Psychological Test and Assessment Modeling, Volume 64, 2022 (4), 445-470

Considerations in Using XGBoost Models with SHAP Credit Assignment to Calculate Student Growth Percentiles

Steven Tang[1] *& Zhen Li*

Abstract

The wealth of student data collected in education enables machine learning to be a promising option to provide further insight into predicting important outcomes in a student's education as machine learning approaches can handle increased data sources and data volume. As a prominent machine learning approach, Gradient Boosted Models (GBMs) have been shown to be a potential alternative methodology in place of the commonly used quantile-regression (QR) based procedure to estimate student growth percentiles (SGP). This study discusses aspects of using GBMs in computing growth percentiles by 1) illustrating the effects of different hyperparameters on model fit, 2) comparing GBM and QR-based SGP agreement across different sets of predictors, 3) using an interpretability method, SHAP (SHapley Additive exPlanations), to show the impact of each predictor on the predictions of the GBM model, and 4) analyzing the effect of sample size on GBM prediction accuracy. The dataset in this study comes from math tests for grades 3 to 8 across 4 years of a state summative assessment

Keywords: Gradient boosted models, student growth percentiles, SHAP

[1] eMetric *Correspondence concerning this article should be addressed to:* Steven Tang, eMetric, 211 N Loop 1604 E Suite 170 San Antonio, TX 78232, USA. steven@emetric.net

Introduction

In recent years, highly effective big data models have shown impressive results in modeling tabular data of many dimensions. One class of these models is called Gradient Boosted Models (GBM). GBMs do not make strong assumptions about distributions, can model multiple predictors, are robust to data irregularities including missing data, and are computationally very fast.

The Student Growth Percentiles (SGP) approach is currently one of the most popular methods for calculating growth scores in K-12 education (Betebenner, 2008). A student's growth in academic performance is defined as the percentile rank of this student in the conditional distribution of all students with the same score on the previous grade or administration. SGP have been used for measuring students' annual growth in many states. SGP are commonly computed using a procedure based on quantile regression (QR) in an R package: "SGP" (Betebenner, Van Iwaarden, Domingue, & Shang, 2022; R Core Team, 2022). This package computes growth percentiles using a computationally demanding procedure. It requires the estimation of 100 B-spline based quantile regression lines and a search of the thresholds on each percentile regression line for every combination of the previous years' test scores. Students' growth percentiles are obtained by comparing their current-year scale scores to these thresholds among the regression lines. Additionally, SGP estimates may fluctuate as the B-spline function changes, such as increasing or decreasing the polynomial degree in the B-spline parameterization (Castellano K. E., 2011). In the past decade, quite a few studies have investigated the accuracy and reliability of SGP calculation (Castellano K. E., 2011; Monroe & Cai, 2015). As an analog to the quantile regression approach, an ordinary least squares (OLS) regression approach to calculate the growth percentiles, known as the percentile rank of residuals (PRRs), was found to perform better than SGP in many situations (Castellano K. E., 2011). PRR was found to be more similar to the gold standard, the empirical conditional percentile ranks, in multivariate normal distributed data. PRR was also more computationally efficient than SGP and more robust for small and sparse samples. This PRR approach can calculate growth percentiles accurately when the assumptions of linear regression models are not violated.

In educational measurement, even though many problems are still addressed using traditional linear regression methods, interest in leveraging newer methods like GBM remains high, given that traditional methods may not be able to explain complex interactions and may also not be appropriate when the assumption of linearity between the predictors and the outcome is not valid (Sinharay, 2016). The method "**SGP(gbm)**" has been proposed as a GBM based method (Tang & Li, 2019) to compute student growth percentiles. SGP(gbm) has been shown to have good computational speed, good prediction accuracy, and a built-in potential to add additional covariates to the prediction model without requiring assumptions of independence or assumptions about distributions. GBMs are well known for their accurate performance, but do not have the inherent interpretability that simpler models might have. SHAP (**SH**apley **A**dditive ex**P**lanations) were developed for the purpose of explaining

the output of machine learning models such as GBMs (Lundberg & Lee, 2017). Using SHAP values, users can better interpret model results from models like GBMs.

This paper consists of 4 sub-studies. Study 1 provides details about optimizing XGBoost, a popular implementation of the GBM approach, for growth percentiles. Parameter optimization is a crucial step in deploying GBMs. Study 2 compares SGP(gbm) to QR-based SGP or SGP(QR) results, varying aspects such as the number of predictors and covariates. QR-based SGP is chosen as the baseline because it is usually considered as the rule-of-thumb for SGP calculation in states' accountability systems. Study 3 illustrates how SGP(gbm) results can be interpreted by assigning a contribution value to each predictor in the model. This illustration is possible by utilizing the SHAP interpretation package. The ability to interpret complex machine learning models can greatly enhance practical use cases based on GBM results. Study 4 investigates the effect of sample size on SGP(gbm) prediction accuracy, providing insight into minimum sample sizes needed to perform SGP computation. Together, the 4 studies provide an overview of practical considerations in applying SGP(gbm) towards computing growth percentiles, as well as considerations in applying interpretation packages such as SHAP on GBM results.

Methodology

The XGBoost Model

In recent years, XGBoost, short for eXtreme Gradient Boosting (Chen & Guestrin, 2016), has proven to be an increasingly popular method for a variety of predictive modeling tasks, especially of tabular data. XGBoost is an open-source implementation of GBM, known to be efficient, highly scalable, and easy to implement out-of-the-box. XGBoost also handles missing data innately by treating missing values as its own category for predictors. XGBoost models have been found to produce more accurate predictions with less sensitivity to the preprocessing of features compared to generalized linear models (Benjamin, et al., 2017).

XGBoost iteratively builds a collection of simple regression trees; regression trees are decision trees that predict continuous outcomes. The term boosting**,** in machine learning terms, generally refers to a broad class of models that attempt to create a "strong" model from many "weak" models. Weak is a somewhat subjective term, but essentially refers to models that are generally very simple.

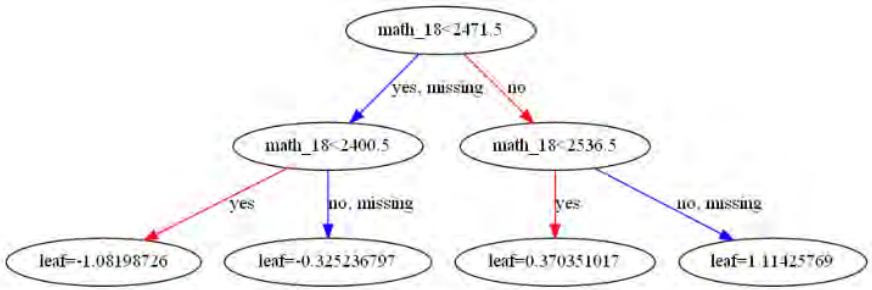

Figure 1.

