The SNU Journal of Education Research
December 2017, Vol.26, No.4, pp.43-61.

TIMSS 2015 Korean Student, Teacher, and School Predictor Exploration and Identification via Random Forests*

Yoo, Jin Eun**

Rho, Minjeong

Korea National University of Education

ARTICLE INFO

Article history:
Received Dec 4 2017
Revised Dec 22 2017
Accepted Dec 27 2017

Keywords:
Random forests, decision trees, machine learning, large-scale data, TIMSS, mathematics achievement

ABSTRACT

Previous TIMSS studies have employed conventional statistical methods, focusing on selected few indicators. The purpose of this study was to explore and identify important variables to predict students' mathematics achievement, utilizing as many student, teacher, and school variables as possible via random forests, a popular machine learning technique. TIMSS 2015 Korean 8th graders' student, teacher, and school datasets were merged to extract important predictors for students' mathematics achievement. The prediction accuracy, sensitivity, and specificity of the model were 78%, 83%, and 73%, respectively. Among 413 TIMSS variables explored, variables identified as having the highest variable importance were all student variables, consistent with previous research. Scientific importance of the study was discussed as well as further research topics.

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A2A1A01025538).

** Corresponding author: jeyoo@knue.ac.kr

I. Introduction

Decision trees are known for capabilities to handle non-linear effects as well as higher-order interactions, but at the same time are infamous for instability or high variance. The random forests method, consisting of multiple trees, solves this instability problem of trees but still keeps the merits of the tree method by maximizing randomization. As an ensemble method, random forests yield models of low bias and reduced variance, combining the results of fully-grown trees obtained from bootstrapped samples (Yoo, 2015).

With the advent of big data era, random forests have been particularly popular with so-called 'large p, small n' problems, where there are more predictors than observations (Strobl et al., 2007), as well as with collinear data (Gregorutti et al., 2017). Although random forests have been actively investigated in the fields of study including statistics and engineering, there has been little research in education. Especially with educational large-scale data such as TIMSS (Trends in International Mathematics and Science Study), random forests can provide a breakthrough to the current literature, as random forests can handle hundreds of variables in one model and capture non-linear, higher-order interactions of them.

Previous TIMSS studies have utilized only a handful of variables, selected based on theory or previous research or a mix of the two, and have handled mostly main effects with conventional statistical methods such as SEM (Structural Equation Model) or HLM (Hierarchical Linear Model). This practice partly relates to the conventional methods' difficulty to solve nonconvergence and/or overfitting problems resulting from putting hundreds of variables in one model (Yoo, 2017a). Aforementioned, random forests can identify important variables assuaging such problems, and can provide researchers with insights from a machine learning perspective.

This study explored a total of 413 TIMSS 2015 student, teacher, and school variables to predict Korean 8th graders' mathematics achievement, and successfully identified 17 variables of highest importance, using random forests. Notably, item-parceling convention was unnecessary with random forests, and missing data resulting from non-applicable or non-administered responses were properly handled with a built-in random forests imputation technique.

Specifically, non-applicable or non-administered (NA) responses have plagued TIMSS, frequently with Likert-type items. Unlike conventional research which treated Likert-type items as continuous and deleted their NA responses, this study handled them as nominal and appropriately conducted imputation to minimize data loss. Relatedly, this study did not adopt the item-parceling convention of averaging a set of variables for scale development. Item-parceling is unnecessary in machine learning, primarily for data reduction purposes to prevent nonconvergence or overfitting. Moreover, item-parceling for

scale development requires the inconvenient unidimensionality assumption and consequently sufficient levels of internal consistency, which is not always satisfied in practice (Yoo, 2017a).

In summary, this study did not use the item-parceling convention, properly handled NA responses, and successfully identified important student, teacher, and school predictors for mathematics achievement, using hundreds of variables TIMSS provides. Lastly, by employing random forests, this study considered non-linear, higher-order interactions, which have been largely ignored in educational research using parametric statistics.

II. LITERUTURE REVIEW

Previous TIMSS research has found that student-level predictors have more influence to students' academic achievement than school-level predictors (Kim, 2013; Park Kim, Oh, Chung, & Kim, 2014; Park, 2008; Lee, Park, & Huh, 2012; Chung, Lee, & Kim, 2014). Particularly, students' gender (Kim, 2013; Park et al., 2014), parents' educational level (Kim, 2013; Chung et al., 2014; Lee at al., 2012; Park et al., 2014; Yoo, 2017b), both students' and parents' educational aspiration (Chung et al., 2014; Lee et al., 2012; Park, 2008; Park, 2014), subject-specific interests (Chung et al., 2014; Kim, 2013; Park, 2008; Park et al., 2014; Yoo, 2017b), amount of books at home (Kim, 2013; Lee et al., 2012; Park et al., 2013; Yoo, 2017b), and time spent on homework (Kim, 2013; Lee et al., 2012; Yoo, 2017b) are known to be significant student-level predictors.

