teaching (labeled as 1) or leave the profession (labeled as 0). Figure 6 illustrates the data integration pipeline. Of the 42,086 public teachers who participated in the SASS 1999-2000, only 4,156 (less than 10%) participated in the TFS 2000-2001, comprising 2,477 current teachers and 1,679 former teachers. The data set includes 76.6% of schools with at least one Teacher participating in both SASS and TFS. We excluded 301 current teachers and 215 former teachers who did not have TFS data on principal and school associations, resulting in the labeled data set. Initially, 124 attributes, including 107 categorical and 17 numerical attributes, represented 3,640 teachers. These attributes include 70 public teachers, nine public principals, and 45 public schools. Selected Teacher Attributes in the SASS dataset are described in *Supplemental Material* section in the Table 4.

Data analysis reveals that (i) female teachers comprise a two-thirds majority, (ii) male teachers exhibit higher turnover rates, (iii) white non-Hispanic teachers form the majority racial/ethnic group in public schools, and (iv) this group also experiences the highest attrition rate. Surprisingly, this over 20-year-old data analysis reveals that (v) teachers working more than three years and those teaching STEM subjects have the highest annual attrition rates. In the *Supplemental Material* section, the Figure 5 illustrates these data exploratory analysis findings.

Attribute Importance Scoring

We introduce an easily interpretable suite of approaches for analyzing attribute importance in multi-source data analysis. Here, we introduce a novel method to overcome the challenges of working with large-scale survey data containing noise, missing values, and potential data quality issues, often called the ”Garbage In Garbage Out” (GIGO) problem. The SASS and TFS data sets, which provide extensive information with a mix of numerical and categorical data, also exhibit significant overlap (National Center, 2001). To address these challenges and enhance the interpretability of our models, we employ a filter method that identifies correlated attributes. This filtering process allows us to construct a quasi-orthonormal attribute



	label_Current	age_T	earnings_total	earnings_school	base_salary	num_dependents	dependents_age5	region_Midwest	region_Northeast	region_South	region_West	married
label_Current	1	-0.29	-0.16	-0.16	-0.15	0.075	0.023	0.0032	0.04	0.027	-0.067	-0.075
age_T	-0.29	1	0.47	0.5	0.52	0.029	-0.28	0.0024	0.025	-0.055	0.04	0.12
earnings_total	-0.16	0.47	1	0.93	0.85	0.029	-0.13	-0.026	0.19	-0.14	0.034	0.022
earnings_school	-0.16	0.5	0.93	1	0.9	0.0075	-0.14	-0.026	0.19	-0.14	0.035	0.038
base_salary	-0.15	0.52	0.85	0.9	1	-0.01	-0.14	-0.06	0.2	-0.11	0.022	0.042
num_dependents	0.075	0.029	0.029	0.0075	-0.01	1	0.42	0.029	-0.024	-0.019	0.011	0.33
dependents_age5	0.023	-0.28	-0.13	-0.14	-0.14	0.42	1	0.048	-0.012	-0.0034	-0.036	0.2
region_Midwest	0.0032	0.0024	-0.026	-0.026	-0.06	0.029	0.048	1	-0.24	-0.45	-0.32	0.046
region_Northeast	0.04	0.025	0.19	0.19	0.2	-0.024	-0.012	-0.24	1	-0.31	-0.22	-0.018
region_South	0.027	-0.055	-0.14	-0.14	-0.11	-0.019	-0.0034	-0.45	-0.31	1	-0.42	-0.027
region_West	-0.067	0.04	0.034	0.035	0.022	0.011	-0.036	-0.32	-0.22	-0.42	1	-0.0015
married	-0.075	0.12	0.022	0.038	0.042	0.33	0.2	0.046	-0.018	-0.027	-0.0015	1

Figure 1. Correlation of SASS attributes with the TFS attributes.

space, enabling us to observe correlations between different attributes or between a feature and our target label. By identifying and aggregating linearly related attributes, we prevent artificial weighting of attributes during the modeling step. To achieve this, we expand several categorical attributes into multiple binary attributes. Through this expansion, we discover that multiple separate categories capture highly overlapping data, further enhancing the granularity and accuracy of our analysis. 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 $\overline{x}$ and attribute feature vector $y = (y_1, ...y_n)$ with mean $\overline{y}$ is obtained via a Least-Squares fit as defined in Eq. 1.

