Data Driven Teacher Attrition Modeling

June Yu, Li Feng, Jelena Tešić

[a] *Texas State University, San Marcos, 78666, TX, USA*

Abstract

Teacher attrition in public schools has reached a critical juncture, with many educators leaving the profession. To address this pressing issue, we conducted a large-scale analysis of public data from the National Center for Education Statistics (NCES) to provide data-driven insights into teacher attrition challenges. We developed an open-source end-to-end educational data modeling pipeline tailored for large-scale analysis to examine interpretable teacher attrition; and adapted state-of-art AI/ML approaches to model the survey data for two tasks: (i) identify the essential factors for teacher attrition using multi-view feature importance analysis and (ii) derive a reliable predictive model that outcomes the probability that the teacher will leave their position in the next year. For the first task, we discovered that the race and sex of the principal, the type of school, and the school's location impact teacher retention rates the most. For the second task, we observed that modeling historical data resulted in a predicted attrition rate of over 10%, aligning closely with the current prevalent attrition rates in the USA. This finding implies a concerning persistence of attrition rates over the past several decades, despite the various changes in the educational landscape. These findings enable policymakers to make data-driven decisions.

Preprint submitted to Information Processing & Management *July 3, 2023*

1. Introduction

Teacher attrition is a pressing issue in education, with significant implications for the quality of education and the well-being of students. Defined as the percentage of teachers leaving the profession within a given school year, teacher attrition rates play a crucial role in shaping the effectiveness of public schools worldwide. While a moderate turnover rate of 6% to 8% is considered natural and desirable, low and excessively high attrition rates can adversely affect educational outcomes [1]. The impact of teacher attrition becomes evident when we examine its consequences. A school with an attrition rate below 5% will likely stagnate, needing more fresh perspectives and ideas from new educators. On the other hand, when attrition rates exceed 10%, the detrimental effects on a public school's effectiveness become increasingly pronounced. Therefore, addressing the factors contributing to teacher attrition and seeking strategies to mitigate its negative consequences is crucial. 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 [2]. 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 [3]. The COVID-19 pandemic has exacerbated the problem of teacher attrition in K-12 education worldwide [4]. 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 [5]. 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 [5].

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 [6]. 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 [7, 8]. 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. Through the implementation of a large-scale educational data modeling pipeline, we integrate, clean, and analyze educational data. Additionally, we employ automated attribute importance analysis to identify meaningful conclusions and develop a suite of interpretable teacher retention prediction models that utilize open-source data. In this article, we present our research findings and recommend next steps. Specifically, Section 3 provides background information on public education data in the United States and outlines the exploratory data analysis conducted. Section 4 introduces automated approaches to identify the most relevant attributes for teacher retention. Subsequently, Section 5 summarizes the state-of-the-art modeling comparison and presents the results of our experiments. Finally, in Section 6, we conclude the article by summarizing our findings and offering insights into future directions for addressing the challenges of teacher hiring and retention.

2. 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 [9]. 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 [9]. 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 [10]. 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 [11]. Demographic, academic, family/personal, and internal assessment attributes were commonly employed to predict student performance across various contexts [12]. 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) [13]. 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 [14]. 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 [15, 16, 17]. 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 [18, 19, 20]. However, there is no consensus on whether deep learning surpasses GBDT in tabular data, as standard benchmarks and open-source implementations have been limited [21, 22]. 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 [21, 23, 24]. The field of education economics has extensively analyzed teacher turnover, attrition, retention, and recruitment on a global scale [2]. 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 [25, 26, 27, 28, 29, 30, 31, 32, 33]. However, no principal data-driven study has comprehensively identified the attributes that explain teacher attrition.