An example of a simple regression tree in the boosted tree model.

Figure 1 shows an example of a weak regression tree. The predictor is students' 2017-18 math scale scores (math_18) from a state summative test, and the predicted variable is their 2018-19 math scale scores. The depth of this tree is 2. A decision tree starts at the top of the tree. As an example, assume there is a math_18 score of 2420. Looking at the top of the tree in Figure 1, the value of 2420 is compared to 2471.5. Since 2420 is less than 2741.5, the "yes" path is followed since that was the condition in the first tree node. Then, since 2420 does not meet the next condition in the next node, the "no" path is followed. Finally, there are no more conditions. The end of a path is known as a "leaf" in tree model terminology. Since a leaf has been reached, this is the final model output for this simple tree. The output value is -0.325236797. The XGBoost model will combine the output from multiple simple models to reach the final output; combining the outputs is as simple as summing up all outputs from the full collection of weak or simple models and adding it to the mean value of the predicted data.

Weak regression trees are constructed by searching through many potential split values among all input variables and finding the splits that minimize prediction error. After one tree is constructed, the XGBoost algorithm continues to build an additional tree. This additional tree is of the same structure as all previous trees, but the latest tree is tasked with minimizing the residual errors of all previous regression trees. The process of iteratively creating new trees that minimize the residual error of the model thus far continues until a stopping criterion is met. XGBoost uses a gradient descent algorithm to minimize the loss when adding new models.

Computation of SGP(gbm)

The calculation of SGP(gbm) consists of several steps: 1) Data preparation. XGBoost requires numerical values, so non-numerical variables need to be converted to numerical variables. For this study, students are removed from the analysis if they do not have complete data in both the predictor and the outcome data. The requirement of full data is to maintain comparability when using different sets of predictors as well as comparing SGP(gbm) results to QR-based SGP. 2) Hyperparameter tuning. Hyperparameters impact the learning process and affect the prediction accuracy of XGBoost. Tuning these parameters involves testing model performance using different values for each hyperparameter and searching for the best fitting model. Tuning involves the use of training and test datasets. The XGBoost package has a default set of hyperparameter values that can generally work well on many datasets. 3) Final model training and prediction. All students are included in the training of the final prediction model. The reason for this choice has two folds. First, the model hyperparameters have been tuned in step 2 to avoid model overfitting or underfitting. Second, each observation should be equally influential in the training process, as the predicted score will be calculated for all students. This prediction model will be used to generate students' predicted scores.

After a prediction model is tuned, growth percentiles can be computed using the metric of Percentile Rank of Residual (PRR, Castellano K. E., 2011). The residual for student *i* will be defined as the model-predicted "expected score" minus the observed score:

$$residual_i = expected\_score_i - observed\_score_i. \quad (1)$$

For student *i*, PRR_i is the percentage of residual values smaller than or equal to the residual value for student *i*. If more than one student has the same residual value, PRR_i takes the average of their ranks.

$$PRR_i = \frac{\#\, residual \leq residual_i}{N} \times 100. \quad (2)$$

PRR_i is regarded as the student growth percentile in the SGP(gbm) methodology. A high PRR_i means that student *i* score higher than expected given his/her previous academic achievements. SGP(gbm) and QR-based SGP are two different approaches to estimating the conditional percentiles of students' current-year scores among students with the same prior scores. The empirical conditional percentiles are difficult to calculate because the number of conditional groups can be large, and many groups may have sparsity issues. Both SGP(gbm) and QR-based SGP are located towards the nonparametric end on a parametric continuum. SGP(gbm) generally makes no assumptions on the input data, while QR-based SGP utilizes a B-spline parameterization for the quantile regression lines.

SHAP Values

The growing availability of big data has led to a rise in the use of complex models. Consequently, more attention has been paid to being able to interpret the outputs of complex models. In 2017, SHAP (Lundberg & Lee, 2017) was introduced as a unified framework for interpreting predictions. One of the SHAP package's main contributions is the ability to compute SHAP values for each input feature or predictor of a model. SHAP values adhere to the property of "additive feature contribution." This means that the SHAP values of the input features for an individual sample sum up to the model's prediction for that individual sample. Since SHAP values sum up to a model's prediction, SHAP values are on the same scale as the outcome variable, allowing for direct interpretability of a SHAP value. Technically, SHAP algorithms compute an approximation of the Shapley value, a well-known and theoretically sound concept to interpret black box machine learning models (Wagner, 2022).

SHAP values are computed using one sample or one individual at a time. For example, if there are 20,000 students, then SHAP values will be computed individually for all 20,000 samples. The SHAP formulation is as follows:

$$\phi_{it}(f,x) = \sum_{R\in\mathcal{R}} \frac{1}{M!}[f_x(P_t^R \cup t) - f_x(P_t^R)]. \quad (3)$$

In Equation 3, $\phi_{it}(f,x)$ is the SHAP value for student *i* and feature *t*, *f* is the prediction model, *x* is the input vector for the current prediction, $\mathcal{R}$ is the set of all feature orderings, P_t^R is the set of all features that come before feature *t* in ordering *R*, and *M* is the number of input features for the model. f_x is the conditional expectation function of the model's output. Given the prediction model *f* and input vector x, the SHAP value for each feature in a prediction model can be calculated by iteratively computing the conditional expectation f_x for all subsets of features with or without the target feature *t* as shown in Equation 3. The calculation of SHAP values for tree models can be simplified by utilizing TreeExplainer (Lundberg, et al., 2020), which has the advantage of calculating SHAP values in polynomial time. The SHAP value for a boosted tree model is the sum of SHAP values on each tree. For a single tree, the algorithm finds all the subsets of features on each node and their corresponding weights. $\phi_i(f,x)$ is then obtained by keeping track of these subsets and approximately calculating f_x using the weights and predictions for each node. For an intelligible yet thorough explanation of how the SHAP algorithm approximates Shapley values for tree models, interested readers can refer to Lundberg et al. (2020).

In the current study, the TreeExplainer method is used to compute SHAP values for evaluating the importance of predictors in XGBoost models. For a specific input feature, the absolute SHAP values computed from all samples are then averaged; this "mean absolute SHAP value" is one way to determine the overall contribution of a particular input feature across all samples.

Data

This study utilizes test scores from grades 3 through 8 of a state's summative tests and students' demographics from four consecutive school years, namely, 2015-16, 2016-17, 2017-18, 2018-19. The subjects include ELA and Mathematics. Missing values exist in the data file. For each prediction model, observations with missing values in the predicted variable or observations which had missing values in any of the predictors were removed. In this study, data from one cohort is analyzed: students in Grade 8 in 2018-19. The predicted variable is Grade 8 math scale score in 2018-19, which ranges from 2265 to 2802, with an average standard error of 29. In total, 36,591 students were included in the analysis.