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

In our analysis, we assign a value of 1 to indicate a perfect positive relationship between variables, -1 for a perfect negative relationship, and 0 when there is no relationship between variables. We aggregate attributes with high correlation coefficients to reduce redundancy and enhance interpretability, as they exhibit linear dependence on each other. We select the one with the highest correlation with our target label from these overlapping attributes. Additionally, we combine all binary dummy-coded variables from related categories as a set during variable selection. This consolidation approach reduces the dimensionality of the attribute set, allowing for improved interpretability and understanding of attribute importance. Before calculating correlation coefficients and identifying linearly dependent attributes, we pre-process categorical attributes with high cardinality. For instance, we convert categorical attributes with numerous categories, such as 80 categories representing the major codes for teachers' BA or MA degrees, into STEM or non-STEM majors. We then expand the remaining categorical attributes into multiple binary attributes to identify highly overlapping data patterns. Our expanded attribute set comprises 134 categorical attributes and 17 numerical attributes. Among these, 78 attributes pertain to public teachers, 17 to public principals, and 56 to public schools. With the correlation coefficients of our data, we combined all binary variables to see if they could be related categories as a set. Finally, we reduce the dimensionality of the data set to 53 attributes, with 39 categorical and 14 numerical. Figure 1 shows the high correlation coefficients among attributes; for example, the *base_salary* highly correlates with *earnings_school* and *earnings_total* attributes. In the *Supplemental Material* Section, Table 3 summarizes examples of how we aggregated sets of attributes.

Feature importance analysis is crucial to machine learning as it has several benefits. It aids in *model interpretation*, allowing us to identify the most influential attributes and understand their relative importance in contributing to the model's predictions. This interpretation is valuable for gaining insights, making informed decisions, and building trust in the model's outputs. Feature importance helps identify the most critical attributes in *feature selection*. The model can generalize, improve performance, and reduce noise by focusing on these high-importance features. Furthermore, feature importance provides valuable insights into the *underlying relationships* within the data. It helps domain experts understand which attributes are crucial in determining the outcome and uncovers meaningful patterns and dependencies. This knowledge can drive further research, guide feature engineering efforts, and inform decision-making processes. Feature importance analysis can help detect *data issues* such as missing values, outliers, or incorrect labels. By examining the importance of features, we can identify any issues or anomalies in the data. Feature Importance scoring allows us to address data quality problems before building the model, ensuring better performance and reliability. We propose a novel fusion of six distinct approaches for feature importance analysis. *Logistic Regression with Lasso Regularization* is the popular baseline used for feature importance analysis beyond data science. Using the L1 penalty term, Lasso Regularization minimizes the loss function during the training of the logistic regression model by shrinking the coefficients. Attributes with non-zero coefficient values are considered and selected for the final set. *Variance Threshold*(VT) allows us to evaluate data quality problems. The method removes attributes with low variance by thresholding the training of the dataset. The top characteristics with the highest variance are selected for the final set (Ghojogh et al., 2019).

Random Forests classification and regularization machine learning algorithm provide attribute importance measures through the Gini importance score. In this paper, we set the threshold at the 50th percentile of attribute importance, and attributes with importance scores above this threshold are included in the final set (Speiser et al., 2019). *Recursive Feature Elimination (RFE)*: RFE starts by fitting a model to the complete attribute set. The algorithm eliminates attributes with the smallest coefficients and removes characteristics that worsen the 10-fold cross-validation score of the models (ridge regression and random forest) on the training data. The final set consists of attributes that do not compromise the model's generalizability (Abe, 2005). *Permutation Feature Importance (PFI)*: PFI measures the difference in accuracy score or other performance metrics between a baseline dataset and a permuted dataset where the values of a feature are replaced with random noise. The attributes with positive importance mean are included in the final set, as the method returns positive and higher values. PFI addresses limitations related to impurity-based attribute importance but can be influenced by feature correlations (Hooker and Mentch, 2019).