Since the late 20's century, school accountability has been a grave issue in Korea (Chung et al., 2014), and therefore empirical research is a necessity to provide policy-makers with evidence-based research results (Chung et al., 2014; Kim & Kim, 2012). Nevertheless, there have not been many Korean TIMSS studies to examine teacher and/or school predictors for students' academic achievement. Out of seven Korean TIMSS studies since 2000, only three and two studies examined teacher-level (Chung et al., 2014; Lee & Chung, 2011; Park, 2008) and school-level predictors (Kim, 2013; Park et al., 2014), respectively.

Specifically, teachers' background variables including age, gender, years of experience, degree, and major as well as teachers' participation in professional development and perception toward their school were examined as teacher-level predictors to students' achievement (Chung et al., 2014; Lee & Chung, 2011; Park, 2008). Variables relating to school location, school climate, and school resources were also attempted to pinpoint school-level effect on students' achievement (Kim, 2013; Park et al., 2014). However, these teacher and school predictors investigated in the previous research

generally showed statistically insignificant or quite limited effect, if any, compared to student predictors.

Consequently, most of the previous TIMSS research reviewed in this study articulated the need to sort out important teacher and/or school predictors for students' achievement (Chung et al., 2014; Kim, 2013; Park, 2008; Park et al., 2014). Following the suggestion of the previous research, this study utilized all possible student, teacher, and school predictors that TIMSS provides in a single model for the purpose of finding important predictors for students' mathematics achievement, using a machine learning technique, random forests.

Ⅲ. TREE-BASED MACHINE LEARNING METHODS

A. DECISION TREES

The decision tree method has been one of the most popular techniques in classification and regression, particularly famous for its intuitive, inverted tree-like graphical representation of the results. As a nonparametric method, a tree starts with the root node, where the first split takes place, and recursively partitions observations in predictor space until predetermined criteria such as maximum tree size are reached. Parent nodes are immediately followed by daughter nodes, and each branch of the tree has terminal nodes where the last split occurs.

Interestingly, the tree methods have been continually reinvented in various contexts in the names of AID (Automatic Interaction Detection; Morgan & Sonquist, 1963), THAID (Theta AID; Morgan & Messenger, 1980), CHAID (CHi-squared Automatic Interaction Detection; Kass, 1980), and CART (Classification And Regression Trees; Breiman, Friedman, Olshen, & Stone, 1984) as well as C4.5 (Quinlan, 1993) and later C5.0 (Quinlan, 1996). Notably, the earlier methods such as AID and THAID were developed in social sciences survey research, the primary motivation of which was to capture interactions among predictors, complementing the conventional linear, additive statistical models (Ritschard, 2013). On top of this, the later methods such as CART and C5.0 explicitly aim at classification and prediction by maximizing homogeneity within each subgroup via impurity measures such as Gini index and entropy.

B. RANDOM FORESTS

1. Overview

In spite of its popularity in applied settings, the decision tree method is infamous for instability, as a small change in data may result in substantially different trees. This was the original motivation of Breiman, so-called the father of trees, who also invented bagging (*b*ootstrap *agg*regat*ing*; 1996) and later random forests (2001). Specifically, the variance of random forests gets more reduced with decreasing pairwise tree correlation or increasing number of trees (Breiman, 2001; Cafri, 2013). In the attempt to get smaller pairwise tree correlation, random forests randomly select predictors at each split, which ensures more diverse trees so that important predictors are not missed out. This number of predictors considered at each split serves as the only tuning parameter in random forests. Breiman (2001) suggests the square root of the total number of predictors for categorical response variable, which is the default in most statistical software programs including R's randomForest.

Another characteristic of random forests relates to bootstrapping. Approximately two-thirds of the original sample is retained in the bootstrapped samples, which serves as the training (or learning) data. The remaining one-third is called OOB (out-of-bag) data, and can automatically serve as the test (or new, independent) data. In machine learning, training data is used for model building, and test data validates that model built from the training data. For continuous and categorical response variables, averaging and majority vote is applied, respectively.

2. Variable importance measure and partial dependence plots

Unlike trees, random forests do not provide a straightforward, intuitive graph, as each bootstrapped sample may result in a different tree. To alleviate this challenging task of interpreting random forests results, variable importance is obtained, using the OOB data (Yoo, 2015). More specifically, the prediction accuracy of each tree is calculated. Next, each predictor is permuted to decouple its relation with the response variable, and the prediction accuracy is again calculated. The differences of the before-after permutation prediction accuracy are averaged over the trees, and divided by the standard error (Breiman, 2001). This serves as the variable importance measure.

Partial dependence plots also can be useful to interpret the random forests results (Yoo, 2015). A partial dependence plot of an explanatory variable (or a predictor) uses the predicted mean probability of the explanatory variable to the response variable after deleting the effects of the other explanatory variables (Hastie et al., 2009). The X-axis of

partial dependence plots is the range of the explanatory variable. The Y-axis, the partial dependence, is calculated as $\frac{logit(\pi)}{2}$, if the response variable is binary (Cutler et al., 2007).

IV. Methods

A. Data merging

For student, teacher, and school data merging, this study used the IDB Analyzer (version 4.0) provided by IEA. Out of the initial 5,547 entries with 725 variables after merging, 238 entries (or students) had two matching teachers. These duplicate entries were deleted, keeping the first teacher of each duplicate student, and thus a total of 5,309 students remained after merging.