3. 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

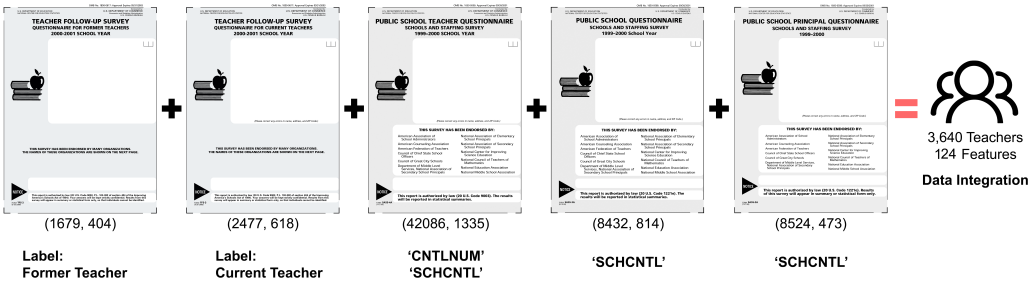

Figure 1: NCES Data from five sources for teachers, schools, and districts is aggregated for the 3640 teachers with over 124 attributes that have the outcome labels 1(stayed) and 0 (left).

research community to inform policy and practice [34]. 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 elementary and secondary education aspects. 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 [35]. 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 [35].

For this research, we analyze the available data from the 1999-2000 SASS and 2000-2001 TFS, explicitly focusing on public schools, teachers, and principals [35]. The TFS data provide binary labels indicating whether teachers decided to continue teaching (labeled as 1) or leave the profession (labeled

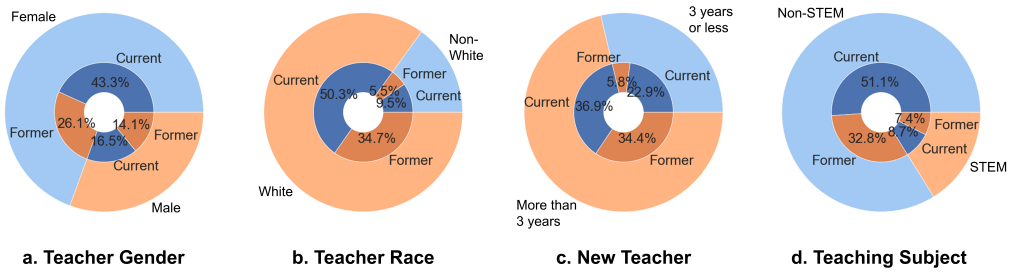

Figure 2: SASS and TFS Exploratory Retention Analysis for (a) gender, (b) race/ethnicity, (c) new teacher, and (d) teaching field.

as 0). Figure 3 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. Data analysis, as shown in Figure 2, reveals several interesting findings: (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.