Table 1. Nine approaches selected the number of features, and the selection is illustrated in Figure 2 with distinguished bar colors marked in the Color column.

Label	Approach	Method	# Attributes	Color
VT	Variance Threshold	Filter	34	🟫
LR	Lasso Regularization	Embedded	38	🟧
RF FI	RF Feature Importance	Embedded	27	🟧
PFI RF	PFI - Random Forests	Wrapper	28	🟨
PFI RR	PFI - Ridge Regression	Wrapper	33	⬜
RFE RR	RFE - Ridge Regression	Wrapper	18	🟪
RFE RR	RFE - Random Forests	Wrapper	49	🟪
SFS KNN	SFS - KNN	Wrapper	26	🟪
SFS RR	SFS - Ridge Regression	Wrapper	26	🟪

Sequential Feature Selection (SFS): SFS sequentially selects an optimal set of attributes by exhaustively searching through all possible combinations. Each subset adds one predictor at a time and is evaluated based on the 5-fold cross-validation score of ridge regression and KNN models. The method selects half of the provided attributes for the final set (Saha et al., 2020). Six distinct approaches produce nine total scorecards on feature importance, as the last three approaches implement two measures per method. Table 1 presents the number of attributes each approach selects. Among them, RFE with ridge regression resulted in the most miniature set, consisting of 18 attributes. At the same time, RFE with random forests produced the most extensive set, with 49 attributes, as illustrated in Figure 2. We find that all nine attribute importance ranking approaches consistently ranked the following **four** attributes as the most impactful: (1) *remain_teaching* - Teacher responded to the survey question on the likelihood of remaining in teaching); (2) *public_pt_exp* number of years of part-time teaching experience in public schools); (3) *public_ft_exp* - number of years of full-time teaching experience in public schools) and (4) *level_Elementary* - level of school in teaching is elementary, as illustrated in Figure 2. Eight methods agree on the following five most impactful attributes (Figure 2), etc. In this section, we have demonstrated a data-driven way to uncover the most impactful attributes (in positive and negative senses) related to the Teacher's decision to stay or leave the job. Note that race and gender appear to be picked by two or three methods only in Figure 2. In Figure 3, the main attributes of Random Forest and Random Forest Permutation are summarized 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 before becoming a principal), *num_dependents* (number of dependent teachers), *age_P* (age of a principal), *new_teacher* (teachers who teach three 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.