Study Design

This section outlines the four studies that are conducted. Study 1 focuses on parameter optimization of GBM; Study 2 answers research questions about the comparison between SGP(gbm) and QR-based SGP; Study 3 illustrates SHAP analysis for interpreting SGP(gbm); Study 4 explores requirements on sample sizes for the method.

Study 1: The Optimization of XGBoost Hyperparameters

The purpose of Study 1 is to investigate the impact of hyperparameters on model performance. Three influential hyperparameters for XGBoost are learning rate, maximum depth of trees, and number of trees (Wen, Ye, & Gao, 2020). As such, different levels of each hyperparameter are tested on model performance and model fit.

Learning rate, also known as "shrinkage factor", controls the weighting of new trees added in the model. The value ranges from 0 to 1. Lower learning rate slows the learning process and requires a greater number of trees. Learning rate is commonly set at a number between 0.001 and 0.5. Nine levels are selected for the learning rate: 1, 0.5, 0.3, 0.2, 0.1, 0.05, 0.01, 0.005, and 0.001.

The number of trees, also known as the number of estimators or rounds, plays an important role in balancing computation efficiency and the complexity of an XGBoost model. If the number of trees is too small, the learning process might be terminated before an optimized model is built. If the number of trees is too large, not only does the model take a long time to train, but the model can also become very complex and overfitted. The choice of the number of trees is related to the learning rate. Considering that some small learning rates are used in this study, a few relatively large numbers of trees are selected. Eleven levels are selected for the number of trees: 50, 100, 200, 300, 400, 500, 1000, 1500, 2000, 3000, and 5000.

Maximum depth of trees constrains the depth of each tree in the model. Lower depth of trees, leading to simpler trees, is often used to avoid model overfitting. Two levels are selected for studying the maximum depth of trees: 2 and 4.

Optimization of hyperparameter tuning usually starts with splitting the data into a training set and test set. The model will be trained on the training data with varying sets of hyperparameters. The predictors and predicted variable in this study are selected according to the state's current SGP model for grade 8 students. The predictors are these students' previous math scale scores in 2015-16, 2016-17 and 2017-18. The predicted variable is the Grade 8 cohort's math scale scores in 2018-19.

The prediction accuracy of XGBoost models will be evaluated by Root Mean Squared Error (RMSE). RMSE is the square root of the average of squared errors between the model's predicted current-year math scale scores and students' observed scores. RMSE is a well-known indicator for prediction errors. Lower RMSE means higher prediction accuracy. Model overfitting can be indicated by the difference of RMSE on the training data and test data. When the prediction model is more accurate on the training data set compared to the test data set, a model overfitting issue is likely to exist. In contrast, model underfitting happens when the prediction accuracy is low on both training and test data. The best fit model performs well on both training and test data. Usually, a cross validation procedure is used to find the best set of hyperparameters for a well-fitted model.

Study 2: SGP(gbm) comparisons with traditional Quantile Regression SGP

Growth Percentiles calculated by SGP(gbm) are compared with quantile-regression based SGP. The hyperparameters for the prediction model of SGP(gbm) are chosen by a 5-fold grid-search cross-validation procedure. Based on the findings from study 1, the levels of the candidate hyperparameters are chosen from narrower ranges: a) number of trees: 500, 1000, 1500, and 2000; b) learning rate: 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, and 0.010; c) maximum depth of trees: 2 and 4. In the grid-search cross-validation procedure, the full data is randomly partitioned into 5 equal sized subsamples, named "folds". XGBoost is iteratively trained on 4 of the subsamples (training data set) and tested on the other subsample (validation or test data set). For each combination of hyperparameters, the prediction accuracies of XGBoost on the test data sets is averaged across 5 folds. The average prediction error is calculated for all sets of the hyperparameters of interest. The set of hyperparameters with the lowest prediction error is regarded as the "best" set of hyperparameters. A final model is trained on the full data using the "best" hyperparameters.

The QR-based SGP was calculated by the widely-used SGP package in R (Betebenner et al., 2022). The following evaluation criteria are used to compare the two methods:

1) Pearson's correlation coefficients
2) The average of the absolute differences

3) The density of absolute differences
4) The percentage of students who have similar SGP estimates, where similarity is defined to be an absolute difference between 1 to 20 units.

In addition, some other factors are considered in the comparison between SGP(gbm) and QR-based SGP: the number of prior years' scores as predictors and covariates in the prediction model of SGP(gbm). For grade 8 students in the current study, three prior years' scores are available for calculating the growth scores, which is also the applied rule for SGP calculation as specified in the state's growth model documentation. However, the number of prior years' scores might be varying across years and states. For states who have only collected two consecutive years of data, only one prior year's scores can be used as the predictor in growth modeling. Also, the number of prior years' scores increases as students' test scores are collected from lower grades to higher grades. This study also investigates how the difference between SGP(gbm) and QR-based SGP changes as different numbers of prior years' scores are included. The impact of including demographic variables in models with one prior year and multiple prior years is also illustrated.

Study 3: SHAP analysis for interpreting prediction models

After an SGP(gbm) model is fit to a particular set of input features and an outcome, SHAP values can be used to assign a contribution weight to each input feature, quantifying how important each predictor is to predicting the outcome. In this study, SHAP analysis for an SGP(gbm) model are presented and discussed.

Study 4: Investigating impact of sample sizes on the accuracy of estimation for the prediction models

Currently, growth scores are often computed for a large population annually. The data typically includes all students in a state. There may be times when it could be appropriate to compute growth scores for a smaller population, like a district or a group of students who need special education (Castellano & Ho, 2013). This experiment investigates model performance of SGP(gbm) when using smaller data sizes. The data sets with various sample sizes are randomly sampled from the state-level data file. The sample size (n) levels included are: 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000, 4000, and 5000. A previous study (Castellano K. E., 2011) shows that QR-based SGP can provide stable estimates when the sample size reaches 5000. Therefore, 5000 is chosen as the maximum n count for the small data analysis. The minimum sample size of 500 is close to the n count of a grade in a medium-size district or a large school. 100 replications were conducted for each sample size level.

For small data sets, model overfitting and low prediction accuracy are two concerns. 5-fold grid-search cross validation, like the procedure used in study 2, is carried out on the training dataset to tune the prediction models on the small data sets. The levels of candidate hyperparameters for small datasets are different from those for large data sets, as simpler models may be preferred for smaller datasets. The levels of number of trees are: 50, 75, 100, 200, 300, 400, and 500; learning rates: 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, and 0.10. The maximum depth of trees is fixed at 2. After the best set of hyperparameters is chosen, the final model is trained on the full training data set. Finally, the average RMSE of predicted scores by GBM on the training and test data sets across 100 replications are compared at different sample sizes (n). Study 4 focuses on the prediction accuracy of GBM, which cannot be compared to quantile regression. Therefore, a simple OLS linear regression model is used as a baseline for comparison.

