Applying XGB Regression Trees to Produce Growth Percentiles

Steven Tang, Zhen Li

eMetric LLC

Paper written for the 2019 meeting of the National Council on Measurement in Education, Toronto, Canada. The views expressed in this paper are solely those of the authors and they do not necessarily reflect the positions of eMetric LLC.

Correspondence concerning this paper should be addressed to Steven Tang, eMetric, 211 N Loop 1604 E, Suite 170, TX 78232. Email: steven@emetric.net.

Abstract

This study compares percentile rank residuals using an XGBoost regression tree model to quantile regression based SGP. Results indicate that with default hyperparameters, the XGB tree based approach can exactly replicate standard SGP, and that the XGB method may be further tuned to potentially predict more accurately.

Keywords: Gradient boosted regression tree, growth percentile ranking, student growth percentile

Background

In recent years, big data methods such as gradient boosted decision trees and deep neural network architectures have shown great promise in tackling a variety of prediction modeling tasks, often surpassing the results from traditional methods or even solving previously unsolvable prediction tasks. In this study, we investigate the potential for applying gradient boosted regression trees, enabled through the XGB statistical package, to the prediction task of computing student growth measures, under the hypothesis that the favorable statistical properties of XGB models may allow for an alternative procedure to compute growth measures similar to quantile regression based student growth percentiles (SGP).

SGP has been used for measuring students' annual growth in many states. In theory, an SGP describes a student's relative progress with respect to his/her academic peers, who are students beginning at the same place (Betebenner, 2008, 2018). Quantile regression is commonly used to estimate the conditional growth percentiles of current-year scores based on prior year scores.

Castellano & Ho (2013) explored using percentile rank residuals (PRR) based off of ordinary least squares (OLS) regression and found that the OLS regression method proved to be a promising alternative to the quantile-regression based SGP method, as the OLS regression PRR method recovers the true conditional status percentile ranks better in certain situations. However, OLS regression is known to have strict assumptions such as homoskedasticity of the errors and gaussian distributions of the covariates, et al.

In this study, eXtreme Gradient Boosting (XGB) regression trees using the PRR method are applied to two case study datasets as an alternative to the quantile-regression based SGP approach. Both XGB and quantile regression relax the homoskedasticity assumption, but XGB goes a step further and makes no assumption that data distributions need to be gaussian or that relationships must be linear. Moreover, XGB regression trees have favorable properties such as high predictive accuracies with many possible input variables, a tweakable and tunable training procedure, fast computation, and an interpretable decision-tree structure that can be illuminated after training. XGB approaches can be prone to issues of overfitting, therefore requiring special consideration in model construction and interpretation.

This paper proposes an XGB PRR approach to generate growth percentile ranks using recent data from a state summative assessment program. Given that quantile regression based SGP is relatively common, we will compare XGB PRR results to quantile regression SGP. We will then explore adding additional input features to improve predictive accuracy.

Research Methods

Data

Scores from math test administrations for grade 7 2016 and grade 8 2017 from a state summative assessment were used as the data for the first part of this study. shows the descriptive statistics of the data. To obtain SGP estimates using the quantile regression method, students must have a previous-year score, and students with incomplete records are omitted from the analysis. In total, 24926 students are included.

Table 1

Descriptive Statistics of the First Dataset

	N	Mean Scale Score	Minimum	Maximum
2016 Grade 7 Math	24926	2486	2250	2778
2017 Grade 8 Math	24926	2499	2265	2802

The second dataset analyzed in this study contains real student scores for both math and ELA test administrations from 2016 to 2018. Students in grades 5-8 have two prior years' scale scores. In total, five cohorts of students' growth measures were calculated by XGB PRR and SGP for an extensive comparison. Table 2 presents the descriptive statistics of the second dataset.

Table 2

Descriptive Statistics of 2018 Mathematical and ELA Test Data

			Mathematics			ELA		
Cohort	Year	Grade	N	Mean	S.D.	N	Mean	S.D.
1	2017	3	37803	2426	81	37868	2418	84
	2018	4	38311	2465	81	38309	2467	85
	2016	3	37626	2423	79	37682	2420	83
2	2017	4	38089	2463	81	38099	2461	87
	2018	5	38684	2489	89	38776	2498	90
	2016	4	36241	2459	79	36306	2462	86
3	2017	5	36804	2488	85	36868	2499	90
	2018	6	37459	2500	99	37527	2512	89
	2016	5	35475	2485	84	35518	2499	85
4	2017	6	36105	2497	98	36147	2511	89
	2018	7	36528	2509	106	36585	2539	96
	2016	6	34631	2498	97	34684	2508	84
5	2017	7	34654	2506	101	35361	2539	95
	2018	8	35502	2524	111	35577	2555	98