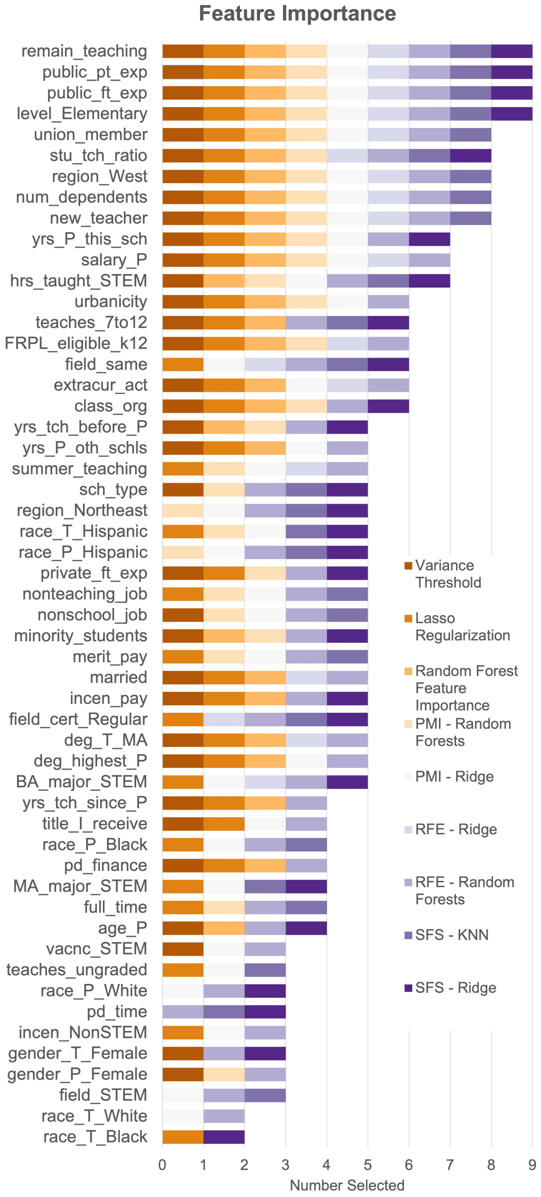


Figure 2. All nine methods select the 4 attributes *remain_teaching, public_pt_exp, public_ft_exp, level_Elementary* as the most important features.

Vanilla Random Forest has 27 attributes with an impact score greater than 0.011. Both methods select *public_ft_exp* as the most significant characteristic: the years of full-time teaching experience in public schools. If teachers work longer years as full-time teachers in public schools, our prediction of teacher retention improves.

Analysis and Prediction Modeling of Teacher Attrition

Prediction Leave Decision Modeling

We have established five baseline models, including the ridge regression as the most common logistic classification model, Support Vector Machines (SVM) and K Nearest Neighbors (KNN) for nonlinear and non-separable data, and two decision tree-based ensemble methods: Random forests and gradient boost.

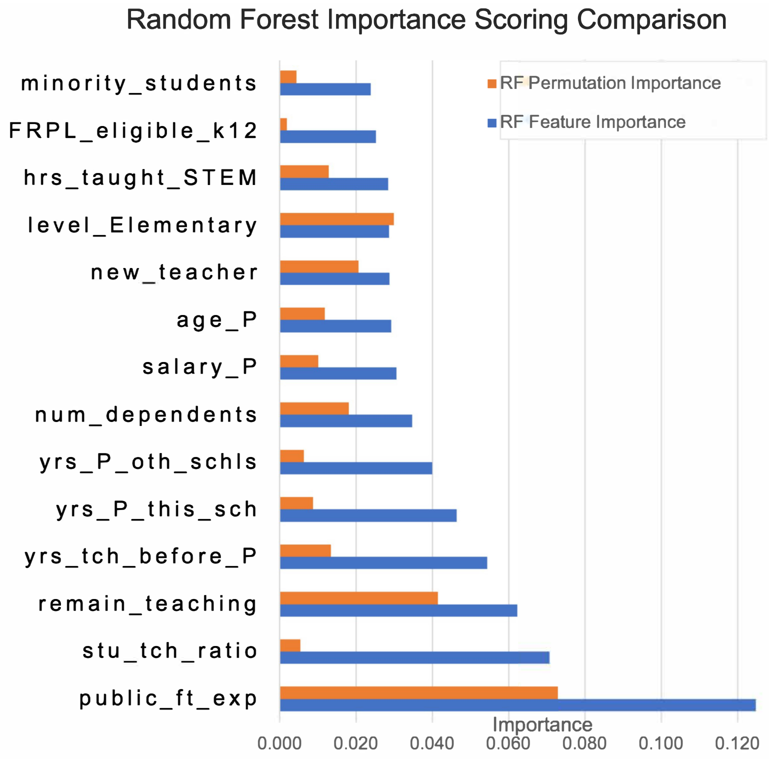


Figure 3. Random Forest Feature Importance and Permutation Attribute ranking comparison

Each model runs with a 10-fold cross-validation of grid search to find optimal hyper-parameters. Training data is 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. The training set contains randomly selected 2,192 teacher instances (80%), and the test set includes 728 teacher instances (20%). The feature reduction methods produced different attributes: the whole set contains 53 attributes, and 18, 26, 26, 27, 28, 33, 34, and 38 attributes are selected by the nine feature selection methods. The ensemble models based on decision trees, gradient boost, and random forests training 27 and 28 chosen attributes by the importance of random forest characteristics and PFI with random forests, respectively, are the models with the highest accuracy (77%) and F1 (82%). The metrics, accuracy, F1, and MCC generally show steady performance across all models except KNN and feature sets. The performance of the five state-of-the-art models in the test set, organized by the number of attributes, is illustrated in Figure 7.