Results

Study 1 Results: Hyperparameter Tuning and Model Comparison

To show the impact of XGBoost hyperparameters on model fitting, several prediction models with varying hyperparameters are built for the grade 8 cohort's math scale scores. After removing students with missing values, 29471 students with complete data on 4 years' math scale scores are included in study 1. The full data was randomly split into a training data set (67%) and a test data set (33%). For each trained model, RMSE of the predicted scores was computed for both the training and test data set.

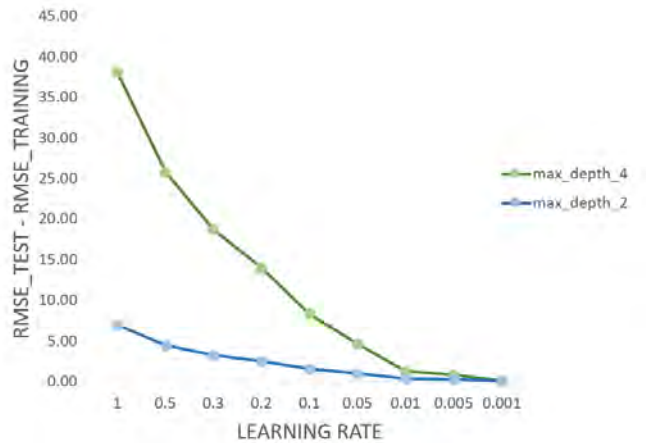

Figure 2.

Training and Test set RMSE differences across learning rates and maximum depth of trees.

In Figure 2, the difference between RMSE on test data and RMSE on training data is shown. A smaller difference means that the model is performing more similarly on the seen training data and the unseen test data, which means that the model's predictive accuracy is generalizing to unseen data well. From the figure, the difference decreases as the learning rate decreases from 1 to 0.001, indicating that the GBM model is less overfitted with a lower learning rate. When the maximum depth of trees decreased from 4 to 2, the difference between the two sets is relatively lower, indicating less overfitting. In Figure 2, the number of trees was fixed to 1000. When the number of trees is very small, a higher learning rate might lead to model underfitting issues.

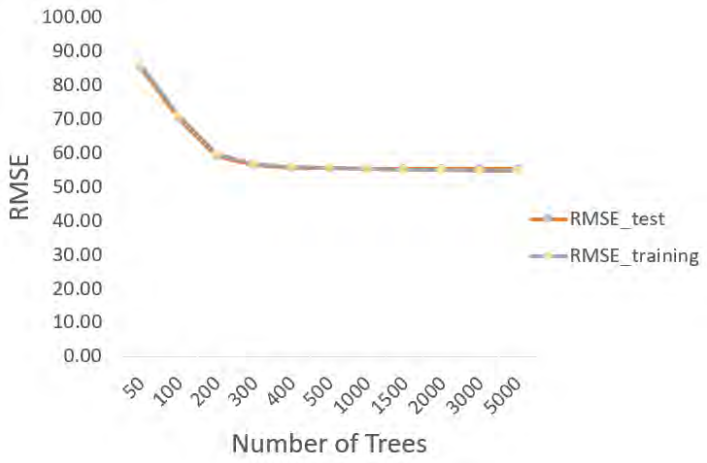

Figure 3.

The Influence of the number of trees on RMSE.

Next, the influence of number of trees is visualized in Figure 3. The learning rate is fixed to 0.01 and the maximum depth of trees is fixed to 2. Figure 3 shows that RMSE on both training and test set mainly decreases as the number of trees increases, although RMSE seems to plateau at a certain point. The y-axis is the RMSE of predicted scores, while the x-axis is the number of trees in the XGBoost models. When the number of trees is only 50, the model is underfitted, with high RMSE values for both training and test data.

Table 1:

Examples of Underfitting, Overfitting, and Proper Fit

Number of trees	Learning rate	RMSE_test	RMSE_train	RMSE_diff	Label
100	0.01	69.38	70.13	-0.75	Underfitted
1000	0.30	59.05	41.80	17.25	Overfitted
1000	0.01	55.20	54.38	0.82	Fit well

Table 1 highlights several conditions where the models are either underfitted, overfitted or fit well, as shown in the Label column. The underfitted model has higher RMSE on both training and test data. The overfitted model has a much lower RMSE on the training data, comparing to the test data. RMSE difference (i.e., the RMSE_diff column) shows the difference between RMSE on the test data and training data. The overfitted model has a high value of RMSE_diff. After increasing the number of trees from 100 to 1000 and decreasing the learning rate from 0.3 to 0.01, the overfitting problem is alleviated. Based on these results, hyperparameter tuning is essential to ensuring that the XGBoost models fit well to be used to calculate growth scores.

Study 2 Results: Comparisons Between SGP(gbm) and QR-based SGP

In study 2, SGP(gbm) is compared with QR-based SGP. First, SGP(gbm) is compared with the SGP calculation used in the state's accountability report in 2019. The state's SGP values are calculated by a QR-based procedure implemented in R (Betebenner et al., 2022). For grade 8 students, 3 previous years' math scale scores are included in quantile regression to calculate math SGP in this year. SGP(gbm) also includes three previous years' scores as the predictors. The hyperparameters for the XGBoost model were selected by a 5-fold cross-validation procedure. The best set of hyperparameters was used for model training and prediction of SGP(gbm). Specifically, the number of trees was 1000, the maximum depth of trees was 2, and the learning rate was 0.01. Furthermore, percentiles can take values between 0 and 100, but QR-based SGP values are integers between 1 and 99. The value of 0 was converted to 1, while 100 was converted to 99. To compare with the QR-based SGP, SGP(gbm) is also rounded as an integer and converted to values between 1 and 99.

Results show that the correlation coefficient between SGP(gbm) and QR-based SGP was very high (0.99-1.00, see **Table 3**). The following two figures are presented to illustrate the distribution of the absolute difference between individual SGP(gbm) and traditional SGP.

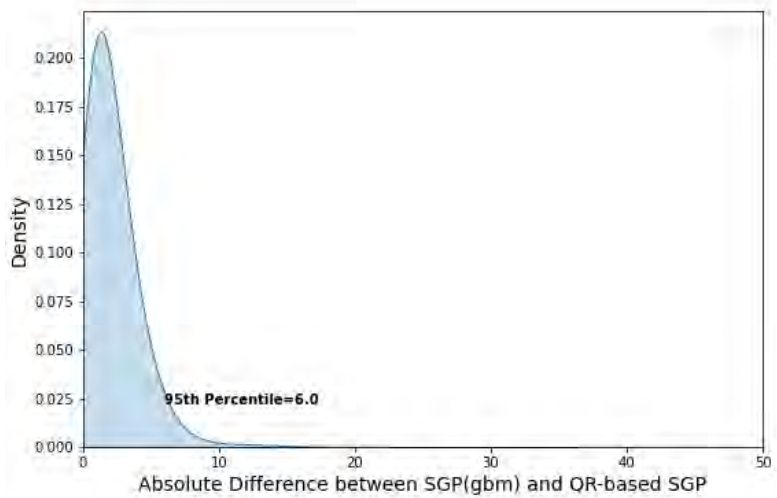

Figure 4.