XGB Regression Trees

The XGB Regression Tree approach relies on iteratively building a collection of simple regression trees; regression trees are decision trees that predict continuous outcomes. The iterative process starts by first creating an extremely simple predictive regression tree; such a tree might only have between 2 to 16 leaf nodes. This initial regression tree is constructed by searching through a large number of potential split values among all input variables and finding the splits that minimize prediction error. The iterative process continues by constructing an additional regression tree of the same structure, but this time constructed to minimize the *residual errors* of the first regression tree. The next iterative tree is then constructed to minimize the residuals of the full model thus far, and the process of iteratively creating new trees continues until stopping criteria is met. As the name implies, gradient boosting uses gradient descent to find the next regression tree to add to the ensemble. At the end of the building process, the predictions are given by the sum of the outputs of all trees. This process of building a gradient boosted regression tree was optimized in the XGB package allowing for very fast computation of gradient boosted trees as well as many opportunities for additional model tuning (Benjamin, Fernandes, Tomlinson, Ramkumar, VerSteeg, Miller, & Kording, 2014).

For a predictive model $\hat{y}_1 = f_1(X)$, where X indicates input variables, $\hat{y}_1$ indicates predications by the first tree and y indicates the observed output variable, a loss function can be defined between the prediction and the observed outcome: $l(\hat{y}_1, y)$. During training, the first tree can be estimated by minimizing the following objective:

$$L_1 = \sum l(\hat{y}_1, y) + \Omega(f_1) \tag{1}$$

Ω is a regularizing function to avoid overfitting. Then a second tree $f_2(X)$ will be constructed by predicting the residuals of the first tree. The objective to minimize is as follows:

$$L_2 = \sum l(\hat{y}_1 + f_2(X), y) + \Omega(f_2) \tag{2}$$

The process continued sequentially for a fixed number of trees (*N*). Total loss will be progressively decreased with each additional tree. In the end, the prediction for y will be the sum of the predictions of all trees:

$$\hat{y} = \sum_{k}^{N} f_k(X) \tag{3}$$

Compared to linear regression and quantile regression, XGB regression tree require completely different assumptions. For example, linear regression has a basic assumption that the sum of its residuals is 0. XGB regression tree, through its boosting process, instead attempts to find and model patterns in the residuals and strengthen the model with weak learners that exploit these patterns. This approach has shown to be extremely powerful in big data tasks, winning a variety of competitions where predictions need to be made based on a wide set of predictors.

Procedure of Applying XGB to Produce Percentile Ranks of Residual

To produce XGB PRR, the following steps were carried out: 1) Train a XGB prediction model with two or more years of consecutive scale scores for one cohort of students; 2) Use the prediction model to generate a predicted score, which is regarded as the expected score that a student should have got in the current year; 3) Compute a current-year residual score by subtracting the predicted score from the current-year observed score; 4) Calculate PRR, the percentage of students whose residual scores are lower than or equal to the score of interest in the population. A function "rankdata" from a python package "scipy.stats.mstats" is used to compute ranks (order statistics) of each residual score. When the residual scores are tied, the average rank is used. Then the following formula is applied to compute percentile ranks.

$$PRR = round(100 \times \frac{rank_x - 1}{N}) \tag{4}$$

Equation (4) is slightly different from the equation (4) in Castellano & Ho's (2013) article, where they calculated PRR as the percentage of residual scores that are smaller or equal to the score of interest. Another definition of percentile rank is the percentage of residual scores less than the target score plus 0.5 of the percentages of ties in all residual scores. The different definitions of percentile ranks might lead to slightly different outcomes, but these differences should be minor after we round the percentages to integers. In addition, PRR is forced to be located within [1,99] to compare to SGP.

The XGB results presented in this study use the XGB package (Chen & Guestrin, 2016) implemented in Python. SGP results are obtained using the SGP package (Betebenner, 2018) in R. Results from two studies are presented in the following section.

Results and Discussion

The first result comes from comparing XGB PRR and SGP using just two years of scale scores for a state mathematical test. In **Error! Reference source not found.**, four different models' results are shown, each trained to incorporate different input variables. A hyperparameter grid search was performed to mitigate overfitting concerns. Results show that the base model, where only grade 7 math is used to predict grade 8 math scores, can achieve a 0.997 correlation with standard SGP.