The modeling pipeline applied the advanced gradient boosting models, XGBoost, LightGBM, CatBoost, and HistGradientBoosting to the dataset. Gradient-boosting approaches are optimized for faster and more efficient fitting using a data sparsity-conscious histogram-based algorithm approximating gradient creates estimates by creating a histogram for tree splits. This algorithm handles the data's sparsity, especially for tabular data with missing values and one-hot encoded categorical features. For example, XGBoost uses sparsity-aware split finding that directs the tree split in each non-leaf tree node (Chen and Guestrin, 2016). Additionally, LightGBM provides the Gradient-Based One-Side Sampling technique, which filters data instances with a large gradient to adjust the influence of sparsity, and Exclusive Feature Bundling combining attributes with non-zero values to reduce the number of columns (Guolin Ke, 2017). Handling categorical attributes is challenging when 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 splitting with these methods often result in unbalanced trees and data sparsity, especially for high-cardinality categorical features.

The four gradient boost models implement and suggest optimal methods for processing categorical attributes to optimize numerous boost steps for computing time and memory consumption. LightGBM, HistGradientBoosting, and XGBoost use the optimal split method (Fisher, 1958) 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(Pedregosa et al., 2011). CatBoost proposes 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 (Prokhorenkova et al., 2017). 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. As the boosting algorithm trains weak learners iteratively, early stopping reduces training time and avoids overfitting. At every boost round, the model evaluates and decides whether to stop or continue the training when the model shows no more improvement for a certain number of consecutive rounds in terms of the evaluation metric specified as the fit parameter. For early stopping, we use the validation set and set the number of early stopping rounds to 10% of the maximum number of boosting iterations. CatBoost trained with 27 attributes selected by the importance of random forests attributes has the best performance with the best accuracy (78%), F1 (83%), and MCC (54%). We provide the performance details on the four gradient-boosting models in Figure 8 and Table 6. As you can see, the performance of the four gradient-boosting algorithms is similar to and does not exceed the vanilla gradient boost implementation, as the difference in accuracy between them is equal to or less than 1%. In conclusion, **the reduction in dimensionality does not** influence the machine learning models, and the gradient-boosting algorithms perform slightly better than the other baseline models. To improve the gradient-boosting models, we can penalize and regularize the algorithm by hyperparameter tuning so that we aim to increase accuracy and avoid overfitting. To begin with, constraining tree structures reduces the growth of complex and more extended 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, usually less than 0.5, allows weighting trees to slow the learning by a small amount at each iteration to reduce errors. Furthermore, setting the optimal L1 and L2 regularization terms penalizing the sum of the leave weights improves the models by simplifying the complexity and size of the model (Chen and Guestrin, 2016). These hyperparameters are searched with a 5-fold cross-validation randomized search with the number of iterations that is 20% of parameter distributions of each model. For example, XGBoost is supposed to explore 100 distributions of the parameters; the number of iterations for RandomizedSearch is 20 times.

Teacher Retention Prediction and Analysis

This experiment shows how skewed the training data is and whether the classification of teachers (1 if it stays, 0 if it leaves) is skewed by a high percentage of former teachers. Our labeled data (3,640 teachers) is small, and we label 2,176 as current and 1,464 as retired teachers. The attrition rate in the labeled data is 40%, much higher than the retention rate of under 10% in the USA. The exploratory data analysis of the labeled dataset shows the same characteristics as the exploratory data analysis of the dataset, which consisted of teachers who took the SASS survey but did not follow up with the TFS survey. The entries without principal and school associations were removed from the test set. We have fitted the best-performing models to the training dataset, and the result on the testing dataset shows that the CatBoost modeling using 27 attributes is the most robust model among advanced gradient boosting models with 78.3% accuracy and 83/2% F1 score. The breakdown of the results for the top 4 models is presented in Table 6 in the *Supplemental Material* Section.