Density of the absolute difference between SGP(gbm) and QR-based SGP.

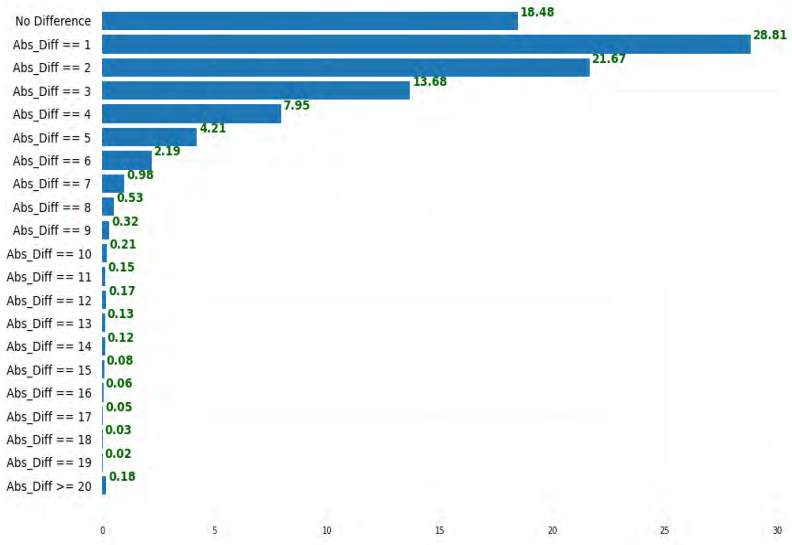

Figure 5.

Plot of percentages of the absolute difference (Abs_Diff) between SGP(gbm) and QR-based SGP.

As shown in Figure 4, the 95th percentile of the distribution is 6. This means that 95 % of students received very similar growth scores comparing SGP(gbm) and QR-based SGP, with a difference of 6 or less. Results in Figure 5 show the percentages of students at varying levels of absolute difference between QR-based SGP and SGP(gbm). According to existing studies (Lockwood & Castellano, 2017), the standard errors for individual SGPs by the traditional quantile regression method could be as large as 20, largely due to the measurement errors of scale scores. The marginal standard error of measurement for grade 8 math scores in the current test is 29. Therefore, SGP estimates by different methods are regarded as similar when the absolute difference is within the range between 0 and 20. The percentage of students with the absolute difference of each value in this range are shown in the bar plot. The highest percentage (28.81 %) belong to students with an absolute difference of 1, and the second highest percentage is attributed to those with an absolute difference of 2, 21.67 %. Meanwhile, 18.48 % students received the same SGP estimates by the two methods. The average absolute difference of SGP estimates for all students is very low: 2.11.

Numbers of Previous Years' Predictors

The estimates of growth percentiles by SGP(gbm) and QR-based SGP are compared across prediction models using different numbers of previous years' predictors. For students in lower grades or students who only take the test in two consecutive years, the number of predictors is limited. For grade 8 students in 2019, most of them have three previous years' scores: grade 5 in school year 2015-16, grade 6 in school year 2016-17, and grade 7 in school year 2017-18. The students' SGP could be estimated by 1-prior, 2-prior, or 3-prior years' scale scores. In this study, the influence of the number of predictors on SGP(gbm) estimates is investigated. 1-prior means that only the math scale score from the 2017-18 school year is used as the predictor. 2-prior means that the predictors included are the math scale scores from the 2016-17 and 2017-18 school years. 3-prior includes math scale scores from 2015-16, 2016-17, and 2017-18.

Table 2:

Compare SGP Estimates of 1-Prior, 2-Prior and 3-Prior Models for Each Method

		Correlation coefficient	Average absolute difference
SGP(QR)	3-prior vs 2-prior	1.00	1.55
	2-prior vs 1-prior	0.96	5.70
	3-prior vs 1-prior	0.96	5.98
SGP(gbm)	3-prior vs 2-prior	1.00	1.71
	2-prior vs 1-prior	0.96	5.73
	3-prior vs 1-prior	0.96	6.04

Table 2 shows the correlation coefficients between SGP estimates with different number of predictors by SGP(gbm) and quantile-regression based SGP separately. In general, for both approaches, the estimates of student growth percentile are very similar when the number of priors decreases from 3 to 2. However, the change of SGP estimates is much more evident when the number of priors decreases from 2 to 1. This could indicate that SGP estimates are not linearly influenced by the number of years of scores included in the predictors. Increasing the number of prior years' scores in the prediction model causes less impact on SGP when the number of prior years is higher.

Table 3:

Compare SGP Estimates from SGP(gbm) and SGP(QR) with Different Predictors

		Correlation coefficient	Average absolute difference
SGP(gbm) vs SGP(QR)	1_prior	1.00	1.95
	2_prior	0.99	1.98
	3_prior	0.99	2.11

The Pearson's correlation coefficients between the SGP estimates from SGP(gbm) and QR-based SGP were slightly smaller when the number of priors change from 1 to 3, while the average absolute difference increased. This means that the SGP estimates from SGP(gbm) and QR-based SGP are less similar when more previous years' scale scores are included in the prediction model. A possible reason is that more predictors increase model complexity, allowing for more opportunity for divergence. This finding is consistent with a previous study (Castellano K. E., 2011), where the author examined the effect of the number of previous years' scale scores included in different SGP estimation approaches on the recovery of the estimated growth percentile metrics of a benchmark. It was demonstrated that when there was only one prior year, the differences between the estimated growth percentiles and their benchmark were smaller.

Other Covariates

In SGP(gbm), the prediction model is flexible, as the models can take additional covariates, such as students' scale scores in other subjects and their demographics as inputs. It is advised that careful consideration should always be given to the variables included in a model, even when it is computationally easy and convenient to include whatever is available. This study further investigated the inclusion of the demographic variables: students' gender, Free Reduced Lunch (FRL), English Language Learner

(ELL), Individual Education Plan (IEP) and ethnicity are added as extra predictors in addition to English language arts/literacy (ELA) and math scale scores from the three prior years. Student's gender was re-coded into a binary variable, 1 for female and 0 for male. FRL, ELL, and IEP were binary variables, indicated by 0 and 1 showing the 'No' vs. 'Yes' statuses. Ethnicity was a categorical variable and re-coded into 6 binary variables: American Indian, Asian, Black, Hispanic, Pacific Islander, and White with 1 for 'Yes' in a specific ethnicity group and 0 for 'No' in other ethnicity groups.