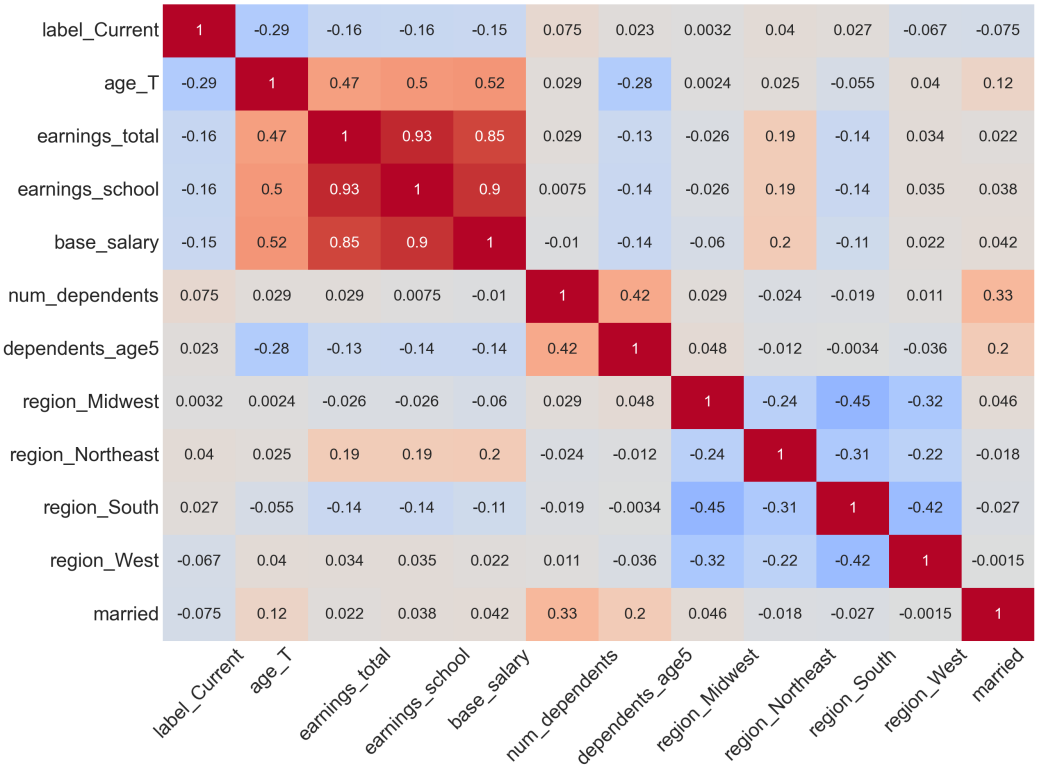

Figure 3: Correlation of SASS attributes with the TFS attributes.

4. Attribute Importance Scoring

4.1. Attribute Filtering by Mutual Correlations

This section presents a unique and easily interpretable suite of approaches for analyzing attribute importance. Our goal is 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 [35]. To address these challenges and enhance the inter-

pretability of our models, we employ a filter method that identifies correlated attributes. This filtering process allows us to construct a quasi-orthonormal attribute space, enabling us to observe correlations between different features 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 $\bar{x}$ 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.

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

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

Table 1: Example of aggregated attributes filtered by correlations.

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

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. The correlation coefficients of the expanded data are illustrated in Figure 3, e.g. *base_salary* is highly correlated with *earnings_school* and *earnings_total* attributes. With the correlation coefficients of our data, we combined all binary variables if they can be related categories as a set, as summarized in Table 1. Finally, the dimensionality of the data set has been reduced to 53 attributes, with 39 categorical and 14 numerical.

4.2. Multi-view Attribute Importance Analysis

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 features 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 important features 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

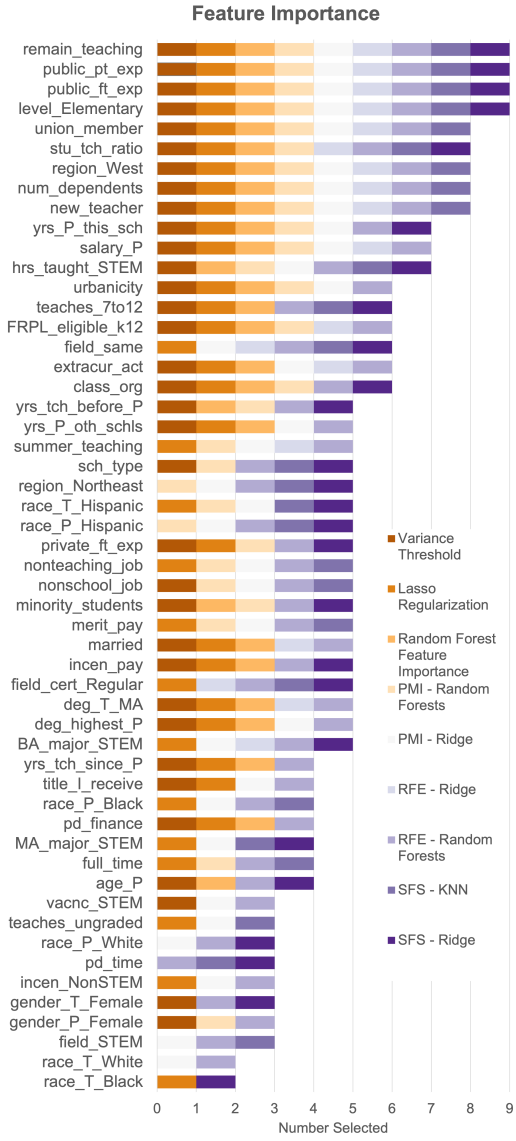

Figure 4: All nine methods select the 4 features *remain_teaching, public_pt_exp, public_ft_exp, level_Elementary* as the most important features.

in the data. This allows us to address data quality problems before building the model, ensuring better performance and reliability.