The labeled data of 3,640 teachers becomes a training set, and our new unlabeled test set is a set of 33,198 teachers and their attributes. The dataset that contains attributes for teachers who took the SASS survey but did not follow up with the TFS survey also does not include information on the Teacher's marital status and the number of dependents. Thus, we exclude those attributes from the training dataset and fit the XGBoost model, which is the best gradient boosting model for the complete feature set of 51 characteristics in the training data set, with the best hyperparameters: *'n_estimators'*: 200, *'min_child_weight'*: 0, *'max_depth'*: 6, *'learning_rate'*: 0.2, *'lambda'*: 10, *'gamma'*: 0.1, *'alpha'*: 10. Next, we rank predictions in the test set. To account for the bias in the training set that favors former teachers, we raise the confidence in the model threshold to 0.8. The model predicts that 3,399 teachers from the unlabeled SASS data set have also left education (80%+ model confidence), that is, **10.24%** predicted attrition rate for the entire population that did not respond to the follow-up survey. The breakdown of the predictions shown in Figure 4 is aligned with our primary EDA for the labeled data: (i) female teachers are the majority; (ii) the turnover rate is higher for male teachers; (iii) white non-Hispanic teachers are the majority race/ethnicity group and have the highest attrition rate; (v) the highest attrition yearly rate is for teachers working more than three years and for teachers teaching STEM subjects.

The open-source data were have used for this AI/ML analysis at scale is 20+ years old. Yet, the predicted attrition rate for the entire population that did not respond to the follow-up survey is**10.24%**. The attrition rate in the U.S.A. today is around the same percentage. Our open-source data analysis supports the claim that the attrition rate among U.S.A. teachers has been consistent and has not increased in this decade. Attrition rate modeling that predicts consistency in historical data is the most exciting finding of the entire study, as it points to the need for the data-driven support of qualitative claims at a larger scale in educational research.

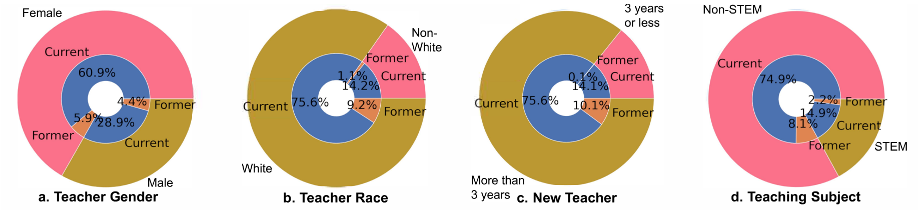


Figure 4. Teacher attrition prediction analysis per school and principal attributes.

Conclusion and Future Work

This study leverages open-source historical data to elucidate the most significant factors contributing to teacher attrition in the USA. By employing a multi-view feature importance analysis, we provide a comprehensive assessment of the intrinsic relationships and patterns within educational datasets. Our findings consistently identify the willingness of teachers to participate in surveys, their years of teaching experience in public schools, and the elementary level of their schools as the predominant factors influencing attrition. Our application of gradient-boosting models to raw input data demonstrates a marked improvement in predicting school-level teacher attrition, particularly for unlabeled data. Notably, enhancements through data alignment and imputation offer limited additional benefit within this analytical framework. The attrition rate for teachers who completed the Schools and Staffing Survey (SASS) but not the Teacher Follow-Up Survey (TFS) exceeds 10%, which is consistent with current national attrition rates. Our experimental procedures are fully reproducible and accessible for further scrutiny. Future work will involve expanding the dataset to incorporate additional years of public-use SASS and TFS data, as well as restricted-use data from specific years and the National Teacher and Principal Survey (NTPS). This broader dataset will enable us to validate our methodology and extend our analysis to global educational contexts. The ultimate goal is to equip policymakers with the tools to allocate resources more effectively to schools and teachers at high risk of attrition.

In summary, this research offers critical insights into the factors driving teacher attrition and introduces a robust analytical framework. Our work advances the understanding of teacher retention dynamics and lays the groundwork for future explorations aimed at enhancing educational stability.