However, as more input variables are incorporated, the correlation with SGP goes down, but R^2 with realized scores correspondingly increases. This means that the XGB PRR model with more input variables disagrees more with SGP, but has better model predictive accuracy relative to realized scores. This provides evidence that it is operationally easy for the XGB PRR approach to replicate standard (quantile regression) SGP results, but that incorporating additional explanatory variables can increase model accuracy and correspondingly decrease correlation with standard SGP.

A trained XGB regression tree model can also be inspected to better understand how the model is making decisions. There are numerous metrics that can be used.

Figure 1 depicts the most important features used in the most complex model, which used grade 7 math, grade 7 reading, and demographic variables as predictors.

Previous studies (Castellano & Ho, 2013; Lockwood & Castellano, 2015) found that alternative estimation methods (OLS based SGP or Logit model based SGP) can provide SGP estimates closer to the SGP calculated using empirical conditional distributional functions (ECDF). We didn't use ECDF in the current study, although this may be useful to look at in future studies. Lockwood and Castellano (2015) also showed that even if the correlation between the estimates by different methods are very high, the small difference between individual SGP estimates can cause significant effect for teacher evaluation, which is based on group-level SGP.

Table 3

XGB Model Results in the Pilot Study (2016-2017 Mathematical Test)

Input to XGB Model	Hyperparameters (All but the first model was chosen via 5-fold Cross-Validation)	Correlation with SGP	R^2
G7 Math	Estimators = 100, Max Depth = 1, Learning Rate = .1	.997	.619
G7 Math + Demographics	Estimators = 700, Max Depth = 1, Learning Rate = .04	.985	.628
G7 Math + G7 Reading	Estimators = 600, Max Depth = 1, Learning Rate = .03	.951	.650
G7 Math + G7 Reading + Demographics	Estimators = 700, Max Depth = 1, Learning Rate = .04	.945	.653

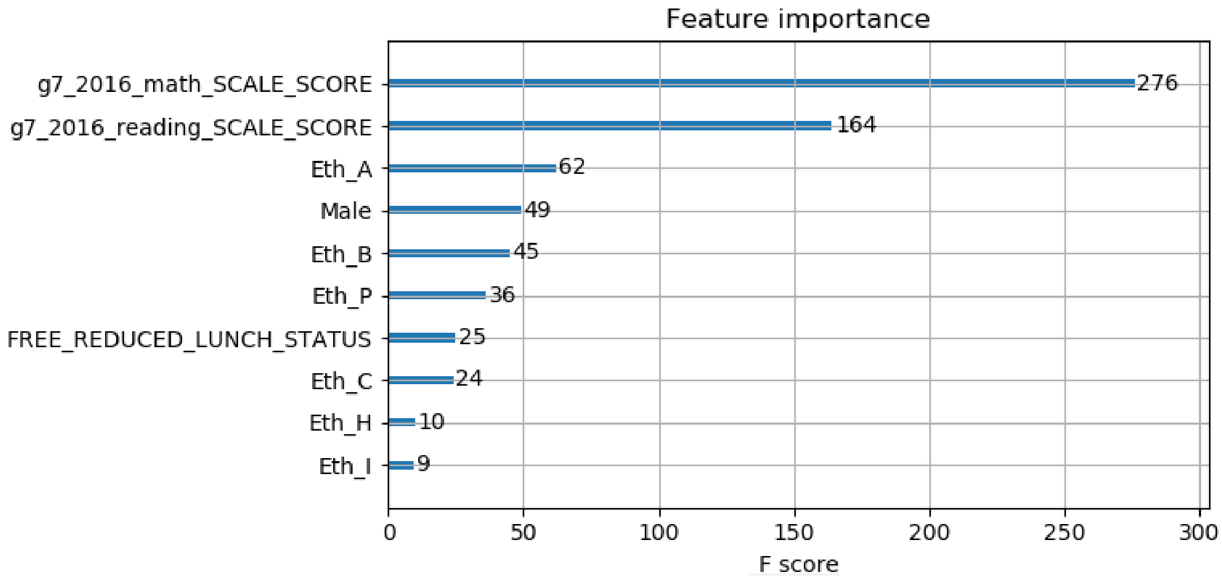

Figure 1 XGB Feature Importance from the Pilot Study

Next, using 2016-2018 students' scale score data from both mathematics and ELA test administrations, we compared the two models with more prior years' scale scores. For XGB PRR, we apply a simple XGB regression tree model with most hyperparameters set as default values. The number of estimators was fixed to 125 and max depth was fixed as 4 for all prediction models.