B. Response variable

TIMSS provides 5 PVs (plausible values) for students' academic achievement. The PVs are then grouped in 5 levels: 1 (Below Low), 2 (Low), 3 (Intermediate), 4 (High), and 5 (Advanced), and these classified variables are named as categorical benchmark variables. A majority vote was employed to create a single class for each student's math achievement, using the five categorical benchmark variables. For instance, if the benchmark variables of a student (BSMIBM01 to BSMIBM05) were 5, 4, 4, 5, 5, then the student's class was coded as 5. Among the 5,309 Korean 8th graders after merging using the IDB Analyzer (version 4.0), 14 of them (0.2%) had ties (Table 1). A total of 5,295 students after excluding the ties served as the final sample of this study.

<Table 1> Majority Vote Result with TIMSS 2015 Korean 8th Grader's Math

Level	**1**	**2**	**3**	**4**	**5**	**Total**
Observations	63	324	912	1,707	2,289	5,295
(percent)	(1.2%)	(6.1%)	(17.2%)	(32.2%)	(43.1%)	(100%)

As the proportions of the levels were highly unbalanced, this study collapsed the first four levels. Whether the student reached the 'High' level (Level 5; coded as '1') or not (Levels 1, 2, 3, 4; coded as '0') served as the response variable of the study. Each group consisted of 2,289 and 3,006 students, respectively.

C. Explanatory variables

Originally, the merged dataset had 725 variables, and a total of 312 variables were deleted as the following. First, thirty-two variables relating to IDs (e.g., population ID, school ID, etc.), weights (e.g., total student weight, etc.), and file maintenance (e.g., date of testing, file creation date, etc.) were deleted from the explanatory variable pool. Second, 136 variables of 100% missingness (e.g., BSBB21, BSBB22A, BSDBWKHB, BSDPWKHP, etc.) were removed. Third, numerical scale score variables of categorical indices for constructs such as 'home resources for learning' (e.g., BTBGEAS, BSBGHER, BCBGMRS, etc.) were removed, as the numerical scale scores and categorical indices convey basically the same information. Fourth, fifteen duplicate variables after merging were deleted (e.g., BSBS37A, BSBM38BA, BSBS39AB, etc.). Lastly, ninety PVs and benchmark variables were ineligible to be explanatory variables, as the response variable was directly related to these variables, and inclusion of them would dominate the random forests model without useful information.

As results, the cleaned dataset had 413 variables of 5,295 students, comprising 147 student, 175 teacher, and 91 school variables. Among them, only 14 variables were continuous variables the responses of which were hours and numbers (e.g., BTBG01, BTBM16, BCBG07A, BTDM21NU) and the remaining 399 variables were Likert-type scaled. In this study, the Likert-type explanatory variables were analyzed as nominal to include 'not applicable' or 'not administered' responses in the analysis. Previous studies used averages of sums of Likert-type variables. More explanation is given in the next section.

D. Handling Missing Data

Missing data plague TIMSS analyses. This is particularly due to the fact that tens of countries with various educational curriculum have participated in TIMSS and it is not feasible to invent 'one size fits all' kind of questionnaire items for all the participating countries worldwide. TIMSS differentiates this kind of missingness caused by different educational curriculum from the missingness caused by participants' carelessness. The former type of missingness is marked as 'logically not applicable' or 'not administered' (NA) in TIMSS. This study kept these responses of categorical variables as another group of response. Therefore, the categorical variables were treated nominal. If the variables were treated as continuous, the value of NA responses was difficult to determine. Finally, the latter type of missingness caused by carelessness, omitted or invalid, was deleted throughout the data set.

As a result, the 413 explanatory variables had an average rate of 0.1% missingness and a maximum of 7%. This study utilized the built-in 'rfImpute' function in the randomForest package for missing data imputation, which successfully kept all the 5,295 observations with 413 variables. The rfImpute starts replacing missing values with median or mode values, depending on the variable type, and updates them using the proximity matrix obtained from random forest as a weight matrix (Breiman, 2003). Specifically, the $N \times N$ elements (N is the number of observations) in the proximity matrix are increased by one, if two cases end up in the same terminal node in a tree. The normalized elements over the trees in the random forest model work as weights, with which the random forest model is fit iteratively. In this study, 5 iterations with 100 trees were used with the rfImpute.

E. Evaluation Criteria

All the programs were written in R 3.3.3. Particularly, the randomForest package was used as well as caTools and caret packages. Detailed instructions on random forests can be found at Breiman & Cutler (n. d.). The evaluation criteria were prediction accuracy, specificity, and sensitivity. Accuracy indicates the correctly classified rate in the predicted and actual counts. Specificity calculates the rate of correctly classified rate in the actual counts of '0,' while sensitivity calculates the correct rate in the actual counts of '1.' Besides, variable importance measure and partial dependent plots were provided for further interpretation of the results.