References

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

Abutbul, A., Elidan, G., Katzir, L., and El-Yaniv, R. (2020). Dnf-net: A neural architecture for tabular data. *CoRR*, abs/2006.06465.

Arik, S. ö. and Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6679–6687.

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

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

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

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

Carlsson, R., Lindqvist, P., and Nordänger, U. K. (2019). Is teacher attrition a poor estimate of the value of teacher education? a swedish case. *European Journal of Teacher Education*, 42(2):243–257.

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

Casely-Hayford, J., Björklund, C., Bergström, G., Lindqvist, P., and Kwak, L. (2022). 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.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Information Fusion*, pages 785–794.

Dill, K. (2022). School's out for summer and many teachers are calling it quits. Technical report, The Wall Street Journal.

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53(284):789–798.

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

Greufe, M. (2020). *Evaluating Teacher Turnover Rates in America, Canada, and Finland*. Honor's undergraduate thesis, University of Nebraska-Lincoln.

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

Gunn, T. M. and McRae, P. A. (2021). Better understanding the professional and personal factors that influence beginning teacher retention in one canadian province. *International Journal of Educational Research Open*, 2:100073.

Guolin Ke, Qi Meng, T. F. e. a. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *NIPS'17*, pages 3149–3157.

Han, E. (2022). The gendered effects of teachers' unions on teacher attrition: Evidence from district-teacher matched data in the us. *Feminist Economics*, pages 1–33.

Hooker, G. and Mentch, L. (2019). Please stop permuting features: An explanation and alternatives. *arXiv e-prints*, pages arXiv–1905.

Joseph, M. (2021). Pytorch tabular: A framework for deep learning with tabular data.

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

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

National Center, E. S. (2001). 1999-2000 sass public-use data and documentation & 2000-01 tfs public-use data and documentation. `https://nces.ed.gov/surveys/sass/dataprod9901.asp`.

National Center, E. S. (2022). The national center for education statistics (nces). `https://nces.ed.gov`.

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

OECD (2021). *Education at a Glance 2021*. Organisation for Economic Co-operation and Development.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pham, L. D., Nguyen, T. D., and Springer, M. G. (2021). Teacher merit pay: A meta-analysis. *American Educational Research Journal*, 58(3):527–566.

Popov, S., Morozov, S., and Babenko, A. (2019). Neural oblivious decision ensembles for deep learning on tabular data. *CoRR*, abs/1909.06312.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2017). Catboost: unbiased boosting with categorical features.

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

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

Saha, P., Patikar, S., and Neogy, S. (2020). A correlation - sequential forward selection based feature selection method for healthcare data analysis. In

2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), pages 69–72.

Shrestha, R. K. (2022). Teacher retention in private schools of nepal: A case from bhaktapur district. *KMC Journal*, 4(2):167–183.

Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

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

Sorensen, L. C. and Ladd, H. F. (2020). The hidden costs of teacher turnover. *AERA Open*, 6(1).

Speiser, J. L., Miller, M. E., Tooze, J., and Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, 134:93–101.

UNESCO (2017). Global education monitoring report 2017/8: Accountability in education–meeting our commitments.

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

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

Supplemental material

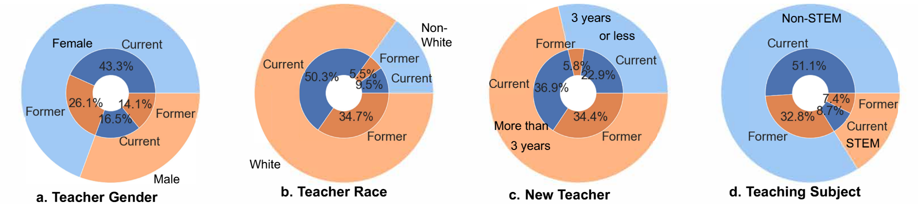


Figure 5. SASS and TFS Exploratory Retention Analysis for (a) gender, (b) race/ethnicity, (c) new Teacher, and (d) teaching field.