Table 4

XGB Model Results for 2018 Test Data

Output	Input Variables	Correlation with SGP	R^2
G8 Math	G6 Math+G7 Math	.991	.769
G8 Reading	G6 Reading+G7 Reading	.993	.778
G7 Math	G5 Math+G6 Math	.990	.812
G7 Reading	G5 Reading+G6 Reading	.991	.772
G6 Math	G4 Math+G5 Math	.989	.787
G6 Reading	G4 Reading+G5 Reading	.993	.763
G5 Math	G3 Math+G4 Math	.992	.781
G5 Reading	G3 Reading+G4 Reading	.992	.768
G4 Math	G3 Math	.996	.759
G4 Reading	G3 Reading	.995	.723

Table 5

XGB Model Results with more Input Variables

Output	Input Variables	Correlation with SGP	R^2
G8 Math	G6 Math+G7 Math+G6 Reading+G7 Reading + Demographics	.954	.788
G8 Reading	G6 Math+G7 Math+G6 Reading+G7 Reading + Demographics	.957	.794
G7 Math	G5 Math+G6 Math+G5 Reading+G6 Reading + Demographics	.958	.824
G7 Reading	G5 Math+G6 Math+G5 Reading+G6 Reading + Demographics	.956	.788
G6 Math	G4 Math+G5 Math+G4 Reading+G5 Reading + Demographics	.939	.810
G6 Reading	G4 Math+G5 Math+G4 Reading+G5 Reading + Demographics	.960	.779
G5 Math	G3 Math+G4 Math+G3 Reading+G4 Reading + Demographics	.971	.791
G5 Reading	G3 Math+G4 Math+G3 Reading+G4 Reading + Demographics	.958	.784
G4 Math	G3 Math+G3 Reading + Demographics	.966	.775
G4 Reading	G3 Math+G3 Reading + Demographics	.934	.756

Results from Table 4 shows that the correlation coefficients between XGB PRR and SGP range from .989 to .996. The correlation coefficients are equivalently high across all grades and subjects. Results in Table 5 shows that when incorporating additional input variables (more subjects and demographics), the correlation between XGB PRR and standard SGP decreased and R^2 increased. These results closely mimic the trend found from Table 3, where adding more data to the XGB model decreased correlation to SGP results but increased overall R^2.

Furthermore, results from 2018 data analysis show that the difference between XGB-PRR and SGP are higher at the extreme previous year scale scores. This effect is very significant for Grade 4 tests, where the input variables only include one prior year test data. When the number of prior years increase, this pattern is not as clear. As shown in Figure 2, for grade 8 ELA and math, the largest difference occurs for extreme scoring students, but also shows a little bit in the middle. This effect was also discovered in the 2017 data analysis and in a previous study (Castellano & Ho, 2013).