In order to investigate the influence of covariates on SGP(gbm), students' demographics and ELA scale scores from previous years were added as additional predictors to the prediction models for calculating SGP(gbm) in Math. The number of prior years was also manipulated across the tested models. Three of the model hyperparameters (learning rate, number of trees, maximum depth of trees) were tuned by cross validation for each prediction model. The SGP estimates by SGP(gbm) are compared to QR-based SGP with the same number of priors. QR-based SGP calculation doesn't take ELA scale scores or the demographics into consideration.

Table 4:

SGP(gbm) Models for Math with and without Additional Covariates

Number of Prior Years	Predictors	RMSE	Relation to SGP(QR) without Covariates	
			Correlation Coefficient	Average Absolute Difference
		Models With Math Scale Scores		
1-prior	math_18	56.64	1.00	1.97
1-prior	math_18, Demographics	55.60	0.98	4.74
2-prior	math_17, math_18	54.51	0.99	1.99
2-prior	math_17, math_18, Demographics	53.71	0.98	4.43
3-prior	math_16, math_17, math_18	54.37	0.99	2.11
3-prior	math_16, math_17, math_18, Demographics	53.53	0.98	4.47
		Models With Math and ELA Scale Scores		
1-prior	math_18, ELA_18	54.64	0.96	6.30
1-prior	math_18, ELA_18, Demographics	54.11	0.95	6.90
2-prior	math_17, math_18, ELA_17, ELA_18	53.15	0.97	5.51
2-prior	math_17, math_18, ELA_17, ELA_18, Demographics	52.72	0.96	6.10
3-prior	math_16, math_17, math_18, ELA_16, ELA_17, ELA_18	52.98	0.97	5.53
3-prior	math_16, math_17, math_18, ELA_16, ELA_17, ELA_18, Demographics	52.48	0.96	6.13

Table 4 presents the RMSE of the predicted scores, the correlation coefficients, and average absolute difference between SGP(gbm) and QR-based SGP. After adding the covariates, RMSE of the predicted scores is slightly lower, indicating higher prediction accuracy of the prediction models. For example, the RMSE decreased from 56.64 to 55.60 for the model with 1-prior years' math scale scores (math_18) as the predictor. The model with all the covariates in the predictors has the lowest RMSE of 52.48. This means that, comparing to the model with no additional covariates, the prediction accuracy of the model with covariates is slightly higher, while the difference between SGP(gbm) and QR-based SGP also increased. Moreover, adding ELA scale scores in the prediction model seems to be more influential to SGP(gbm) than adding the demographics. The importance of the predictors is further investigated in study 3, where SHAP is applied to measure the contribution of the predictors in the prediction model.

Computation Efficiency of SGP(gbm) Comparing to the QR-based SGP

The estimation of SGP by the QR-based procedure can be time consuming, as it involves the creation of numerous regression lines that relate prior with current students' scale scores. Specifically, the SGP is commonly known as an integer, ranging from 1 to 99. That means 99 quantile regression lines are estimated for the 1st to 99th conditional percentiles of the current-year test scores. In addition, B-spline parameterization instead of a linear model is used to fit the regression lines, meaning that more parameters need to be estimated. An individual's SGP is obtained by comparing his/her current score to the curvilinear quantile regression lines. In this study, QR-based SGP was simultaneously estimated for all subjects and grades in a state summative test using the R package. Each grade/subject has about 30,000 students in the data set. It took about 15 minutes for the "SGP" package in R to estimate all students' SGP after the data and meta data were prepared in the right format. The computation time is expected to be higher for a larger state.

In contrast, SGP(gbm) requires the training of one XGBoost model, which is often fast. The total training time might be influenced by XGBoost hyperparameters. However, even if the number of trees is set to 1000, the model could be trained within 10 seconds for the state data for about 30,000 students. Therefore, SGP(gbm) is a preferable choice when growth scores need to be calculated on-the-fly. If a XGBoost model is built for calculating SGP(gbm) for a new data set, some extra time is needed for hyperparameter tuning. Usually, only some of the hyperparameters need to be adjusted to obtain a well fitted model. A grid search for hyperparameter tuning should be carried out when XGBoost is fit to a new data set. After the best set of parameters are found, it is not necessary to repeat hyperparameter tuning for that dataset.

In this study, the computation time of SGP(gbm) is about 1-2 seconds for one subject at one grade, including the computation of model evaluation statistics and growth percentiles. The calculation of SGP(gbm) with hyperparameter tuning could take a few

minutes to several hours for the full data, depending on the granularity of the hyperparameter grid search space.

Study 3 Results: The Feature Importance of Predictors in GBM

Study 3 generates and compares the feature importance values of predictors for the prediction models used in SGP(gbm). Figure 6 shows the feature importance of each predictor in a prediction model with all 3 prior years' scale scores and students' demographics as the predictors. Feature importance is the mean absolute SHAP value computed from each individual student's data. Math scale score in school year 2017-18 is the most important predictor in predicting math scale score in school year 2018-19. The second most important predictor is math scale score in school year 2016-17. Meanwhile, students' demographic variables play a less important role in the prediction model compared to the other predictors. Among all the demographic variables, gender has the highest feature importance value.

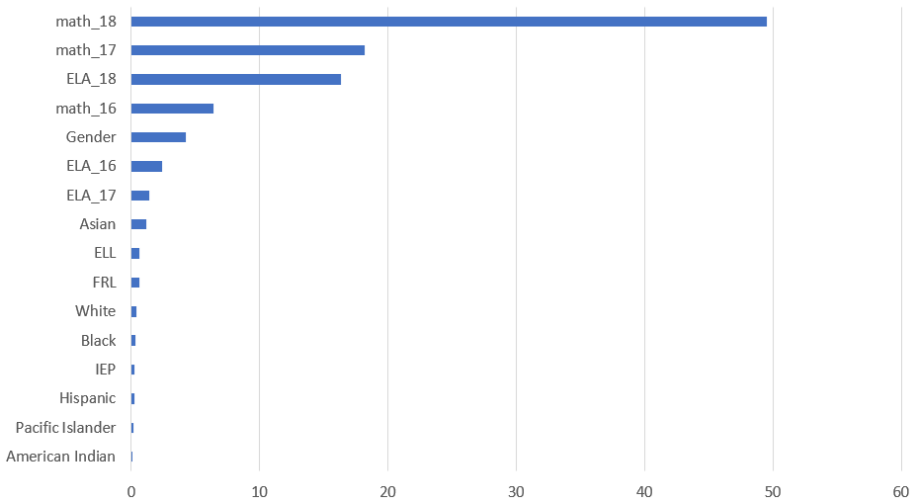

Figure 6.

Feature importance (mean absolute SHAP value) of predictors in the full prediction model.