In this paper, we propose 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* allows us to evaluate data quality problems. The method removes attributes with low variance by applying a threshold, such as 0.8*(1-0.8), to the training dataset. Attributes with a similarity of 80% or more are eliminated, and the top-k attributes with the highest variance are selected for the final set [36]. *Random Forests* classification and regularization machine learning algorithm provide attribute importance measures through the Gini importance or mean decrease impurity. 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 [37].

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 [38]. *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. Features 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 [39]. *Sequential Feature Selection (SFS)*: SFS sequentially selects an optimal set of features 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 is set to select half of the provided attributes for the final set [40].

Six distinct approaches produce nine total scorecards on feature importance, as the last three approaches implement two measures per method to illustrate the difference and sensitivity. Table 2 presents the number of attributes each approach selects. Among them, RFE with ridge regression resulted in the smallest set, consisting of 18 attributes, while RFE with random forests produced the largest set, with 49 attributes, as illustrated in Figure 4. 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 4. Eight methods agree on the next five most impactful attributes (Figure 4), 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 4.

Table 2: Nine approaches selected the number of features, and the selection is illustrated in Figure 4 with distinguished bar colors marked in the Color column.

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

Figure 5 indicates the main attributes of 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 before becoming a principal), *num_dependents* (number of dependent teachers), *age_P* (age of a principal), *new_teacher* (teachers who teach 3 years or less), *level_Elementary* (teachers teaching in an elementary school), and *hrs_taught_STEM* (hours of teaching STEM subjects per week) are the only eight overlapping highly impactful attributes. Vanilla Random Forest has 27 features 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