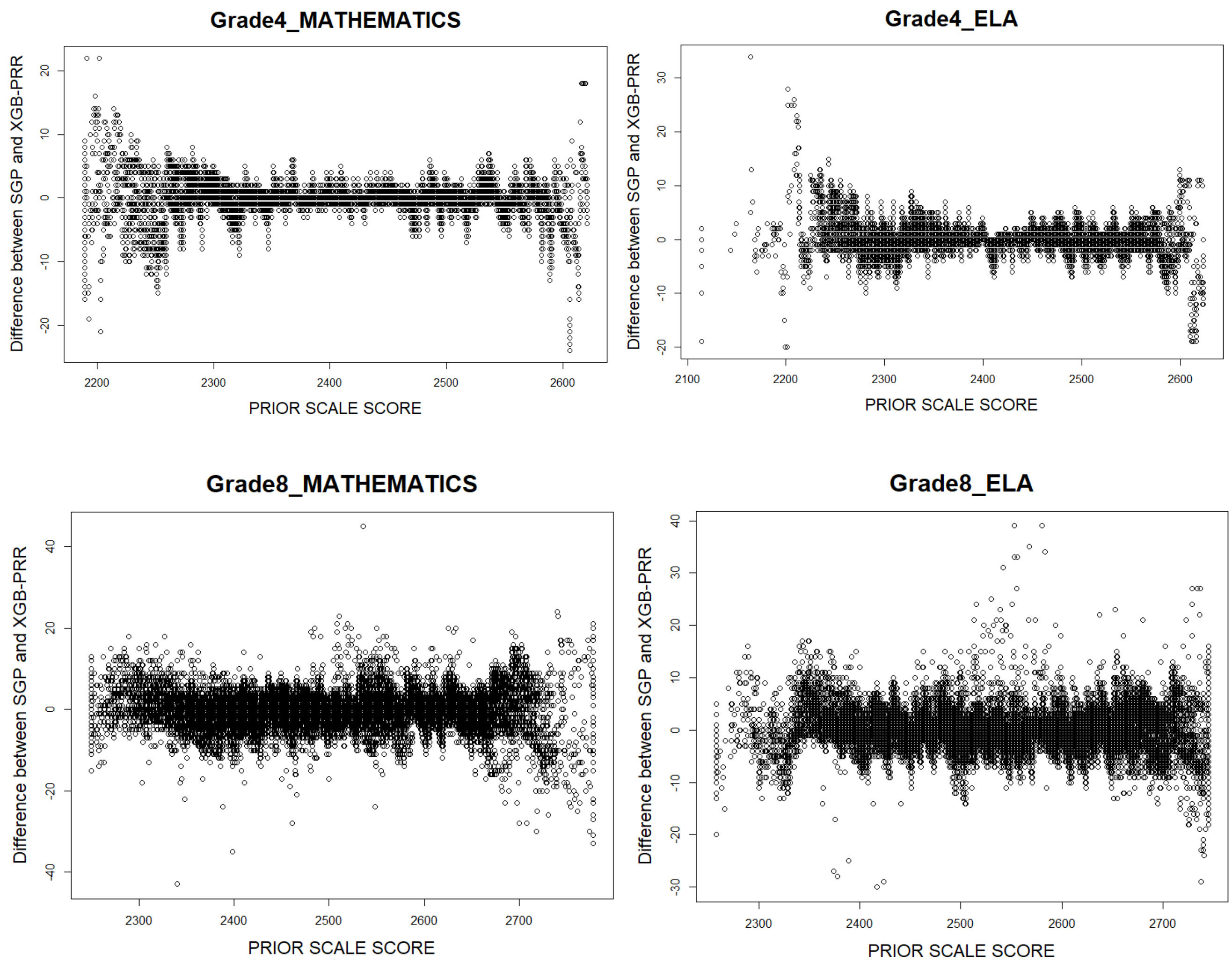

Figure 2. The Difference between two Growth Measures (SGP-XGBPRR) across Prior-year Scale Scores

Conclusion

The practical purpose of this study is to see if XGB PRR could be a feasible alternative statistical framework to quantile regression SGP. The results of this paper indicate that an XGB based model of ranking student growth can, at a minimum, replicate ranks produced by quantile regression based SGP. However, the XGB model has additional statistical properties that may make it preferable, such as being able to model more input features to achieve better predictive accuracies. Additionally, the XGB framework is easy to operationalize, is robust to missing data, and is relatively easy to interpret and analyze.

To establish the XGB PRR as a useful and viable alternative will take additional research, but given how successfully the XGB approach has been applied to many other big data prediction tasks, this line of research appears to be quite promising. There are numerous avenues for future exploration to utilize the expressive and robust properties of the XGB decision tree methodology for prediction. Additionally, other prediction problems in educational statistics, such as making useful forecasts of other results besides growth measures, may also be addressed by modern statistical frameworks like XGB regression trees. The results presented in this study can contribute to a fuller understanding of how modern statistical methods can solve or improve on problems of prediction in large scale measurement.

References

Benjamin, A.S., Fernandes, H.L., Tomlinson, T., Ramkumar, P., VerSteeg, C., Miller, L., & Kording, K.P. (2018). Modern machine learning far outperforms GLMs at predicting spikes. *Frontiers in Computational Neuroscience, 12 (56), 1-13.*

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York, NY: Taylor & Francis.

Betebenner, D. W. (2018). SGP: Student growth percentile and percentile growth Trajectories [R package version 1.8-0.0].

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.

Chen, T. & Guestrin, C. (2016). *XGB: A Scalable Tree Boosting System.* Paper presented in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco.

Lockwood, J. R. & Castellano, K. E. (2015). Alternative statistical frameworks for student growth percentile estimation, *Statistics and Public Policy*, 2:1, 1-9