Table 5:

The Change of Feature Importance Values Across Models with Decreasing Number of Predictors

Predictors	Model					
	1st	2nd	3rd	4th	5th	6th
math_18	55.75	56.07	56.15	56.78	59.12	82.26
math_17	21.98	22.11	22.40	22.82	26.09	
math_16	6.98	6.96	7.19	6.25		
Gender	6.80	6.70	6.83			
FRL	1.53	1.33				
ELL	0.99	0.97				
IEP	0.28	0.02				
Asian	1.44					
Black	0.59					
White	0.52					
Hispanic	0.34					
American Indian	0.20					
Pacific Islander	0.02					
Total	97.42	94.16	92.57	85.85	85.21	82.26

Table 5 shows the feature importance in the 6 models with a decreasing number of input features. The 1st model contains all 13 predictors that were available in this study's dataset. The predictors are shown in descending order of feature importance values, except that the ethnicity features are grouped together even if "Asian", "Black", "White" and "Hispanic" are more important than "IEP". The results of this table show how the contribution scores of the assigned SHAP values might shift when adding or removing predictors from the model. Each model after the 1st model has several of the predictors removed. The last row of the table, the "Total" value of SHAP values, decreases as predictors are removed. This indicates that the model loses some predictive accuracy as predictors are removed.

After removing the ethnicity features from the 1st model, the 2nd model contains 7 features, among which math scale scores in school year 2017-18 is still the most important predictor, with a feature importance value of 56.07, while the order of feature importance of other predictors remains consistent from model 1 to model 2. The 3rd, 4th, and 5th models continue to contain fewer features by removing the least important features of the preceding model. The trend in which feature importance changes by variable addition or removal can be observed in the table.

The last row in **Table 5** shows that the sum of SHAP values of all predictors is decreasing from the 1st model to the 6th model, as the prediction models are slightly less

accurate when the number of predictors is smaller. However, in model 6, where there is only one predictor, the feature importance value of 2017-18 math scale scores increased to 82.26 compared to a value of 59.12 from model 5. This indicates that, as expected, 2017-2018 math scale score is correlated with the other predictors used in the previous models, such as 2016-2017 math scale scores used as a second feature in model 5. When multiple correlated input features are used in a prediction model, SHAP values will be relatively lower than if those correlated input features were in a prediction model without the other correlated input features.

Study 4: Investigating impact of sample size on the accuracy of estimation for prediction models

This study shows how well an XGBoost model can perform on small data sets. 1-prior and 3-prior year covariate prediction models are evaluated. The prediction accuracy of the trained model on the training and test data set is calculated. The average RMSE across 100 replications of a XGBoost model is compared to that of a linear regression.

Table 6:

The Best Hyperparameters and Prediction Accuracy for Small Data Sets (1-Prior)

Sample Size	Best Hyperparameter Set		Prediction Accuracy			
	Learning Rate	Number of Trees	RMSE_test SGP(gbm)	RMSE_diff SGP(gbm)	RMSE_test (Linear)	RMSE_diff (Linear)
500	0.09	75	57.99	5.99	58.56	0.14
600	0.09	75	57.53	4.98	58.29	-0.15
700	0.09	75	57.50	4.58	58.23	-0.19
800	0.09	75	57.80	4.75	58.85	0.64
900	0.09	75	57.76	4.33	59.03	0.73
1000	0.07	100	57.21	3.48	58.34	-0.13
1500	0.09	75	57.10	2.57	58.49	-0.03
2000	0.05	150	57.03	2.08	58.66	0.10
2500	0.05	150	57.41	2.32	58.99	0.51
3000	0.05	150	56.94	1.55	58.67	0.10
4000	0.05	150	57.12	1.56	58.91	0.42
5000	0.05	150	56.82	0.97	58.60	-0.04

Note: The explored levels of number of trees include 50, 75, 100, 200, 300, 400, and 500; learning rates include 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, and 0.10. The maximum depth of trees is fixed at 2.

Table 6 depicts results of using sample sizes ranging from 500 to 5000. Looking at the Best Hyperparameter Set column, these results represent the best performing hyperparameter set for each sample size across 100 replications of a grid search using the specific sample size. In the Prediction Accuracy columns of the table, four different RMSE results are presented. RMSE_test for both the XGBoost model and the OLS linear regression model are presented. RMSE difference (RMSE_diff) is also presented, which is the difference between the test set and training set RMSE. Low RMSE_test is better, as the model is more accurate. Additionally, RMSE difference that is closer to 0 is better, since the prediction model can generalize well from seen training data to unseen test data. In the table, RMSE_diff for the GBM model decreases as sample size increases. RMSE_diff of the linear model appears stable across sample sizes, seemingly unrelated to sample size. The prediction accuracy of GBM is always higher than that of the linear models for 1-prior models, indicated by lower RMSE values (range from 56.82 to 57.99) on the test data sets. RMSE_diff is 5.99 when the 1-prior prediction model is optimized for sample size of 500 and decreases to 0.97 when the sample size is 5000. This means that the XGBoost model is more likely to be overfitted when the sample size is smaller. RMSE_diff in all the tested conditions is relatively small, considering the scale of the predicted variable. Compared to a linear model, the prediction accuracy of GBM is mostly higher, in spite of the challenge of model overfitting.

Table 7:

The Best Hyperparameters and Prediction Accuracy for Small Data Sets (3-Prior)

Sample Size	Best Hyperparameter Set		Prediction accuracy			
	Learning Rate	Number of Trees	RMSE_test SGP(gbm)	RMSE_diff SGP(gbm)	RMSE_test (Linear)	RMSE_diff (Linear)
500	0.09	75	56.41	8.85	56.19	0.47
600	0.09	75	55.90	7.49	55.99	0.21
700	0.09	75	55.87	6.80	55.89	0.14
800	0.09	75	56.17	6.86	56.33	0.71
900	0.07	100	55.94	6.04	56.49	0.80
1000	0.09	75	55.48	5.11	55.78	-0.10
1500	0.07	100	55.28	3.85	55.94	-0.01
2000	0.05	150	55.15	3.12	56.09	0.13
2500	0.05	150	55.47	3.10	56.37	0.48
3000	0.05	150	55.01	2.27	56.06	0.06
4000	0.05	150	55.21	2.18	56.31	0.39
5000	0.05	150	54.92	1.51	56.06	-0.03

Note: The explored levels of number of trees include 50, 75, 100, 200, 300, 400, and 500; learning rates include 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, and 0.10. The maximum depth of trees is fixed at 2.

Table 7 displays similar patterns as observed in **Table 6**, with the inclusion of 3-prior years' covariates. Comparing RMSE_test between the two tables, both the GBM and the Linear Regression models have lower RMSE values with the 3-prior model, indicating that prediction accuracy improves when including more priors. The prediction accuracy on the test sets of GBM is slightly lower than the linear model when $n = 500$ for the 3-prior prediction models. The RMSE_test of GBM is 56.41, which is slightly higher than that of a linear model: 56.19. However, as the sample size gets larger, the prediction accuracy of GBM improves, while the prediction accuracy of the linear regression model appears to remain relatively stable without clear improvement. RMSE_test for the GBM model becomes more accurate than the linear model after the sample size reaches 600. Similar to the trend observed from 1-prior models, the trend in the results from the 3-prior models is that RMSE_diff decreases as the sample size increases. Having a larger sample size improves both RMSE_test and RMSE_diff.