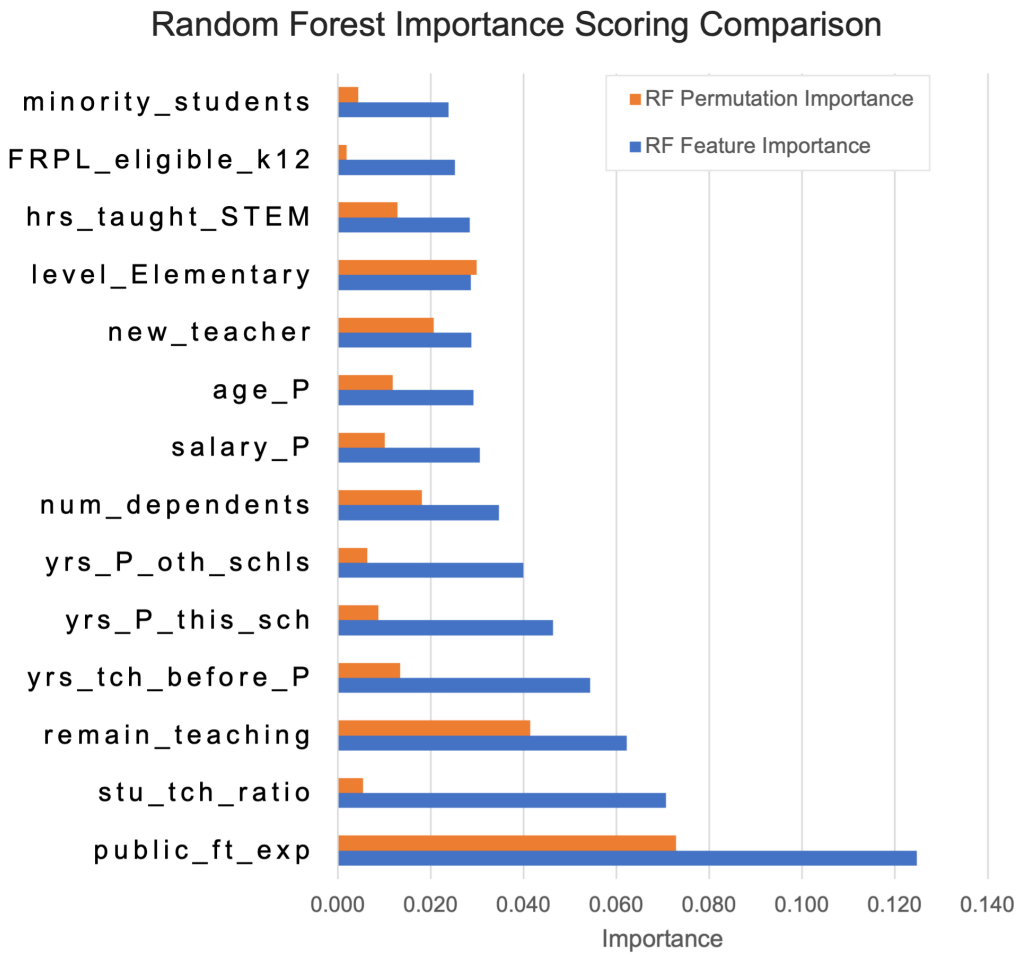

Figure 5: Random Forest Feature Importance and Permutation Attribute ranking comparison

experience in public schools. Specifically, since teachers work longer years as full-time teachers in public schools, we can better predict teacher retention.

5. Analysis and Prediction Modeling of Teacher Attrition

5.1. 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. 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.

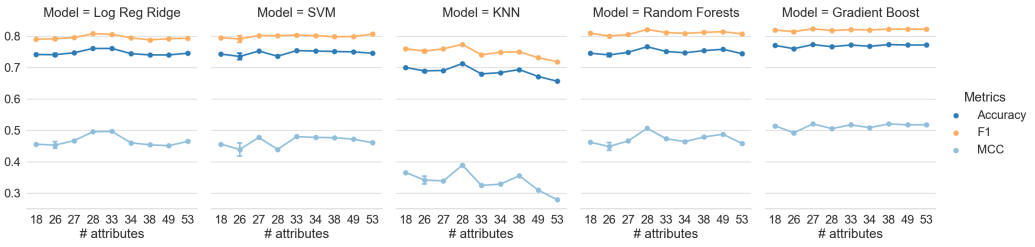

Figure 6: Five machine learning models fitted to the training and test sets with 10-fold cross-validation of hyper-parameter grid search. Test set accuracy, F1, and MCC results show stable performance for all models except KNN.

We randomly split the data into a training set (2,192 teacher instances, 80%) and a test set (728 teacher instances, 20%) with shuffling and stratification on the label. The feature reduction methods produced 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 performance of the five state-of-the-art models in the test set, organized by the number of attributes, is illustrated in Figure 6. The ensemble models

Table 3: Best model of the five state-of-the-art machine learning models is gradient boosting training 27 features.

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

based on decision trees, gradient boost, and random forests training 27 and 28 attributes selected by the importance of random forest characteristics and PMI with random forests, respectively, are the model 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 modeling pipeline was repeated for the advanced gradient boosting models, XGBoost, LightGBM, CatBoost, and HistGradientBoosting. 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 defining a default direction of the tree split in each tree node [15]. 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 features with non-zero values to reduce the number of columns [16].

Handling categorical features. Handling categorical features is challenging in building a machine-learning model for tabular data. While there are several ways to process representing categorical features, such as one-hot and ordinal encoding, tree building and 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 features to optimize numerous boost steps for computing time and memory consumption. LightGBM, HistGradientBoosting, and XGBoost use the optimal split method [41] 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[42]. 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 [43]. All four models accept hyperparameters to handle categorical features, such as categorical feature indices or thresholds to control one-hot encoding or the number of tree split points.

Early stopping rounds. As the 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, a validation set, the split test set at the beginning of the modeling process, and the number of early stopping rounds set to 10% of the maximum number of boosting iterations are provided.

Next, we compared the four gradient boosting models with our best-performing baseline models, gradient boosting. Although the boosting models remain stable for all sets of attributes regarding their test precision, F1 and MCC, as shown in Figure 7, the most robust boosting model performs best is CatBoost trained 27 features selected by the importance of random forests features with the best accuracy (78%), F1 (83%) and MCC (54%). Furthermore, the performance of the four gradient-boosting algorithms is similar to and not exceeding 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.

Table 4: CatBoost fitting 27 features is the most robust model among Advanced gradient boosting models.