V. Results

A total of 500 bootstrap samples were generated. The number of variables tried at each split was 21. This was the square root of the number of predictors in the model, following Breiman (2001). Figure 1 shows that the error rates of random forest converged fewer than 100 replications. OOB samples served as test data in random forests. The accuracy, sensitivity, and specificity of the OOB samples were 78%, 83%, and 73%, respectively. In other words, random forests yielded a quite stable model of prediction accuracy around 80%.

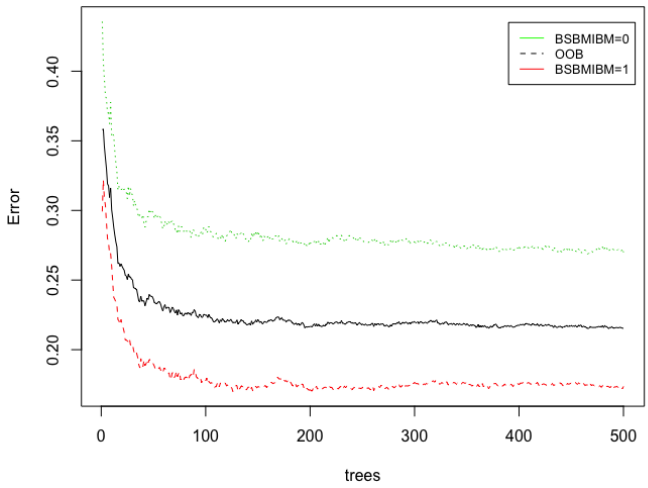


Figure 1. Error rates with 500 replications

This study used a total of 413 TIMSS variables (147 students, 175 teachers, and 91 school variables) to predict Korean students' mathematics performance. Out of 413 variables, 17 had variable importance higher than 20, and all of them were student variables (see Table 2 and Figure 2). Among the 17 items, 9 of them measured students' mathematics self-efficacy, confidence, and interest, 2 items were about mathematics extra lessons, and 3 items were about science self-efficacy, confidence, and time spent on science homework. The other 3 items measured educational aspiration, amount of books at home, and father's educational level.

<Table 2> Top 17 Variables of Importance Higher than 20

Order	**Variable**	**Variable Label**	**MeanDcreaseGini**
1	BSBM19A	MATH\AGREE\USUALLY DO WELL IN MATH	80.309
2	BSBM19B	MATH\AGREE\MATHEMATICS IS MORE DIFFICULT	73.461
3	BSBM26AA	MATH\EXTRA LESSONS LAST 12 MONTH\MATHEMATICS	56.139
4	BSDGSCM	Student Confident in Mathematics/IDX	55.696
5	BSBM26BA	MATH\EXTRA LESSONS HOW MANY MONTH\MATHEMATICS	52.474
6	BSBM19C	MATH\AGREE\MATHEMATICS NOT MY STRENGTH	50.448

7	BSBM19F	MATH\AGREE\GOOD AT WORKING OUT PROBLEMS	44.930
8	BSBS23A	SCI\AGREE\USUALLY DO WELL IN SCIENCE	36.439
9	BSBG08	GEN\HOW FAR IN EDU DO YOU EXPECT TO GO	34.808
10	BSBM19D	MATH\AGREE\LEARN QUICKLY IN MATHEMATICS	31.061
11	BSDGSCS	Student Confident in Sciences/IDX	30.104
12	BSBM19G	MATH\AGREE\I AM GOOD AT MATHEMATICS	29.969
13	BSBM19H	MATH\AGREE\MATHEMATICS HARDER FOR ME	26.120
14	BSBM17G	MATH\AGREE\LIKE MATH PROBLEMS	25.126
15	BSBG04	GEN\AMOUNT OF BOOKS IN YOUR HOME	23.050
16	BSBG07B	GEN\HIGHEST LVL OF EDU OF FATHER	22.226
17	BSBS25BB	SCI\HOW MANY MINUTES SPENT ON HOMEWORK/SCIENCE	21.129

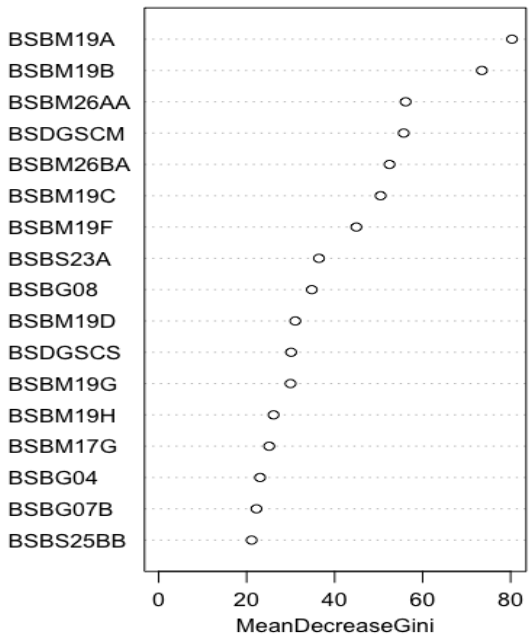


Figure 2. Variable importance of top 17 variables

Easing the variable importance criterion to importance higher than 10 resulted in 53 variables, but only three teacher-level (BTBG12, BTBG01, BTBM25) and three school-

level (BCDG07HY, BCBG10, BCBG19) variables were included among the 53 selected variables (Appendix A). The additionally identified 36 variables of importance between 10 and 20 are presented in Appendix A. To summarize, the highest contributing variables turned out to be student variables, while teacher and school variables had relatively limited effect on students' achievement, consistent with previous research.