Discussion

This paper showed, through empirical data analyses, the procedure of training the proposed SGP(gbm) model on datasets of varying predictors and of varying sample sizes, and comparing the results to traditional growth percentile approaches. SHAP values, as delivered by the TreeExplainer package, were shown to increase model interpretability of GBM by assigning a "feature importance" score to each predictor.

In study 1, hyperparameter tuning was shown to be essential in obtaining a stable SGP(gbm) calculation. Specifically, the number of trees, learning rates, and the maximum depth of trees play a key role in avoiding model overfitting or underfitting. For example, when the number of trees is increased to 1000 and learning rate is decreased to 0.01, the procedure fits well on both the training and test data in one of tested conditions. Study 2 results showed that students' growth scores are highly consistent between SGP(gbm) and QR-based SGP for some of the tested conditions. The difference between SGP(gbm) and QR-based SGP was influenced by adding more covariate variables. Adding students' scale scores in other subjects and students' demographic variables in the prediction model improved the model's prediction accuracy but decreased the similarity between SGP(gbm) and QR-based SGP. Results from study 3 demonstrated that feature importance values provide useful information when interpreting XGBoost models. These feature importance values are consistent across models and shown to be significantly influenced by the predictors' collinearity as well. In summary, study 1 to 3 introduce how the GBM approach can provide an alternative to the QR-based approach for estimating student growth percentiles.

The fourth study focused on the sample size requirements for SGP(gbm). Although it is understood that machine learning models prefer a large data size, this study demonstrated to what extent XGBoost models can be overfit for small-size data sets. As sample sizes increase, the XGBoost model becomes both more accurate and less overfit. Compared with QR-based SGP, SGP(gbm) has the advantage of providing direct evidence on prediction accuracy and model fit evaluation. The results of this study indicate that it is important that stakeholders decide whether to use GBM or a simpler model for small datasets, taking into consideration the assumptions of each model type and model fitting issues.

Regression trees are interpretable, simple, and can easily find nonlinear relationships. In the past, it may have been computationally and algorithmically difficult to find the right ways to construct the best ensemble of trees. However, recent algorithmic and computational improvements have enabled regression trees to be far more practical to implement, effectively minimizing prediction error while maintaining reasonable model complexity and being very fast to train through programming optimizations. This study showed that SGP(gbm) can generate consistent SGP values with a well-trained prediction model which makes no assumptions on the data. The method of SHAP TreeExplainer can generate easy-to-understand feature importance values for the prediction model. In practice, SGP(gbm) could provide efficient and accurate estimates of SGP for state- and district- level tests. When more covariates are added into the model, the estimated expected scores become more accurate, which is currently unavailable in the QR-based SGP approach.

The current study is limited in several respects, however. First, only one set of testing data was used for evaluating the performance of SGP(gbm). The format and characteristics of this data might not represent other data sets. Future studies could test this method in other large-scale assessment data. Second, quantile-regression based SGP was treated as somewhat a gold standard in the current study. However, this traditional approach is not without its own limitations. It might also be worth conducting simulation studies where the true growth percentiles are known in order to evaluate the accuracy of SGP(gbm). More importantly, the accuracy of SGP(gbm) on small data sets could be further evaluated using a dataset where the “true” SGP is known. In the current empirical analyses, the true SGP is unknowable. Further work could involve a simulation study with the true SGP known.

References

Banerjee, P. (2022, August 28). A guide on XGBoost hyperparameters tuning. Retrieved from Kaggle: https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning/notebook

Benjamin, A. S., Fernandes, H. L., Tomlinson, T., Ramkumar, P., VerSteeg, C., Miller, L., & Kording, K. P. (2017). Modern machine learning far outperforms GLMs at predicting spikes. Frontiers in Computational Neuroscience. doi:https://doi.org/10.1101/111450

Betebenner, D. (2008). Toward a normative understanding of student growth. In K. Ryan, & L. Shepard, The future of test-based educational accountability (pp. 155-170). New York: NY: Taylor & Francis. doi:https://doi.org/10.4324/9780203895092

Betebenner, D., Van Iwaarden, A., Domingue, B., & Shang, Y. (2022). SGP: Student Growth Percentiles & Percentile Growth Trajectories. Retrieved from https://github.com/Center-ForAssessment/SGP

Castellano, K. E. (2011). Unpacking student growth percentiles: Statistical properties of regression-based approaches with implications for student and school classificaitons. Iowa: ProQuest Dissertations and Theses Database (UMI No. 3461371). doi:10.17077/etd.jviwgs6q

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. Journal of Educational and Behavioral Statistics, 38(2), 190-215. doi:https://doi.org/10.3102/1076998611435413

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 785-794). San Francisco. doi:https://doi.org/10.1145/2939672.2939785

Li, Z., & Tang, S. (2019). Forecasting students' future academic performance using big data analytics. annual meeting of the National Council on Measurement in Education (NCME). Toronto, Canada. Retrieved from https://www.emetric.net/Content/pdf/Manuscript-Forcasting%20Students'%20Future%20Performance.pdf

Lockwood, J. R., & Castellano, K. E. (2017). Estimating true student growth percentile distributions using latent regression multidimensional IRT models. Educational and Psychological Measurement, 77(6), 917-944. doi:10.1177/0013164416659686

Lundberg, S. M., & Lee, S.-I. (2017). A United Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30. Long Beach, CA. doi:https://doi.org/10.48550/arXiv.1705.07874

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., . . . Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2522-5839. doi:https://doi.org/10.48550/arXiv.1905.04610

Monroe, S., & Cai, L. (2015). Examining the reliability of student growth percentiles using multidimensional IRT. Educational Measurement: Issues and Practice, 21-30. doi:10.1111/emip.12092

R Core Team. (2022). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sinharay, S. (2016). An NCME Instructional Module on Data Mining Methods for Classification and Regression. NCME Educational Measurement Issues and Practice, (pp. 1-17). doi:https://doi.org/10.1111/emip.12115

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 41(3), 647-665. Retrieved from https://doi.org/10.1007/s10115-013-0679-x

Tang, S., & Li, Z. (2019). Applying XGB Regression Trees to Produce Growth Percentiles. annual meeting of the National Council on Measurement. Toronto. Retrieved from https://www.emetric.net/Content/pdf/Manuscript-XGB%20Growth%20modeling.pdf

Wagner, L. (2022, May 20). Shap's partition explainer for language models. Retrieved from https://towardsdatascience.com/shaps-partition-explainer-for-language-models-ec2e7a6c1b77

Wen, L., Ye, X., & Gao, L. (2020). A new automatic machine learning based hyperparameter optimization for workpiece quality prediction. Measurement and Control, 1088-1098. doi:https://doi.org/10.1177/0020294020932347