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

Hyperparameter optimization. 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 [15]. These hyperparameters are searched with a 5-fold cross-validation RandomizedSearch with the number of iterations that is 20% of parameter distributions of each model. For example, XGBoost is supposed to explore 100 distributions of the parameters; the number of iterations for RandomizedSearch is 20 times.

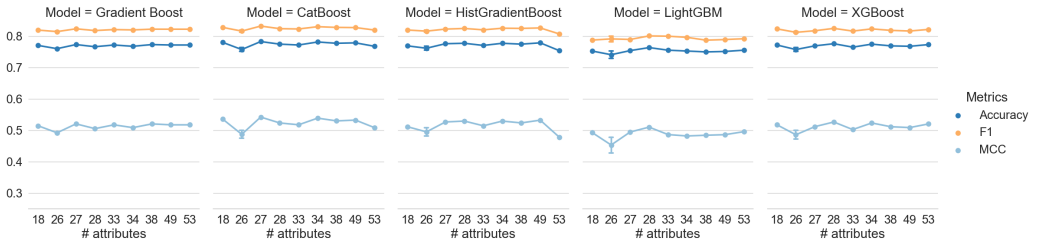

Figure 7: Five gradient boosting models fitted to the training and test sets with 5-fold cross-validation of RandomizedSearch. We show that the dimensionality reduction plays no role in the performance of the models w.r.t accuracy, F1, and MCC. *TODO: make graphs larger and in 2 rows – June*

5.2. Teacher Retention Prediction and Analysis

In this section, we evaluate how skewed the training data is and does 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 labels 2,176 as current and 1,464 as former 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 consisting

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. 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 that 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 features 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.

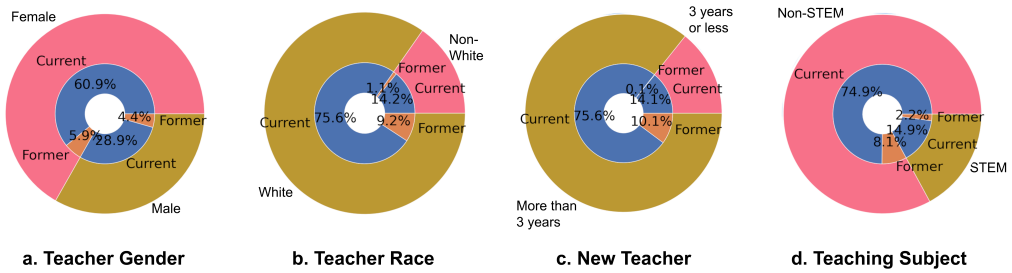

Figure 8: Teacher attrition prediction analysis per school and principal attributes.

As a result, our 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. This result aligns with the teacher attrition

rate in the USA, demonstrating that our training data are not skewed. The breakdown of the predictions shown in Figure 8 is aligned with our primary EDA (Figure 2) 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.

6. Conclusion and Future Work

This paper utilizes open-source historical data to model the most impactful attributes of teacher attrition in the USA, and introduces the multi-view feature importance analysis for robust assessment of the intrinsic connections and patterns in the education data. The results show a consensus among these methods, highlighting the teacher's willingness to respond to a survey question, years of teaching experience in public schools, and the school's elementary level as the most influential attributes. Furthermore, we demonstrate that gradient-boosting models applied to raw input data yield superior performance in predicting teacher attrition at the school level for unlabeled data. Interestingly, data alignment and imputation do not significantly improve the modeling performance within this framework. The predicted attrition rate for teachers who completed the SASS survey but not the TFS follow-up survey surpasses 10%, aligning with current attrition rates in the USA. Our experiments are reproducible and available for reference. Moving forward, our next step involves expanding the dataset to include public-use SASS and TFS data from other years, as well as restricted-

use data such as SASS and TFS (2003-2004, 2007-2008, and 2011-2012), and National Teacher and Principal Survey (NTPS) data (2015-2016 and 2017-2018). This expansion will allow us to validate our pipeline and explore open-source educational data worldwide, enabling policymakers to allocate resources effectively to schools and teachers at high risk of leaving the system. In conclusion, this research sheds light on the influential factors of teacher attrition and presents a robust data analysis and modeling pipeline. It contributes valuable insights into the understanding of teacher attrition and provides a framework for future investigations in this field.