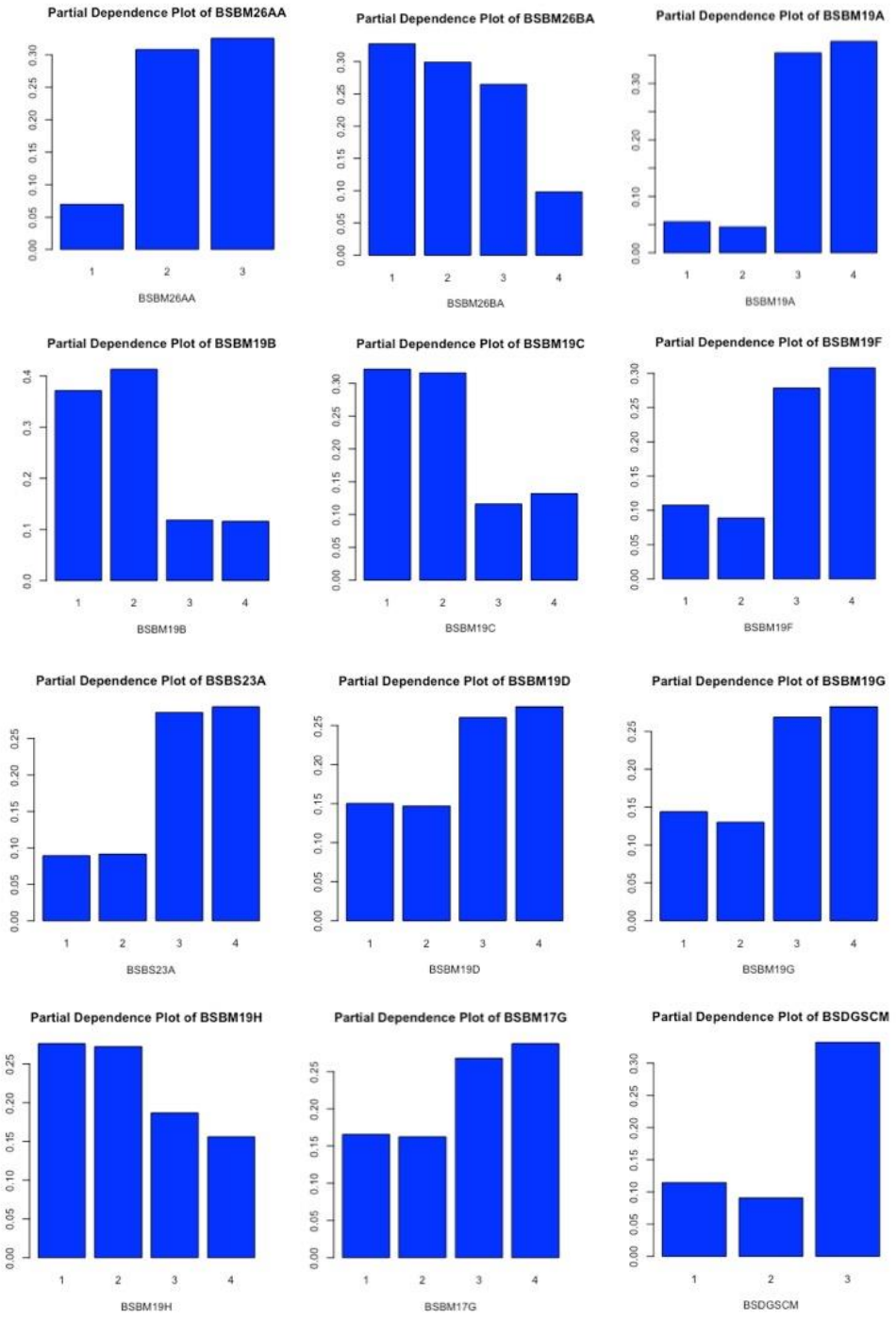
Partial dependence plots of the 17 variables of highest importance are presented in Figure 3. Figure 3 intentionally grouped variables of similar contents for comparison purposes. Most of the variables were measured on ordinal (or Likert-type) scales. Among the 17 plots, the first fifteen plots had positive and negative slopes, nine and six, respectively, and the last two plots showed non-linear patterns. The positive and negative slopes of Likert-type items simply relate to the fact that this study did not employ reverse-coding, as Likert-type scales were treated as nominal.

Specifically, the first two items were about students' extra lessons or tutoring the last 12 months. Students who attended extra lessons or tutoring to excel in class (coded as 1) showed better performance than those who did not attend extra lessons or tutoring (coded as 3) or did to keep up in class (coded as 2; BSBM26AA). Students who did more than 8 months' extra lessons or tutoring (coded as 4) had a higher chance of reaching the Advanced level than others, but there was no markedly visible difference when the extra lessons or tutoring lasted 8 months or fewer (coded as 1, 2, 3; BSBM26BA).