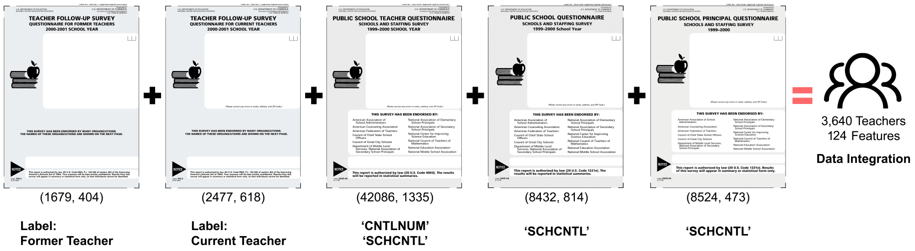


Figure 6. NCES Data from five sources for teachers, schools, and districts is aggregated for the 3640 teachers with over 124 attributes that have the outcome labels 1(stayed) and 0 (left).

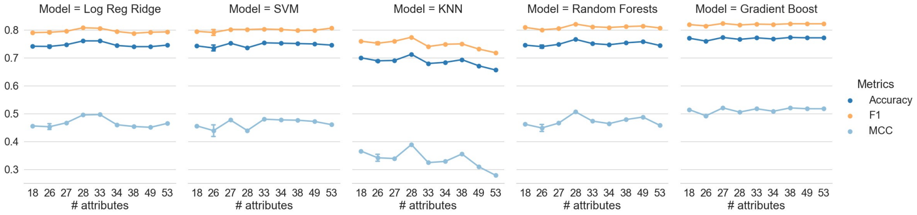


Figure 7. Five machine learning models fitted to the training and test sets with 10-fold cross-validation of hyper-parameter grid search. Test set accuracy, F1, and MCC results show stable performance for all models except KNN.

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 2. Selected 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.

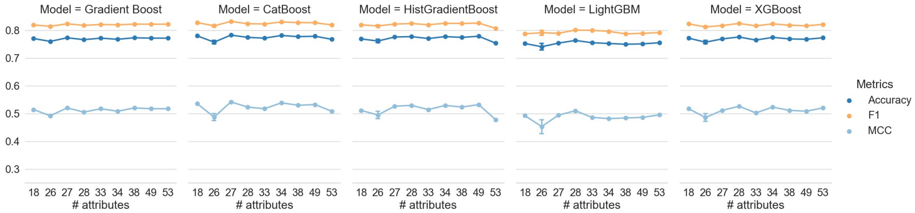


Figure 8. Five gradient boosting models fitted to the training and test sets with 5-fold cross-validation of randomized search. We show that the dimensionality reduction plays no role in the performance of the models w.r.t accuracy, F1, and MCC.

Table 3. Example of aggregated attributes filtered by correlations.

New Label	From Label Set	New Label	From Label Set
teaches _Xth $X \in [7, 12]$	teaches_7to12 Teaching 7 to 12th grades (1 0)	deg_P_Y $Y \in$ [Ass, Bach, Masters, Edu, Ph.D.] Principal's highest degree (5 categories)	deg_highest_P
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)	urbanicity_LargeCity urbanicity_SmallTown urbanicity_MidCity	urbanicity: Urbanic Locale
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 in salary (binary)	incen_STEM_comp_sci incen_STEM_math incen_STEM_phys_sci incen_STEM_biology	incen_STEM: Pay incentives in STEM fields (binary)

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)		

Table 4. Selected Teacher Attributes in the SASS dataset. Value (1 0): If the statement is true, the attribute value is 1. Otherwise it is 0.

Table 5. Out of the five state-of-the-art machine learning models, gradient boosting training, 27 attributes perform the best.

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

Table 6. Performance measure of 4 different models predicting teacher attrition

Model	Best	Selection Method	Acc [0,1]	F1 [0,1]	MCC [-1,+1]
CatBoost	27	RF FE	0.783	0.832	0.543
HistGradientBoost	49	RFE RF	0.779	0.826	0.533
LightGBM	28	PFI RF	0.764	0.801	0.51