References

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

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

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

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

[5] K. Dill, School's Out for Summer and Many Teachers Are Calling It Quits, Technical Report, The Wall Street Journal, 2022.

[6] L. C. Sorensen, H. F. Ladd, The hidden costs of teacher turnover, AERA Open 6 (2020). doi:10.1177/2332858420905812.

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

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

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

[10] A. R. Rao, Y. Desai, K. Mishra, Data science education through education data: an end-to-end perspective, in: 2019 IEEE Integrated STEM Education Conference (ISEC), 2019, pp. 300–307. doi:10.1109/ISECon.2019.8881970.

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

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

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

[14] K. Yan, Student performance prediction using xgboost method from a macro perspective, in: 2021 2nd International Conference on Computing and Data Science (CDS), 2021, pp. 453–459. doi:10.1109/CDS52072.2021.00084.

[15] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, Information Fusion (2016) 785–794. doi:https://arxiv.org/abs/1603.02754.

[16] T. F. e. a. Guolin Ke, Qi Meng, Lightgbm: A highly efficient gradient boosting decision tree, NIPS'17 (2017) 3149–3157. doi:10.5555/3294996.3295074.

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

[18] S. ö. Arik, T. Pfister, Tabnet: Attentive interpretable tabular learning, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 6679–6687. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16826. doi:10.1609/aaai.v35i8.16826.

[19] A. Abutbul, G. Elidan, L. Katzir, R. El-Yaniv, Dnf-net: A neural architecture for tabular data, CoRR abs/2006.06465 (2020). URL: https://arxiv.org/abs/2006.06465. arXiv:2006.06465.

[20] S. Popov, S. Morozov, A. Babenko, Neural oblivious decision ensembles for deep learning on tabular data, CoRR abs/1909.06312 (2019). URL: http://arxiv.org/abs/1909.06312. arXiv:1909.06312.

[21] R. Shwartz-Ziv, A. Armon, Tabular data: Deep learning is not all you need, Information Fusion 81 (2022) 84–90. doi:10.1016/j.inffus.2021.11.011.

[22] M. Joseph, Pytorch tabular: A framework for deep learning with tabular data, 2021. URL: https://arxiv.org/abs/2104.13638. doi:10.48550/ARXIV.2104.13638.

[23] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, G. Kasneci, Deep neural networks and tabular data: A survey, 2021. URL: https://arxiv.org/abs/2110.01889. doi:10.48550/ARXIV.2110.01889.

[24] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on tabular data?, 2022. URL: https://arxiv.org/abs/2207.08815. doi:10.48550/ARXIV.2207.08815.

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

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

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

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

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

[30] E. Han, The gendered effects of teachers' unions on teacher attrition: Evidence from district-teacher matched data in the us, Feminist Economics (2022) 1–33. doi:10.1080/13545701.2022.2105375.

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

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

[33] R. K. Shrestha, Teacher retention in private schools of nepal: A case from bhaktapur district, KMC Journal 4 (2022) 167–183. doi:10.3126/kmcj.v4i2.47776.

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

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

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

[37] J. L. Speiser, M. E. Miller, J. Tooze, E. Ip, A comparison of random

forest variable selection methods for classification prediction modeling, Expert systems with applications 134 (2019) 93–101.

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

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

[40] P. Saha, S. Patikar, S. Neogy, 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), 2020, pp. 69–72. doi:10.1109/GUCON48875.2020.9231205.

[41] W. D. Fisher, On grouping for maximum homogeneity, Journal of the American Statistical Association 53 (1958) 789–798. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501479. doi:10.1080/01621459.1958.10501479. arXiv:https://www.tandfonline.com/doi/pdf/10.1080/01621459.1958.10501479.

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

[43] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, Catboost: unbiased boosting with categorical fea-

tures, 2017. URL: https://arxiv.org/abs/1706.09516. doi:10.48550/ARXIV.1706.09516.

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

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 6: 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.