Consistent with previous research, students who answered more positively to mathematics self-efficacy and interest items (BSBM19A, BSBM19B, BSBM19C, BSBM19D, BSBM19F, BSBM19G, BSBM19H, BSBM17G) had a higher chance of achieving the Advanced level. Interestingly, one science self-efficacy item (BSBS23A) showed a similar pattern as well as the mathematics and science self-confidence indices (BSDGSCM, BSDGSCS). Also in line with the previous research, the amount of books at home was positively related to students' mathematics achievement (BSBG04).

Students whose educational aspiration was equivalent to or beyond bachelor's degree (coded as 5 or 6; BSBG08) had a higher chance of achieving the Advanced level than the others. Likewise, students who answered "I don't know" (coded as 8) or high school or below (coded as 1 to 5) to father's completed educational level had a lower chance of achieving Advanced than those who answered bachelor's or beyond (coded as 6 or 7; BSBG07B). That is to say, the "I don't know" response to father's completed educational level had similar effect on students' mathematics achievement to responses, high school or below.

Interestingly, time spent on science homework was another important predictor to students' mathematics achievement (BSBS25BB). Students who spent more than 90 minutes (coded as 6) or those who answered their science teachers never gave homework (coded as 1) had a lower chance of reaching the Advanced level than others.


Partial Dependence Plot of BSBM26AA
BSBM26AA
Partial Dependence Plot of BSBM26BA
BSBM26BA
Partial Dependence Plot of BSBM19A
BSBM19A
Partial Dependence Plot of BSBM19B
BSBM19B
Partial Dependence Plot of BSBM19C
BSBM19C
Partial Dependence Plot of BSBM19F
BSBM19F
Partial Dependence Plot of BSBS23A
BSBS23A
Partial Dependence Plot of BSBM19D
BSBM19D
Partial Dependence Plot of BSBM19G
BSBM19G
Partial Dependence Plot of BSBM19H
BSBM19H
Partial Dependence Plot of BSBM17G
BSBM17G
Partial Dependence Plot of BSDGSCM
BSDGSCM

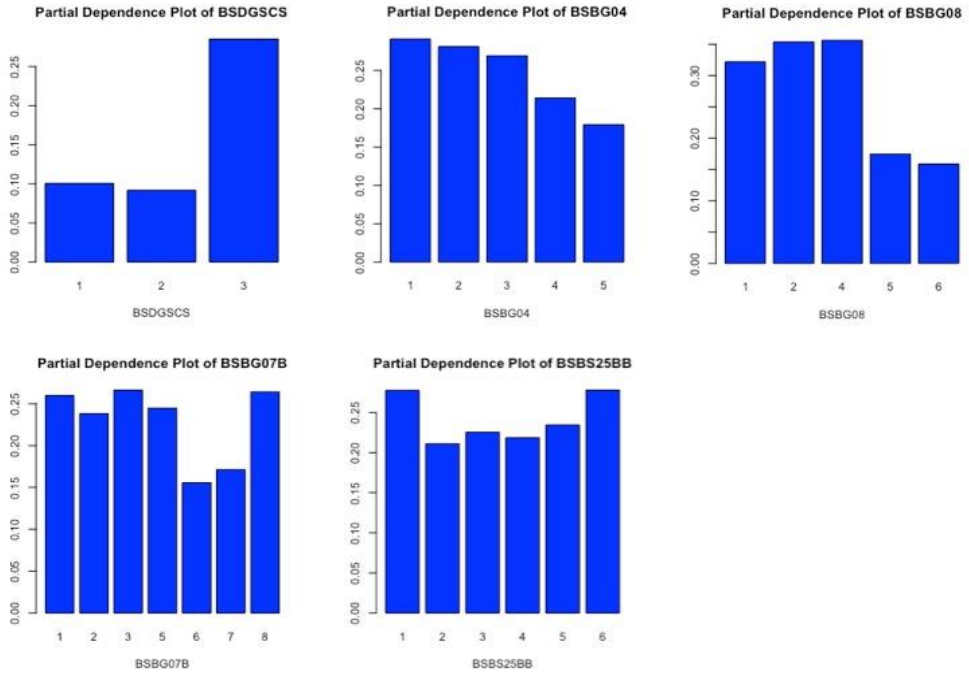


Figure 3. Partial dependence plots of top 17 variables of highest importance

VI. Discussion

Using random forests, this study explored a total of 413 TIMSS 2015 student, teacher, and school variables to predict Korean 8th graders' mathematics achievement, and successfully identified 17 variables of the highest importance. All the 17 variables were student variables, in line with previous research. Easing the variable importance criterion resulted in more variables, but only a handful of teacher and school variables were additionally identified. This was in accordance with previous research that student variables exerted more influence over students' achievement than teacher and school variables. Although this was not surprising at all, studies need to be furthered on to explore and identify teacher and school variables, particularly in the age of increasing emphasis on teacher and school accountability. Teachers and schools are held more responsible for their students' academic achievement than ever, and educational researchers have to support teachers and schools with evidence-based research. It was the basic intention of this empirical study.

Apart from this, the study has several scientifically important features, as one of the first random forests studies with educational large-scale data. First, this study explored as many predictors as possible in one statistical model via random forests. Second, a proper imputation technique was applied to the cleaned dataset. A proximity matrix was obtained and utilized for missing data imputation, as suggested by Breiman (2003) and Breiman & Cutler (n. d.), minimizing data loss. Third, conventional research has used averages (or sums) of Likert-type variables, treating them as continuous. This study treated them as nominal, and thus the 'not applicable' or 'not administered' (NA) responses of them were properly imputed via the proximity matrix. Fourth and relatedly, this study did not use item-parceling, as there was no need for data reduction, particularly to prevent nonconvergence or overfitting. The conventional practice of averaging a set of variables holds, if the set of variables meets the unidimensionality assumption and shows sufficient level of internal consistency, which is not always satisfied in practice (Yoo, 2017a). Fifth and lastly, this study considered non-linear, higher-order interactions in predicting students' mathematics achievement via random forests, an ensemble of decision trees.

Interestingly, this study with random forests newly identified three science variables such as science self-efficacy, confidence, and time spent on science homework among the 17 variables of the highest importance. These science variables were rarely investigated as predictor candidates for mathematics achievement, as previous research with conventional methods has been confined to the same discipline. Notably, one TIMSS 2011 study by Yoo (2017b) using LASSO, another popular machine learning technique, also identified science self-efficacy and homework as important variables to predict Korean 8th graders' mathematics achievement. At least, the TIMSS studies with machine learning techniques yielded similar results that mathematics and science have something in common. As TIMSS provides mathematics and science variables altogether, further research is warranted to investigate the relationship of these two disciplines as well as their impact on students' academic achievement, which will shed light on the ongoing discussion about cross-curricular integration.

Random forests are superior to decision trees in terms of bias, variance, and prediction accuracy, and at the same time retain most of the merits of trees such as the capabilities to detect non-linear and/or higher-order effects as well as to handle 'large p, small n' problems. Thus, random forests have been actively investigated in the fields including statistics and engineering, but they have not received enough attention they deserve, particularly in educational research. With large-scale data such as TIMSS, conventional statistical methods may fail to converge with only main effects. A logistic regression model of the 413 predictors' main effects simply did not converge, for instance. In contrast, random forests yielded a model without convergence problems, and its prediction accuracy, specificity, and sensitivity were close to 80%. The set of identified predictors is in itself of

importance, casting implications to practitioners who have been in search of important predictors for students' higher mathematics achievement. At the same time, the identified predictors can be utilized in subsequent analyses such as HLM and SEM, and the models should be built in a way to keep the nonlinear, higher-order interaction effects of the random forests.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32.

Breiman, L. (2003). *Manual for setting up, using, and understanding random forest V 4.0.* Retrieved from http://oz.berkeley.edu/users/breiman/Using_ random_forests_v4.0.pdf

Breiman, L., & Cutler, A. (n. d.). *Random forests*. Retrieved from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Breiman, L., Freidman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth.

Cafri, G. (2013). *Predictors of students switching out of STEM majors.* Unpublished Master's Thesis. San Diego State University: San Diego, CA.

Choe, S. H., Park, S., & Hwang, H. J. (2014). Analysis of the current situation of Affective Characteristics of Korean Students Based on the Results of PISA and TIMSS. *Journal of the Korean School Mathematics Society, 17*, 23-43.

Chung, J. Y., Lee, H., & Kim, S. (2014). A hierarchical analysis of the factors influencing on student achievement - Using the teacher and student factors of TIMSS 2011. *The Journal of Korean Teacher Education, 31*(2), 53-75.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess. K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology, 88*, 2783-2792.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing, 27*, 659-678.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*, 119-127.

Kim, S. (2013). The compositional effects of middle school in Korea, English, Japan, Singapore, and USA: using TIMSS 2007. *Education Research Institute, 14*(2), 51-74.

Kim, J. Y. (2008). Differences in test scores among Korean middle schools: Evidence from the TIMSS data. *Korean Journal of Public Finance, 1*(13), 53-77.

Kim, S. H., & Kim, S. (2012). Analysis of trends of mathematics Education in Korean classes based on TIMSS. *The Korean Journal for History of Mathematics, 25*(4), 139-155.

Lee, A. (2017). An analysis of the effect of class size on academic achievement. *The Journal of Economics and Finance of Education, 26, 1-26.*

Lee, H. S., & Chung, J. Y. (2011). An analysis of the influence of teachers' traits on student achievement-focusing on teachers' efforts to enhance professionality in TIMSS 2007. *The Journal of Korean Teacher Education, 28*, 243-266.

Lee, H. J., Park, C. G., & Huh, N. (2012). Effect of contextual variables on mathematics achievement - Based on analysis of TIMSS 2007 using path analysis. *Journal of the Korean School Mathematics Society, 15*(3), 585-603.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association, 58*, 415-434.

Morgan, J. N., & Messenger, R. C. (1973). *THAID a sequential analysis program for analysis of nominal scale dependent variables.* Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor.

Park, C. (2008). Effect of educational contextual variables on mathematics achievement of the different ability groups in Korea. *Journal of Educational Evaluation, 21*(3), 23-41.

Park, S. Y., Kim, J. Y., Oh, E. B., Chung, D. B., & Kim, S. H. (2014). The effects of school principal's accountability mechanism on school outcomes with TIMSS 2011. *Korean Journal of Educational Administration, 32*(1), 159-185.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research, 4*, 77-90.

Ritschard, G. (2013). CHAID and earlier supervised tree methods. In J.J. McArdle & G. Ritschard (eds.), *Contemporary issues in exploratory data mining in behavioral sciences* (pp. 48-74). Routledge: New York.

Strobl, C., Boulesteix, A, Zeiles, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*, 25. http://doi.org/10.1186/1471-2105-8-25.

Yoo, J. E. (2015). Random forests, an alternative data mining technique to decision tree. *Journal of Educational Evaluation, 28*, 427-448.

Yoo, J. E. (2017a). *TIMSS student and teacher variables through machine learning: Focusing on Korean fourth graders' mathematics achievement.* Paper presented at American Educational Research Association. San Antonio: TX.

Yoo, J. E. (2017b). TIMSS 2011 predictors relating to Korean 8th graders' mathematics achievement, explored via machine learning. *Korean Journal of Teacher Education, 33*(1), 43-56.

Authors

Jin Eun Yoo, 1st and corresponding author
Korea National University of Education
jeyoo@knue.ac.kr

Minjeong Rho
Korea National University of Education
minjeong019@gmail.com

Appendix A

<Table A1> Variables of Importance between 10 and 20 (continued from Table 2)

Order	Variable	Variable Label	MeanDecreaseGini
18	BSBM17E	MATH\AGREE\LIKE MATHEMATICS	19.823
19	BSBS24I	SCI\AGREE\IMPORTANT TO DO WELL IN SCI CLASS	19.161
20	BSBM20H	MATH\AGREE\PARENTS THINK MATH IMPORTANT	19.016
21	BSBS24G	SCI\AGREE\MORE JOB OPPORTUNITIES	18.753
22	BSBM25BA	MATH\HOW MANY MINUTES SPENT ON HOMEWORK/MATHEMATICS	18.323
23	BSBM19I	MATH\AGREE\MAT MAKES CONFUSED	18.132
24	BSBS23B	SCI\AGREE\SCIENCE IS MORE DIFFICULT	17.827
25	BSBM17I	MATH\AGREE\FAVORITE SUBJECT	17.764
26	BSBG07A	GEN\HIGHEST LVL OF EDU OF MOTHER	17.67
27	BSBS23F	SCI\AGREE\I AM GOOD AT SCIENCE	17.117
28	BSDGEDUP	Parents' Highest Education Level	15.539
29	BSBS23C	SCI\AGREE\SCIENCE NOT MY STRENGTH	15.327
30	BSBS24F	SCI\AGREE\GET AHEAD IN THE WORLD	14.393
31	BSDGHER	Home Educational Resources/IDX	14.123
32	BSBM17C	MATH\AGREE\MATH IS BORING	13.935
33	BSDGSLM	Students Like Learning Mathematics/IDX	13.597
34	BSBM20I	MATH\AGREE\IMPORTANT TO DO WELL IN MATH	12.839
35	BSBM20G	MATH\AGREE\MORE JOB OPPORTUNITIES	12.804
36	**BTBG12**	**GEN\NUMBER OF STUDENTS IN THE CLASS**	**12.669**
37	**BCDG07HY**	**Total Instructional Hours per Year**	**12.498**
38	BSDGSVM	Students Value Mathematics/IDX	12.419
39	BSBS24H	SCI\AGREE\PARENTS THINK SCI IMPORTANT	12.308
40	BSBG12	GEN\HOW OFTEN BREAKFAST ON SCHOOL DAYS	12.265

41	**BTBG01**	**GEN\YEARS BEEN TEACHING**	**11.834**
42	BSBS24C	SCI\AGREE\NEED SCI TO GET INTO <UNI>	11.777
43	BSBS23E	SCI\AGREE\GOOD AT WORKING OUT PROBLEMS	11.773
44	BSBM25AA	MATH\HOW OFTEN TEACHER GIVE YOU HOMEWORK/MATHEMATICS	11.762
45	BSBM20C	MATH\AGREE\NEED MATH TO GET INTO <UNI>	11.582
46	BSBG13C	GEN\HOW OFTEN USE COMPUTER TABLET\OTHER	11.341
47	BSBM17A	MATH\AGREE\ENJOY LEARNING MATHEMATICS	11.109
48	**BCBG10**	**GEN\TOTAL NUMBER COMPUTERS**	**10.981**
49	**BCBG19**	**GEN\YEARS PRINCIPAL ALTOGETHER**	**10.649**
50	BSDGSVS	Students Value Science/IDX	10.530
51	BSBM18B	MATH\AGREE\TEACHER IS EASY TO UNDERSTAND	10.390
52	**BTBM25**	**MATH\<PROF DEVELOPMENT> HOURS**	**10.226**
53	BSBG13A	GEN\HOW OFTEN USE COMPUTER TABLET\HOME	10.195

Note: Teacher and school predictors are in bold.